Synthesis of Nonverbal Listener Vocalizations

Sathish Pammi
DFKI GmbH, Saarbrücken, Germany
Sathish.Pammi@dfki.de

Abstract

Listener vocalizations play an important role in communicating listener intentions while the interlocutor is talking. Synthesis of listener vocalizations is one of the focused research areas to improve emotionally colored conversational speech synthesis. The major objective of the work presented in this paper is providing a new functionality to text-to-speech synthesis system that can synthesize nonverbal listener vocalizations. As synthesis of listener vocalizations is a new topic in conversational speech synthesis, many research questions are raised. A methodology is proposed to conduct research on those questions which can provide solutions to build a system to generate nonverbal listener vocalizations. We discuss the work done so far according to proposed working strategy and tentative plans for future work.

Keywords: nonverbal listener vocalizations, back-channel, multi-modal interaction, speech synthesis

1 INTRODUCTION

In multimodal human-computer interaction, the ability of systems to generate listener vocalizations (Gardner, 2002) is an important requirement for generating affective interaction.

Listener vocalizations include back-channel utterances (Yngve, 1970; Ward and Tsukahara, 2000) related to the flow of the conversation as well as affect vocalizations (Schröder et al., 2006) based on the listener's affective state (Scherer, 2003). For example, nonverbal listener vocalizations like *mm-hm* or *uh-huh* can be used as back-channel utterances to keep the floor open for the current speaker to continue speaking. Listener vocalizations can also transmit affective states like excited, bored, confused, surprised, etc. For example, *wow* can be used for both back-channel and to communicate affective meaning. Listener vocalizations also include non-linguistic vocalizations like laughter or sigh as well as some response tokens like *yes*, *right*, *really* or *absolutely*.

Nowadays, speech synthesis systems are providing high quality synthetic reading speech. Synthesis of nonverbal listener vocalizations, a new functionality to text-to-speech synthesis systems, provides an opportunity to build interactive synthesis systems suitable to multi-modal interaction systems. Database collection, annotation and realization of speech waveform are crucial steps in building speech synthesis systems. Above three major steps need more investigation in case of the new functionality. For example, traditional speech synthesis databases including expressive speech material were recorded in a studio environment with a single speaker using predefined recording scripts, but this traditional recording setup is not suitable to capture listener vocalizations as they are natural only in a conversation. Success in generation of listener vocalizations depends on the answers to the following questions:

- How to collect a database of listener vocalizations?
- What kinds of meanings are expressed through listener vocalizations?
- What form is suitable for a given meaning?
- How to annotate meaning and behavior (form) of a listener vocalization?
- How to realize the form using a technological framework?

Many listener vocalizations are short and nonverbal in nature. As synthesis of nonverbal vocalizations is a new topic in synthesis, we are not aware of any technological framework to synthesize these vocalizations. In the level of realization, some technological research questions should be answered like:

- What kind of technology is suitable to synthesize nonverbal vocalizations? Unit-selection, HMM-based or other.

- If it is Unit-selection, what strategy would be better to select a unit?
- If it is HMM-based, how to model and realize nonverbal vocalizations?
- How to get advantage from signal modification algorithms?

The major objective of this work is not only providing answers to the above research questions, but also building a system, which will be integrated into SEMAINE (SEMAINE, 2008) multi-modal interaction system, to synthesize nonverbal listener vocalizations. The system has to be robust and it has to use standard representation like eXtensible Markup Language (XML) formats in the view of future inter-module communication. A possibility is there to raise more research questions when we try to evaluate our final system as part of a real-time SEMAINE demonstration system.

A methodology is proposed in Section 2 to conduct research on synthesis of nonverbal vocalizations. We describe the results of the data collection and annotation in Section 3 and this section also explains our baseline system. In Section 4, we discuss our tentative plans and proposals to build a system for realization of listener vocalizations in a speech synthesis framework.

