Speech and Emotion Research

An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis

Dissertation
zur Erlangung des Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von

Marc Schröder
aus Bad Kreuznach

Dekan: Prof. Dr. Klaus Martin Girardet
Berichterstatter: Prof. Dr. William Barry und Prof. Dr. Peter Roach
Tag der letzten Prüfungsleistung: 24. 10. 2003

Zusammenfassung

Die vorliegende Arbeit besteht aus zwei Teilen. Im ersten Teil wird der Versuch einer systematischen Aufstellung derjenigen Konzepte gemacht, die für das Forschungsgebiet Sprache und Emotionen relevant sind. Der zweite Teil widmet sich der Entwicklung und Untersuchung eines neuen Ansatzes zum Ausdruck von Emotionen in der Sprachsynthese mit Hilfe von Emotionsdimensionen.

Wie es in einem kürzlich abgehaltenen ISCA-Workshop (Cowie et al., 2000b) deutlich geworden ist, tut ein konzeptueller Unterbau für die Forschung an Sprache und Emotionen not. Aufgrund der Schwierigkeiten, das Konzept Emotionen zu fassen und zu beschreiben, existieren eine Reihe von Forschungstraditionen, die auf unterschiedlichen Grundannahmen basieren. Dementsprechend werden gleiche Ideen verschieden benannt und gleiche Wörter für Unterschiedliches gebraucht. Aufbauend auf einer Sichtung der existierenden Literatur zu dem Thema bemüht sich der **erste Teil** der Dissertation um eine klärende Darstellung und Einordnung der relevanten Konzepte.

Kapitel 1 beginnt mit einer Definition des verwendeten Vokabulars: *voll entwickelte Emotionen*, die üblicherweise Gegenstand von Emotionstheorien sind, werden *zugrundeliegenden Emotionen* gegenübergestellt, die als eine Komponente jedes mentalen Zustands zu finden sind. Beide gemeinsam werden als *emotionale Zustände* bezeichnet, und von *mit Emotionen verwandten Zuständen* unterschieden. Anschließend wird eine anerkannte Klassifikation von Theorien über voll entwickelte Emotionen (Cornelius, 1996) zusammengefasst, sowie eine Erläuterung der weiteren definierten Begriffe vorgenommen. Schließlich wird die Erforschung der vorgestellten Arten von Emotionen unterschieden von der Erforschung der Emotionskonzepte von Laien.

Auf diesem Grundverständnis aufbauend stellt **Kapitel 2** eine Reihe von Beschrei-

bungswerkzeugen für Emotionen vor. Bei dem einfachsten und am weitesten verbreiteten Beschreibungswerkzeug, den Emotionskategorien, werden evolutionär motivierte Basisemotionen unterschieden von hierarchisch geordneten Emotionskategorien und von Listen wichtiger Emotionswörter für die Beschreibung von Alltagssituationen. Es wird die Problematik der genauen Definition einer Emotionskategorie diskutiert.

Eine Reihe weiterer Beschreibungsmethoden wird vorgestellt: Prototypen-Beschreibungen, physiologiebasierte Beschreibungen, Beschreibungen basierend auf kognitiven Bewertungen (*appraisal*), und Zirkumplex-Modelle. Schließlich werden, in Vorbereitung auf den zweiten Teil, Emotionsdimensionen genauer behandelt, einschließlich eines historischen Überblicks und einer kritischen Diskussion des Beschreibungsgegenstandes von Emotionsdimensionen. Die Problematik der Subjektivität bei der Benennung von Emotionsdimensionen wird diskutiert, und es wird ein Sprachgebrauch festgelegt: Die drei in vielen verschiedenen Studien als grundlegend erkannten Emotionsdimensionen werden in dieser Arbeit als Aktivation, Evaluation und Power bezeichnet.

Widmeten sich die ersten Kapitel der Beschreibung der Emotionen an sich, so wendet sich **Kapitel 3** einer Analyse des Kommunikationsvorgangs zu. Anhand des Brunswikschen Linsenmodells in der von Scherer (1978) verwendeten Form werden der Sprecher, das Gesprochene und der Hörer illustriert, und daran anknüpfend *sprecherzentrierte*, *akustisch-auditive* und *hörerzentrierte* Forschungsfragen und Anwendungsgebiete unterschieden. Es wird darauf hingewiesen, dass je nach Forschungsausrichtung unterschiedliche Beschreibungswerkzeuge geeignet sein können.

Die folgenden drei Kapitel widmen sich der Erforschung der drei unterschiedenen Aspekte des Kommunikationsvorgangs. **Kapitel 4** stellt eine Reihe von Quellen emotionalen sprachlichen Materials vor: Darstellungen von Schauspielern, gelesene emotionale Texte, hervorgerufene und natürlich vorkommende Emotionen, sowie synthetisierte Sprache. Wiederum kommt es auf die Forschungsfrage an, welche der Quellen geeignet sind. **Kapitel 5** zeigt die Komplexität der Frage auf, durch welche Stimmparameter Emotionen ausgedrückt werden. Den zumeist erforschten graduellen prosodischen Effekten werden Beispiele von kategorischen prosodischen Effekten gegenübergestellt. Neben diesen beiden Arten des Ausdrucks von Emotionen durch sprachbegleitende prosodische Variationen existieren aber auch lokale Ereignisse, sogenannte Affektlaute oder emotionale Interjektionen. Als eine Möglichkeit der Erklärung

beobachteter Phänomene werden der Frequenzcode und andere evolutionär geprägte Codes vorgestellt. Schließlich wird die Forschung an multimodalem Emotionsausdruck berührt, in dem Sprache gemeinsam mit anderen Kanälen die Emotion kommuniziert. **Kapitel 6** widmet sich Perzeptionstest-Methodologien, und unterscheidet Identifikationstests, Präferenztests, Ähnlichkeitsbewertungen, Lokalisierung auf Emotionsdimensionen, und physiologische Messungen am Hörer.

Als letztes Kapitel des ersten Teils formuliert **Kapitel 7** die Problematik der Untersuchung von Authentizität. Dabei wird Scherers “push–pull” Unterscheidung dargestellt, und es wird argumentiert, dass willentliche Kontrolle ein weiterer Faktor ist, der von dieser Unterscheidung nicht abgedeckt ist. In Hinblick auf experimentelle Handhabbarkeit werden sprecherzentrierte und hörerzentrierte Aspekte der Natürlichkeit unterschieden.

Der **zweite Teil** der Dissertation befasst sich mit der praktischen Erforschung eines neuen Ansatzes zur emotionalen Sprachsynthese, in dem die auszudrückenden emotionalen Zustände mit Hilfe von Emotionsdimensionen modelliert werden.

Kapitel 8 motiviert zunächst das Forschungsvorhaben durch den Wunsch nach mehr Natürlichkeit in der Sprachsynthese. Emotionsdimensionen werden dabei als eine Möglichkeit dargestellt, einen neuen Grad an Flexibilität in die Ausdrucksfähigkeit der synthetischen Stimme zu bringen. Desweiteren wird das Vorhaben sorgfältig entlang der im ersten Teil entwickelten Konzepte positioniert. Es handelt sich um die Erforschung gradueller Emotionen, die mit Hilfe von Emotionsdimensionen beschrieben werden. Die Anwendung Sprachsynthese ist offensichtlich hörerzentriert. Die Aufgabe ist es, einen Zusammenhang zwischen akustischen Parametern und der wahrgenommenen Emotion herzustellen. Zu diesem Zweck wird ein Korpus emotionaler Sprache untersucht, das auf quasi-natürlicher Konversation basiert. Aufgrund technischer Begrenzungen werden nur graduelle prosodische Parameter analysiert. Das resultierende emotionale Sprachsynthesesystem wird mittels eines Präferenztests evaluiert.

Der aktuelle Stand der Forschung im Bereich emotionaler Sprachsynthese wird in **Kapitel 9** dargestellt. Der zugehörige **Anhang A** fasst in tabellarischer Form die in der Literatur beschriebenen Prosodieregeln für den Ausdruck von Emotionen zusammen. Alle existierenden Ansätze haben gemeinsam, dass sie Emotionen mit Hilfe von Emotionskategorien beschreiben.

Die existierende Literatur im Bereich der stimmlichen Korrelate von Emotionsdimensionen wird in **Kapitel 10** vorgestellt. Dabei wird deutlich, dass es klare, wiederkehrende Muster für die Korrelate der Aktivationsdimension gibt, dass die Befunde zu Korrelaten der Evaluations- und der Powerdimension hingegen lückenhaft und widersprüchlich sind.

Vor diesem Hintergrund wird die Korpusanalyse in **Kapitel 11** beschrieben. Das verwendete Datenmaterial, die Belfast-Datenbank natürlicher emotionaler Sprache, sowie die vorhandenen akustischen Analysen und emotionalen Bewertungen, werden vorgestellt. Aus den zur Verfügung stehenden akustischen Parametern werden solche ausgewählt, die für die Sprachsynthese relevant sind, und es werden lineare Regressionsanalysen durchgeführt, in denen jeweils ein akustischer Parameter durch die drei Emotionsdimensionen Aktivation, Evaluation und Power vorhergesagt wird. In einem kleinen Exkurs wird dargelegt, wie aufgrund der sich ergebenden Linearkoeffizienten die akustischen Korrelate von Emotionskategorien vorhergesagt werden können, und wie darüber hinaus akustische Ähnlichkeiten zwischen Emotionskategorien auf emotionale Ähnlichkeiten der Kategorien zurückgeführt werden können.

In Vorbereitung auf die Implementierung der Prosodieregeln wird in **Kapitel 12** zunächst das Sprachsynthesesystem MARY vorgestellt. Eine Besonderheit des Systems ist die intern verwendete, XML-basierte Repräsentationssprache MaryXML, die mittels geeigneter Annotation eine “ferngesteuerte” Kontrolle über die Verarbeitungskomponenten erlaubt. Nach einer detaillierten Beschreibung der einzelnen Module und einer Benutzerschnittstelle für Experten wird dargelegt, warum das System für den Ausdruck von Emotionen besonders geeignet ist. Insbesondere wird dabei auch beschrieben, dass Diphonstimmen zum Einsatz kommen, die im NECA-Projekt speziell zu diesem Zweck erstellt wurden.

Kapitel 13 stellt schließlich die implementierten Prosodieregeln vor (Tabelle 13.1, S. 158). Die dabei modellierten Syntheseparameter werden erläutert, und der gewählte Ansatz für die Implementierung wird beschrieben. **Anhang B** gibt den Quellcode des verwendeten XSLT-Stylesheets wieder. Desweiteren wird “EmoSpeak” vorgestellt, eine graphische Oberfläche, die es dem Benutzer ermöglicht, interaktiv die prosodischen Korrelate der verschiedenen Koordinaten auf den Emotionsdimensionen zu erforschen.

Eine perzeptive Evaluierung des Systems wird in **Kapitel 14** beschrieben. Emotional möglichst eindeutige Situationsbeschreibungen wurden mittels eines geschriebenen Bewertungstests aus einer Liste von 36 Kandidaten (siehe **Anhang C**) ausgewählt. Die für jede Situationsbeschreibung als passend vorhergesagte Prosodie wurde berechnet und mit allen Situationsbeschreibungen, passenden wie unpassenden, kombiniert. Die kombinierten Stimuli wurden Hörern vorgelegt und von diesen in Bezug auf die Frage bewertet, wie gut der Klang der Stimme zum begleitenden Text passte. Die Ergebnisse zeigten, dass in den meisten Fällen solche Stimuli als passend bewertet wurden, in denen die durch Prosodie und Text ausgedrückten Aktivationswerte ähnlich waren, wohingegen Stimuli mit sehr unterschiedlichen Aktivationswerten in Prosodie und Text als unpassend bewertet wurden. Dies kann als eine Bestätigung der Grundannahmen des Ansatzes gewertet werden.

Der Arbeit ist eine 56 Seiten umfassende **annotierte Bibliographie** beigefügt.

Personal Preface

I would never have written this thesis if it weren't for the boring synthesis voices of current speech synthesis systems. Somehow, the frustration of listening to such a voice made me wish I could make the synthetic voice expressive, so that it could express the emotions contained in the text. The long-term goal, I told myself, would be to have a synthetic voice read a fairy tale in a way that you would like listening to it.

Since I first had this motivation, quite a few years have passed, and they have shaped my carreer considerably. After doing undergraduate studies in physics in Saarbrücken, Germany, I went on to study phonetics, first in Saarbrücken, then in Grenoble, France, in order to move towards research in the field of speech synthesis. In my final year of maîtrise studies in Grenoble, my supervisor Véronique Aubergé gave me the opportunity to explore the vast field of vocal emotion expression and emotion theory. She also encouraged me to ask myself intriguing questions about the nature of the mechanisms leading to emotion expression. The experiments we conducted during that time included attempts to distinguish voluntarily controlled from spontaneous emotion expression.

In the following year, I moved back to Saarbrücken, in order to pursue PhD studies, and was given the opportunity to take over the development of the MARY text-to-speech synthesis system. Since then, I have shaped the architecture of the system in the most modular and flexible way I could, so that rules for emotion expression would have access to all the levels of abstraction that might be relevant.

In parallel, I investigated some fundamental technical questions, such as the role of voice quality in vocal emotion expression, and how to describe the relationships between emotions, which I felt varied in their degree of similarity, but were all treated alike in categorical descriptions of emotions. Being able to express extreme anger, fear, sadness,

and joy, I thought, would not be sufficient for the fairy tale reading which was still in the back of my mind, so a descriptive framework was needed for representing shades of emotions. I also continued to think about theoretical aspects such as the role of volition in emotion expression. This last point led me to undertake an investigation of "affect bursts", or emotional interjections, where the complex interaction between evolutionary and cultural, between automatic and controlled factors seemed a bit less opaque than in emotion expression accompanying speech.

An interlude in Belfast, Northern Ireland, again had a strong impact on my way of thinking about emotions. I was given the opportunity to work for seven months with Roddy Cowie and Ellen Douglas-Cowie, who had been building expertise in dealing with non-extreme emotional states, and had done considerable conceptual work on how to characterise and represent these states. Their representation of emotion dimensions, the activation-evaluation space, and the Feeltrace tool used for rating stimuli according to this descriptive framework, gave me the type of representation which I had felt was needed for the more flexible expressive speech synthesis I had in mind. I also had the chance to get access to the Belfast Naturalistic Emotion Database, which currently is probably the largest audio-visual database of spontaneous emotional speech. Going back to Saarbrücken, I now had the means to formulate, and then implement, prosody rules for emotion expression using a descriptive framework allowing for non-extreme emotional states.

The present thesis is the result of this odyssey of mine. The multitude of aspects under which the general topic of speech and emotion can be addressed seemed so complex and often chaotic to me that I decided to put together an overview of these aspects, which has become the first part of this thesis. In it, I have tried to structure, according to my best understanding, the diverse approaches, theories and techniques I encountered. As after five years, I am still a relative newcomer to this research area, it is probable that some people will find some aspects incomplete or suboptimally presented. Still, I hope that some readers might find a bit of orientation and structure through this overview, so that they can more easily take conscious decisions about the frameworks and methods to use in order to address their research questions.

The second part of this thesis is dedicated to emotional speech synthesis, and presents the novel approach I am proposing for expressing gradual emotional states in a more flexible way than has previously been possible. I still consider this state of affairs to

be far from the goal of satisfactory fairy-tale reading from which I have started; but if I have successfully proposed an alternative to the purely categorical approach to emotion representation for speech synthesis, I think that is a start.

I want to thank many people who have been helpful during the research leading to this thesis. First of all Bill Barry, my first examiner, who has given me the freedom to follow the research direction I considered most interesting, and who has always been very helpful and encouraging when I was uncertain about how to proceed. I also want to express my deep gratitude to Roddy Cowie and Ellen Douglas-Cowie, who have been a precious source of inspiration regarding the wealth of possibilities in speech and emotion research, and who have had a great impact on my research approach. I want to thank Véronique Aubergé who accompanied me and helped me find an orientation during my first steps in the domain of scientific research on speech and emotion. Finally, many people have helped me with their comments and interesting discussions – among many others Jürgen Trouvain, Martine Grice, Hans Uszkoreit, Jacques Koreman, Stefan Baumann, Akemi Iida, and Edelle McMahon. To these and the many whom I have not named go my warmest thanks.

Contents

List of Figures

List of Tables

Introduction: Aims and methods

This dissertation consists of two main parts. The first part is an extended review of the literature, with the aim of providing some structure to the complex field of speech and emotion research. The second part proposes a novel approach to the synthesis of emotional speech, by means of emotion dimensions.

Frameworks in speech and emotion research

Speech and emotion is a fascinating but difficult research domain. Systematic research in this field can be traced back at least to the Fairbanks studies at the end of the 1930's (Fairbanks & Pronovost, 1939; Fairbanks & Hoaglin, 1941), and many research and review articles have been published in the meantime. However, most of today's researchers in the domain of speech and emotion would probably agree that the knowledge gained up to today does not nearly cover the wide variety of relevant phenomena, and that many interesting and challenging aspects have yet to be investigated.

At first sight, it may seem clear how to go about studying speech and emotion: Produce speech utterances in which emotions are systematically varied while all other factors are held constant, and measure the acoustic parameters affected by the different emotions as well as the perceptual effects of these acoustic parameters.

A closer look, however, reveals a multitude of aspects relevant for studying speech and emotion, which in their complexity may well seem dispiriting. The complexity starts with the fuzziness of the emotion concept itself. No unified and generally accepted theory of emotions is currently available; instead, multiple approaches stress different interesting aspects, but they cannot easily be integrated and sometimes seem to contradict one another. Even the boundaries of what should count as emotion are not generally agreed upon. The

methods for describing and measuring emotions vary widely.

Attempts to study emotion *expression* face additional problems. Depending on the theoretical and application context, different aspects of the expression process may be put into focus, such as its evolutionary origins and functions, the biological systems influencing the expression, or the social rules triggering or inhibiting emotion expression. The fact that emotion is expressed by a mixture of biological and cognitive, of automatic and consciously controlled factors, motivates questions related to the role of volition and consciousness. The facial, vocal and other signs identified as expressing emotions may serve other communicative or non-communicative purposes as well, which makes their interpretation context-dependent.

While for a given research project, some of these questions may be irrelevant and can safely be left unanswered, many questions will most likely be relevant throughout the research, starting from the formulation of hypotheses worth testing, via the methodology employed for testing them, the collection of data to be analysed, to the analysis methods, and maybe even the conclusions drawn from the results.

Due to the inherent difficulties outlined above, the literature on speech and emotion research is highly fragmented. In the words of Scherer (1986), "there has been neither continuity nor cumulativeness in the area of the vocal communication of emotion" (p. 143). Although a large number of interesting contributions exist, an overview of the choices available in the various sub-aspects of this research field remains a challenge in its own right. The existing reviews of the field concentrate on specific questions such as the link between physiological and acoustic variables (Scherer, 1986), the acoustic realisations of emotion categories in view of speech synthesis (Murray & Arnott, 1993), or the conceptual issues involved when designing an emotion recognition system (Cowie et al., 2001). What can be considered common knowledge in the research domain is only a very limited subset of the available possibilities. It seems, therefore, that an attempt to present an overview of the available choices regarding different aspects of speech and emotion-related research might be a useful contribution.

In its first part, this dissertation aims to present such an overview of existing frameworks which have been or might be used in research on speech and emotion. The respective qualities of different approaches are outlined, and the conclusion is drawn that the most appropriate descriptive framework depends on the type of application and the

questions addressed. Chapter 1 starts by presenting some vocabulary definitions and an overview of the different meanings with which the word "emotion" is associated. It also gives a very short overview of four major research traditions in emotion theory. Chapter 2 presents available descriptive frameworks for emotions, and discusses their links to particular emotion types as introduced in Chapter 1. Chapter 3 introduces a distinction between several aspects of the expression-perception process, into speaker-centered, transmission-related and listener-centered aspects. Chapters 4–6 provide more in-depth discussions of these three aspects. Finally, Chapter 7 touches the important topic of authenticity or naturalness in speech and emotion research, presenting some conceptual tools for tackling this difficult issue.

Emotional speech synthesis using emotion dimensions

The second part of the dissertation proposes a novel way of modelling emotional states in speech synthesis applications, by means of gradual dimensions rather than emotion categories. This is motivated by the wish to contribute to the improvement of the naturalness of synthetic speech.

Currently, the biggest factor improving the naturalness of synthetic speech is the synthesis method called unit selection. It is not knowledge-driven, but obtains its naturalness from the *lack* of interference with speech data produced by a human speaker. While reproducing the naturally produced way of speaking, the method does not model any properties of the speaking style realised, which results in a very inflexible voice. A more demanding and ambitious task, reaching beyond the naturalness of the synthetic voice in one communication situation, is the adaptability of the voice to a variety of communication situations. This requires explicit models of the factors determining the speaking style, as well as models of the effects of these factors. One of the factors leading to a different speaking style is undoubtedly emotion.

Requirements for an emotional synthetic voice in typical application settings are formulated. It is argued that in many envisageable scenarios of Human-Computer Interaction, as in human communication, the emotional message to be expressed by the computer system is determined jointly by the voice and other "channels" such as verbal content, and possibly facial and bodily expression, as well as by the situational context. Therefore, it

does not seem crucial for the voice prosody to contain by itself all the information necessary to determine the details of an emotional state. It is sufficient, at least as a first approximation, if the voice fits roughly with the type of emotion required by the situation and expressed via the other channels such as verbal content and, possibly, facial expression. In this respect the approach presented here is less ambitious than other work published in the past, which intended to establish the capacity of prosody to convey emotions when no other cues were available.

A different aspect of emotion expression, however, is considered most important: The possibility to represent, and to express via the voice, gradual emotions, including weak emotional states as well as emotional states changing over time. The rationale behind this requirement is the idea that in many states which would not themselves qualify as emotions, an emotional colouring is present, which has been called an “underlying emotional state” (Cowie & Cornelius, 2003).

In the light of the considerations in the first part of this thesis, the descriptive framework for representing emotional states which seems most appropriate for meeting these requirements is a dimensional one: Essential aspects of emotional states are captured by the dimensions of *activation* (the readiness to take some action), *evaluation* (in terms of positive and negative), and *power* (in terms of dominance and submission). Due to their gradual nature, these dimensions can represent states that differ only slightly from a neutral state. Gradual changes over time can easily be represented. The fact that these dimensions do not capture all relevant aspects of all emotions is in line with the observation made above: that the voice prosody need not allow the identification of the exact emotional state, as long as it is compatible with the general emotional colouring expressed.

In this thesis, the principal feasibility of using emotion dimensions for emotional speech synthesis is demonstrated. In the analysis of a spontaneous emotional speech database, robust correlations between emotion dimensions and acoustic parameters are found. These findings are formulated as rules for determining the acoustic correlates of emotion dimensions, and implemented in the text-to-speech synthesis system MARY. Finally, an evaluation of perceived appropriateness of the synthesised emotional prosody in emotionally defined contexts is carried out.

The second part of this thesis starts, in Chapter 8, with a motivation from the point of view of speech synthesis research and development, explaining why emotions are interesting to model in speech synthesis, and why emotion dimensions are a promising approach to do so. Thereafter, the starting point of the current research is described: Chapter 9 presents a review of emotional speech synthesis research to date, and Chapter 10 reviews previous research on the vocal correlates of emotion dimensions. A corpus analysis through which quantified correspondences between emotion dimensions and acoustic speech parameters are determined is described in Chapter 11, along with the results of that analysis. In view of the implementation of the results obtained from the corpus analysis and from the literature review, the speech synthesis system MARY is presented in Chapter 12, and the reasons for which this system is particularly well-suited for the task at hand are outlined. Chapter 13 formulates prosody rules for emotional speech synthesis in an implementable form and describes the technical realisation of that implementation, including a graphical user interface allowing the user to interactively explore the emotional characteristics of the synthetic voice. Chapter 14 describes a perceptual evaluation of the system. The Conclusion, finally, summarises the work presented in the preceding chapters and points towards directions in which subsequent research should be headed.

Part I

Frameworks in speech and emotion research

Chapter 1

Multiple meanings of the word “emotion”

The word “emotion” is used in the literature in a number of different ways. This chapter points out some of the different meanings, and gives an overview of some major research traditions focussing on different aspects of the complex emotion phenomenon.

1.1 Definitions

A recent paper by Roddy Cowie and Randolph Cornelius (2003) introduced a number of useful terms and concepts that will be adopted in this thesis.

First of all, the term “**fullblown emotion**” is used for the type of fully developed emotion episode which is typically the object of emotion theories. Without prejudice to particular emotion theories, this term denotes the fully developed form of an emotion, which is typically intense, and incorporates most or all of the aspects or facets considered relevant for the emotion syndrome (see below). Where “emotion” is used in this sense, it is natural to formulate tight criteria which a state needs to fulfill in order to qualify as an emotion. The particular criteria depend on the emotion theory.

Second, the term “**underlying emotion**” denotes the type of emotional colouring which is part of most or all mental states. An example given by Cowie and Cornelius is that of friendliness, which is not itself an emotion, but implies some positive underlying emotion. The phenomena described by this term are usually not central to emotion theories, and seem more difficult to describe. However, lay people seem to consider underlying emotion as relevant in communication, which makes the concept a useful one in

the context of speech and emotion.

While on the conceptual level, the distinction between fullblown emotions and underlying emotion seems useful, it can be expected that a wide range of intermediate states exist which are too emotional to be merely underlying but not developed enough to qualify as fullblown. The term proposed by Cowie and Cornelius for the entire range from underlying to fullblown emotions is “**emotional states**”.

Finally, the term “**emotion-related states**” covers such states that are not themselves emotions, but which have certain aspects in common with emotions, e.g. moods, states of arousal, or attitudes.

Starting with this vocabulary, the following sections aim to present an overview of what different persons and research traditions consider important about emotions.

1.2 Fullblown emotions as multi-faceted syndromes

There appears to be reasonable agreement in the scientific community that fullblown emotions are multi-faceted syndromes (Cowie & Cornelius, 2003; Cornelius, 1996; Plutchik, 1994; Sokolowski, 1993). The facets involved in a fullblown emotion episode include at least the following:

- appraisal of a stimulus situation or event, i.e. the evaluation of the meaning of the stimulus for the individual, including a valence (Arnold, 1960);
- physiological adjustments, such as increased heart beat rate and muscle tension (Scherer, 1986);
- action tendencies such as attack or flight patterns (Frijda, 1986);
- subjective feeling (Russell, 1980);
- and expressive behaviour such as facial expression (Ekman, 1993), bodily expression, and speech including verbal (Whissell, 1989) and non-verbal, suprasegmental aspects (see Chapter 5).

It becomes clear from this short and probably incomplete list that no one study can deal with all relevant aspects of an emotion. Depending on the research tradition in which a study is based, different subsets of these facets are typically investigated. In the following section, an overview of some major traditions in emotion theory is given, which might be useful for understanding the theoretical background on which specific studies draw.

1.3 Four perspectives on fullblown emotions

In psychology, theories of emotion are grouped into four main traditions (Cornelius, 1996), each making different basic assumptions about what is central to the nature of emotion. Hereafter, the basic ideas in each of the traditions are briefly presented, closely following overviews given by Randolph Cornelius (1996; 2000).

1.3.1 The Darwinian perspective

Starting with Charles Darwin’s work laid down in his 1872 book *The Expression of Emotion in Man and Animals*, emotions are seen as reaction patterns shaped through evolution. Natural selection favoured responses that had a survival value: The emotion serves a function leading to a selection advantage.

In this perspective, emotions are seen as common to humans as a species, which implies that all humans should have more or less the same emotions. Furthermore, humans should share a certain range of emotions with other mammalians.

A concept typically associated with the Darwinian tradition is that of *basic emotions* – a small number of emotions which are evolutionarily shaped in order to fulfill specific, survival-benefiting functions (see also 2.1.1, p. 22).

The survival-benefiting value of the emotion pattern can be of different kinds. The main function may be some sort of a biological activation, making the individual fitter for dealing with a given situation, possibly including the tendency or urge to perform a given action; but it may also be a display directed towards an external observer, be it friend or foe, influencing the observer’s behaviour, such as in threat or submission displays. This was pointed out by John Ohala (1996), who also noticed that in the two cases, the emotion *expression* has entirely different roles: In the first case (biological activation), the emotion may “leak out” and become perceivable by an external observer as a possibly undesirable side effect of its benefiting function. In the second case, however, the expression of the emotion is in itself a benefiting function, in that it influences the observer.

An important finding in the Darwinian tradition is the universality of of facial ex-

pressions of emotions, demonstrated by Paul Ekman (1993). He showed that at least six emotions (happiness, sadness, anger, fear, surprise, and disgust) were expressed in the face and recognised from the face in much the same way in many different cultures.

While the emotions themselves and their facial expressions are seen as largely universal, Ekman describes culture-specific *display rules* defining which emotions can be expressed in a given situation, and which are considered inappropriate and must be concealed. In Ekman’s words, “A display rule specifies who can show what emotion to whom, when.” (Ekman, 1977b, p. 62) The term *leakage* is used by Ekman to describe the situation “when emotional responses escape attempts to conceal them” (Ekman, 1977b, p. 63). When the concealment attempt is a deliberate act, leakage is more likely to occur than when the concealment attempt is habitual.

1.3.2 The Jamesian perspective

In the tradition of thought about emotions founded by William James in his 1884 article “What is an emotion?”, the body is seen as essential for the emotion. It is through the proprioceptive experience of bodily changes that the emotion arises. As in the Darwinian tradition, the bodily changes follow some stimulus more or less automatically. Emotion arises through the perception of these changes.

As a consequence, without the perception of the body there could not be emotion. There seems to be some support for this claim, e.g. from the neurologist Antonio Damasio (1994), who describes patients suffering from anosognosy, who cannot feel their body and who are unable to experience emotions.

An aspect of the Jamesian perspective directly related to experimental approaches is the so-called *facial feedback hypothesis* (Cappella, 1993). It states that there is a small but reliable effect of a person’s facial expression on his or her subjective emotional experience, at least as far as the valence of the emotion is concerned. For example, a person assuming the facial muscle configuration corresponding to a happy face will report feeling happier than a person assuming the facial configuration of an anger expression.[1]

[1] The effect exists whether the subject is aware of posing an emotional facial display or not: Strack et al. (1988) had subjects rate the funniness of cartoons while they held a pen either in their lips (corresponding to a facial configuration similar to a frown) or between their teeth (facial configuration similar to a smile). Subjects holding the pen between their teeth rated the cartoons as funnier than subjects holding the pen in their lips.

1.3.3 The cognitive perspective

In cognitive emotion theories, the central concept is *appraisal*, a term coined by Magda Arnold (1960). It denotes an evaluation of a stimulus through relatively low-level, automatic cognitive processes. The appraisal of a stimulus determines the significance of the stimulus for the individual, and triggers an emotion as an appropriate response.

Details about how and according to which criteria the perceived stimuli are evaluated and which reactions are triggered have been worked out by a number of researchers. The most notable for a speech and emotion researcher is the *component process model* developed by Klaus Scherer (1984b), from which Scherer has made detailed physiological predictions about the vocal changes associated with certain emotions (Scherer, 1986), which in their large majority were verified experimentally (Banse & Scherer, 1996).

Scherer’s component process model (1984b) specifies the appraisal process as a series of *stimulus evaluation checks* (SECs) effected in a given temporal order, from the simplest to the most complex: novelty check, intrinsic pleasantness check, goal/need significance check, coping potential check, and norm/self compatibility check. Each SEC triggers an appropriate, survival-benefiting response in the various facets of emotions (see 1.2). An emotional state as denoted by a verbal label such as happiness, sadness etc. is represented in the component process model as a given configuration of SEC outcomes.

Another cognitive emotion model, detailing the presumed appraisal structure leading to the multitude of emotions, was proposed by Andrew Ortony, Gerald Clore and Allan Collins (1988). In this so-called OCC model, emotions are seen as valenced reactions to three types of stimuli: Events, agents, and objects. Central to appraising events is their desirability, with respect to goals; central to appraising agents is the praiseworthiness of their actions, with reference to standards; and central to appraising objects is their appealingness determined by attitudes. The model is formulated in a way permitting its implementation in AI systems. Several conversational agent systems have adopted the model, in a so-called “affective reasoning” module: On the basis of domain knowledge, a situation is appraised using the OCC model, and the appropriate emotional reaction is determined, influencing subsequent actions and/or expressive behaviour of the agent (e.g., André et al., 1999).

1.3.4 The social constructivist perspective

The “youngest” among the views on what emotions are is attributed by Cornelius (1996) to James Averill (1980). Here, emotions are seen as socially constructed patterns that are learned and culturally shared. They fulfill a social purpose, regulating in various ways the interactions between the individuals. Not only the expression of the emotions, but the emotions themselves including the subjective experience are seen as culturally constructed. This differs in particular from Ekman’s notion of *display rules*, which constitute socially shared filters for the expression of biologically “hard-wired” emotions. Social constructivists recognise the existence of biological foundations for emotions; however, they consider their importance as secondary compared to the socially constructed mechanisms.

1.3.5 Discussion

At first sight, the claims made and evidence found in the four traditions may seem contradictory. In particular, the central idea in the Darwinian and Jamesian traditions that there are universal “basic” emotions seems irreconcilable with the central conception in social constructivism that emotions are culturally created “scripts”. However, a closer look reveals that the four perspectives actually focus on different aspects of emotions. In Cornelius’ words,

> emotions may be seen as being organized on a variety of levels. Neurophysiologists are interested—almost by definition—in the neural organization of emotion, Darwinians are interested in the evolutionary organization of emotion, Jamesians are interested in the bodily organization of emotion (for want of a better term), cognitive-emotion theorists are interested in the psychological organization of emotion, and social constructivists are interested in the social-psychological and sociological organization of emotion. (Cornelius, 1996, p. 211)

In this view, the use of different research paradigms and methods in the different traditions is a natural consequence of the nature of the phenomena being investigated. Such a reconciliatory perspective regards the multitude of approaches to the study of emotions

not as a source of conflict, but rather as a wealth of possibilities from which a researcher can draw, e.g., for planning an experiment.

At the same time, the fact that the various traditions use different methodologies, appropriate for their research goals, makes clear the importance of a conscious decision in experimental design, a conscious identification of the tradition in which one’s research questions are based. It is one aim of this dissertation to present an overview of the available choices.

Depending on the particular aspects a given theory emphasises, descriptive frameworks have emerged which capture those aspects particularly well. A large variety of description systems have been developed for capturing the essential aspects of fullblown emotions according to a given theory (see Chapter 2). Naturally, as fullblown emotions are maximally different from one another, their descriptions capture these differences very well, often in categorical ways.

1.4 Underlying emotions

As Cowie & Cornelius (2003) have pointed out, most (or all) mental states contain an emotional aspect, an underlying emotion. This is recognised by emotion researchers such as Richard Lazarus and Ross Buck, when they write “every way of experiencing the world involves a stance that is emotional” (Lazarus, 1999a, p. 11) and “we are always ... happy, angry, and so forth, to some extent ... one can always ask how loving, attached, or affiliative one feels” (Buck, 1999, p. 325).

While these non-extreme, underlying emotions may be investigated to a lesser extent in the context of emotion theories, it seems nevertheless that they are highly interesting for speech and emotion research: Most situations in which speech is produced and perceived do not involve fullblown emotions, but are likely to involve some underlying emotion.

Descriptive frameworks have been developed which are capable of describing underlying emotions (see Chapter 2). Naturally, as underlying emotions are often low in intensity and lack the typicality of fullblown emotions, the descriptive systems designed for measuring them are better at capturing gradual general tendencies than category-specific configurations.

1.5 Emotional states

The relationship between fullblown emotions and underlying emotions does not seem to be a fixed piece of shared scientific knowledge yet. If fullblown emotions are more or less delimitable *episodes*, triggered by clearly identifiable events (Scherer, 1994), then some of their properties are likely to be structurally different from underlying emotions accompanying many other mental states. In particular, the idea of fullblown emotions as an “alarm system” interrupting normal behaviour, making the organism deal with certain stimuli (Sloman, 1998), does not seem to be easily scalable to less intense emotional states.

Other aspects of fullblown emotions, however, may be easier to connect to underlying emotions. It seems reasonable to assume that most of the facets of fullblown emotions (see 1.2) may allow gradual changes. For instance, physiological arousal or the appraisal of a stimulus are likely to be a matter of degree rather than allowing only binary settings.

These considerations suggest it might in principle be possible to describe the entire spectrum of emotional states, from underlying emotions to fullblown emotions, with a single descriptive framework. Such a system would have to capture all relevant aspects of the emotional state, including general tendencies for the weaker and specific properties for the more intense states. Currently, however, no single unifying system seems to be available; instead, different systems can be used according to the aspect relevant for specific research questions.

1.6 Emotion-related states

In whichever way the boundaries around the concept of emotion are drawn, there are states sharing some properties with emotions. Where exactly the emotion ends and where the emotion-related state starts is likely to be of secondary relevance for some research questions in the field of speech and emotion, as long as the shared property of both has an impact on speech. Therefore, a number of neighboring states are mentioned hereafter.

In the temporal domain, *moods* are “longer lasting feeling states that need not be about anything in particular ... They are relatively mild, mundane affective experiences that are neither distracting nor disruptive, but do influence a variety of behaviors” (Guerrero et al., 1998, p. 7). This description shows a strong similarity between moods and underlying

emotions. However, moods differ from underlying emotions in that they are “frequently described as more diffuse and nonspecific” (Guerrero et al., 1998, p. 7). Also, moods are “most often described by their valence—either positive or negative, happy or sad, good or bad or neutral” (Guerrero et al., 1998, p. 7). Thus, while sharing subjective feeling and valence with underlying emotion, moods do not usually seem to be associated with an arousal component or action tendencies.

Similarly, *attitudes* seem to involve a valenced appraisal, but not necessarily arousal, as can be deduced from what Cowie & Cornelius (2003) call a “standard definition of attitude”: An attitude entails “categorisation of a stimulus object along an evaluative dimension” (Zanna & Rempel, 1988, p. 319).

On the other hand, states exist to which *arousal* is essential while valence does not play a major role, such as “stressed”, “excited” or “sleepy”.

If the effects of emotions on speech are the research topic, it seems relevant whether the properties which emotions share with such related states have an impact on speech. Investigating that question may involve systematic variation of such common properties, like valence or arousal, within or across the boundaries of what is considered emotion.

1.7 Lay people’s emotion concepts

In psychology, a research tradition exists which is not concerned with the emotion phenomenon itself, in the fullblown or underlying sense described above, but rather with the properties of lay people’s mental emotion concepts. This research tradition usually employs questionnaires in which subjects report their understanding of the emotional meaning of a situation (e.g., Mehrabian & Russell, 1974) or of an emotion word (e.g., Scherer, 1984a; Whissell, 1989). According to James Russell, self-reports of one’s own felt state also make use of this mental system of emotion concepts: “[T]he cognitive structure that is utilized in interpreting the meaning of verbal messages or of facial expressions from others is the same structure utilized in the process of conceptualizing one’s own state ... Self-report of one’s affective state is thus a task like the labelling of photographs of faces and can be taken as a means of revealing the way in which emotions are conceptualized.” (Russell, 1980, p. 1176–1177).

It seems useful to outline this approach as differing from methods in research on full-

blown emotions themselves, where aspects of the actual emotion episode would be controlled and/or measured.[2] The implications of this distinction will be discussed in more detail in the chapters dealing with descriptive frameworks (Chapter 2), sources of emotional speech data (Chapter 4), and methodologies for perception tests (Chapter 6).

[2] A typical example illustrating this distinction is the measurement of the physiological effects of a controlled varation of appraisal components (e.g., Smith, 1989; Johnstone & Scherer, 1999).

Chapter 2

Descriptive frameworks for emotions expressed in speech

This chapter presents a number of descriptive frameworks which have been used or could be used in speech and emotion research. It draws partly on the insightful article by Roddy Cowie and Randolph Cornelius (2003). An extended discussion is provided for the dimensional framework, in order to form a solid basis for its use in the second part of this thesis.

2.1 Emotion categories

The most straightforward description of emotions is the use of emotion-denoting words, or category labels. Human languages have proven to be extremely powerful in producing labels for emotional states: Lists of emotion-denoting adjectives exist that include at least 107 English (Whissell, 1989) and 235 German (Scherer, 1984a) items; according to Cowie & Cornelius (2003), the Semantic Atlas of Emotion Concepts (Averill, 1975) lists 558 words “with emotional connotations”, i.e. which contain at least some underlying emotion.

It can be expected that not all of these terms are equally central. Therefore, for specific research aims, it seems natural to select a subset fulfilling certain requirements. In the following, a number of approaches to selecting such lists are outlined.

Lazarus (1999b)	Ekman (1999)	Buck (1999)	Lewis & Haviland (1993)	Banse & Scherer (1996)	Cowie et al. (1999a)
anger	anger	anger	anger / hostility	rage / hot anger irritation / cold anger	angry
fright	fear	fear	fear	fear / terror	afraid
sadness	sadness	sadness	sadness	sadness / dejection grief / desperation	sad
anxiety		anxiety	anxiety	worry / anxiety	worried
happiness	sensory pleasure	happiness	happiness	happiness	happy
				elation (joy)	
	amusement satisfaction contentment		humour		amused pleased content
		interested curious surprised			interested
	excitement				excited
		bored		boredom / indifference	bored
					relaxed
		burnt out			
disgust	disgust contempt	disgust scorn	disgust	disgust contempt / scorn	
pride	pride	pride arrogance	pride		
jealousy		jealousy			
envy		envy			
shame	shame	shame	shame	shame / guilt	
guilt	guilt embarrassment	guilt	guilt embarrassment		
					disappointed
relief	relief				
hope					
					confident
gratitude					
love			love		loving affectionate
compassion		pity moral rapture moral indignation			
aesthetic					

Table 2.1: Recent lists of key emotions, reproduced from Cowie & Cornelius (2003, Table 2).

2.1.1 Basic emotions

Especially in the Darwinian and the Jamesian traditions of emotion research, there is general agreement that some fullblown emotions are more basic than others. From a Darwinian point of view, the basic emotions correspond to specific, evolutionarily shaped functions benefiting survival. Consequently, these emotions are expected to be universally found in all humans. A Jamesian addition to this conception is to expect specific patterns and possibly sub-systems for these emotions in peripheral and brain physiology.

A number of criteria have been applied for identifying the basic emotions. In an overview chapter, Robert Plutchik mentions the following approaches to proposing basic emotion lists: Evolutionary approaches, neural approaches, a psychoanalytic approach, an autonomic approach, facial expression approaches, empirical classification approaches, and developmental approaches (Plutchik, 1994, p. 58).

As the numbers of basic emotions are usually small (in earlier studies, less than ten; in recent proposals, between 10 and 20; see Table 2.1), it is possible to characterise each basic emotion category by its properties, according to those facets of the emotion syndrome (see 1.2, p. 10) that are considered relevant. As the states considered are usually fullblown emotions, they are sufficiently specific to be clearly and categorically distinguished by such a characterisation.

2.1.2 Superordinate emotion categories

Basic emotions are thought of as being categories defined by highly specific functional or physiological patterns. Alternatively, emotion categories have been proposed as more fundamental than others on the grounds that they *include* the others. An example may clarify the idea: Shaver et al. (1987) proposed five prototypes underlying all emotion categories: Anger, love, joy, fear, and sadness. Joy, for example, would be subdivided into pride, contentment, and zest. Cowie & Cornelius (2003) give a short overview of recent proposals of such lists (see also 2.2).

The idea seems compatible with cognitive accounts of emotions, which describe emotions in terms of the cognitive evaluation components leading to them. Both Scherer (1984b) and Ortony et al. (1988) suggest that an emotion A is a more general form of another emotion B if the appraisal components defining the emotion A form a subset of

the appraisal components defining the emotion B. An example given by (Scherer, 1984b, p. 309) is that of the general emotion “anger” being subdivided into “just anger” and “blind anger” depending on the outcomes of particular stimulus evaluation checks not specified for the general “anger”.

2.1.3 Essential everyday emotion terms

A pragmatic approach is to ask for the emotion terms that play an important role in everyday life. The approach is exemplified by the work of Cowie et al. (1999a), who proposed a Basic English Emotion Vocabulary. Starting from lists of emotion terms from the literature, subjects were asked to select a subset which appropriately represents the emotions relevant in everyday life. A subset of 16 emotion terms emerged (see Table 2.1).

2.1.4 Agreement on the meaning of category labels

In an experimental setting, the definition of the states under study is of major importance. In a review of the literature on vocal emotion expression, Klaus Scherer attributes some of the apparently contradictory results of existing studies to the lack of clear definition of the emotional states studied and of the way they were induced (Scherer, 1986). As a remedy, recently used in major studies (Banse & Scherer, 1996; Leinonen et al., 1997), a definition of emotion words by frame stories is given, e.g., to actors expressing the emotions.

2.2 Prototype descriptions

A description of lay people’s emotion concepts in terms of clear-cut boundaries between non-overlapping categories is difficult. An alternative description has been proposed by a number of researchers (Fehr & Russell, 1984; Shaver et al., 1987; Russell, 1997), namely a prototype-based definition of emotion concepts. Using Wittgenstein’s notion of *family resemblance*, they proposed that emotional states would be recognised as members of the same emotion by comparison to a prototypical mental image of the emotion. In their accounts, no necessary or sufficient criteria are required for the definition of an emotion; instead, membership in an emotion class is considered a matter of degree based on the similarity with the corresponding emotion prototype. Consequently, a given emotional state can be a member of several emotion classes to varying extents.

It has to be noted that prototype theories make statements explicitly about emotion concepts rather than emotions (see also 1.7, p. 17). Prototype descriptions may be interesting especially when people's emotion concepts are the research object, such as in listener inferences of a speaker's emotional state (see also Chapter 3).

A research strategy for identifying emotion prototypes was applied by Shaver et al. (1987). Starting from similarity ratings, they used a hierarchical cluster analysis technique rather than the classical multidimensional scaling technique used in studies finding a dimensional organisation of emotion concepts (see 2.6). The hierarchical analysis allowed them to identify five basic-level emotion categories (anger, love, joy, fear, and sadness) and 25 subcategories.

No prototype-based descriptions seem to have been explicitly applied in the field of speech and emotion (maybe with the exception of Amir et al. (2000), see 3.2, p. 37); therefore, the details of such descriptions remain to be worked out. A simple first step for introducing prototype-based ideas into, e.g., perception tests might be to replace forced choice tests with scale ratings measuring degree of category membership. Instead of having to make an all-or-nothing decision between a number of category labels, subjects would determine for each of the category labels the degree of appropriateness for the stimulus.

2.3 Physiology-based descriptions

From a Jamesian point of view, the essential aspect of an emotion is the bodily state associated with it. Variables which are typically measured include peripheral physiological measures such as skin conductance and heart rate (Smith, 1989), as well as brain physiological measures (Alter et al., 2000).

The natural use of physiological measures in speech and emotion research is the establishment of correlations between physiological measures and vocal changes. Scherer (1986) made detailed predictions about the vocal effects of emotions based on emotion-related physiological changes. These predictions were verified in a large experiment (Banse & Scherer, 1996).

2.4 Appraisal-based descriptions

For cognitive emotion theories, it is natural to describe emotions in terms of the appraisals involved. As mentioned above (p. 13), such approaches include the ones by Scherer (1984b) and Ortony et al. (1988).

As an example, Scherer (1984b) describes "anger" as involving the following stimulus evaluation checks: highly relevant; obstructive to own goals; unjust; low control; high power; and low norm compatibility. In other words, anger is triggered when something is perceived as an important, unfair obstruction to one's goals, about which one is able to do something.

While in their current formulations, these models seem to refer rather to fullblown than to underlying emotions, it seems principally possible to extend them by conceiving of appraisals as gradual rather than dichotomic. In the example of anger, one could propose that the appraised degree of relevance has a direct effect on the intensity of the anger emotion: all other things being equal, a situation appraised as relatively irrelevant might produce a minor annoyance, while the same situation interpreted as centrally important may produce a fullblown rage.

2.5 Circumplex models

Several researchers have come to the conclusion that emotion concepts can be represented adequately by a circular structure (Schlosberg, 1941; Russell, 1980; Plutchik, 1980). Proximity of two emotion categories in the circumplex represents conceptual similarity of the two categories.

The circular pattern was originally proposed by Schlosberg (1941) in the context of explaining the recognition errors in a facial expression rating task: If categories were ordered in a particular circular structure, confusions mainly occurred between adjacent categories. In more recent research, the circumplex pattern was confirmed by several different methods for characterising emotion words: Russell (1980) used a categorisation and sorting method specifically designed for testing circularity, a grouping task providing similarity measures used for multidimensional scaling, and direct positioning on pleasure and arousal dimensions by means of semantic differential scales. All three methods yielded nearly identical circular patterns. Plutchik (1980) had his subjects rate similarity to three

emotion words (“accepting”, “angry”, and “sad”) which seemed sufficiently different and converted these similarity ratings to angular locations on a circle. In addition, he used factor analysis of ratings on semantic differential scales. From the domain of self-report of emotional state, Russell (1980) obtained similar distributions: Subjects reported their current emotional state on bipolar scales representing the emotion dimensions as well as on unipolar scales representing emotion adjectives. Regression analysis (prediction of the value on the adjective scale from the values on the emotion dimensions) as well as principal component analysis of the adjective ratings provided patterns very similar to the circular structure obtained in the concept characterisation tasks.

A major benefit of circumplex models is that they provide an explicit notion of the degree of similarity between emotion categories. Adjacent categories in the circle are very similar, while opposite categories are maximally different from each other.

2.6 Emotion dimensions

The description of emotional states by means of emotion dimensions is discussed here in more detail than the other methods, because in the second part of this thesis, emotion dimensions are used as the descriptive framework.

While they may not be very well known in the speech community, dimensional descriptions are well established in the psychological literature. In the following, an overview over their historic development is given, and their role in relation to emotion theory is discussed.

2.6.1 A historical overview

The idea that emotional experience is organised according to a small number of basic dimensional properties can be tracked at least as far back as to Wilhelm Wundt (1896), who writes:

> In the diversity of feelings consisting of a plethora of different and most finely graded qualities, various main directions are to be distinguished, that stretch between certain feeling opposites of dominating character. Such main directions of feeling can therefore always be expressed by means of two terms indicating opposites. Thereby, each term has to be considered as a collective expression encompassing an infinite amount of individually varying feelings.
>
> In this sense, three main directions can now be stated: we want to call them the directions of pleasure and displeasure, of exciting and depressing and finally of tensing and releasing feelings. An individual feeling can show either all of these directions, or only two of them, or it can also belong to only one of them. (Wundt, 1896, pp. 97–98, translation MS) [1]

In today’s terms, Wundt proposes three bipolar, independent scales pleasure-displeasure, excitation-depression and tension-release, which however do not seem to be based on reported experimental evidence.

It was Harold Schlosberg who introduced the idea of basic dimensions of emotionality into the field of experimental psychology, in the context of perceived emotional content of facial expressions (Schlosberg, 1941). He asked subjects to judge a number of facial expressions along a scale consisting of six emotion categories. His motivation was the following:

> If it were possible to arrange facial expressions along a continuum, instead of in an indefinite number of categories, it would be possible to obtain some numerical measure of divergence in judgments. (Schlosberg, 1941, p. 498)

The results of his study confirmed the assumed ordering of the categories, as the vast majority of answers generally fell into adjacent categories. Additionally, the results suggested the scale to be circular, with the two ends of the scale showing the same type of answer distribution as was observed between adjacent categories. A two-dimensional structure such as a circle needs two criteria (or dimensions) for the distinction between the categories composing it – with only one distinguishing criterion, the circle would become a flat line. Interpreting the meaning of the categories on the circle, Schlosberg came to the conclusion that the most important variable distinguishing the categories he proposed

[1] German original: “An der so aus einer Fülle verschiedener und auf das feinste abgestufter Qualitäten bestehenden Mannigfaltigkeit der Gefühle sind jedoch verschiedene Hauptrichtungen zu unterscheiden, die sich zwischen gewissen Gefühlsgegensätze von dominirendem Charakter erstrecken. Solche Hauptrichtungen des Gefühls können daher immer durch je zwei Bezeichnungen ausgedrückt werden, die Gegensätze andeuten. Dabei ist aber jede Bezeichnung wieder als ein Collectivausdruck anzusehen, der eine unendliche Menge individuell variirender Gefühle umfasst.
In diesem Sinne lassen sich nun drei Hauptrichtungen feststellen: wir wollen sie die Richtungen der Lust und Unlust, der erregenden und beruhigenden (excitirenden und deprimirenden) und endlich der spannenden und lösenden Gefühle nennen.”

was the Pleasantness-Unpleasantness dimension, whereas for the second axis, he proposed "something like 'Attention-Rejection' " (Schlosberg, 1941, p. 506).

Schlosberg saw the description of facial expressions by means of two dimensions as a reduced account:

> To give a complete description of all facial expressions would call for the addition of quite a few factors or dimensions of decreasing importance. (Schlosberg, 1941, p. 507)

Affective dimensions underlying all concepts were identified by Osgood et al. (1957) by means of the *semantic differential* technique. The affective properties inherent in natural language concepts are assessed by means of a number of paired adjective scales on which each concept is positioned. Using 50 paired adjective scales (e.g., heavy–light, sweet–sour, bright–dark, etc.), 20 concepts (e.g., lady, boulder, lake, etc.) were rated by subjects. In a factor analysis of the responses, three factors were identified as being fundamental for the characterisation of the concepts. Osgood et al. named them *evaluation*, *potency*, and *activity*. While interpreting these factors as most important in their subjects' judgments of the concepts, Osgood et al. were aware of the fact that they did not capture the entirety of a concept's meaning: "the representational state indexed by the semantic differential is not the only determinant operating in lexical encoding. It is a necessary but not a sufficient condition." (Osgood et al., 1957, p. 323–324)

Albert Mehrabian and James Russell (1974) presented collected evidence that there were three dimensions underlying emotionality. Evidence comes from a variety of domains, including intermodality responses, synesthesia[2], physiological reactions, and Osgood *et al.*'s semantic differential. The three emotion dimensions they proposed were pleasure, arousal and dominance, and thus corresponded to the three dimensions evaluation, activity and potency found by Osgood et al. (1957).[3] They presented a list of 18 paired-adjective scales (six per emotion dimension) particularly suited for measuring the three emotion dimensions in semantic differential style rating tests (see Table 6.1, p. 69).

The same authors provided additional evidence (Russell & Mehrabian, 1977) for these three dimensions, by means of two different methodologies. On the one hand, 200 situations described by short paragraphs of text were rated on scales measuring the three emotion dimensions and on emotion adjective scales. Multiple regression analysis showed that the three emotion dimensions, along with a response style variable, "accounted for almost all of the reliable variance in the various scales of emotion investigated" (Russell & Mehrabian, 1977, p. 280). On the other hand, subjects defined 151 terms denoting emotional states by means of the 18 rating scales measuring pleasure, arousal and dominance, which provided coordinates of these terms on the three emotion dimensions and showed that "all eight possible combinations of high and low values along each of the three dimensions actually occur as components of various emotional states" (Russell & Mehrabian, 1977, p. 292).

[2] The stimulation in one sense affects perception in another.

[3] See 2.6.2 for a discussion of the different names given to dimensions by different researchers.

In 1975, Rex Green and Norman Cliff investigated the interpretation of emotional speech stimuli in terms of emotion dimensions (Green & Cliff, 1975). One actor expressed 11 emotions speaking letters of the alphabet. The stimuli were presented in pairs to subjects who rated their similarity. Multi-dimensional scaling techniques yielded a three-dimensional interpretation with the dimensions "pleasant-unpleasant", "excitement", and "yielding-resisting". Each of the stimuli was also rated on seven semantic differential adjective scales measuring tone-of-voice. A principal component factor analysis suggested two dimensions, the first being related to the thin-thick and high-low pitch sound of the voice, the second representing pleasant-unpleasant feelings. The two dimensions found in semantic differential scales corresponded strongly to the first two dimensions found in similarity judgments. Stimuli fell in a V-shape in two-dimensional space: Stimuli that were either highly pleasant or unpleasant were also excited, while stimuli unmarked in pleasantness were low in excitement.

A few years later, Russell (1980) presented evidence supporting the first two dimensions, pleasure and arousal. Additionally, he found emotion concepts to be located along the periphery of a circle (see 2.5), in a way similar to Schlosberg's (1941) original idea.

While the particular concepts he used, representing fullblown emotions, were ordered in the shape of a circle in the two-dimensional space, he remarked that

> a disk or a wheel would be a better image, since affective states of moderate intensity would fall toward the middle of the space, with the origin presumably corresponding to adaptation level or a neutral feeling. (Russell, 1980, p. 1170)

In 1984, Klaus Scherer (1984a) presented results from a multidimensional scaling experiment with 80 terms out of a list of 235 relatively synonym-free emotion-denoting adjectives, for which similarity judgments had been obtained. The results of the multidimensional scaling, yielding two dimensions, were consistent with an interpretation in terms of a "positive/negative evaluation" dimension and an "activity" dimension. However, Scherer proposed an alternative interpretation as "low versus high discrepancy of actual and desired state" and "low versus high control and/or power", distinctions important in the stimulus evaluation checks in his component process model (see 1.3.3, p. 13).

David Watson and Auke Tellegen (1985) proposed a variant of a two-dimensional structure, based on a re-analysis of a large number of studies of self-reported mood. They considered two dimensions named Positive Affect and Negative Affect more appropriate than activation and evaluation. Their dimensions are rotated by 45° compared to the activation and evaluation dimensions. This is an interesting variant demonstrating the uncertainty of interpretation of a dimensional structure "emerging" from the data (see also 2.6.2 below) – it would seem difficult to decide which of the proposed structures is the "right" one, and it may well be that both interpretations are valid and vary only in their suitability for a given application domain. Watson et al. (1988) later published a concise measure of these dimensions, the so-called PANAS (Positive and Negative Affect Schedule), in which each dimension is represented by ten adjective scales.

Margaret Bradley, in an article investigating emotional memory by means of emotion dimensions (Bradley, 1994), reported data from rating studies of emotion pictures and sounds collected at University of Florida (called the International Affective Picture System and the International Affective Digitized Sounds, respectively). Stimuli were rated for pleasantness, arousal and dominance. The results on the first two dimensions were presented graphically in a two-dimensional plot, which showed that "the number of stimuli occurring in each quadrant of space is roughly equivalent, with, perhaps, the exception of the unpleasant, calm portion of space." (Bradley, 1994, p. 101) With respect to the dominance dimension, Bradley states that "when stimuli clearly involve some aspect of social interaction (e.g., narrative or actual events described by text or film), the dominance dimension—sometimes construed as social control or aggression—accounts for significant variability" (Bradley, 1994, p. 100).

The cross-cultural universality of emotion dimensions was investigated, e.g., by Timothy Church et al. (1998). After establishing a list of 256 Filipino emotion-related adjectives, these terms were rated in a similarity judgment task by native speakers. Multidimensional scaling provided a solution in which up to six dimensions were required for representing the data optimally. In the two-dimensional solution, in which the most essential aspects of similarity were represented, the dimensions were clearly identifiable as pleasantness and arousal. In addition, the emotion terms formed a near-perfect circumplex. In the three-dimensional solution, the first two dimensions were again identifiable as pleasantness and arousal. The third dimension, however, was not dominance, but related to certainty/uncertainty in combination with negative emotions.

An interesting recent description of two emotion dimensions was given by Roddy Cowie et al. (2001). They described emotional states in terms of a two-dimensional, circular space, the axes of which were labelled "evaluation" (from negative to positive) and "activation" (from passive to active). They interpreted evaluation as a simplified description of the appraisal concept seen as central in cognitive emotion theories (see 1.3.3, p. 13), where the valence of stimuli is considered a fundamental aspect of all appraisals (e.g., Ortony et al., 1988). The activation dimension was presented by Cowie et al. as a simplified representation of action tendencies, which are one of the facets of emotions (see 1.2, p. 10).

2.6.2 What is measured by emotion dimensions?

The previous overview made it clear that the two or three dimensions of emotional meaning are commonly identified in rating experiments. Subjects report their cognitive representation of the emotional colouring of some item, be it their own current emotional state, the emotion concept denoted by a verbal label, a facial expression, or a situation. It seems remarkable that in a large number of different studies applying different rating methods, structures emerge that are consistent with accounts using basically the same emotion dimensions.

From the methods leading to their establishment, it seems clear that emotion dimensions represent the essential properties of lay people's emotion concepts (see 1.7, p. 17). Therefore, it cannot be claimed that they represent the "objective" properties of emotion phenomena directly.

