The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching

Marc Schröder

DFKI GmbH, Saarbrücken, Germany
schroed@dfki.de

Jürgen Trouvain

Institute of Phonetics, University of the Saarland
trouvain@coli.uni-sb.de

http://mary.dfki.de

Abstract

The German Text-to-Speech Synthesis system MARY is presented. An interface allowing to access and modify intermediate processing steps without the need for a technical understanding of the system is described, along with examples of how this interface can be put to use in research, development and teaching.

1. Introduction

This article presents the German text-to-speech system MARY (**M**odular **A**rchitecture for **R**esearch on speech s**Y**nthesis) as a tool for research, development and teaching in the domain of text-to-speech synthesis. It is aimed both at readers who have little experience with the internal workings of a text-to-speech (TtS) system and at specialists who want to know about the particularities of the MARY system.

MARY allows 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 [1], also for German. However, apart from displaying the intermediate processing results, our system also allows their modification by the user. Thereby, the user is given the 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 [2]. These features are also found in FESTIVAL [3], 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 [4][5]. 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 want to get into the technical details of the system.

The article is structured as follows. First, a detailed account of the system structure is given, including a short presentation of each module. After that, the user interface is described which allows to display and edit intermediate processing results. Finally, examples are given to show the use of such an interface for teaching, TtS development, and research.

2. Structure of the TtS System

The architecture of the MARY TtS system is similar to a typical TtS architecture as described by Dutoit [6]. Figure 1 shows the individual processing modules, the flow of information and intermediate results.

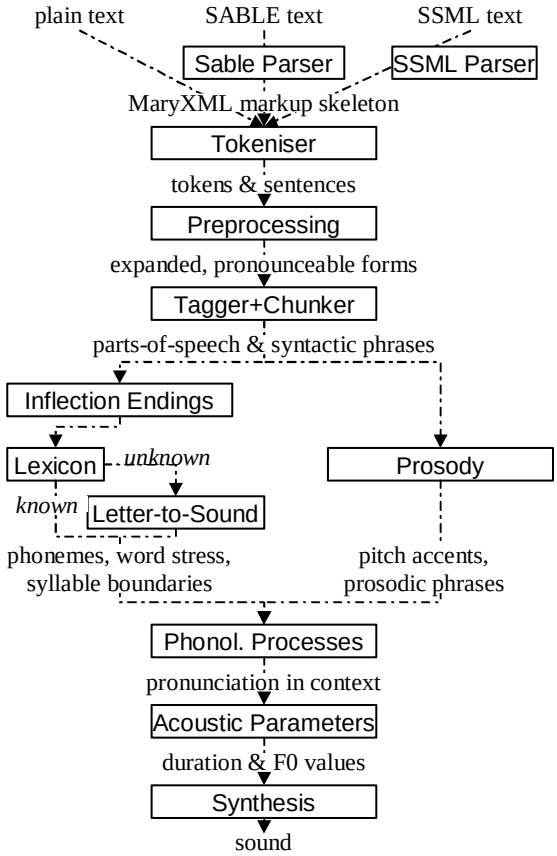

Figure 1. The architecture of the MARY TtS system.

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

2.1. 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 markup language such as SABLE.

The input markup language, presently SABLE, is translated into an internal, lower-level markup that we call MaryXML, which reflects the modelling capabilities of this particular TtS system. MaryXML is based on XML (eXtensible Markup Language) [7]. A DTD (Document Type Definition) formally specifies the structure of a correct MaryXML document.

As an example, an <EMPH>...</EMPH> SABLE tag requesting moderate emphasis for the enclosed words is translated into a raised F0 level, reduced speed, and an obligatory pitch accent for every enclosed word. These prosodic settings are meant to realise the abstract concept of emphasis. They 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 or concretises information. E.g., 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 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 [7]. This technique allows a very simple adaptation to new markup languages such as the upcoming W3C speech synthesis markup language [8], as only the stylesheet defining the translations will need to be adapted.

2.2. Tokeniser

The tokeniser cuts the text into tokens, i.e. words and punctuation marks. It uses a set of rules determined through corpus analysis to label the meaning of dots based on the surrounding context. 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, the punctuation is used to determine start and end of sentences which are marked using the MaryXML ... tag enclosing a sentence.