2 METHODOLOGY

The SEMAINE system, a demonstration of audiovisual Sensitive Artificial Listener(SAL) (Douglas-Cowie et al., 2008), aims to build a virtual dialog partner who intends to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. Different 'action proposers' in the system produces different 'action commands' to synthesize a meaningful agent behavior. Simulation of a convincing audiovisual listener behavior is one major part of the system. According to the project plans, an action proposer, with the help of multi-modal inputs, will be planning the intention of the listener as well as the timing information to trigger the behavior. The description of listener intention uses standard XML representation ('Multi-modal XML input' in the Figure 2). Our part of the work is mainly focusing on modules for synthesis of appropriate listener vocalizations when the intended meaning behind the listener intention is given.

This section describes conceptual model of our proposed methodology to build a framework for synthesis of nonverbal vocalizations. The proposed work consists of three different levels (as shown in Figure 1): Data collection, Annotation and Realization.

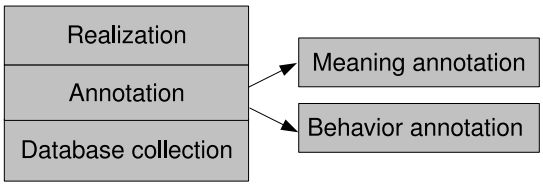

Figure 1: Major aspects of proposed work

2.1 DATABASE COLLECTION

As the traditional way of recording setup is not useful to capture nonverbal listener vocalizations, we propose to record a natural dialog speech between an actor and his dialog partner in an anechoic studio because listener vocalizations seem to be natural only in a conversation. According to the new proposed recording setup, the actor and his dialog partner will sit in different rooms and hear each other using headphones, so that we can record each speaker's voice on a different channel without interference of the other speaker's speech. As we are aiming to capture listener vocalizations, the actor will be instructed to participate in a free dialog, but to take predominantly a listener role.

2.2 ANNOTATION

To know different kinds of meanings expressed through listener vocalizations, the intended meaning behind each vocalization should be annotated. Similarly, the annotation of behavioral properties will be useful to know suitable behavior for a given meaning. Initially, we do not know how many meaning or behavior categories can be used to annotate all listener vocalizations, so we propose to annotate all nonverbals using informal descriptions to make sure that we are not guided by any pre-existing set of categories. Pre-existing

sets of categories may or may not be suitable to represent all listener vocalizations available in our data. So informal descriptions will be helpful to understand better the structure of both behavior and meaning. Subsequent grouping of these descriptions will help to understand the types of behavior and meaning of listener vocalizations, at least for the speaker we studied. In the later stages, a suitable limited set of categories that capture the essence of meaning as recorded in informal descriptions will be identified.

The sequence of steps involved in the proposed annotation scheme is: Firstly, start-end time labels will be annotated for all listener vocalizations made by the actor. Secondly, informal descriptions will be provided for each labeled segment in three different levels: content, behavior, sub-texts. In latter stages, suitable meaning category will be identified for each vocalization with the help of informal descriptions. Finally, annotation for behavioral properties like intonation, voice quality etc.. will be provided.

2.3 REALIZATION

The conceptual model for the realization system, as shown in Figure 2, contains off-line and runtime processing modules. Data analysis on annotated speech samples is a crucial step in off-line processing which provides relations between behavior and meaning. The experience from this analysis will let us know whether the relation between meaning and behavior is one-one mapping or a single behavior can be usable to simulate multiple intended meanings. A thorough research is expected in the level of technological framework to realize a nonverbal listener vocalization. For example, we have to find a way to model and generate nonverbal vocalizations if we choose Hidden Markov Model (HMM) based synthesis as a technological framework.

