# Annotating meaning of listener vocalizations for speech synthesis

Sathish Pammi and Marc Schröder
DFKI GmbH
Saarbrücken, Germany
sathish.pammi@dfki.de, marc.schroeder@dfki.de

## Abstract

*Generation of listener vocalizations is one of the major objectives of emotionally colored conversational speech synthesis. Success in this endeavor depends on the answers to three questions: What kinds of meaning are expressed through listener vocalizations? What form is suitable for a given meaning? And, in what context should which listener vocalizations be produced? In this paper, we address the first of these questions. We present a method to record natural and expressive listener vocalizations for synthesis, and describe our approach to identify a suitable categorical description of the meaning conveyed in the vocalizations. In our data, one actor produces a total of 967 listener vocalizations, in his natural speaking style and three acted emotion-specific personalities. In an open categorization scheme, we find that eleven categories occur on at least 5% of the vocalizations, and that most vocalizations are better described by two or three categories rather than a single one. Furthermore, an annotation of meaning reference, according to Bühler's Organon model, allows us to make interesting observations regarding the listener's own state, his stance towards the interlocutor, and his attitude towards the topic of the conversation.*

## 1. Introduction

Synthesis of listener vocalizations is one of the focused research areas to improve emotionally colored conversational speech synthesis. Listener vocalizations play an important role in communicating listener intentions while the interlocutor is talking.

Listener vocalizations include backchannel utterances [12, 13] related to the flow of the conversation as well as affect vocalizations [9] based on the listener's affective state [7]. Backchannel vocalizations are the listener's response tokens [4] towards the conversation. They include acknowledgment messages like 'I am listening' and 'I am with you', but also other types of token, such as continuer, newsmarker, clarification, etc. [4]. For example, the continuer, like *mm-hm* or *uh-huh*, keeps the floor open for the current speaker to continue speaking. Listener vocalizations can also transmit affective states like excited, bored, confused, surprised, etc. [9]. Indeed, it seems that a single listener vocalization can at the same time function as a backchannel to keep the dialogue going *and* communicate affective meaning. For example, *'wow'* as a listener response can confirm reception of the speaker's message and at the same time express affective meaning like wonder or pleasure. Exactly which meaning listener vocalizations can convey does not seem clear yet. In particular, there does not seem to be any prior research on differentiating affective meaning in listener vocalizations into different kinds of affective states, such as interpersonal stance and attitudes [7].

The relationship between form and meaning of listener vocalizations is also not fully explored. With a pragmatic analysis, Gardner [4] argued that short interactive vocalizations such as *uh-huh*, *oh*, *mm*, *yeah* and *mm-hm* do not have a meaning in the conventional dictionary sense, but their meaning depends on the form of the vocalizations like prosodic shape, phonetic form, the timing within the conversation, etc. In an analysis of non-lexical utterances, Ward [11] proposed a compositional relationship between form and pragmatic meaning. He found syllabification, duration, pitch height, loudness and creaky voice to convey the lack of desire to talk, amount of thought, degree of interest, confidence and assertion of authority, respectively.

Appropriate rules for the use of listener vocalizations also remain to be specified. As one among the few studies formulating concrete prediction rules, Ward *et al*. [10, 12] suggested to predict backchannel responses in a conversation based on low pitch regions in the interlocutor's speech.

Synthesis of listener vocalizations is a crucial aspect of interactive speech synthesis, and to communicate the intended meaning, it requires answers to different research questions: Where to synthesize a listener vocalization? What form should be used to convey a given meaning? And, even more basically: What kinds of meaning are conveyed through listener vocalizations?

The present paper addresses the latter question. We at-

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

tempt to identify relevant categories of meaning for listener vocalizations in a German dialog corpus, which was recorded in view of interactive speech synthesis as a long-term research goal. Section 2 explains our attempt to capture different types of listener vocalizations from a German professional actor in a recording studio. An open annotation scheme is proposed in Section 3 to describe the meaning of listener vocalizations. We discuss results of the data collection and annotation in Section 4.

## 2. Database

Traditionally, speech synthesis databases, including expressive speech material, are recorded in a studio environment with a single speaker using predefined recording scripts. However, listener vocalizations appear natural only in conversation. Considering these issues, we opted to use a different strategy for database collection.

### 2.1. Method for database collection

We recorded dialog speech in a studio environment to get a good quality and anechoic speech corpus. Our speaker was a professional male German actor with whom we had 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. The actor was instructed to participate in a free dialogue, but to take predominantly a listener role. We encouraged him to use “small sounds that are not words”, such as *mm-hm*, where it felt natural, in order to keep his interlocutor talking for as long as possible. However, he was also allowed to “say something” and therefore to become the speaker in the conversation where this “felt natural” to keep the dialogue going.

Recordings were made in several different stages. 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 [3]: Spike is always aggressive, Obadiah is always gloomy, Poppy is always happy. These characters have been designed to represent different quadrants of the arousal-valence plane, and the actor was acquainted with their definitions from previous recordings. Sessions lasted about 20 minutes each. Their durations vary slightly according to the actor’s ability to maintain a consistent personality during the conversation.