An important question is whether one believes emotion dimensions to capture all relevant properties of emotion concepts or whether they are seen as a simplified and reduced description. Russell & Mehrabian (1977) consider the three dimensions emerging from their factor analysis "sufficient to define all the various emotional states", based on the observation that the three dimensions "accounted for almost all of the reliable variance in the 42 scales studied" (Russell & Mehrabian, 1977, p. 292). The opposite position is held by Richard Lazarus, who writes: "Much of value is lost by putting [emotions] into dimensions, because the simplifying or reductive generalizations wipe out important meanings about person–environment relationships, which the hundreds of emotion words were created to express... Anger, then, becomes only a kind of unpleasant activation, when in reality it is a complex, varied, and rich relational pattern between persons" (Lazarus, 1991, p. 63–64).

I tend to agree with Lazarus in that emotion dimensions are a reduced account of emotions, in which many important aspects of emotions are ignored, such as eliciting conditions and specific action tendencies. Nevertheless, I consider a dimensional description as a useful representation, capturing those aspects that appear as conceptually most important in many different methodologies, and providing a means for measuring similarity between emotional states. In particular, a dimensional description is particularly well-suited for the task undertaken in Part 2 of this thesis, namely generating a voice prosody which is *compatible* with an emotional state expressed through a different channel, rather than fully *defining* that emotional state. A reduced description is sufficient for this purpose, and does not preclude being used in conjunction with richer descriptions.

Another important property of emotion dimensions, arising from the methodologies employed, is the relative arbitrarity of the names the dimensions are given. This point is emphasised by Scherer, who formulates a general "criticism leveled at factor analysis and multidimensional scaling, namely that the reality and interpretation of the factors/dimensions exist only in the eye of the beholder." (Scherer, 1984a, p. 53) This remark seems particularly relevant to studies finding emotion dimensions "emerging" from the data. The structures identified in the data and particularly the names they are given post hoc are strongly shaped by the conceptual background of the researcher.

One of the questions that can be asked in this context is whether the dimensions of "activation" and "arousal" are to be distinguished or whether they should be treated as

synonymous or identical. Given the relative arbitrarity of the names assigned to the emotion dimensions, it is likely that results of factor analysis or scaling experiments would be compatible with either interpretation as long as the two names refer to related properties. In the case of activation and arousal, this seems to be the case: Activation points out the action aspect of emotion; arousal stresses the physiological aspect. Indeed is increased arousal, in terms of heart rate, muscle tension etc., certainly one of the things that occur in preparation of an action. Therefore, it seems reasonable to treat "arousal" and "activation" as the same emotion dimension.

2.6.3 Relations to the "real world"

Cowie et al. (2001) suggest that emotion dimensions do not only describe emotion concepts, but may in addition be linked to the emotion phenomenon itself. They suggest that activation is a simplified representation of action tendencies (and, it could be added following the discussion above, of physiological arousal). Evaluation, on the other hand, can be considered a simplified representation of the appraisal process.

From a different perspective, Scherer proposes appraisals closely related to the emotion dimensions in the context of his component process model:

> I had noticed the close correspondence between the ubiquitous three dimensions found for affect expression and three major factors characterizing an organism's processing of antecedent stimulus events that seemed jointly to determine the nature or type of emotional reaction. Specifically, the positive/negative evaluation dimension was seen to result from the intrinsic or inherent pleasantness or unpleasantness of a stimulus, the activity dimension from a mismatch between goal-/plan-related expectations and the actual state (requiring action), and the potency dimension from the organism's estimate of how well it would be able to cope with the particular stimulus event and its consequences. (Scherer, 1984a, p. 38)

It seems likely that some relations exist between emotion dimensions and the "objective" emotion phenomenon, simply because people's concepts are shaped through their experience of the objective phenomena. At the same time, it seems advisable to remember that emotion dimensions are *directly* related only to lay people's emotion concepts, and that the degree to which these overlap with reality is not immediately obvious.

2.6.4 Summary

In summary, emotion dimensions are particularly well suited for measuring people's emotion concepts. They represent the properties of emotion concepts which are considered most important by subjects. While they are not based on evidence from the domain of emotion phenomena, it seems that they can still be considered simplified descriptions of several aspects of emotion phenomena, namely action tendencies, physiological arousal, and cognitive appraisal.

It was pointed out that the names given to emotion dimensions are somewhat arbitrary; in particular, it would seem that a distinction between an "arousal" and an "activation" dimension does not make sense. The names used for the emotion dimensions in the remainder of this thesis are *activation* (also known as activity or arousal), *evaluation* (also called pleasure or valence), and *power* (also known as dominance or control).

Chapter 3

Research orientation towards the expression and perception of emotion

In this chapter, different possible orientations in investigating speech and emotion are presented. As a visual anchor for the various aspects involved, the Brunswikian lens model as used by Klaus Scherer is introduced.

3.1 The Brunswikian lens model

The various orientations which are possible in speech and emotion research can be illustrated using a graphical representation of the expression and perception process, in which one person infers another person's emotional attributes from external speech markers. In a paper reporting a study on vocal correlates of personality (Scherer, 1978), Klaus Scherer presented an adapted version of Egon Brunswik's lens model, originally presented in 1956 (Brunswik, 1956). Figure 3.1 shows a version of the lens model which was slightly adapted and simplified for the current discussion.

The figure reads as follows. A speaker state C[1] is expressed through a number of objectively measurable parameters, the so-called "distal indicator cues". In the case of speech and emotion, this corresponds to a speaker emotion, expressed through acoustic parameters. Both are in principle to be measured objectively. In the first step of the perceptual inference process, the distal cues are perceived by a listener and internally

[1] C stands for "criterion", which is relevant on the "operational level" omitted here for conciseness.

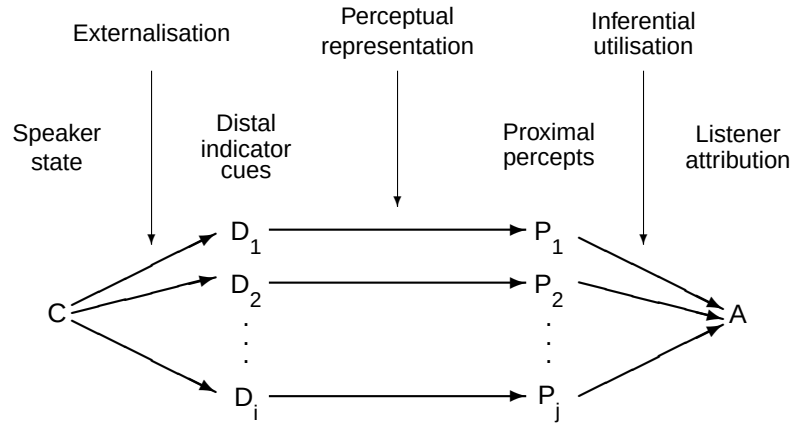


Figure 3.1: The Brunswikian lens model, adapted from Scherer (1978). This figure only shows the "phenomenal level" of the inference process, describing the objects and processes involved. The "operational level" in Scherer (1978), linked to an experimental methodology based on the lens model, was omitted for conciseness.

represented as "proximal percepts". These percepts, finally, are used for "attribution" by the listener, i.e. for inferring the speaker's state. In speech and emotion, examples of proximal percepts are subjectively perceived pitch or voice quality, while the attribution is the perceived speaker emotion.

The lens model provides visual support for a distinction between possible research orientations in speech and emotion research. Research can be *speaker-centered* ("cause-oriented" in the terms of Cowie & Cornelius (2003), "encoding studies" in the terms of Scherer (2000a)) or *listener-centered* ("effect-oriented", "decoding and inference studies"). In the first case, the goal is to collect objective data in order to establish a link between a speaker's emotional state and emotion-specific modifications of his speech. In the second case, the factors leading to a listener's percept are the topic of interest. Scherer's further differentiation into decoding studies and inference studies allows the specification of the originally modified variables: In inference studies, the distal cues are controlled, while in decoding studies, the speaker state changes.

Many studies in the field of speech and emotion seem to have combined encoding and decoding aspects. This is the case when acoustic properties of emotional speech are

measured (encoding aspect) and, independently, perception tests of the emotional speech material are performed (decoding aspect).

Some studies, however, focus particularly on the one or the other aspect. In many cases, this seems to be motivated by the type of application targeted. In the following, a few such studies are presented as examples of the respective research orientations. In addition, the descriptive frameworks employed are discussed.

3.2 Speaker-centered studies and applications

In speaker-centered studies, the goal is to establish a link between the emotional state of the speaker and measurable speech parameters. The natural type of application which requires this information is emotion recognition from speech: On the basis of a speech sample, a computer needs to make a decision regarding the emotional state of the speaker.

Three very different example studies from this domain may clarify the issue. In a study entitled "Modelling drivers' speech under stress", Raul Fernandez and Rosalind Picard elicited stress in their subjects by asking math questions while they were driving in a driving simulator (Fernandez & Picard, 2000). The stress level was operationalised through driving speed (fast/slow) and frequency of the math questions (frequent/infrequent). Speech was analysed using spectral subbands, and automatic classifiers were trained to predict the stress condition on the basis of the speech analysis.

Noam Amir et al. (2000) developed an automatic emotion recogniser. They collected speech samples for five basic emotions (anger, fear, sadness, joy and disgust) and neutral speech, by asking their speakers to recall an emotional event appropriate for that emotion category. Three physiological variables (electromyogram of the corrugator (the "frown muscle" between the eyebrows), heart rate and galvanic skin response) were measured, and used as exclusion criteria if they deviated too much from the category mean. The remaining speech utterances were analysed using a number of speech variables. For each emotion category, a reference point in the N-dimensional space spanned by these speech variables was calculated. In order to determine the emotion with which a test utterance was spoken, a normalised distance measure in the N-dimensional space (the so-called Mahalanobis distance) was calculated between the feature vector representing the test utterance and each of the reference points. This distance measure was interpreted in terms

of fuzzy category membership, i.e. it indicated the degree to which the utterance was considered part of the different emotion categories.

Tom Johnstone and Klaus Scherer (1999) elicited emotions in their subjects by having them play a computer game. In the game, eliciting conditions for appraisals were manipulated based on Scherer's component process model (enemy / friend spaceship $\Rightarrow$ goal obstruction / conduciveness; firepower $\Rightarrow$ coping potential). During the game, the subjects were asked to give a spoken report of their current emotional state, including one or more of eight possible emotion categories. A number of physiological variables were also recorded. Both speech and physiological variables were analysed with respect to the appraisals under which they were produced.

All three examples are speaker-centered studies, in that they control the speaker state in one way or the other and that they measure speech variables in order to link them to the speaker state. The particular descriptions of emotion employed in the studies differ, however. Fernandez and Picard model stress, an emotion-related state; Amir et al. model basic emotion categories; and Johnstone and Scherer model individual appraisal components. In addition, the way speech is used to infer the speaker state differs: Fernandez and Picard's classifiers make a binary class membership decision; Amir et al. calculate fuzzy membership in each possible category. Johnstone and Scherer do not attempt to predict emotional state from speech in this study.

3.3 Identification of relevant cues and percepts

An important research question in its own right is to find the particular acoustic cues and their auditory correlates that carry information about emotion. Indeed, variables carrying emotional meaning need to be identified before they can be studied more systematically.

This was a major research question in the annotation process of the Reading/Leeds Emotional Speech Database, consisting of 4 1/2 hours of spontaneous emotional speech. Recent work by Peter Roach and colleagues (Roach et al., 1998; Roach, 2000) presents their approach. An annotation system was developed consisting of information about perceived emotional content and linguistic descriptions annotated by expert transcribers. The linguistic descriptions contained an orthographic transcription, an intonation transcription using the ToBI system, and a very detailed annotation of prosodic and paralinguistic effects, such as pitch level, voice quality, loudness, and speech tempo. Several types of emotion annotation (Greasley et al., 2000; Stibbard, 2001) were created independently of the linguistic/paralinguistic annotations, with no a priori hypotheses about correlations between the two. These emotional annotations included verbal labels, lexical valency, and coded appraisals according to the OCC model (Ortony et al., 1988). Subsequent analyses of this data by Richard Stibbard (2001) showed the difficulty to find reliable patterns in this type of data. One major cause for this difficulty may have been the fine degree of detail in both the emotional and the paralinguistic transcriptions, leading to a lack of recurring configurations (e.g., most of the 73 verbal labels used occurred only one or two times in the database (Stibbard, 2001, Table 3.4)). In order to find systematic patterns in which other factors such as, e.g., regional speaker accents are levelled out, huge amounts of data would probably have been required.

An alternative way to proceed is to review available data and to collect evidence in order to build hypotheses concerning the types of speech variables carrying emotional meaning. Two recent studies follow this path: A corpus of Chinese spontaneous speech was built by Li-chiung Yang and Nick Campbell (2001), in which stretches expressing emotions and emotion-related states were identified and evidence was collected concerning the prosodic means through which they were expressed. Similarly, Ellen Douglas-Cowie examined the Belfast Naturalistic Emotion Database in view of the types of vocal signs communicating emotions. The evidence gathered in such studies can serve as the basis for subsequent, more formal investigations of the same material, for new collections of specific types of material, or for the design of specific experiments.

3.4 Listener-centered studies and applications

In listener-centered studies, the emotional meaning inferred on the basis of speech variables ("distal cues") is investigated. This is naturally connected with simulation tasks such as emotional speech synthesis: Whatever their origin, the speech parameters are modelled with the goal to convey a certain emotional state.

Again, a number of example studies are briefly presented in order to illustrate listener-centered methodologies. A study by Christer Gobl and Ailbhe Ní Chasaide investigated emotion attribution on the basis of voice quality (Gobl & Ní Chasaide, 2000). A trained

phonetician produced a number of phonetically defined voice qualities (modal, creaky, breathy, whispery, tense and lax), which were analysed using a voice source model (the so-called LF model), and copy-synthesised using a formant speech synthesis system. The copy synthesis method ensures that only the modelled parameters (in this case: the voice source characteristics as represented in the LF model) are retained from the original recordings. All other parameters are replaced by synthesis defaults. The synthesised utterances were presented in a perception test, and rated on scales measuring perceived affective content (relaxed/stressed, content/angry, friendly/hostile, sad/happy, bored/interested, timid/confident and afraid/unafraid). This measurement method allowed them to conclude that voice qualities typically had an effect on most of the proposed scales, rather than on single ones. The non-modal voice qualities clustered into two main groups perceived as signalling aroused and agressive vs. relaxed and unagressive states.

Felix Burkhardt and Walter Sendlmeier systematically tested the effect of acoustic parameters on emotion classification (Burkhardt & Sendlmeier, 2000). They systematically varied five acoustic parameters which could be controlled in the speech synthesis system they used (mean F0, F0 range, speech rate, phonation type, and vowel precision) in a number of perception experiments. Listeners were asked to classify the speech samples as one of five emotion categories (fear, anger, joy, sadness and boredom) or as neutral speech. The results allowed them to propose acoustic parameter configurations best perceived as the given emotion categories.

Juan Montero et al. (1999a) investigated the relative contribution of prosody and voice quality in emotion recognition. From an actor's portrayal of four emotions (happiness, sadness, cold anger and surprise) and neutral speech, they created small diphone synthesis databases (one per speaking style) and measured the prosodic variables for a given sentence. They then created mixed stimuli, either with emotion-specific prosody and neutral diphones or with neutral prosody and emotion-specific diphones. These stimuli were presented in a perception test. The results gave clear indications that anger was recognised mainly on the basis of voice quality and that surprise was recognised predominantly on the basis of prosody.

These examples are clearly listener-centered studies, as they are centrally interested in the perceptual effect which certain stimuli, in which distal cues are controlled, have on listeners. The parameters modelled in the stimuli are obtained either by systematic variation without relying on any human speech (Burkardt and Sendlmeier), or they process and filter human speech recordings in a way that allows the contributions of individual parameters to be identified (Gobl and Ní Chasaide, Montero et al.). The type of perceptual effects which can be found depends on the descriptive framework used for assessing the perceived emotion. For example, it is probable that Gobl and Ní Chasaide would have obtained very different results had they asked their subjects to classify their voice qualities as one of a number of emotion categories.

3.5 Some reflections on research orientation and descriptive frameworks

In Chapter 2, the descriptive frameworks presented were shown to rely to varying extents on lay persons' emotion concepts. Prototype, circumplex, and dimensional descriptions, in particular, mainly rely on reported properties of emotion concepts, while functional, physiological and appraisal models seem to be based on objectively specifiable aspects.

It seems valid to ask whether any a priori reasons exist why certain descriptive frameworks would be better suited for certain tasks. In this chapter, this question can be reformulated as whether certain descriptive frameworks seem more appropriate for speaker-centered or listener-centered studies than others.

Speaker-centered studies target the emotional state as an objective fact, independently of how people might perceive it. It would appear, therefore, that descriptions based on objective aspects, such as functional, physiological and appraisal models, are best suited for this task. Conversely, the type of descriptive framework relying basically on properties of emotion concepts is not ideally suited for the task, as it describes the objective phenomenon only to the degree to which people's emotion concepts overlap with objective measures. In particular, concept-based descriptions are likely not to accurately describe aspects of emotion such as brain activity, which are typically not perceivable and are therefore not part of everyday emotion knowledge.

On the other hand, listener-centered studies are usually concerned with the conscious percept evoked by a stimulus.[2] This percept seems to be most naturally operationalised in

[2] There are exceptions to this, such as the study by Morris et al. (1998) in which unconscious learning by means of masked angry faces is investigated. The term "listener-centered" as used in this thesis is not meant to include this type of study.

terms of people's emotion concepts, which suggests that concept-based descriptions such as prototypes, circumplex models, and emotion dimensions, are particularly well-suited for assessing the perceived emotional state in listener-centered studies.

Chapter 4

Sources of emotional speech data

One of the biggest challenges in speech and emotion research is to obtain authentic emotional speech data. Ideally, it would be desirable to have a closely controlled corpus of spontaneously emotional speech, in which all aspects are held constant except those influenced by the emotion. This, however, does not seem feasible, as the contradiction between "controlled" and "spontaneous" indicates.

Therefore, researchers have developed a number of different strategies for obtaining emotional speech data, each with their advantages and shortcomings. The most commonly used methods are described below.

4.1 Actors

The oldest and still most frequently used method of obtaining emotional speech is to have it produced by actors (e.g., Fairbanks & Pronovost, 1939; Banse & Scherer, 1996; Leinonen et al., 1997). The great advantage of this method is to have control over the verbal and phonetic speech content, i.e. all emotional states can be produced using the same sentence(s) (e.g., Mozziconacci, 1998), pseudo-sentence(s) consisting of nonsense-words (e.g., Banse & Scherer, 1996) or single word(s) (e.g., Leinonen et al., 1997). This production strategy allows direct comparisons of the phonetic, prosodic and voice quality realisations for the different emotional states expressed.

A second major advantage of using actor simulations is the ease with which it is possible to obtain expressions of extreme, fullblown emotions.

Critics of the actor-based approach question the ecological validity of the actor portrayals. There may be a risk of actors producing stereotypic forms of a given emotion, and their speech may not reflect what people would spontaneously produce. Banse and Scherer challenge these questions on two different grounds. First, they mention a method called the Stanislavski method, used by actors for auto-induction of emotional states by means of imagination or other mental or muscular techniques (Banse & Scherer, 1996, p. 619). If they succeed, actors using this method might actually experience a state similar to the intended emotion while they are producing their utterances. Banse and Scherer's second argument in favor of actor portrayals is that due to the control required by the social context in which affect expression occurs in everyday life, " 'natural' vocal affect expression is also staged" (Banse & Scherer, 1996, p. 619). Still, actors seem to be better at performing convincing emotion portrayals than non-actors (e.g., Schröder, 2003), and it seems unclear in how far the affect expression control can be tranferred from socially embedded situations onto a stage or into a recording studio.

The use of actor portrayals for the expression of underlying emotions seems to be less common. It is not immediately clear how actor portrayals would compare to spontaneous expressions in this domain. The questions regarding ecological validity, raised above for fullblown emotions, are certainly worth asking in this domain as well. In addition, there could be a tendency to over-amplify the intensity of the emotion expression, due to actor training, given that for actors, recognisability of the expression is central.

4.2 Expressive reading of emotional material

A variant of the actor approach was proposed by Nick Campbell (2000), under the heading "stimulation" of emotions. Rather than using emotionally neutral "carrier sentences", he proposed to have speakers read texts with a verbal content appropriate for the emotion to be expressed. He reported experiences in joint work with Akemi Iida in which this method seemed to evoke genuine emotion.

A similar idea was followed in the creation of the Belfast Structured Emotion Database (Douglas-Cowie et al., 2003). For each of four emotions (anger, fear, sadness, happiness) and a neutral expression, two short paragraphs were written in a way as to express the intended emotion. In addition, two transcripts from emotional episodes in the Belfast

Naturalistic Emotion Database (see 4.4 and 11.2, p. 104) were used. 38 speakers produced each of the texts. The speech material was then evaluated in view of the perceived naturalness using a tool called Validtrace, derived from Feeltrace (see 6.4.2, p. 70).

The disadvantage of this approach compared to the carrier sentence approach is of course that the text material will not be identical for different emotions, making comparisons less straightforward. Depending on the research question, it is, however, conceivable to carefully construct appropriate texts sharing the properties to be studied, such as, e.g., vowel class and probable stress and pause patterning.

A compromise between appropriate semantic content of the text and the constancy of the text across emotions was used by Joel Davitz (1964a). He embedded two emotionally neutral sentences into each of 14 emotion-specific paragraphs, which were read as a whole by the speakers.

4.3 Emotion elicitation

The elicitation of authentic emotions in subjects in the laboratory is certainly difficult and restricted to non-extreme states, not least for ethical reasons. However, a number of ingenious methods have been developed for emotion elicitation. In a review of experimental inductions of emotional states, Astrid Gerrards-Hesse, Kordelia Spies and Friedrich Hesse (1994) collected evidence from a large number of studies inducing elation and depression in the laboratory. They distinguished five groups of what they call mood induction procedures (MIPs):

1. MIPs based on the free mental generation of emotional states, including Hypnosis and Imagination, where subjects are instructed to imagine and re-experience situations or events;

2. MIPs based on the guided mental generation of emotional states, including the Velten method ("self-referent statements describing positive or negative self-evaluations and bodily sensations", (Gerrards-Hesse et al., 1994, p. 56)), as well as the Film/Story+ and the Music+ method (where the '+' indicates the explicit instruction to "get into" the corresponding emotional state);

3. MIPs based on the presentation of emotion-inducing material, including the Film/Story and Music methods, this time without the instruction to feel involved

with the material presented, as well as the Gift method, where the subject is offered an unexpected gift;

4. MIPs based on the presentation of need-related emotional situations, including the Success/Failure method where subjects are given false-positive or false-negative feedback concerning their performance in a test, and a Social Interaction method; and

5. MIPs aiming at the generation of emotionally relevant physiological states, including the Drug method where subjects are injected a drug such as epinephrine or a placebo, and the Facial Expression method based on the facial feedback hypothesis (see 1.3.2, p. 12).

Comparing the methods' effectiveness, Gerrards-Hesse et al. concluded that the Film/Story and Gift MIPs were most effective in inducing elation, while the Imagination, Film/Story, Success/Failure and Velten MIPs were most effective at inducing depression. They noticed that the classification presented above, based on methodological principles, was not reflected in the effectiveness:

> Interestingly, ... a MIP's effectiveness does not covary with the classification into a certain group. Apparently other characteristics of a MIP beyond the kind of stimuli used to influence subjects or the announcement of the purpose of emotion induction are responsible for the degree of effectiveness. (Gerrards-Hesse et al., 1994, p. 69)

Few studies in the field of speech and emotion seem to have used induction techniques for emotion elicitation. Among the few are Amir et al. (2000), who have used an Imagination task, and Johnstone & Scherer (1999), who have used a computer game. Both are speaker-centered studies (see 3.2, p. 37).

A social interaction scenario was used by Roland Kehrein (2002), who elicited spontaneous emotion using a collaborative task. Two subjects were to collaboratively build a Lego item, communicating only via their voices. One of the subjects indicated what to build by spoken descriptions and commands. Emotionally significant situational factors were introduced by the experimenter in a controlled way, e.g. through manipulation of the Lego kit, the time allowed, or the expectation of success. The two subjects were seated

in two different soundproof rooms and communicated with each other via microphones and headphones, which made it possible for their speech to be recorded in optimal quality even during overlapping passages.

4.4 Natural occurrences

A maximally natural but uncontrolled type of emotional speech occurs in spontaneous human interaction. The study of naturally occurring emotional speech seems to gain importance with the increasing prospects of technical applications.

One study working with naturally occurring emotional speech was conducted by Klaus Scherer, Bob Ladd and Kim Silverman (1984). They recorded interviews between social agency workers and actors mimicing the clients. Only the social agency workers' speech was used (see also Chapter 5 for an extended discussion of that study).

Two recent English language database projects aimed at collecting spontaneous emotional speech: The Emotion in Speech project in Reading and Leeds (Roach et al., 1998; Roach, 2000; Greasley et al., 2000; Stibbard, 2001) pointed out the complexity of the task. In the project, naturally occurring emotional speech was recorded from television and radio documentaries as well as from video diaries. 4 $\frac{1}{2}$ hours of speech material were collected. A detailed transcription system for prosodic and paralinguistic features of the emotional speech was developed (Roach, 2000), and the perceived emotions were obtained in free response as well as forced choice perception tests (Greasley et al., 2000). According to Stibbard (2001), the project did not find systematic correlations between emotions and paralinguistic annotations, because it was not possible "to devise a system for the classification of the emotion variables which was on the one hand both sufficiently fine-grained and sufficiently systematic ... while on the other hand remaining sufficiently broad that replication might be achieved." (Stibbard, 2001, p. 204) Stibbard also identified a problem concerning the paralinguistic annotations, namely that the "large quantity of annotation types rendered analysis problematic" (Stibbard, 2001, p. 205). The difficulty in finding systematic patterns seems to be linked to the fact that very detailed descriptions were used.

The Belfast Naturalistic Emotion Database (Douglas-Cowie et al., 2000; Cowie et al., 2001) consists of recordings of spontaneously occurring emotion from interviews and

from TV programs. For each of 124 speakers, one or more emotional clips as well as a relatively neutral reference clip were included. The emotionality was assessed by dimensional ratings in activation-evaluation space, using the Feeltrace tool (see 6.4.2, p. 70), as well as by verbal labels. Acoustic properties were determined using a semi-automatic analysis tool basically measuring gradual acoustic properties. See also 11.2, p. 104 for a more extensive description of the Belfast database. The investigation of correlations between the dimensional ratings and the acoustic measures is part of the practical work carried out in this thesis, see Chapter 11.

A corpus of Chinese conversational speech was collected by Li-chiung Yang and Nick Campbell (2001). They noted that their reason for studying spontaneous speech was that "it is only in such speech that we will encounter the complex emotions occurring in real life". In the corpus consisting of six hours of recorded conversational speech, they identified stretches where emotions and emotion-related states were present, and collected evidence for typical prosodic phenomena associated with these states.

4.5 Synthesised speech

Finally, a method may gain importance which allows very close control, namely the use of synthesised speech. In terms of the Brunswikian lens model (Figure 3.1, p. 36), the use of synthesised speech skips the speaker state, and directly operates on the level of the distal indicator cues. This method is of course not appropriate for speaker-centered studies, but it may become increasingly useful in listener-centered studies concerned with the perceptual effect of specific speech parameters. In particular, when specific hypotheses have been developed, e.g. in a database inspection, they can be tested using speech synthesis. The technique allows the modification of a single parameter, all other things being equal.

An interesting example of a question which could be investigated using this technique is the often-reported influence of loudness or intensity on emotion recognition (Murray & Arnott, 1993; Cowie et al., 2001). It can be hypothesised that listeners might not actually interpret emotion on the basis of the absolute intensity, but rather on the basis of voice quality changes linked to speaker effort (Eriksson & Traunmüller, 1999; Holmberg et al., 1988; Lehiste & Peterson, 1959). If so, then changing the loudness of a synthesised utterance while keeping the voice quality constant should have a smaller perceptual effect than changing the voice quality according to vocal effort while keeping the absolute loudness constant.

4.6 Summary

This chapter has given an overview of the large variety of means which have been used for obtaining recordings of emotional speech. These methods vary in the extent to which they allow for control over the speech signal: from speech synthesis allowing the closest control, via actor portrayals of standard sentences, reading of emotional texts, to emotion elicitation and observation of natural speech which give the researcher little control over the speech produced.

It was mentioned that these methods are suitable for different research orientations. In particular, studies using acted material often investigate fullblown emotions rather than low-intensity underlying emotions, which are typically found in naturally occurring material. Speaker-centered studies often work with emotion elicitation, while listener-centered studies can benefit from the control provided by speech synthesis techniques.

Chapter 5

Speech parameters expressing emotion

> During evolution, language and speech were superimposed on a primitive, analog vocal signaling system. Because speech uses the same voice-production mechanism and many of the same acoustic features as the more primitive nonverbal system, we find an intriguing intermeshing of verbal and nonverbal aspects in human sound production. (Scherer, 1982, p. 138)

As becomes clear from Klaus Scherer's statement, identifying the parameters through which emotions are expressed in speech is a complex task. Non-verbal emotion expression is part of the "primitive analog vocal signaling system" onto which speech was superimposed during evolution. Many acoustic parameters carry information either according to the non-verbal system or according to a linguistic function. An investigation of these parameters must take that multi-functionality into account.

It is important to note that the particular parameters which should be investigated are by no means clear. The fact that many studies investigate similar parameters may be linked to their technical accessibility rather than theoretical motivations. This becomes clear by looking at the example of voice quality, which is considered important by most researchers in speech and emotion (e.g., Roach, 2000; Cowie et al., 2001; Montero et al., 1999a), but which is often not measured due to the technical difficulties involved.

It should also be mentioned that speaker-centered and listener-centered studies may not be interested in the same parameter set: Those parameters in which emotion-related variation is reliably detected by humans may be difficult to capture using current speech

analysis algorithms, and vice versa (see the hypothesised role of voice quality vs. loudness, p. 48).

While the most commonly studied types of effects concern gradual, para-linguistic influences of emotion on prosody, a few studies exist which demonstrate the existence of other types of voice use, such as a configurational use of linguistic categories, or affect bursts. They will be reviewed briefly in this chapter.

5.1 Gradual "para-linguistic" use of prosody

Most studies into the speech correlates of emotions have investigated global, gradual effects on prosodic settings, such as F0 mean and range, overall speech tempo, and mean intensity. This type of voice use has been called "para-linguistic", and the influence of emotions on the voice parameters was said to be following the "covariation model" (Scherer et al., 1984).

The studies investigating this type of vocal emotion effects are too numerous to be listed here, and there does not seem to be an urgent need for this, given the existence of several thorough reviews of the literature on vocal emotion expression (Scherer, 1986; Murray & Arnott, 1993; Cowie et al., 2001; Stibbard, 2001).

However, two important recent studies are described here in order to illustrate their methods. Rainer Banse and Klaus Scherer (1996) carefully selected well-recognised actor portrayals for 14 emotions (see Table 2.1, p. 21, for a list). Actors produced each of the 14 emotions with two pseudo-sentences consisting of logatomes. Each emotion was defined by two different scenarios, or frame stories. In a pre-selection step, experts rated actor portrayals for authenticity and recognisability; in a second step, the best-rated portrayals were presented to subjects in a forced-choice recognition test. The mean recognition rate was 55%, clearly above chance level. Recognition rates served as a further exclusion criterion, such that only the best-recognised portrayals of each emotion were acoustically analysed. The parameters calculated in the acoustic analysis were utterance-global. They included F0 mean, standard deviation, and quartiles; mean energy; speech rate measured by the duration of articulation periods and by the duration of voiced periods; and a number of frequency bands describing, separately, the voiced and the unvoiced long-term average spectrum. In a multiple regression analysis, these acoustic parameters were correlated

to the independent variables sentence, sex of actor, actor identity, emotion and scenario. Banse and Scherer summarise: "After the variance accounted for by sentence type, gender of actor, and idiosyncratic speaker differences is removed, emotion still explains a large and highly significant proportion of variance in the majority of the acoustic variables. It amounts to 55% for mean energy and to 50% for mean fundamental frequency." (Banse & Scherer, 1996, p. 623)

Lea Leinonen et al. investigated the expression of emotions with a single-word utterance (Leinonen et al., 1997). Actors produced the Finnish name "Saara" with ten "emotional-motivational connotations", defined by frame stories: Naming, astonished, pleading, content, admiring, sad, commanding, scornful, frightened, and angry. After "poor" examples had been discarded in a pre-selection, a perception test and acoustic analysis were carried out using the remaining examples. In the perception test, the mean recognition rate was 50%. Due to the short duration of the utterance, acoustic analysis could examine individual segments, as well as differences between them, for duration, F0 level and range, and sound pressure level. In addition, visual evaluation was carried out for sound pressure curves, F0 contours, and spectral models using self-organising Kohonen neural-net maps. The measured acoustic parameters were related to the intended emotions.

In addition, the study by Leinonen et al. is one of the few which also explicitly consider listener-centered aspects (another example being Heuft et al., 1996), by relating the acoustic measures to the emotion categories chosen by the listeners, thus addressing "listener expectations", or the "perceptual representation and inferential utilisation" in terms of the Brunswikian lens model (see Figure 3.1, p. 36).

Most studies solely employ speaker-centered methods (see 3.2, p. 37) for analysing speech parameters, i.e. they measure speech parameters as a function of produced emotion. As a result, they collect information about the expression or "externalisation" aspect (see the Brunswikian lens model, p. 36).

5.2 Categorical "linguistic" use of prosody

Apart from gradually changing parameters such as mean F0 level, configurations of linguistic categories can play a role in emotion expression and recognition. Accordingly,

this type of voice use for emotion expression was called a “configuration model” (Scherer et al., 1984). Studies investigating the use of prosody according to the configuration model are much scarcer than studies presupposing the covariance model.

Maybe the first study directly investigating the contribution of covariance and configuration factors in emotion perception was conducted for German by Klaus Scherer, Bob Ladd and Kim Silverman (1984). In perception tests using spontaneously produced material from social agency worker interviews, they found evidence for both the covariance and the configuration model. Evidence for the latter consisted in an interaction between sentence type (wh-question vs. yes/no-question) and utterance-final F0 contour type: only in conjunction with yes/no-questions were final falling contours perceived as reproachful and aggressive.

In a study investigating the linguistic function of intonation, emotional meaning was found as a “side effect”. Bistra Andreeva and William Barry investigated the role of the intonation contour in distinguishing between checks (confirmation questions) and statements in Bulgarian, using sentences in which sentence mode was not syntactically marked (Andreeva & Barry, 1999). A hypothesis for the typical question and statement contours was derived from a corpus of spontaneous speech. These contours were presented, in a factorial setting, in contexts indicating questions and statements, respectively. Subjects indicated on a scale how good the utterance fitted with the context. Among other things, Andreeva and Barry found that the question intonation, presented in a context requiring a statement, was accepted as a statement, but subjects remarked that the speaker sounded angry or unhappy (Andreeva & Barry, 1999, p. 8).

Sylvie Mozziconacci investigated the influence of intonation contours on perceived emotion for Dutch (Mozziconacci, 1998). In a series of perception tests, she found, among other things, that stimuli containing certain intonation contours were judged more frequently as some emotions than others. This allowed her to formulate “recommendations” of intonation contour types to use and to avoid in order to cause the perception of certain emotions.

Richard Stibbard, in an analysis of the Reading/Leeds corpus of spontaneous emotional speech (Stibbard, 2001, Chapter 5), searched for systematic co-occurrences of ToBI accents and terminal contours with emotions. The only systematic effect he found was that the low accent tone, L*, occurred twice as often with sad speech than with the other

emotions. It would seem that if emotion-specific effects existed, the broad statistical analysis, on the limited data base available, was not able to discern these effects. As Stibbard points out himself, some factors known to affect intonation were not controlled for, such as regional variation between speakers. It would also have been desirable to take sentence structure into account – apparently, no attempt was made to distinguish between statements and questions, for example.

These studies show the challenge involved in identifying configurational effects of intonation on emotion. The three studies by Scherer et al. (1984), Andreeva & Barry (1999) and Mozziconacci (1998), who used carefully selected or resynthesised material in closely controlled perception settings, found effects of intonation on emotion recognition, the first two in interaction with sentence type. The statistical analysis carried out by Stibbard (2001) apparently faced too many confounding factors in the data, so that nearly no systematic effects could be found.

A more promising approach to finding emotional effects of intonation contours might be that of Douglas-Cowie as well as of Yang and Campbell (see 3.3, p. 38), namely an informal investigation of the natural speech data in order to identify individual occurrences of emotional material. On the basis of these observations, hypotheses can be built regarding the specific parameters expressing emotions. Such hypotheses can subsequently be investigated, most easily in a listener-centered approach, by selecting appropriate naturally occuring speech material (Scherer et al., 1984) or by generating appropriate material using synthesis or re-syntheis techniques (Andreeva & Barry, 1999; Mozziconacci, 1998).

5.3 Affect bursts

An interesting type of voice use for emotion expression is what Klaus Scherer called “affect bursts” (Scherer, 1994). He defined them as “very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events” (Scherer, 1994, p. 170). This comprises both physiological “raw affect bursts” (or primary interjections in the vocabulary of Kainz (1940)), dominated by push effects, and quasi-verbal “affect emblems” (or secondary interjections), dominated by pull effects (for a description of push and pull effects, see 7.2, p. 74).

I have conducted an experimental study of affect bursts and affect emblems (Schröder,

2003). In the first part of the study, the recognition of emotions from affect bursts was addressed. A list of affect bursts was compiled from a variety of sources, and a tentative list of 10 emotion categories was established which seemed to be typically expressed through the affect bursts in the list: admiration, threat, disgust, elation, boredom, relief, startle, worry, contempt, and hot anger. Two affect bursts were selected from the list for each of the 10 categories, and produced by actors, along with an affect burst spontaneously associated to each emotion category by each actor. The recordings were pre-selected in an expert rating task (see 6.1.1, p. 66), and the best-rated versions in each category were used as stimuli in a forced-choice perception test. The mean recognition rate was found to be 81%, which is clearly higher than the 50–60% usually obtained in emotion expression through speech prosody (see 5.1). When affect bursts within an emotion category were grouped into "affect burst classes" on the basis of phonetic similarity, it became evident that many affect burst classes were recognised with more than 90% accuracy.

In the second part of the study, the role of the segmental form of affect burst classes was investigated. Transcribed forms of the affect burst classes were presented to subjects, in a forced-choice test using the same 10 emotion categories as before. Most affect burst classes were found to be clearly identifiable based on their segmental transcription; however, some affect burst classes were not identified from the transcription alone, indicating a crucial role of prosody and voice quality in their recognition. Transcription variability between non-expert transcribers was proposed as an experimental criterion for the gradual distinction between physiological "raw affect bursts" and conventionalised "affect emblems". Conventionalised emblems, I argued, should correspond to a mental reference pattern influencing perception towards a standardised form, thus leading to a lower transcription variability than for unconventionalised raw bursts. On the basis of this criterion, candidates for typical emblems and raw bursts were proposed.

Affect bursts may play a more important role in expressing some emotions than others. Disgust is the prime example of an emotion that appears to be typically expressed through affect bursts. In Banse & Scherer (1996), the recognition rate for disgust, when expressed through prosody and voice quality accompanying speech, is particularly low (15%). The possible explanation given by Banse & Scherer (1996) is that disgust is not likely to be expressed through long utterances, but rather through short affect bursts. The results in Schröder (2003), with disgust being conveyed through affect bursts with 93% accuracy, lend support to that assumption.

5.4 Evolutionary perspective: Frequency code and other codes

According to the push-pull distinction (see 7.2, p. 74), two different mechanisms of evolutionarily shaped emotion expressions can be distinguished: Those that "leak out" as by-products of physiological adaptation processes (push effects), and those for which the expression itself has a survival-benefiting function by influencing a receiver.

John Ohala has proposed that voice F0 is used within the latter mechanism. He called the phenomenon "frequency code" (Ohala, 1994). The idea is that low F0, which physically tends to be associated with large individuals, is used to convey the *impression* of a large signaller. Accordingly, low F0 and rough voice is used by speakers to convey the impression of, and perceived by listeners as denoting, a strong, large, and self-confident speaker. Correspondingly, high F0 and melodious voice signal small, non-threatening individuals, maybe reflecting infant-mimicry. Ohala presents diverse pieces of evidence from ethology and cross-language studies to support his proposal.

Recently, Carlos Gussenhoven (2002) proposed that two related codes evolved from speaker effort and vocal production constraints, which he named "effort code" and "production code", respectively. The natural phenomenon underlying the effort code is that "increases in the effort expended on speech production will lead to greater articulatory precision, but also a wider excursion of the pitch movement." (Gussenhoven, 2002) Exploiting this phenomenon in terms of a code, speakers can use, and listeners can understand, high F0 and precise articulation as signalling high effort, and consequently importance or agitation. In a similar fashion, the production code exploits the fact that empty lungs lead to low energy and low F0 and consequently uses these features to signal finality. Gussenhoven proposes, among other things, an affective interpretation of the effort code (but not of the production code): 'Surprise' as well as 'helpfulness' are affective meanings expressed through the effort code. As to the latter, by using the effort code (e.g. by making large pitch movements), the speaker indicates to the listener that he is actually making an effort to help. It is a signal frequently found in speech directed towards children.

5.5 Emotion expression in speech and other channels

Emotion expression typically occurs not only through the voice, but also through verbal content, facial expression, and body language. This section presents some of the few

studies investigating such joint effects, and formulates potentially relevant issues in this largely unexplored field, distinguishing production and perception aspects.

5.5.1 In production

In introducing the concept of affect bursts (see 5.3), Klaus Scherer stresses the joint expression of a given emotion through face and voice, stating: "The most important aspect of this co-occurrence of facial and vocal activity in what I call *affect bursts* is the *integration and synchronization* of the two expressive modalities" (Scherer, 1994, p. 180). In Scherer's model, such integration and synchronisation is due to physiological factors building up over time as consequences of a sequence of stimulus evaluation checks (see p. 13). Both modalities are expected to be influenced simultaneously by these physiological effects. Voluntary simulations, as in actor portrayals, are likely not to show the same patterns of integration and synchronisation.

Similarly, interaction patterns between voice prosody and other channels, such as text structure, are to be expected. It is likely that the type of utterance produced depends on the emotion expressed, including sentence and phrase length. If that is so, then the prosodic phenomena accompanying these emotion-specific phrasing patterns are lost in a constant-text setting, which is typical for many actor-based studies investigating the prosodic effects of emotions (see 4.1, p. 43). The consequence of studying prosody in an "all other things being equal" setting is that interactions of prosody with these "other things" are excluded from the study. While of course such studies are necessary and perfectly valid, it seems important to keep in mind that they investigate only a subset of potentially relevant phenomena.

5.5.2 In perception

The interpretation of prosody in terms of the emotion expressed has been shown to depend at least in part on other modalities and on situational factors.

With regards to the verbal channel, this question was examined by Iain Murray and John Arnott (1995) in the evaluation of their HAMLET system of emotional speech synthesis. They combined neutral and emotional texts with neutral and emotional prosody[1],

[1] For stimuli with emotional text and emotional prosody, the same emotion was expressed through both channels; no contradictory stimuli were created.

and presented the four types of stimuli in perception tests. The subjects were asked to base their judgment on the voice only, not on the words spoken. The recognition results were analysed separately for each of the four sets. After that, for each type of verbal content (neutral vs. emotional), the *improvement* in recognition due to emotional prosody was calculated. A clear tendency was found that the perceptual effect of emotion-specific prosody increased when the verbal content matched the expressed emotion; in other words, they observed a synergetic effect of prosody and verbal content on emotion recognition, their joint effect being larger than the sum of the two individual effects. At the same time, the fact that the text had an influence despite the instruction to base the judgment on the voice only, shows the unavoidability of processing the two aspects together.

The same idea was followed by John Stallo (2000), also in the context of speech synthesis. In a way similar to Murray and Arnott, he combined emotionally neutral and emotionally expressive verbal contents with neutral and emotional prosody, in a forced choice test. In a different test, he also investigated multimodal effects (Stallo, 2000). He used a talking head combining facial emotion expression with either neutral or emotional synthetic speech, and presented the two versions in a preference task (see 6.2, p. 67). The version where face and voice both expressed the emotion was very clearly preferred.

Richard Cauldwell impressively demonstrated the context-dependency of vocal emotion perception (Cauldwell, 2000). A stimulus was perceived as expressing anger when presented in isolation, but as emotionally neutral when presented in context (a father talking to his child). The effect persisted even if the stimulus was first heard in context and then heard in isolation.

Klaus Scherer, Bob Ladd and Kim Silverman showed that certain prosodic means of emotion expression, such as terminal contour direction, only have an effect in conjunction with certain linguistic parameter settings such as question type (Scherer et al., 1984). As was mentioned above (see 5.2), they found downward terminal contours for questions to be perceived as impolite only for yes-no questions, not for wh-questions.

Dominic Massaro investigated audio-visual emotion perception uing a framework developed for audio-visual speech perception (Massaro, 2000). In each of two modalities, a continuum was created between two categories (e.g., anger and happiness expressions). In a factorial design, all combinations of settings in the two modalities were presented in a two-alternative forced choice perception test. The so-called fuzzy logical model of per-

ception (FLMP) was used for modelling the perceptual process, distinguishing between the modality-specific evaluation of the stimulus, the integration of the modalities, and the decision based on a comparison of the available alternatives. Once the model is fitted to the data, it can be used to predict the perceptual effect of a given stimulus. Massaro applied this technique to facial and vocal expressions of anger vs. happiness. Facial emotions were expressed using the talking head Baldi; vocal emotions were obtained by recording a human speaker. In each modality, the "continuum" consisted of three steps, happy, neutral, and angry. Both modalities were shown to contribute to the recognition, with the facial expression being more important than the vocal expression.

Beatrice de Gelder and Jean Vroomen investigated the integration of facial and vocal information in emotion perception (de Gelder & Vroomen, 2000). They created a continuum between two photographs of faces expressing happiness and sadness, using morphing techniques. Each face was presented, in a factorial design, paired with happy, sad, and no voice. Subjects indicated whether they perceived happiness or sadness. In a first experiment, subjects were instructed to attend both to the face and the voice; in the second, they were told to ignore the voice and to attend only to the face. In both experiments, both the face and the voice had an effect on responses as well as on response timing: Subjects responded more slowly to ambiguous stimuli. In the condition where subjects were told to ignore the voice, the effect of the voice was smaller but significant, indicating the partly non-voluntary nature of the integration of the two modalities. In a third experiment, subjects were presented with a vocal continuum between happiness and fear, and happy or fearful faces. They were told to ignore the faces and judge the perceived emotion solely based on the voice. Again, a small but significant effect of the facial emotion was found in addition to the vocal emotion.

Dominic Massaro and Michael Cohen analysed de Gelder and Vroomen's results using the FLMP (Massaro & Cohen, 2000). After fitting the model to the average results of de Gelder and Vroomen's perception test, they could show that the FLMP makes good predictions of subjects' perceptual performances. The smaller influence of the modality which subjects had been told to ignore is reflected in FLMP parameter values assigning less weight to that modality. Apart from that, the same basic integration mechanism, as presumed by the FLMP, can account for the observed perception patterns.

These few examples are intended to demonstrate the importance of the effects stem-

ming from interaction between speech and emotion on the one hand and other channels of emotion expression as well as situational context on the other hand. It is probable that convincing perceptual effects, e.g. in an emotionally expressive computer interface, can only be obtained when these effects are appropriately taken into account. At present, most of this field is largely unexplored.

Chapter 6

Perception test methodologies

The capability of listeners to recognise emotion from speech is assessed through perception tests. This chapter discusses the methodologies used in speech and emotion research, as well as some methodologies from related domains which might be interesting to use in this field.

6.1 Identification tasks

The perception test paradigm which seems to be most widespread in the speech and emotion community is the use of forced-choice identification tasks. A number of stimuli are presented to a listener in randomised order, and each stimulus has to be assigned to one (and only one) of a number of emotion categories usually identified by single-word labels.

Forced-choice identification tasks are applied in many small studies, e.g. evaluating emotional speech synthesis (see also 9.4, p. 93), but also in recent larger studies (Banse & Scherer, 1996; Leinonen et al., 1997). A major pre-requisite for their use is the control over the channels carrying emotional information. In the simplest case, this means that the parameters in the channel under study are systematically varied, while all other channels convey no emotionally relevant information. In particular, the verbal content must not carry emotional meaning. It must either be held constant (the neutral carrier sentences often used in actor portrayals) or masked through filtering or other methods (e.g., Scherer et al., 1984). It should be noted that it is difficult to isolate individual vocal parameters in this way. With recorded emotional speech, it is difficult at best to separate voice quality,

F0 level, F0 range, speech rate etc.; only with speech synthesis techniques is it possible to modify these parameters individually in a largely controlled way.[1] Therefore, forced choice recognition tasks employing recorded speech usually assess the overall effect of all vocal parameters taken together.

A number of measures are usually employed for presenting the results of a forced-choice identification task: The *recognition rate*, i.e. the percentage of answers "as intended" in a given category, provides a simple measure of success. The recognition rates are compared to *chance level*, i.e. the percentage of correct answers which would be obtained when each category was chosen with the same probability.[2]

Apart from the "correct" answers (recognition as intended) measured by recognition rates, it is also interesting to look at the "wrong" answers (confusions). These inform about perceptual similarities, either on the level of the "proximal percepts" (see Figure 3.1, p. 36), or on the level of the emotion concepts themselves. The former corresponds to a situation where two different emotion categories lead to very similar proximal percepts; in the latter case, the proximal percepts might be very different, but the emotion categories attributed by the listener on their basis are conceptually similar.

The method for investigating perceptual similarity in forced-choice paradigms is the use of a *confusion matrix*. It informs about whether confusions were symmetrical, or whether some categories were more attractive than others, such as in cases where the "neutral" category seems to serve as a default answer when stimuli are difficult to classify (Edgington, 1997).

It may be worth mentioning that the original version of the circumplex model (see 2.5, p. 25) was proposed by Harold Schlosberg (1941) on the basis of the confusion matrix of a facial expression categorisation task. He noticed that when he ordered his emotion categories in a specific way, the upper and lower secondary diagonals (corresponding to confusions between neighboring categories) contained the vast majority of the "false" responses, thus indicating a perceptual similarity between these neighboring categories.

Several shortcomings of forced-choice identification tests have been noted. Rainer Banse and Klaus Scherer pointed out that if only a small number of categories is involved,

[1] It must be kept in mind, however, that depending on the synthesis technique, parameters are not fully independent. For example, the signal manipulation algorithms employed in concatenative (diphone) synthesis may introduce unwanted distortions. In particular, methods for modifying fundamental frequency, which alter the duration of a period by shortening the low-energy *open phase* of the period, inevitably modify the *open quotient*, which is a measure of voice quality (Ní Chasaide & Gobl, 1997).

[2] The chance level, therefore, is 100% divided by the number of categories presented to the listener.

judges are likely to *discriminate* between the possible answers rather than *identifying* the category: "It is doubtful whether studies using 4–6 response alternatives in a vocal emotion recognition study actually study recognition or whether, more likely, the psychological process involved is *discrimination* among a small number of alternatives ... using exclusion and probability rules" (Banse & Scherer, 1996, p. 615).

A second, maybe more serious problem, related to the first one, is the impossibility to account for perceptual categories not represented in the answer set. In a pilot study (Schröder, 1999a,b), I copy-synthesised four emotion categories (anger, fear, happiness, sadness) and a neutral speaking style using specially created diphone material, and presented them in a forced choice perception test. Subjects remarked that they would have wanted the answer category "disappointment". A subsequent free-response test revealed a tendency towards "disappointment" answers for most stimuli, possibly due to the particular text or diphone voice quality employed. The effect was relatively independent of the prosodic parameters modelled. In strict forced-choice perception tests, such unexpected effects potentially remain undiscovered.

An alternative to free response tests for discovering at least some non-intended categories is the introduction of "distractor" categories in the response set. This method was originally employed in the Fairbanks studies (Fairbanks & Pronovost, 1939; Fairbanks & Hoaglin, 1941), and more recently in studies by Murray & Arnott (1995) and Stallo (2000).

The limitations of identification tasks in accounting for the precise emotional state perceived by listeners was demonstrated by Peter Greasley, Carol Sherrard and Mitch Waterman (2000). They presented audio clips of naturally occurring emotional speech both in a free response perception test and in a forced choice test with five basic emotion categories (anger, disgust, fear, sadness, happiness). The clips included the original, possibly emotion-specific verbal content. In the forced choice test, agreement between subjects was limited: While very high agreement was found for clips rated as "happiness" (the only "positive" emotion category provided), only a minority of the "negative" stimuli were "pure" in the sense that they were identified as one given emotion category by a statistically significant majority of the listeners. The authors argue that the remaining, "mixed" stimuli could only have been described appropriately if several answer categories had been allowed. For the "pure" emotions, 70% of the free-response answers were representative of the chosen basic emotion, but differing in quality and intensity.

6.1.1 Pre-selection

Depending on the research question, it may be desirable to use pre-selected speech material for acoustic analyses and/or perception tests.

If the goal is the study of "good examples", i.e. optimal or best-case scenarios, then a careful selection of speech samples is necessary. This goal seems appropriate for studying the acoustic correlates of fullblown emotions (Banse & Scherer, 1996) and for establishing the communicative power of a given expressive channel, as I have done for affect bursts (Schröder, 2003). It seems evident that no claim regarding typicality and generalisability can be made on the basis of such a study; rather, it seems analogous to a "proof of existence" in mathematics. A high recognition rate on such material proves that stimuli exist which are well-recognised. A low recognition rate, on the other hand, is not a proof for the non-existence of well-recognised stimuli, but only indicates that in that particular study, and despite efforts to select good examples, no such examples could be found.

A very clear example of pre-selection methodology is the work of Banse & Scherer (1996). After obtaining 1344 emotion portrayals from professional actors, they had experts (advanced students from a professional acting school) rate the portrayals according to recognisability and authenticity, based on the audio, visual, and combined audio-visual channel. The items were presented grouped per emotion, and individually rated using one scale for authenticity and one for recognisability. Based on these ratings, 280 items were chosen for use in the perception test and later in the acoustic analyses.

6.1.2 Interactions between different channels

The single-channel approach described above investigates the capability of the vocal channel to express emotion by itself. A different focus lies in the question how the vocal channel can *interact* with an emotional message expressed via a different channel.

A *factorial design* is suited for addressing this type of question: Several parameters are modified in a controlled way independently of one another. As each setting in one parameter is combined with each possible setting in the other parameters, the number of stimuli becomes very big when more than three or four parameters are included in the study. The method has the advantage of allowing the study of interactions between parameters without requiring a priori hypotheses. The requirement for precise control in each of the parameters is the same as for studies of individual parameters.

If the contributions of the various parameters involved in speech prosody are to be studied, speech synthesis techniques are well suited for this task (the study by Burkhardt & Sendlmeier (2000) is a good example, see 3.4, p. 39).

Interactions with other channels have also been studied, including verbal content (Murray & Arnott, 1995; Stallo, 2000), situational context (Cauldwell, 2000), linguistic sentence type (Scherer et al., 1984), and facial expression of emotion (de Gelder & Vroomen, 2000; Massaro, 2000; Massaro & Cohen, 2000). Section 5.5.2, p. 58 presents short summaries of these studies.

6.2 Preference tasks

Preference tasks have the advantage of measuring the type of information that forced choice tests obscure: The question whether, and to what extent, the stimuli sound natural and convincing. Although preference tasks do not seem to be generally used in speech and emotion research, it is conceivable that they might be useful in certain conditions. In this kind of task, subjects are asked for overall preference, or more specifically naturalness, intelligibility etc. of stimuli presented either one by one or in pairs. If stimuli are presented individually, scales can be used for their evaluation; in pairwise presentation, one of the stimuli is chosen over the other one. This latter method is likely to allow more fine-grained distinctions due to the presence of a reference stimulus.

In the context of multi-channel emotion expression, one or more channels for which appropriate emotion-specific configurations are known can serve as an emotion-defining frame. In order to find appropriate parameter settings in an additional channel, parameters can be varied in that channel (e.g., prosody) and combined with the emotion-defining frame consisting of the fixed settings in the other channels (e.g., verbal content, facial expression). The parameter setting perceived as most natural in a preference task corresponds to the perceptually optimal setting in the tested channel with respect to the reference provided by the other channels.

A first step in this direction was made by John Stallo in his Honour's thesis (Stallo, 2000). He combined synthesised speech with emotion-specific facial expressions in a talking head. An expressive passage of Lewis Carroll's "Alice's Adventures in Wonderland"

was rendered in two versions: Manually tuned facial expressive behaviour of the talking head was accompanied, in the first version, by neutral synthetic speech, and in the second, by expressive synthetic speech with prosody manually tuned through speech synthesis markup. The stimuli were presented in pairs and rated by the subjects for understandability, expressiveness, naturalness, and interest. On all scales, the version combining facial expressivity with expressive prosody was very clearly preferred over the version with facial expressivity and neutral prosody.

6.3 Similarity judgments

Similarity judgments can be used to gain information about the criteria by which people judge similarity without asking them to make this knowledge explicit (which possibly they cannot). The technique has been used, e.g., with emotion words (Russell, 1980; Scherer, 1984a) which were sorted into groups based on similarity, but also with emotional speech stimuli (Green & Cliff, 1975) which were presented pairwise, and for which the subjects indicated the dgree of similarity. Subsequent multi-dimensional scaling determines an arrangement of the tested items in an n-dimensional space which best approximates the similarity measures obtained. In the case of Russell's and Scherer's work with emotion words, a two-dimensional space was best suited for representing the data. In the case of Green and Cliff's work with emotional speech, a three-dimensional space was considered appropriate. After the scaling procedure is completed, the experimenter can then try to identify a meaning in the dimensions.[3]

While the method seems not to have been used extensively with vocal emotion expression, with the notable exception of Green & Cliff (1975), it seems potentially useful for establishing, without theoretical prejudices, the most relevant features leading to perceptual similarity among emotional speech stimuli.

6.4 Placement on emotion dimensions

While preference and similarity judgment tasks do not presuppose a particular descriptive framework, category identification tasks seem most appropriate for category-based emotion descriptions. Similarly, some test paradigms seem most appropriate for dimensional emotion descriptions (see 2.6, p. 26).

[3] This is one of the methods leading to the proposal of emotion dimensions, see 2.6, p. 26.

	Pleasure	
Happy	___:___:___:___:___:___:___	Unhappy
Pleased	___:___:___:___:___:___:___	Annoyed
Satisfied	___:___:___:___:___:___:___	Unsatisfied
Contented	___:___:___:___:___:___:___	Melancholic
Hopeful	___:___:___:___:___:___:___	Despairing
Relaxed	___:___:___:___:___:___:___	Bored
	Arousal	
Stimulated	___:___:___:___:___:___:___	Relaxed
Excited	___:___:___:___:___:___:___	Calm
Frenzied	___:___:___:___:___:___:___	Sluggish
Jittery	___:___:___:___:___:___:___	Dull
Wide-awake	___:___:___:___:___:___:___	Sleepy
Aroused	___:___:___:___:___:___:___	Unaroused
	Dominance	
Controlling	___:___:___:___:___:___:___	Controlled
Influential	___:___:___:___:___:___:___	Influenced
In control	___:___:___:___:___:___:___	Cared-for
Important	___:___:___:___:___:___:___	Awed
Dominant	___:___:___:___:___:___:___	Submissive
Autonomous	___:___:___:___:___:___:___	Guided

Table 6.1: Semantic differential scales for measuring emotion dimensions, as proposed by Mehrabian & Russell (1974, Appendix B, p. 216).

6.4.1 The semantic differential

Scale judgments according to the semantic differential (Osgood et al., 1957) are widespread in the psychological literature (e.g., Mehrabian & Russell, 1974; Bradley, 1994). The method assesses perceived emotional content by means of a set of paired adjective scales. Albert Mehrabian and James Russell proposed a set of six scales for each of their three emotion dimensions (Mehrabian & Russell, 1974), which are reproduced in Table 6.1.

The use of semantic differential measures for emotional speech or, more generally, sounds, does not seem to be very wide-spread. Margaret Bradley (1994) briefly reports on a system of International Affective Digitized Sounds (IADS) developed by Peter Lang and herself at the University of Florida, a collection of "environmental sounds, including people laughing, crying, screaming, and coughing, as well as sounds from animals, machines,

sports events, household objects, and so forth" (Bradley, 1994, p. 100). These sound files were rated for their pleasantness, arousal, and dominance.

In the domain of speech and emotion, Rex Green and Norman Cliff (1975) had their emotional speech stimuli rated on seven adjective scales, in which a principal component analysis identified two factors, the first "highly related to the thin-thick and high-low pitch scales", the second "highly related to the pleasant-unpleasant and warm-cold scales" (Green & Cliff, 1975, p. 433).