2.3. Preprocessing

In the preprocessing module, those tokens for which the spoken form does not entirely correspond to the written form are replaced by a more pronounceable form[1].

2.3.1. 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 the ordinal belongs to. In the preprocessing 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" ("firstly") 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 2.5.1)[2].

2.3.2. 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 appreviations are correctly pronounced by spelling rules.

The second group is pronounced using an expansion table, containing a graphemic and optionally a phonemic expansion. The latter is expecially useful for foreign abbreviations, such as "FBI" which is pronounced as the English spelling [ɛ f-bi: -'ʔ aɪ] in German.

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 2.5.1).

2.4. Part-of-speech tagger / chunk parser

Part-of-speech tagging is performed with the statistical tagger TnT [10], using the Stuttgart-Tübingen Tagset (STTS) [11], and trained on the manually annotated NEGRA corpus [12]. A chunk parser [13] is used to determine the boundaries of noun phrases, prepositional phrases and adjective phrases.

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 presently used encoding with the XML structure representing prosodic structure and syntactic structure "squeezed" into the token tags seems to be a viable solution.

2.5. Phonemisation

The SAMPA phonetic alphabet for German [14] 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.

[1] An excellent overview of the phenomena that need to be accounted for in German preprocessing has been given by Breitenbücher [5].

[2] A different solution for this problem, employing a sentence grammar, is used in the SVOX system [9].

2.5.1. *Inflection endings*

This module deals with the ordinals and abbreviations which have been marked during preprocessing (see 2.3) 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 [15].

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 to determine the appropriate values of gender, number and case for the ordinal or abbreviation and thus the correct ending, which is added to the expanded form.

For example, in "mein 2. Angebot" ("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. 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".

2.5.2. *Lexicon*

The pronunciation lexicon is derived from CELEX [16]. It contains the graphemic form, a phonemic transcription, a special marking for adjectives, and the inflection information mentioned above (see 2.5.1).

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. An optional "+s+" bounding morph, typical for German noun compounds, is 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.

2.5.3. *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 to be neighbours. This had been trained on data extracted from CELEX [16]. The resulting morpheme chain is compared to a list of affixes which have a predictable effect on word stress position, either attracting the stress or shifting the stress, or with no effect on stress [17].

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

The syllabification of the transcribed morphemes is based on standard phonological principles such as the sonority hierarchy of phonemes, the maximal onset principle, the obligatory coda principle and the phonotactic restrictions for the German language (see also [18]).

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

2.6. Prosody rules

Prosody is modelled using GToBI [19], an adaptation of ToBI ("Tones and Break Indices") for German. ToBI 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 2.8), these are translated into concrete F0 targets and pause durations.

The prosody rules were derived through corpus analysis and are mostly based on part-of-speech and punctuation information. 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 aptitude to receive 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 [20]. 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.

2.7. 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 [21].

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.

2.8. Calculation of acoustic parameters

This module performs the translation from the symbolic to the physical 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 [22] adapted to German [18]. A classification and regression tree (CART) trained on a corpus of German read speech [18] will replace that module.

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 [ʼfʊn] of the sequence [gə-ʼfʊn-dən] ("found") is realised as a target on the baseline at the start of the Schwa of [gə], followed by a target on the topline in the middle of the [ʊ] in [ʼfʊn]. Obviously, the actual frequency values of the topline and baseline 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.

2.9. Synthesis

At present, MBROLA [23] is used for synthesising the utterance based on the output of the preceding module. The diphone sets of two German MBROLA voices (one male, one female) are presently 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.

3. The interface to partial processing results

An interface allows to comfortably traverse only parts of the MARY architecture tree (see Figure 1). Besides plain text and SABLE-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. In the future, the XSLT stylesheet performing that translation is to be made editable from within the interface, allowing the experimentation with realisation strategies for SABLE 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 2 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. As a well-formed and valid XML document, it contains some header information, 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 [7]).

Output of a given type can be obtained by simply selecting the desired output format (in this case, the output of the prosody module) 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.

4. Use of the interface

This section gives examples demonstrating the usefulness of the MARY interface in the domains of teaching, development and research, respectively.