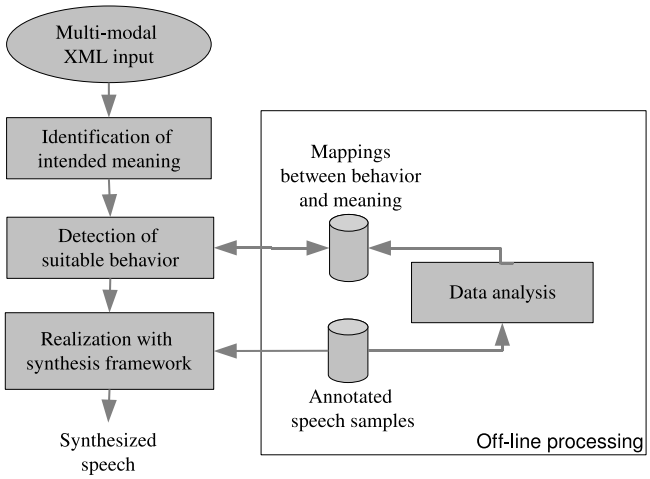

Figure 2: Conceptual model of proposed realization system

The proposed runtime system will work as follows: Initially, an XML front-end processing module will identify the intended meaning behind requested nonverbal vocalizations. The next module will be finding suitable behavior to the requested meaning category with the knowledge of relations between behavior and meaning. Finally, another module will realize appropriate behavior with a synthesis technology like Unit-selection or HMM-based.

3 RESULTS SO FAR

The work has been progressing on all three levels described in Section 2. This section explains the results of the work done so far.

3.1 DATABASE COLLECTION

We recorded dialog speech in a studio environment as described in Section 2.1. Our speaker is a professional male German actor with whom we had already recorded expressive speech synthesis databases in the past. Using this speaker was essential for being able to use the recorded vocalizations with our synthesis

voices in the future. Recordings were made in several stages and in sessions of about 20 minutes each. In the initial stage, we instructed the actor to "be himself" (not to act) and in the later stages, he was instructed to act like one of three characters representing different emotionally colored personalities (Douglas-Cowie et al., 2008): Spike is always aggressive, Obadiah is always gloomy, Poppy is always happy. Two female student assistants took turns as the dialog partner, and tried to keep the actor in listening mode for a maximum amount of time while they were talking to the actor about a topic of their choice. The dialogue partners were sitting in separate rooms and hearing each other using headphones. Each speaker's voice was recorded on a separate channel.

As a result of the database collection exercise, we obtained around six hours of German dialog speech. Table 1 provides statistics of dialog speech material.

The actor status	Corpus duration (in minutes)	Number of listener vocalizations
Natural	190	568
Obadiah	45	181
Poppy	45	93
Spike	70	238
Total	350	1080

Table 1: Corpus duration in minutes when the actor is being himself (natural) or acted like an emotional character.

3.2 ANNOTATION

So far in this project, we have worked on meaning annotation only. A detailed version of results in meaning annotation were reported in (Pammi and Schröder, 2009), but an overview of those results were shortly discussed in this section. As outlined in Section 2.2, so far in this work, we have worked on informal description and meaning annotation only.

3.2.1 INFORMAL DESCRIPTIONS

In order to get a fuller picture of the data, we use a detailed informal description of each vocalization before trying to find suitable categories to represent the meaning and behavior observed. An informal description in this work contains an annotator's description of the form, content and subtext of each listener vocalization using his/her own vocabulary. The form provides information about phonetic segments, voice quality, duration and/or intonation. Similarly, the content and subtext tiers describe the meaning and, optionally, a suitable text substitution.

3.2.2 INSTRUCTIONS GIVEN TO ANNOTATE MEANING

We used the Baron-Cohen (Baron-Cohen et al., 2004) set of 33 categories describing epistemic-affective states as a starting point for our tag set. Annotators were instructed to use only those categories from the set that seemed appropriate, and to add categories that seemed necessary to describe the data but were not contained in the Baron-Cohen set. They could use categories from the Geneva Emotion Wheel (Scherer, 2005) or propose their own category labels as they felt appropriate.

According to informal descriptions provided from annotators, listener vocalizations seem to differ with respect to their reference: self expression, stance towards the other, attitude towards the topic. Bühler's (Bühler, 1934) Organon model provides a structure that distinguishes the above three types. So, we instructed annotators that they could optionally indicate the reference according to the Organon model: (S)elf reference, (O)ther reference, or (T)opic reference.

3.2.3 RESULTS OF MEANING ANNOTATION

Annotators used 24 out of the 33 Baron-Cohen categories to annotate meaning. They added nine out of the 40 categories of the emotion wheel (Geneva, 2005), as well as four custom categories. The 37 categories used are shown in Table 2. The number of frequently used categories is much smaller, though.