More recently, I have used three scales representing the arousal, valence and control dimensions in conjunction with category judgments, in a perception test investigating the recognition of affect bursts (Schröder, 2003). Subjects assigned each stimulus to one of ten emotion categories, and placed the stimulus at the same time on the three emotion dimensions. Subsequent analyses were thereby able to establish the similarity of the emotion categories and therefore the relevance of category confusions. Confusions between "similar" categories, i.e. categories which were close in the three-dimensional emotion space, were considered less problematic than confusions between distant categories.

6.4.2 Feeltrace

A new, dynamic method for the assessment of emotion dimension ratings, not relying on adjective scales, is provided by the Feeltrace tool developed by Roddy Cowie et al. (2000a). In contrast to static scale judgments, it allows the tracking of emotional content continuously over time. It thus has the potential to capture changes in emotional state as they are perceived during the time-course of an utterance.

Feeltrace presents a graphical representation of two emotion dimensions, *activation* (from passive to active) and *evaluation* (from negative to positive), in which the subject can locate the currently perceived stimulus emotion by means of a coloured cursor (see Figure 6.1). Emotion words are printed in the two-dimensional space at their coordinates as determined by previous rating experiments, serving as landmarks giving some orientation to the subjects. The cursor leaves a trail of shrinking circles behind, thus informing about recent cursor movements in the activation-evaluation space. As an example, the trail in Figure 6.1 shows a recent cursor movement from the active/negative to the passive/positive region of the space. Inspired by circumplex models of emotion (see 2.5, p. 25), the space is presented as circular, with a relatively neutral state in the centre of the circle and the

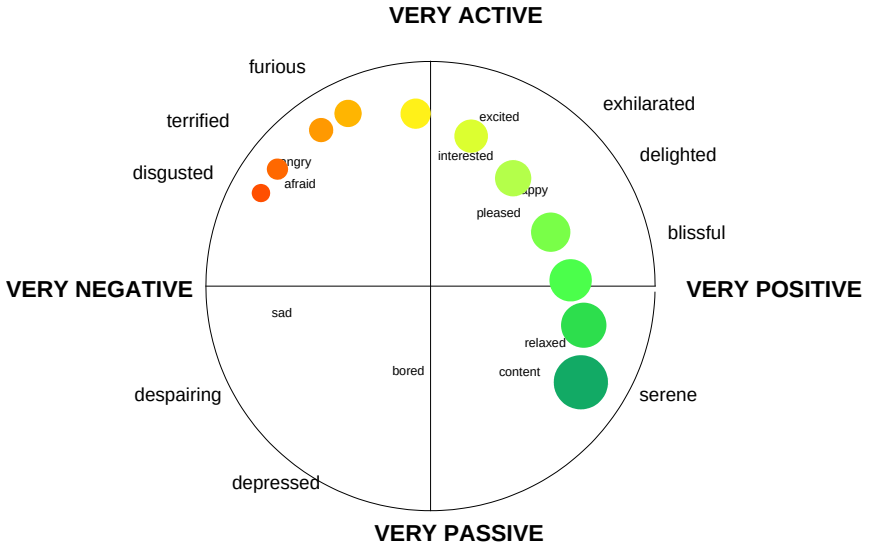


Figure 6.1: The Feeltrace tool for simultaneous rating of stimuli on two emotion dimensions (Cowie et al., 2000a).

maximum intensity of emotional states corresponding to the circle radius.

Feeltrace has been used for describing the emotional content of the audio-visual clips in the Belfast Naturalistic Emotion Database (Douglas-Cowie et al., 2000). The subjects tracked the emotion expressed in a clip as they were watching it; every few milliseconds, the cursor position was recorded and saved with the time stamp obtained from the video clip.

The reliability of the Feeltrace tool was experimentally verified (Cowie et al., 2000a), using 16 subjects who had undergone a training procedure assuring they had understood how to use Feeltrace. Audio-visual stimuli, taken from the Belfast database, covered the centre and all four quadrants of the activation-evaluation space. For all stimuli, the standard deviation of rating distributions across subjects was about 1/5 of the circle diameter, for both activation and evaluation. The conclusion was drawn that "FEELTRACE in its current form has considerable resolving power. On the crudest summary, it is comparable to an emotion vocabulary of 20 words or more." (Cowie et al., 2000a, p. 24).

6.5 Physiological measures on the listener

If the research goal is to by-pass the conscious judgment of perceived emotion, an interesting path to pursue might be the measurement of physiological changes in listeners, as consequences of stimulus presentations.[4] A hypothesis could be that through a process related to empathy, an emotional state similar to the stimulus emotion is induced in the listener. Some evidence seems to exist on which such a hypothesis may be grounded, notably mimicry behaviour on which Joseph Cappella has based his "interpersonal facial feedback hypothesis" (Cappella, 1993).

The approach would have the advantage of by-passing listeners' conscious conceptual representations. The fact that at least neuro-physiological reactions can be triggered by masked emotional stimuli which are not consciously perceived has been shown by J. Morris et al. (1998), who found neural activity changes in the amygdala, as a reaction to a short (30 ms) presentation of an angry face followed by a neutral face. Subjects reported seeing only the neutral face.

While the approach is likely to open interesting new possibilities, it seems clear that interpretations linking physiological reactions to emotions are not straightforward. In particular, it seems impossible to distinguish the listener's perception of the stimulus emotion (perceived speaker emotion) from the listener's reaction to the stimulus (induced listener emotion).

Currently, this approach does not seem to have been applied to speech and emotion research.

[4] Thanks to Jan van Santen for pointing out this option.

Chapter 7

Authenticity and related questions

In the context of sources of emotional speech data (Chapter 4), the question of naturalness of emotion portrayals was raised, and the question of actors' and lay persons' ability to simulate emotions was formulated. This chapter is intended to add some conceptual background to the discussion.

7.1 Naturalness

One of the problems mentioned in conjunction with emotional speech material obtained from actors (see 4.1, p. 43) is the danger of the material being stilted or unnatural.

It seems tempting to consider stiltedness as a direct consequence of voluntary, non-felt expression, and naturalness or authenticity as a direct consequence of the involuntary expression of an emotional state. However, Rainer Banse and Klaus Scherer argue (Banse & Scherer, 1996) that the simulation of an emotion expression is not limited to explicit actor portrayals, but rather occurs as a normal part of daily life. They suggest that socially shared display rules require people to suppress emotions they feel and express emotions they do not feel as a normal part of social behaviour.

In the following, the difference between physiologically and socially determined emotion expression is made explicit in the terms of Klaus Scherer's "push" and "pull" effects. It is then argued that voluntary control is an additional factor not covered by the push-pull distinction.

7.2 Scherer's push-pull distinction

Most emotion theorists seem to agree that emotion expression is caused by a mix of biological and social effects, with the relative importance assigned to each depending on the research tradition. In the Darwinian tradition, for example, Paul Ekman sees the biologically determined emotion expression as "filtered" by means of socio-culturally determined "display rules" (see 1.3.1, p. 11). In the social constructivist tradition, on the other hand, the biological aspect, while acknowledged, is considered secondary, and the major role in the creation of emotional states as well as their expression is attributed to social rules.

In order to capture this distinction, Klaus Scherer has coined the terms 'push-effects' and 'pull-effects'. In an article discussion the symbolic function of vocal affect expression, he proposes a

> differentiation between 'push effects', in which physiological processes such as muscle tone push vocalisations into a certain direction, and 'pull effects', where external factors such as the expectations of the listener may pull the affect vocalisation into a different direction. (Scherer, 1988, p. 82)

Generally, push effects correspond to the biological, evolutionarily shaped emotion system. Emotional expression occurs as a "side effect" of other survival-benefiting functions such as increased muscle tension, heart rate etc. As displaying these conditions may not serve the interests of the signaler, signs under these circumstances "leak out", as John Ohala formulated it (Ohala, 1996). As an example for a situation where strong push effects are present, one can imagine a public presentation: Despite himself, the speaker finds his hands trembling and his voice sounding squeaky, as side effects of the physiological arousal helping him to stay alert in a demanding situation.

Pull effects, in Scherer's terminology, are externally based, in the sense that an externally defined vocal target is aimed at, be it for optimal signal transmission (a voice which "carries well") or for social communication. Social communication includes the complex of rules prescribing how to "behave oneself" in a given culture, including which type of emotion expression is appropriate in a given situation, and in which form, intensity, and duration; therefore, pull effects should be considered the most essential part of emotion expression according to social constructivists. In addition, ethological research suggests that evolution has also shaped emotion *expression* in a way which benefits survival, by

influencing a receiver (Ohala, 1996). Ohala lists behaviours falling into that category, including smiles and frowns, eyebrow movements, crying and tears, gaze behaviour, and F0 use according to the frequency code (see 5.4, p. 57). This type of expression seems to share more properties with pull effects than with push effects.[1]

In Scherer's conception, push and pull effects never occur independently of each other; both are always present. They only vary in the degree to which they are dominant in a given situation. Push effects are largely involuntary, according to Scherer. This does not, however, mean that pull effects are necessarily voluntary, intentional, and consciously controlled (see next section).

Establishing the degree to which a given speech expression is based on push effects and which is the influence of pull effects seems highly problematical, if at all feasible. An area which might have reasonable prospects for studying at least some aspects of the interplay between push and pull effects, however, is that of "affect bursts", or emotional interjections (see 5.3, p. 55).

7.3 Control, volition, and the role of automation

Push effects have been said to be largely involuntary, and pull effects to employ control in order to create the externally defined target. Scherer employs an image to exemplify this tendency in push and pull effects:

> A good analogy for push and pull effects is that of a man standing on a hillside trying to move a stone. If he pushes it, the stone will roll down, but the pusher cannot control exactly where it goes; if, however the stone is at the bottom of the hill and the man pulls it up, the resting point is clearly defined, and he can, to a certain extent, determine the way the stone comes up. Similarly, if one is very aroused, one cannot control the details of one's vocal production mechanisms, whereas if one desires to mimic the ecstatic

[1] It may be worth mentioning in this context that from a neuropsychological perspective, Antonio Damasio has made a proposal regarding a simulation mechanism in the brain (Damasio, 1994). Beside the Jamesian "body-loop", in which nervous activation of bodily changes is fed back into the brain through proprioception, Damasio also proposes an "as-if-body-loop" in which certain proprioception-related brain regions can be stimulated from within the brain, bypassing the body. This "as-if-body-loop" would constitute a simulation mechanism, acquired through repeated association between certain situations and bodily reactions in these situations. It is conceivable that such a simulation mechanism, if it exists, could play a role in the pull effect type of emotion experience and its expression.

> howls of an enthusiastic juvenile, one can refer to an external model which is independent of the production mechanism used. (Scherer, 1988, p. 83)

While the distinction between push and pull effects certainly is a useful one and helps us to understand the various aspects involved in speech and emotion, the situation regarding the control of push and pull effects seems to be more complex than in the illustrating examples given above.

Push effects, in their impact on the vocalisation apparatus, do follow a certain inevitability. Attempts can be made to suppress the effects, but everyday experience suggests this does not succeed beyond a certain point. As speech in the sense of entire sentences always requires a certain amount of control, relatively unadulterated fullblown emotions, dominated by strong push effects defying control, are unlikely to occur in conjunction with speech. This pattern can actually be observed in the Belfast Naturalistic Emotion Database: Examples of emotional speech showed consistently higher duration of "degraded communication" (speech overlap, non-speech vocalisations, silence, and incoherent speech) than neutral examples (Cowie & Cornelius, 2003). Pull effects, however, even if they start off as culturally learned targets, are likely to undergo automation due to repetition, to become automatic associated habits rather than being consciously controlled (Cohen et al., 1990).[2] If that is so, then pull effects are likely to occur as automatically as push effects, unless conscious control intervenes in favour of a modification of habitual behaviour. This would suggest, then, that conscious control, as is expected to occur in simulation and deception, is an additional dimension of complexity not covered by the push-pull distinction.

7.4 Speaker-centered and listener-centered aspects of naturalness

The above discussion has shown that putting questions of authenticity or naturalness of emotion expression in speaker-centered terms is a non-trivial undertaking. It would seem that push and pull effects as well as voluntary intervention might play a role in emotion expression. However, it is not easy to delimit the contribution of either factor.

[2] An example of such an automated, but learned expression is the interjection expressing pain: It is language-specific ("Aua" in German, "Ouch" in English, "Aïe" in French), but everyday experience suggests it is produced involuntarily upon presentation of an appropriate stimulus.

Thinking along the lines of *perceived* naturalness, a listener-centered notion, may provide a measure which is easier to operationalise. Independent of the way in which stimuli were produced, listeners can be asked how stilted or how natural they sound.

A direct assessment of this question is possible by means of the Validtrace tool, a modified version of Feeltrace (see 6.4.2, p. 70). In Validtrace, the horizontal and vertical axes correspond to perceived stiltedness and emotion intensity rather than evaluation and activation. As they listen to a clip, subjects can indicate how authentical the emotion expression is in their perception, and how intense the emotion is. The tool has been put to use in validating the Belfast Structured Emotion Database (see 4.2, p. 44).

In an early experiment, I adressed the question whether, in an audiovisual setting, observers are able to perceive the difference between spontaneous and acted emotional material (Schröder, 1998; Schröder et al., 1998). The results showed that sentences uttered with spontaneous amusement were distinguished from sentences uttered with acted amusement at better than chance level. Interestingly, one group of subjects was quite reliable at recognising the difference, while another group seemed unable to discern it.

Part II

Emotional speech synthesis using emotion dimensions

Chapter 8

Motivation: Natural speech synthesis

This chapter motivates the practical work reported in the second part of this thesis. Starting from the goal to improve the naturalness of synthetic speech, it is explained why emotions are an important aspect of such naturalness. Subsequently, the new approach to modelling emotions for speech synthesis, by means of emotion dimensions, is outlined and the decisions it involves are indicated.

8.1 Why emotions?

The generation of natural-sounding synthetic speech is a long-term goal in speech technology. Speech synthesis systems have existed for several decades now, but they are still applied only in niche domains such as screen readers for the blind. The mass market seems to require much higher quality from synthetic voices than what is currently available.

The quality of synthetic speech is usually considered to have two aspects, *intelligibility* and *naturalness*. Major development has taken place regarding both aspects, from the rule-based or formant synthesisers of the 1970's and 1980's, via the diphone concatenation synthesisers of the 1990's, to the latest corpus-based or unit selection synthesisers. Especially in limited domains, such as weather forecasts, where the type of text to be spoken is known in advance, the latest synthesis systems approach the quality of human speech.

Still, the human ear appears to be very critical when judging the acceptability of speech utterances. This is different for the human eye, which is easily betrayed by a simulation. Computer graphics are perceived as sufficiently convincing to be used by the

film industry, leading to entire movies such as “Final Fantasy”. However, the voices in all animated or cartoon movies continue to be those of actors dubbing the virtual characters. The shortcomings of current speech synthesis technologies are obviously so severe that the movie industry does not consider them a viable option.

There are a number of emerging interactive technologies in which synthetic voices will be required, such as telephone-based dialogue systems and conversational agent applications. With rising complexity of the spoken content, it is not possible anymore to use pre-recorded speech – when unpredictable, free text is to be spoken, the use of text-to-speech synthesis becomes unavoidable. While the first criterion for speech output systems in such applications is of course intelligibility, user acceptance seems to depend on the naturalness aspect of the speech in the given situation. As soon as the dialogue topic or application domain is not a purely fact-based one, but includes emotional aspects, the requirement of naturalness includes the need for the synthetic voices to become expressive. Therefore, one of the most central challenges for speech synthesis in the next decade will be to add expressivity to the synthetic voices.

It can be argued that such expressivity is not necessarily emotional. A synthetic voice may need to express things such as the degree of certainty that the information provided is true, general friendliness, a sense of urgency to act, or empathy when giving bad news. It is true that when emotions are conceived of as *fullblown emotions* as defined in Chapter 1, then the ability to express emotions will not be of much use for these applications. When, however, emotions are understood in the broad sense of *emotional states*, ranging from weak, underlying emotional states to fullblown emotions, then the expression of emotions will be able to contribute substantially to the types of expressivity mentioned above.

8.2 Why emotion dimensions?

When modelling emotions for synthetic speech, two main strategies can be identified. One is to model a few well-defined emotional states or other types of expressivity as closely as possible. Recent work using unit selection synthesis has chosen this method: Akemi Iida et al. (2000) recorded three entire unit selection databases, one spoken in an angry tone of voice, one in a happy and one in a sad tone. Similarly, Lewis Johnson et al. (2002) modelled military speech recording limited domain databases for shouted

commands, shouted conversation, spoken commands and spoken conversation. This type of method guarantees the best possible quality of the selected expressions. On the other hand, no generalisation to other expressions is possible, due to the fact that the expressivity is not explicitly modelled, but simply reproduced from the speech database recordings. These properties make the approach best-suited for limited or specialised applications.

An alternative strategy is to build a general-purpose expressive speech synthesis system. Such a system needs to be able to express any type of emotional or emotion-related state, which requires an explicit modelling of the link between emotions and their prosodic realisations. It is this second strategy which is explored in this thesis. As will be developed below, it seems appropriate to model emotions by means of emotion dimensions in order to provide the required flexibility.

It is important to think carefully about the type of expressivity which is required for likely applications. In a dialogue, an emotional state may build up rather gradually, and may change over time as the interaction moves on. Consequently, a speech synthesis system should be able to gradually modify the voice in a series of steps towards an emotional state. In addition, it seems reasonable to assume that most human-machine dialogues will require the machine to express only mild, non-extreme emotional states. Therefore, the need to express fullblown emotions is a marginal rather than a central requirement, while the main focus should be on the system’s capability to express a large variety of emotional states of low to medium intensity.

In applications, the synthetic voice will often not be the only channel conveying an emotional message. As was stated in the first part of this thesis (see 5.5, p. 57), the perception of emotions is based not only on voice prosody and voice quality, but also takes into account the verbal content of the spoken utterance, facial expression, gestures, body stance, and even the situational context. While neither the interactions between these channels nor their respective importance can be considered established knowledge, there are first indications suggesting different speaker strategies in the use of the expressive means at their disposal. For example, in a pilot study investigating the role of voice quality for emotion expression, I found that different speakers seemed to rely to a varying extent on voice quality for expressing different emotions (Schröder, 1999a). Therefore, it is not necessary for the voice prosody to fully express the emotional state, but it is sufficient if it only approximates the emotional state. For the exact specification of the emotion, other

channels, such as situational context, verbal content or visual aspects are then required.[1]

Emotion dimensions, as presented in 2.6, p. 26, are a representation of emotional states which fulfils the requirements formulated above: They are naturally gradual, and are capable of representing low-intensity as well as high-intensity states. While they do not define the exact properties of an emotional state in the same amount of detail as a category label, they do capture the essential aspects of the emotional state. It can be hypothesised that the approximated modelling of voice prosody based on emotion dimensions will contribute to the recognition of the overall emotional message and to the perceived naturalness.[2]

8.3 Positioning

The second part of this thesis explores the use of emotion dimensions for a general-purpose emotional speech synthesis system. Before outlining the method, I want to point out the properties of this research endeavour, along the concepts set out in the first part of the thesis.

As the type of emotion to be studied here is not exclusively fullblown, it is not trivial to link the research question to one of the four basic traditions of emotion theory (see 1.3, p. 11). Actually, certain links can be made out between the research and each of the four traditions. A minor link to the Darwinian tradition is the frequency code hypothesis (see 5.4, p. 57), according to which low voice pitch is used to convey dominance as a result of evolutionary processes. A link to the Jamesian tradition is the central role of the activation or arousal dimension, which is reliably identified in the literature as the emotion dimension showing the strongest correlations with acoustic parameters (see Chapter 10). The research is related to the cognitive tradition in that it is interested in emotion concepts as they are evoked in the listener's mind, although the appraisals leading to the emotion are not relevant because the task is not to induce an emotion, but only to trigger its recognition. Finally, the work is linked to the social constructivist perspective in that it is interested in the emotional states which are actually relevant in everyday communication.

A descriptive framework is used which allows for the representation of non-extreme emotional states, namely emotion dimensions (see 2.6, p. 26).

[1] Consequently, success in a perception test needs to be thought of in terms of naturalness, i.e. as a successful contribution to a multi-channel emotion expression.

[2] It will be seen later that much of what is known about the prosodic parameter settings reported for the expression of individual emotion categories can be explained by their location on emotion dimensions, see 11.6, p. 130.

Dealing with emotions in speech synthesis is obviously a listener-centered research orientation (see 3.4, p. 39) – the goal being to evoke the intended perception in the listener's mind. The task is to find a correspondence between the acoustic parameters (the distal indicator cues in terms of the Brunswikian lens model, see Figure 3.1, p. 36) and the perceived emotion (the listener attribution), expressed in terms of emotion dimensions.[3]

The study uses a corpus of emotional speech data which is based on quasi-natural conversation (see 4.4, p. 47), thus respecting the social constructivist requirement to study emotional states that are relevant in everyday communication.

The speech parameters investigated are of the gradual, "para-linguistic" type (see 5.1, p. 52), assuming a relative independence between the emotionally relevant acoustic parameters and configurational, linguistic variables (see 5.2, p. 53).

Finally, the perception test uses a preference task methodology (see 6.2, p. 67) in order to assess the overall improvement of a multimodal emotional stimulus due to the emotional prosody rules implemented in the text-to-speech system.

8.4 Outline

The research proceeds as follows. In Chapter 9, the current state of the art in the synthesis of emotional speech is reviewed. The prosody rules for emotion expression reported in the literature are summarised in Appendix A.

As no publications seem to exist which report in detail on the use of emotion dimensions in speech synthesis (with the exception of a short mention in Murray & Arnott (1996)), the wider literature on the vocal correlates of emotion dimensions is reviewed in Chapter 10. The information summarised there gives a solid orientation for what can be expected from the corpus analysis, and provides some more detailed but less solidly confirmed indications about specific parameters which go beyond what can be extracted from the corpus analysis.

In Chapter 11, a statistical analysis of the Belfast Naturalistic Emotion Database is reported. This large corpus of spontaneous emotional speech contains acoustic analyses of

[3] It does not seem essential for the question investigated here to explicitly investigate the proximal percepts, i.e. the auditory correlates of the acoustic parameters. Although the question how acoustic parameters are auditorily perceived is of general interest, it does not seem evident how this additional knowledge would help to address the question whether emotion dimensions are a suitable means for describing emotional states in a speech synthesis application.

the speech utterances as well as the perceived speaker emotions described using emotion dimensions. The corpus analysis investigates correlations between these two types of variables and performs regression analyses allowing the prediction of acoustic parameter settings from the emotion dimensions. It shows that the correlations show highly stable trends across 124 speakers for some acoustic variables (notably variables related to the fundamental frequency, F0), but that other variables for which stable predictions could be made from the literature (notably mean intensity and speech rate) showed only very low correlations. These results are discussed in view of the literature and in relation to the application in speech synthesis.

The properties of the text-to-speech and markup-to-speech synthesis system MARY are presented in Chapter 12, in order to motivate why this system is particularly well-suited for the expression of emotions. One main system feature which is highlighted is the accessibility of intermediate processing results due to the XML-based data representation in the system, using the specially created representation language MaryXML. A second major feature is the use of a diphone voice database providing all German diphones at three different levels of vocal effort.

The actual formulation and implementation of prosody rules for the acoustic expression of emotion dimensions is reported in Chapter 13. The manner in which this implementation draws upon the results from the corpus analysis and from the literature is explained, and practical aspects arising from the selected implementation strategy are described and discussed. A graphical user interface named "EmoSpeak" is presented which allows the user to interactively investigate the acoustic effects of changing the emotional state on the emotion dimensions.

A perceptual evaluation of the resulting emotional speech synthesis system, reported in Chapter 14, investigates the basic usefulness of the dimensional approach, by obtaining goodness-of-fit ratings for matching and non-matching combinations of emotional texts and emotional prosody.

Finally, the Conclusion summarises the work carried out, the main contributions and the information gained. In addition, a number of possible next steps arising from this work are outlined.

Chapter 9

A review of emotional speech synthesis to date

Attempts to add emotion effects to synthesised speech have existed for more than a decade now. Several prototypes and fully operational systems have been built based on different synthesis techniques, and quite a number of smaller studies have been conducted. This chapter aims to give an overview of what has been done in this field, pointing out the inherent properties of the various synthesis techniques used, summarising the prosody rules employed, and taking a look at the evaluation paradigms. Finally, an attempt is made to discuss interesting directions for future development.[1]

9.1 Introduction

With the intelligibility of synthetic speech approaching that of human speech, the need for increased naturalness becomes more palpable. One of the aspects of naturalness most obviously missing in synthetic speech is appropriate emotional expressivity. This observation has been motivating attempts to incorporate the expression of emotions into synthetic speech for more than a decade, and such attempts seem to have gained popularity in recent years. While advances in other aspects of naturalness of synthetic voices have been made, notably with unit selection techniques, the synthesis of emotional speech still has a long

[1] An earlier version of this chapter was published as M. Schröder (2001). Emotional speech synthesis: A review. *Proceedings of Eurospeech 2001*, Aalborg, Denmark, Vol. 1, pp. 561–564.

way to go.

In the studies concerned with the expression of emotion in synthetic speech that can be found in the literature, an interesting variety of approaches has been employed. This chapter will try to give an overview of these studies, and work out the differences and similarities in approach, technique and underlying assumptions. First, the studies are presented in groups according to the type of synthesis technique employed, which coincides in many cases with similarities in the approach. Next, prosody rules employed for expressing emotions are reported, and the paradigms used for evaluation are discussed.

Finally, a number of points will be discussed related to possible directions for future development. These are in part inspired by the ISCA Workshop on Speech and Emotion, recently held in Northern Ireland (Cowie et al., 2000b), which for the first time brought together researchers interested in speech and emotion from a large variety of backgrounds. This fruitful exchange showed, among other things, that our understanding of the way emotion is expressed in speech can be improved along two axes: On the one hand, the description of the vocal correlates of emotions ("How is a given emotion expressed in speech?"); on the other hand, the description of the emotional states themselves (Cowie, 2000) ("What are the properties of the emotional state to be expressed? What is the relation between this state and another state?"). Some implications for future research in the synthesis of emotional speech are proposed in the discussion section.

9.2 Existing approaches and techniques

The modelling of emotion in speech relies on a number of parameters like, among others, fundamental frequency (F0) level, voice quality, or articulatory precision (see 9.3 below). Different synthesis techniques provide control over these parameters to very different degrees.

9.2.1 Formant synthesis

Formant synthesis, also known as rule-based synthesis, creates the acoustic speech data entirely through rules on the acoustic correlates of the various speech sounds. No human speech recordings are involved at run time. The resulting speech sounds relatively unnatural and "robot-like" compared to state-of-the-art concatenative systems, but a large

number of parameters related to both voice source and vocal tract can be varied quite freely. This, of course, is interesting for modelling emotional expressivity in speech.

Several larger undertakings (Cahn, 1989, 1990; Murray, 1989; Murray & Arnott, 1995; Burkhardt, 2000; Burkhardt & Sendlmeier, 2000) have used Formant synthesisers because of the high degree of control that they provide. These include the first ones, from 1989: Janet Cahn's Affect Editor (Cahn, 1989, 1990), and Iain Murray et al.'s HAMLET (Murray, 1989; Murray & Arnott, 1995). Both used DECtalk as a formant synthesis system, providing dedicated processing modules which adapt their input according to the acoustic properties of a number of emotions. In both cases, the acoustic profile for each emotion category was derived from the literature and manually adapted. While the Affect Editor requires the input to be manually annotated, HAMLET processes its input entirely by rule.

Within the VAESS project ("Voices, Attitudes and Emotions in Speech Synthesis"), which ran from 1994 to 1996, emotional expressivity was to be added to a formant synthesiser. Montero et al. (1998) report reasonable success for the modelling of three emotions (hot anger, happiness, and sadness) in Spanish using global prosodic and voice quality parameter settings.

Felix Burkhardt, in his PhD (Burkhardt, 2000; Burkhardt & Sendlmeier, 2000), has also chosen to use formant synthesis, despite the reduced naturalness, because of the high degree of flexibility and control over acoustic parameters that this technique provides. His systematic, perception-oriented approach to finding good acoustic correlates of emotions for German consisted of two main steps. In a first step, he systematically varied five acoustic parameters known to be related to emotion, without using prior knowledge from the literature about the best parameter values for a given emotion. The resulting stimuli were presented in a perception test, providing perceptually optimal parameter values for each emotion studied. In a second step, these optimal values were taken as the basis for the exploration of a wider set of parameters, inspired from the literature. The resulting variants were presented in another perception test, leading to the formulation of refined prosody rules for the synthesis of the emotions studied.

9.2.2 Diphone concatenation

In concatenative synthesis, recordings of a human speaker are concatenated in order to generate the synthetic speech. The use of diphones, i.e. stretches of the speech signal

from the middle of one speech sound (“phone”) to the middle of the next, is common. Diphone recordings are usually made on a monotonous pitch. At synthesis time, the required F0 contour is generated through signal processing techniques which introduce a certain amount of distortion, but with a resulting speech quality usually considered more natural than formant synthesis.

In most diphone synthesis systems, only F0 and duration (and possibly intensity) can be controlled. In particular, it is usually impossible to control voice quality.

Fundamental to every attempt to use diphone synthesis for expressing emotions is the question whether F0 and duration are sufficient to express emotion, i.e. whether voice quality is indispensable for emotion expression or not. Interestingly, very different results were obtained by different studies. While Vroomen et al. (1993), Edgington (1997), Montero et al. (1999a), Schröder (1999a), Murray et al. (2000) and Stallo (2000) report that synthesised emotions can be recognised at least reasonably well, Heuft et al. (1996) and Rank & Pirker (1998) report recognition rates close to chance level. The reason may be that there is no simple general answer: Montero et al. (1999a) reported that for a given speaker, the relative contribution of prosody and voice quality to emotion recognition depends on the emotion expressed, and Schröder (1999a) has found evidence that this may, in addition, be speaker-dependent. In other words, there seem to be speaker strategies relying mostly on F0 and duration for expressing some emotions, and these can be successfully modelled in diphone synthesis. Whether this is true for all types of emotion is not clear yet.

One approach to emotional speech synthesis with diphones, used by Vroomen et al. (1993), Heuft et al. (1996), Edgington (1997), Montero et al. (1999a), Schröder (1999a), Boula de Mareüil et al. (2002), and Bulut et al. (2002), is copy synthesis: F0 and duration values are measured for each speech sound in a given utterance (usually an actor’s portrayal of an emotion), and used for synthesising the same utterance from diphones. The result is a synthetic utterance with the same F0 and duration values as the actor’s speech, but the voice quality determined by the diphones. This technique is suitable for modelling what humans do as closely as possible with the given parameter set. Whether that is the best way to obtain perceptually optimal, believable expressions can be questioned, though: for example in the domain of animated characters, it has been observed that features occurring in human expression need to be exaggerated in synthetic expression in order to be believable (Bates, 1994; Oudeyer, 2002).

A more ambitious approach is the formulation of prosody rules for emotions (Mozziconacci, 1998; Mozziconacci & Hermes, 1999; Rank & Pirker, 1998; Iriondo et al., 2000; Murray et al., 2000; Stallo, 2000; Boula de Mareüil et al., 2002), which is discussed in 9.3.

9.2.3 Unit selection

The synthesis technique often perceived as being most natural is unit selection, or large database synthesis, or speech re-sequencing synthesis. Instead of a minimum speech data inventory as in diphone synthesis, a large inventory (e.g., one hour of speech) is used. Out of this large database, units of variable size are selected which best approximate a desired target utterance defined by a number of parameters. The criteria used for selecting the units are usually of a symbolic nature, such as the phoneme symbol, stress and accent status, sentence type, or position in the sentence. The actual prosodic parameters, such as phoneme duration and F0, are not directly controlled with this approach, but are considered appropriately set as a consequence of the afore-mentioned selection criteria. As a consequence, it is not trivial to precisely control prosody in unit selection synthesis. The weights assigned to the selection parameters influence which units are selected. If well-matching units are found in the database, no signal processing is necessary. While this synthesis method often gives very natural results, the results can be very bad when no appropriate units are found.

Limited domain synthesis is a special type of unit selection synthesis in which the speech corpus is especially designed to cover a given limited target domain, such as weather forecasts. Only utterances from the chosen domain can then be generated using that synthesis voice. Careful corpus design can make sure that suitable units for all possible sentences in the domain exist. Therefore, these voices are very natural for the given domain, to the point that they are sometimes perceived as natural human speech.

The feature of unit selection synthesis to preserve the properties of the recorded speech very well has been exploited by Akemi Iida et al. (2000) for the synthesis of emotional speech. For each of three emotions (anger, joy, and sadness), an entire unit selection database was recorded by the same speaker. In order to synthesise a given emotion, only units from the corresponding database are selected. The emotions in the resulting synthe-

sised speech are well recognised (50-80%).

Lewis Johnson et al. (2002) pursued a similar approach. They employed limited domain synthesis for the generation of convincing expressive military speech, in the framework of the Mission Rehearsal Exercise project. The styles, each recorded as an individual limited domain speech database, were shouted commands, shouted conversation, spoken commands, and spoken conversation. In a perception test, the perceived naturalness of the intonation of the synthesised utterances approached that of natural utterances, more so for commands than for conversational speech.

A different, theoretically more demanding approach is to select the material appropriate for the targeted emotion from one database. The equivalent of prosody rules is then used as selection criteria. This has been attempted by Marumoto & Campbell (2000) and Campbell & Marumoto (2000), who used parameters related to voice quality and prosody as emotion-specific selection criteria. The results indicated a partial success: Anger and sadness were recognised with up to 60% accuracy, while joy was not recognised above chance level.

9.3 Prosody rules employed

In the literature concerned with emotional speech synthesis, global prosodic parameters are often treated as universal, culture-independent cues for emotion. While this claim can certainly be the subject of debate, there seems to be some support for it (Chung, 1999; Tickle, 2000; Scherer et al., 2001), at least as long as the number of available emotion categories is small.

At least in formant and diphone synthesis, prosody rules are at the heart of automatically generated emotional expressivity in synthetic speech. Such rules have been obtained in a number of ways by different authors. Cahn (1990), Murray & Arnott (1995), Rank & Pirker (1998), Murray et al. (2000), and Stallo (2000) have extracted rules from the literature; Montero et al. (1998), Mozziconacci & Hermes (1999), Iriondo et al. (2000), Campbell & Marumoto (2000), and Boula de Mareüil et al. (2002) have carried out their own corpus analysis; and Mozziconacci (1998) and Burkhardt & Sendlmeier (2000) have obtained perceptually optimal values by systematic parameter variation in synthesis.

The types of parameter modelled vary greatly between different studies. All studies

agree on the importance of global prosodic settings, such as F0 level and range, speech tempo and possibly loudness. Some studies try to go into more detail about these global settings, modelling e.g. steepness of the F0 contour during rises and falls (Cahn, 1990; Murray & Arnott, 1995; Iriondo et al., 2000; Stallo, 2000), distinguishing between articulation rate and the number and duration of pauses (Cahn, 1990; Murray & Arnott, 1995; Montero et al., 1999a; Iriondo et al., 2000), or modelling additional phenomena like voice quality (Cahn, 1990; Murray & Arnott, 1995; Rank & Pirker, 1998; Iriondo et al., 2000; Murray et al., 2000; Campbell & Marumoto, 2000; Burkhardt & Sendlmeier, 2000) or articulatory precision (Cahn, 1990; Murray & Arnott, 1995; Rank & Pirker, 1998; Burkhardt & Sendlmeier, 2000). A further step is the consideration of interactions with linguistic categories, like further distinguishing between the speech tempo of vowels and consonants (Murray & Arnott, 1995; Rank & Pirker, 1998; Stallo, 2000), or of stressed and unstressed syllables (Murray & Arnott, 1995; Burkhardt & Sendlmeier, 2000; Stallo, 2000), or the placement of pauses within utterances (Cahn, 1990). The influence of linguistic prosodic categories, like F0 contours (Mozziconacci & Hermes, 1999; Burkhardt & Sendlmeier, 2000), is only rarely taken into account, although these have been shown to play an important role in emotion recognition (Mozziconacci & Hermes, 1999; Burkhardt & Sendlmeier, 2000).

Table 9.1 presents a short overview of prosody rules that have been successfully employed to express a number of emotions. Instead of a reduced summary of all the rules employed in different studies, one successful modelling example per emotion is presented in detail, along with the recognition rate obtained. See Appendix A for a more extensive matrix of the prosody rules employed in the listed publications.

9.4 Evaluation paradigms

There seems to be a de-facto standard for the evaluation of synthetic emotional speech, i.e. a methodology employed in the vast majority of the studies. However, whether that method is actually the most suitable is open to discussion.

The typical way of evaluating the quality of the resulting synthetic emotional speech is through a forced choice perception test including the emotion categories actually modeled, employing a small number of semantically neutral carrier sentences (Cahn, 1990;

Emotion Study Language Rec. Rate	**Parameter settings**
Joy Burkhardt & Sendlmeier (2000) German 81% (1/9)	**F0 mean**: +50% **F0 range**: +100% **Tempo**: +30% **Voice Qu.**: modal or tense; “lip-spreading feature”: F1 / F2 +10% **Other**: “wave pitch contour model”: main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness Cahn (1990) American English 91% (1/6)	**F0 mean**: “0”, reference line “-1”, less final lowering “-5” **F0 range**: “-5”, steeper accent shape “+6” **Tempo**: “-10”, more fluent pauses “+5”, hesitation pauses “+10” **Loudness**: “-5” **Voice Qu.**: breathiness “+10”, brilliance “-9” **Other**: stress frequency “+1”, precision of articulation “-5”
Anger Murray & Arnott (1995) British English	**F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear Burkhardt & Sendlmeier (2000) German 52% (1/9)	**F0 mean**: “+150%” **F0 range**: “+20%” **Tempo**: “+30%” **Voice Qu.**: falsetto
Surprise Cahn (1990) American English 44% (1/6)	**F0 mean**: “0”, reference line “-8” **F0 range**: “+8”, steeply rising contour slope “+10”, steeper accent shape “+5” **Tempo**: “+4”, less fluent pauses “-5”, hesitation pauses “-10” **Loudness**: “+5” **Voice Qu.**: brilliance “-3”
Boredom Mozziconacci (1998) Dutch 94% (1/7)	**F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C, avoid final patterns 5&A and 12

Table 9.1: Examples of successful prosody rules for emotion expression in synthetic speech. Recognition rates are presented with chance level for comparison. Sadness and Surprise: Cahn uses parameter scales from –10 to +10, 0 being neutral; Boredom: Mozziconacci indicates intonation patterns according to a Dutch grammar of intonation, see Mozziconacci (1998) for details.

Montero et al., 1998; Burkhardt & Sendlmeier, 2000; Mozziconacci & Hermes, 1999; Vroomen et al., 1993; Heuft et al., 1996; Edgington, 1997; Rank & Pirker, 1998; Montero et al., 1999a; Schröder, 1999a; Iida et al., 2000; Campbell & Marumoto, 2000). It can be argued, though (Banse & Scherer, 1996, p. 615), that this corresponds rather to a discrimination task than an identification task, especially when the number of categories involved is small. The advantages of such a forced-choice test are that it is relatively easy to carry out, provides a simple measure of recognition relative to chance level and allows at least a limited comparability between studies. However, a forced choice test provides no information about the quality of the stimulus in terms of naturalness or believability. Therefore, a number of studies assess the degree of naturalness, believability or overall preference of the emotion expression in addition to the forced choice rating, often on a five-point scale (Cahn, 1990; Rank & Pirker, 1998; Schröder, 1999a; Iida et al., 2000). In addition, the intensity of the emotion (Cahn, 1990) or the synthetic speech intelligibility (Iida et al., 2000) have been assessed.

Another possibility, especially suited for finding phenomenons not expected by the experimenter, are free response tests (Murray & Arnott, 1995; Schröder, 1999a). A subsequent grouping of the responses into meaningful classes can be performed using validated word lists (Murray & Arnott, 1995).

An interesting alternative evaluation paradigm was employed by Murray & Arnott (1995) and recently adopted by Stallo (2000). First, a number of “distractor” response categories are introduced in the perception test, as well as a category “other”. In addition, semantically neutral as well as semantically emotional texts are used, both synthesised with neutral and emotional prosody. The difference in recognition between the version with neutral prosody and the version with emotional prosody is then taken as the measure for the perceptive impact of the prosody rules. Interestingly, the recognition *improvement* due to prosody was bigger for emotional texts than for neutral texts.

In an audio-visual context, a talking head visually expressing emotion (Stallo, 2000) was presented with neutral and with emotional synthetic speech. Subjects rated which version they perceived as more natural, more understandable, etc. The version with emotional speech was clearly preferred.

9.5 Discussion

Emotional speech synthesis is not yet applicable in many real life settings. A number of structural problems which seem to contribute to that are discussed in the following.

In most studies, a number of between three and nine discrete, extreme emotional states are modelled. The often implicit assumption that the expression of a few basic or primary emotion categories is most important to model, and that other emotional states can somehow be derived from that, has been questioned by Roddy Cowie (2000). He argued that systems should be able to express less intense emotions more suitable for real life applications. For a perception-oriented task such as the synthesis of emotional speech, a listener-oriented taxonomy like emotion dimensions may be a suitable starting point for describing non-extreme emotional states. A first step in that direction is briefly mentioned in Murray & Arnott (1996).

Besides the gradual, global parameter settings such as F0 mean, overall speech tempo etc., it is well known that linguistic categories such as F0 contour can have an effect on emotion perception in interaction with other linguistic information like sentence type (Scherer et al., 1984; Andreeva & Barry, 1999, p. 8). Such effects, most likely language-specific in nature, are not yet appropriately accounted for in emotional speech synthesis.

As pointed out earlier (see 9.2), synthesis techniques currently seem to show a trade-off between flexibility of acoustic modelling and perceived naturalness. In order to express a large number of emotional states with a natural-sounding voice, either the rule-based techniques need to become more natural-sounding (see e.g. Kasuya et al., 1999), or the selection-based techniques must become more flexible (Campbell & Marumoto, 2000) and reliable (Mokhtari & Campbell, 2002).

Finally, evaluation techniques should be developed that are more suitable for assessing the appropriateness of acoustic parameter settings for a given communication situation. This might be achieved by moving away from forced-choice tests using abstract emotion words towards tests measuring the perceived naturalness of an utterance given an emotion-defining context.

Chapter 10

Vocal correlates of emotion dimensions in the literature

This chapter examines the literature on the vocal correlates of emotion dimensions. It will be seen that although only a limited number of studies are available, a pattern emerges that can serve as a relatively stable baseline reference.

10.1 The evidence

The studies are presented in this section in temporal order. As can be expected when knowledge is accumulated, the first studies start with relatively ad hoc approaches, and based on their experiences, later studies develop more well-defined research questions and techniques for investigating them.

One of the first experiments investigating the link between emotions as measured by emotion dimensions and intonation was carried out by Uldall (1960). On a number of neutral carrier sentences, she modified the intonation contour artificially. Perceptual ratings were carried out using semantic differential scales chosen on the basis of a pilot experiment. A factor analysis showed that scales related to pleasantness accounted for most of the variance. Unfortunately, aspects of the intonation contour related to fundamental frequency level and range and to contour shape were not clearly separated. In addition, the generalisability of the results was later questioned by Pakosz (1983) on the grounds that the scales used by Uldall were not well-chosen for clearly representing individual emotion

dimensions. An additional problem is the fact that the intonation contours were apparently not chosen according to any intonation model, which makes it likely that some contours sounded rather unnatural. Rating scales which were designed to measure pleasantness may actually have been used by listeners to indicate perceived naturalness.

Davitz (1964a) had seven speakers produce two semantically neutral standard sentences in each of 14 emotional tones. The sentences were embedded into emotion-specific paragraphs read by the speakers. Three types of ratings were obtained. First, the standard sentences were presented in a forced choice recognition task; second, they were rated on scales representing the auditory variables loudness, pitch level, timbre and speech tempo. Third, the emotion-specific paragraphs were rated on the three semantic differential dimensions proposed by Osgood et al. (1957): valence, strength, and activity. Correlations were calculated between the auditory ratings and the dimensional ratings of the 14 emotions. Correlations were found to be strongly significant for activity, but not for valence and strength. Expressions of more active emotions were rated louder, higher-pitched, more "blaring" in timbre, and faster. The author suggests that valence and strength may be conveyed by more subtle auditory cues than the ones he studied.

Huttar (1968) recorded the spontaneous speech of one speaker in classroom lectures and discussions. 30 selected utterances were presented to listeners and rated by means of semantic differential scales measuring "degree of emotion" (which, according to Huttar's description, seems to correspond closely to the activation dimension) and some individual emotions. These ratings were correlated with acoustic measurements and with auditory ratings of prosodic variables. Significant correlations were found, in the sense that more active emotions corresponded to a higher maximum F0, higher F0 range, higher maximum intensity, and to a faster perceived speech rate.

Scherer (1974) and Scherer & Oshinsky (1977) reported on experiments with sine wave tone sequences resembling a short sentence. In a factorial design, a number of acoustic parameters (amplitude variation, pitch level, pitch contour direction, pitch variation, tempo, envelope (attack-decay ratio), and low-pass filter cut-off level) were varied. The resulting stimuli were rated in a perception test for their location on the emotion dimensions pleasantness/evaluation, activity and potency. Inter-rater agreement was highest for the ratings on the activity dimension. Multiple regression analyses were carried out with the acoustic variables as predictors and the emotion dimensions as the predictees, in

order to test the predictive strength of the acoustic variables. Two-thirds to three-quarters of the variance in the emotion ratings could be explained by the acoustic variables. Tempo was strongly and positively correlated with activity, and to a lesser extent with pleasantness and potency. The low-pass filter cut-off level (number of harmonics, related to voice quality) showed a strong positive correlation with potency, a negative correlation with pleasantness and a weak positive correlation with activity. Pitch level also correlated positively with activity and potency, and negatively with pleasantness. Pitch variability (range) and envelope (attack-decay ratio) correlated positively with pleasantness.

Green & Cliff (1975) used similarity judgments of emotional speech stimuli in order to identify emotion dimensions (see also 2.6.1, p. 26). Alphabet recitations, which were used as carriers for the simulation of 11 emotional states, were rated according to their perceptual similarity, and according to auditory properties on seven scales. Correlations between the emotion dimensions that had emerged from the similarity ratings (see 2.6.1, p. 26) and the tone-of-voice scales indicated a correspondence between pleasantness and "warmth" of the voice, as well as between excitement and increased "thinness of voice" and vocal pitch.

Pakosz (1982; 1983) discussed the link between intonation contour and emotional meaning. Pakosz (1983) reviewed some earlier publications on the vocal expression of emotions, and came to the conclusion that intonation[1] only conveys information about the activation dimension, not about individual emotion categories.[2] This "Relative Strength Hierarchy" hypothesis is expanded upon in Pakosz (1982), where it is supported by a number of examples. According to this hypothesis, an intonation contour does not directly express a denotative emotional meaning, but only the abstract position on the activation dimension of emotional meaning. Only in combination with other semantically defined factors such as the verbal or situational context can this activation level be interpreted as a concrete emotion or attitude. Suggestions are made as to the placement of intonation contours on the activation dimension for English.

In a review of the literature on vocal expression of emotions, Tischer (1993) found the following links between emotion dimensions and acoustic parameters. The activation

[1] It is not entirely clear how Pakosz used the term "intonation", whether this denotes only the shape of the intonation contour or also encompasses global settings such as pitch level and pitch range.

[2] From today's perspective, this claim certainly must be considered too strong. At least since the studies by Banse & Scherer (1996) and Leinonen et al. (1997), it is established that a number of 10–14 emotion categories can be distinguished on the basis of non-verbal aspects of speech alone, with a mean reliability of 50–60%.

is mirrored in fast speaking rate, high intensity and pitch level, large pitch range, and a high relative energy in the high frequency spectrum. The power dimension cannot be considered orthogonal to the activation dimension with respect to the vocal expression of emotions, as their vocal correlates correspond for most parameters. An exception are the findings regarding pitch level, which is reported to be positively correlated with power by some studies, and negatively by others. The frequency code hypothesis (see 5.4, p. 57), according to which high power should be expressed through low pitch, is thus only confirmed by a part of the studies. As to the evaluation dimension, Tischer's review found inconclusive evidence in the literature, due to the small number of studies into the vocal correlates of the evaluation dimension.

In the practical part of his study, Tischer (1993) investigated listener-centered aspects of the links between emotion categories and dimensions and acoustic parameters, and how these evolve during the time course of an utterance. An emotionally unspecified text was produced with 14 emotional tones by four speakers. The utterances were rated, among other things, on scales measuring the three emotion dimensions. In the following, some of the correlations which he found between acoustic parameters and emotion dimensions are listed: vowel duration correlated positively with power and evaluation; the relative amount of pausing in the utterance correlated negatively with activation and power; the smaller the differences between the maximum intensity in an utterance and the mean syllable intensity, the lower are activation and power; intensity maxima which occurred late in the voiced part of a syllable corresponded to high activation; the more the intensity increased during the course of a "sense unit" (consisting of three to four words in the test utterance), the higher the activation and the more negative the evaluation; mean fundamental frequency correlated positively with activation and power; the slope of fundamental frequency rises between syllable maxima corresponded to a high activation and power.

Ohala (1994) described a frequency code type of voice use (see also 5.4, p. 57), according to which higher power is expressed through lower mean F0.

Banse & Scherer (1996) analysed their data, among other things (see 5.1, p. 52), in terms of "intense" emotions such as despair, hot anger, panic fear, and elation, for which the highest F0 mean (Banse & Scherer, 1996, p. 624) and mean energy (p. 627) were found. Banse and Scherer also observed that low coping potential (power) led to a high energy proportion in low frequency bands (a steep negative spectral slope).

Pereira (2000) conducted a small experiment in which two actors (one male, one female) produced two standard sentences in five emotions (happy, sad, hot/cold anger, neutral). Listeners placed these utterances on the three dimensions arousal, pleasure, and power. Correlations with acoustic analyses showed that perceived arousal correlated with higher F0, larger F0 range, and increased loudness. In addition, for the male speaker, the power dimension correlated with the acoustic variables in the same way as the arousal dimension (higher F0, larger F0 range, and increased loudness).

Trouvain & Barry (2000) examined the prosodic correlates of excitement in three horse race commentaries. The following trends became obvious from this data: As excitement built up during the races, most pauses became shorter; inter-pause and inter-breath stretches became shorter, indicating an increase in breathing; speaking rate did not increase; fundamental frequency level did rise by about an octave; intensity increased slightly; and spectral tilt decreased, indicating higher vocal tension.

10.2 Discussion

An unambiguous agreement exists concerning the link between the activation dimension and the most frequently measured acoustic parameters (Tischer, 1993): Activation is positively correlated with mean F0, mean intensity, and, in most cases, with speech rate. The evidence in this point is so strong that it can serve as a solid baseline for the analysis reported in the next chapter. Additional parameters positively correlated with activation are pitch range (Huttar, 1968), "blaring" timbre (Davitz, 1964a), high-frequency energy (Scherer & Oshinsky, 1977), late intensity peaks (Tischer, 1993), intensity increase during a "sense unit" (Tischer, 1993), and the slope of F0 rises between syllable maxima (Tischer, 1993). Higher activation also corresponds to shorter pauses (Tischer, 1993; Trouvain & Barry, 2000) and shorter inter-pause and inter-breath stretches (Trouvain & Barry, 2000).

The evidence for evaluation and power is less stable. There seems to be a tendency that studies which take only a small number of acoustic parameters into account do not find any acoustic correlates of evaluation and/or power (Davitz, 1964a; Pereira, 2000). Other studies only investigated activation, but not evaluation or power (Huttar, 1968; Trouvain & Barry, 2000).

The limited evidence regarding the vocal correlates of power indicates that power is

basically recognised from the same parameter settings as activation (high tempo, high F0, more high-frequency energy, short or few pauses, large intensity range, steep F0 slope), except that sometimes, high power is correlated with lower F0 instead of higher F0 (Tischer, 1993), and power is correlated with vowel duration (Tischer, 1993).

There is even less evidence regarding the acoustic correlates of evaluation. Positive evaluation seems to correspond to a faster speaking rate, less high-frequency energy, low pitch and large pitch range (Scherer & Oshinsky, 1977); a "warm" voice quality (Green & Cliff, 1975); and longer vowel durations and the absense of intensity increase within a "sense unit" (Tischer, 1993).

It can be noted that Tischer (1993) provides the most detailed information currently available about acoustic parameters that correlate with emotion dimensions. A main reason is that he investigated a large number of potentially relevant acoustic parameters, out of which he selected the most relevant parameters, i.e. those which correlated most strongly with the emotion dimensions. It is unfortunate that the book does not contain the numerical values of the correlations or of linear predictions derived from these correlations – such numerical values could have been used for the prediction of acoustic parameter settings appropriate for conveying a given emotional colouring in speech synthesis.

10.3 Conclusion

The evidence collected in this chapter can serve two purposes. On the one hand, where the corpus analysis investigates the same or similar parameters, the findings in that study can be compared to the previous work. In particular the correlation of activation with the most frequently measured acoustic parameters – mean F0, mean intensity, and speech rate – can be considered established knowledge due to the comparatively large number of sources confirming it. These correlations can be considered hypotheses for the corpus analysis.

On the other hand, where technical limitations do not allow the corresponding parameters to be measured in the corpus analysis, the evidence collected here can serve itself as the basis for the formulation of tentative prosody rules in the synthesis system.

Chapter 11

Corpus analysis: Quantitative assessment of vocal correlates of emotion dimensions

In this chapter, a statistical analysis of the Belfast Naturalistic Emotion Database is presented.[1] The Belfast database is introduced, along with the emotion ratings and acoustic analyses available for it. The approach to determining correlations between emotion dimensions and acoustic parameters is described, and the results are discussed in the light of the literature survey of Chapter 10. In addition, an alternative approach to database analysis using a normalisation technique is explored. Finally, a method is proposed through which the results found in this study can be put in relation to acoustic correlates of emotion categories.

11.1 Introduction

The literature survey in Chapter 10 made it clear that there are certain stable patterns of correspondences between emotion dimensions and acoustic parameters. However, the studies presented there were not concerned with speech synthesis, and therefore did not attempt to quantify the correlations in the form of a rule set. Such a rule set should predict the acoustic parameters on the basis of the emotional state as specified by the position on the emotion dimensions.

[1] An earlier version of this chapter appeared as: M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk & S. Gielen (2001). Acoustic correlates of emotion dimensions in view of speech synthesis, in Proc. Eurospeech 2001, Aalborg, Denmark, Vol. 1, pp. 87–90.

This chapter investigates a database of spontaneous emotional speech, the Belfast Naturalistic Emotion Database (Douglas-Cowie et al., 2000, 2003). This database is one of the largest corpora of spontaneous emotional speech available today, and it is unique in being rated according to emotion dimensions. Therefore, analysing this database currently appears to be the best means available for obtaining a broadly based set of correspondences between acoustic parameters and emotion dimensions.

It is true that a database containing few emotion-specific samples from many different speakers is not the optimal basis for an analysis targeted towards speech synthesis. Rather than trying to find the "lowest common denominator" of many speakers, in speech synthesis one tries to model the typicalities of one specific speaker. This means that the optimal database to be used in this analysis would contain large amounts of emotional and non-emotional speech data from a single speaker, all rated for the perceived position on emotion dimensions. No such database is currently available, although there are promising developments in this direction in the JST/CREST ESP project (Douglas-Cowie et al., 2003; Mokhtari & Campbell, 2002). In the analysis of a multi-speaker database such as the one used here, it must be kept in mind that due to inter-speaker variability, the correlation and regression coefficients obtained are likely to be smaller than when modelling a single speaker.

11.2 The Belfast database of spontaneous emotional speech

The Belfast Naturalistic Emotion Database contains audio-visual recordings of 124 English speakers exhibiting relatively spontaneous emotion (Douglas-Cowie et al., 2000, 2003), 78% of which are female. The material contains TV recordings of chat shows and religious programs, as well as interviews recorded in a studio. In terms of scale and range of emotions, it is one of the largest collections of natural emotional speech available.

11.2.1 Perceptual ratings on emotion dimensions

The database has been perceptually annotated with respect to emotional content using the Feeltrace tool (see 6.4.2, p. 70). After a necessary training phase, subjects can locate the emotional tone of a clip in the 2-dimensional activation-evaluation space, continuously over time. In addition, each clip is labelled using the words of a Basic English Emotion Vocabulary (Cowie et al., 1999a). The rich characterisation of these emotion words obtained in previous experiments (Cowie et al., 1999a) allows the addition of the power dimension, as associated with the emotion word, to each clip. Thus each clip is positioned on the three dimensions, with activation and evaluation changing over time during the clip and power remaining static.

Seven subjects have provided Feeltrace ratings for each clip in the database. Figure 11.1 shows the coverage of the activation-evaluation space. It can be seen that the neutral region is particularly densely covered, due to the fact that the database contains one "neutral" clip for each speaker. All four quadrants are reasonably well-covered, with a slight predominance of the negative-active quadrant, reflecting the type of emotions typically displayed in chat shows. The pattern is similar for female and male speakers, with the male space being less densely covered than the female space due to the larger number of female speakers included in the database.

Inter-rater agreement was measured by means of pairwise correlations between two subjects' mean ratings of each clip, separately for activation and evaluation. The results are shown in Table 11.1. It can be seen that all subjects agreed to a very high degree on the evaluation of the clips, with a typical correlation of above .8; the agreement on activation was slightly lower, but stable at a typical correlation coefficient of about .6. None of the subjects were markedly different from the others, so that it seems appropriate to include all subjects' ratings in the analyses.

11.2.2 Acoustic analyses

The acoustic analyses of the database were performed semi-automatically using the ASSESS system (Cowie et al., 1995). It generates a simplified core representation of the speech signal based mainly on the F0 and intensity contours. In order to compare recordings from different sources, an intensity normalisation is attempted, based on the amplitude difference between neutral speech and background noise. Key 'landmarks' are then identified, including peaks and troughs in the contours as well as boundaries of pauses and frication events. Measuring the 'pieces' between these landmarks gives rise to a range of variables called 'piecewise'. They provide a rich description of the way contours (of pitch and intensity) behave over time. Variables, piecewise and others, are then summarised in an array of statistics (covering central tendency, spread and key centiles). Additional

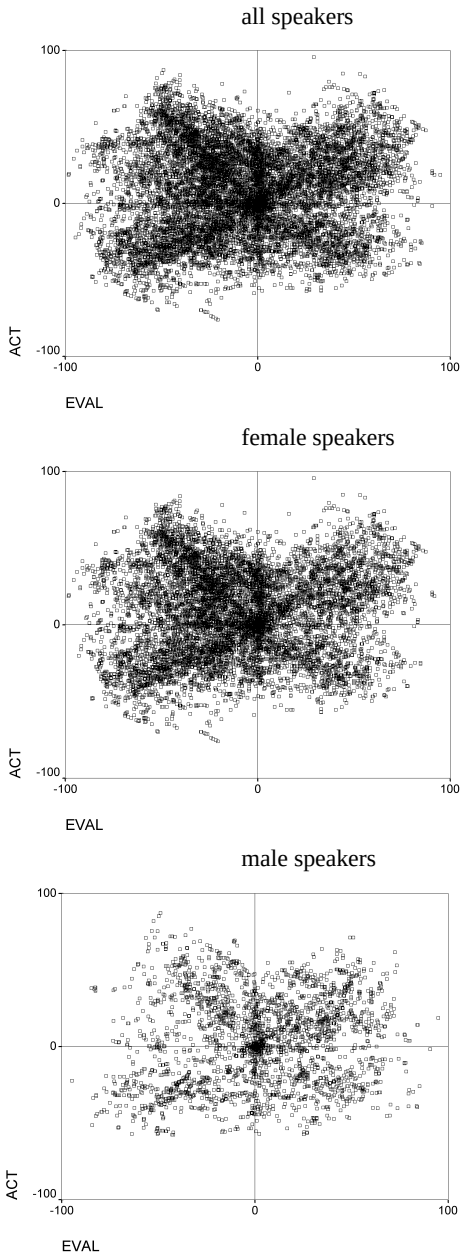


Figure 11.1: Activation-evaluation space coverage of the Belfast Naturalistic Emotion Database.

	Subject	ao	ed	ma	bo	de	jo	li
Activation	**ao**	1	.546	.614	.547	.571	.596	.501
	ed		1	.628	.647	.594	.640	.647
	ma			1	.684	.678	.672	.654
	bo				1	.704	.667	.654
	de					1	.642	.620
	jo						1	.614
	li							1

	Subject	ao	ed	ma	bo	de	jo	li
Evaluation	**ao**	1	.858	.842	.865	.882	.855	.757
	ed		1	.840	.845	.855	.864	.803
	ma			1	.833	.824	.812	.769
	bo				1	.873	.851	.773
	de					1	.859	.757
	jo						1	.769
	li							1

Table 11.1: Pairwise correlations of rated mean coordinates of each clip on the activation and evaluation dimensions, demonstrating the degree of inter-rater agreement when using the Feeltrace tool.

measures deal with properties of ‘tunes’ (i.e. segments of the pitch contour bounded at either end by a pause of 180 ms or more) as well as with spectral properties.