4.1. Teaching

The interface allows students to explore the workings of the individual modules in the TtS system. This can be done as a presentation performed by a teacher or interactively by the students themselves.

In order to disentangle the various components of a TtS system, it is helpful for a first insight that the students walk through the individual modules from the very beginning to the very end. After each module, they can see the information that this module has added.

As an example, the screenshot in Figure 2 shows an intermediate step in the processing of the sentence "Ich fliege nach Schottland" (lit. "I fly to Scotland"). What is visible in Figure 2 is input data of type "MaryXML tagged" on the left-hand side and output data of type "MaryXML intonation" on the right-hand side. They represent tagger output and prosody module output, respectively. As can be seen in the MARY system architecture (Figure 1), only the prosody module is needed to perform that transformation. In order to show the information added by this processing step, the differences between input and output are highlighted. In this case, the added information represents the beginning and end of intonation phrases, the location of prosodic boundaries with their strengths, as well as the location and type of pitch accents and boundary tones.

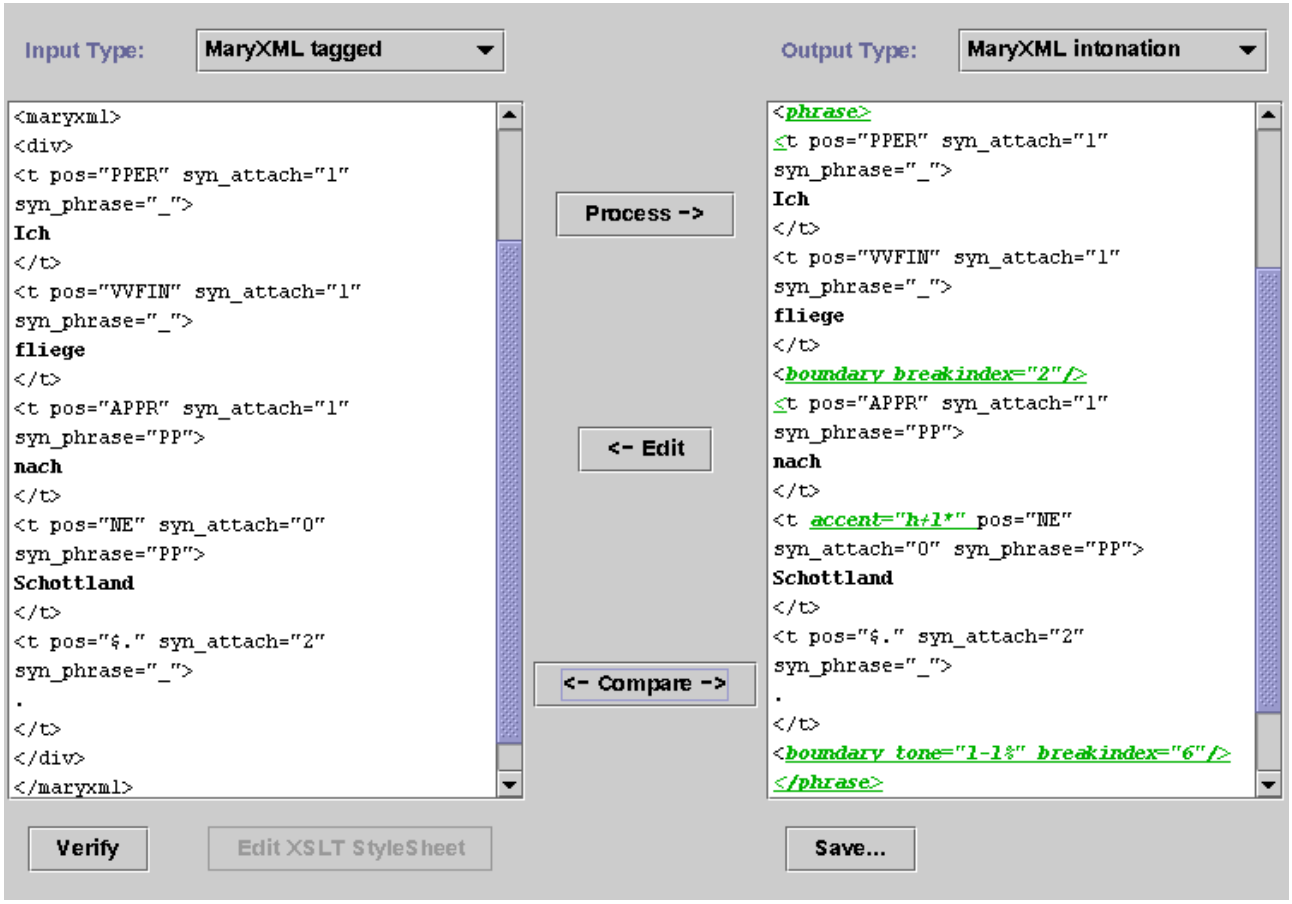

Figure 2. Example of partial processing with the MARY interface. See text (3.) for explanations.

More advanced students can explore the functioning of a particular module in more detail by modifying specific pieces of information in that module's input and observe the changes in the output. In the example of Figure 2, the effect of changing a token's part-of-speech on accenting can be observed by changing, e.g., the part-of-speech of the token "fliege" from VVFIN (finite full verb) to NN (noun).

4.2. Development

A possible development task could lie in the domain of speech synthesis markup realisation, i.e. the interpretation of high-level markup (e.g. SABLE) in terms of lower level internal MaryXML markup. As an example, one might be interested in an appropriate rendering of "strong" emphasis, which can be expressed in SABLE using the tag `<EMPH LEVEL="strong">`.

Take again the sentence "Ich fliege nach Schottland" with an `<EMPH>` tag around "fliege" in order to emphasise the fact that one is flying and not driving to Scotland. The most obvious realisation is to give the originally unaccented "fliege" a pitch accent. Another possibility could be to look beyond the portion of the utterance to be emphasised and to de-accent tokens after this portion within the given intonation phrase. But what about giving particular emphasis to the already accented "Schottland", e.g. in the context of contrastive focus ("You fly to England?" – "I fly to *Scotland*.")? Prosodic parameters can be modified to that end, such as an increase of F0 level and/or range, as well as lengthening of sound segments. In addition, increased articulation precision may be useful in some cases[3].