Baron-Cohen categories	**anticipating**, cautious, concerned, confident, contemplative, decisive, defiant, **despondent**, **doubtful**, **friendly**, hostile, insisting, **interested**, nervous, playful, preoccupied, regretful, serious, suspicious, **tentative**, **thoughtful**, uneasy, upset, worried
Emotion wheel categories	**amused**, angry, compassionate, disgusted, happy, **irritated**, relieved, **scornful**, **surprised**
Custom categories	depressed, excited, ironic, outraged

Table 2: The list of categories used for annotation. Frequently used categories ($> 5\%$) are highlighted in bold, and most frequent categories ($> 10\%$) are underlined. (Pammi and Schröder, 2009)

The full descriptions of meaning are summarized in terms of meaning categories associated with types of functional reference. The results show that Baron-Cohen's affective-epistemic categories are not sufficient to describe our data – it is necessary to add a number of categories from the Geneva Emotion Wheel as well as some custom categories. The results from reference annotation according to Bühler's Organon model suggest that distinguishing the reference in addition to affective-epistemic meaning categories is a useful means to gain insights regarding a character's mood or personality (Self reference), interpersonal stance (Other reference) and attitude towards a topic (Topic reference).

A subset of 102 listener vocalizations from the non-acted part of the dialog corpus was annotated by both annotators with meaning and reference categories for inter-rater agreement. As described in (Pammi and Schröder, 2009), we computed Kappa for each meaning category and each reference type. The Kappa values for the most frequently used meaning categories friendly, interested and amused were 0.02, 0.41 and 0.82 respectively. Among the less frequent categories, Kappa values for decisive, confident, tentative, doubtful and surprised scores range between 0.22 and 0.43, whereas anticipating, thoughtful, ironic, irritated, outraged, angry show nearly no agreement between two annotators. For reference categories, there is no consistent agreement between the two annotators. It remains to be seen whether this is due to an intrinsic ambiguity or due to insufficient instructions.

3.3 REALIZATION

A base-line system was implemented in MARY (Schröder and Trouvain, 2003; Schröder et al., 2008) Text-To-Speech(TTS) framework for synthesis of nonverbal listener vocalizations. This simple system can generate nonverbal listener vocalizations based on an XML request. It stores all nonverbal listener vocalizations in the form of datagrams in a single time-line waveform file and a corresponding unit file containing index numbers and start-end timestamps of each vocalization to retrieve efficiently. We can request a nonverbal vocalization with or without index number. When the XML request does not have an index number then the system will select any one among the vocalizations existing in the database. The baseline system was integrated to the first version of the open source SEMAINE (Schröder and et al., 2008) demonstration system for generating back-channel vocalizations when requested.

An example XML request:

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml xmlns="http://mary.dfki.de/2002/MaryXML" version="0.4" xml:lang="de">
    <voice name="spike">
        <nvv variant="6"/>
    </voice>
</maryxml>
```

4 FUTURE WORK

So far the results of the work related to database collection and meaning annotation were described. This section proposes our plans for behavior annotation, realization strategies and evaluation.

4.1 ANNOTATION OF BEHAVIOR

Behavior annotation is one of the crucial tasks as this part of the work directs the way to surface level realization of nonverbal listener vocalization.

The following elements are expected in the behavior annotation:

1. A representation of intonation
2. A suitable phonetic segmental form in alignment with the waveform
3. Aspects of volume, para-language and voice quality

The intonation of a nonverbal vocalization can be extracted automatically from any pitch tracking algorithm available in computer programs like Praat (Boersma and Weenink, 2005) and can be stored as a set of points of the pitch contour or a set of polynomial coefficients which can represent the pitch contour of the nonverbal vocalization. A suitable phonetic segmental form of a nonverbal vocalization in alignment with the waveform should be annotated manually as we do not have any immediate procedure to do that automatically. The phonetic segmental form is useful for lip synchronization of the visual synthesis system, when we integrate with audiovisual synthesis system like GRETA (Poggi et al., 2005). A suitable set of descriptors should be identified to annotate aspects of volume, para-language and voice quality. A pilot study (Douglas-Cowie et al., 2003) was conducted on the Belfast naturalistic database (Douglas-Cowie et al., 2003) for the description of naturally occurring emotional speech. The descriptors, as shown in Table 3, provided from the study will be a starting point to annotate aspects of volume, para-language and voice quality.

Para-language Descriptors	Laughter, Sobbing, Break in Voice, Tremulous Voice, Gasp, Sigh, Exhalation and Scream
Voice Quality Descriptors	Creak, Whisper, Breathy, Tension and Laxness
Volume Descriptors	Raised Volume, Lowered Volume and Excessive Stressing

Table 3: A set of descriptors which are considered to be strongly indicative of emotion (Douglas-Cowie et al., 2003)

4.2 Relation between meaning and behavior

The system has to identify a suitable behavior for surface-level realization whenever the multi-modal interaction system requests a nonverbal vocalization with an intended meaning. In order to provide this functionality, we must carry out research on the relation between the meaning and the behavior of nonverbal vocalizations. The data analysis on annotated samples might provide an answer to the question whether the relation between meaning and behavior is one-one mapping pattern or any other. If the relation is having one-one mapping pattern, a simple lookup table will be able to find an appropriate behavior.

4.3 Realization with different technologies

This section outlines our plans regarding the technological realization of nonverbal vocalizations. Nowadays Unit-selection (Hunt and Black, 1996) and HMM-based (Tokuda et al., 2000; Black et al., 2007) speech synthesis technologies are the most popular. The corpus-based unit selection approach can produce near-natural high quality speech; it simply relies on runtime selection and concatenation of units from a speech database using explicit matching criteria. HMM-based speech synthesis provides an efficient model-based parametric method for speech synthesis that is based on a statistical framework of HMMs. In the scope of this work, we propose to perform experiments with both technologies to identify the pros and cons of the different technologies for the task at hand.

4.3.1 Unit-selection speech synthesis

In general, the selection of a unit at runtime is a crucial task in unit-selection synthesis framework. In MARY TTS, the unit can be a diphone or a half-phone. But here the unit is a nonverbal listener vocalization. One challenge in this framework is to find a way to choose a nonverbal vocalization with the help of behavioral properties identified from the mappings between meaning and behavior.

We can propose two possible solutions regarding the selection of a unit: One possibility could be finding a suitable nonverbal vocalization with appropriate behavior descriptors using explicit matching criteria. Another possibility could be training a classification tree to find the index of a nonverbal vocalization with a given set of behavioral properties. In the latter case, it is possible to choose a vocalization with closest but not exact behavior. Signal modification algorithms may be useful to realize exact behavior.

4.3.2 HMM-BASED SPEECH SYNTHESIS

We do not yet have a clear view regarding the realization of nonverbal vocalizations in the HMM-based synthesis framework. A simple starting point would be a copy-synthesis mechanism using the MLSA (Mel Log Spectrum Approximation) filter (Tokuda et al., 2002), which would have to support external prosody specification. The sequence of steps involved in the simple proposal system is: 1. Extract Mel Frequency Cepstral Coefficients (MFCCs) of each vocalization and store them as one of the behavior properties. 2. Re-synthesize the vocalization using the MLSA filter with external prosody specification according to requested behavior.

5 EVALUATION