11.2.3 Expected correspondences

When thinking about the correspondences between emotional and acoustic variation in the corpus, it is important to keep in mind that due to the spontaneous nature of the data and the large number of different speakers, there is a large number of variables which it is impossible to control for (see 4.4, p. 47). These include a wide range of extra-linguistic factors such as age, health, etc., as well as linguistic factors such as dialogue management situation, sentence mode, utterance length etc. Indeed, it is a serious danger that the emotion-specific variation gets lost in the “noise” of these other factors (Stibbard, 2001). Therefore, it is all the more meaningful if a significant correlation between emotional and acoustic measures emerges from the data: if *despite* the large number of uncontrolled variables, general tendencies can be found for correlations between emotion dimensions and acoustic variables, these are likely to be stable and generalisable effects.

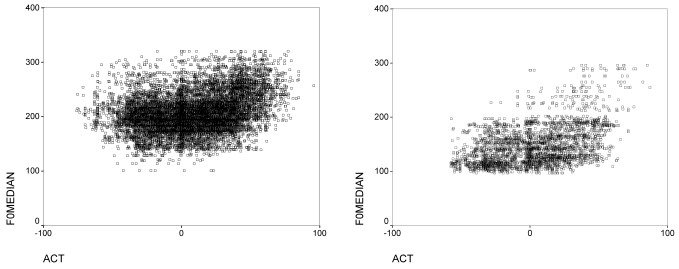


Figure 11.2: Scatterplot activation – F0 median for female speakers (left) and male speakers (right).

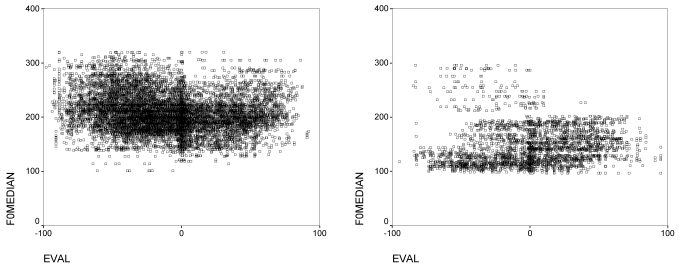


Figure 11.3: Scatterplot evaluation – F0 median for female speakers (left) and male speakers (right).

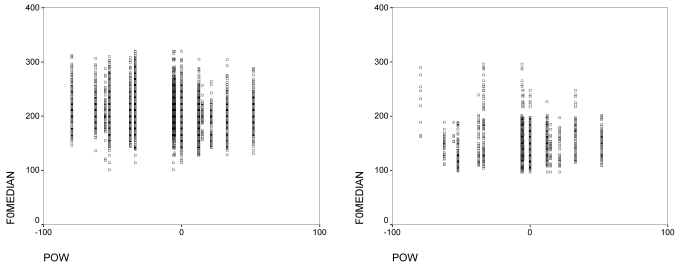


Figure 11.4: Scatterplot power – F0 median for female speakers (left) and male speakers (right). Note that power is derived from the emotion category labels, for which reason it only takes a limited number of discrete values.

In order to illustrate the difficulty of the task, Figures 11.2 – 11.4 show scatterplots of each of the emotion dimensions with F0 median, the parameter most strongly correlated to emotion dimensions. Even in this strongly affected acoustic parameter, the correlation with the emotion dimensions is barely visible. It is obvious that many factors besides the emotion dimensions influence the acoustic parameter.

11.3 Prosodic parameters relevant for speech synthesis

For speech synthesis, the most interesting acoustic variables are those which can be controlled in speech synthesis systems. Table 11.2 lists the prosodic variables which were selected and gives a detailed account of the ASSESS measures used for operationalising these prosodic variables. It is clear that not all interesting information can be provided by a semi-automatic analysis tool like ASSESS. For example, as no phone boundaries are detected, only approximate measures of articulation rate are available, such as the number of intensity peaks per second which are expected to correspond roughly to syllable nuclei, or the number of fricative stretches per second. Other interesting information, such as the location of pauses with respect to sentence structure, or articulation precision, is not available at all, because it would require linguistic information ASSESS does not have.

Intensity descriptions as fine-grained as for F0 would have been available from ASSESS, but were not considered due to the fact that usually, the current concatenative speech synthesis systems do not allow the same degree of control over intensity as over F0.

Spectral slope is a simple spectral measure of the relation of high-frequency and low-frequency energy. It is used as a simple approximation of the “harshness” vs. “softness” of the voice quality. In the ASSESS system, spectral slope is determined by splitting the spectrum in third-octave bands and fitting a line to their relative energies.

Approximations of the Hammarberg indices were used as another set of coarse measures of spectral properties related to voice quality. Britta Hammarberg and her co-workers (Hammarberg et al., 1980) showed that differences in voice quality were correlated to differences in the maximum intensity in each of three spectral bands. They used bands ranging from 0–2kHz, 2–5kHz and 5–8kHz. The closest-matching ASSESS bands were b2, ranging from 0.7–1.8kHz, b3, ranging from 1.8–3.6kHz, and b4, ranging from 3.6–10kHz.

	prosodic parameter	ASSESS measure
fund. freq.	F0 mean	F0 median
	F0 range	90th - 10th F0 percentile
	accent structure	magnitude of F0 rises & falls
		duration of F0 rises & falls
		slope of F0 rises & falls (magnitude / duration)
		number of F0 rises & falls per second
tempo	phrasing	duration of pauses
		duration of 'tunes' (inter-pause stretches)
	articulation rate	number of intensity peaks per second
		number of fricative bursts per second
intens.	intensity mean	intensity median
	intensity range	90th - 10th intensity percentile
	dynamics	median intensity for intensity maxima - overall intensity median
voice quality	spectral slope	spectral slope in long-term average spectrum for non-fricatives
	vocal effort	Hammarberg 'effort' = b3mx
	breathy voice	Hammarberg 'breathy' = (b2mx-b3mx) - (b3mx-b4mx)
	"head" vs. "chest" register	Hammarberg "head" = b2mx - b4mx
	coarse voice	Hammarberg "coarse" = b2mx - b3mx
	unstable voice	Hammarberg "unstable" = b3mx - b4mx

Table 11.2: Prosodic variables selected for investigation in view of speech synthesis, and the ASSESS measures as which they were operationalised. See text for an explanation of the Hammarberg measures.

The measures used (b2mx, b3mx, b4mx) are the maximum energies in the respective spectral bands, i.e. the maxima in a frequency–energy diagram. The measures are listed in Table 11.2 in decreasing order of relevance: Hammarberg et al. (1980) reported that the amount of variance in expert ratings explained by the spectral measure was highest for the 'effort' measure (the maximum energy in the spectral band from 2–5kHz accounted for 53% of the variance in expert ratings related to a factor named "hyper-hypofunctional", interpreted as "high-low vocal effort"), and lowest for the 'unstable' measure (accounted for only 6% of the variance in 'unstable' ratings).

11.4 Analysis in terms of absolute values

The dimensional ratings, on scales from -100 to 100, were aligned with the acoustic analyses of the individual 'tunes' (inter-pause stretches) within each clip. The number of valid data points resulting from this alignment are listed in Table 11.3. The numbers vary between the acoustic measures because the amount of audio data required for providing reliable measures differ.

11.4.1 Method

Correlation and linear regression analyses of the data were performed using SPSS. The data was analysed separately for male and female speakers. The three emotion dimensions Activation (A), Evaluation (E) and Power (P) were used as independent variables, predicting in turn each of the acoustic variables selected. The regression coefficients calculated by this analysis predict the value of the acoustic variable at a given point in the 3-dimensional emotion space.

It is important to note that the regression coefficients for different independent variables are influenced by each other. As a consequence, the set of coefficients obtained is only optimal in precisely this combination. For example, if the input to be used for calculating acoustic variables specified activation and evaluation, but not power, then it would not be optimal to use the coefficients calculated here.[2]

For the purpose of interpretation, only those correlations that were significant at the level of $p < .05$ or better were considered. As the number of data points available for female speech was about three to four times larger than for male speech (depending on the acoustic variable), more significant correlations were found for female than for male speech.

[2] For this reason, separate analyses using in turn all possible combinations of the three emotion dimensions as independent variables were calculated. In addition, analyses were conducted in which the squares of the emotion dimensions (A^2, E^2, P^2) were added as predictor variables, which is a method of accounting for simple non-linear effects (Tabachnick & Fidell, 2001, p. 113). Indeed, the explained variance increased when the squared terms were included. However, there are two important disadvantages of using squared terms: First, the resulting correlations are less easily interpretable, and second, the squared terms in the equation have a tendency to predict unreasonably large effects in regions of missing data. Therefore, when using squared predictors, it is particularly important to only apply the resulting equations in regions well supported with corpus data. In the present case, this means that the predictions made by equations including squared terms should be treated with caution for extreme emotional states, i.e. for extreme positions on the emotion dimensions. For these reasons, the model which will be discussed in this chapter does not include any squared predictors. For space reasons, and as the synthesis rules developed in Chapter 13 only partially rely on the corpus analysis reported in this chapter, the other analyses are not reproduced here.

	Acoustic variable	Condition					
		absolute values		normalised		normalised restr.	
		female	male	female	male	female	male
fundamental frequency	F0 median	12550	3493	12543	3493	2626	845
	F0 range	12536	3486	12529	3486	2623	844
	med. magn. F0 rises	11591	3059	11584	3059	2418	761
	rng. magn. F0 rises	6217	987	6212	987	1305	270
	med. magn. F0 falls	11591	3087	11584	3087	2429	751
	rng. magn. F0 falls	5692	903	5687	903	1231	256
	med. dur. F0 rises	11591	3059	11584	3059	2418	761
	rng. dur. F0 rises	7979	1701	7974	1701	1672	429
	med. dur. F0 falls	11591	3087	11584	3087	2429	751
	rng. dur. F0 falls	7454	1589	7449	1589	1587	425
	med. slope F0 rises	11591	3059	11584	3059	2418	761
	med. slope F0 falls	11591	3087	11584	3087	2429	751
	F0 rises p. sec.	12025	3213	12018	3213	2507	783
	F0 falls p. sec.	12025	3213	12018	3213	2507	783
tempo	duration pauses	10298	2898	10292	2898	2153	700
	'tune' duration	13395	3766	13388	3766	2806	920
	intensity peaks p. sec.	13353	3773	13346	3773	2785	920
	fricat. bursts p. sec.	13353	3773	13346	3773	2785	920
intens.	intensity median	12892	3577	12885	3577	2691	871
	intensity range	12599	3493	12592	3493	2635	844
	dynamics at peaks	12599	3500	12592	3500	2635	845
voice quality	spectral slope non-fric.	12207	3423	12200	3423	2555	828
	Hamm. 'effort'	12207	3423	12200	3423	2555	828
	Hamm. 'breathy'	12200	3423	12193	3423	2554	828
	Hamm. 'head'	12200	3423	12193	3423	2554	828
	Hamm. 'coarse'	12207	3423	12200	3423	2555	828
	Hamm. 'unstable'	12200	3423	12193	3423	2554	828

Table 11.3: Number of data points used for the regression analyses. For the differences between the "normalised" and "normalised restricted" conditions, see text in 11.5.1.

11.4.2 Results

The most fundamental result is that nearly all of the acoustic variables show substantial correlations with the emotion dimensions (Tables 11.4 and 11.5). The null hypothesis that no systematic correlation between emotion dimensions and acoustic variables exists can thus be rejected.

When looking at the different groups of acoustic variables, it is obvious that the F0-related variables show the strongest effects. The measures related to tempo and to voice quality are less strongly affected, but some stable patterns emerge. Given the difficulty in measuring tempo-related variables with the ASSESS tool, and of finding voice quality-related acoustic variables in general, it is not surprising to find the correlations to be more moderate. The intensity-related variables, however, which would have been expected to be equally affected by the emotion dimensions, show hardly any correlations at all. It seems probable that this is due to the use of automatic gain control in TV recordings, which automatically maximises the sound amplitude independently of the original sound pressure level. It would seem that the intensity normalisation mechanism in ASSESS, which attempts to identify the background noise level and carries out a normalisation based on the assumption of a globally constant level of background noise, is not able to compensate for the amplitude distortions introduced by the automatic gain control of the recordings.

In the following discussion, four strengths of correlations are distinguished: *very strong* correlations correspond to a correlation coefficient $\beta > 0.25$; *strong* correlations show a β between 0.14 and 0.25; *weak* correlations correspond to a β between 0.08 and 0.14; and *very weak* correlations correspond to a β between 0.04 and 0.08. Even smaller beta values are considered too weak to be discussed and are usually not highly significant.[3]

The emotion dimension showing the most numerous and strongest correlations is the **activation** dimension. Most acoustic variables correlate with activation.

The patterns found in common for female and male speakers are the following. Very strong effects are found for F0 median, F0 range (and consequently the magnitude of F0 rises and falls), and the variability of the magnitude of F0 rises and falls within a tune, as

[3] The selected category limits are relatively arbitrary quantifications intended to structure and simplify the following presentation. They do not claim to represent any objective classification.

	Acoustic variable	Correlation Coefficients			R^2
		Activation	Evaluation	Power	
fundamental frequency	F0 median	**0.313**	**-0.056**	**-0.170**	0.130
	F0 range	**0.258**	**-0.072**		0.070
	med. magn. F0 rises	**0.195**	(-0.032)		0.039
	range magn. F0 rises	**0.304**	**-0.064**		0.095
	med. magn. F0 falls	**0.242**	**-0.085**		0.064
	range magn. F0 falls	**0.319**	**-0.121**		0.108
	med. dur. F0 rises	**0.067**	**0.058**	0.040	0.014
	rng. dur. F0 rises	**0.049**	(-0.034)	(0.032)	0.003
	med. dur. F0 falls	**0.090**	(0.030)	(0.031)	0.012
	rng. dur. F0 falls	**0.051**	**-0.087**	**0.062**	0.006
	med. slope F0 rises	**0.183**	**-0.095**	**-0.051**	0.049
	med. slope F0 falls	**0.218**	**-0.102**	**-0.045**	0.063
	F0 rises p. sec.	**-0.081**	-0.038	(-0.029)	0.011
	F0 falls p. sec.	**-0.089**	**-0.054**		0.014
tempo	duration pauses	**-0.100**	(0.035)	**-0.066**	0.013
	'tune' duration	**0.064**	(-0.024)		0.005
	intensity peaks p. sec.				0.000
	fricat. bursts p. sec.	**-0.063**		0.036	0.006
intens.	intensity median	0.029	**-0.048**	0.037	0.002
	intensity range	**0.030**			0.001
	dynamics at peaks	**0.082**	**-0.051**		0.010
voice quality	spectral slope non-fric.	**0.265**	**-0.067**	0.035	0.073
	Hamm. 'effort'	**0.108**		**0.045**	0.014
	Hamm. 'breathy'	**-0.077**			0.006
	Hamm. 'head'	**-0.059**		**-0.145**	0.025
	Hamm. 'coarse'	**-0.094**		**-0.064**	0.015
	Hamm. 'unstable'	**0.032**		**-0.099**	0.008

Table 11.4: Significant correlation coefficients for acoustic variables predicted by emotion dimensions, for **female** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

	Acoustic variable	Correlation Coefficients			R^2
		Activation	Evaluation	Power	
fundamental frequency	F0 median	**0.441**	**0.115**	(-0.056)	0.203
	F0 range	**0.294**			0.090
	med. magn. F0 rises	**0.171**			0.033
	range magn. F0 rises	**0.217**			0.046
	med. magn. F0 falls	**0.148**	0.079		0.029
	range magn. F0 falls	**0.256**		-0.121	0.077
	med. dur. F0 rises		**0.147**		0.027
	rng. dur. F0 rises	**0.113**	(0.073)		0.014
	med. dur. F0 falls		**0.170**		0.034
	rng. dur. F0 falls	**0.126**	**0.119**		0.033
	med. slope F0 rises	**0.176**	-0.078		0.034
	med. slope F0 falls	**0.190**	**-0.110**		0.045
	F0 rises p. sec.	-0.051	**-0.160**		0.034
	F0 falls p. sec.	-0.049	**-0.106**		0.023
tempo	duration pauses	**-0.141**	-0.090		0.045
	'tune' duration	**0.143**	0.066	0.075	0.045
	intensity peaks p. sec.			**-0.153**	0.019
	fricat. bursts p. sec.		**0.093**		0.006
intens.	intensity median	(0.039)			0.003
	intensity range				0.002
	dynamics at peaks	0.046	**-0.087**	**0.131**	0.012
voice quality	spectral slope non-fric.	**0.232**	**0.090**	**-0.099**	0.052
	Hamm. 'effort'	**0.176**	**0.119**	**-0.108**	0.033
	Hamm. 'breathy'		**-0.165**	**0.157**	0.014
	Hamm. 'head'	**-0.193**			0.035
	Hamm. 'coarse'	**-0.141**	**-0.170**	**0.178**	0.029
	Hamm. 'unstable'	**-0.105**	**0.101**	-0.083	0.018

Table 11.5: Significant correlation coefficients for acoustic variables predicted by emotion dimensions, for **male** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

well as the spectral slope. All these correlations are positive, i.e. higher activation corresponds to higher F0 median and range, larger F0 excursions, larger differences between the excursion sizes within a tune, and a flatter spectral slope (more high-frequency energy). Less extreme, but still strong correlations with activation were found for the slope of F0 contours (higher activation = steeper contours), pause duration (higher activation = shorter pauses), and the Hammarberg 'effort' measure. Weaker correlations were found with the number of F0 rises and falls per second (higher activation = fewer rises and falls), tune duration (higher activation = longer tunes), dynamics (higher activation = increased dynamics), and the Hammarberg measures 'head' and 'coarse' (higher activation = more chest tone and less coarse). The intensity median, which would have been expected to correlate with activation as strongly as the F0 level, showed only a very weak correlation in the expected direction.

There were relatively few differences between female and male speakers with respect to the activation dimension. For female speakers, a number of relatively weak correlations indicate that higher activation corresponds to an increased median duration of F0 rises and falls, a smaller number of fricative bursts per second, less breathiness of the voice, and, even more tentatively, an increased intensity range. For male speakers, the negative correlation of activation with the Hammarberg 'head' measure was much stronger than for female speakers, and a weak correlation with the 'unstable' measure was observed.

In summary, active emotion is accompanied by higher F0 median and range, larger, steeper and more varied F0 rises and falls, longer phrases, shorter pauses, increased intensity dynamics, a flatter spectral slope and a spectral distribution indicating higher vocal effort.

Correlations with the **evaluation** dimension are also numerous, but less strong. They also show a considerably higher degree of differences between male and female speakers.

The common patterns in speakers of both genders are a few rather weak correlations. More positive evaluation corresponds to an increased duration and a flatter contour of F0 rises and falls, a reduced number of F0 rises and falls per second, and reduced dynamics at intensity peaks.

In addition, for female speakers, some weak correlations indicate that positive evaluation is accompanied by reduced F0 range, F0 falls of reduced magnitude, and more regularity (= less variability) in the magnitude of F0 rises and falls. A number of very

weak correlations indicate that positive evaluation also seems to correspond to a lower F0 median, more regularity in the duration of F0 rises and falls, fewer F0 rises per second, a lower intensity median, and a more negative spectral slope.

The male voices show quite a different pattern. Relatively strong correlations indicate that positive evaluation is accompanied by F0 rises and falls with a longer duration, clearly fewer F0 rises per second, and clear negative effects on the Hammarberg measures 'breathy' and 'coarse'. Weak correlations indicate that for male speakers, positive evaluation is accompanied by a higher F0 median, larger F0 falls, increased variability in the duration of F0 rises and falls, shorter pauses, more fricative bursts per second, a flatter spectral slope, and increased 'effort' and 'unstable' measures.

In summary, expression of positive emotions corresponds to longer and flatter F0 movements, fewer F0 rises and reduced intensity dynamics. Male speakers, in addition, show a clear spectral pattern, with positive emotion sounding less breathy and less coarse. While for male speakers, positivity seems to correspond to a higher F0 level, larger falls and more variability, the opposite seems to be the case for female speakers: if anything, more positive emotion seems to correspond to lower F0 level and range, smaller F0 falls, and more regularity of rises and falls.

As to correlations with the **power** dimension, these are clearly less numerous, and again show a large number of differences between female and male speakers.

The only effects in common for speakers of both genders are a negative correlation of power with the F0 median and with the 'unstable' measure, i.e. higher power corresponds to a lower and more stable voice.

For female speakers, the correlation with F0 median is strong. Another strong negative correlation is found with the 'head' measure. Several very weak correlations hint that higher power could be associated with an increased variability of the duration of F0 falls, a flatter F0 contour slope for rises and falls, shorter pauses, and a less coarse voice.

For male speakers, the correlation between power and F0 median is very weak. Instead, strong correlations indicate that high power corresponds to fewer intensity peaks per second (indicating a slower speech rate), and higher 'breathy' and 'coarse' measures. Weak correlations indicate high power to be correlated to more regularity in the magnitude of F0 falls, increased intensity dynamics, a steeper negative spectral slope, and a reduced 'effort' measure. An additional very weak positive correlation is found between power

and the duration of tunes.

In summary, higher power is accompanied by a lower F0 median. In addition, female speakers seem to speak more in a chest tone, while male speakers seem to speak slower and with a more breathy and coarse voice.

Table 11.6 sums up these effects. Tables 11.7 and 11.8 show the regression coefficients corresponding to these correlations, for female and male speakers respectively.

11.4.3 Discussion

Many of the findings detailed above are in agreement with what was found in the literature survey of Chapter 10.

The present database analysis is in agreement with the literature that high activation is accompanied by higher mean F0 and F0 range, a steeper slope of F0 contours, shorter pauses, and more high-frequency energy (a flatter spectral slope). If "blaring" timbre corresponds to a flatter spectral slope, more vocal effort or a less coarse voice, then the effect described by Davitz (1964a) that higher activation corresponds to a more "blaring" timbre can be confirmed as well. The only conflicting result that was found is that the duration of tunes (inter-pause stretches) increased with activation in this analysis, while Trouvain & Barry (2000) found a decrease in duration.

Some of the results reported in the literature could be neither verified nor falsified. For speech tempo, no satisfactory measures were available in this study, and the measures used as rough approximations (intensity peaks per second and fricative bursts per second) provided no conclusive results. In addition, intensity and related measures did not produce the expected strong results. It seems likely that this is due to the automatic gain control recording technique used in TV recordings. Some measures reported in particular by Tischer (1993) could not be repeated due to limitations of the analysis package ASSESS, such as timing of intensity peaks or intensity increase during "sense units".

The present study is in line with the findings in the literature that correlations of acoustic parameters with evaluation and power are much less stable than correlations with activation. Indeed, important differences between male and female speakers were found.

The few findings from the literature concerning the acoustic correlates of evaluation do not agree well with the results found in this study. The faster speaking rate corresponding to positive evaluation found by Scherer & Oshinsky (1977) may be reflected in

	Acoustic variable	Correlations					
		Activation		Evaluation		Power	
		female	male	female	male	female	male
fundamental frequency	F0 median	⇑	⇑	↓	↑	↓	↓
	F0 range	⇑	⇑	↓			
	med. magn. F0 rises	↑	↑				
	range magn. F0 rises	⇑	↑	↓			
	med. magn. F0 falls	↑	↑	↓	↑		
	range magn. F0 falls	⇑	⇑	↓			↓
	med. dur. F0 rises	↑		↑	↑		
	rng. dur. F0 rises	↑	↑		↑		
	med. dur. F0 falls	↑			↑		
	rng. dur. F0 falls	↑	↑	↓	↑	↑	
	med. slope F0 rises	↑	↑	↓	↓	↓	
	med. slope F0 falls	↑	↑	↓	↓	↓	
	F0 rises p. sec.	↓	↓		↓		
	F0 falls p. sec.	↓	↓	↓	↓		
tempo	duration pauses	↓	↓		↓	↓	
	'tune' duration	↑	↑		↑		↑
	intensity peaks p. sec.						↓
	fricat. bursts p. sec.	↓			↑		
intens.	intensity median			↓			
	intensity range						
	dynamics at peaks	↑	↑	↓	↓		↑
voice quality	spectral slope non-fric.	⇑	↑	↓	↑		↓
	Hamm. 'effort'	↑	↑		↑	↑	↓
	Hamm. 'breathy'	↓			↓		↑
	Hamm. 'head'	↓	↓			↓	
	Hamm. 'coarse'	↓	↓		↓	↓	↑
	Hamm. 'unstable'		↓		↑	↓	↓

Table 11.6: A summary representation of the main correlation effects found for female and male speakers. Upward arrows indicate a positive correlation, downward arrows a negative correlation. ⇑ = very strong correlation ($\beta > 0.25$); ↑ = strong correlation ($0.14 < \beta < 0.25$); ↑ = weak correlation ($0.08 < \beta < 0.14$); ↑ = very weak correlation ($0.04 < \beta < 0.08$).

	Acoustic variable	Unit	Linear Regression Coefficients			
			Const.	Activation	Evaluation	Power
fundamental frequency	F0 median	Hz	200.1	**0.370**	**-0.0523**	**-0.190**
	F0 range	Hz	28.45	**0.243**	**-0.0531**	
	med. magn. F0 rises	Hz	14.57	**0.107**	(-0.0140)	
	range magn. F0 rises	Hz	20.94	**0.212**	**-0.0352**	
	med. magn. F0 falls	Hz	19.46	**0.177**	**-0.0488**	
	range magn. F0 falls	Hz	24.68	**0.244**	**-0.0727**	
	med. dur. F0 rises	sec	0.2140	**0.000379**	**0.000258**	0.000216
	rng. dur. F0 rises	sec	0.1914	**0.000305**	(-0.000164)	(0.000180)
	med. dur. F0 falls	sec	0.2448	**0.000537**	(0.000140)	(0.000173)
	rng. dur. F0 falls	sec	0.1924	**0.000311**	**-0.000409**	**0.000338**
	med. slope F0 rises	Hz/sec	76.35	**0.403**	**-0.166**	**-0.106**
	med. slope F0 falls	Hz/sec	85.85	**0.516**	**-0.191**	**-0.101**
	F0 rises p. sec.	1/sec	2.712	**-0.00593**	-0.00217	(-0.00202)
	F0 falls p. sec.	1/sec	2.486	**-0.00578**	**-0.00273**	
tempo	duration pauses	sec	0.4367	**-0.00122**	(0.000332)	**-0.000775**
	'tune' duration	sec	1.424	**0.00355**	(-0.00105)	
	intensity peaks p. sec.	1/sec	4.090			
	fricat. bursts p. sec.	1/sec	1.410	**-0.00464**		0.00247
intens.	intensity median	cB	531.2	0.0513	**-0.0667**	0.0615
	intensity range	cB	103.2	**0.149**		
	dynamics at peaks	cB	25.89	**0.0525**	**-0.0256**	
voice quality	spectral slope non-fric.	db/oct	-7.396	**0.0147**	**-0.00293**	0.00181
	Hamm. 'effort'	-	32.98	**0.0229**		**0.00898**
	Hamm. 'breathy'	-	8.084	**-0.0235**		
	Hamm. 'head'	-	24.68	**-0.0121**		**-0.0282**
	Hamm. 'coarse'	-	16.38	**-0.0178**		**-0.0114**
	Hamm. 'unstable'	-	8.297	**0.00569**		**-0.0170**

Table 11.7: Significant linear regression coefficients for acoustic variables predicted by emotion dimensions, for **female** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$.

	Acoustic variable	Unit	Linear Regression Coefficients			
			Const.	Activation	Evaluation	Power
fundamental frequency	F0 median	Hz	143.4	**0.552**	**0.117**	(-0.0676)
	F0 range	Hz	20.66	**0.218**		
	med. magn. F0 rises	Hz	15.54	**0.102**		
	range magn. F0 rises	Hz	19.50	**0.127**		
	med. magn. F0 falls	Hz	17.48	**0.0942**	0.0409	
	range magn. F0 falls	Hz	20.13	**0.163**		-0.0742
	med. dur. F0 rises	sec	0.3321		**0.00122**	
	rng. dur. F0 rises	sec	0.1714	**0.000995**	(0.000529)	
	med. dur. F0 falls	sec	0.3468		**0.00146**	
	rng. dur. F0 falls	sec	0.1609	**0.00101**	**0.000794**	
	med. slope F0 rises	Hz/sec	59.86	**0.385**	-0.139	
	med. slope F0 falls	Hz/sec	62.04	**0.331**	**-0.155**	
	F0 rises p. sec.	1/sec	2.038	-0.00302	**-0.00766**	
	F0 falls p. sec.	1/sec	1.982	-0.00293	**-0.00512**	
tempo	duration pauses	sec	0.4334	**-0.00163**	-0.000835	
	'tune' duration	sec	1.305	**0.00658**	0.00246	0.00334
	intensity peaks p. sec.	1/sec	3.841			**-0.0126**
	fricat. bursts p. sec.	1/sec	0.9083		**0.00316**	
intens.	intensity median	cB	531.1	(0.0640)		
	intensity range	cB	124.1			
	dynamics at peaks	cB	27.22	0.0272	**-0.0420**	**0.0764**
voice quality	spectral slope non-fric.	db/oct	-7.752	**0.0117**	**0.00369**	**-0.00484**
	Hamm. 'effort'	-	32.45	**0.0346**	**0.0191**	**-0.0207**
	Hamm. 'breathy'	-	5.967		**-0.0425**	**0.0483**
	Hamm. 'head'	-	27.67	**-0.0481**		
	Hamm. 'coarse'	-	16.82	**-0.0245**	**-0.0240**	**0.0301**
	Hamm. 'unstable'	-	10.85	**-0.0235**	**0.0185**	-0.0182

Table 11.8: Significant linear regression coefficients for acoustic variables predicted by emotion dimensions, for **male** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$.

the increased number of fricative bursts per second found for male speakers in this study. Where they found less high-frequency energy, we find a small tendency for the same effect for female speakers, but a slightly larger opposite effect for male speakers. The same holds for the correspondence of positive evaluation with low pitch found by Scherer & Oshinsky (1977): we find a small tendency in this direction for female speakers, but a stronger opposite effect for male speakers. The correspondence of positive evaluation and large pitch range reported by Scherer & Oshinsky (1977) cannot be confirmed in our data. Qualitative descriptions of voice quality are always difficult to interpret, but it seems unlikely that the "warm" voice quality reported as a correlate of positive evaluation by Green & Cliff (1975) corresponds well with the flatter spectral slope, higher 'effort' and 'unstable' and lower 'breathy' and 'coarse' measures we find for male speech. Finally, our analysis does not allow any comparison with the measures used by Tischer (1993), i.e. vowel durations and intensity increase during a "sense unit".

However, a relatively conclusive pattern is found in this study which was not reported in the literature: The slope of F0 rises and falls is negatively correlated with evaluation. This seems to be achieved by making slower F0 movements in male speech, and by reducing the magnitude at least of F0 falls in female speech.

Findings for power are even more difficult to compare to the literature. For example, Tischer (1993) calculated individual correlations between acoustic variables and emotion dimensions. He found power and activation to be relatively redundant, with power showing basically the same correlation patterns with many acoustic variables as activation. In the present study, however, the emotion dimensions were entered commonly as predictors into a linear regression analysis. Therefore, the correlations calculated are *partial* correlations, i.e. they represent the *unique* contribution of the predictor to the predicted variable after removing the co-variation between the predictors. It is therefore not surprising to find a correlation pattern for power which differs entirely from the activation pattern.[4] The effect reported by Tischer (1993) that high power sometimes corresponds to low F0 can be found in our data, as a strong effect for female speakers and a weak effect for male speakers. This is in line with the frequency code assumption (see 5.4, p. 57), stating that signaling behaviour reflects typical physical properties, in this case that large

[4] Analyses in which only power was used as a predictor showed that the pattern is still different from the activation pattern. This is an additional conflict between the present analysis and the results reported in the literature, but does not invalidate the argument put forward here.

individuals are typically powerful and also have a low voice. This explanation seems to fit well with the observation we made for female speech that power is negatively correlated to the 'head' measure: low power would correspond to a "head" voice, high power to a "chest" voice. It is less easy to reconcile this assumption with the voice quality-related observations made for male speech, indicating a breathier and coarser voice and a more negative spectral slope for higher power. We also found fewer, but more dynamic intensity peaks for high power in male speech. Again, the vowel duration measure used by Tischer (1993) was not available in this analysis, so that the positive correlation of vowel duration with power could be neither verified nor falsified.

In summary, the present study replicates the basic patterns of correlations between emotion dimensions and acoustic variables. In addition, it provides numerical linear regression coefficients which can serve as a starting point for the formulation of quantified emotion prosody rules. It was shown that the acoustic correlates of the activation dimension are highly stable, while correlates of evaluation and power were smaller in number and magnitude and showed a high variability between male and female speakers.

11.5 Normalisation relative to neutral speech

One way to try to reduce uncontrolled variation in the acoustic parameters is the normalisation relative to the neutral speech of the speaker. This possibility is provided by the Belfast database at least in principle, because for each speaker, a relatively neutral clip is included in the database.

11.5.1 Method

In principle, the normalisation should be carried out as follows. The ideal neutral state of the speaker, corresponding to the zero value on all emotion dimensions, would be determined, and the corresponding acoustic parameters would be measured. Normalisation of an acoustic parameter would then simply be a substraction of the neutral reference value from every emotion-specific value for the corresponding acoustic parameter, and the resulting difference values could be correlated with the emotion dimensions.

Unfortunately, this is not realistic with the data in the database. The vast majority of the "neutral" clips in the database are not perceived as entirely neutral, but only as more or

less close to the centre of the activation-evaluation space. Even if a relatively lax criterion is applied (for a clip to be accepted as representing a neutral state, the mean Feeltrace rating of a clip must be inside a circle with a radius of 25 around the neutral centre, where 100 is the radius of the entire Feeltrace circle), only 44.8% of the clips included as neutral speaker references actually qualify as neutral.

At the same time, many of the "emotional" clips are not extremely emotional, but are distributed over a wide range of emotional intensities, some of them relatively close to the neutral centre. Had the emotional clips been very far from the centre, then treating the "neutral" clips as ideally neutral would have been a legitimate approximation, because the distance between the neutral clips and the centre would have been very small compared to the distance between the emotional clips and the neutral clips. This, however, is not the case here, which makes the simple approach to normalisation untenable.

A different approach to normalisation seems more promising. Instead of correlating acoustic parameter deltas (emotional parameter value minus neutral parameter value) with emotion dimensions, it may be possible to correlate acoustic parameter deltas with emotion dimension deltas. What would that mean? It would build upon the assumption that the effect of a change in an emotion dimension on an acoustic parameter is independent of the absolute location on the emotion dimension. Only the magnitude of the change, but not the starting point of the change, would influence the acoustic parameter. A mathematical model fulfilling this constraint is a linear model.

A linear regression analysis in its basic form also starts from a linearity assumption. In the analysis of the data in absolute terms, this model was extended by using the squared versions of the emotion dimensions as additional predictors (see footnote on p. 111). In order to fulfil the linearity constraint formulated above, the use of squared terms cannot be allowed for this normalised analysis.

The linearity assumption implies that the difference between two emotional states can be described by the same mechanism as the difference between an emotional and a neutral state. If the assumption is justified, the location of the two states would be irrelevant for their differences in acoustic parameter values, only their distance on each of the emotion dimensions would have an effect. In order to test this assumption, the normalised analysis of the entire database was accompanied by a second, restricted analysis, in which only those "neutral" reference clips were considered which were close to the neutral state,

more precisely not further than 15/100 away from the centre of the activation-evaluation space. As this criterion was fulfilled only by 22.1% of the "neutral" clip ratings, the data basis for this normalised analysis is considerably smaller than for the unrestricted analysis (see Table 11.3). If the assumption holds, then the results in the second, restricted test should not be better than in the first, unrestricted test.

11.5.2 Results and discussion

It is to be expected that a smaller data set is easier to model, and therefore the correlations for the restricted test are expected to be larger than for the unrestricted test. This can indeed be observed when comparing Table 11.9 with Table 11.11 for female speech and Table 11.10 with Table 11.12 for male speech. The different data set sizes therefore make it impossible to draw any conclusions from a direct comparison of the normalised correlation coefficients in the two tests. It is, however, possible in each test to compare the normalised analysis with an analysis of the same data set in terms of absolute values.

It is obvious from Tables 11.9 and 11.10 that for the unrestricted test, the delta measures lead in nearly all cases to worse results than the absolute measures. The correlations and the explained variance R^2 are smaller than for the analysis of the same data set in terms of absolute values.

The normalised analyses based on a restricted data set show a different picture (Tables 11.11 and 11.12). The correlations and explained variance R^2 of the normalised analysis approach those for the analysis of the same restricted data set in terms of absolute values. In particular for male speech (Table 11.12), several variables actually show larger correlations and explained variance R^2 than for the absolute analysis.

These observations indicate that the linearity assumption in the strong form as formulated above, postulating a total independence of the acoustic effect from the absolute location on the emotion dimensions, does not seem to be supported by the data. First, the basic aim of the normalisation procedure, namely an improvement of the model with respect to the unnormalised analysis, is not met in the unrestricted analysis; second, the model improves (relative to an unnormalised analysis of the same data) when only a restricted set of good examples of "neutral" states are admitted as reference values.

Only in the smallest data set, namely the restricted male data (Table 11.12), can a real improvement compared to the unnormalised analyses be observed for 13 out of the 27

	Acoustic variable	Absolute Values				Normalised (Delta) Values			
		Correlation Coefficients			R^2	Correlation Coefficients			R^2
		Act.	Eval.	Pow.		ΔAct.	ΔEval.	ΔPow.	
fundamental frequency	F0 median	**0.313**	**-0.056**	**-0.170**	0.130	**0.280**	**-0.059**	-0.030	0.081
	F0 range	**0.258**	**-0.072**		0.070	**0.192**	-0.031	**0.069**	0.042
	med. magn. F0 rises	**0.195**	(-0.032)		0.039	**0.135**	**-0.052**		0.020
	range magn. F0 rises	**0.304**	**-0.064**		0.095	**0.208**	**-0.092**		0.047
	med. magn. F0 falls	**0.242**	**-0.085**		0.064	**0.151**	**-0.074**		0.026
	range magn. F0 falls	**0.319**	**-0.121**		0.108	**0.191**	**-0.079**	0.051	0.039
	med. dur. F0 rises	**0.067**	**0.057**	0.041	0.014	**0.047**			0.004
	rng. dur. F0 rises	**0.049**	(-0.034)	(0.032)	0.003		**-0.080**	**0.076**	0.005
	med. dur. F0 falls	**0.090**	(0.030)	(0.031)	0.012	**0.062**			0.004
	rng. dur. F0 falls	**0.051**	**-0.087**	**0.062**	0.006	0.035	**-0.054**	**0.065**	0.004
	med. slope F0 rises	**0.183**	**-0.095**	**-0.051**	0.049	**0.112**	**-0.074**	(-0.023)	0.019
	med. slope F0 falls	**0.218**	**-0.102**	**-0.045**	0.063	**0.116**	**-0.059**		0.017
	F0 rises p. sec.	**-0.081**	-0.037	(-0.029)	0.011	**-0.065**	-0.035		0.006
	F0 falls p. sec.	**-0.089**	**-0.053**		0.014	**-0.088**	(-0.023)		0.010
tempo	duration pauses	**-0.100**	(0.035)	**-0.066**	0.013				
	'tune' duration	**0.064**	(-0.024)		0.005	-0.024		**-0.072**	0.006
	intensity peaks p. sec.				0.000			(-0.022)	0.001
	fricat. bursts p. sec.	**-0.063**		0.036	0.006				
intens.	intensity median	0.029	**-0.048**	0.037	0.002	**0.126**	**-0.098**		0.022
	intensity range	**0.030**			0.001				0.000
	dynamics at peaks	**0.082**	**-0.051**		0.010	**0.050**	-0.036	**0.047**	0.004
voice quality	spectral slope non-fric.	**0.266**	**-0.067**	0.035	0.073	**0.188**	(-0.027)		0.035
	Hamm. 'effort'	**0.108**		**0.045**	0.014	**0.183**	**-0.037**	0.031	0.035
	Hamm. 'breathy'	**-0.077**			0.006	**-0.077**	**-0.047**		0.011
	Hamm. 'head'	**-0.059**		**-0.145**	0.025	-0.029			0.002
	Hamm. 'coarse'	**-0.095**		**-0.064**	0.015	**-0.081**	**-0.042**	(-0.028)	0.011
	Hamm. 'unstable'	**0.032**		**-0.099**	0.008	**0.052**	**0.038**		0.005

Table 11.9: Comparison of correlations in absolute and normalised variables based on an **unrestricted** data set, for **female** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

	Acoustic variable	Absolute Values				Normalised (Delta) Values			
		Correlation Coefficients			R^2	Correlation Coefficients			R^2
		Act.	Eval.	Pow.		ΔAct.	ΔEval.	ΔPow.	
fundamental frequency	F0 median	**0.441**	**0.115**	(-0.056)	0.203	**0.334**	**-0.105**	**-0.230**	0.168
	F0 range	**0.294**			0.090	**0.254**	(-0.055)	**-0.094**	0.070
	med. magn. F0 rises	**0.171**			0.033	**0.123**	**-0.103**		0.027
	range magn. F0 rises	**0.217**			0.046		**-0.177**	-0.108	0.064
	med. magn. F0 falls	**0.148**	0.079		0.029	**0.115**	**-0.078**	-0.067	0.026
	range magn. F0 falls	**0.256**		-0.121	0.077	**0.182**			0.041
	med. dur. F0 rises		**0.147**		0.027	-0.051	**-0.129**		0.016
	rng. dur. F0 rises	**0.113**	(0.073)		0.014			-0.094	0.016
	med. dur. F0 falls		**0.170**		0.034				0.000
	rng. dur. F0 falls	**0.126**	**0.119**		0.033	**0.096**			0.011
	med. slope F0 rises	**0.176**	-0.078		0.034	**0.146**	(-0.049)		0.024
	med. slope F0 falls	**0.190**	**-0.110**		0.045	**0.125**	**-0.118**	**-0.078**	0.040
	F0 rises p. sec.	-0.051	**-0.160**		0.034		**-0.100**		0.014
	F0 falls p. sec.	-0.049	**-0.106**		0.023				0.006
tempo	duration pauses	**-0.141**	-0.090		0.045				
	'tune' duration	**0.143**	0.066	0.075	0.045		**0.208**	**0.099**	0.077
	intensity peaks p. sec.			**-0.153**	0.019		(-0.046)	**-0.123**	0.024
	fricat. bursts p. sec.		**0.093**		0.006				
intens.	intensity median	(0.039)			0.003	**0.113**			0.013
	intensity range				0.002				0.000
	dynamics at peaks	0.046	**-0.087**	**0.131**	0.012		**-0.098**	**0.083**	0.007
voice quality	spectral slope non-fric.	**0.232**	**0.090**	**-0.099**	0.052	**0.148**			0.022
	Hamm. 'effort'	**0.176**	**0.119**	**-0.108**	0.033	**0.182**			0.034
	Hamm. 'breathy'		**-0.165**	**0.157**	0.014	**-0.072**		(-0.049)	0.013
	Hamm. 'head'	**-0.193**			0.035	**-0.103**	(0.051)	**0.120**	0.030
	Hamm. 'coarse'	**-0.141**	**-0.170**	**0.178**	0.029	**-0.132**			0.017
	Hamm. 'unstable'	**-0.105**	**0.101**	-0.083	0.018		(0.052)	**0.099**	0.019

Table 11.10: Comparison of correlations in absolute and normalised variables based on an **unrestricted** data set, for **male** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

	Acoustic variable	Absolute Values				Normalised (Delta) Values			
		Correlation Coefficients			R^2	Correlation Coefficients			R^2
		Act.	Eval.	Pow.		ΔAct.	ΔEval.	ΔPow.	
fundamental frequency	F0 median	**0.357**	**-0.090**	**-0.258**	0.220	**0.377**	**-0.134**		0.165
	F0 range	**0.260**	(-0.067)		0.078	**0.255**	**-0.083**		0.070
	med. magn. F0 rises	**0.189**		(-0.063)	0.039	**0.199**	**-0.093**		0.050
	range magn. F0 rises	**0.313**	(-0.086)	(-0.079)	0.115	**0.305**	(-0.075)		0.098
	med. magn. F0 falls	**0.241**	-0.093		0.073	**0.239**	**-0.104**		0.072
	range magn. F0 falls	**0.318**	**-0.145**		0.124	**0.212**	-0.116		0.052
	med. dur. F0 rises	0.065			0.008	**0.084**	(-0.063)		0.010
	rng. dur. F0 rises	(0.048)			0.003		-0.097		0.006
	med. dur. F0 falls	**0.079**			0.007	**0.071**			0.006
	rng. dur. F0 falls		(-0.085)	(0.073)	0.004		**-0.114**	**0.132**	0.010
	med. slope F0 rises	**0.193**	(-0.068)	**-0.105**	0.061	**0.168**	-0.072		0.038
	med. slope F0 falls	**0.223**	**-0.113**	-0.079	0.080	**0.192**	**-0.103**	(-0.056)	0.058
	F0 rises p. sec.	**-0.085**			0.008	**-0.073**			0.006
	F0 falls p. sec.	**-0.092**			0.009	**-0.086**			0.008
tempo	duration pauses	**-0.116**			0.016				
	'tune' duration				0.001				0.003
	intensity peaks p. sec.	(-0.041)			0.002				0.001
	fricat. bursts p. sec.	**-0.082**		(0.056)	0.010				
intens.	intensity median	0.051	**-0.100**	0.076	0.008	**0.074**	**-0.090**		0.011
	intensity range				0.001				0.001
	dynamics at peaks		**-0.148**	0.092	0.012		-0.086	**0.095**	0.006
voice quality	spectral slope non-fric.	**0.260**	**-0.095**		0.081	**0.244**	**-0.170**	(0.065)	0.079
	Hamm. 'effort'	**0.083**			0.009	**0.190**	**-0.173**	**0.101**	0.055
	Hamm. 'breathy'			0.079	0.003	**-0.124**	(0.054)		0.018
	Hamm. 'head'			**-0.155**	0.028	(-0.045)	**0.096**	**-0.150**	0.015
	Hamm. 'coarse'				0.006	**-0.125**	**0.093**	**-0.113**	0.024
	Hamm. 'unstable'			**-0.154**	0.016	**0.084**			0.008

Table 11.11: Comparison of correlations in absolute and normalised variables based on a **restricted** data set, for **female** speech. Only "neutral" clips with a distance $< 15/100$ from the center are included. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

	Acoustic variable	Absolute Values				Normalised (Delta) Values			
		Correlation Coefficients			R^2	Correlation Coefficients			R^2
		Act.	Eval.	Pow.		ΔAct.	ΔEval.	ΔPow.	
fundamental frequency	F0 median	**0.458**			0.205	**0.499**	**-0.238**	**-0.215**	0.376
	F0 range	**0.384**			0.144	**0.386**	-0.155		0.183
	med. magn. F0 rises	**0.284**			0.081	**0.260**	-0.177		0.113
	range magn. F0 rises	**0.324**		-0.200	0.136			-0.263	0.053
	med. magn. F0 falls	**0.249**	**0.202**	**-0.178**	0.075	**0.225**			0.080
	range magn. F0 falls	**0.341**	(-0.158)		0.172	**0.384**			0.153
	med. dur. F0 rises		(0.109)		0.010		-0.168		0.025
	rng. dur. F0 rises	0.136			0.020				0.004
	med. dur. F0 falls		0.157		0.028				0.003
	rng. dur. F0 falls	**0.175**			0.046	**0.196**			0.049
	med. slope F0 rises	**0.255**			0.067	**0.260**			0.075
	med. slope F0 falls	**0.283**		**-0.164**	0.084	**0.143**		(-0.122)	0.058
	F0 rises p. sec.				0.006	**-0.170**			0.030
	F0 falls p. sec.	(-0.084)	-0.149		0.024	-0.124			0.021
tempo	duration pauses	**-0.198**		-0.139	0.056				
	'tune' duration	**0.131**		(0.104)	0.043	**0.110**	**0.322**		0.153
	intensity peaks p. sec.			**-0.181**	0.024	(-0.087)	(-0.103)		0.047
	fricat. bursts p. sec.				0.005				
intens.	intensity median	0.106	**-0.163**		0.054	**0.200**		-0.157	0.049
	intensity range				0.021	**-0.128**			0.030
	dynamics at peaks			**0.180**	0.025				0.003
voice quality	spectral slope non-fric.	**0.171**			0.030	**0.139**	(-0.124)		0.030
	Hamm. 'effort'	**0.184**		**-0.244**	0.062	**0.273**			0.084
	Hamm. 'breathy'		**-0.257**	**0.299**	0.056			(-0.137)	0.017
	Hamm. 'head'	**-0.185**	(0.115)		0.043	-0.117	0.149	(0.122)	0.072
	Hamm. 'coarse'		**-0.162**	**0.240**	0.029	-0.118	(0.133)		0.026
	Hamm. 'unstable'	**-0.142**	**0.254**	**-0.253**	0.068			0.165	0.037

Table 11.12: Comparison of correlations in absolute and normalised variables based on a **restricted** data set, for **male** speech. Only "neutral" clips with a distance $< 15/100$ from the center are included. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$. R^2 is the fraction of the total variance in the acoustic parameter explained by the emotion dimensions.

variables. As the amount of data upon which this pattern is based is small when compared to the other analyses where no such pattern was found, it would seem premature to claim an improvement of the analysis as a result of the normalisation procedure.

In summary, it would appear that the simple approaches to a speaker normalisation of the data which were explored here can not improve the model reliably. The fact that many of the clips included as "neutral" speaker references were perceived as relatively emotional made then unsuitable as simple references. The alternative which was explored, namely the assumption of a linear model and a subsequent correlation of acoustic delta measures with emotion delta measures, led to poor results, and it only approached (and, for the smaller male data set, sometimes exceeded) the unnormalised analyses when the model was restricted to good examples of neutral states serving as references. It would therefore seem that the best data source to be used in subsequent chapters are the unrestricted, unnormalised analyses as calculated in 11.4.

11.6 Excursion: Accounting for the acoustic correlates of emotion categories

Although it might not be relevant for the main line of investigation in this thesis, it may be worth following an interesting link to the wider literature on vocal emotion expression, which can be established through the positioning of emotion categories in the 3-dimensional emotion space studied here. Table 11.13 shows the positions for the categories of a Basic English Emotion Vocabulary proposed by Cowie et al. (1999a). For activation and evaluation, these positions are mean Feeltrace rating positions for clips which were assigned the given verbal emotion label by the same rater; for power, the value was associated with the verbal emotion label in Cowie et al. (1999a).

The linear regression coefficients presented in Tables 11.7 and 11.8 can predict the values of the acoustic variables for each point in the 3-dimensional emotion space. In particular, they can predict these values at the co-ordinates of a given emotion category. The resulting prediction can be compared to the findings in literature about the acoustic properties of that emotion category. This allows the distinction between acoustic settings that are due to general features of emotion, as expressed in the emotion dimensions, and acoustic settings that are due to more specialised features of that emotion.

Emotion	Activation	Evaluation	Power
neutral	1.8	-1.7	0
bored	-6.8	-17.9	-55.3
disappointed	2.4	-24.9	-37.2
sad	-17.2	-40.1	-52.4
worried	4.6	-26.3	-62.3
afraid	14.8	-44.4	-79.4
angry	34.0	-35.6	-33.7
interested	16.8	16.6	-6.1
excited	36.1	30.5	-5.8
loving	1.2	33.3	14.9
affectionate	0.7	37.3	21.4
pleased	19.0	38.6	51.9
confident	13.8	14.1	32.9
happy	17.3	42.2	12.5
amused	23.4	16.8	-5.0
content	-14.9	33.1	12.2
relaxed	-18.5	25.7	-5.2

Table 11.13: Positions on the three emotion dimensions for some emotion categories

For example, anger is described in literature as showing a very much higher pitch average compared to neutral speech, much wider pitch range, abrupt pitch changes on stressed syllables, slightly faster speech rate, higher intensity, "chest tone" voice quality and tense articulation (Murray & Arnott, 1993). The position of 'angry' in the emotion dimensions (A=34.0, E=-35.6, P=-33.7) predicts higher F0 median[5]; wider F0 range; steeper F0 rises and falls; higher intensity; and a flatter spectral slope, i.e. more high frequency energy, corresponding to a tenser voice quality. Thus, much of what is presented as typical for anger can be predicted based on the three emotion dimensions.

The regression coefficients linking acoustic parameters to emotion dimensions also provide the possibility to track acoustic similarities between emotions to shared emotional characteristics. Comparing anger to fear, for example, it can be seen that the two share acoustic properties as well as emotional properties. Acoustically, according to Murray & Arnott (1993), fear is similar to anger in pitch average, pitch range, speech rate and articulation precision. Fear differs from anger in that pitch changes are not steeper than for neutral, the speech rate is even faster, the intensity is only normal, and voicing is irregular.

[5]Calculation example for female speech: F0 median = 200.1 + 0.370*A - 0.0523*E - 0.190*P = 220.9 Hz, compared to 200.1 Hz for neutral speech. See Table 11.7.

On the three emotion dimensions, the two emotions are very close on the activation and evaluation dimension, but differ in power.

A stepwise application of the regression rules can now account for some of the acoustic similarities in terms of emotion dimensions: In a first step, the acoustic correlates of the activation and evaluation levels, relatively similar for anger and fear, are calculated. This leads to an increase of F0 median and range compared to neutral, steeper F0 rises and falls, increased intensity, and a flatter spectral slope. In a second step, the rules for the power dimension are applied using the different values for anger and fear. The power value for the latter being more extreme, the effects for fear are stronger than for anger. According to the regression rules, this leads to an even higher F0 median for fear compared to anger; even steeper F0 rises and falls for female voices; and a lower intensity for female voices. Most of this is in accordance with the acoustic profiles as summarised in Murray & Arnott (1993).

Similarly, the acoustic similarities of anger and happiness (higher pitch average, pitch range and intensity according to Murray & Arnott (1993)) can be "explained" by the similarity in activation level.

The method sketched here can account for some of the acoustic correlates of an emotion category in terms of the position of that category on emotion dimensions. It is undisputed that the emotion is only partially characterised by these emotion dimensions, and that human listeners can distinguish fine categorical distinctions between emotion categories (Banse & Scherer, 1996; Leinonen et al., 1997). It is not clear, however, to what extent acoustic measures capture the perceptive cues which characterise the emotion. The method proposed above can identify acoustic parameter settings which are generic in the sense of general properties of the emotional state as captured by emotion dimensions. If all known acoustic properties of an emotional state can be accounted for by this method, then either the state lacks typical identifying acoustic properties, or these are not yet determined.

11.7 Conclusions

This study has explored an emotional speech database in view of correlations between emotional characteristics in terms of emotion dimensions, obtained by listener judgement,

and acoustic properties measured using a semi-automatic tool. The correlations have been shown to be numerous, which in itself is an interesting result in a corpus as heterogeneous as the one used. In addition, in the light of the findings reported in the literature, the linear regression coefficients obtained seem to allow reasonable predictions of acoustic variable values.

In view of the application of these results for speech synthesis, it seems highly probable that activation can be successfully conveyed through the parameters found in this study, if they are complemented by the results from the literature regarding speech tempo and intensity. It seems rather improbable that the same can be achieved for evaluation and power.

Chapter 12

The MARY text-to-speech system

One major goal of the present work is to put the ideas formulated in the previous two chapters into practise, in the context of the automatic generation of synthetic emotional speech from text annotated with the appropriate emotion. Such a task requires an adequate tool, capable of controlling the acoustic parameters influenced by the emotional state. This chapter explains why the MARY system is a suitable tool for that purpose. The overall system architecture is described before the aspects of particular relevance for emotion expression are pointed out.

12.1 Introduction

This chapter[1] presents the German text-to-speech system MARY (**M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis), which is a flexible tool for research, development and teaching in the domain of text-to-speech (TTS) synthesis.[2]

MARY allows for a step-by-step processing with an access to partial processing results. In this respect, MARY is similar to the TTS system and interface DRESS developed in Dresden (Hoffmann et al., 1999), also for German. However, as the MARY system uses an XML-based data representation, it does not only display the intermediate processing results, but also allows for their modification by the user. Thereby, the user is given the

[1] An earlier version of this chapter is accepted for publication as: Schröder, M. & Trouvain, J. (to appear). The German text-to-speech synthesis system MARY: A tool for research, development and teaching, in *International Journal of Speech Technology*.

[2] The system is accessible online under `http://mary.dfki.de`. Notice that the web interface is visually different from the interface described in this chapter, but provides nearly identical functionality.

opportunity to interactively explore the effects of a specific piece of information on the output of a given processing step.

MARY is composed of distinct modules and has the capability of parsing speech synthesis markup such as SABLE (Sproat et al., 1998). These features are also found in FESTIVAL (Black et al., 1999), an open source TTS system designed for multi-lingual use. The modular design of FESTIVAL allows everybody to write their own modules which can be plugged into the system. For German, a text normalisation and pre-processing module for FESTIVAL is provided by IMS Stuttgart[3] (Breitenbücher, 1999). FESTIVAL is excellent for getting an in-depth understanding of the technical aspects of text-to-speech synthesis. In contrast, MARY provides a web interface accessible from everywhere with no need to install the system locally. This makes it more suitable for those with an interest in the linguistic aspects of the input and output of the individual modules who do not need access to the technical details of the system.

The article is structured as follows. First, the properties of the system-internal XML representation are described. Then, a detailed account of the system structure is given, including a short presentation of each module. After that, the user interface which allows the user to display and edit intermediate processing results is described. Finally, the suitability of the MARY system for emotion expression is motivated.

12.2 The MaryXML markup language

Throughout the MARY system, an internal, low-level markup called MaryXML is used, which reflects the modelling capabilities of this particular TTS system. MaryXML is based on XML (eXtensible Markup Language) (Harold, 1999). A DTD (Document Type Definition) formally specifies the structure of a correct MaryXML document.

As the use of this internal XML language is fundamental for the flexibility of the MARY system, its properties are discussed before the system as such is presented.

12.2.1 Positioning the markup language

Because of the growing number of XML-based markup languages related in different ways to speech synthesis, it may be necessary to position MaryXML with respect to these existing markups.

[3] http://www.ims.uni-stuttgart.de/phonetik/synthesis

One group of markup languages provides relatively high-level markup functionality for speech synthesis input, intended for the use of non-experts. Early examples for this group include the original SSML (speech synthesis markup language, (Taylor & Isard, 1997)) and STML (spoken text markup language, (Sproat et al., 1997)) as well as Sun Microsystems' JSML (Java speech markup language (JSML, 1999)). Out of these, SABLE (Sproat et al., 1998) was developed, for which parsers exist e.g. in the Bell Labs system (Sproat, 1997) and in FESTIVAL. More recent additions to this family of high-level markups are the XML-based markup language coming with Microsoft's SAPI (Speech API) 5 (Microsoft, 2002) and the new W3C SSML (speech synthesis markup language, (Walker & Hunt, 2001)) which is still in draft status. All of these markup languages are, except for superficial syntactic differences, functionally similar: They aim at giving a non-expert user the possibility to add information to a text in order to improve the way it is spoken. These markup languages are (at least in principle) independent of any particular TTS system. A specific system is assumed to parse the markup language in its input and translate the information contained in it into a system-internal data representation format which in most cases is not XML-based.

A recent addition to the landscape of markup languages, with a huge commercial potential, is VoiceXML (VoiceXML, 2000). This markup language combines parts of the functionality of speech synthesis markup languages such as SABLE with speech recognition, dialogue management and touchtone dialing functionalities. Its main focus is to provide the necessary tools for a speech access to the World Wide Web.

MaryXML belongs to a different category, and might rather be called a representation language than a markup language. Its purpose is to serve as the data representation format inside the TTS system. For that reason, the concepts represented in it are low-level, detailed, and specific to the design decisions, modules, and scientific theories underlying the TTS system. By means of the Document Object Model (DOM), a standardised object-oriented representation of an XML document, the TTS system modules operate directly on the XML document, interpreting and adding information. Currently the MARY system as well as the BOSS system (Klabbers et al., 2001) follow this approach.

12.2.2 Advantages and disadvantages

The system-internal XML representation enables the MARY system to provide access to intermediate processing results. Technically, this is realised as follows. Through what is called *serialisation*, a standard operation on DOM XML representations, the current state of the document can be made externally visible in the form of a textual XML document at any stage of processing. As the external XML document contains the complete data which was available at the intermediate processing step where serialisation occurred, the inverse process is also possible: *deserialisation*, i.e. a textual XML document corresponding to an intermediate processing step is parsed by the system, and processing can continue from that step onwards. An expert can edit the textual XML document before feeding it back into the system and thus control all aspects of system data.