All these changes can be requested by using appropriate MaryXML tags.

4.3. Research

Speech synthesis allows the controlled creation of stimuli for perception experiments, be it for applied research (system improvement) or basic research (knowledge increase). The MaryXML markup makes the linguistic units used at any stage of processing accessible. Researchers wanting to modify these units can do this in a controlled way.

For example, Brinckmann & Trouvain [18] compared the perceptual relevance of the symbolic string (phonemes, word stress, pitch accent, prosodic phrase boundaries) in combination with two different models predicting segment durations. An outcome of that study is that modelling the symbolic string seems to be more important for perceptual preference than modelling the duration predictor.

An example for basic research is given by Baumann & Trouvain [26]. For a perception test with read telephone numbers, they created stimuli varying in pitch accent and pause structure. The findings of this study supported the idea that strategies for reading telephone numbers found in human

[3] More extreme forms of emphasis can be observed in hypercorrections, e.g. after misunderstandings in spoken man-machine communication, see e.g. [24][25].

speech production are preferred over strategies currently employed in telephone inquiry systems.

The advantage of the MARY web interface is that it delivers a comfortable way of preparing the stimuli for such perception tests from everywhere with no need to install the system locally.

5. Summary

An overview of the processing components of the German Text-to-Speech Module MARY has been given. It has been attempted to give a rough idea of how an XML representation can be used to make partial processing results accessible and editable, and the benefits of that possibility for TtS development, teaching and research have been sketched.

Acknowledgements

Many persons have contributed to the development of the MARY system over the years. We wish to thank, in alphabetical order, Ralf Benzmüller, Caren Brinckmann, Martine Grice, Tilman Jaeger, Kerstin Klöckner, Brigitte Krenn, Annette Preissner, Diana Raileanu, and Henrik Theiling. We would also like to thank Markus Becker and Bettina Braun for their comments on an earlier draft of this paper.