The system will be implemented with based on the annotations of form and meaning described above and it will use all nonverbal listener vocalizations available from the dialog speech corpus. The evaluation of the system is perhaps the most significant challenge. One major objective of the system is the generation of nonverbal listener vocalizations that support effective human-computer interaction. Therefore, a subjective evaluation of the dialog system with and without support for generation of nonverbal listener vocalizations would be a promising strategy.

6 DISCUSSION AND CONCLUSION

The term 'nonverbal vocalizations' does not quite describe the types of vocalizations that this work aims to cover. Not all of them are nonverbal. The listener responses like *yes*, *absolutely*, *really*, etc. are actual words that can be found in a dictionary. Several terms are considered to describe the types of vocalizations, but we did not find such single and appropriate one. For example, if the term 'epistemic vocalizations' is taken, the term does not describe continuers' like *mhm* or *uh-huh* since no epistemic stance seems to be involved. So the term 'nonverbal' is a place holder for the moment. However, finding a proper term that describe the types of vocalizations in the scope of this work is an open issue.

An emotionally colored conversational synthesis system is required to synthesize not only listener nonverbal vocalizations, but also speaker's nonverbal vocalizations with it's context speech. For example, sentences like '*Oh! My dear daddy*' and '*Wow! It is wonderful*'. Though the topic of speaker's nonverbal vocalizations is not relevant to the discussion so far, the annotation and technological realization strategies are expected to be same as we discussed in this paper. But this topic raises another interesting question, namely how to realize behavior of nonverbal vocalization which matches the context speech. For example, do we see any similar patterns of behavior in a nonverbal vocalization (*ex: Wow!*) and it's context speech (*ex: It is wonderful*)?. We have not yet confirmed whether the dialog speech recorded with the strategy used for data collection provides sufficient coverage of speaker's nonverbal vocalizations as we do not have annotation for them. However, we will be able to extend this work to synthesize all kinds of non verbal vocalizations if there is no data coverage problem regarding speaker's nonverbal vocalizations.

To conclude, the solutions identified from the proposed research work will lead us towards expressive conversational speech synthesis. The main contribution of this research work is not only providing technological solutions to generate nonverbal listener vocalizations, but also building a real-time system that can be integrated with the SEMAINE project demonstration system which is aiming to build an audiovisual SAL system.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE). Thanks to Dr. Marc Schröder for useful discussions and suggestions.

REFERENCES

Baron-Cohen, S., Golan, O., Wheelwright, S., and Hill, J. (2004). *Mind Reading: The Interactive Guide to Emotions*. Jessica Kingsley Publishers, London.

Black, A. W., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *in Proc. ICASSP, 2007*, pages 1229–1232.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer. Computer program, available: *http://www.praat.org/*.

Bühler, K. (1934). *Sprachtheorie*. Gustav Fischer Verlag, Stuttgart, Germany.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60.

Douglas-Cowie, E., Cowie, R., Cox, C., Amir, N., and Heylen, D. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Proceedings of LREC*, pages 1–4, Marrakech, Morocco.

Douglas-Cowie, E., Cowie, R., and Schröder, M. (2003). The description of naturally occurring emotional speech. In *In Proceedings of the 15th International Conference on Phonetic Sciences*, Barcelona, Spain.

Gardner, R. (2002). *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins Publishing Co.

Geneva (2005). Geneva emotion wheel. *http://www.unige.ch/fapse/emotion/resmaterial/gew.zip*, Accessed 6 April 2009.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference*, pages 373–376, Washington, DC, USA.

Pammi, S. and Schröder, M. (2009). Annotating meaning of listener vocalizations for speech synthesis. In *Proc. Affective Computing and Intelligent Interaction (ACII) 2009*, Amsterdam, The Netherlands.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and de Carolis, B. (2005). Greta. a believable embodied conversational agent. pages 27–45.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Schröder, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In *Proc. Blizzard Challenge 2008*, Brisbane, Australia.

Schröder, M. and et al. (2008). Semaine deliverable d1b : First integrated system. *http://semaine.sourceforge.net/SEMAINE-1.0/D1b20system.pdf*.

Schröder, M., Heylen, D., and Poggi, I. (2006). Perception of non-verbal emotional listener feedback. In *Proc. Speech Prosody 2006*, Dresden, Germany.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech technology*, 6:365–377.

SEMAINE (2008). Semaine project page, *http://www.semaine-project.eu*.

Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318.

Tokuda, K., Zen, H., and Black, A. (2002). An HMM-based speech synthesis system applied to English. In *Proc. of 2002 IEEE SSW*, Santa Monica, CA, USA.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistic Society. Papers from the 6th regional meeting*, volume 6, page 567.