A second benefit of using a system-internal XML representation is that it is very easy to parse a speech synthesis input markup language such as SABLE, as this amounts to the translation of one XML format into another (see 12.3.1).

A structural limitation inherent to XML in general should be mentioned. XML documents enforce an unambiguous tree structure, in which one element is always fully embedded in another one. Therefore, it is not possible to represent two non-embedding structures (e.g., syntactic and prosodic structure) simultaneously via structural XML elements. One method of circumventing this problem is to represent one of the structures via hierarchically structured elements while representing the other structure in a "flat" form (see 12.3.4).

12.2.3 Syntax

The syntax (in a formal technical, not in a linguistic sense) of a MaryXML document reflects the information required and provided by the modules in the TTS system. Those units of information which can also be encoded in speech synthesis input markup languages, such as sentence boundaries and global prosodic settings, are represented by the same tags as in a standard representant of that group of markup languages. At the time of system design, SABLE was chosen as a model for these tags; in the future, if the W3C SSML becomes an established standard, the MaryXML tags should be adapted to their SSML equivalents.

Most of the information to be represented in MaryXML, however, is too detailed to be expressed using tags from input markup languages, for the reasons outlined in 12.2.1 above. Specific MaryXML tags need to represent the low-level information required during various processing steps. This encompasses mainly *tokens* along with their textual form, part of speech, phonological transcription, pitch accents etc., as well as prosodic phrasing. In order not to clutter up this paper with technical details, only a selection of tags is introduced as the modules requiring them are discussed. A full DTD for MaryXML is available online.[4]

12.2.4 Future

The emergence of system-internal markup languages in recent systems such as MARY, BOSS and possibly others opens interesting new lines of thought geared towards connecting TTS systems. If it were possible to define at least a minimal standard TTS architecture with clearly defined XML-based data representations at the interfaces, this would open up the possibility to interconnect modules from different TTS systems and thus work towards a "plug-and-play" TTS architecture. Many problems regarding the details of such work can be anticipated, as each system will differ substantially with respect to the types of data represented internally, both fundamentally (e.g., target-based vs. contour-based descriptions of intonation) and in detail (e.g., the tag sets used for part-of-speech annotation). Still, it would seem worthwhile pursuing this idea even if only a subset of system-internal information is transferable via such standardised interfaces.

12.3 Structure of the TTS system

In principle, the modular design of the MARY system allows arbitrary system architectures. An architectural frame is defined via the notion of *data types*, specifying the data format serving as input and/or output to processing modules. Each module knows about the data type it requires as its input and the data type it produces as its output. Two modules can be "plugged" together if the first module's output type is equal to the second module's input type.

[4] The DTD can be found at `http://mary.dfki.de/lib/MaryXML.dtd`. An XML Schema-based definition of MaryXML is planned; the Schema will reside at `http://mary.dfki.de/lib/MaryXML.xsd`.

Using this frame, an example architecture has been implemented for German TTS. Nothing limits the system from being extended to other languages: It suffices to define new data types corresponding to the intermediate processing steps sensible for that language (e.g., *text, preprocessed, phonemised* and *audio*), and to provide a chain of processing modules connecting these new data types (e.g., *preprocessor, phonemiser* and *waveform synthesiser*).

In the following, the current German TTS architecture within the MARY system is described. Not surprisingly, it is similar to a typical TTS architecture as described by Dutoit (1997). Figure 12.1 shows the individual processing modules, the flow of information and the intermediate results corresponding to data types defining the interfaces between the modules.

In the following, each of the modules will be briefly presented.

12.3.1 Optional markup parser

The MARY text-to-speech and markup-to-speech system accepts both plain text input and input marked up for speech synthesis with a speech synthesis input markup language such as SABLE.

The input markup language, presently SABLE and the W3C draft version of SSML, is translated by this module into the system-internal MaryXML format, upon which subsequent modules will operate.

As an example, an `<EMPH>...</EMPH>` SABLE tag requesting moderate emphasis for the enclosed words is translated into low-level settings such as, e.g., a raised F0 level, reduced speed, and an obligatory pitch accent for every enclosed word.[5] These settings are expressed in the MaryXML annotation and reflect the capabilities of the following modules to influence the utterance realisation. This module only determines the fact *that*, e.g., a pitch accent must be present, whereas the corresponding specialised module will determine at a later stage *which* accent to realise on that word.

The realisation indications expressed in the input markup are considered as supplements to the modules' text-to-speech analysis of the input. Each module adds new or

[5] These prosodic settings are meant to realise the abstract concept of emphasis, which does not seem to be a clearly defined concept and which appears to encompass concepts as different as contrasting accentuation and paralinguistic intensification. Not least because of these conceptual difficulties, the parameters selected for the realisation of emphasis are currently based on linguistic intuition rather than hard scientific evidence.

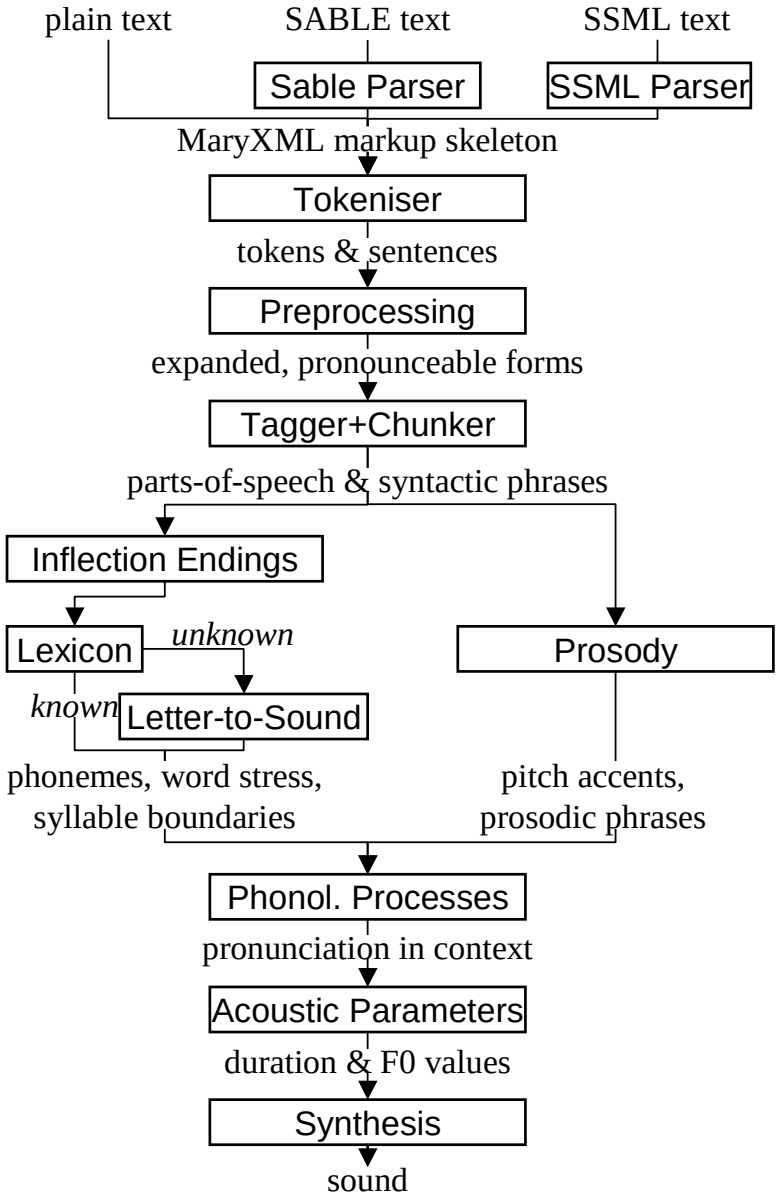


Figure 12.1: The architecture of the MARY TTS system.

more detailed information. For example, if the prosody module does not get information from its input on the locations and types of accents and boundaries, it will use its default rules (see 12.3.6) to determine them. If it finds partial information in its input, such as the location, but not the type of an accent, it will apply its rules to fill in the missing piece of information.

Technically, the markup parser's task of translating one XML format into another is performed using a specialised XSLT (eXtensible Stylesheet Language Transformation) stylesheet (Harold, 1999). This technique allows a very simple adaptation to new markup languages such as the upcoming W3C Speech Synthesis Markup Language SSML (Walker & Hunt, 2001), as only the stylesheet defining the translation into MaryXML needs to be adapted.

12.3.2 Tokeniser

The tokeniser cuts the text into tokens, i.e. words, numbers, special characters and punctuation marks. It uses a set of rules determined through corpus analysis to label the meaning of dots based on the surrounding context. In order to disambiguate the meaning of dots, which can be sentence-final periods, decimal number delimiters, parts of ordinal numbers, or abbreviation points, the rules collect evidence from the surrounding context on the role(s) which the dot can or cannot fulfill. For example, a dot preceded directly by a number and followed by whitespace and a lower-case character is not a sentence-final period.

Each token is enclosed by a `<t>...</t>` MaryXML tag. All local information about a token determined by subsequent processing steps is added to that token's `<t>` tag as attribute/value pairs. In addition, punctuation signs, including those dots which are identified as sentence-final periods, are used to determine start and end of sentences, which are marked using the MaryXML `<div>...</div>` tag enclosing a sentence.

12.3.3 Text normalisation

In the text normalisation module, those tokens for which the spoken form does not entirely correspond to the written form are replaced by a more pronounceable form.[6]

[6] An excellent overview of the phenomena that need to be accounted for in German text normalisation has been given by Breitenbücher (1999).

Numbers

The pronunciation of numbers highly depends on their meaning. Different number types, such as cardinal and ordinal numbers, currency amounts, or telephone numbers, must be identified as such, either from input markup or from context, and replaced by appropriate token strings.

While the expansion of cardinal numbers is straightforward, the expansion of ordinal numbers poses interesting problems in German, because of their inflections. On the one hand, the expansion of an ordinal number depends on its part-of-speech (adverb or adjective); on the other hand, for adjective ordinals, the inflection ending depends on gender, number and case of the noun phrase which the ordinal belongs to. In the text normalisation module, none of that information is available, so the ordinal number is simply marked as such, and a stem expansion is given. For example, the ordinal "1." would become "erstens" (Engl. "first (adverb)") in adverbial position ("denn 1. ist das...") and "erste/ersten/erstes/erster" in adjectival position. This module adds the information `ending="ordinal"` and `sounds_like="erste"` to the ordinal's `<t>` tag. Based on this markup, the correct ending will be selected during phonemisation (see 12.3.5 below).[7]

Abbreviations

Two main groups of abbreviations are distinguished: Those that are spelled out, such as "USA", and those that need expansion. The first group of abbreviations are correctly pronounced by spelling rules.

The second group is pronounced using an expansion table, containing a graphemic and optionally a phonemic expansion. The graphemic expansion is used for abbreviations such as "bzw.", expanded as "beziehungsweise" (Engl. "respectively"), or "BAföG" (a German government scholarship), expanded as "Bafög" and left to be treated by the default letter-to-sound conversion algorithm (see 12.3.5 below). The phonemic expansion is useful for non-standard pronunciations such as "SFOR" (pronounced [ˈɛs-foːɐ]), and for foreign abbreviations, such as "FBI" which is pronounced as the English spelling [ɛf-biː-ˈaɪ] in German.

[7] A different solution for this problem, employing a sentence grammar, is used in the SVOX system (Traber, 1993).

One group of abbreviations, such as “engl.”, pose a problem similar to ordinal numbers: Depending on the context, they can be adverbs (“englisch”), or to-be-inflected adjectives (“englische/n/s/r”). This group is specially marked in the expansion table and consecutively in the markup (`ending="adjadv" sounds_like="englisch"`) for later processing (see 12.3.5 below).

Tokens which are identified as abbreviations but for which no entry in the expansion table is found are either spelled out, if they consist of no more than five characters, or left to be pronounced like normal words by the phonemisation component (see 12.3.5) if they are longer.

12.3.4 Part-of-speech tagger / chunk parser

Part-of-speech tagging is performed with the statistical tagger TnT (Brants, 2000), using the Stuttgart-Tübingen Tagset (STTS) (Schiller et al., 1995), and trained on the manually annotated NEGRA corpus (Skut et al., 1997). A chunk parser (Skut & Brants, 1998) is used to determine the boundaries of noun phrases, prepositional phrases and adjective phrases. In addition to punctuation and part-of-speech, this information about syntactic phrasing is useful for determining correct prosodic phrasing (see 12.3.6). Furthermore, syntactic phrases are used to delimit the domain for morphological unification, a prerequisite for assigning the correct inflection ending to expanded abbreviations and ordinal numbers (see 12.3.5).

Part-of-speech and chunking information is added to each token’s `<t>` tag. For the chunking information, this is not actually a very satisfactory solution, as the local syntactic structure can hardly be considered a property of the individual token. However, the more logical representation of syntactic structure as an XML tree structure would possibly conflict with the prosodic structure, due to the fact that syntactic and prosodic structure cannot be guaranteed to coincide in all cases. As XML only allows for a proper tree structure, with no crossing edges, the only alternative seems to be to give up XML representation in the present form in favour of, e.g., a chart representation allowing more flexible edges. However, the encoding used at present, with the XML structure representing prosodic structure and syntactic structure “squeezed” into the token tags, seems to be a viable solution.

12.3.5 Phonemisation

The SAMPA phonetic alphabet for German (Wells, 1996) is used for the phonemic transcription. An extensive lexicon deals with known words, and a letter-to-sound conversion algorithm with unknown words; but first, a dedicated module adds inflection endings to ordinals and abbreviations.

Inflection endings

This module deals with the ordinals and abbreviations which have been marked during text normalisation (see 12.3.3) as requiring an appropriate inflection ending. The part-of-speech information added by the tagger tells whether the token is an adverb or an adjective. In addition, information about the boundaries of noun phrases has been provided by the chunker, which is relevant for adjectives.

In the lexicon, all entries occurring in noun phrases (determiners, adjectives, and nouns) are annotated with their possible value combinations for the morphological inflection information gender, number and case. In addition, determiners are marked as definite or indefinite. This information was obtained from the morphological analyser MMORPH (Petitpierre & Russell, 1995).

When the inflection endings module finds an ordinal or an abbreviation with an adjectival role, it performs a unification of the morphological variables over the known tokens in the noun phrase to which the ordinal or abbreviation belongs. In many cases, this allows the appropriate values of gender, number and case to be determined for the ordinal or abbreviation, so that the correct ending can be selected and added to the expanded form.

For example, in “mein 2. Angebot” (Engl. “my second offer”), the words “mein” and “Angebot” are looked up in the lexicon, their associated values for gender, number and case are compared, and only the common ones (gender=neutral, number=singular, case=nom./acc.) are retained. Further disambiguation is not necessary, as all remaining possibilities (neutral/singular/nom. and neutral/singular/acc.) correspond to the same adjective ending (“-s” with indefinite determiner “mein”), so the correct adjective ending can be added to the ordinal: “zweites”.

Lexicon

The pronunciation lexicon is derived from CELEX (Baayen et al., 1995). It contains the graphemic form, a phonemic transcription, a special marking for adjectives, and the inflection information mentioned above (see 12.3.5).

As the inflection of adjectives is quite regular in German, only the stem form of an adjective is contained in the lexicon, while all inflected forms are generated by the lexicon lookup program.

The lexicon performs a simple compound treatment. If a word is not found in the lexicon but is the concatenation of two or more lexicon entries, the corresponding phonemic forms are concatenated. Optional bounding morphs (*Fugen* or infixes) such as "+s+", "+es+", "+n+", "+en+", and "+e+", typical for German noun compounds, are also allowed. For all parts of a compound except the first, primary word stress is reduced to secondary stress, i.e. the first part is considered the dominant one, which seems to be the default for German. Exceptions to this rule, such as "Bundesinnenminister", "Oberverwaltungsgericht", are part of the lexicon.[8]

Letter-to-sound conversion

Unknown words that cannot be phonemised with the help of the lexicon are analysed by a "letter-to-sound conversion" algorithm. This algorithm is more complex than a simple application of letter-to-sound rules: on the one hand, correct phonemisation relies in many cases on a correct identification of morpheme boundaries. On the other hand, for the phoneme string to be properly uttered, syllabification and word stress information needs to be added.

First, a morphological decomposition is attempted using a statistical morpheme "parser" based on the probability of two adjacent morphemes being neighbours. This had been trained on data extracted from CELEX (Baayen et al., 1995). The resulting morpheme chain is compared to a list of affixes which have a predictable effect on word stress position, either attracting or shifting the stress, or with no effect on stress.

The remaining morphemes are subjected to a set of generic letter-to-sound rules for German.

[8] A more elaborate approach to productive compounding in German, including morphological decomposition, using weighted affix and stem lexicons, can be found e.g. in Möbius (1999).

The syllabification of the transcribed morphemes is based on standard phonological principles such as the sonority hierarchy of phonemes, the maximum onset principle, the obligatory coda principle and the phonotactic restrictions for the German language (see also Brinckmann & Trouvain, 2003).

Finally, a word stress assignment algorithm decides which syllable receives the primary lexical stress. No rule-based secondary stress assignment is attempted at present.

12.3.6 Prosody rules

Prosody is modelled using GToBI (Grice et al., 2002), an adaptation of ToBI ("Tones and Break Indices") for German. ToBI (Silverman et al., 1992) describes intonation in terms of fundamental frequency (F0) target points, distinguishing between accents associated with prominent words and boundary tones associated with the end of a phrase. The size of a phrase break is encoded in break indices. Within MARY, break indices are used as follows: "2" is a potential boundary location (which might be "stepped up" and thus realised by some phonological process later on); "3" denotes an intermediate phrase break; "4" is used for intra-sentential phrase breaks; "5" and "6" (not part of GToBI) represent sentence-final and paragraph-final boundaries.

The prosody rules module assigns the symbolic GToBI labels. In a later step (see 12.3.8), these are translated into concrete F0 targets and pause durations and are taken into account for accentual lengthening and phrase-final lengthening in the duration module.

The prosody rules were derived through corpus analysis and are mostly based on part-of-speech and punctuation information. Prosodic boundaries are inserted at punctuation signs, conjunctions which are not part of co-ordinated noun phrases, and after the *Vorfeld* in statements, i.e. just before the first finite verb in a statement. The syntactic information which comes as an output of the chunk parser is used as an additional source for assigning prosodic boundaries, e.g. for special speaking styles. In Trouvain (2002) it has been shown that for slow speech, syntactic phrasing information is very useful to determine appropriate locations where to insert additional pauses.

Some parts-of-speech, such as nouns and adjectives, always receive an accent; the other parts-of-speech are ranked hierarchically (roughly: full verbs > modal verbs > adverbs), according to their propensity for receiving an accent. This ranking comes into play where the obligatory assignment rules do not place any accent inside some intermediate

phrase. According to a GToBI principle, each intermediate phrase should contain at least one pitch accent (Benzmüller & Grice, 1997). In such a case, the token in that intermediate phrase with the highest-ranking part-of-speech receives a pitch accent.

After determining the location of prosodic boundaries and pitch accents, the actual tones are assigned according to sentence type (declarative, interrogative-W, interrogative-Yes-No and exclamative). For each sentence type, pitch accent tones, intermediate phrase boundary tones and intonation phrase boundary tones are assigned. The last accent and intonation phrase tone in a sentence is usually different from the rest, in order to account for sentence-final intonation patterns.

12.3.7 Postlexical phonological processes

Once the words are transcribed in a standard phonemic string including syllable boundaries and lexical stress on the one hand, and the prosody labels for pitch accents and prosodic phrase boundaries are assigned on the other hand, the resulting phonological representation can be re-structured by a number of phonological rules. These rules operate on the basis of phonological context information such as pitch accent, word stress, the phrasal domain or, optionally, requested articulation precision. Currently, only segment-based rules apply, such as the elision of schwa in the endings “–en” and “–em”, the backward assimilation of articulation place for nasal consonants, and the insertion of glottal stops before vowels of pitch-accented syllables with a free onset. For the future it is planned to take into account some re-structuring on the prosodic level, e.g. reducing the number of pitch accents and phrase boundaries for fast speech (Trouvain & Grice, 1999).

The output of this module gives the maximally rich MaryXML structure, containing all the information added to the structure by all of the preceding modules.

12.3.8 Calculation of acoustic parameters

This module performs the translation from the symbolic to the parametrical domain. The MaryXML structure is interpreted by duration rules and GToBI realisation rules.

The duration rules are at present a version of the Klatt rules (Klatt, 1979; Allen et al., 1987) adapted to German (Brinckmann & Trouvain, 2003). They have been shown to yield perceptual results only sightly inferior to a classification and regression tree (CART)

trained on a corpus of German read speech (Brinckmann & Trouvain, 2003), while having the advantage of being readily interpretable e.g. by students of speech science.

The realisation of GToBI tones uses a set of target points for each tone symbol. These targets are positioned, on the time axis, relative to the nucleus of the syllable they are attached to; on the frequency axis, they are positioned relative to a descending pair of topline and baseline representing the highest and lowest possible frequency at a given moment. The fact that these lines are descending accounts for declination effects, i.e. overall F0 level is higher at the beginning of a phrase than close to the end. As an example, the GToBI accent “L+H*”, associated with the syllable [fʊn] of the sequence [gə-'fʊn-dən] (Engl. “found”) is realised as follows:

- first, the “L+” part is realised by a target on the baseline at the start of the nucleus of the preceding syllable (the schwa of [gə]);
- second, the “H*” part is realised by a target on the topline in the middle of the nucleus of the accented syllable (the [ʊ] in ['fʊn]).

As is illustrated in Figure 12.2, this allows concrete frequency target values to be calculated if the segment durations and the frequency values for the start and end points of the topline and baseline are known. Obviously, the latter values need to be set appropriately for the voice to be used during synthesis, in particular according to the sex of the speaker.

The output produced by this module is no longer a MaryXML structure, but a list containing the individual segments with their durations as well as F0 targets. This format is compatible with the MBROLA `.pho` input files.

12.3.9 Synthesis

At present, the MBROLA diphone synthesiser (Dutoit et al., 1996) is used for synthesising the utterance based on the output of the preceding module. MBROLA was selected because of the comparatively low degree of distortions introduced into the signal during signal processing. All available German MBROLA voices can be used. Due to the modular architecture of the MARY system, any synthesis module with a similar interface could easily be employed instead or in addition.

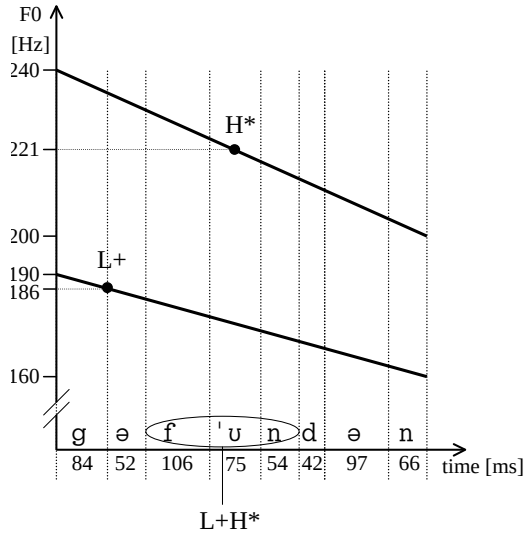


Figure 12.2: Illustration of the calculation of frequency parameters for target points realising the GToBI accent L+H*.

12.4 An interface for expert users

An interface has been designed which allows the user to easily investigate parts of the MARY architecture tree (see Figure 12.1). Besides plain text and SABLE- or SSML-annotated text, each intermediate processing result can serve as input, and any subsequent processing result can be output.

In particular, it is possible to only investigate the translation of SABLE into MaryXML, i.e. the interpretation of high-level markup in terms of low-level markup.

Individual processing steps can be carried out, allowing the user to understand the function of each module, or to investigate the source of an error. In addition, the intermediate results can be modified by hand, experimenting which input to a given module yields which output.

Figure 12.3 shows an example of such partial processing. The input text pane on the left side contains a partially processed version of the utterance “Ich fliege nach Schottland.” (lit. “I fly to Scotland.”), more precisely the output of the tagger/chunker module (corresponding to the *data type* “MaryXML tagged”). As a well-formed and valid XML

document, it contains some header information (not shown in Figure 12.3), followed by the document body enclosed in `<maryxml>...</maryxml>` tags. In this example, the document consists of a single sentence (`<div>...</div>`) containing five tokens (four words and one punctuation mark). The tokens have already been enriched with some part-of-speech and syntactic information encoded as attribute/value pairs of the respective `<t>` tags. A “Verify” button allows the user to perform a validating XML parse of the input, making sure that the input is well-formed and valid (i.e., conforms to the MaryXML DTD (Harold, 1999)).

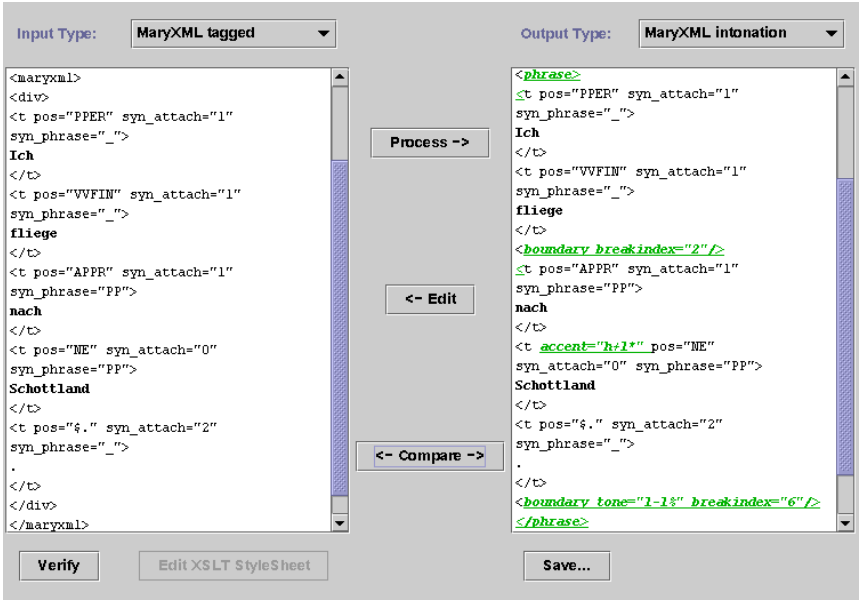


Figure 12.3: Example of partial processing with the MARY interface. See text (12.4) for explanations.

Output of a given type can be obtained by simply selecting the desired output format (in this case, the output of the prosody module, “MaryXML Intonation”) and pressing the “Process” button. If both input and output are MaryXML, the “Compare” button allows the differences between the two versions of the document to be highlighted, which correspond to the information added by the selected processing steps.

If the output obtained in this step is to be used as input for subsequent processing steps, it can be transferred into the input text pane using the “Edit” button.

12.5 Suitability for emotion expression

The modular architecture of the MARY system, based on an internal XML representation, is useful beyond the expert user interface presented above. An example for the benefits of the system structure is the ease with which prosodic parameters can be modified, e.g. for the synthesis of emotional speech.

12.5.1 Accessibility of prosodic parameters

As was noted in the chapter discussing the speech parameters used for expressing emotions in the voice (Chapter 5), the prosodic parameters used for emotion expression are basically the same ones as those used in the linguistic signalling system. Therefore, many of the parameters required for an acoustic modelling of emotions are already present in a speech synthesis system and can be modelled for the purpose of emotion expression provided that an appropriate access is guaranteed.

This is a problem in commercial systems that can only be treated like a “black box”. Existing speech synthesis markup formats such as the Microsoft Speech API markup, SABLE, or the new W3C SSML only provide a very coarse-grained, imprecise specification of prosody which is intended for use by non-experts. As they are intended to be independent of an actual speech synthesis system, they do not provide the possibility to model prosody explicitly by means of a concrete model such as ToBI.

In contrast, the MaryXML format used in the MARY system is an attempt to make all intermediate data accessible and modifiable. This can be done either manually by modifying intermediate processing results, or automatically by inserting additional modules. All concepts used to calculate prosody at various stages of processing can thus be accessed, taken into account and modified if desired. This includes high-level concepts such as accentuation status of words, GToBI accent and boundary labels, or syllable stress, as well as low-level acoustic parameters such as segment durations and F0-time target points containing the concrete fundamental frequency specified in Hertz. In addition, it is possible to adjust the topline and baseline used in the GToBI model (see Figure 12.2) as well as the

speech rate not only for entire phrases, but also recursively for groups of words or individual words within a phrase. The parameters which are used for the synthesis of emotional speech are described more thoroughly in Chapter 13.

12.5.2 Voices for emotion expression

As was discussed briefly in 9.2, p. 88, a fundamental difficulty in using diphone synthesis for emotion expression is the role of voice quality. While it is well-known that it plays an important role in emotion expression, voice quality is usually impossible to model in diphone synthesis, as the voice quality with which the diphone set was recorded cannot be modified. Opinions differ on whether voice quality is absolutely necessary or whether at least some emotions can be convincingly expressed without modelling voice quality (see the discussion in 9.2.2, p. 89). It seems to be undisputed, though, that voice quality would be a desirable feature to model.

This limitation of concatenative synthesis in general and of diphone synthesis in particular was addressed in the recordings of a male and a female diphone voice in the NECA project (IST-2000-28580). Against the background of the findings reported in Chapters 10 and 11, we decided that vocal equivalents of the activation dimension were most promising to model. As was described in detail by Scherer (1986), activation affects phonation, among other things via the tension in the laryngeal muscles. The resulting change in vocal effort is reliably perceived by listeners (Eriksson & Traunmüller, 1999; Lehiste & Peterson, 1959). We therefore decided to record the full diphone set in each of three levels of vocal effort: low, medium and high.

The recordings were carried out using an adapted version of the Festvox toolkit[9]. A list of nonsense words was compiled which contains all German diphones. In a sound-treated room, each nonsense word was shown to the speaker on a silent TFT display and played as a prompt produced on a monotonous pitch by an existing synthesis voice. The prompts were delivered via loudspeakers, using a playback volume which was adjusted according to the targeted vocal effort, i.e. for high vocal effort recordings, the volume was increased, and for low vocal effort recordings, it was reduced. After hearing the prompt, the speaker produced the word, which was recorded directly into a laptop computer. An AKG D330 BT microphone was used. As the signal-to-noise ratio of the laptop computer’s line-in

[9] `http://www.festvox.org`

port was much better than that of its microphone port, a minidisc recorder was used as a pre-amplifier, amplifying the microphone signal to the level required by the line-in port and at the same time recording all material including the informal speech during recording pauses. As the number of aspects to be controlled in these recordings was particularly high (constant level of vocal effort, monotonous pitch, phonetic quality, position relative to the microphone), a total number of four phonetic experts were present, verifying the recordings. Directly after each word was recorded, it was played to the experts via headphones.[10] If any of the experts had any reservations, the word was re-recorded immediately.

The phonetic labelling of the recordings was first carried out automatically, using the Festvox tools, and then corrected manually. The diphones were sent to the MBROLA team for coding into the MBROLA format (Dutoit et al., 1996). The resulting voice databases (one male, one female) contain a full diphone set for each of the three levels of vocal effort: low, medium, and high.

A perceptual evaluation was carried out (Schröder & Grice, 2003) for the male voice database, addressing two hypotheses: (I) The three diphone sets are perceived as belonging to the same speaker, and (II) the vocal effort is perceived from synthesised material as intended in the recordings. Both hypotheses could be verified.

In addition to the diphone material, the voice databases contain a number of affect bursts (emotional interjections, see 5.3, p. 55 and Schröder (2003)). These were recorded separately, as it was expected that this would lead to a more natural segmental structure than what would be obtained by synthesising them from diphones. In particular for affect bursts containing non-phonemic sounds, such as a rapid intake of breath, it would be totally impossible to reconstruct such affect bursts from diphones.

At first, we had intended to combine the unmodified recordings of affect bursts with the synthetic speech. However, informal listening tests invalidated this approach, because the voice change due to the MBROLA signal manipulation made the synthetic voice sound too different from the original voice for the combination of synthesised and unsynthesised material to sound acceptable. Therefore, the affect bursts were also MBROLA-encoded and resynthesised with their original duration and F0 contour. This allowed the combination of the affect bursts with the diphone synthesis.

[10] In order not to confuse the speaker, the signal was not played via the loudspeakers. Technically, this was achieved by the use of a specially-made cable sending the left stereo channel to the loudspeakers and the right stereo channel to the headphones.

The two voice databases resulting from this work represent a new step towards ecologically valid expression of emotion in diphone synthesis. They provide the user with the possibility to change the voice quality as far as it is related to vocal effort. This opens new possibilities in the synthesis of expressive speech including the voice quality channel, which is known to be important for emotion expression, but which in the past had been impossible to model in diphone speech. The fact that the recorded voice qualities are not specific to individual emotion categories, but generally usable, is in line with the approach to emotional speech synthesis formulated in 8.2, p. 82, namely a flexible, generally usable synthesis for which it is acceptable if it does not fully specify the exact emotional state expressed, but only its basic properties. In this context, the voices must be considered an important building block of a convincing emotional speech synthesis system modelling emotions by means of emotion dimensions.

12.6 Summary

An overview of the processing components of the German text-to-speech system MARY has been given. It has been described how MaryXML, a system-internal XML-based data representation, can be used to make partial processing results available outside the system. The advantages of MaryXML are three-fold:

1. *All* intermediate processing results can be made visible;
2. these intermediate results can be modified and fed back as input into the system;
3. via the WWW, the interface is accessible from everywhere without a local installation of the system.

These features are very helpful for teaching purposes and for non-technical users.

The relevance of these system features for the implementation of emotional prosody rules in the system was pointed out, and two new diphone voices were described which provide full diphone sets expressing low, medium and high vocal effort, as well as affect bursts, which makes them optimally suitable for the synthesis of emotional speech using diphone material.

Chapter 13

Prosody rules for emotional speech synthesis

In this chapter, the results from the literature reviews in Chapters 9 and 10 and from the data analysis in Chapter 11 are re-formulated in a way that allows them to be implemented in the MARY text-to-speech system. The approach chosen for the technical realisation of this implementation is described, and a graphical user interface is presented which allows the user to interactively explore the expressive capabilities provided by the emotional speech synthesis system.

13.1 Generative formulation of prosody rules

Table 13.1 presents the essential data required to express emotions in a speech synthesis system using emotion dimensions. The columns represent the emotion dimensions, while the rows list all the acoustic parameters for which emotion effects are modelled.

The numeric data fields represent the linear coefficients quantifying the effect of the given emotion dimension on the acoustic parameter, i.e. the change from the neutral default value. As an example, the value 0.5% linking Activation to `rate` means that for an activation level of +50, rate increases by +25%, while for an activation level of -30, rate decreases by -15%.

The parameters represented in the rows of Table 13.1 were selected because they can be manipulated in the MARY system. Several types of parameters can be distinguished:

	Prosodic parameter	Coefficients		
		Activation	**Evaluation**	**Power**
fundamental frequency	pitch	0.3[a]	0.1[a]	-0.1[a]
	pitch-dynamics	0.3%		-0.3%[b]
	range	0.4[a]		
	range-dynamics	1.2%		0.4%
	accent-prominence	0.5%[a]	-0.5%[a]	
	preferred-accent-shape		E≤-20: falling -20<E≤40: rising E>40: alternating[c]	
	accent-slope	1%[a]	-0.5%[a]	
	preferred-boundary-type			P≤0: high P>0: low[b,c]
tempo	rate	0.5%[b]	0.2%[b]	
	number-of-pauses	0.7%[b]		
	pause-duration	-0.2%[b]		
	vowel-duration		0.3%[b]	0.3%[b]
	nasal-duration		0.3%[b]	0.3%[b]
	liquid-duration		0.3%[b]	0.3%[b]
	plosive-duration	0.5%	-0.3%	
	fricative-duration	0.5%	-0.3%	
	volume	0.33%[b]		

Table 13.1: Emotion dimension prosody rules as implemented in the MARY system. Origin of the parameters: a: corpus analysis in Chapter 11; b: literature on acoustic correlates of emotion dimensions, reviewed in Chapter 10; c: literature on emotional speech synthesis, reviewed in Chapter 9 and summarised in Appendix A; the remaining parameters were chosen by the author during interactive exploration. All parameters were manually fine-tuned to adapt them to the MARY system using the male NECA voice.

- "standard" global parameters: `pitch` and `range`, the mean value of the GToBI *baseline* and the mean distance between GToBI *topline* and *baseline*, respectively (see 12.3.8, p. 148 and Figure 12.2, p. 150); `rate`, the speech rate, specifying the articulation rate as well as the number and duration of pauses (higher rate = fewer and shorter pauses); and `volume`, a combined parameter affecting the amplitude of the speech signal and the selection of a diphone set with appropriate vocal effort-related voice quality (see 12.5.2, p. 153).

- non-standard global parameters: `pitch-dynamics` and `range-dynamics` describe the slope of the GToBI *baseline* and *topline*, quantified as the difference between the respective Hertz values at the start and the end of an intonation phrase. Their default values (not shown in Table 13.1) are negative, reflecting the declination effect usually encountered in neutral speech;

- rules applicable to specific entities:

  – GToBI accents: `accent-prominence` is aimed at the perceptual salience of accented syllables, through a combination of F0 target over-/undershoot, segment duration, loudness, and vocal effort-related voice quality; `preferred-accent-shape` is a categorical setting indicating the rough type of accent tones to be realised on accented syllables; and `accent-slope` influences the steepness of F0 rises or falls for bitonal pitch accents, by determining the placement on the time-axis of the "+" target relative to the "*" target (see Figure 12.2, p. 150).

  – GToBI boundaries: `preferred-boundary-type` is a categorical parameter expressing a preference as to the type of boundary tones to be assigned; `number-of-pauses` influences the number of pauses to realise, and thus opens the possibility to deviate from the default behaviour linked to speech rate; and similarly, `pause-duration` provides the possibility to modify the pause duration relative to the default as determined by the speech rate.

  – Individual phoneme classes: five groups of phonemes are distinguished in order to provide the possibility to selectively increase or decrease their duration. In particular, `vowel-duration` allows for the implementation of Tischer's

(1993) finding of a link between vowel duration and the evaluation and power dimensions.

As in other studies (e.g., Cahn, 1990; Murray & Arnott, 1995), the parameter values listed in Table 13.1 are not directly and uniquely taken from the literature or from experimentation; instead, these sources are usually taken as a starting point, and an iterative process of listening and manual fine-tuning is required to adapt the rules to the system. In the present case, when there was clear evidence from the corpus analysis in Chapter 11, the corresponding parameters were used. As is stated earlier, this was made possible by the quantified description using a linear regression model. The indications from the literature survey in Chapter 10 were used when the corpus analysis did not produce conclusive evidence (as in the case of speech rate and intensity), as well as for parameters which can be modelled in speech synthesis but which could not be studied in the corpus analysis (e.g., vowel duration). Inspiration for categorical changes in accent and boundary type were taken from Chapter 9 and the associated Appendix A. Table 13.1 documents the origin of each parameter value.

At the time of writing, only the male voice created in the NECA project is available, while recordings for the female voice are still ongoing. Other publicly available voices do not provide the possibility to model a changing voice quality, a feature considered important for the implementation of convincing emotion rules. For this reason, the values in Table 13.1 are tuned to the male NECA voice. It is currently unclear whether the rules for a female voice should differ from this rule set adapted for a male voice. In particular, it is unclear whether the different set of parameters correlating with the evaluation and power dimensions in the corpus analysis of Chapter 11 should also be used in a speech synthesis system. In any case, the adaptations required to use the rules with a female voice should be small, given the fact that all quantified effects are expressed as relative changes from a (voice-specific) default value. The rules formulated in Table 13.1 can certainly be used as a starting point for exploring this question in future work.

As high a number of prosodic parameters as possible was used, in order to exploit all the means of expressivity provided by the system. It is to be noted, however, that currently each individual rule is very simple. In accordance with the minimalistic model used in the corpus analysis described in Chapter 11, a linear correlation between the emotion dimensions and the acoustic parameters was assumed. This is likely to be an oversimplification. In particular, it does not allow for a distinction between the effect of raising, e.g., activation above the neutral state and of lowering it below the neutral state. This forces the model to predict, e.g., F0 reductions for passive states which are as large as the F0 increases predicted for active states, which can lead to unreasonably low F0 values for very passive states. Similarly, it may be that the effects are of different strengths for near-neutral and for extreme emotional states, or there may be effects occurring only in certain regions of the emotional space, which would correspond to interactions between the emotion dimensions.

In this simple first model, all these complex phenomena were ignored. It should therefore be regarded as a first proof-of-concept rather than a fully optimised model, which would be beyond the scope of this dissertation.

13.2 Implementation: Technical realisation

A major design feature in the technical realisation of the emotional speech synthesis system was that the acoustic effects of emotions should be specified in one single module. This module adds appropriate MaryXML annotations to the text which are then realised by the respective modules within the MARY system. As a consequence, all of the parameters are global in the sense that they will be applied to all enclosed text. This approach is considered the most transparent, as the link between emotions and their acoustic realisations is not hidden in various processing components, and the easiest to maintain and adapt, as all rules are contained in one document.

For maximum conformity with the existing MARY system architecture (see Figure 12.1, p. 141), the emotion realisation module was realised at the same level as the "Sable Parser" and "SSML Parser" modules. As a proof of concept, a simple emotion dimension markup language (EDML) was created, annotating text using a single `<emotion>` tag in which the positions on emotion dimensions are specified as the values of the `activation`, `evaluation` and `power` attributes. An example EDML document is the following:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<emotion activation="+30" evaluation="+70" power="+20">
Wie wunderbar!
</emotion>
```

In the same way as for the Sable and SSML parsers, which use an XSLT stylesheet for transforming an input markup language into the MARY-internal MaryXML, the emotion realisation module uses an XSLT stylesheet (see Appendix B) implementing the rules listed in Table 13.1. Applying that XML transformation to the above example creates the following MaryXML document:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         version="0.3" xml:lang="de">
<voice name="de6">
<prosody accent-prominence="-20%" accent-slope="-5%"
   fricative-duration="-6%" liquid-duration="+27%"
   nasal-duration="+27%" number-of-pauses="+21%"
   pause-duration="-6%" pitch="124" pitch-dynamics="-12%"
   plosive-duration="-6%" preferred-accent-shape="alternating"
   preferred-boundary-type="low" range="37" range-dynamics="+4%"
   rate="+29%" volume="60" vowel-duration="+27%">
Wie wunderbar!
</prosody>
</voice>
</maryxml>
```

Note that the voice is set to `de6`, the male NECA voice, and that the attributes of the prosody tag, though listed here in alphabetical order rather than grouped by their meaning, correspond exactly to the rows of Table 13.1. As motivated above, the MaryXML document retains no information about the emotional state, but all information required for the desired acoustic realisation of that emotional state. Like any other MaryXML document, the above example can be processed by the standard MARY modules.

13.3 EmoSpeak: A graphical interface to emotional speech synthesis

A graphical user interface was programmed to allow for an interactive exploration of the emotional module described above.

The interface, shown in Figure 13.1, allows the user to type in any text, and to specify the emotional state with which the text is to be spoken. The position on the activation and evaluation dimensions is specified simultaneously, by locating a green cursor in a two-dimensional space modelled after the Feeltrace circle (see 6.4.2, p. 70). The third dimension, power, is set independently. Using these positions on the three emotion dimensions as well as the text to be spoken, an EDML emotion markup document is created. The emotion-to-maryxml transformation stylesheet, described above, transforms this into a MaryXML document, which is displayed for the interested user in a second window under the EmoSpeak interface (not shown). Simultaneously, the MaryXML document is sent to the MARY server, which synthesises a corresponding audio file and sends it back to the interface. All of these transformations are carried out continuously as the user modifies the emotional state or the text. By clicking on the “Play” button, the user can hear the result.

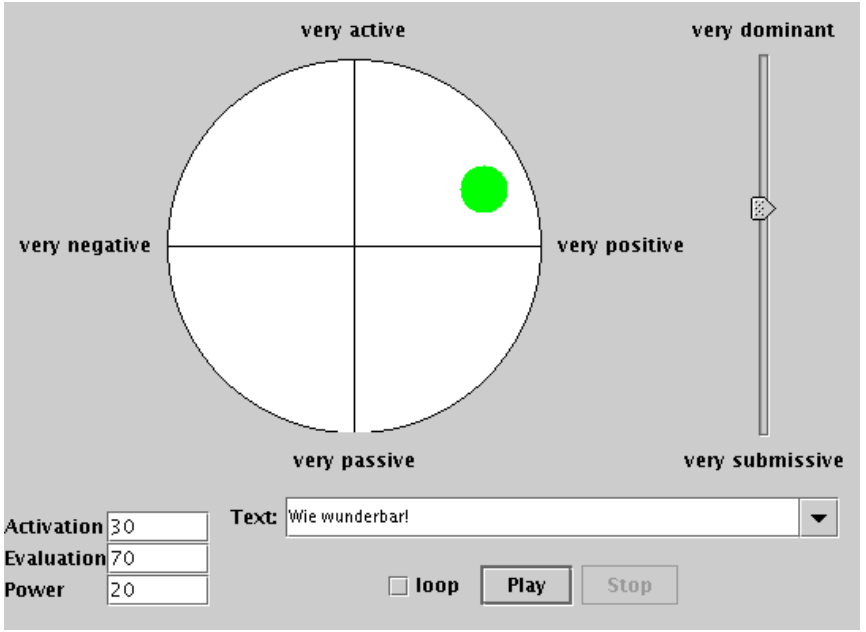


Figure 13.1: The EmoSpeak interface to emotional speech synthesis

13.4 Summary

The results of the literature surveys and the experimental work, which were presented in the preceding chapters, were put to use in this chapter to create an emotional speech

synthesis system. A maximally modular and transparent approach was selected, relying on an early specification of all the acoustic correlates of the emotional state in one single step. This was made possible by the special architecture of the MARY system which allows for a "remote-control" type of specification of module behaviour through the representation language MaryXML.

An XSLT stylesheet was programmed, implementing the transformation from a simple emotion markup language into MaryXML. This approach makes the transformation very flexible to use in many different application settings. In addition, it is simple to adapt the stylesheet to "real-world" input markup languages in which emotion is only one out of many annotations.

A concrete set of rules for mapping emotion dimensions onto acoustic parameters, summarised in Table 13.1, was proposed. This rule set, based on the literature and on experimental corpus analysis, was manually fine-tuned by the author.

A graphical user interface was created for interactively exploring the effect which changing the emotional state has on the synthetic voice.

Chapter 14

Perceptual evaluation

The appropriateness of the generated emotional prosody and voice quality is assessed in a perception test. As has been argued before (see 8.2, p. 82), this appropriateness is to be thought of in terms of coherence with other channels expressing the emotion, such as verbal content, possibly visual channels, and the situational context. For applications, it is necessary for the speech prosody to "fit with" the rest of the message, in order not to put off the users. This question seems more important than the question what emotional information can be conveyed by the speech prosody alone.

Consequently, any perception test methodology aiming at the assessment of the perceived emotional state is suboptimal for the current task. This includes the identification task methodology widely used for the evaluation of synthesised emotional speech (see 6.1, p. 63), as well as methods for describing the perceived emotional state in terms of emotion dimensions, such as semantic differential ratings (see 6.4.1, p. 69) or the Feeltrace tool (see 6.4.2, p. 70). Instead, a preference task methodology (see 6.2, p. 67), using coherent and contradictory multi-channel emotional messages as stimuli, is explored as a promising alternative.

14.1 Overview and design

The evaluation proceeds as follows. A number of emotional states are defined using non-vocal channels of emotion expression. A good coverage of the emotion space should be aimed for. A configuration of prosodic settings is predicted for each of the targeted

emotional states. The prosodic and the non-vocal configurations are cross-combined in a factorial setting, and preference ratings are obtained for the resulting stimuli.

In order to limit the number of variables to be controlled, the "other channels" of emotion expression serving as an emotion reference were kept to a minimum by eliminating the visual channel. The emotion reference was designed to consist of a written situation description along with the verbal content of the utterance to be spoken.

Furthermore, in order to allow for a good coverage of the emotion space including both intense and moderate emotional states while keeping the total number of states manageable, a two-dimensional space was used, leaving out the power dimension for which the least conclusive results were found in Chapter 11.

A first necessary step is the establishment of the emotion conveyed by the reference channel, i.e. a written situation description along with the text of an utterance produced in that situation. A first, written test was carried out in order to identify suitable situation descriptions and to establish their perceived emotional content. For a number of candidate situation descriptions, subjects indicated the emotion conveyed, on an evaluation and an activation scale. Situation descriptions were selected based on low ambiguity as to the emotion conveyed (i.e., high agreement among raters) and wide distribution across the activation-evaluation space.

The selected situation descriptions were then used in the actual evaluation test. For each of the emotional states as defined by the situation descriptions, the corresponding prosodic parameter settings were predicted by the emotional prosody rules (see Table 13.1, p. 158). Each text was then combined with each prosodic setting, and the resulting stimuli were presented along with the written situation description. Subjects indicated on a scale how well the "tone of voice" fitted with the situation.

The following result patterns can be expected. If the emotional prosody rules were optimal, prosodic settings corresponding to the emotion in the text would be preferred over the other prosodic settings. This would correspond to a clear preference for "matching" emotion messages in prosody and text. However, it is well-known that even human vocal emotions are only recognised to a certain extent (e.g., Banse & Scherer, 1996), so that confusions are bound to occur. What patterns can be expected for these confusions? Previous evidence (see Chapters 10 and 11) indicated that prosody and voice quality best convey the activation level of the emotion, and that it is unclear whether or to what extent

the evaluation level is conveyed. Therefore, prosodic settings corresponding to similar activation levels can be expected to lead to similar preference ratings, while a difference in activation level should also lead to a difference in preference. Due to the continuous nature of the emotion dimensions, this effect is expected to be gradual, i.e. the larger the difference in activation level, the larger the difference in preference. It is to be seen whether a similar pattern can also be found for differences in evaluation.

14.2 Written texts as emotion references

The Belfast Naturalistic Emotion Database (see 11.2, p. 104) was used as a source of emotional situation descriptions. Based on the Feeltrace ratings of the audio-visual clips, textual transcriptions of the clips were grouped into nine categories covering the activation-evaluation space: the neutral centre, and each of the four quadrants in a moderate and an intense version (moderate active negative, intense active negative, moderate active positive, etc.). From each group, four transcriptions were selected based on which German situation descriptions were formulated. The resulting 36 situation descriptions are listed in Appendix C.

As the original ratings were not based on the text alone, but on audio-visual clips, and also because of the translation and re-formulation, it was necessary to assess the emotion conveyed by the written situation descriptions. To that end, the texts were presented on paper in pseudo-randomised order to 15 subjects (mostly students, 4 male, 11 female). The subjects indicated for each situation the speaker's emotional state on the two scales "positiv-negativ" (positive-negative, the evaluation dimension) and "aktiv-passiv" (active-passive, the activation dimension). In view of the later use with the emotional prosody rules assuming continuous dimensions, quasi-continuous scales were used on which values ranged from -10 to +10. Half the subjects rated the texts in reverse order.

As in the analysis of the Feeltrace ratings of the Belfast database, consistency between subjects was verified by means of pairwise correlations between all subjects' activation and evaluation ratings (see 11.2.1, p. 104 and Table 11.1, p. 107). The mean correlation of a subject's ratings with all other subjects' ratings was calculated as a simple measure of subject consistency. The overall mean evaluation correlation among subject ratings was 0.90, with no marked differences between subjects. The overall mean activation corre-

	Name	Act.	Eval.	s.d. Act.	s.d. Eval.
A	neutral1	-5.9	-0.4	3.0	3.3
B	modactneg2	5.1	-2.5	2.5	4.4
C	modactneg1	1.3	-2.9	3.8	2.7
D	modactpos2	0.1	8.2	3.9	1.4
E	intactpos3	8.1	8.6	2.4	1.3
F	modpasneg3	-4.3	-6.2	2.3	1.5
G	intpasneg1	-7.7	-8.7	3.7	1.8
H	modpaspos4	-5.5	6.6	3.4	2.8

Table 14.1: Co-ordinates of selected situation descriptions in activation-evaluation space. Identifiers A–H match those in Figure 14.1.

lation among subject ratings was considerably lower, at 0.34. Two subjects (one male, one female) showed particularly low mean correlations, at 0.20 and 0.21, respectively. These subjects' ratings were completely removed from the analysis. The mean correlation between the remaining 13 subjects' activation ratings was 0.39.

Based on the 13 "reliable" subjects, mean values and standard deviations were calculated for activation and evaluation for each situation description. The lower correlations between activation ratings manifested in greater standard deviations compared to evaluation ratings. Appendix C reproduces these values for each of the situation descriptions.

"Good" examples of situation descriptions, showing low ambiguity, i.e. a low standard deviation for both activation and evaluation ratings, were then selected. An arbitrary maximum value of 4.0 for activation and evaluation standard deviations was set as a minimal requirement for "good enough" situation descriptions. Eight situation descriptions fulfilled this requirement.[1] Their co-ordinates and standard deviations in activation-evaluation space are shown in Table 14.1 and in Figure 14.1.

[1] Only text B/modactneg2 shows an evaluation standard deviation slightly above the arbitrary threshold. The text was nevertheless selected because of the high agreement in activation and the improved coverage of activation-evaluation space it provided.

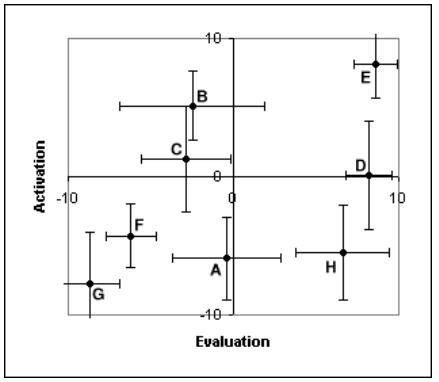


Figure 14.1: Co-ordinates of selected situation descriptions in activation-evaluation space. Error bars show one standard deviation in each direction. Identifiers A–H match those in Table 14.1.

14.3 Listening test

14.3.1 Method

Stimulus creation

64 audio stimuli were created as follows. The eight acoustic parameter configurations corresponding to the mean activation and evaluation co-ordinates of the eight situation descriptions, as shown in Table 14.1, were predicted using the emotional prosody rule system described in Chapter 13. The last sentence of each situation description, corresponding to the last part of the speaker's utterance in each situation, was synthesised in eight versions, i.e. with each of the eight acoustic parameter configurations.

Test setup selection

A pre-test was carried out in order to select a suitable test setup. Two versions of the test were prepared and presented to four native German subjects, each of whom performed both tests on a single text and the eight associated audio stimuli and then commented on the two versions of the test.

The first version of the test was a "classical" scale-based rating test. The eight synthesised versions of the text were presented individually, in random order. Subjects could

listen to a stimulus as many times as they wanted to, and then placed it on the scale “passt optimal” – “passt überhaupt nicht” (fits optimally [with the text] – does not fit at all) before they moved on to the next stimulus.

The second version of the test presented all eight synthesised versions of the text simultaneously, randomly assigned to icons labelled from 1 to 8. Subjects could to listen to all stimuli as many times as they wanted to, and placed them on the same scale as in the first version of the test. The main difference was that they were able to compare the stimuli, and make corrections to their first placement of one stimulus after hearing other stimuli. Figure 14.2 shows a screenshot of the corresponding graphical user interface.

All four subjects who took part in the comparatory pre-test indicated a strong preference for the second version of the test, as it allowed them to “make finer distinctions and give more precise answers”. Therefore, the second version of the test was used.

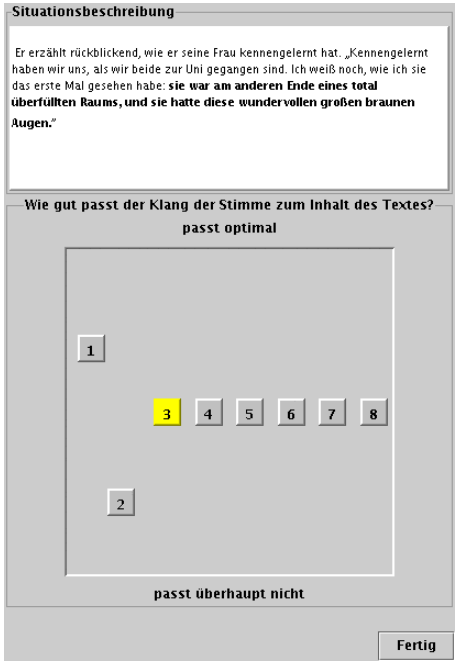


Figure 14.2: A screenshot of the graphical user interface used in the evaluation test. The part of the text in bold is provided in eight synthesised versions. The user is currently listening to version number 3.

Test

20 native German subjects (11 male, 9 female; mean age 30.4 years) who had not taken part in the written test or in the pre-test rated the 64 stimuli using the graphical user interface selected in the pre-test (see Figure 14.2). Four of the subjects (2 male, 2 female) were speech synthesis experts, while the other 16 subjects had little to no experience with speech synthesis.

Each subjects took the test individually, listening to the stimuli via headphones. Both the order of the eight texts and the assignment of the eight audio versions per text to the graphical icons labelled 1 to 8 were automatically randomised for each subject.

Most subjects remarked that some situation descriptions were ambiguous as to the activation of the speaker intended in the textual situation description – the dispute between brother and sister in text C could be uttered in a heated or disappointed way, the complaint of a divorced father in text B could be aggressive or sad, and the despair in text G could be depressed or upset. Apparently, despite the care taken in selecting emotionally unambiguous texts, there remained space for interpretation of the texts, notably regarding the activation dimension.

Inter-rater agreement was assessed by calculating the mean pairwise correlations between subjects. The overall mean correlation was .423; the lowest mean correlation of one subject’s ratings with all other subjects’ ratings was .282, the highest was .550. Agreement was slightly higher among men (mean pairwise correlation between male subjects: .454) than among women (.376), higher among experts (.525) than non-experts (.410).

14.3.2 Results

The pattern of “goodness-of-fit” evaluations of a given prosodic configuration for the various emotional states as defined by the texts may best be illustrated using graphs showing the ratings for that prosody as a function of the position in activation-evaluation space. Figure 14.3.2 shows such graphs for the eight prosodic configurations used in the test. The exact numbers for the mean ratings are shown in Table 14.2.

Audio	Text								
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Mean**
A	51.4	29.6	28.5	46.5	21.4	43.6	70.7	44.5	42.0
B	44.6	66.3	75.4	25.2	68.6	40.3	42.2	41.7	50.5
C	69.6	65.5	72.7	65.9	48.4	66.1	59.7	74.1	65.2
D	47.8	69.5	70.3	49.3	67.0	50.8	49.2	54.9	57.3
E	16.7	64.0	69.4	24.7	71.8	25.0	31.1	31.1	41.7
F	51.3	23.4	26.1	43.1	15.9	47.9	71.1	43.4	40.3
G	27.1	7.2	11.8	18.6	4.8	26.9	43.7	19.8	20.0
H	47.3	44.3	44.8	54.8	37.6	52.5	57.8	45.0	48.0
Mean	44.5	46.2	49.9	41.0	41.9	44.1	53.2	44.3	45.6

Table 14.2: Evaluation test results, showing the mean goodness-of-fit rating for each combination of prosodic configuration and emotional text. Identifiers A–H match those in Figure 14.1. Row means express overall preference of a given prosodic configuration; column means express overall preference of all prosodic configurations for a given emotional text. This data is graphically presented in Figure 14.3.2.

Before the individual ratings are discussed in more detail, a few general observations are made.

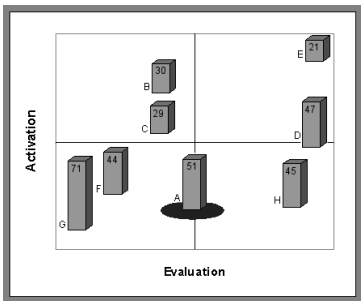
Generally, the activation was perceived as intended: There is a highly significant negative correlation between the difference in activation between text and prosody on the one hand and the goodness-of-fit ratings on the other (partial correlation controlling for difference in evaluation: $r = -.331$, one-tailed $p < .001$) – the more different the activation expressed in the text was from that expressed through the prosody, the lower the ratings for their combination.

A similar, but much smaller correlation pattern was found for the evaluation dimension: the distance between text and prosody on the evaluation dimensions showed a negative correlation with the ratings (partial correlation controlling for difference in activation: $r = -.079$, one-tailed $p = .002$). Clearly, the prosody succeeded in expressing the activation dimension and, to a very limited extent, the evaluation dimension.