6. References

[1] Hoffmann, R., Kordon, U., Kürbis, S., Ketzmerick, B., & Fellbaum, K. (1999). An Interactive Course on Speech Synthesis. *Proc. ESCA/SOCRATES Workshop MATISSE*, p. 61-64.

[2] Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K. & Edgington, M. (1998). SABLE: A Standard for TTS Markup. *Proc. ICSLP Sydney*, p. 1719-1724.

[3] http://www.cstr.ed.ac.uk/projects/festival

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

[5] Breitenbücher, M. (1999). Textvorverarbeitung zur deutschen Version des Festival Text-to-Speech Synthese Systems [Text preprocessing for the German version of the Festival Text-to-Speech synthesis system]. http://elib.uni-stuttgart.de/opus/volltexte/1999/225

[6] Dutoit, T. (1997). An Introduction to Text-to-Speech Synthesis. Dordrecht: Kluwer.

[7] Harold, E. R. (1999). *XML Bible*. Hungry Minds, Inc. http://www.ibiblio.org/xml/books/bible

[8] W3C Speech Synthesis Markup Language Specification. http://www.w3.org/TR/speech-synthesis

[9] Traber, Ch. (1993). Syntactic Processing and Prosody Control in the SVOX TTS System for German. *Proc. Eurospeech Berlin*, p. 2099-2102.

[10] Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. *Proc. 6th Applied Natural Language Processing Conference*, Seattle, WA, USA. http://www.coli.uni-sb.de/~thorsten/publications

[11] Schiller, A., Teufel, S. & 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

[12] Skut, W., Krenn, B., Brants, T. & Uszkoreit, H. (1997), An Annotation Scheme for Free Word Order Languages. *Proc. 5th Conf. Applied Natural Language Processing*, Washington D. C. http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html.

[13] Skut, W. & Brants, T. (1998). Chunk Tagger – Statistical Recognition of Noun Phrases. *Proceedings of the ESSLLI Workshop on Automated Acquisition of Syntax and Parsing, Saarbrücken, Germany.* http://www.coli.uni-sb.de/~thorsten/publications

[14] SAMPA Phonetic Alphabet for German. http://www.phon.ucl.ac.uk/home/sampa/german.html

[15] Petitpierre, D. & Russell, G. (1995). MMORPH – The Multext Morphology Program. *MULTEXT deliverable report,* ftp://issco-ftp.unige.ch/pub/multext/mmorph.doc.ps.tar.gz

[16] R. H. Baayen, R. Piepenbrock & L. Gulikers (1995), The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[17] Jessen, M. (1999). German. In *Word Prosodic Systems in the Languages of Europe* (H. van der Hulst, ed.). Berlin & New York: Mouton de Gruyter. p. 515-545.

[18] Brinckmann, C. & Trouvain, J. (2001). The Role of Duration Prediction and Symbolic Respresentation for the Evaluation of Synthetic Speech. *This volume.*

[19] Grice, M., Baumann, S. & Benzmüller, R. (to appear). German ToBI. In *Prosodic Typology and Transcription: A Unified Approach* (S.-A. Jun, ed.).

[20] Benzmüller, R. Grice, M. (1997): Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI. Research Report Phonetics Saarbrücken *Phonus* 3, p. 9-34.

[21] Trouvain, J. & Grice, M. (1999). The Effect of Tempo on Prosodic Structure. *Proc. 14th ICPhS San Francisco*, p. 1067-1070.

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

[23] Dutoît, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High Quality Speech Synthesisers Free of Use for Non Commercial Purposes. *Proc. ICSLP Philadelphia*, p. 1393-1396.

[24] Fischer, K. (1999). Discourse Effects on the Prosodic Properties of Repetitions in Human-Computer Interaction. *ESCA Workshop on Dialogue and Prosody Eindhoven*, p. 123-128.

[25] Pirker, H. & Loderer, G. (1999). I Said "TWO TICKETS": How to Talk to a Deaf Wizard. *ESCA Workshop on Dialogue and Prosody Eindhoven*, p. 181-186.

[26] Baumann, S. & Trouvain, J. (2001). On the Prosody of German Telephone Numbers. *Proc. Eurospeech 2001, Aalborg,* Vol 1, pp. 557-560.