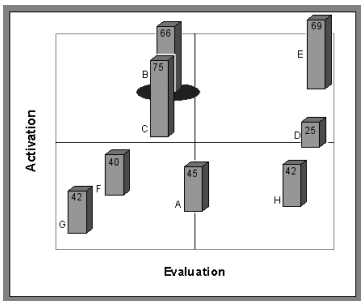
From the row totals in Table 14.2, overall preference patterns can be interpreted. It becomes clear that the prosody G was generally considered inappropriate, and only just acceptable for the corresponding text G. Several subjects had described the prosodic settings for G, in particular the slow speaking rate, as too extreme and therefore unnatural. Conversely, the prosody for state C, the emotional state closest to the neutral centre of the emotion space, was perceived as appropriate for most texts.

Regarding the details of the preference patterns, the extreme cases are easiest to characterise (see Figure 14.3.2). The prosody for the extremely active and positive case E was perceived as fitting best with the corresponding text E, while the prosody for the extremely negative and passive case G fitted best with the corresponding text G. For both, the ratings are relatively similar to the best value for texts close in activation, and generally decrease with increasing difference in activation. For the other prosodic configurations, the preference patterns are more complex.

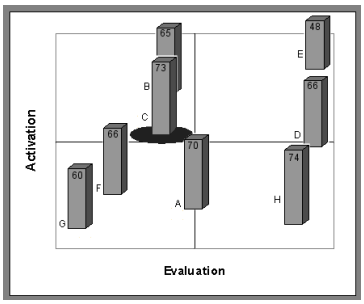
The moderate states B, C, A and F showed a clear pattern of high acceptability ratings for neighbouring states in addition to the targeted state. Prosody B was perceived as fitting well with text B, but even slightly better with the texts expressing the two neighbouring emotional states C and E. Similarly, prosody F was perceived as more appropriate for the states G and A, neighbouring F, than for text F itself. Prosody A had ratings extremely similar to those for prosody F. For prosody C, text C is narrowly rated second-best while



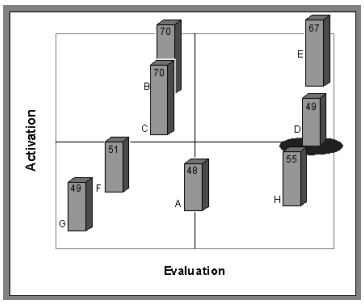
prosody A, all texts



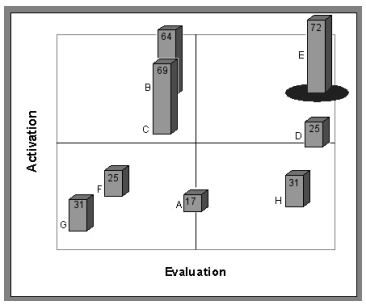
prosody B, all texts



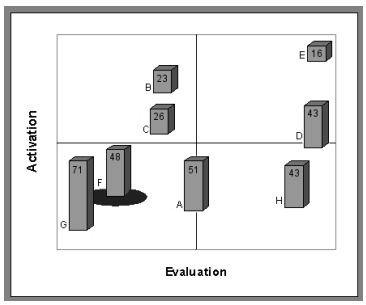
prosody C, all texts



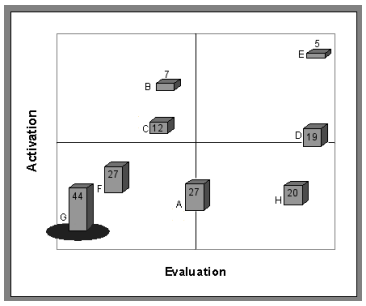
prosody D, all texts



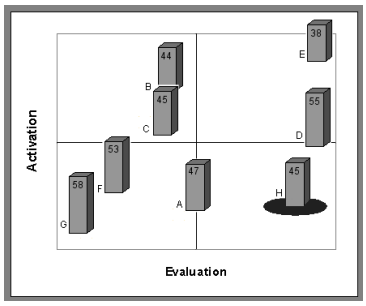
prosody E, all texts



prosody F, all texts



prosody G, all texts



prosody H, all texts

Figure 14.3: Evaluation test results. Each graph shows the mean goodness-of-fit ratings for one configuration of prosodic settings with all eight emotional texts. The height of a pillar represents the mean rating for the given combination of text and prosody, from 0 (worst) to 100 (best). The emotion intended to be expressed in the prosody is marked with a grey disk. See also Table 14.2.

the most distant texts E and G are rated least compatible. It is unclear, however, why prosody C was rated as best-fitting for text H.

The least conclusive results are found for the two positive and non-active emotional states, D and H. Prosody D fitted moderately well with most texts, and particularly well with the active texts B, C and E. Prosody H also fitted moderately with most states, with the exception of the most active state E. Apparently, these two prosodic configurations lack typicality for the type of emotional state they were meant to express.

14.4 Discussion

These results support the usefulness of the approach to synthesis of emotional speech by means of emotion dimensions. As expected, the prosodic configurations succeed best at conveying the activation dimension. Moreover, the appropriateness of a prosodic configuration for a given emotional state was shown to depend on the *degree* of similiarity between the emotional state intended to be expressed by the prosody and that in the textual situation description.

In agreement with previous findings for human speech, the evaluation dimension was found to be more difficult to convey through the prosody. Only a very small partial correlation was found between the distance between text and prosody on the evaluation dimension and the goodness-of-fit ratings. In particular, for two of the three positive emotions (D and H), no conclusive patterns were found in the results.

While the emotional prosody model used in this experiment (see Table 13.1, p. 158) was shown to roughly convey at least the activation aspect of the intended emotional state, it can not yet be expected to be optimal. Although most of the parameters modelled were taken from the literature, most were also manually fine-tuned. It would be most astonishing if this parameter configuration were already the optimal combination. The perception-based optimisation of this model would be a new, time-consuming task in itself (see Conclusion for some ideas on how to proceed). Still, the model was shown to be a suitable point of departure – only one prosodic configuration out of the eight (the extremely negative and passive emotion G) was rated unacceptable independently of the text.

Several aspects of the methodology applied in this evaluation test have been used for the first time in emotional speech synthesis research. First of all, a new criterion was used to measure the quality of the synthesised emotional speech, namely the question of appropriateness of the prosody for an emotional state defined through a textual situation description. This required texts to be identified which conveyed an unequivocal emotional state. The results of the written test (see 14.2) as well as the comments of the subjects during the listening test showed the difficulty of unambiguously defining the position on the activation dimension in a written text. It would appear that different aspects of an emotional message are differentially expressed through different "channels": Evaluation seems to be expressed more easily through text (see 14.2) and facial expression (Schlosberg, 1941), while activation is more easily conveyed through the voice. Still, this predominance does not preclude the perceptual experience of consistent and inconsistent emotional messages conveyed through different channels, as this listening test has shown.

A second novel aspect about the methodology employed in this test is the fact that the emotional states expressed included non-extreme emotional states, as well as a measure of similarity among them, due to their being modelled using emotion dimensions. The corresponding hypothesis, namely that emotional states similar to the intended state would be rated as more consistent than more distant states, was generally confirmed, with the exception of the prosodic configurations intended to express non-active and passive states.

The test setup (see Figure 14.2) departed from the "standard" rating scales by allowing subjects to listen to all eight prosodic variants of a given text repeatedly and to adjust their respective positions on the answer scale relative to the other stimuli. This method was chosen because the subjects in a pre-test unanimously preferred it over the more traditional test setup. It is clear that the general reliability of this test method has not been formally evaluated, and it remains to be seen whether the subjects' opinion that this test setup allows the subjects to give more reliable answers can be confirmed. Such a formal evaluation would require a large number of subjects performing both types of test. Clearly, this is beyond the scope of the present work.

Conclusion and outlook

The present dissertation has attempted to provide a conceptual frame facilitating an orientation in speech and emotion research. A number of concepts related to emotions, their description, and various aspects of their expression and perception were presented, illustrated by means of examples from the literature, and discussed in relation to the "bigger picture". It was argued that such a conceptual frame is a useful and even necessary prerequisite for taking conscious decisions as to the research orientation most promising for a given research question or task.

In the second, practical part of the thesis, one particular, applied research question was investigated: The expression of emotions in synthesised speech. The task was identified as listener-centered by nature. Likely application scenarios motivated the requirement of a gradual modelling of non-extreme emotional states, for which emotion dimensions were proposed as a suitable representation formalism.

Based on a literature review and an analysis of a corpus of spontaneous emotional speech, a set of emotional prosody rules was proposed and implemented in the MARY text-to-speech synthesis system. An evaluation of the resulting system was carried out, based on the question of appropriateness of the prosody for a given emotional situation. The results confirmed the general assumption underlying the approach: That at least for the activation dimension, different points on a continuum can be conveyed, and that the degree of similarity of two emotional states is reflected in the degree of appropriateness of a prosodic configuration for these states.

There are many possible next steps arising from this research.

A conceptual challenge is the integration of the highly specific properties of fullblown

emotions and the general tendencies of gradual emotional states into a single framework. Both can be conveyed through speech, but a descriptive framework capable of representing that full range remains to be developed.

On the methodological level, it would be necessary to perform a formal verification of the experimental setup used in the evaluation of the emotional speech synthesis system, to verify my subjects' claim that the common presentation of all prosodic variants of a given text will lead to more reliable scale ratings than their individual presentation in a classical one-at-a-time rating scale task. More generally, experimental methods based on the comparative presentation of auditory stimuli could be evaluated, possibly including an auditory variant of a grouping task for assessing similarity ratings.

A question relevant both from a scientific point of view and from an application usability point of view is that of perceived naturalness or authenticity. Currently, the conceptual toolkit available for investigating that question is small, and the existing experimental work appears to be even sparser. For customer acceptance of a voice based service expressing emotions, it is vital for an expressed emotion to be perceived as "honest" and not "fake". Badly simulated friendliness or empathy may be worse than a neutral speaking style. Currently, the cues leading to the perception of an emotion expression as "dishonest" are not yet known.

As to the emotional speech synthesis system developed in this dissertation, a natural next step would be the perceptual optimisation of the individual parameters to use in the emotional prosody rules, in a systematic, listener-centered way as done by Burkhardt (2000) for a categorical description of emotions. The current rule set could be used as a starting point, and individual parameters could be systematically varied. A serious problem is the large number of parameters modelled, which makes it impossible to systematically cross-combine different settings for all parameters. Instead, a step-by-step procedure would have to be applied, investigating only a few parameters at a time.

Regarding the expression of evaluation in the synthetic voice, it is still unclear whether more can be achieved than some unreliable, small effects. The available reports on natural human speech indicate that this may be difficult. However, the fact that smiling and frowning is audible (Tartter, 1980; Tartter & Braun, 1994) suggests that at least some progress should be possible.

A further starting point for future research is the field of multi-modal integration. As I found out as a "side effect" of the synthetic evaluation, it seems to be relatively difficult to express activation in written text. This suggests a partial complementarity of the different channels of emotion expression in a multi-channel emotional message. A question worth asking is that of the relative contributions of the different channels. More information on this topic would also be needed to clarify how far the idea of a voice which only needs to "fit to" the emotional message defined by other channels can be maintained. That this model works to a certain extent was shown in the evaluation experiment in this thesis; however, the same experiment also suggested that the voice is better than text in expressing activation. Answers to these questions should make it clearer in what respects a given expressive channel, such as the voice, needs to be well modelled, and which aspects of an emotional message may be handled more leniently, as they are better conveyed by a different channel anyway.

Finally, it is clear that the diphone synthesis technology used in this research, even if enhanced with the possibility to express vocal effort through the selection of different diphone inventories (Schröder & Grice, 2003), is a compromise, trading in a certain amount of naturalness for the possibility to explicitly model prosodic parameters. In the long term, different synthesis technologies will need to be investigated, be it formant synthesis systems with improved naturalness (Carlson et al., 2002) or unit selection systems with improved flexibility.

Annotated Bibliography

Albrecht, I., Haber, J., Khler, K., Schröder, M., and Seidel, H.-P. (2002a). "May I talk to you? :-)" – Facial animation from text. In *Proceedings of Pacific Graphics 2002*, pages 77–86.

System description. Integration of the MPI photorealistic facial animation system with the Mary text-to-speech system, including a mechanism to translate emoticons (smileys etc.) into emotion tags interpreted by the facial animation system.

Albrecht, I., Haber, J., and Seidel, H.-P. (2002b). Automatic generation of non-verbal facial expressions from speech. In *Proceedings of Computer Graphics International*, pages 283–293.

System description. Analyse speech signal (F0, intensity, pauses) to generate more natural, non-verbal facial expressions (eyebrows, gaze, head tilt, etc.).

Allen, J., Hunnicutt, S., and Klatt, D. H. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, UK.

Chapter 9: Rules for synthesis. Duration rules in close-to-algorithmic form. For English. Segment lengths are stretched/shrunk from "inherent" values as a function of their context, stress, word type etc.

Alt, S. (1997). Prosodie attitudinale de l'allemand: Indices sur la globalité des réalisations intonatives de quelques particules illocutoires. Mémoire de DEA, Sciences du Langage, Université Stendhal – Grenoble 3.

Experiment. Study of illocutionary particles ("ja", "doch", etc.) in German. Extensive literature overview.

Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., and Friederici, A. D. (2000). Accentuation and emotions – two different systems? In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 138–142, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Utterances produced with three emotions were acoustically analysed. Event-related brain potentials (ERPs) induced in listener subjects were measured. Conclusion: Accentuation and emotion seem to be two different systems.

Amir, N. and Ron, S. (1998). Towards an automatic classification of emotions in speech. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pages 555–558, Sydney, Australia.

Experiment, test system. Recognition of emotions in speech by distance to 5 reference points in a parameter space. Gradual, not binary membership in an emotion class. Calculating in relatively small windows (20/s), i.e. not the entire utterance is judged at once.

Amir, N., Ron, S., and Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 29–33, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Collected speech samples for five basic emotions (anger, fear, sadness, joy and disgust) and neutral speech, by asking their speakers to recall an emotional event appropriate for that emotion category. Three physiological variables (electromyogram of the corrugator, heart rate and galvanic skin response) were measured, and used as exclusion criteria if they deviated too much from the category mean. The remaining speech utterances were analysed using a number of speech variables. For each emotion category, a reference point in the N-dimensional space spanned by these speech variables was calculated. In order to determine the emotion with which a test utterance was spoken, a normalised distance measure in the N-dimensional space (the so-called Mahalanobis distance) was calculated between the feature vector representing the test utterance and each of the reference points. This distance measure was interpreted in terms of fuzzy category membership, i.e. it indicated the degree to which the utterance was considered part of the different emotion categories.

André, E., Klesen, M., Gebhard, P., Allen, S., and Rist, T. (1999). Integrating models of personality and emotions into lifelike characters. In *Proceedings of the Workshop on Affect in Interactions – Towards a new Generation of Interfaces*, pages 136–149, Siena, Italy. http://www.dfki.de/~klesen/archive/iwai-99.ps.gz

System prototypes. Three projects: Presence (Cyberella), Inhabited Marketplace, Puppet. Five-factor model of personality, cognitive model of emotion. Emotion dimensions Valence and Arousal. Two levels of emotion processing: reactive, behaviour-based layer; deliberative, goal-oriented layer.

André, E., Rist, T., van Mulken, S., Klesen, M., and Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In Cassell, J., Prevost, S., Sullivan, J., and Churchill, E., editors, *Embodied Conversational Agents*, pages 220–255. MIT Press, Cambridge, MA, USA.

Introduction to embodied conversational agents. Behaviour from script or (semi-) autonomous, centrally planned or distributed. Behaviour selection filtered, among other things, by personality and emotion (valence and arousal dimensions). Benefit of presentation teams. Overview of work at DFKI and elsewhere.

Andreeva, B. and Barry, W. J. (1999). Intonation von Checks in der Sofia-Varietät des Bulgarischen. *Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland*, pages 1–13.

Experiment. The authors investigated the role of intonation contour in distinguishing between checks (confirmation questions) and statements in Bulgarian, using sentences in

which sentence mode was not syntactically marked. A hypothesis for the typical question and statement contours was derived from a corpus of spontaneous speech. These contours were presented, in a factorial setting, in contexts indicating questions and statements, respectively. Subjects indicated on a scale how good the utterance fit with the context. Among other things, Andreeva and Barry found that the question intonation, presented in a context requiring a statement, was accepted as a statement, but subjects remarked that the speaker sounded angry or unhappy.

Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press, New York.

Pioneering work. Introduced the concept of *appraisal*, thus laying the basis for cognitive emotion theories.

Aubergé, V. (1992). Developing a structured lexicon for synthesis of prosody. In Bailly, G., Benoît, C., and Sawalli, T. R., editors, *Talking Machines: Theories, Models, and Designs*, pages 307–321. Elsevier Science Publishers, Amsterdam.

Model. Basic idea: "surface-level"-intonation as a superposition of contours, hierarchically structured in different levels. Analysing an in vitro corpus to obtain a contour lexicon that can be used for synthesis. See also Aubergé & Bailly (1995).

Aubergé, V. and Bailly, G. (1995). Generation of intonation: a global approach. In *Proceedings of Eurospeech 1995*, pages 2065–2068, Madrid, Spain.

Model. Complementary to Aubergé (1992). Prosody does not directly reflect syntactic structure, but rendez-vous points occur between intonation and other linguistic structures.

Aubergé, V., Grépillat, T., and Rilliard, A. (1997). Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. In *Proceedings of Eurospeech 1997*, Rhodes/Athens, Greece.

Experiment. Applied the gating paradigm (playing only the first x seconds of a stimulus) to attitude recognition from speech. Results: 1. Some of the 6 ICP attitudes/modalities are not well chosen: "incredulous question" and "suspicious irony" are easily mixed up. "Declaration", "Question" and "Evidence" are well distinguished. 2. Gating: non-linearity of prediction

Averill, J. R. (1975). A semantic atlas of emotional concepts. *JSAS Catalog of Selected Documents in Psychology*, 5:330. Ms. No. 421.

Resource. According to Cowie & Cornelius (2003), contains a list of 558 words "with emotional connotations".

Averill, J. R. (1980). A constructivist view of emotion. In Plutchik, R. and Kellerman, H., editors, *Emotion: Theory, research and experience*, volume 1, pages 305–339. Academic Press, New York.

According to Cornelius (2000), a key reference for the constructivist tradition in emotion theory.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database (CDROM)*. Linguistic Data Consortium, University of Pennsylvania, PA, USA.

Resource. A lexical database in several languages, containing among other things a phonetic transcription and morpheme boundary information.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57.

Overview. Short overview over the field of speech and emotion. Notion that arousal and valence are encoded in the voice, arousal more reliably than valence. Inter-individual differences in emotion expression. Few but well-selected publications.

Bailly, G. (2001). Visual synthesis. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland. http://www.ssw4.org

Overview. Model-based approaches: geometry-based; biomechanical models. Image-based approaches: visual or audio-visual unit selection synthesis. Control models: How to get from one viseme to another. Cohen&Massaro's coarticulation model; All-data triphone models. Audiovisual synchrony. 6 degrees of freedom shown at ICP.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

Experiment. Designed to verify the predictions formulated in Scherer (1986) about the vocal characteristics of emotions, based on the component process model. 14 emotions: identification test and acoustic / statistical analysis. Confusion matrix analysis. 3 dimensions of similarity between emotions: quality, intensity and valence. Push vs. pull factors of emotion expression. Hypothesis about prototype-based emotion recognition.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37:122–125.
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/oz/web/papers/ba-and-emotion.ps

Artistic recommendations. Conveying emotions is crucial for expressing personality and for believability. Inspired by Disney suggestions:

- internal emotional state must be clearly defined
- emotional state ⇒ thought process ⇒ visible in actions
- accentuate the emotion
  - –exaggeration
  - – anticipation
  - – staging (= on-stage co-ordination of characters' actions, in order not to confuse the spectator).

Batliner, A., Kießling, A., Kompe, R., Niemann, H., and Nöth, E. (1997). Can we tell apart intonation from prosody (if we look at accents and boundaries)? In *Proceedings of the ESCA Workshop on Intonation*, Athens, Greece.

Experiment. Theoretical questions: distinctive/redundant features, trading relations, criticising in vitro experiments. Short description of the prosodic recognition component in Verbmobil. Result: Best classification (= automatic prosodic labelling) when using *all* 276 features, no features can be neglected.

Baumann, S. and Trouvain, J. (2001). On the prosody of German telephone numbers. In *Proceedings of Eurospeech 2001*, pages 557–560, Aalborg, Denmark.

Experiment. Investigated the prosodic realisation of German telephone numbers. In a production experiment, subjects read telephone numbers. Typical prosodic patterns were identified. On the basis of these findings, a perception test was conducted in which several prosodic realisations were proposed. Clear preferences could be stated.

Benzmüller, R. and Grice, M. (1997). Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI. *Phonus 3, Research Report of the Institute of Phonetics, University of the Saarland*, pages 9–34.

Annotation system. Description of a the GToBI system for intonation annotation, including examples from a corpus of spontaneous German speech.

Biemans, M. and van Bezooijen, R. (1999). Biological gender and social gender in relation to voice quality. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 1249–1252, San Francisco, USA.

Experiment. Communication model: production → ... → perception. Some strong correlations between naive listeners' voice quality judgments and naive listeners' personality trait attribution. Nearly no correlation to speakers' self-assessed personality traits.

Black, A. and Lenzo, K. (2002). Festvox: Building synthetic voices, edition 1.6. Technical report, Language Technologies Institute, Carnegie Mellon University, PA, USA.
http://www.festvox.org

An open framework for building speech synthesis voices, closely linked to the Festival speech synthesis system (Black et al., 1999).

Black, A., Taylor, P., and Caley, R. (1999). Festival speech synthesis system, edition 1.4. Technical report, Centre for Speech Technology Research, University of Edinburgh, UK. http://www.cstr.ed.ac.uk/projects/festival

The famous open source text-to-speech synthesis system.

Black, A. W. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. In *Proceedings of Eurospeech 1995*, volume 1, pages 581–584, Madrid, Spain.

Algorithm. One of the first systems doing unit selection (CHATR). No clustering; phone-size units.

Black, A. W., Lenzo, K., and Pagel, V. (1998). Issues in building general letter to sound rules. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Algorithm. Letter-to-sound conversion using decision trees trained on pronunciation dictionaries, requiring only minimal manual interference. A grapheme can be mapped to zero, one or two phonemes. The context window is of limited size. Stress assignment can be trained by distinguishing stressed and unstressed vowels. Test results are presented. An implementation of the algorithm is available in the Festival system.

Black, A. W. and Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of Eurospeech 1997*, volume 2, pages 601–604, Rhodes/Athens, Greece.

Algorithm. Reference paper for the “clunits” cluster unit selection module in Festival. Describes the clustering and selection strategy employed.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of SIGGRAPH’99*.

System description. System creating 3D models of faces from 2D pictures by morphing a general 3D model.

Boëffard, O. (2001). Variable-length acoustic units inference for text-to-speech synthesis. In *Proceedings of Eurospeech 2001*, Aalborg, Denmark.

Algorithm. Apply the EM algorithm in order to find the best subset of n-multigrams (= 1-grams, 2-grams, ..., n-grams) representative of a big corpus. Very mathematical. Based solely on phonetic symbols for describing the “atoms”, not F0, position, etc.

Boula de Mareüil, P., Célérier, P., and Toen, J. (2002). Generation of emotions by a morphing technique in English, French and Spanish. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. http://www.lpl.univ-aix.fr/sp2002/papers.htm

Experiment. Actors produced speech material in six emotions (anger, disgust, fear, joy, surprise, and sadness), in English, French and Spanish. Using an automatised copy-synthesis technique (which the authors call a “morphing” technique), synthetic utterances with the same prosodic characteristics as the actor recordings are generated. Based on the prosodic properties of these utterances, language-specific rules for the generation of emotion-specific prosody were formulated. An outlook towards a perceptual evaluation is given.

Bradley, M. (1994). Emotional memory: A dimensional analysis. In van Goozen, S. H. M., van de Poll, N. E., and Sergeant, J. A., editors, *Emotions: Essays on emotion theory*, pages 97–134. Lawrence Erlbaum, Hillsdale, NJ.

Experiment. Investigated the effect of emotional connotation, as measured by the emotion dimensions *pleasure* and *arousal*, on memory.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle, WA, USA. http://www.coli.uni-sb.de/~thorsten/publications

Reference paper for TnT.

Breazeal, C. (1999). Robot in society: Friend or appliance? In *Proceedings of the Agents99 workshop on emotion-based agent architectures*, Seattle, WA, USA. http://www.ai.mit.edu/projects/kismet/Breazeal-Agents99.ps.gz

System description. A social robot with emotion modelling. Affective space consisting of three dimensions arousal, valence and stance (open–closed). Emotions correspond to regions in that space, and influence behaviour.

Breitenbücher, M. (1999). Textvorverarbeitung zur deutschen Version des Festival Text-to-Speech Synthese Systems. Technical report, IMS Stuttgart. http://elib.uni-stuttgart.de/opus/volltexte/1999/225

Algorithm. Thorough discussion of the phenomena to be accounted for in text normalisation in German TTS.

Brinckmann, C. and Trouvain, J. (2001). The role of duration prediction and symbolic representation for the evaluation of synthetic speech. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland. http://www.ssw4.org

Experiment. Two models for duration prediction (a CART and an adapted version of the Klatt rules) were combined with perfect and with imperfect symbolic representations (phoneme chains), and presented in a perception test. With a perfect symbolic input, the CART performed slightly better than the Klatt rules, but with an imperfect symbolic string, the difference vanished.

Brinckmann, C. and Trouvain, J. (2003). The role of duration models and symbolic representation for timing in synthetic speech. *International Journal of Speech Technology*, 6(1):21–31.

Experiment. Comparison of two duration models with optimal and with sub-optimal symbolic strings as input. The effect of sub-optimal symbolic input was found to outweigh the differences between the duration models.

Brøndsted, T., Nielsen, T. D., and Ortega, S. (1999). Classification of emotional attitudes in pet-directed speech. In *Proceedings of DALF*, Centre for Language Technology, Copenhagen. http://www.cpk.auc.dk/~tb

System. Emotion recognition: approval, disapproval, mainly based on pitch variation and vowel length; integrated with gesture and eye-tracking. Tamagochi-like application where a pet (dog)-like multimedia agent has to be influenced in its behaviour by the user’s emotion expression.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley.

Methodology. The book contains an early version (Figure 17, p. 69) of what Scherer (1974) later called the “Brunswikian lens model”. The model makes an explicit distinction between distal (objective) stimuli, their proximal (perceived) correlates and the perceptual impression resulting from them. It is proposed here in the context of vision and perceived size constancy.

Buck, R. (1999). Biological affects: A typology. *Psychological Review*, 106:301–336.

According to Cowie & Cornelius (2003), contains a mention of the idea of underlying emotions.

Bühler, K. (1934). *Sprachtheorie*. Gustav Fischer Verlag, Stuttgart, Germany. 2nd ed. 1965.

Theory. Very short description of the Organon model. Representation (symbol) as most important function of language (of the language sign). Expression (symptom) and appeal (signal) only mentioned marginally and anecdotally. All three functions are present simultaneously, but can vary in their salience.

Bühler, K. (1936). Das Strukturmodell der Sprache. *Travaux du Cercle Linguistique de Prague*, 6:3–12. Reprint 1968, Nendeln, Liechtenstein: Kraus Reprint.

A sort of reply. Short mention of the Organon model and of the "pointing function" of language (deixis).

Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesiser. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA.

Experiment. Copy-synthesis of three emotions (happy, angry, sad) + neutral, cross-combining prosody and diphone inventory. Inventories were recorded using emotionally expressive carrier sentences for the three emotions. Test sentences were semantically neutral. Results: Sadness is predominantly recognised from prosody, anger from the segmental properties in the inventory (voice quality and phonetic characteristics). Happiness was less-well recognised.

Burkhardt, F. (1999). Simulation des emotionalen Sprecherzustands „Freude" mit einem Sprachsyntheseverfahren. Magisterarbeit, TU Berlin.

Experiment, Implementation. Acoustic analysis of a subset of the Berlin SFB corpus, deduction of rules. Stimuli created using resynthesis (factorial design: F0, duration, ...), perception test "How happy (1–7)?" Rules deduced and implemented in speech synthesis, manipulating the MBROLA input format on the parameter level. Perception test ⇒ well recognised (binary answer, +/- happy).

Burkhardt, F. (2000). *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*. PhD thesis, TU Berlin. http://www.kgw.tu-berlin.de/~felixbur/publications/diss.ps.gz

Experiment. A number of detailed prosody rules for expressing different emotions in German synthesised speech were determined in perception tests. In a first step, relevant variables were systematically varied in a factorial design, without assumptions about the correct settings for a given emotion. The values found optimal in this step were then combined with different combinations of additional parameters, different for different emotions. Prosody rules are reported in detail. See also English language summary in Burkhardt & Sendlmeier (2000).

Burkhardt, F. and Sendlmeier, W. F. (1999). Simulation emotionaler Sprechweise mit konkatenatierender Sprachsynthese. In *Proceedings of the 30. Jahrestagung der GAL*, Frankfurt/Main, Germany.

Experiment. Comparison of copy synthesis and emotion rules ⇒ stimuli generated by rule are generally recognised much better. Description of the rules for the generation of anger, joy, fear, sadness.

Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 151–156, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment, system (prototype). KLSYN88-based formant synthesis.

1. systematic variation of many parameters, perception test ⇒ systematic assessment of their effects on emotion judgments.
2. emotion "prototypes" and variants (with added features) ⇒ good recognition (relatively diagonal confusion matrix); additional info regarding "helpful" parameters.

Cahn, J. E. (1989). Generating expression in synthesized speech. Master's thesis, MIT Media Lab. http://www.media.mit.edu/~cahn/masters-thesis.html

Experiment. A detailed account of Janet Cahn's early work in emotional speech synthesis. See Cahn (1990) for a summarised version.

Cahn, J. E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19.

Experiment. The first and most frequently cited reference for a prototype on synthesis of emotional speech. Parameters for DECTalk for six emotions. Prototype "Affect Editor" like a little TTS (and not just a TTS backend), i.e. has access to high-level linguistic concepts such as phrases and accents. Derived one setting of parameters per emotion from the literature. Recognition rates: 42–52% (sad: 91%). Parameter values for many parameters.

Cahn, J. E. (1998). Generating pitch accent distributions that show individual and stylistic differences. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Algorithm. A memory model (periodic 50x50 space, attention focus does a random walk, "recognition" of an item if a corresponding memory entry is found within a search radius around the attention focus) is used to determine intonation patterns by judging the read item as "known" or "new". Three styles (child, adult, knowledgable) are realised by varying the search radius size.

Campbell, N. (2000). Databases of emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 34–38, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Overview. Presentation of methods for obtaining databases of emotional speech: Acting, stimulation of emotions through reading expressive paragraphs, elicitation, and "found speech", i.e. emotional expression found in material originally recorded for a different purpose. Also contains a short presentation of the objectives of the JST CREST project.

Campbell, N. and Iida, A. (1999). Multi-level labelling of speech for synthesis. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 499–502, San Francisco, USA.

System description. Short description of the mechanism used by CHATR. Labelling: phonetic: broad phonetic, automatic; prosodic: prominence + boundaries, automatic; phonatory: under development, to distinguish happy/sad/angry voice quality.

Campbell, N. and Marumoto, T. (2000). Automatic labelling of voice-quality in speech databases for synthesis. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China.

English-language version of Marumoto & Campbell (2000).

Cappella, J. N. (1993). The facial feedback hypothesis in human interaction. *Journal of Language and Social Psychology*, 12:13–29.

Overview. Short review of the evidence supporting the FFH. Conclusion: Reliable effect, small in magnitude, only proven to work for general valence (not enough evidence to decide about individual emotion categories); can initiate as well as modify subjective emotional states. In addition, Cappella proposes the Interpersonal FFH by combining the FFH with imitation patterns in interaction.

Carlson, R., Granström, B., and Nord, L. (1992). Experiments with emotive speech – acted utterances and synthesized replicas. In *Proceedings of the 2nd International Conference of Spoken Language Processing*, Banff, Canada.

Experiment. One neutral sentence produced by two speakers with three emotions (happy, sad, angry) and neutrally. Resynthesis cross-combined intonation and timing from one emotion to the utterance (speech material) of another emotion ⇒ tendency for prosody to be more important than voice quality, very clear for sadness, less clear for happy and anger. Timing alone, on a neutral sentence, is not sufficient to convey the emotion.

Carlson, R., Sigvardson, T., and Sjölander, A. (2002). Data-driven formant synthesis. Progress Report 44, KTH, Stockholm, Sweden. http://www.speech.kth.se/qpsr/tmh.

Experiment. A new method for driving a formant synthesiser was explored. The acoustic parameters for units to be synthesised were extracted from a database. A fully data-driven method and a hybrid method (in which only the vowel formants were data-driven, while the consonants were rule-based) were compared to a rule-based system and rated in a perception test for clarity and naturalness. The hybrid system was rated best.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of Computer Graphics (SIGGRAPH'94)*, pages 413–420.

Overview. Description of the domain, the problems, and this system.

Cauldwell, R. T. (2000). Where did the anger go? The role of context in interpreting emotion in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 127–131, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Two short utterances (falling wh-questions starting at high pitch) are interpreted as "anger" when presented in isolation, but as "neutral" when presented with preceding context (talking to a three-year-old child). Effect independent from order of presentation (even when in-context version is presented first, the effect remains).

Chung, S.-J. (1999). Vocal expression and perception of emotion in Korean. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 969–972, San Francisco, USA.

Experiment. Simple valence and emotion intensity perception test; valence not gradual (+1, 0, -1 only); emotion intensity and activation dimension treated as synonymous. Relatively high (inter-cultural) listener agreement for negative states; differences for positive emotion.

Church, A. T., Katigbak, M. S., Reyes, J. A. S., and Jensen, S. M. (1998). Language and organisation of Filipino emotion concepts: Comparing emotion concepts and dimensions across cultures. *Cognition & Emotion*, 12(1):63–92.

Experiment. Large experiment investigating Filipino emotion concepts, comprising three main studies. 1) Establishment of a list of emotion terms, out of which a list of 256 Filipino emotion-related adjectives was compiled; 2) Comparison of the Filipino and English emotion lexicons, using several cluster-based techniques; and 3) Conceptual organisation of Filipino emotion terms. In the third study, similarity judgments of emotion terms were analysed in a way similar to Shaver et al. (1987): Using hierarchical cluster analysis, concept clusters or categories were identified; using multi-dimensional scaling, emotion dimensions were established. The two analysis methods provided complementary information about the organisation of emotion concepts. The dimensions in the two-dimensional solution could be identified as pleasantness and arousal. Emotion terms formed a near-perfect circumplex structure in two-dimensional space. In the three-dimensional solution, the third dimension was not dominance, but rather related to certainty/uncertainty in combination with negative emotions. Altogether, the authors conclude that "the results provided better support for the cross-cultural comparability of emotion concepts and experience than for a strong social constructivist view" (p. 63).

Clore, G. L. and Ortony, A. (1991). What more is there to emotion concepts than prototypes? *Journal of Personality and Social Psychology*, 60(1):48–50.

Theory of emotion concepts. An answer to a 1990 article by Russell entitled "In defense of a prototype approach to emotion concepts". Both "classical" definitions including necessary and sufficient conditions and prototypes are suggested to be useful means for describing emotion concepts. Interesting, explicit distinction between concepts and the phenomena which they are concepts of, including examples concerned with birds and fish. Suggestion that whether emotion concepts are relevant for emotion theories depends on the type of theory: "whereas a view based solely on physiological and cognitive concepts might have no place for lay concepts in a conception of emotion, a view based on a social constructivist perspective ... would reserve for them a critical role. From such a perspective, the presence or absence of emotion in some situation would depend on the nature of the social context and on the conception of emotion held by that community." (p. 49)

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed account of the Stroop effect. *Psychological review*, 97(3):332–361.

Model. Introduce the distinction automatic/controlled. Propose to see the distinction as a continuum, on which a task can change place due to practise. Description of the Stroop task. A neural net type model with learning, and processing time, "attention" as the facilitation of one processing path, "learning" enhances connection strength.

Cornelius, R. R. (1996). *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Prentice-Hall, Upper Saddle River, NJ.

Overview. An overview of the four major traditions in emotion theory: Darwinian, emphasising evolutionary origins of emotions; Jamesian, emphasising bodily involvement in emotion; cognitive, emphasising the appraisal process leading to emotions; and social constructivist, emphasising the culturally constructed nature of emotion phenomena. Includes, as an appendix, an introduction into the neurophysiology of emotion.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 3–10, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Theory. Overview of the four theoretical perspectives on emotion. Darwinian: emotion expression evolutionarily shaped. Jamesian: same idea, + subjective experience through body. Cognitive: appraisal as link between stimulation & response; Social constructivist: emotions fulfil social functions, are culturally shaped. First three can be integrated more or less, fourth less easily. Plea for awareness of where one's work is situated.

Cowie, R. (2000). Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 11–18, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Overview. Domain of speech and emotion: Fullblown emotions vs. emotional states; cause-type (speaker-centered) and effect-type (listener-centered) descriptions. Categories and the need for a complementary representation putting categories in relation to each other. Types of representations: biological; cognitive appraisal; dimensional (valence; activation; others). Timing. Mixing, masking, simulation. Humour. Proposals: Feeltrace, basic English emotion vocabulary, richer representations.

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication Special Issue on Speech and Emotion*, 40(1–2):5–32.

Review. Thorough overview of the descriptive frameworks available for speech and emotion research. Starts by defining "fullblown emotion", "underlying emotion", "emotional states" and "emotion-related states". States different goals and orientations that speech emotion research can focus on. Points out that subtle and complex emotional states are commonly expressed in speech, rather than extreme fullblown emotions. A description needs to be found for the emotional colouring implied in emotion-related states such as attitudes. Then goes on to discuss the different descriptive framesworks. Category labels are discussed starting with the richness of everyday emotion terms and emotion-related states, before approaches for reduction to basic emotion categories and superordinate categories are presented. A point is made regarding the usefulness of embedding categories in underlying structures. A number of available structures are discussed, namely biological, abstract dimensional, and cognitive-structural descriptions. Aspects adding to the complexity of the domain are mentioned: The temporal properties of the phenomenon studied, and restraint, ambivalence and simulation effects which may occur in social interaction. Finally, available descriptive tools are presented, ranging from lists of key emotions over cause-oriented descriptions (self report, antecedent conditions, physiological measures, facial expression, and subsequent actions) to effect-oriented descriptions (forced-choice perception tests, structural assignment, and dimensional descriptions including Feeltrace).

Cowie, R. and Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proceedings of the 4th International Conference of Spoken Language Processing*, pages 1989–1992, Philadelphia, USA.

Experiment, tool. Linked to McGilloway et al. (1995). Term: "augmented prosodic domain" as the scope of ASSESS analysis. Deafened people's speech. Flattened affect in schizophrenia.

Cowie, R., Douglas-Cowie, E., Appolloni, B., Taylor, J., Romano, A., and Fellenz, W. (1999a). What a neural net needs to know about emotion words. In Mastorakis, N., editor, *Computational Intelligence and Applications*, pages 109–114. World Scientific & Engineering Society Press.

Experiment. Empirical classification of emotion words. Basic English emotion vocabulary. Placement of English adjectives in activation-evaluation space. Richer information ("schema") about the adjectives, including characteristics of the situation (self, other, past, present, future) and action tendencies.

Cowie, R., Douglas-Cowie, E., and Romano, A. (1999b). Changing emotional tone and its prosodic correlates. In *Proceedings of the ESCA Workshop on Dialogue and Prosody*, pages 41–46, Eindhoven, The Netherlands.

Experiment, tools. Feeltrace: Tool for continuous emotional perceptual evaluation in two-dimensional activation-evaluation space. Colour coding for emotions from Plutchik. ASSESS: pointwise measures (mean, s.d.) vs. piecewise measures (rise duration, etc.). Four types of piece: F0 rise, F0 fall, F0 level, pause.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000a). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Rating tool. The paper introduces the Feeltrace tool for rating the emotion expressed in audio-visual stimuli. Emotions are represented as a two-dimensional space representing the emotion dimensions *activation* (from very passive to very active) and *evaluation* (from very negative to very positive). The perceived emotional state can be indicated as it changes over time, by moving a cursor in the two-dimensional space.

Cowie, R., Douglas-Cowie, E., and Schröder, M., editors (2000b). *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Northern Ireland 2000*, Northern Ireland. Textflow, Belfast. http://www.qub.ac.uk/en/isca/proceedings

Proceedings of the first ISCA workshop bringing together people interested in Speech and Emotion research from different backgrounds.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

Overview. An overview of the issues involved in automatic emotion recognition from face and voice. Reviews descriptive frameworks of emotional states, as well as the literature

on speech and emotion correlates. Presents methods for the automatic analysis of facial features from video. Discusses how these can be used for emotion recognition.

Cowie, R., Sawey, M., and Douglas-Cowie, E. (1995). A new speech analysis system: ASSESS. In *Proceedings of the 13th International Conference of Phonetic Sciences*, volume 3, pages 278–281, Stockholm, Sweden.

System description. The paper introduces the ASSESS system for semi-automatic analysis of acoustic speech parameters. The system performs a largely automatic and robust analysis of a large number of speech parameters. It generates a simplified core representation of the speech signal based mainly on the F0 and intensity contours. Key 'landmarks' are then identified, including peaks and troughs in the contours as well as boundaries of pauses and fricatives. Measuring the 'pieces' between these landmarks gives rise to a range of variables called 'piecewise'. They provide a rich description of the way contours (of pitch and intensity) behave over time. Variables, piecewise and others, are then summarised in an array of statistics (covering central tendency, spread and key centiles). Additional measures deal with properties of 'tunes' (i.e. segments of the pitch contour bounded at either end by a pause of 180 ms or more) as well as with spectral properties.

Cronk, A. and Macon, M. (1998). Optimized stopping criteria for tree-based unit selection in concatenative synthesis. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 5, pages 1951–1955, Sydney, Australia.

Algorithm; introduction. Very good introduction into CART use for unit selection using clustering. Alternative algorithm for determining best tree size by using unseen data.

Daelemans, W. and van den Bosch, A. (1996). Language-independent data-oriented grapheme-to-phoneme conversion. In van Santen, J., Sproat, R. W., Olive, J., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 77–90. Springer Verlag, New York. http://cnts.uia.ac.be/~walter/papers/1996/db96.ps

Algorithm. Description of an algorithm for grapheme-to-phoneme conversion. The algorithm consists of three main steps: Alignment of graphemic and phonemic forms of words; compression of the aligned data into an "IG-tree" (IG = information gain); and lookup. The IG tree is a decision tree with the graphemes and their minimal disambiguating context as the edges and the phonemes as the leafs. In order to allow for extrapolation to unseen data, a "best guess" phoneme is saved at each non-final node, based on the most frequent transcription in the training data. The algorithm requires a one-grapheme-to-one-phoneme alignment, allowing for epsilon phonemes. It is not possible to align multiple phonemes with a single grapheme.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam, New York.

Overview. An influential book about the role of emotions, written by a neurologist in the Jamesian tradition of emotion theories. Presents evidence regarding the importance of emotions in descision-making: Patients with certain brain lesions do not feel emotions and are unable to make decisions. Introduces the idea of "somatic markers", bodily experiences associated with emotionally relevant situations. Patients who do not perceive their body do not experience emotions. Apart from this "body loop" of emotion experience, suggests an "as if" body loop in the brain simulating the physical state.

Davitz, J. R. (1964a). Auditory correlates of vocal expressions of emotional feeling. In Davitz, J. R., editor, *The Communication of Emotional Meaning*, pages 101–112. McGraw-Hill, New York.

Experiment. One of the first experiments studying the correlation between emotion dimensions and speech variables. 7 speakers produced two semantically neutral standard sentences in each of 14 emotional tones. The sentences were embedded into emotion-specific paragraphs read by the speakers. Three types of ratings were obtained. First, the standard sentences were presented in a forced choice recognition task; second, they were rated on scales representing the auditory variables loudness, pitch level, timbre and speech tempo. Third, the emotion-specific paragraphs were rated on the three semantic differential dimensions proposed by Osgood et al. (1957), valence, strength, and activity. Correlations were calculated between the auditory ratings and the dimensional ratings of the 14 emotions. Correlations were found to be strongly significant for activity, but not for valence and strength. The correlations were in the sense that expressions of more active emotions were also rated louder, higher-pitched, more "blaring" in timbre, and faster. The author suggests that valence and strength may be conveyed by more subtle auditory cues than the ones he studied.

Davitz, J. R. (1964b). A review of research concerned with facial and vocal expressions of emotion. In Davitz, J. R., editor, *The Communication of Emotional Meaning*, pages 13–29. McGraw-Hill, New York.

Review. Short overview about the early work in facial and vocal emotion research.

de Gelder, B. and Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, 14(3):289–311.

Experiment. A continuum between two photographs of faces expressing happiness and sadness was generated using morphing techniques. Each face was presented, in a factorial design, paired with happy, sad, and no voice. Subjects indicated whether they perceived happiness or sadness. In a first experiment, subjects were instructed to attend both to the face and the voice; in the second, they were told to ignore the voice and to attend only to the face. In both experiments, both the face and the voice had an effect on responses as well as on response timing: Subjects responded slower to ambiguous stimuli. In the condition where subjects were told to ignore the voice, the effect of the voice was smaller but significant, indicating a non-voluntary aspect to the integration of the two modalities. In a third experiment, subjects were presented with a vocal continuum between happiness and fear, and happy or fearful faces. They were told to ignore the faces and judge the perceived emotion solely based on the voice. Again, a small but significant effect of the facial emotion was found in addition to the vocal emotion.

de Rosis, F. and Grasso, F. (2000). Affective text generation. In Paiva, A., editor, *Affective Interactions*, volume 1814 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag. http://www.csc.liv.ac.uk/~floriana/papers/iwai99.ps.gz

System description. NLG in view of the hearer's emotional state. Some basic questions. Examples from drug prescription/medical domain. Not about expressing a virtual talker's emotion. Interesting references.

Dietz, R. B. and Lang, A. (1999). Æffective agents: Effects of affect on arousal, attention, liking & learning. In *Proceedings of the 3rd International Cognitive Technology Conference*. http://www.cogtech.org/CT99/Dietz.htm

System, experiment. Polara project: non-verbal emotional behaviour of an "alien robot" agent. Three emotion dimensions: arousal, valence, control. Picture of "usual" points in this three-dimensional emotion space. Agent's emotion display makes users feel more emotional; other effects did not reach significance.

Dines, J., Sridharan, S., and Moody, M. (2001). Application of the trended Hidden Markov Model to speech synthesis. In *Proceedings of Eurospeech 2001*, Aalborg, Denmark.

Algorithm. Trended hidden Markov models = each state is not a constant, but a polynomial function. Very mathematical. Clustering of the HMMs. Intelligibility below diphone synthesis.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication Special Issue Speech and Emotion*, 40(1–2):33–60.

Overview. Discussion of the issues arising in the creation of state-of-the-art emotional speech databases, written by representatives from the three largest emotional speech database projects to date: The Reading/Leeds Emotion in Speech Project, the Belfast database, and the JST/CREST expressive speech processing project. Discussion of the main issues to take into account; a long list of existing "datasets" of emotional speech; and a summary of the particularities of each of the three large databases created by the authors' teams.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2000). A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 39–44, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Resource description. Focus on ecologically valid emotion expression. Audio-visual. Against the "benign interpolation hypothesis" underlying the study of archetypal emotions. Short review of existing databases (A/V). Sources: from interviews and TV. Archetypal emotion expressions are rare. Associated descriptions: Feeltrace + categories (= perceived emotional content); ASSESS; relevant features.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht.

Textbook. Introductory book presenting the various processing components involved in TTS. Written by the author of the MBROLA speech synthesis algorithm.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesisers free of use for non commercial purposes. In *Proceedings of the 4th International Conference of Spoken Language Processing*, pages 1393–1396, Philadelphia, USA.

System description. Short description of the MBROLA project. Description of how to join as a user/developer/database provider.

Edgington, M. (1997). Investigating the limitations of concatenative synthesis. In *Proceedings of Eurospeech 1997*, Rhodes/Athens, Greece.

Experiment. One actor, several sentences, 5 emotions + neutral. No preselection. Natural recordings well recognised, synthesised replica worse (mean=42%). Synthesised versions: Neutral was the default answer; anger, happiness, boredom: recognised; fear, sadness: chance level. F0, duration and energy modelled.

Ehlich, K. (1986). *Interjektionen*. Max Niemeyer Verlag, Tübingen.

Resource. (Subjective) linguistic description of the German interjections

- hm
- ah, oh, ih, ai, au
- na

with several tonal variants each and description of the meaning.

Ekman, P. (1977a). Bewegungen mit kodierter Bedeutung: Gestische Embleme. In Posner, R. and Reinecke, H.-P., editors, *Zeichenprozesse*, pages 180–198. Athenaion, Wiesbaden.

Theory. Emblems = verbalisable, generally known meaning, voluntarily employed. Culture specific, but mostly iconic (imitates the corresponding movement). Emotional emblems: clearly distinguishable from real emotion expression, because of different muscle tension, different duration (longer or shorter). Role of emblems in dialogue: Listener-response (turn-taking).

Ekman, P. (1977b). Biological and cultural contributions to body and facial movement. In Blacking, J., editor, *The anthropology of the body*, pages 39–84. Academic Press, London.

Overview article. Describes emblems, body manipulators, illustrators, and emotional expressions, the latter in some detail. In particular, a relatively detailed description of display rules is given.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4):384–392.

Overview. Short historic presentation of Ekman's work. Summarises contributions of facial expression research to emotion theory according to nine points: 1. Study emotion 2. Consider both nature and nurture 3. Search for emotion-specific physiology 4. Specify the events that precede emotions 5. Examine ontogeny 6. Examine more than verbal behavior 7. Consider emotions as families 8. Consider emotions to be discrete states 9. Consider expression in determining how many emotions there are. Interesting detail: Mentions that several positive emotions share the same facial expression; they are differentiated by an observer using context.

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Power, M. J., editors, *Handbook of Cognition & Emotion*, pages 301–320. John Wiley, New York.

According to Cowie & Cornelius (2003), contains a list of basic emotions.

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, 58(2):342–353.

Experiment. Duchenne smile (with the orbicularis oculi):

- related to enjoyment (subjective experience) and brain asymmetry;
- amount of D-smiling can distinguish which of two experiences is more positive;
- more left-sided anterior temporal and parietal brain activation than for other smiles.

Other smiles: not related to enjoyment; right frontal, anterior temporal, and parietal activation (similar to withdrawal pattern).

Eriksson, A. and Traunmüller, H. (1999). Perception of vocal effort and speaker distance on the basis of vowel utterances. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2469–2472, San Francisco, USA.

Experiment. Speakers utter single vowels at different distances to an addressee, with normal phonation and whispered. These recordings are played at different amplitudes to listeners who judge (1) the distance between the speaker and the addressee, independently of their own distance from the speaker (thus measuring perceived vocal effort); and (2) their own distance from the speaker, independently of the distance of the addressee (thus measuring perceived distance). For normally phonated vowels, perceived vocal effort was more strongly correlated to original vocal effort (as measured by recorded sound pressure level) than to listening amplitude, while perceived distance was more strongly correlated to listening amplitude than to original vocal effort. This was not the case for whispered vowels. These findings support the idea that listeners infer vocal effort from voice quality more than from signal amplitude.

Fairbanks, G. and Hoaglin, L. W. (1941). An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monograph*, 8:85–91.

Experiment. One of the "Fairbanks studies", often cited as one of the earliest experimental investigations of emotional speech.

Fairbanks, G. and Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monograph*, 6:87–104.

Experiment. One of the "Fairbanks studies", often cited as one of the earliest experimental investigations of emotional speech.

Fehr, B. and Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113:464–486.

Theory, experiment. Apply prototype theory to emotion concepts. The authors are explicitly investigating emotion *concepts*, not the emotions themselves. According to prototype theory, concepts are not defined by a set of necessary and sufficient properties. Instead, category membership is a matter of degree, i.e. categories have an internal structure. The best examples are called prototypes; other instances vary in their degree of family resemblance to the prototypes. Fehr and Russell carried out a number of experiments for verifying whether prototype theory was appropriate for describing people's emotion concepts, and found the predictions confirmed. In addition, category boundaries are "fuzzy", i.e. there are unclear cases, items for which category membership cannot be agreed upon. Therefore, Fehr and Russell suggest that prototype-based descriptions might be appropriate for describing lay people's emotion concepts. In addition, they may or may not be usefully applied in the scientific concepts used in investigations of the emotion phenomena themselves.

Fernandez, R. and Picard, R. (2000). Modelling drivers' speech under stress. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 219–224, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Elicited stress in their subjects by asking math questions while driving in a driving simulator. The stress level was operationalised through driving speed (fast/slow) and frequency of the math questions (frequent/infrequent). Speech was analysed using spectral subbands, and automatic classifiers were trained to predict the stress condition on the basis of the speech analysis.

Fried, I., Wilson, C. L., MacDonald, K. A., and Behnke, E. J. (1998). Electric current stimulates laughter. *Nature*, 391:650.

Experiment. Electric stimulation in the anterior part of the SMA can elicit laughter, accompanied by the sensation of merriment. At low currents only a smile was present, while at higher currents, duration and intensity of the laughter increased.

Frijda, N. H. (1986). *The Emotions*. Cambridge University Press, Cambridge, UK.

Theory. Typically cited as *the* reference for action tendencies. Action tendencies, "states of readiness to execute a given kind of action" (p. 70), are introduced as a defining aspect of emotion. The general term "activation" refers to "an organism's state of readiness for action" (p. 90). The book contains a broad overview of the emotion phenomenon, providing a wealth of information about behavioural, physiological and experiential aspects of emotions, as well as an extensive discussion of emotion antecedents. The last chapter outlines an emotion theory.

Frijda, N. H. and Mesquita, B. (1994). The social roles and functions of emotions. In Kitayama, S. and Markus, H. R., editors, *Emotion and Culture*, pages 51–87. United Book Press, Baltimore, MD, USA.

Theory. 1. Emotion Theory. 2. Social interaction: Emotions are recognized by others; inform about the emotional relevance of something ($\Rightarrow$ associations accompanied by emotion are learned more rapidly); influence interpersonal relationships (anger $\rightarrow$ social correction; shame $\rightarrow$ conformity); evoke responses from others. Some emotions (anger) regulate when felt; other emotions (shame) regulate by anticipation/avoidance of feeling them. These emotions (and not only their display) exist to fulfil regulatory social functions.

Gerrards-Hesse, A., Spies, K., and Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: A review. *British Journal of Psychology*, 85:55–78.

Overview. Different methods of induction of emotions in the laboratory. The methods that showed to be effective in the literature.

Gobl, C. and Ní Chasaide, A. (2000). Testing affective correlates of voice quality through analysis and resynthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 178–183, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Very controlled creation of seven voice qualities using the LF model and resynthesis of one utterance with the KLSYN88a formant synthesiser. Perceptual judgments of these voice qualities on seven 7-point scales: stressed/relaxed, content/angry, ... Results: two groups of voice qualities: tense/harsh $\rightsquigarrow$ aggression, breathy/creaky/... $\rightsquigarrow$ relaxed, unagressive.

Greasley, P., Sherrard, C., and Waterman, M. (2000). Emotion in language and speech: Methodological issues in naturalistic approaches. *Language and Speech*, 43(4):355–375.

Experiment. Two experiments showing the need for more complex descriptions in order to account for the exact emotional meaning perceived in naturally occurring emotional speech. First, a sample of 89 audio clips of naturally occurring emotional speech was presented both in a free response perception test and in a forced choice test with five basic emotion categories (anger, disgust, fear, sadness, happiness). The clips included verbal content. In the forced choice test, agreement between subjects was limited: While very high agreement was found for clips rated as "happiness" (the only "positive" emotion category provided), only 54% of the "negative" stimuli were "pure" in the sense that they were identified as one given emotion category at a statistically significant level (67%). The authors argue that the remaining, "mixed" stimuli could only have been described appropriately if several answer categories had been allowed. For the "pure" emotions, 70% of the free-response answers were representative of the chosen basic emotion, but differing in quality and intensity. In the second experiment, gradual valency ratings of emotion words were obtained with and without (aural, verbal and situational) context. The valency of the context was found to influence the rated valency of the emotion words, in the sense that a positive or neutral word was rated more negatively when in a negative context than when rated in isolation.

Green, R. S. and Cliff, N. (1975). Multidimensional comparisons of structures of vocally and facially expressed emotion. *Perception & Psychophysics*, 17(5):429–438.

Experiment. Similarity judgments of emotional speech stimuli led to emotion dimensions. One actor expressed 11 emotions speaking letters of the alphabet. The stimuli were presented in pairs to subjects who rated their similarity. Multi-dimensional scaling techniques yielded a three-dimensional interpretation with the dimensions "pleasant-unpleasant", "excitement", and "yielding-resisting". Each of the stimuli was also rated on seven semantic differential adjective scales measuring tone-of-voice. A principal component factor analysis suggested two dimensions, the first being related to the thin-thick and high-low pitch sound of the voice, the second representing pleasant-unpleasant feelings. The two dimensions found in semantic differential scales corresponded strongly to the first two dimensions found in similarity judgments. Stimuli fell in a V-shape in two-dimensional space: Stimuli that were either highly pleasant or unpleasant were also excited, while stimuli unmarked in pleasantness were low in excitement. A comparison with an earlier study of facially expressed emotion showed similarities in the first two dimensions.

Grice, M., Baumann, S., and Benzmüller, R. (2002). German intonation in autosegmental-

metrical phonology. In Jun, S.-A., editor, *Prosodic Typology*. Oxford University Press.

Intonation theory. Short overview of German intonation models. Description of GToBI. Comparison of GToBI with other autosegmental-metrical intonation models for German.

Guerrero, L. K., Andersen, P. A., and Trost, M. R. (1998). Communication and emotion: Basic concepts and approaches. In Andersen, P. A., editor, *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, pages 3–27. Academic Press, New York.

Theory. Distinction affect-emotion-mood. Three description frameworks for emotions: Discrete or basic emotions, emotion dimensions including circumplex and Plutchik's model, and emotion prototypes.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. http://www.lpl.univ-aix.fr/sp2002/papers.htm

Theory. Introduction to effort code and production code. Three codes which have biological origins play a role as universals in pitch variation: Frequency code, Effort code and Production code. Grammaticalisation will often mimic form-function relations according to these codes. Frequency code: Low F0 signals a large speaker, and consequently assertiveness or dominance. Effort code: High F0 and precise articulation signal high speaker effort, and consequently importance or agitation. Production code: Low energy and lowering F0 signal that the lungs are empty, and consequently finality (and no affective meaning). Peak delay can substitute peak height perceptually; high register can substitute pitch span. Examples for grammatical meaning.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice quality. *Acta Otolaryngologica*, 90:441–451.

Experiment. A short story was produced by each of 17 pathological voices, and was rated by expert listeners on scales representing 28 terms used for voice description. Factor analysis of these ratings yielded five bipolar factors: "unstable–steady", "breathy–overtight", "hyperfunctional–hypofunctional", "coarse–light", and "head register–chest register". Acoustic measures of mean F0 as well as of the long-term average spectrum were taken. In three spectral bands (0–2kHz, 2–5kHz, 5–8kHz), the maximum sound pressure level (the peak in a frequency–SPL diagram) was determined, and difference values of these SPL levels were calculated. Correlations were calculated between the factors which had emerged from the perceptual ratings and the acoustic measures. All factors were correlated with mean F0. In addition, the factor "breathy" was correlated with (SPL_{0-2}–SPL_{2-5})–(SPL_{2-5}–SPL_{5-8}); "hyper-hypofunctional" was correlated with SPL_{2-5}; "coarse" was correlated with SPL_{0-2}–SPL_{2-5}; and "head" was correlated with SPL_{0-2}–SPL_{5-8}.

Harold, E. R. (1999). *XML Bible*. Hungry Minds, Inc. http://www.ibiblio.org/xml/books/bible

Textbook. A good introduction and reference to XML. In particular, a chapter on XSL transformations is freely accessible via the web.

Henton, C. and Edelman, B. (1996). Generating and manipulating emotional synthetic speech on a personal computer. *Multimedia Tools and Applications*, 3:105–125.

System description. An easy-to-use visual interface to prosodic parameter control in the MacinTalk speech synthesiser is described. Volume and duration of a word are represented by character height and width respectively, while emotion is represented by font colour. A list of prosodic parameter configurations for each of 10 emotions is presented, based on informal listening and parameter adaptation. No perception tests are undertaken because there "is tacit acknowledgement that the perception of synthesized emotions is not necessarily predictable and may not yet be a precise science" (p. 108).

Heuft, B., Portele, T., and Rauth, M. (1996). Emotions in time domain synthesis. In *Proceedings of the 4th International Conference of Spoken Language Processing*, Philadelphia, USA.

Experiment. Recorded emotionally coloured read paragraphs without the speaker knowing the goal of the recording ⇒ naturally sounding, unexaggerated stimuli. Synthesised with F0, duration and energy ⇒ only fear and neutral recognised. Tried to correlate perception test results with prosodic parameters.

Hoffmann, R., Kordon, U., Kürbis, S., Ketzmerick, B., and Fellbaum, K. (1999). An interactive course on speech synthesis. In *Proceedings of the ESCA/SOCRATES Workshop MATISSE*, pages 61–64.

System description. Web-based TTS introduction/overview:

- short description of what each module does;
- more detailed section on signal segmentation for building a diphone database;
- stepwise processing of text input in the TTS (but no manipulation of intermediate results).

Holmberg, E. B., Hillman, R. E., and Perkell, J. S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voices. *Journal of the Acoustic Society of America*, 84(2):511–529.

Experiment. Soft voice ⇒ greater open quotient; loud voice ⇒ smaller open quotient.

Holzapfel, M. and Campbell, N. (1998). A nonlinear unit selection strategy for concatenative speech synthesis based on syllable level features. In *Proceedings of the 5th International Conference of Spoken Language Processing*, Sydney, Australia.

Algorithm description (ATR CHATR / Siemens Papageno). Acoustic clustering (quasi CART). Syllable-based, phonetics-type features. Suitability function (bell-shaped) around the target used in distance measure (close values are suitable, distant values are unsuitable). Suitabilities are multiplied (logical AND).

Huber, R., Nöth, E., Batliner, A., Buckow, J., Warnke, V., and Niemann, H. (1998). You BEEP machine – emotion in automatic speech understanding systems. In *Proceedings of the Workshop on Text, Speech and Dialog*, Brno, Czech Republic.

Experiment. First experiments with recognition of emotions from speech, distinguishing angry and neutral speech using acted material. In a brute force approach, 276 acoustic prosodic features are used for training multi-layer perceptrons. Correct identification of angry vs. neutral speech is of the order of 90%.

Huttar, G. L. (1968). Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech and Hearing Research*, 11:481–487.

Experiment. Spontaneous speech of one speaker was presented to listeners and rated by means of semantic differential scales measuring "degree of emotion" (the activation dimension) and some individual emotions. These ratings were correlated with acoustic measurements of prosodic variables and with listener ratings of prosodic variables. Significant correlations were found, in the sense that more active emotions correspond to a higher maximum F0, higher F0 range, higher maximum intensity, and to a faster perceived speech rate.

Hyman, S. E. (1998). A new image for fear and emotion. *Nature*, 393:417–418.

Review. Amygdala is critical in emotional learning. Explanation of what is (Pavlovian) "conditioning". Neuro-imaging is still too slow (seconds–minutes) for subtle effects in the nervous system.

Igarashi, T. and Hughes, J. H. (2001). Voice as sound: Using non-verbal voice input for interactive control. In *Proceedings of the 14th Annual Symposium on User Interface Software and Technology (ACM UIST'01)*, Orlando, FL, USA.

System description. Voice as sound used for controlling computer applications on a low level (continuity, pitch, discrete sounds). Unnatural use of the voice, but robust.

Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M. (2000). A speech synthesis system with emotion for assisting communication. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 167–172, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

System description. "CHATAKO", based on CHATR. Uses three corpora of emotional speech per speaker: anger, joy, sadness. Speech material: $>$ 400 sentences of read text. Mean recognition rates on 5 semantically neutral sentences: female: joy 51%, ang. 60%, sad 82%; male: joy 52%, ang. 51%, sad 74% (chance level 33%). Good intelligibility and subjective evaluation.

Iriondo, I., Guaus, R., Rogríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., and Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 161–166, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Emotional speech actor corpus (2 texts, 8 actors, 7 emotions, 3 intensities). Perceptual test for pre-selection (disgust discarded due to low recognition). Acoustic analysis of good examples ⇒ rules for emotion expression, reported in detail. Implemented rules with TTS output. Perceptual evaluation of synthesised utterances yet to be done.

Johnson, W. L., Narayanan, S. S., Whitney, R., Das, R., Bulut, M., and LaBore, C. (2002). Limited domain synthesis of expressive military speech for animated characters. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA.

System description. Experiments with limited domain synthesis for emotion and speaking style in a military training application (Mission Rehearsal Exercise). Styles were: Shouted commands, shouted conversation, spoken commands, and spoken conversation. In a perception test, natural and synthesised utterances were mixed. The perceived naturalness of the intonation of the synthesised utterances approached that of natural utterances, more so for commands than for conversational speech. Perceived emotion was not reported in detail.

Johnstone, T., Banse, R., and Scherer, K. R. (1995). Acoustic profiles in prototypical vocal expressions of emotion. In *Proceedings of the 13th International Conference of Phonetic Sciences*, volume 4, pages 2–5, Stockholm, Sweden.

Supplementary to Banse & Scherer (1996). Calculated separate profiles for well vs. poorly recognised portrayals of each emotion. Similarity of these profiles is interpreted as existence of well-defined acoustic profiles. Well-defined profile, well-recognised: hot anger, boredom. Actors might have used the sadness profile to express shame. Limited recognition of actor portrayals might be due to the inability of actors to control voluntarily certain involuntary physiological changes. Interest might be conveyed by F0 contour.

Johnstone, T. and Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2029–2032, San Francisco, USA.

Experiment. Induced emotions using a computer game. Measured physiological parameters and acoustic signal. Imagination technique and EGG measures. Emotion-typical long-term average voice parameters. Planning glottal analysis with the LF-model.

JSML (1999). Java speech markup language 0.6. Technical report, Sun Microsystems. http://java.sun.com/products/java-media/speech/forDevelopers/JSML

Reference for Sun's JSML markup language.

Kainz, F. (1940). *Psychologie der Sprache*, volume 1. Ferdinand Enke Verlag, Stuttgart, Germany. 2nd edition 1954.

Theory. Functions of language: Information, appeal, representation (= dialogic functions); expression → role of interjections and "nature sounds". Idea of a development from nature sounds towards language through affect control during the development of civilisation.

Kasuya, H., Maekawa, K., and Kiritani, S. (1999). Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2505–2512, San Francisco, USA.

Model, method, synthesis. Novel method (similar to LPC) for voice source *and* vocal tract estimation. Needs user to label voiced segment first. So far only applied to a short slowly-spoken utterance: [e::ki::] with three emotions: neutral, suspicious, and disappointed.

Kehrein, R. (2002). The prosody of authentic emotions. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. http://www.lpl.univ-aix.fr/sp2002/papers.htm

Experiment. Elicited spontaneous emotion using a variant of a Maptask: Two subjects were to collaboratively build a Lego item, communicating only via their voices; emotionally significant factors were introduced by the experimenter in a controlled way. Emotionally expressive episodes were identified by three raters, and assigned a category label. Auditory and acoustic analyses suggested common prosodic properties in different emotion categories; a tentative explanation of this phenomenon was given in terms of the emotion dimensions activation, dominance and valence as well as an "unexpected" quality.

Kienast, M., Paeschke, A., and Sendlmeier, W. (1999). Articulatory reduction in emotional speech. In *Proceedings of Eurospeech 1999*, pages 117–120, Budapest, Hungary.

Experiment. In 4 sentences produces by 3 speakers with 4 emotions + neutral, measured articulatory segment reduction following Kohler. Results: Anger: high accuracy of articulation, consonants short, vowels long; Fear: Articulatory reduction, fast speech, short segments, aspirated stops; Sadness: Articulatory reduction, slow speech, long consonants, pauses; Happiness: Long voiced fricatives and phrase-stressed vowels.

Kitayama, S. and Markus, H. R. (1994). Introduction to cultural psychology and emotion research. In Kitayama, S. and Markus, H. R., editors, *Emotion and Culture*, pages 1–19. United Book Press, Baltimore, MD, USA.

Introduction. Emotion as fully culture-shaped: Emotion = "socially shared scripts composed of physiological, subjective, and behavioral processes" ("precomputed" scripts). Even physiological processes are influenced by culture. "Family resemblance" between the "same" emotions in different cultures, i.e. common components. "Emotionalization" as one interpretive schema locating internal sensations within social etc. interaction (vs. "somatization").

Klabbers, E., Stöber, K., Veldhuis, R., Wagner, P., and Breuer, S. (2001). Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proceedings of Eurospeech 2001*, pages 521–524, Aalborg, Denmark.

System description. Overview of the basic ideas in the BOSS system.

Klatt, D. H. (1979). Synthesis by rule of segmental durations in English sentences. In Lindblom, B. and Öhman, S., editors, *Frontiers of Speech Communication*, pages 287–299. Academic, New York.

Algorithm. Rules for synthesis. Klatt duration rules (the rules themselves are literally reproduced in Allen et al. (1987)). More info than Allen et al. (1987) about input format (and some phonological and F0 contour rules).

Klein, J. and Picard, R. W. (1997). Support for human emotional needs in Human-Computer Interaction. In *Proceedings of the CHI'97 Workshop on Human Needs and Social Responsibility*. http://www.media.mit.edu/affect/AC_research/projects/needs_paper.ps

Humans have two basic kinds of emotional needs: Emotional skill needs (emotional self-awareness, affect perception and management, ...) and experiential needs (feel understood, accepted, attention, security, ...). Computers could help fulfill these needs: Affect tutor for autistic children; "Active Listener"; "pet" (tamagochi); imaginary friend for a shy child (= help learning "normal" emotional interaction).

Krenn, B., Pirker, H., Grice, M., Piwek, P., van Deemter, K., Schröder, M., Klesen, M., and Gstrein, E. (2002). Generation of multimodal dialogue for net environments. In *Proceedings of Konvens*, Saarbrücken, Germany. http://www.ai.univie.ac.at/NECA

System description. Description of the NECA architecture and RRL.

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. *Journal of the Acoustic Society of America*, 78(2):435–444.

Experiment. Continuation of Scherer et al. (1984). Resynthesis of natural spoken utterances → perception tests. No interaction effects between contour type, F0 range, and voice quality. Change of F0 range is perceived in a continuous way, not as categorical changes. Results for contour and voice quality are less conclusive.

Lang, P. J. (1994). The motivational organization of emotion: Affect-reflex connections. In van Goozen, S. H. M., van de Poll, N. E., and Sergeant, J. A., editors, *Emotions: Essays on emotion theory*, pages 61–93. Lawrence Erlbaum, Hillsdale, NJ.

Theory. Emotions = action dispositions ⇒ autonomous / behavioral effects. Can be cued, in man, by mental representations. Even verbally cued, associative processes leading to emotion are automatic and not controlled; they may defy intentionality. Emotion networks in the brain differ from other knowledge structures in that they include direct connections to the primary motivational system. Two dimensions of emotion: Valence, arousal. Arousal-related stimuli are better recalled. Startle reflex: Organism in aversive state ⇒ strong reaction to negative stimuli. Organism in appetitive state ⇒ weak reaction to negative stimuli and vice versa.

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press, New York.

Overview. Presents Lazarus' cognitive-motivational-relational theory. Many links to the literature. Contains a criticism of dimensional descriptions which he considers too reduced to be meaningful.

Lazarus, R. S. (1999a). The cognition-emotion debate: a bit of history. In Dalgleish, T. and Power, M. J., editors, *Handbook of Cognition & Emotion*, pages 3–19. John Wiley, New York.

According to Cowie & Cornelius (2003), contains a mention of the idea of underlying emotions.

Lazarus, R. S. (1999b). *Stress and Emotion: A new synthesis*. Springer, New York.

According to Cowie & Cornelius (2003), contains a list of basic emotions.

Lehiste, I. and Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, 31:428–435.

Experiment. Investigated relationship between vocal speaker effort, acoustic signal amplitude, and perceived loudness. Vowels produced with the same vocal effort are perceived as having the same loudness, but they differ in amplitude (roughly: [i], [u] 6 dB below [a]). If vowels are produced such that the pressure levels are equal, some require more vocal effort. These are almost invariably identified as louder by listeners. In addition, they were described as sounding 'tense' or 'harsh'.

Leinonen, L., Hiltunen, T., Linnankoski, I., and Laakso, M.-L. (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustic Society of America*, 102(3):1853–1863.

Experiment. 10 emotions expressed in a one-word utterance are well recognised. Actor portrayals from frame stories; preselected; presented to 73 judges; acoustic analysis including F0 contours and sound pressure curves, and spectral analysis with a self-organising Kohonen map. Searched for acoustic properties of the different emotion expressions. Thorough statistical analyses, clear presentation of results, many references to others' results.

Lewis, M. and Haviland, J. M., editors (1993). *Handbook of Emotions*. Guilford Press, New York.

A collection of 44 acticles on emotion, grouped into five sections: Interdisciplinary foundations; Biological and neurophysiological approaches to emotion; Basic psychological processes in emotion; Social processes related to emotion; and more specific discussions of Selected emotions.

Love, S., Foster, J., and Jack, M. (2000). Health warning: Use of speech synthesis can cause personality changes. In *Proceedings of the IEE Electronics and Communications Seminar on the State of the Art in Speech Synthesis*, London.

Experiment. Difference in perceived personality between an actor's voice and the synthesis voice created using that actor's voice, using five-factor model rating scales. On all five scales, the synthetic version performed worse. This effect was amplified when the synthesised version was played first, attenuated when the actor was played first.

Macon, M. W., Cronk, A. E., and Wouters, J. (1998). Generalization and discrimination in tree-structured unit selection. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 225–230, Jenolan Caves, Australia.

Overview. Good overview of unit selection algorithms employed, very clear. Perceptually optimal distance measures (see Wouters & Macon (1998)). Stopping criteria in CART building (see Cronk & Macon (1998)).

Marcel, A. J. (1983a). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology*, 15:238–300.

Theory. Proposes a model of perceptual processes and consciousness taking into account the results of Marcel (1983b): The representations produced by perceptual processes are not identical to the phenomenal experience.

Marcel, A. J. (1983b). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15:197–237.

Experiment. Pattern masking of written words allows for graphical and semantical similarity judgements without word detection ⇒ perceptual analysis without consciousness

(Orientation of response by partial activation of graphical / semantic circuits). Stroop effect: word influence even if unaware. Subliminal perception is possible with pattern masking, but not with energy masking. Identity of perception analysis output and conscious representation is untenable.

Markel, N. (1998). *Semiotic Psychology*. Peter Lang, New York.

Review of work from the 1950s. Speech as an index of emotions and attitudes. The goal is to account for internal states (attitudes) during social interaction. The method is a detailed analysis of speech; methods and concepts of nowadays-lost research from the 1950s are presented. Seems to work with *written transcripts* of spontaneous speech. No acoustic analysis.

Markus, H. R. and Kitayama, S. (1994). The cultural construction of self and emotion: Implications for social behavior. In Kitayama, S. and Markus, H. R., editors, *Emotion and Culture*, pages 89–130. United Book Press, Baltimore, MD, USA.

Social psychology. The cultural framework forms the understanding of the self ⇒ learning emotional behaviour, what feels "good" etc. Distinction: *independent* (American) cultural frame (⇒ fulfilling own needs creates "good" emotions) vs. *interdependent* (Asian) cultural frame (⇒ group conformity feels "good"). Importance to see the cultural frame / self to understand emotional reactions.

Marumoto, T. and Campbell, N. (2000). Control of speaking types for emotion in a speech re-sequencing system. In *Proceedings of the Acoustic Society of Japan, Spring meeting 2000*, pages 213–214. In Japanese.

System: Algorithm and evaluation. First attempt for unit selection based synthesis of emotions using emotion-specific selection criteria. HMMs for voice source and prosody were trained; in addition, simple prosody rules were formulated. 3 emotions: anger, joy, sadness ⇒ only anger and sadness recognised better than chance. Synthesis methods using different selection criteria were compared. Method without HMMs, only using prosody rules, performed best ⇒ voice quality parameters do not capture the perceptually relevant emotion-specific cues yet. English language version see Campbell & Marumoto (2000).

Massaro, D. W. (2000). Multimodal emotion perception: Analogous to speech processes. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 114–121, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Model. Application of his Fuzzy Logic Model of Perception (FLMP) to bimodal emotion perception. Idea of the FLMP:

1. Evaluation. (degree to which a mode supports an alternative – e.g. happy), happens independently for each mode.
2. Integration. Calculate the overall degree of support for a given alternative by *multiplying* the values from 1.
3. Decision. Normalise the values from 2. by dividing through the sum of all degrees of support for all alternatives.

Massaro, D. W. and Cohen, M. M. (2000). Fuzzy logical model of bimodal emotion perception: Comments on "The perception of emotions by ear and by eye" by de Gelder & Vroomen. *Cognition & Emotion*, 14(3):313–320.

Model test. The data of de Gelder & Vroomen (2000) was analysed using the fuzzy logical model of perception (FLMP). After being fit to the average results of de Gelder and Vroomen's perception test, the FLMP made good predictions of subjects' perceptual performances. The smaller influence of the modality which subjects had been told to ignore was reflected in FLMP parameter values assigning less weight to that modality. Conclusion: The same basic integration mechanism, as presumed by the FLMP, can account for the observed perception patterns.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60:175–215.

Model. Model of personality: 5 factors/dimensions determine/compose a person's personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness to experience. Empirically supported both by personality questionnaires and lexical synonym clustering. A factor represents a group of traits that covary, but which are not necessarily interchangeable.

McGilloway, S., Cowie, R., and Douglas-Cowie, E. (1995). Prosodic signs of emotion in speech: Preliminary results from a new technique for automatic statistical analysis. In *Proceedings of the 13th International Conference of Phonetic Sciences*, volume 1, pages 250–253, Stockholm, Sweden.

Experiment. ASSESS acoustic analysis system applied to emotional speech. Read emotional speech (4 emotions + neutral). Passages with emotional content constructed by the authors, read by 40 speakers.

McMahon, E. and Cowie, R. (2001). Describing emotion in music: Validation of the Feeltrace device. In *Proceedings of the 12th Irish Conference on Artificial Intelligence and Cognitive Science*, pages 285–295, Maynooth, Co. Kildare, Ireland.

Experiment. Rated pieces of music with Feeltrace. Ratings of pieces of music differed most consistently according to emotional orientation (angular measure in polar coordinates), less according to intensity.

Mehrabian, A. and Russell, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA, USA; London, UK.

Theory. Informative review of evidence from different domains (intermodality responses, synesthesia (=stimulation in one sense affects perception in another), physiological reactions, semantic differential) supporting the idea that there are three emotion dimensions: Pleasure, Arousal, and Dominance.

Meyer, W.-U., Schützwohl, A., and Reisenzein, R. (1997). *Einfürung in die Emotionspsychologie. Band II: Evolutionspsychologische Emotionstheorien*, chapter 5: Kritik der Grundannahmen der Basisemotionstheorien. Verlag Hans Huber, Bern, Göttingen, Toronto, Seattle.

Review. Critical discussion of all aspects of basic emotion theories: Facial expression; peripheral physiological changes; action tendencies; feeling experience; cognitive appraisals. Criticise the idea that secondary emotions are composed of primary emotions.

Microsoft (2002). *SAPI 5: Microsoft Speech API 5.1*. http://www.microsoft.com/speech

XML-based application interface. Functionality more or less unchanged from SAPI 4 with respect to prosody control: only one emphasis level, pitch level and rate can be modified (= a subset of SABLE functionality).

Mixdorff, H. and Fujisaki, H. (1999). The influence of focal condition, sentence mode and phrase boundary location on syllable duration and the F0 contour in German. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 1537–1540, San Francisco, USA.

Experiment. Fujisaki model for German. Narrow focus brings an F0 peak and a syllable lengthening. Presenting duration "contours" as z-scores: centered and reduced syllable durations, giving the deviation from the mean duration of this syllable over all utterances in s.d. units ⇒ graphically presentable.

Mizuno, H., Abe, M., and Nakajima, S. (1999). Development of speech design tool "Sesign99" to enhance synthesized speech. In *Proceedings of Eurospeech 1999*, Budapest, Hungary.

System description. GUI allows some manipulations of synthesised speech, e.g. prosody transplantation.

Möbius, B. (1995). Components of a quantitative model of German intonation. In *Proceedings of the 13th International Conference of Phonetic Sciences*, volume 2, pages 108–115, Stockholm, Sweden.

Model. Fujisaki's model, adapted for German. Discussion of tone sequence vs. contour superposition approaches. Fujisaki's model: Analysis-by-synthesis (= adapting the model's parameters for optimal approximation); phrase component + accent component. Möbius interprets Fujisaki's model components in linguistic terms, introducing new linguistically motivated constraints in analysis.

Möbius, B. (1999). The Bell Labs German text-to-speech system. *Computer Speech and Language*, 13:319–357.

System description. Thorough description of the German version of the Bell Labs TTS system. Extensive use of weighted finite state transducers (WFST) for all labelling and transformation tasks. Many linguistic explanations for design decisions. Text analysis component: Stem and affix lexicon, inflection classes for nouns, adjectives and verbs; but no part-of-speech tagging or syntactic analysis. Duration model: van Santen's sums-of-products model. Intonation model: a variant of Fujisaki's contour superposition model. Concatenation units: mostly diphones.

Möbius, B., Sproat, R., van Santen, J., and Olive, J. (1997). The Bell Labs German text-to-speech system: An overview. In *Proceedings of Eurospeech 1997*, volume 5, pages 2443–2446, Rhodes/Athens, Greece.

System description. An overview of the TTS approach implemented in the Bell Labs system for German.

Mokhtari, P. and Campbell, N. (2002). Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 2015–2018, Las Palmas de Gran Canaria, Spain.

Algorithm. Towards the automatic determination of voice quality features in very large speech corpora. An array of analyses was combined to determine reliable locations for formant estimation. First, quasi-syllabic nuclei are identified on the basis of high energy in the spectrum below 3000 Hz. Second, delta cepstra are used for measuring the steadiness of the spectrum, and formant calculation reliability is determind by first estimating formants and then using them to calculate a cepstrum which is compared to the actually measured cepstrum. If all three measures (sonority, spectral steadiness, reliable formant detection) coincide in a given location, this location is retained as a centre of reliability, where subsequent voice quality and articulation parameter estimations can be calculated.

Montero, J. M., Gutiérrez-Ariola, J., Palazuelos, S., Enríquez, E., Aguilera, S., and Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pages 923–926, Sydney, Australia.

Experiment. In the framework of the VAESS project. Corpus (neutral, sad, happy, anger, surprise) with 1 actor and much text (3 passages, 15 short sentences, 30 words). Perception test with natural recordings; analysed phoneme duration, F0 contour, pauses. Formant synthesis ⇒ very good recognition after an accustomisation phase. Concatenative resynthesis mixing voice quality x prosody ⇒ voice quality was crucial only for anger.

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., and Pardo, J. M. (1999a). Analysis and modelling of emotional speech in Spanish. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 957–960, San Francisco, USA.

Experiment. Mix diphones / prosody copy synthesis ⇒ anger and happiness are recognised on the basis of diphones (voice quality), surprise and sadness on the basis of prosody.

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Macías, J., Enríquez, E., and Pardo, J. M. (1999b). Development of an emotional speech synthesiser in Spanish. In *Proceedings of Eurospeech 1999*, Budapest, Hungary.

Short system description. In addition to the copy synthesis experiments reported in Montero et al. (1999a), automatic prosody rules are implemented in a TTS system (but not reported in detail). First experiments with unit selection are mentioned, using low-level prosody-based target costs and cepstrum-based concatenation costs. A mention is made of smiled voice quality for happy speech.

Morlec, Y., Aubergé, V., and Bailly, G. (1995). Evaluation of automatic generation of prosody with a superposition model. In *Proceedings of the 13th International Conference of Phonetic Sciences*, volume 4, pages 224–227, Stockholm, Sweden.

Model. General idea of the superposition model of intonation. Two automatic learning procedures, structured lexicon of contours and sequential neural networks ⇒ comparable results in perception test (complete "real" sentences like "Rappelez monsieur Dupont jeudi.").

Morlec, Y., Bailly, G., and Aubergé, V. (1996). Un modèle connexioniste modulaire pour l'apprentissage des " gestes " intonatifs. In *Proceedings of the XXIes Journées d'Etude sur la Parole*, pages 207–210.

Model. Model sentence-level intonation by recurrent neural networks. Perception with expectations: Deviation from an expected form (="normal" intonation). Contrasts exist at every level of linguistic description. The contribution of the different levels is weighted; it varies with attitude. The prosodic "movement" expands with more syllables; prototypical final phase and start phase. Additional modulations beyond 5–6 syllables.

Morlec, Y., Bailly, G., and Aubergé, V. (1997). Apprentissage automatique d'un module de génération multistyle de l'intonation. In *Proceedings of the 1es JST Francil*, pages 407–412, Avignon.

Model. Builds upon Morlec et al. (1996). Rhythm: mean and s.d. of phonemes; GIPC as rhythmic unit.

Moroni, V. (1997). Enquête sur les attitudes du français: Définition et interprétation. T.E.R. de maîtrise, Sciences du Langage, Université Stendhal – Grenoble 3.

Resource. List of words for emotions/attitudes in the voice. List of "little words" accompanying emotion production.

Morris, J. S., Öhman, A., and Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, 393:467–470.

Experiment. Emotional visual stimuli (angry faces) can be masked (not consciously perceived) when shown for 30 ms only (and followed by another face). Conditioned (noise-associated) *masked* stimuli elicit *right* amygdala reaction. Conditioned *unmasked* stimuli elicit *left* amygdala rection. Conditioned vs. unconditioned: different reaction ⇒ emotional learning. Difference left/right: Conscience / verbal activity seems to inhibit right and to enhance left amygdala reaction.

Mozziconacci, S. J. L. (1998). *Speech Variability and Emotion: Production and Perception*. PhD thesis, Technical University Eindhoven.

Experiment. The link between prosody and emotions is studied from production and perception perspectives. Seven emotion categories are studied. For each emotion, perceptually optimal global settings are identified for pitch level, pitch range and speech rate. The probability of particular intonation contours to co-occur with particular emotions is identified in production and perception tests. The intonation contours are described using the IPO Grammar of Dutch intonation.

Mozziconacci, S. J. L. (2000). The expression of emotion considered in the framework of an intonation model. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 45–52, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Theory. Embeds study of emotion and intonation in a theoretical frame (what are the functions of intonation? what types of intonation models exist? value of a model). Covariance vs. configuration. View towards speech synthesis.

Mozziconacci, S. J. L. and Hermes, D. J. (1998). Pertinence perceptive des configurations intonatives en parole émotionelle. In *Proceedings of the XXIIes Journées d'Etudes sur la Parole*, pages 163–166, Martigny, Switzerland.

Experiment. 7 emotions (315 utterances) labelled with the IPO system of intonation labels. Two neutral sentences resynthesised with the abstracted intonation patterns (no duration modelling). Perception test ⇒ did not present recognition rates, but perceptive similarity of the intonation patterns as well as evaluation *tendencies* ("this pattern often leads to emotion X or Y; for Z this pattern should be avoided").

Mozziconacci, S. J. L. and Hermes, D. J. (1999). Role of intonation patterns in conveying emotion in speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2001–2004, San Francisco, USA.

Experiment. Modified only intonation, not duration or voice quality. Intonational categories (IPO system). Optimal mean F0 for 7 emotions. Over- and under-represented intonation patterns in perceptual emotion judgments. "Standard" pattern "1&A" (= early rise, late fall on one syllable).

Murray, I. R. (1989). *Simulating emotion in synthetic speech*. PhD thesis, University of Dundee, UK.

System description. The HAMLET system generating emotional synthetic speech.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93(2):1097–1108.

Review. Difficult definition of emotional qualities. "Hardwired" basic = primary = innate vs. secondary = learned emotions (but the limit varies substantially with authors). Pakosz's variations (speakers, judges). 3 dimensions: Activity, Valence, Strength. Emotion caracteristics: many findings on the vocal expression of the different emotions.

Murray, I. R. and Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16:369–390.

System description, experimental evaluation. Implementation of emotion rules from the literature (with heuristic improvements). Fully automatic TTS with emotion simulated entirely by rule, "plugged into" DECTalk. Interesting evaluation methodology: neutral (emotionally undetermined) and emotional texts, spoken with neutral and emotional prosody; measure the difference in identification rate caused by emotional prosody.

Murray, I. R. and Arnott, J. L. (1996). Synthesizing emotions in speech: Is it time to get excited? In *Proceedings of the 4th International Conference of Spoken Language Processing*, pages 1816–1819, Philadelphia, USA.

System description. Short summary of the HAMLET system and the problems inherent in the domain. Mentions that newer versions of HAMLET have incorporated a three-dimensional emotion model for synthesising shades of emotions.

Murray, I. R., Arnott, J. L., and Rohwer, E. A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, 20:85–91.

Review. Very general summary of the state of the art of research in emotional expression in speech and synthesis of emotional speech. Need for evaluation of "naturalness".

Murray, I. R., Edgington, M. D., Campion, D., and Lynn, J. (2000). Rule-based emotion synthesis using concatenated speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 173–177, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Prototype. Emotion synthesis using concatenative synthesis. Four emotions were modelled: anger, happiness, sadness, and fear. The LAERTES system sets global pitch and tempo by rule. Manual post-processing was performed to add intensity / freq. filter / vibrato / tremolo; .pho files (MBROLA input format) were edited. A pilot evaluation test is reported (recognition "as expected" / quality rating). LAERTES performed as good as the more sophisticated, formant-synthesis based HAMLET system. Hand-edited files were clearly better recognised and obtained slightly better quality ratings.

Ní Chasaide, A. and Gobl, C. (1997). Voice source varation. In Hardcastle, W. J. and Laver, J., editors, *Handbook of Phonetic Sciences*, pages 427–461. Blackwell, Oxford, UK.

Overview / teaching. Very good introduction, precise, good overview. Inverse filtering. Glottal airflow vs. differentiated glottal airflow. Voice source models. Measuring with "source model matching". Parameters describing the source signal: dynamic leakage RA = weakness of the (closure) excitation; glottal frequency RG = 1/the length of the opening phase relative to the total period; glottal symmetry / skew RK; open quotient OQ; Aspiration noise AH.

Nöth, E., Batliner, A., Kießling, A., Kompe, R., and Niemann, H. (1997). Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. In Paulus and Wahl, editors, *Mustererkennung 1997*, Informatik aktuell, pages 37–52. Springer, Heidelberg.

System description. Prosody in the automatic speech recognition in Verbmobil. Classical neural net, continuous speech recognition augmented with prosodic property vectors → drastically reduces parsing effort (syntax). Particularly useful for finding sentence boundaries. Also: semantical disambiguation, dialogue act splits. Mode for translation ("transfer"): Adaptation of speech synthesis to the speaker to be translated.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In Hinton, J., Nichols, J., and Ohala, J. J., editors, *Sound Symbolism*, pages 325–347. Cambridge University Press, Cambridge, UK.

Theory. Description of the frequency code. Low F0 and rough voice = strong, large, self-confident. High F0 and melodious voice: small, non-threat, maybe infant-mimicry. Presents diverse pieces of evidence from ethology and cross-language studies.

Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice. In *Proceedings of the 4th International Conference of Spoken Language Processing*, Philadelphia, USA.

Theoretical/philosophical deduction from ethology. 3 types of emotion/attitude expression:

- influencing the receiver, make-believe (beneficial for survival) e.g. smile;
- expression of inner state (not beneficial, "leak out") e.g. fearful trembling;
- attitudes (acquired, learned).

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotion*. Cambridge University Press, Cambridge, UK.

Model. A detailed model of the appraisal components involved in emotions. Emotions are seen as valenced reactions to three types of stimuli: Events, agents, and objects. Central to appraising events is their desirability, with respect to goals; central to appraising agents is the praiseworthiness of their actions, with reference to standards; and central to appraising objects is their appealingness determined by attitudes (i.e., liking/disliking). The model is formulated in a way permitting its implementation in AI systems.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97(3):315–331.

Theory. Discussion on emotion theory. Criticise conceptions of basic emotions, on various grounds. Propose an alternative, component-based account in which the components out of which the emotions are built are basic, instead of the emotions themselves.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press, Urbana, USA.

Measurement system. The book proposes the *semantic differential* technique to measure the most important aspects of meaning. Paired adjective scales (strong–weak, loud–soft, red–green, etc.) are used by subjects to characterise their understanding of all types of concepts. Subsequent factor analyses of these ratings lead to three most important factors, which the authors named *evaluation*, *potency* and *activity*. The technique, although introduced here as measuring the essentials of meaning for all types of concepts, is used later for determining the structure and dimensionality of "emotion space" (e.g., Mehrabian & Russell (1974)). Osgood et al. consider the factors or dimensions structuring meaning as a reduced account: "the representational state indexed by the semantic differential is not the only determinant operating in lexical encoding. It is a necessary but not a sufficient condition." (p. 323–324)

Oudeyer, P. (2002). The synthesis of cartoon emotional speech. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. http://www.lpl.univ-aix.fr/sp2002/papers.htm

System description, experiment. Using MBROLA, nonsense baby-like cartoon speech expressing emotions was generated. Five emotional states (calm, anger, sadness, comfort and happiness) were modelled, representing a neutral state and each of the four quadrants in arousal-valence space. Settings for the acoustic parameters were taken from the literature and then manually adapted to yield an exaggerated realisation appropriate for cartoon-style speech. Audio examples are available online.

Paeschke, A., Kienast, M., and Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 929–932, San Francisco, USA.

Experiment. Berlin emotional speech corpus. Measures of many aspects of F0 contours.

Pakosz, M. (1982). Intonation and attitude. *Lingua*, 56:153–178.

Theory. The link between intonation contour and emotional meaning is discussed. A "Relative Strength Hierarchy" hypothesis is presented, and supported by means of a number of examples. According to this hypothesis, intonation contour does not directly express a denotative emotional meaning, but only the abstract position on the activation dimension of emotional meaning. Only in combination with other semantically defined factors such as the verbal or situational context can this activation level be interpreted as a concrete emotion or attitude. Concrete suggestions are made as to the placement of intonation contours on the activation dimension (Figure 1 (a) and (b), p. 171).

Pakosz, M. (1983). Attitudinal judgements in intonation: Some evidence for a theory. *Journal of Psycholinguistic Research*, 12:311–326.

Review. Review of some earlier publications on vocal expression of emotion. Pakosz concludes that intonation only conveys information about the activation dimension, not about individual emotion categories. (It is not entirely clear what exactly Pakosz means with "intonation", whether this encompasses only the shape of the intonation contour or also global settings such as pitch level and pitch range).

Pelachaud, C., Badler, N., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, 20:1–46.

System description. Description of a system for generative facial animation determined by lip shape (incl. coarticulation), conversational signals, punctuators, regulators, and manipulators. Affect is modelled using Ekman's six basic emotions.

Pereira, C. (2000). Dimensions of emotional meaning in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 25–28, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Two actors produce two standard sentences in five emotions (happy, sad, hot/cold anger, neutral). Listeners place these utterances on three dimensions arousal, pleasure, power. Correlation with acoustic analysis: arousal $\Leftrightarrow$ higher F0, larger F0 range, louder; for man only: power $\Leftrightarrow$ higher F0, larger F0 range, louder.

Pereira, C. and Watson, C. (1998). Some acoustic characteristics of emotion. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pages 927–929, Sydney, Australia.

Experiment. Acoustic analysis of 5 emotions: hot anger, cold anger, sadness, happiness, neutrality. Measured F0 mean, F0 range, F0 variability, F0 contour using 4 points, mean energy, rate of articulation, using ESPS/Waves.

Petitpierre, D. and Russell, G. (1995). MMORPH – the Multext morphology program. Deliverable report, MULTEXT. ftp://issco-ftp.unige.ch/pub/multext/mmorph.doc.ps.tar.gz

Reference paper for MMORPH.

Piot, O. (1999). Experimental study of the expression of emotions and attitudes in four languages. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 369–370, San Francisco, USA.

Experiment. Interesting ideas in theoretical framework: frequency code, arousal, motivation. Very short though.

Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S., and Pirker, H. (2002). RRL: A Rich Representation Language for the description of agent behaviour in NECA. In *Proceedings of the AAMAS Workshop Conversational Agents*, Bologna, Italy.

System description. An overview of the design issues and capabilities of the NECA RRL, tied to the NECA system architecture.

Ploog, D. (1979). Phonation, emotion, cognition, with reference to the brain mechanisms involved. In *Brain and Mind, CIBA Foundation Symposium*, pages 79–98.

Experiment/Hypothesis. Brain centers involved in sound production. Hierarchical organisation of voicing:

- ponto-mesencephalic area $\rightarrow$ species-specific vocal gestures;
- anterior limbic cortex $\rightarrow$ control the readiness to react vocally (primates only);
- cortical larynx and facial area $\rightarrow$ learning of articulatory movements (humans only).

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Harper and Row, New York.

Theory. Overview of approaches to emotion. Plutchik's proposal of a circular structure of emotion, based on perceived similarity of emotion words and semantic differential ratings. A colour metaphor for emotions, including the idea that complex emotions are obtained by mixing primary emotions.

Plutchik, R. (1994). *The Psychology and Biology of Emotion*. HarperCollins College Publishers, New York.

Overview. An overview of emotion theories and descriptions. Among many other things, presents circumplex models for the description of emotion concepts.

Posner, M. I. (1992). Attention as a cognitive and neural system. *Current Directions in Psychological Science*, 1:11–14.

Overview. Two attention networks in the brain: Posterior (parietal lobe; also the pulvinar (thalamus) and the superior colliculus (midbrain)) and anterior (anterior cingulate gyrus). Posterior is implicated in visual word form recognition, anterior in semantic word analysis. Vigilance system (= long-term attention) in right frontal lobe.

Posner, M. I., Rothbart, M. K., and Harman, C. (1994). Cognitive science's contributions to culture and emotion. In Kitayama, S. and Markus, H. R., editors, *Emotion and Culture*, pages 197–216. United Book Press, Baltimore, MD, USA.

Overview. Semantic networks: activation without attention. Emotions might serve to coordinate different cognitive processors. Analogy color perception vs. emotion perception: Cultural influence. Attention serves to amplify particular areas of the perception system. Interaction attention ↔ emotion: possibly superior colliculus and anterior cingulate gyrus. Distracting infants (attention focussed on something else) is used in Western culture to calm them.

Provine, R. R. (1996). Laughter. *American Scientist*, 84:38–45.

Overview. Laughter: short vowels of 75 ms, repeated at 210 ms intervals; m = 276 Hz, f = 502 Hz; decrescendo. Chimpanzee laughter: 1 laughter = 1 inspiration or 1 expiration. Laughter has a social function (only 20% due to jokes; mutual playfulness, in-group feeling, positive emotion). Laughter rarely interrupts a sentence. Speaker laughs more than audience; women more than men.

Rank, E. (1999). Erzeugung emotional gefärbter Sprache mit dem VieCtoS-Synthesizer. Technical Report 99-01, ÖfAI. http://www.ai.univie.ac.at/cgi-bin/tr-online?number+99-01

System prototype. More detailed version of Rank & Pirker (1998). More detailed description of the manipulated parameters.

Rank, E. and Pirker, H. (1998). Generating emotional speech with a concatenative synthesizer. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 3, pages 671–674, Sydney, Australia.

Experiment. Using mainly voice quality parameters from the literature (Cahn, Klasmeyer). Synthesised (LPC, semi-syllable inventory) 5 sentences with 4 emotions (anger, sadness, fear, disgust). Results: Only sadness is well recognised (69%), anger 40%, fear and disgust chance level. Apparently, did not model F0.

Reinke, K. (2000). Ein Babylon der Emotionen? Das Problem der kultur- und sprachenübergreifenden Erforschung der phonetischen Emotionssignale. *Deutsch als Fremdsprache*, 37(2):62–72.

Overview. An overview of problematic aspects of listener-centered aspects of vocal emotion communication from a foreign language teacher's point of view. Question of universality vs. culture-specificity. Language-specific emotion concepts.

Roach, P. (2000). Techniques for the phonetic description of emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 53–59, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Resource description. Annotation system for the paralinguistic annotation of the speech files in the Emotion in Speech Project in Reading/Leeds. This complex annotation scheme has the goal to "allow exhaustive and unambiguous coding" (p. 53) of speech features potentially relevant for emotion. The multi-tier annotation contains a ToBI annotation of intonation, other prosodic features such as tempo and pitch range, and paralinguistic features such as voice quality.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., and Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, 28:83–94.

Transcription system. Frame: large natural emotional speech corpus, annotation of this corpus. Corpus will be linguistically annotated (words, ToBI, prosody) and psychologically annotated (by perception tests). In this article: Presentation of the prosodic/paralinguistic annotation tier.

Robson, J. and MackenzieBeck, J. (1999). Hearing smiles – perceptual, acoustic and production aspects of labial spreading. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 219–222, San Francisco, USA.

Experiment. Labial spreading on whole sentences is heard as more "smiling" than neutral facial expression in 90% of the cases. Also visual recordings ⇒ greater horizontal expansion for spreading. Acoustics: F2 and F3 raised. 3 sentences differing in amount of segments predicted by Laver to increase audibility of spreading ⇒ non-conclusive results.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Model. Emotion concepts are proposed to be organised according to a circular structure, a "circumplex", in a two-dimensional space of pleasure-displeasure and degree of arousal. Russell reviews the evidence from the literature before presenting his own experiments. The circumplex pattern was confirmed by several different methods for characterising emotion words: Russell used a categorisation and sorting method specifically designed for testing circularity, a grouping task providing similarity measures used for multidimensional scaling, and direct positioning on pleasure and arousal dimensions by means of semantic differential scales. All three methods yielded nearly identical circular patterns. From the domain of self-report of emotional state, similar distributions were obtained: Subjects reported their current emotional state on bipolar scales representing the emotion dimensions as well as on unipolar scales representing emotion adjectives. Regression analysis (prediction of the value on the adjective scale from the values on the emotion dimensions) as well as principal component analysis of the adjective ratings provided patterns very similar to the circular structure obtained in the concept characterisation tasks.

Russell, J. A. (1997). How shall an emotion be called? In Plutchik, R. and Conte, H., editors, *Circumplex Models of Personality and Emotion*, pages 205–220. APA, Washington.

Emotion theory. Six necessary properties of emotions. 1. An emotion belongs to a category (indeed to many categories). 2. Category membership is gradual. 3. Emotion categories are related to each other as described by a circumplex. 4. Emotions are located on continua (dimensions), e.g. intensity, pleasure and arousal. 5. Emotion categories correspond to a script, i.e. a prototypical sequence of causally connected and temporally ordered constituents. 6. Emotion categories are embedded in a fuzzy hierarchy.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294.

Experiment. Subjects rated the emotional states described by short situation descriptions, on the semantic differential scales proposed by Mehrabian & Russell (1974) for measuring the pleasure, arousal and dominance dimensions, as well as on 42 emotion adjective scales. For each of these adjective scales, nearly all of the "reliable variance"

was accounted for by the three dimensions, along with a response-style variable. The authors consider this "strong evidence for the sufficiency of these three dimensions" (p. 292) for the definition of emotions. In a second study, 151 terms (words or phrases) were rated using the semantic differential scales, leading to a position in the three-dimensional emotion space for each of the words (reported in a table).

Schaub, H. (1995). Die Rolle der Emotionen bei der Modellierung kognitiver Prozesse. In *Workshop Artificial Life*, Sankt Augustin. http://www.uni-bamberg.de/~ba2dp1/private/schaub

System description. A model for action organisation (for robots etc.). Analogy between configurations of the model and emotions ⇒ "a cognitively working system is 'automatically' emotional".

Scherer, K. R. (1974). Acoustic concomitants of emotion dimensions: Judging affect from synthesized tone sequences. In Weitz, S., editor, *Nonverbal Communication: Readings with Commentary*, pages 249–253. Oxford University Press, New York.

Experiment. Sine wave tone sequence resembling a short sentence was varied in pitch level & variance, amplitude level & variance, and tempo. Judgments on emotion dimensions pleasantness/evaluation, activity and potency. Acoustic parameters influenced affective judgments (some detail given).

Scherer, K. R. (1977). Affektlaute und vokale Embleme. In Posner, R. and Reinecke, H.-P., editors, *Zeichenprozesse*, pages 199–214. Athenaion, Wiesbaden.

Theory. Difficulty of classifying affect bursts (according to sender/receiver characteristics, degree of intentionality, functions in dialogue, phonetic constancy). Vocal emblems = decodable independently from context. Summary functions of affect bursts. Summary of the causes and properties of affect bursts.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8:467–487.

Methodology, Experiment. Presentation of an experimental methodology based on E. Brunswik's lens model. Idea: The inference process is subdivided into:

1.the speaker state ("criterion");

2. the objective parameters influenced by that state ("distal indicator cues");

3. the subjectively perceived correlates of these parameters ("proximal percepts");

4. the perceived speaker state ("attribution").

Each of these is measured independently (the first two as objectively as possible, the last two through listener judgments). Correlations between adjacent steps correspond to the effect one step has on the next. Applied this methodology to the study of personality inferences from voice quality. Clearest result: Extraversion is expressed through, and perceived from, vocal effort, nasality and dynamic range. Proposed a method for evaluating the appropriateness of a given lens model (choice of distal cues and proximal percepts) by means of "path analysis".

Scherer, K. R. (1980). The functions of nonverbal signs in conversation. In StClair, R. N. and Giles, H., editors, *The Social and Psychological Contexts of Language*, pages 225–244. Lawrence Erlbaum, Hillsdale, NJ, USA.

Theory. Functions of non-verbal signs:

- semantic: signification, amplification, contradiction, modification;
- syntactic: segmentation, synchronization;
- pragmatic: expression, reaction;
- dialogic: relation, regulation.

Scherer, K. R. (1982). Methods of research on vocal communication: paradigms and parameters. In Scherer, K. R. and Ekman, P., editors, *Handbook of Methods in Nonverbal Behavior Research*, pages 136–198. Cambridge University Press, Cambridge, UK.

Overview. Introduction to research methods in non-verbal vocal communication. Organised according to the Brunswikian lens model: Physiological and phonatory / articulatory level; acoustic level (including some detail regarding objectively measurable parameters); proximal cues / non-experts' perceptual categories; and inferences / attributions by the listener. Appendix contains an introduction to speech physiology and acoustics.

Scherer, K. R. (1984a). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality and Social Psychology*, 5:37–63.

Theory, Experiment. 1. Emotion dimensions and how they influenced the SECs proposed in Scherer's component process model. Evaluation ↔ intrinsic pleasantness and goal conduciveness check; Activation ↔ goal discrepancy and degree of control check; Potency ↔ power check. 2. Short overview over the component process model. 3. A study of natural language labels, clustering 235 non-synonymous German and English terms. Similarity judgements followed by cluster analysis yields 39 clusters. Multidimensional scaling ⇒ Two-dimensional plot, the axes of which can be interpreted as evaluation and activity (but other labels for the axes are also possible). 4. Short description of a large intercultural study of emotion antecedents and reactions.

Scherer, K. R. (1984b). On the nature and function of emotion: A component process approach. In Scherer, K. R. and Ekman, P., editors, *Approaches to emotion*, pages 293–317. Erlbaum, Hillsdale, NJ.

Theory. Reference paper for Scherer's component process model. Emotions are conceived of as consisting of several aspects or components: cognitive appraisal, physiological activation, motor expression, behavior intentions, and subjective feeling. Emotional states can be described as configurations of states of these components. The cognitive appraisal or information processing system is modelled in more detail, as a sequence of stimulus evaluation checks: novelty, intrinsic pleasantness, goal/need conduciveness, coping potential, and norm/self compatibility. "The assumption underlying this sequence ... is that each consecutive SEC further differentiates the emotional state of the organism." (p. 308). Scherer proposes an SEC-based description of emotions as an alternative to "palette theories" according to which mixed emotions consist of blends of primary emotions. In his model, mixed emotions are combinations of SEC outcomes typically associated with different "pure" emotions.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99:143–165.

Theory. The component process model applied to the prediction of vocal effects of emotion, based on a detailed description of the physiological effects of the individual stimulus evaluation checks. Includes a literature review on vocal emotion expression.

Scherer, K. R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, 7:79–100.

Theory. Most signals are multifunctional: symptom, appeal, symbol (Bühler). Primary interjections and push effects; secondary interjections (vocal emblems) and pull effects. Covariation for push (Scherer et al., 1984), configuration for pull (Ladd et al., 1985)? Symbolic function of affect vocalisation = allows a listener to reconstruct the major features of the emotion-producing event in its effect on the speaker.

Scherer, K. R. (1989a). Les émotions: fonctions et composants. In Rimé, B. and Scherer, K. R., editors, *Les émotions*, pages 87–133. Delachaux & Niestlé, Neuchâtel-Paris.

Theory. Functions of emotions (phylogenetic approach). Short overview about different emotion theories. The component process model (sub-systems, the modifications of which determine the affective state).

Scherer, K. R. (1989b). Vocal measurement of emotion. In Plutchik, R. and Kellerman, H., editors, *Emotion: Theory, Research, and Experience*, volume 4: The Measurement of Emotions, pages 233–259. Academic Press, New York.

Overview. Some remarks on phylogenetic history of vocal emotion expression. Mention of vocal correlates of the three emotion dimensions. Distinction between encoding and decoding studies. Presentation of his component process model and how it relates to the voice, summarising Scherer (1986), including a literature overview.

Scherer, K. R. (1994). Affect bursts. In van Goozen, S. H. M., van de Poll, N. E., and Sergeant, J. A., editors, *Emotions: Essays on emotion theory*, pages 161–193. Lawrence Erlbaum, Hillsdale, NJ.

Theory. Affect bursts: Spontaneous, non-linguistic, push-effects. High inter-individual variation, integration of facial and vocal expression triggered by a well-determined stimulus. Affect emblems: intentional, language-specific, symbolic representation of an emotion, pull-effects. Affect bursts/emblems are just the extreme poles of a continuum.

Scherer, K. R. (1996). Adding the affective dimension: A new look in speech analysis and synthesis. In *Proceedings of the 4th International Conference of Spoken Language Processing*, Philadelphia, USA.

Overview. Summary: State of the art 1995. Result summary: Acoustic features for anger, fear, sadness, joy. Theory: Three dimensions of similarity: Quality, intensity, valence. Top-down vs. bottom-up recognition to explain asymetry of confusion matrices. Report: Finding acoustic features relevant to human perception by comparison to statistical analysis. Result: Relative independence between different variables for emotion judgement in synthesis.

Scherer, K. R. (2000a). Emotion effects on voice and speech: Paradigms and approaches to evaluation. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Northern Ireland. No paper, slides from presentation. http://www.qub.ac.uk/en/isca/proceedings

Overview. Very insightful overview over the different aspects of the research domain. Structured on the basis of the Brunswikian lens model.

Scherer, K. R. (2000b). Psychological models of emotion. In Borod, J. C., editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press, New York.

Overview. Defines emotion as an intense episode of synchronised changes in component processes. Reviews major emotion models, and makes a proposal for unifying them by acknowledging the different foci. A central role in this unification process is seen in componential models. Short mention of an interesting new topic for research: “hot cognition”, i.e. “the way in which memory, learning, thinking, and judgment are affected by affective states.”

Scherer, K. R., Banse, R., and Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1):76–92.

Experiment. Universality of emotion recognition from the voice. Four German actors produced two nonsense-word sentences with four emotions (joy/happiness, sadness, fear, anger) and neutrally. 428 listeners from nine countries (Germany, Switzerland, Great Britain, Netherlands, United States, Italy, France, Spain, and Indonesia) took part in perception tests. Result: All emotions were recognised better than chance in all countries; Indonesia a bit less well than the others. No significant difference between the mean of the Western countries and the German reference. Best recognition in countries with Germanic languages.

Scherer, K. R., Johnstone, T., and Sangsue, J. (1998). L’état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole. In *Proceedings of the XXIIes Journées d’Etudes sur la Parole*, Martigny, Switzerland.

Good summary of “push” (social signal, configuration model) vs. “pull” (symptom, covariance model) effects. Speaker verification as an application. Summary of data from Banse & Scherer (1996) with french translation. Arguments for the need of emotions in speech technology.

Scherer, K. R., Ladd, D. R., and Silverman, K. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustic Society of America*, 76(5):1346–1356.

Experiment. “Linguistic” or “paralinguistic” vocal affect expression? *Covariance* of continuous variables vs. *configurations* of category variables. Results: Nonverbal aspects are very important in communication of affect. Affective force *can* be conveyed by the nonverbal features alone (e.g., arousal $\Leftrightarrow$ mean F0, s.d. F0) $\rightarrow$ covariance model. Affective signalling can depend on configurations of category variables (e.g., aggressive $\Leftrightarrow$ final fall for yes/no-question in German) $\rightarrow$ depends on verbal content, configuration model. Followed by Ladd et al. (1985).

Scherer, K. R. and Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1(4):331–346.

Experiment. Follow-up to Scherer (1974). Synthesised tone sequences in a factorial design (amplitude variation, pitch level, pitch contour direction, pitch variation, tempo, envelope, and filtration cut-off level). Perception test with three emotion dimension scales $\Rightarrow$ tempo had a very large effect on activity attributions, and large effects on pleasantness and potency. Filtration cut-off level (number of harmonics, related to voice quality) had a strong effect on potency, and to a lesser extent on pleasantness and activity. Pitch level also influenced judgments on all three dimensions.

Scheutz, M. and Römmer, B. (2001). Autonomous avatars? from users to agents and back. In de Antonio, A., Aylett, R., and Ballin, D., editors, *Proceedings of IVA*, pages 61–71, Madrid, Spain.

System description. User agents for a dating game. Uses a three-dimensional model for emotion *and* personality. Reactive and deliberative layers for action planning. Simple NLG system influenced by the affective state, for reporting to the user what happened while he was off-line.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, IMS-CL, University Stuttgart. http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

Resource. Description of the Stuttgart-Tübingen Tagset for German parts of speech.

Schlosberg, H. (1941). A scale for the judgement of facial expressions. *Journal of Experimental Psychology*, 29:497–510.

Experiment. The first paper introducing the idea of emotion dimensions into experimental psychology. Idea to use a "scale" consisting of 6 categories ⇒ obtain a numeric measure of divergence in judgments. Adjacent categories on the scale are somehow similar. Observed circularity of the scale. Proposed two dimensions capturing the most salient of the properties distinguishing opposite categories on the circular scale: Pleasantness/Unpleasantness and Attention/Rejection. Presents this as a simplified description.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, 61(2):81–88.

Theory. Commonly seen as the paper that popularised the description of emotions in terms of emotion dimensions. Introduced the *activation* dimension as readiness to take any action. Suggested measuring it with physiological variables, in particular galvanic skin response. In addition, the two dimensions *pleasantness/unpleasantness* and *attention/rejection*, proposed by Schlosberg (1941), are proposed for differentiating among the states at the activated end of the activation dimension.

Schoentgen, J. (1999). A random-walk model of jitter. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2441–2444, San Francisco, USA.

Algorithm. Detailed description of jitter properties. Presents formulae allowing to calculate period lengths that vary according to jitter.

Schröder, M. (1998). L'expression vocale de l'amusement : premières expériences audiovisuelles. T.E.R. de maîtrise, Institut de la Communication Parlée, Université Stendhal, Grenoble, France. http://www.dfki.de/~schroed

Experiment. Amused speech was elicited from four speakers on standard sentences. In addition, the same speakers enacted amusement on the same sentences, and spoke the sentences in a neutral way. Audiovisual, visual-only and audio-only perception tests were conducted. The spontaneously amused version was reliably distinguished from the neutral version in all three settings. In an audiovisual perception test pairing spontaneous and acted amusement, only some subjects were able to reliably identify the spontaneous version.

Schröder, M. (1999a). Can emotions be synthesized without controlling voice quality? *Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland*, pages 37–55. http://www.dfki.de/~schroed

Experiment. Copy synthesis, 4 emotions + neutral, 3 speakers, modelled duration, F0 (and energy). Preselection. Forced choice perception test: some stimuli very well recognised, others not at all ⇒ different expression strategies. Free-association perception test ⇒ rated as "disappointment" more or less independent of the stimulus ⇒ text or voice quality influences.

Schröder, M. (1999b). Zur Machbarkeit von Synthese emotionaler Sprache ohne Modellierung der Stimmqualität. In *Proceedings of Elektronische Sprachsignalverarbeitung Görlitz*, volume 16 of *Studientexte zur Sprachkommunikation*, pages 222–229, Dresden. w.e.b. Universitätsverlag. http://www.dfki.de/~schroed

Experiment. German summary of Schröder (1999a).

Schröder, M. (2000). Experimental study of affect bursts. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 132–137, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. An experimental study of affect bursts, or emotional interjections. Affect bursts were produced by actors, and presented in a perception test. Most affect bursts were identified as one of ten emotion categories nearly unambiguously. See also Schröder (2003) for an extended version of this study.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Proceedings of Eurospeech 2001*, volume 1, pages 561–564, Aalborg, Denmark. http://www.dfki.de/~schroed

Review. Overview of existing approaches and techniques used in emotional speech synthesis. Prosody rules employed, evaluation paradigms.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication Special Issue Speech and Emotion*, 40(1-2):99–116. http://www.dfki.de/~schroed

Experiment. Extended version of Schröder (2000). In addition to the work reported there, a criterion was proposed for distinguishing raw affect bursts from affect emblems, namely the transcription variability of non-experts. A tentative list of typical German affect emblems and of raw affect bursts was proposed on that basis. In addition, the role of segmental structure in recognition was addressed in a written perception test, were subjects had to rate orthographic transcripts as one of the ten emotion categories. Many affect bursts were thus shown to use their segmental form for recognition, while some depended crucially on the prosody and could not be recognised from the written transcript.

Schröder, M., Aubergé, V., and Cathiard, M.-A. (1998). Can we hear smiles? In *Proceedings of the 5th International Conference of Spoken Language Processing*, Sydney, Australia. http://www.dfki.de/~schroed

Experiment. English language summary of Schröder (1998).

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of Eurospeech 2001*, volume 1, pages 87–90, Aalborg, Denmark. http://www.dfki.de/~schroed

Corpus analysis. Correlated dimensional emotion descriptions with acoustic measurements. Clear effects that seem compatible with the literature.

Schröder, M. and Grice, M. (2003). Expressing vocal effort in concatenative synthesis. In *Proceedings of the 15th International Conference of Phonetic Sciences*, Barcelona, Spain. to appear.

Experiment. A new diphone database with a full diphone set for each of three levels of vocal effort is presented. A theoretical motivation is given why this kind of database will be useful for emotional speech synthesis. Two hypotheses are verified in perception experiments: (I) The three diphone sets are perceived as belonging to the same speaker; (II) The vocal effort intended during database recordings is perceived in the synthetic voice. The results clearly confirm both hypotheses.

Schröder, M. and Trouvain, J. (2001). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland. http://mary.dfki.de

System description. Description of the TTS system MARY. Interface to intermediate processing results. Examples of use in teaching, research and development.

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52:1061–1086.

Experiment. Followed Fehr & Russell (1984) in exploring a prototype description of emotion concepts. Out of the list by Averill (1975), generated a list of 213 emotion words (nouns) which was rated by subjects for emotion prototypicality (a scale with the end points "I definitely would/would not call this an emotion"). The 135 best-rated emotion words were then used in a similarity judgment task, which was analysed in two ways: First, using a hierarchical cluster analysis, yielding 5 basic level emotion categories (love, joy, anger, sadness, and fear), and 25 subcategories further differentiating the basic categories. Second, the same similarity ratings were analysed using classical multi-dimensional scaling techniques, reproducing earlier findings (Russell & Mehrabian, 1977) that a three-dimensional space represents the data appropriately. (Two- and three-dimensional coordinates for all 135 emotion words are listed.) Hierarchical and dimensional representations show different aspects of the emotion knowledge expressed in the similarity ratings. In a second study, subjects gave narrative accounts of emotion scenarios, which were used to characterise the prototypical emotion episodes.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling english prosody. In *Proceedings of the 2nd International Conference of Spoken Language Processing*, pages 867–870, Banff, Canada.

Reference paper for the ToBI system of intonation transcription.

Skut, W. and Brants, T. (1998). Chunk tagger – statistical recognition of noun phrases. In *Proceedings of the ESSLLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany. http://www.coli.uni-sb.de/~thorsten/publications

Reference paper for Chunkie.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington D.C., USA. http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html

Reference paper for the NEGRA corpus.

Sloman, A. (1998). Damasio, Descartes, alarms and meta-management. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego*, pages 2652–2657. http://www.cs.bham.ac.uk/research/cogaff/Sloman_smc98.pdf

Theory. Layered architecture for both biological and silicon "computers": reactive, deliberative and meta-management layers. Global alarm systems triggered by input on reactive/deliberative layer produce control signals = primary/secondary emotions. Meta-management ⇒ qualia and tertiary emotions.

Sluijter, A. (1995). *Phonetic Correlates of Stress and Accent*, chapter 5: Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in American English, pages 93–131. Holland Academic Graphics, The Hague, The Netherlands.

Thorough experimental study (production study). Source and filter variations with stress and accent. Source: LF model. Used by Marumoto & Campbell (2000) for training HMMs for unit selection synthesis.

Smith, C. A. (1989). Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology*, 56(3):339–353.

Experiment. Thoroughly designed experiment. Starting from appraisal theories of emotion, developed concrete hypotheses concerning the physiological effects of specific appraisals. Experimental manipulation of appraisals in a factorial design, using a directed imagery task. Results: Appraisal of anticipated effort has a strong effect on physiological arousal, as measured by heart rate (and, to a lesser degree, skin conductance). Appraisal of unpleasantness or perceived obstacles triggers a frown.

Sokolowski, K. (1993). *Emotion und Volition*. Hogrefe, Göttingen.

Theory. Emotion processing in the brain happens in a system of its own (i.e., not identical to cognitive processing). Components of emotions: Expressive, cognitive, physiological, behavioural and subjective. "Cold cognitions" (rational) vs. "hot cognitions" (emotionally biased). Motives = person-specific evaluation predispositions. Voluntary influence on hot cognitions is difficult. Two states of action control: convergent with motivation and volitional, against motivation ⇒ subjective experience of strain, less fun. Platos trias: evil black horse "greed", noble white horse "courage", car driver "reason".

Sproat, R., editor (1997). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer, Boston.

Description of the major modules of the Bell Labs TTS system.

Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., and Edgington, M. (1998). SABLE: A standard for TTS markup. In *Proceedings of the 5th International Conference of Spoken Language Processing*, pages 1719–1724, Sydney, Australia. http://www.bell-labs.com/project/tts/sabpap

Conference paper introducing SABLE.

Sproat, R., Taylor, P. A., Tanenblatt, M., and Isard, A. (1997). A markup language for text-to-speech synthesis. In *Proceedings of Eurospeech 1997*, Rhodes/Athens, Greece.

One of the first references for a speech synthesis markup language.

Stallo, J. (2000). Simulating emotional speech for a talking head. Honour's thesis, School of Computing, Curtin University of Technology, Australia. http://www.computing.edu.au/~stalloj/projects/honours

Experiment. Inspired by Murray & Arnott (1995), implemented prosody rules for emotion expression in Festival. Conducted perception tests: a) following Murray and Arnott, a forced-choice perception test using emotionally neutral and emotional text; b) a preference task in a talking head setting. In the latter, a visually expressive talking head was combined once with standard synthetic speech and once with manually tuned expressive synthetic speech. The version with expressive speech was very clearly preferred.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689.

Formalism. Extension of the Combinatory Categorial Grammar (CCG) to information structure, thereby explaining/predicting prosody. Complex formalism. Many references to related work.

Stibbard, R. (2001). *Vocal expression of emotions in non-laboratory speech: An investigation of the Reading/Leeds Emotion in Speech Project annotation data*. PhD thesis, University of Reading, UK.

Experiment. Critical review of the Emotion in Speech Project in Reading/Leeds. Statistically analysed the prosodic/paralinguistic (Roach, 2000) and emotional (Greasley et al., 2000) annotations of the speech material. Nearly no correlations could be found. This is tentatively explained by the too fine-grained level of detail in both types of annotation, while not controlling for non-emotional factors in the spontaneous speech corpus.

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of facial expressions: A non-obtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777.

Experiment. Ingenious test of the facial feedback hypothesis: Subjects are requested, under a non-emotional pretext, to hold a pen either with their teeth (activating the muscles activated during a smile) or with their lips (inhibiting the muscles activated during a smile) while rating the funniness of cartoons. Result: When smile-related muscles are activated while looking at the cartoon, the cartoons are rated funnier, when smile-related muscles are inhibited, the cartoons are rated less funny than the control condition. Second experiment: Subjects who do not hold the pens while looking at the cartoons but only while, afterwards, rating their subjective amusement, show the inverse pattern. Explanation: The current state while rating is taken as a reference state against which the amusement in the past situation is evaluated. More positive current state makes past amusement seem less strong and vice versa. Conclusion: Facial expressions can not only modify an emotional experience, but also initiate one.

Tabachnick, B. G. and Fidell, L. S. (2001). *Using multivariate statistics*. Allyn & Bacon, Boston, fourth edition.

A standard reference on multivariate statistics.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, 27(1):24–27.

Experiment. Smiling is audible in monosyllables and in sentences. Reliable speaker difference. Acoustic analysis: Formants F1, F2, F3 raised; F0 raised; amplitude greater; no reliable change in duration.

Tartter, V. C. and Braun, D. (1994). Hearing smiles and frowns in normal and whispery registers. *Journal of the Acoustic Society of America*, 96(4):2101–2107.

Experiment. Smiles and frowns (= voluntarily produced facial expressions) can be heard in normal and whisper registers, in [hVd] syllables, spoken by male and female American speakers. Smiles increased, frowns decreased F2.

Tassa, A. and Liénard, J. (2000). A new approach to the evaluation of vocal effort by the PSOLA method. *WEB-SLS, The European Student Journal of Language and Speech*. http://www.essex.ac.uk/web-sls/papers/00-01/00-01.html

Experiment. Resynthesised single vowels, modelling prosodic and spectral properties in order to convert modal into soft and loud voice. Partial success.

Taylor, P. and Isard, A. (1997). SSML: A speech synthesis markup language. *Speech Communication*, 21:123–133.

Resource description. Markup to be used by language generation systems. Markup to be used by people who are not experts of linguistics. Separate *logical* and *physical* aspects of a document's structure (= all markup). Some SGML basics: DTD, ... Markup-to-Speech system: 1. Parser for markup; 2. Interpreter for markup. Note: This is not to be confused with the newer W3C SSML (Walker & Hunt, 2001).

Tickle, A. (1999). Cross-language vocalisation of emotion: Methodological issues. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 305–308, San Francisco, USA.

Methodology, experiment. English and Japanese. Using facial expression pictures to be imitated while uttering nonsense words. Well-designed and controlled. Categories or dimensions (activation and valence)? Disinhibit the speaker by a "game".

Tickle, A. (2000). English and Japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 104–109, Northern Ireland. http://www.qub.ac.uk/en/isca/proceedings

Experiment. Nonsense utterances expressing one of four emotions (happy, sad, angry, and fearful) and a calm state were elicited from English and Japanese speakers (see Tickle (1999)) and presented to English and Japanese listeners. English speakers were better recognised than Japanese speakers, possibly due to display rules stigmatising emotion expression in Japanese culture. While English listeners were better at recognising emotions encoded by English speakers than Japanese listeners, English and Japanese listeners showed nearly identical recognition rates for emotions encoded by Japanese speakers.

Tischer, B. (1993). *Die vokale Kommunikation von Gefühlen*, volume 18 of *Fortschritte der psychologischen Forschung*. Psychologie-Verlags-Union, Weinheim, Germany.

Habilitationsschrift. German-language review of the domain of vocal emotion expression, including a review of the literature on vocal expression of emotions taking 53 studies into account. In the practical part, a standard text ("Sag das nochmal. Ich kanns nicht glauben. Was für ein Tag.", Engl. "Say that again. I can't believe it. What a day."), embedded in each of 14 emotional paragraphs, was produced by two male and two female professional speakers. In order to examine the temporal development of emotion attribution by the listener, a gating paradigm was employed in the perception tests: Initially, only the first few syllables of each test utterance were played to the listeners; subsequently, more and more of the utterance was revealed. The changes in ratings indicate the location of the information leading to the emotional judgment. A total of 1211 subjects participated in the various parts of the listening test. The results include detailed correlation analyses between 39 acoustic parameters and categorical and dimensional ratings. Unfortunately, no numerical results are given.

Tomlinson, B. and Blumberg, B. (2001). Social behavior, emotion and learning in a pack of virtual wolves. In *Proceedings of the AAAI Fall Symposium "Emotional and Intelligent II: The Tangled Knot of Social Cognition"*, North Falmouth, MA, USA. http://badger.www.media.mit.edu/people/badger/papers/BTomlinson01.pdf

System description. Model social behaviour of virtual wolves by means of emotions, represented by the three emotion dimensions Pleasure, Arousal, Dominance. Emotion is affected by the environment and can trigger actions. Context-specific emotional memories allow for the learning of emotional reactions to stimuli including appropriate actions such as submission display.

Traber, C. (1993). Syntactic processing and prosody control in the SVOX TTS system for German. In *Proceedings of Eurospeech 1993*, pages 2099–2102, Berlin, Germany.

System description. Description of the approach taken to TTS in the SVOX system.

Trouvain, J. (2001). Phonetic aspects of "speech-laughs". In *Proceedings of Orage*, Aix-en-Provence, France.

Experiment. Studied speech-laughs in the Kiel Corpus. Acoustic / phonetic properties of speech laughs. Perception test does not support the idea of a smile–laugh continuum.

Trouvain, J. (2002). Tempo control in speech synthesis by prosodic phrasing. In *Proceedings of Konvens*, Saarbrücken, Germany.

Experiment. Non-linear tempo modelling for speech synthesis. Prosodic phrase breaks were added for slower speaking rate and deleted for faster speaking rate. In a perception test, these versions were generally preferred over linearly stretched/shrunk versions without phrase adjustments, with the exception of moderately slow versions in which apparently too many pauses were inserted.

Trouvain, J. and Barry, W. J. (2000). The prosody of excitement in horse race commentaries. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 86–91, Northern Ireland.

Experiment. The prosodic correlates of excitement were examined in three horse race commentaries. The following trends became obvious from this data: As excitement built up during the races, most pauses became shorter; inter-pause and inter-breath stretches became shorter, indicating an increase in breathing; speaking rate did not increase; fundamental frequency level did rise by about an octave; intensity increased slightly; and spectral tilt decreased, indicating higher vocal tension.

Trouvain, J. and Grice, M. (1999). The effect of tempo on prosodic structure. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 1067–1070, San Francisco, USA.

Experiment. Reading fast $\Rightarrow$ less and shorter pauses, (often) less and weaker boundaries, simpler tones; but not always increased articulation rate, no systematic changes in pitch range. Reading slow $\Rightarrow$ often but not always, the opposite of speaking faster. Inter-individual differences: Different speakers use parameters to a different extent.

Tsapatsoulis, N., Raouzaiou, A., Kollias, S., Cowie, R., and Douglas-Cowie, E. (2002). Emotion recognition and synthesis based on MPEG-4 FAPs. In Pandzic, I. S. and Forchheimer, R., editors, *MPEG-4 Facial Animation - The standard, implementations, applications*. John Wiley & Sons, Hillsdale, NJ, USA.

Algorithm. Represent emotional states using facial animation parameters (FAPs). Proposes an algorithm for modelling facial expressions of non-archetypal emotions, using Whissell's activation dimension and Plutchik's angular orientation. Emotion words belonging to the same "universal emotion category" are derived from the archetype by stretching the FAP ranges according to activation. Other, non-universal emotion words are expressed by interpolating between the two closest neighbours according to the angular emotion orientation. An example is given how a facial expression for "guilty" is derived from "afraid" and "sad".

Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3:223–234.

Experiment. On a number of neutral carrier sentences, modified intonation contour (level, range and shape combined). Perceptual ratings using Osgood's semantic differential (a number of paired adjective scales) $\Rightarrow$ scales related to pleasantness account for most of the variance.

Vaissière, J. (1983). Language-independent prosodic features. In Cutler, A. and Ladd, D. R., editors, *Prosody: Models and Measurements*, pages 53–66. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo. References pp. 147–156.

Overview. Language-independent similarities: pauses; F0 declination; resetting of the baseline; frequency range decline; rise and fall of F0; final lengthening; intensity and the syllable (normally F0, intensity and duration decrease together). Prosodic differences among languages: stress languages (e.g., English) vs. French: different timing; priority: stress or boundary?; interrelation between F0, intensity and duration. Many references for different languages.

van Hessen, A. J., Jansen, R., and Pols, L. C. W. (1999). Pre-processing input text: Improving pronunciation for Dutch text-to-speech systems. In *Proceedings of the 14th International Conference of Phonetic Sciences*, volume 3, pages 2243–2246, San Francisco, USA.

Algorithm. Treatment of numerals in a dutch TTS.

van Hooff, J. A. R. A. M. and Aureli, F. (1994). Social homeostasis and the regulation of emotion. In van Goozen, S. H. M., van de Poll, N. E., and Sergeant, J. A., editors, *Emotions: Essays on emotion theory*, pages 197–217. Lawrence Erlbaum, Hillsdale, NJ.

Ethology. "Displays" (= social signal) are *ritualised* movements/postures with a communicative function. (some details about this ritualisation) Yawning in macaques = power display (teeth). Reconciliation and redirection (=loser aggresses someone else) after a fight in monkeys.

van Santen, J., Sproat, R. W., Olive, J., and Hirschberg, J., editors (1996). *Progress in Speech Synthesis*. Springer Verlag, New York.

A collection of chapters providing accounts of research in various fields of speech synthesis.

Velásquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pages 10–15, Providence, RI.
http://www.ai.mit.edu/people/jvelas/papers/Velasquez-AAAI-97.ps

System description. MIT's Cathexis system. Cognitive as well as non-cognitive elicitors for emotions: neural, sensorimotor, motivational, and cognitive. Emotions are represented as basic emotions following Ekman. Emotions influence the behaviour system according to Frijda's action tendencies.

VoiceXML (2000). *VoiceXML 1.0 Specification*. VoiceXML Forum.
http://www.voicexml.org

Reference for VoiceXML.

Vroomen, J., Collier, R., and Mozziconacci, S. J. L. (1993). Duration and intonation in emotional speech. In *Proceedings of Eurospeech 1993*, volume 1, pages 577–580, Berlin, Germany.

Experiment. Astonishingly good recognition rates, using only intonation and duration. Preselection from 3 actors x 5 sentences x 13 emotions database. Copy synthesis using 6 emotions + neutral: Recognition mean = 81% (PSOLA dynamic time warping of a monotonous utterance). Acceptable results when modelling only overall speed and stylised intonation patterns according to a Dutch "grammar" of intonation, both copied onto a neutral utterance and synthesised with two different diphone inventories.

Wagner, P. S. (1999). The synthesis of German contrastive focus. In *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, USA.

Experiment. Contrastive focus is measured in terms of "perceptual prominence" of the accentuated syllable. Synthesised contrast is best recognised when using "normal" accent prominence (= intonation + duration) + additional duration + postfocal deaccentuation (= reducing the prominence values of postfocal accented syllables).

Walker, M. R. and Hunt, A. (2001). *Speech Synthesis Markup Language Specification*. W3C. http://www.w3.org/TR/speech-synthesis

Resource description. A proposal for a standardised markup language for relatively high-level speech synthesis input markup. Many features in common with SABLE (Sproat et al., 1998), some improvements. At the beginning of 2003, still in draft status.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of Positive and Negative Affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Experiment. Building on the framework of positive and negative affect dimensions (Watson & Tellegen, 1985), a concise measurement tool is developed and validated. Each of the two dimensions is measured by ten adjective scales.

Watson, D. and Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, 98:219–235.

Model. Reanalysed the data from many studies on self-reported mood. Proposed a two-dimensional interpretation in terms of Positive and Negative Affect dimensions, which are rotated by 45° compared to the pleasantness and arousal dimensions.

Wells, J. C. (1996). *SAMPA Phonetic Alphabet for German*.
http://www.phon.ucl.ac.uk/home/sampa/german.html

Resource description. A phonetic alphabet for German, using only ASCII characters, thus particularly suitable for computer-based automatic use.

Whissell, C. M. (1989). The dictionary of affect in language. In Plutchik, R. and Kellerman, H., editors, *Emotion: Theory, Research, and Experience*, volume 4: The Measurement of Emotions, pages 113–131. Academic Press, New York.

Resource description. A list of 4000 words with associated ratings on the Activation and Evaluation dimensions. 4=mean, s.d.=1, values 1-7. 3000 words are at a distance of more than 1 from the neutral mean value. The word list can be used to objectively measure the emotional colouring associated with free text.

Wouters, J. and Macon, M. W. (1998). A perceptual evaluation of distance measures for concatenative speech synthesis. In *Proceedings of the 5th International Conference of Spoken Language Processing*, volume 6, pages 2747–2750, Sydney, Australia.

Experiment. Perceptual ratings of vowel similarity (vowels taken from different contexts) were correlated with objective measures. Best performed: mel-based cepstral and Itakura distances. Minimal added benefit of utilising weighted distances or delta features.

Wouters, J., Rundle, B., and Macon, M. W. (1999). Authoring tools for speech synthesis. In *Proceedings of Eurospeech 1999*, Budapest, Hungary.

System. Application of SABLE markup for teaching deaf children with implants. Graphical user interface for SABLE annotation, text style representation of SABLE tags. A few extensions to the SABLE standard: Pitch contour, use Worldbet phonetic alphabet, specify segment durations.

Wundt, W. (1896). *Grundriss der Psychologie*. Verlag von Wilhelm Engelmann, Leipzig.

One of the first mentions of the idea that emotions are organised according to three dimensions.

Wurm, L. H. and Vakoch, D. A. (1996). Dimensions of speech perception: Semantic associations in the affective lexicon. *Cognition & Emotion*, 10(4):409–423.

Experiment. The question was examined whether online perceptual processes (as opposed to post-perceptual processing) are influenced by emotional meaning as measured by emotion dimensions. Emotion words whose location on the three dimensions evaluation, activity and potency were known, as well as non-words, were presented in a speeded lexical decision task. Reaction times were significantly correlated to potency, as well as to a three-way interaction of evaluation, potency and activity. The authors give an evolutionary explanation of this finding: Immediate threats require fast processing.

Yang, L. and Campbell, N. (2001). Linking form to meaning: The expression and recognition of emotions through prosody. In *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perthshire, Scotland. http://www.ssw4.org

Corpus analysis. A corpus of six hours of Chinese conversational speech was recorded. An impressionistic analysis of the corpus lead to the identification of a) types of emotion typically expressed in spontaneous speech, and b) the prosodic means through which they are expressed.

Zanna, M. and Rempel, J. K. (1988). Attitudes: A new look at an old concept. In Bar-Tal, D. and Kruglanski, A. W., editors, *The social psychology of knowledge*, pages 315–334. Cambridge University Press, Cambridge, UK.

According to Cowie & Cornelius (2003), contains a “standard definition of attitude”.

Zei Pollermann, B. (2002). A place for prosody in a unified model of cognition and emotion. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France. http://www.lpl.univ-aix.fr/sp2002/papers.htm

Theory. An interesting proposal for organising scientific concepts in this research domain. Cognition and emotion are considered intrinsically inseparable because each cognitive state has the three dimensions valence, arousal and power as its constituents, i.e. each cognitive state is affectively coloured. Such a state is experienced as emotional as soon as the current location on one of the dimensions passes a person-specific “out-of-normal-range” threshold. In addition, the emotion expression or communication is discussed with respect to the intentionality of the signalling, using the symptom-symbol-sign distinction and the push-pull distinction. Within that frame, a characterisation of the five types of affective states as listed by Scherer (2000b) is proposed. Two examples of the author’s own (clinical) research are given, exemplifying studies into the prosodic effects of the arousal and power dimensions, respectively.

Zerling, J.-P. (1995). Onomatopées et interjections en français. *Travaux de l’Institut de Phonétique de Strasbourg*, 25:95–109.

Resource. List of interjections in alphabetical order, with short explanation.

Appendix A

Prosody rules in emotional speech synthesis systems

The following tables summarise the prosody rules used in a number of previous publications on emotional speech synthesis. Some general comments are included with each study in order to summarise the approach. The presentation of the respective prosody rules is organised according to the frequently-used emotion categories “Happiness”, “Sadness”, “Anger”, “Fear”, “Surprise” and “Boredom”.

The first line in each field reproduces the category name used by the author(s), the original non-English category label if applicable, and a recognition rate in percent if the results of a perceptual evaluation were available, followed by an indication of chance level based on the number of available answers.

The acoustic parameters are organised according to the most frequently used parameters, namely “F0 mean”, “F0 range”, “tempo”, “loudness” and “voice quality”. The parameter values were taken from the authors as they were and commented where necessary and possible.

Table A.1: Prosody rules used by Burkhardt & Sendlmeier (2000). Language: German.

General comments	Works with formant synthesis (KLSYN88). Perception test with one test sentence. Percentages are relative to the maximum allowed change (not specified).
Happiness	Joy: G. Freude 81% (1/9), happiness: G. Wohlbefinden, Zufriedenheit 62% (1/9) **F0 mean**: joy "+50%", happiness "+0%" **F0 range**: "+100%" **Tempo**: joy "+30%", happiness "+20%" or "-20%"? **Loudness**: **Voice Qu.**: modal or tense; "lip-spreading feature": F1 / F2 +10% **Other**: "wave pitch contour model": main stressed syllables are raised ("+100%"), syllables in between are lowered ("-20%")
Sadness	Crying Despair: G. Weinerliche Trauer 69% (1/9), Quiet Sorrow: G. Stille Trauer 38% (1/9) **F0 mean**: desp. "+100%", sor. "-20%" **F0 range**: "-20%" "pitch variability" (within syllables) "-20%" **Tempo**: desp. "-20%", sor. "-40%" **Loudness**: **Voice Qu.**: breathy **Other**: "F0 flutter" (jitter): desp. "FL 200", sor. "FL 300"
Anger	Hot anger: G. Wut, Zorn 29% (1/9); cold anger: G. Ärger, genervt 60% (1/9) **F0 mean**: hot "+50%", cold "-20%" **F0 range**: **Tempo**: "+30%" **Loudness**: **Voice Qu.**: tense **Other**: cold: vowel articulation precision (formant overshoot): "+30%" for stressed syllables, "-20%" for unstressed syllables
Fear	Fear: G. Angst 52% (1/9) **F0 mean**: "+150%" **F0 range**: "+20%" **Tempo**: "+30%" **Loudness**: **Voice Qu.**: falsetto **Other**:
Surprise	
Boredom	Boredom: G. Langeweile 71% (1/9) **F0 mean**: "-20%" **F0 range**: "-50%" reduced pitch variability (within syllables) "20%" ($\Rightarrow$ almost flat intonation contour) **Tempo**: "-20%" additional lengthening of stressed syllables (40%) **Loudness**: **Voice Qu.**: modal **Other**: reduced vowel precision (formant undershoot) stressed syllables "-20%", unstressed syllables "-50%"
Other	Neutral 55% (1/9)

Table A.2: Prosody rules used by Cahn (1990). Language: American English.

General comments	Has taken results from 5 studies from 1939-1974. So the empirical basis upon which her observations are built is not very broad and not recent. But the rules she derived have been tested in a perception test. Her system is a prototype. Its structure is interesting, because unlike other systems, it incorporates a simplified version of a text-to-speech system, which allows the treatment of high-level concepts like phrasing and accents. Her values are relative to neutral, from -10 to 10 (0 being neutral)
Happiness	Glad 48% (1/6) **F0 mean**: "-3" reference line (=pitch value betw. accents) "-8", but less final lowering "-4" **F0 range**: "+10" contour slope "+5" (i.e. pitch range is expanding throughout the utterance) accent shape (= steepness of F0 contour at accents) "+10" **Tempo**: "+2" fluent pauses "-5", hesitation pauses "-8" **Loudness**: **Voice Qu.**: breathiness "-5", brilliance "-2" **Other**: stress frequency (no. of accents) "+5"; precision of articulation "-3"
Sadness	Sad 91% (1/6) **F0 mean**: "0" reference line "-1", less final lowering "-5" **F0 range**: "-5" steeper accent shape "+6" **Tempo**: "-10" more fluent pauses "+5", hesitation pauses "+10"; **Loudness**: "-5" **Voice Qu.**: breathiness "+10", brilliance "-9" **Other**: stress frequency "+1" precision of articulation "-5"
Anger	Angry 44% (1/6) **F0 mean**: "-5" reference line "-3", extreme final lowering "+10" **F0 range**: "+10" steep accent shape "+10" **Tempo**: "+8" less fluent pauses "-5", hesitation pauses "-7"; **Loudness**: "+10" **Voice Qu.**: breathiness "-5", brilliance "+10" **Other**: precision of articulation "+5"
Fear	Scared 52% (1/6) **F0 mean**: "+10" reference line "+10", no (or negative?) final lowering "-10" **F0 range**: "+10" steeply rising contour slope "+10" steep accent shape "+10" **Tempo**: "+10" no fluent pauses "-10", but hesitation pauses "+10"; **Loudness**: "+10" **Voice Qu.**: brilliance "+10" laryngealisation "-10" **Other**: stress frequency "+10", loudness "+10"
Surprise	Surprised 44 % (1/6) **F0 mean**: "0" reference line "-8" **F0 range**: "+8" steeply rising contour slope "+10" steeper accent shape "+5" **Tempo**: "+4" less fluent pauses "-5", hesitation pauses "-10"; **Loudness**: "+5" **Voice Qu.**: brilliance "-3" **Other**:
Boredom	
Other	Disgusted 42% (1/6)

Table A.3: Prosody rules used by Gobl & Ní Chasaide (2000). Language: Irish English.

General comments	Controlled formant-synthesis of 7 voice qualities using KLSYN88a: tense voice, breathy voice, whispery voice, creaky voice, modal voice; and, experimentally, harsh voice and lax-creaky voice. Voice quality distinctions between accentuated and unaccentuated syllables. Perception test with seven 7-point scales: relaxed/stressed, content/angry, friendly/hostile, sad/happy, bored/interested, intimate/formal, timid/confident, and afraid/unafraid. Results: Two groups of voice qualities: tense/harsh voice signals aggressive states, breathy/whispery/creaky/lax-creaky voice signals relaxed, unaggressive states.
Happiness	Happy **Voice Qu.**: small effect: tense voice is rather happy than sad.
Sadness	Sad **Voice Qu.**: medium effect: lax-creaky; small effect: breathy/whispery/creaky
Anger	angry **Voice Qu.**: strong effect: tense/harsh voice
Fear	afraid **Voice Qu.**: small effect: whispery voice
Surprise	
Boredom	Bored **Voice Qu.**: strong effect: lax-creaky voice; small effect: creaky voice
Other	Tense/harsh voice strongly signals stressed, angry, hostile, formal, and confident states, mediumly interested; lax-creaky voice strongly signals relaxed, content, bored, and intimate states; mediumly friendly and sad; whispery/breathy voice mediumly signals relaxed, content, friendly and timid states.

Table A.4: Prosody rules used by Heuft et al. (1996). Language: German.

General comments	Copy Synthesis of emotional sentences. Recognition rates close to chance level (except for fear). However, interesting listener-centered approach to acoustic properties: Correlations between acoustic parameters and recognised emotion regardless of the intended emotion.
Happiness	Joy **F0 mean**: **F0 range**: **Tempo**: very fast **Loudness**: **Voice Qu.**: **Other**:
Sadness	Sadness **F0 mean**: low **F0 range**: **Tempo**: fast **Loudness**: **Voice Qu.**: **Other**:
Anger	Anger **F0 mean**: very low **F0 range**: narrow **Tempo**: **Loudness**: **Voice Qu.**: **Other**:
Fear	Fear **F0 mean**: **F0 range**: narrow **Tempo**: very fast **Loudness**: **Voice Qu.**: **Other**:
Surprise	
Boredom	
Other	Neutral: F0 mean low, F0 range narrow, Tempo slow Disgust: F0 range wide

Table A.5: Prosody rules used by Iriondo et al. (2000). Language: Castillian Spanish.

General comments	Corpus recording, validation and analysis, then abstraction of rules and implementation of these rules in a TTS (“EMOVS”) with PSOLA-based diphone and triphone concatenation. Perception tests with the synthesised stimuli are under way.
Happiness	Joy/Happiness **F0 mean**: increased (10-50%) **F0 range**: increased (120%) **Tempo**: decrease of silence duration (20%) **Loudness**: stable intensity **Other**: fast inflections of tone; F0 rise and fall times similar, no high plateaux, pitch-energy relation asynchronous (energy peaks 100 to 150 ms earlier than pitch peaks, sounds like laughter)
Sadness	Sadness **F0 mean**: decreased (10-30%) **F0 range**: decreased (30-50%), less than 30 Hz **Tempo**: duration of silences increased (50-100%) **Loudness**: decreased (10-25%) **Other**: “null” inflections of intonation; discourse fragmentation increased (10%)
Anger	Fury **F0 mean**: **F0 range**: very wide (can exceed 140 Hz) **Tempo**: slower; reduction of the number of silences (25%); increase the duration of silences (10%) **Loudness**: raising intensity from the begin to the end (5-10 dB) **Voice Qu.**: “most characteristic for fury”: increase of energy in 500-636 Hz and 2000-2500 Hz bandwidths (10-15 dB) **Other**: variation of the intonation structure (20-80 Hz); pitch rises faster than falls; monosyllabic high pitch plateaux; downward declination (approx.: topline 245-150 Hz, baseline 190-90 Hz)
Fear	Fear **F0 mean**: increased (5-10%) **F0 range**: decreased (5%) **Tempo**: faster (decrease of duration of phonic groups by 20-25%); decrease duration of silences (10%) **Loudness**: raised intensity (10%); energy globally rising **Other**: fast variations of pitch (by 60-100 Hz in 20-30 ms); bi- or trisyllabic high pitch plateaux; high plateaux rising (approx.: topline 180-250 Hz, baseline 140-140 Hz)
Surprise	Surprise **F0 mean**: increased (10-15%) **F0 range**: increased (15-35%); high inflections of intonation **Tempo**: faster (decreased duration of phonic groups by 10%) **Loudness**: increased (3-5 dB)
Boredom	
Other	Desire; Disgust recorded, but not modelled, because of recognition rate $<$ 50% in preselection test with natural stimuli.

Table A.6: Prosody rules used by Campbell & Marumoto (2000). Language: Japanese.

General comments	First attempt for unit selection based synthesis of emotions using emotion-specific selection criteria (HMMs and prosody rules). 3 emotions. Prosody rules, derived by comparing global mean values across emotional speech corpora, performed better than HMMs. Joy, however, was not recognised.
Happiness	Joy (not recognised: 11% 1/3) **F0 mean**: **F0 range**: reduced (* 0.625) **Tempo**: 5% faster (duration * 0.95) **Loudness**: **Voice Qu.**: **Other**:
Sadness	Sadness:52% (1/3) **F0 mean**: lower (-10 Hz) **F0 range**: reduced (0.6875) **Tempo**: 8% slower (duration * 1.08) **Loudness**: intensity range reduced by 5% **Voice Qu.**: **Other**:
Anger	Anger: 65% (1/3) **F0 mean**: raised (+7 Hz) **F0 range**: wider (* 1.125) **Tempo**: 2% faster (duration * 0.98) **Loudness**: intensity range increased by 10% **Voice Qu.**: **Other**:
Fear	
Surprise	
Boredom	
Other	

Table A.7: Prosody rules used by Montero et al. (1998, 1999a). Language: Spanish.

General comments	One actor produced 3 passages and 15 sentences of semantically neutral text in 4 emotions + neutral. Parameters are quite coded and not sufficiently explained, thus difficult to interpret. Untypical selection of parameters. Recognition rates are for neutral diphones with copy-synthesised emotional prosody.
Happiness	Happy 19% (1/5) **F0 mean**: higher than neutral **F0 range**: **Tempo**: pause duration: ca. half of neutral (both sentence-final and intra-sentence) **Loudness**: **Voice Qu.**: **Other**:
Sadness	Sad 67% (1/5) **F0 mean**: lower than neutral **F0 range**: **Tempo**: pause duration: ca. a third more than for neutral (both sentence-final and intra-sentence) **Loudness**: **Voice Qu.**: **Other**:
Anger	Angry (cold anger) 7% (1/5) **F0 mean**: like neutral **F0 range**: nearly no declination (final peaks as high as initial peaks) **Tempo**: pause duration: roughly 2/3 of neutral (both sentence-final and intra-sentence) **Loudness**: **Voice Qu.**: **Other**:
Fear	
Surprise	Surprise 76% (1/5) **F0 mean**: much higher than neutral (ca. 200 Hz for male speaker) **F0 range**: **Tempo**: pause duration: ca. 60% of neutral (both sentence-final and intra-sentence) **Loudness**: **Voice Qu.**: **Other**:
Boredom	
Other	Neutral 76% (1/5)

Table A.8: Prosody rules used by Mozziconacci (1998); Mozziconacci & Hermes (1999). Language: Dutch.

General comments	One actor produced 2 sentences in 6 emotions + neutral. Perceptually optimal values for each emotion for pitch level, pitch range, and duration were obtained in perception experiments with systematic variation of these parameters. The resulting values were used in a perception test using diphone speech (2 different diphone sets). The influence of pitch contour was studied in a perception test by independently varying pitch contour and pitch level/range.
Happiness	Joy 62% (1/7) **F0 mean**: end frequency 155 Hz (male speech) **F0 range**: excursion size 10 s.t. **Tempo**: duration rel. to neutrality: 83% **Other**: final intonation pattern 1&A or 5&A; avoid final patterns A, EA and 12.
Sadness	Sadness 47% (1/7) **F0 mean**: end frequency 102 Hz (male speech) **F0 range**: excursion size 7 s.t. **Tempo**: duration rel. to neutrality: 129% **Other**: final intonation pattern 3C; avoid final pattern 5&A.
Anger	Anger 51% (1/7) **F0 mean**: end frequency 110 Hz (male speech) **F0 range**: excursion size 10 s.t. **Tempo**: duration rel. to neutrality: 79% **Other**: final intonation pattern 5&A or A or EA; avoid final pattern 1&A or 3C.
Fear	Fear 41% (1/7) **F0 mean**: end frequency 200 Hz (male speech) **F0 range**: excursion size 8 s.t. **Tempo**: duration rel. to neutrality: 89% **Other**: final intonation pattern 12; avoid final pattern A or EA.
Surprise	
Boredom	Boredom 94% (1/7) **F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C; avoid final patterns 5&A and 12.
Other	Neutrality: 83% (1/7) F0 mean: end frequency 65 Hz (male speech) , F0 range: excursion size 5 s.t., Final intonation pattern: 1&A; avoid final patterns 3C and 12. Indignation: 68% (1/7) F0 mean: end frequency 170 Hz (male speech), F0 range: excursion size 10 s.t., Tempo: duration rel. to neutrality 117%, Final intonation pattern: 12; avoid final pattern 1&A.

Table A.9: Prosody rules used by Murray & Arnott (1995). Language: British English.

General comments	HAMLET: One of the first emotional speech synthesis systems. Uses DECtalk with "plugged-in" emotion algorithms, can do full rule-based Text-to-Speech. Interesting evaluation methodology (both neutral and emotional text synthesised with both neutral and emotional prosody), but that makes it non-trivial to compare to other's results.
Happiness	Happiness (not recognised with neutral text) **F0 range**: + 9 s.t.; reduce the amount of pitch fall at end of utterance by 10% of F0 range **Tempo**: + 30 wpm; duration of stressed vowels + 50%; modify phoneme durations for regular stressing ($\Rightarrow$ time between two stressed phonemes = 550 ms or multiple thereof) **Loudness**: + 3 dB **Other**: eliminate abrupt changes in pitch between phonemes.
Sadness	Sadness (well recognised with neutral text) **F0 mean**: -30 Hz **F0 range**: -2 s.t. **Tempo**: -40 wpm **Loudness**: -2 dB **Voice Qu.**: spectral tilt +65% **Other**: decrease articulation precision (reduce "a" and "i" vowels); eliminate abrupt changes in pitch between phonemes; replace upward inflections with downward inflections; add 80 ms pause after each word with more than 3 phonemes.
Anger	Anger (well recognised with neutral text) **F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear	Fear (not recognised with neutral text) **F0 mean**: +20 Hz **F0 range**: +3 s.t.; end F0 baseline fall +100Hz **Tempo**: +20 wpm **Voice Qu.**: laryngealisation +50% **Other**: increase articulation precision (un-reduce vowels; duration of plosives +30%)
Surprise	
Boredom	
Other	Grief (not well recognised with neutral text): F0 mean -40Hz, F0 range -4 s.t., Tempo -60 wpm, loudness -3 dB, Voice Qu. laryngealisation +50%, spectral tilt +65%, Other: decrease articulation precision (reduce "a" and "i"), increase the rate of F0 declination (descend linear (except stress peaks) to 50 Hz), add 80 ms pause after words with more than 3 phonemes. Disgust (not recognised with neutral text): F0 mean +10 Hz, F0 range +9 s.t., Voice Qu. laryngealisation +40%, F4 frequency -50 Hz, Other: increase articulation precision (unreduce vowels; duration of plosives +30%); add downward pitch inflections at word endings (F0 values of last vowel + following cons. in each word are decreased by 10% of F0 range)

Table A.10: Prosody rules used by Murray et al. (2000). Language: British English.

General comments	Prototype using concatenative synthesis (BT Laureate). Better sound quality than DECtalk, emotions as well or better recognised than HAMLET. Two different steps of complexity: 1. Only global parameters set to emotion-specific values; 2. manual adaptations $\Rightarrow$ 1. as good as HAMLET, 2. clearly better.
Happiness	Happiness **F0 mean**: raised, high-pitched **F0 range**: **Tempo**: slightly increased **Loudness**: **Voice Qu.**: **Other**:
Sadness	Sadness **F0 mean**: lowered **F0 range**: **Tempo**: slightly slower **Loudness**: **Voice Qu.**: a bit of artificial "laryngealisation" **Other**:
Anger	Anger **F0 mean**: raised **F0 range**: increased **Tempo**: faster **Loudness**: **Voice Qu.**: artificial "laryngealisation" **Other**:
Fear	Fear **F0 mean**: even more raised than happy ("squeaky") **F0 range**: **Tempo**: very much increased **Loudness**: **Voice Qu.**: **Other**:
Surprise	
Boredom	
Other	

Table A.11: Prosody rules used by Rank & Pirker (1998); Rank (1999). Language: Austrian German.

General comments	Parameter values taken from Cahn (1990), Klasmeyer, and Heuft et al. (1996); primarily voice quality, only limited F0 modeling. Recognition rates above chance level only for sadness and anger. Interesting technical attempt to modify voice quality related parameters in residual excited LPC based synthesis: jitter, creaky voice, shape of glottis signal (but limited because of distortions), additional noise, and articulation precision for vowels. Scale for parameter values unclear.
Happiness	
Sadness	Sadness: G. traurig 69% (1/4) **F0 mean**: -3.0 **F0 range**: 0.5 **Tempo**: duration of vowels: 1.4, voiced cons. 1.4, unvoiced cons. 1.2; pause duration 3.0, pause duration variability 0.5 **Loudness**: amp. vowels 0.7, voiced cons. 0.7, unvoiced cons. 0.8; amp. shimmer 0.0 **Voice Qu.**: creaky rate 0.02, glottal noise 0.4 **Other**: F0 jitter 0.0005; articulation precision 0.95
Anger	Anger: G. zornig 40% (1/4) **F0 mean**: 0.0 **F0 range**: 2.0 **Tempo**: duration of vowels 0.75, voiced cons. 0.85, unvoiced cons. 0.9; pause duration 0.8, pause duration variability 0.0 **Loudness**: amp. vowels 1.3, voiced cons. 1.2, unvoiced cons. 1.1; amp. shimmer 0.0 **Voice Qu.**: **Other**: articulation precision 1.05
Fear	Fear: G. ängstlich 18% (1/4) **F0 mean**: 1.5 **F0 range**: 2.0 **Tempo**: duration of vowels 0.65, voiced cons. 0.55, unvoiced cons. 0.55; pause duration 0.6, pause duration variability 0.5 **Loudness**: amp. vowels 1.2, voiced cons. 1.0, unvoiced cons. 1.1; amp. shimmer 0.05 **Voice Qu.**: creaky rate 0.003; glottal noise 0.5 **Other**: F0 jitter 0.35; articulation precision 0.97
Surprise	
Boredom	
Other	Disgust: G. empört 22% (1/4)

Appendix B

XSLT stylesheet emotion-to-maryxml.xsl

The following XSLT stylesheet was used as an implementation of the emotional prosody rules, converting a simple emotion markup language into MaryXML.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
                version="1.0"
                xmlns="http://mary.dfki.de/2002/MaryXML">
  <xsl:output method="xml"
              encoding="ISO-8859-1"
              indent="yes"/>
  <xsl:strip-space elements="*|text()"/>

  <xsl:template match="/">
    <maryxml xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
             version="0.3" xml:lang="de">
     <voice name="de6">
      <xsl:apply-templates/>
     </voice>
    </maryxml>
  </xsl:template>


  <!-- emotion -->
  <xsl:template match="emotion">
    <xsl:variable name="pitch"
    select="round(110 + 0.3*@activation + 0.1*@evaluation - 0.1*@power)"/>

    <xsl:variable name="pitch-dynamics"
    select="format-number(round(-15 + 0.3*@activation - 0.3*@power),
            '+#;-#')"/>

    <xsl:variable name="_range"
    select="round(25 + 0.4*@activation)"/>
```

```
<!-- A range below 0 does not make sense -->
<xsl:variable name="range">
    <xsl:choose>
        <xsl:when test="$_range &lt; 0">0</xsl:when>
        <xsl:otherwise>
            <xsl:value-of select="$_range"/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:variable>

<xsl:variable name="_range-dynamics"
select="format-number(round(-40 + 1.2*@activation + 0.4*@power),
        '+#;-#')"/>
<!-- A range-dynamics below -100 does not make sense -->
<xsl:variable name="range-dynamics">
    <xsl:choose>
        <xsl:when test="$_range-dynamics &lt; -100">-100</xsl:when>
        <xsl:otherwise>
            <xsl:value-of select="$_range-dynamics"/>
        </xsl:otherwise>
    </xsl:choose>
</xsl:variable>


<xsl:variable name="accent-prominence"
select="format-number(round(0.5*@activation - 0.5*@evaluation),
        '+#;-#')"/>
<xsl:variable name="preferred-accent-shape">
  <xsl:choose>
    <xsl:when test="@evaluation &lt; -20">falling</xsl:when>
    <xsl:when test="@evaluation &gt; 40">alternating</xsl:when>
    <xsl:otherwise>rising</xsl:otherwise>
  </xsl:choose>
</xsl:variable>
<xsl:variable name="accent-slope"
select="format-number(round(1*@activation - 0.5*@evaluation),
        '+#;-#')"/>

<xsl:variable name="preferred-boundary-type">
  <xsl:choose>
    <xsl:when test="@power &gt; 0">low</xsl:when>
    <xsl:otherwise>high</xsl:otherwise>
  </xsl:choose>
</xsl:variable>

<xsl:variable name="rate"
select="format-number(round(0.5*@activation + 0.2*@evaluation),
'+#;-#')"/>

<xsl:variable name="number-of-pauses"
select="format-number(round(0.7*@activation), '+#;-#')"/>
<xsl:variable name="pause-duration"
select="format-number(round(-0.2*@activation), '+#;-#')"/>
```



```
    <xsl:variable name="vowel-duration"
    select="format-number(round(0.3*@evaluation + 0.3*@power),
            '+#;-#')"/>
    <xsl:variable name="nasal-duration"
    select="format-number(round(0.3*@evaluation + 0.3*@power),
            '+#;-#')"/>
    <xsl:variable name="liquid-duration"
    select="format-number(round(0.3*@evaluation + 0.3*@power),
            '+#;-#')"/>
    <xsl:variable name="plosive-duration"
    select="format-number(round(0.5*@activation - 0.3*@evaluation),
    '+#;-#')"/>
    <xsl:variable name="fricative-duration"
    select="format-number(round(0.5*@activation - 0.3*@evaluation),
    '+#;-#')"/>

    <xsl:variable name="volume"
    select="round(50 + 0.33*@activation)"/>

    <prosody pitch="{$pitch}" pitch-dynamics="{$pitch-dynamics}%"
        range="{$range}" range-dynamics="{$range-dynamics}%"
        preferred-accent-shape="{$preferred-accent-shape}"
        accent-slope="{$accent-slope}%"
        accent-prominence="{$accent-prominence}%"
        preferred-boundary-type="{$preferred-boundary-type}"
        rate="{$rate}%" number-of-pauses="{$number-of-pauses}%"
        pause-duration="{$pause-duration}%"
        vowel-duration="{$vowel-duration}%"
        nasal-duration="{$nasal-duration}%"
        liquid-duration="{$liquid-duration}%"
        plosive-duration="{$plosive-duration}%"
        fricative-duration="{$fricative-duration}%"
        volume="{$volume}">
      <xsl:apply-templates/>
    </prosody>

  </xsl:template>

  <xsl:template match="text()">
    <xsl:text>&#10;</xsl:text>
    <xsl:value-of select="normalize-space(.)"/>
    <xsl:text>&#10;</xsl:text>
  </xsl:template>

</xsl:stylesheet>
```

Appendix C

Written situation descriptions

The following written situation descriptions were used as candidates for the definition of emotional states. For each candidate, the mean values and standard deviations for activation and evaluation are given, on scales from -10 to 10.

Situations intended as: Neutral

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
neutral1	-5.9	-0.4	3.0	3.3

Situation: Am Rande einer Auktion: Ein Mann beschreibt ein Gemälde, das seit Generationen im Familienbesitz war. “Es gehörte meiner Schwiegermutter. Sie lebete in Wiesbaden und ich glaube, es wurde während des Kriegs leicht beschädigt. Vor vielen Jahren haben wir es geerbt, und seither hing es bei uns an der Wand.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
neutral2	-2.2	3.0	5.5	2.2

Situation: Zwei Bekannte unterhalten sich. ”Wir haben immer Sommertheater in Koblenz gespielt. Also, jetzt in den letzten zwei Jahren hab’ ich nicht mitgemacht, aber vorher sind wir ziemlich oft hingefahren, so für eine Woche hochgefahren und dann dageblieben und Theater gespielt. Und dann für den letzten Tag sind meistens Hans und Erika dazugekommen.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
neutral3	4.2	4.2	5.2	2.1

Situation: Zwei Studenten unterhalten sich, der eine hatte gerade ein Beratungsgespräch. “Also, und dann könnte ich erst mal diesen ersten Abschluss machen, und dann schon mal etwas Geld verdienen und nebenher das Studium weitermachen. Dann hätte ich innerhalb von 2 Jahren den vollen Abschluss.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
neutral4	-2.4	0.9	5.3	1.3

Situation: Kneipengespräch. “Schon beim Bund habe ich LKWs repariert. Dann habe ich bei der Firma gearbeitet, wo mein Vater Abteilungsleiter war, aber er war in einer ganz anderen Abteilung. Er war in der Vertriebsabteilung, und ich war in der Entwicklungsabteilung.”

Situations intended as: Moderate active negative

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactneg1	1.3	-2.8	3.8	2.7

Situation: Bruder und Schwester hatten sich verkracht und seit mehreren Jahren nicht miteinander gesprochen. Jetzt haben gemeinsame Freunde sie zusammengebracht, damit sie sich versöhnen. Er: “Ich würde gerne, ich würde dir gerne vertrauen, Lisa, aber das musst du dir echt verdienen. Damals, das hätte so ein besonderer Tag werden sollen, und alles ist schiefgelaufen wegen dir. Von meiner eigenen Schwester hätte ich echt mehr Ehrlichkeit erwartet!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactneg2	5.1	-2.5	2.5	4.4

Situation: Er und sie haben sich getrennt. Jetzt geht die Diskussion darum, wie oft er den gemeinsamen Sohn sehen darf. Er: “Du weißt, wie gerne ich ihn sehen will. Aber ich kann nicht so oft kommen, weil ich doch keinen Führerschein habe im Moment. Dann versteh doch, dass ich wenigstens anrufe ab und zu. Er ist doch jetzt fast drei, er kann doch schon sprechen! Aber immer wenn ich anrufe, denkst du, ich will Stress machen. Ich will doch bloß ab und zu mit ihm reden!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactneg3	-0.5	-6.8	4.8	2.0

Situation: Polit-Diskussion. “Die Bürger werden geschröpft, an vielen Stellen. Zum Beispiel die ganze Benzin-Steuer, das ist eine riesige versteckte Steuer. Die bringt der Regierung eine Menge Geld. Von 10 Euro, die man für Benzin ausgibt, sind 7 Euro Steuern!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactneg4	-0.2	-6.3	5.7	2.8

Situation: Zwei Freunde unterhalten sich über ihre Eltern. “Meine Mutter denkt so schwarz-weiß. Das ist echt nervig! Ihre Ansichten über Homosexualität, ihre Ansichten über Alkohol, über Sex... Weißt du, man kann einfach nicht mit ihr reden. Es bringt einfach nichts, also lass’ ich es, vermeide die Themen. Aber es ist echt frustrierend, sie kann einfach nicht akzeptieren, dass es eine andere Art zu denken gibt, oder einen anderen Lebensstil oder so. Bei ihr ist alles entweder falsch oder richtig, da gibt’s nix dazwischen.”

Situations intended as: Intense active negative

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactneg1	2.4	-7.2	5.6	2.2

Situation: Zwei Freunde unterhalten sich über das schlechte Verhältnis zu einem ehemaligen gemeinsamen Freund. “Es ist ein dauerndes Gemeckere. Die ganze Zeit geht das nur noch so. Wir können keine fünf Minuten im gleichen Zimmer sein ohne zu diskutieren und zu streiten. Ehrlich, ich hab einfach keine Lust mehr da drauf!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactneg2	0.9	-6.3	5.2	2.7

Situation: Er ist krank, aber sie besteht darauf, dass er sich das alles nur einredet. Er: “Es ist nicht nur in meinem Kopf. Warum verstehst du das denn nicht? Wie soll ich wieder zu Kräften kommen, wenn du mir die ganze Zeit Vorwürfe machst, dass es alles nur in meinem Kopf ist? Versuch’ doch wenigstens zu verstehen, wie ich mich fühle!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactneg3	1.3	-7.2	6.3	3.0

Situation: Polit-Diskussion. “Diese Regierung ist eine einzige große Unehrlichkeit. Die versprechen das Blaue vom Himmel und halten nichts davon, gar nichts. Kaum ist die Wahl vorbei, sind alle Versprechen vergessen, und stattdessen werden kräftig die Steuern erhöht!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactneg4	1.7	-7.8	5.9	1.9

Situation: Diskussion über die Untätigkeit der Polizei gegenüber Drogendealern. “Ein paar davon wohnen ganz bei mir in der Nähe. Die Polizei weiß Bescheid, der Stadtrat weiß Bescheid, und was tun sie? Nichts! Und wenn du nachfragst, kriegst du nur zu hören: ’Wir beobachten sie.’ Das ist der größte Mist, den ich in meinem ganzen Leben gehört habe!”

Situations intended as: Moderate active positive

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactpos1	-1.5	2.8	4.4	2.1

Situation: Zwei Freunde treffen sich wieder, nachdem sie sich eine Weile nicht gesehen hatten. Der eine hatte einige Probleme gehabt. Der andere fragt ihn, wie es ihm geht. “Mir geht’s gut. Nicht mehr so gestresst, und es ist jetzt auch wieder OK zu Hause. Mein kleiner Bruder hat sich wieder beruhigt, und ich hab eine Ausbildung angefangen.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactpos2	0.1	8.2	3.9	1.4

Situation: Er erzählt rückblickend, wie er seine Frau kennengelernt hat. “Kennengelernt haben wir uns, als wir beide zur Uni gegangen sind. Ich weiß noch, wie ich sie das erste Mal gesehen habe: sie war am anderen Ende eines total überfüllten Raums, und sie hatte diese wundervollen großen braunen Augen.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactpos3	5.2	8.1	4.5	1.7

Situation: Er erzählt, wie er sein Hobby zum Beruf gemacht hat. “Das war es, was ich schon immer machen wollte. Und genau das mache ich jetzt! Es ist echt eine große Sache, es macht mir richtig Spaß, und obwohl es viel Zeit kostet, es lohnt sich wirklich.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modactpos4	0.9	5.5	4.4	2.3

Situation: Ein Reporter kommt vom Auslandseinsatz in Asien zurück nach Hause und erzählt seinen Freunden. “Es war so ein toller Ort, wunderschön und ergreifend. Ich erinnere mich an die Überflutungen an Ostern, die waren wirklich schlimm. Aber wie die Leute sich gegenseitig bei den Rettungsaktionen geholfen haben, das hat mich echt beeindruckt.”

Situations intended as: Intense active positive

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactpos1	6.9	9.1	4.2	1.0

Situation: Er erfährt ganz unerwartet, dass er eine halbe Million Euro gewonnen hat. “Nein. Oh. Was? Das kann ich nicht glauben. Ehrlich? Was. Das wäre... das ist ja irre. Echt? Wahnsinn! Klasse! Du meine Güte!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactpos2	4.5	8.3	4.4	1.4

Situation: Er erzählt, wie es war, als er sie gefragt hat, ob sie ihn heiraten will, und sie ja gesagt hat. “Unglaublich aufregend. Es war wie Weihnachten, Geburtstag und Ostern, alles auf einmal.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactpos3	8.1	8.6	2.4	1.3

Situation: Er musste seit Wochen darum bangen, aus seiner Wohnung geworfen zu werden, und zwar heute. Gerade hat er einen Anruf bekommen, dass nun in letzter Sekunde eine Lösung gefunden wurde. “Wow! Ich muss nicht ausziehen! Ich weiß, wo ich heute Nacht schlafe! Mensch, das muss gefeiert werden! Bin ich erleichtert!”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intactpos4	7.2	6.9	4.6	3.2

Situation: Interview mit einem Bungee-Springer, der gerade den Sprung hinter sich hat. “Ich hab’s getan! Ich hab’s echt getan! Die ganze Zeit haben meine Freunde alle gesagt, du bist der einzige, der noch nicht gesprungen ist. Und ich bin gesprungen, kopfüber ins Wasser! Dein Kopf taucht richtig ins Wasser ein!”

Situations intended as: Moderate passive negative

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpasneg1	-0.7	-2.7	5.6	3.2

Situation: Er hat mit seiner früheren Frau eine erwachsene Tochter, die aber nichts von ihm wissen will. Er ist zwar jetzt wieder verheiratet, aber erzählt einem Freund: “Die ganze Zeit denke ich, was sie wohl mit ihrem Leben anfängt, ob sie wohl glücklich ist. Ich würde sie so gerne mal sehen – aber ich weiß, das wird nicht passieren. Also lebe ich halt mein Leben weiter und kümmere mich um meine neue Familie.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpasneg2	1.5	-6.7	5.1	2.3

Situation: Seine Frau hat sich von ihm getrennt und den kleinen Sohn mit sich nach Portugal genommen. Er erzählt: “Er lernt ja gerade sprechen. Ich habe mehrmals versucht, anzurufen und mit ihm zu reden. Aber mein Portugiesisch ist nicht so gut. Und er versteht nicht mal mehr ‘Papa’, er weiß gar nicht mehr, was ‘Papa’ heißt.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpasneg3	-4.3	-6.2	2.3	1.5

Situation: Er erzählt von Schwierigkeiten, die sie in der Ehe hatten. “Ich weiß, wir waren sehr jung, als wir geheiratet haben. Aber wir hatten alles im Leben, und ich konnte einfach nicht verstehen, warum sie so unglücklich war. Am Ende ging sie sogar zu einem Psychiater. Der gab ihr starke Medikamente, aber in ihrem Fall waren sie unwirksam, haben gar nichts genützt.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpasneg4	-0.2	-2.8	4.2	3.8

Situation: Ein Mann, der in einem Bürgerkriegsgebiet humanitäre Hilfe geleistet hat, berichtet. “Es gab so viele tragische Verluste. Wir haben mit Leuten geredet, die echt gelitten haben, und hoffentlich haben wir ein bisschen geholfen. Sie konnten auch mit uns reden, und glaub mir, das tut weh, wenn du jemanden verlierst, der dir nahesteht. Aber wenn du darüber reden kannst, der Schmerz geht nicht weg, aber es wird ein bisschen anders.”

Situations intended as: Intense passive negative

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpasneg1	-7.7	-8.7	3.7	1.8

Situation: Ein vereinsamter Mann berichtet über sein Leben. “Es ist furchtbar. Furchtbar. Ich kann nicht aus dem Haus gehen. Mich mit Leuten unterhalten. Ich denke, alle hassen mich. Also spreche ich erst gar nicht mit den Leuten, weil ich weiß, die werden sowieso nicht mit mir reden.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpasneg2	-4.6	-7.0	4.3	2.2

Situation: Seine Verlobte hat ihn gerade verlassen. “Sie hat Schluss gemacht, weil ich zuviel in der Kneipe gearbeitet habe, und jetzt ist sie weg. Ich bin krank geworden, total krank, und ich hab ewig nicht meine Familie gesehen. War es das wert? Ich glaube nicht. Nicht für die ganze Arbeit und die Sorgen, die ich durchgemacht habe.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpasneg3	-6.0	-9.0	4.3	1.4

Situation: Der Vater eines Mordopfers berichtet. “Ich war in der Küche, 250 Meter von da, wo meine Tochter ermordet wurde. Ich wusste das zu dem Zeitpunkt gar nicht. Dein Leben ist nie mehr wie vorher ab dem Tag.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpasneg4	-1.5	-7.9	5.0	1.8

Situation: Ein Vater hat es verboten bekommen, seinen Sohn zu sehen. “Einmal kam eine Weihnachtskarte durch. Das war reiner Zufall. Aber sonst ist es total unmöglich. Er ist da draußen und ich darf ihn nicht sehen. Das ist das Schlimmste, ich weiß wo er ist und ich darf nicht hin und ihn sehen.”

Situations intended as: Moderate passiv positive

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpaspos1	-3.2	7.5	4.8	1.7

Situation: Jemand wird gefragt, was das Bedeutsamste an der Liebe ist. “Ich glaube, das Wichtigste ist wohl diese Idee, dass Liebe bedingungslos ist. Das ist so in einer langjährigen Ehe. Dass die Leute das Schlimmste über einander wissen und immer noch totales Vertrauen in ihre Liebe haben.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpaspos2	-0.8	6.9	6.7	2.4

Situation: Ein Mann erzählt, wie er anfing, regelmäßig in die Kirche zu gehen. “Vor fünf Jahren bin ich ab und zu in den Gottesdienst in der Kathedrale gegangen. Am Anfang wollte ich nur der Musik zuhören, und die Musik ist natürlich absolut wundervoll. Aber es war nicht nur die Musik, warum ich immer wiederkommen wollte.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpaspos3	-0.3	8.6	6.3	1.6

Situation: Nach einer Zeit der inneren Einkehr hat er zu sich selbst gefunden. “Früher hatte ich das Gefühl, ich bin 99 Prozent blind und ein Prozent Peter. Das war ein beängstigender Zustand. Heute würde ich eigentlich sagen, ich fühle mich einfach 100 Prozent Peter.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
modpaspos4	-5.5	6.6	3.4	2.8

Situation: Ein passionierter Bergwanderer erzählt, was ihm an den Bergen so gefällt. “Die Berge sind eine wundervolle Gegend. Es gibt so eine Atmosphäre von Ruhe hier, von Frieden.”

Situations intended as: Intense passive positive

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpaspos1	-2.0	8.2	5.8	2.3

Situation: Ein Paar hat mit Hilfe der Heilsarmee die Tochter wiedergefunden, die vor mehreren Jahren spurlos verschwunden war. Viele vorhergehende Versuche, sie zu finden, waren vergeblich geblieben. Der Vater berichtet: “Ich persönlich glaube, dass meine Gebete geholfen haben, dass wir sie wiedergefunden haben. Ich glaube, Gott hört unsere Gebete. Ich glaube, er hat alle meine Gebete gehört, und er hat versprochen, unsere Gebete zu beantworten. Ich bin ihm so dankbar, dass er sie gefunden hat.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpaspos2	-0.8	7.8	5.5	1.2

Situation: Ein Mensch, der vor kurzem seinen Frieden mit sich selbst und mit Gott gemacht hat, erzählt, was ihm Weihnachten bedeutet. “Advent ist eine großartige Zeit der Vorbereitung, wo wir uns auf eins der wunderbarsten Geschenke vorbereiten: Das Kommen von Jesus. Und für mich ist dieses Weihnachten so schön wie vielleicht noch nie. Das liegt daran, dass mein Herz nicht mehr so zerrissen ist.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpaspos3	-2.8	5.9	4.9	1.9

Situation: Ein Dichter beschreibt, was für ihn die Messe bedeutet. “Es beeindruckt mich, wie die Messe abgehalten wird, das Formelle und die Stille. Das Ganze verbindet sich in einer Weise, die für mich sehr schön und bewegend ist. Es findet sogar ein bisschen seinen Weg in meine Gedichte.”

Name	Act.	Eval.	s.d. Act.	s.d. Eval.
intpaspos4	-1.7	6.1	5.1	2.0

Situation: Ein gläubiger Mann, der auch schwere Zeiten hinter sich hat, berichtet über seine Erfahrungen. “Es entsteht so ein Frieden, wenn man sich an Gott wendet. Schließlich ist er der Einzige, wenn es drauf ankommt. Es ist gar nicht wichtig, welcher Religion man angehört. Er gibt einem Kraft, und wenn er eine Tür zumacht, dann öffnet er ein Fenster.”