Two female student assistants, one of whom had worked with the same actor in the past, took turns as the dialog partner, talking to him about various emotionally loaded topics of their choice. The dialogue partners were sitting in separate rooms, but they could see each other through a glass wall and hear each other using headphones, which enabled an audio-visual interaction. Each speaker’s voice was recorded on a separate channel. We also recorded the actor’s face using a standard MiniDV camera, enabling future study of audio-visual synchrony in listener behaviour. Only the analysis of the audio data is reported in this paper.

| The actor status | Corpus duration (in minutes) | Number of listener vocalizations |
|---|---|---|
| Natural | 190 | 568 |
| Obadiah | 45 | 181 |
| Poppy | 45 | 93 |
| Spike | 30 | 125 |
| Total | 310 | 967 |

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

### 2.2. Overview of the database

From the recordings, we obtained six hours of German dialog speech. For more than five hours of data, a three stage annotation has been completed as described below. Only the actor’s listener vocalizations are being used. Table 1 shows the material used in this paper.

## 3. Annotation

The annotation of listener vocalizations in our data progressed in several stages. During an initial screening process, listener vocalizations were identified, their occurrences were marked on the time axis, and a simple initial coarse description of meaning and behavior was carried out using an “ABL” annotation scheme (see below). In a second stage, a fuller analysis was carried out by means of detailed, informal descriptions of each listener vocalization. In a third stage, the full descriptions of meaning were summarized in terms of meaning categories associated with types of reference. The corpus was annotated by the same two student assistants using Praat.

### 3.1. ABL scheme

From the first sight of the corpus, we observed that many of the listener vocalizations could be characterized in terms of three overlapping categories: +/- affect, +/- backchannel and +/- laughter. Different combinations were observed, such as affective backchannels, laughter as backchannels and affective laughters. Therefore, an ABL annotation scheme was used, where A stands for Affect, B stands for Backchannel and L for Laughter, and each can be present or absent. For example, the annotated tag ’AL’ tells that the corresponding vocalization is laughter and it carries affective meaning, but it is not a backchannel. According to this scheme, the annotators had to identify listener vocalizations, mark the occurrence on the time axis, and then provide an ’ABL’ tag. For the annotation, backchannels were operationalized as short utterances like *mm-hm* and *uh-huh* which encourage the speaker to continue talking.

### 3.2. Informal descriptions

In order to get a fuller picture of the data, we used a detailed informal description of each vocalization before trying to find suitable categories to represent the meaning and behavior observed. Subsequent grouping of these descriptions will help to understand the types of form and meaning of listener vocalizations, at least for the speaker we studied. Although the annotation of a detailed informal description for each listener vocalization is a time consuming process, we wanted to make sure that we are not blinded by looking through the pattern of a pre-existing set of categories. Therefore, we had the content, form and subtexts of each listener vocalization annotated with informal descriptions in the annotator's own words, as shown in Figure 1. 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.

### 3.3. Categorical annotation of meaning

In order to abstract away from the detailed, individual descriptions towards a generalized summary view of the meaning conveyed in our data, we used a categorical annotation. Based on the informal descriptions, we aimed for a limited set of categories that capture the essence of the meaning as recorded in the descriptions. Whereas we considered it important for the informal descriptions not to be guided by any pre-existing framework, it seems appropriate to attempt using an existing set of meaning categories from the literature, and to verify to what extent it covers the meaning contained in our data. We used the Baron-Cohen [1] 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 [8] or propose their own category labels as they felt appropriate. No restrictions were made concerning the minimum or maximum number of categories to use. The same annotators who wrote the informal descriptions also assigned the categories, based on the informal descriptions and the recordings.

In addition to the annotation of meaning as such, it became apparent from our informal descriptions that several kinds of reference should be distinguished. Indeed, listener vocalizations seemed to differ with respect to their reference: is the listener providing information about his own internal state (self expression), is he reaffirming the relationship with the speaker (stance towards the other), or is he commenting about the current topic of discussion (attitude towards the topic)? Bühler's [2] Organon model (Figure 2) provides a structure distinguishing these three types of reference of an expression. In his terms, a “symptom” has the function of *expression* of the sender's state; a “signal” serves as *appeal* to a receiver; and a “symbol” is used as a *representation* of objects and facts. According to Bühler, all three functions are co-present in spoken communication, though their relative salience can vary. In our terms, this suggests we should distinguish a *self reference* (in which our listener expresses his own state), a reference towards the *other* (where the vocalization is used to signal the listener's stance towards the speaker), and a reference towards the *topic*. Following Bühler, all three functions can be expected to be co-present.

Annotators were instructed to provide a categorical annotation as follows. For any given listener vocalization, they had to provide at least one category; where the expressed meaning seemed too complex to be covered by a single category, they could use up to three categories. For each category used, they could optionally indicate the reference according to the Organon model: (S)elf reference, (O)ther reference, or (T)opic reference.

## 4. Results and discussion

### 4.1. ABL Scheme

The annotation of 967 listener vocalizations according to ABL annotation scheme were provided in the first phase. Among all listener vocalizations, 51.5% were labelled as affective, 75.5% as backchannel and 20% as laughter. The distribution of A, B and/or L is shown in Figure 3. Among the backchannels, 39.2% were labeled as affective (i.e., A+B or A+B+L), which means that more than one third of vocalizations with backchannel function were also transmitting affective meaning.

### 4.2. Meaning categories

Annotators used 24 out of the 33 Baron-Cohen categories. They added nine out of the 40 categories of the emotion wheel [5], 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. Only five

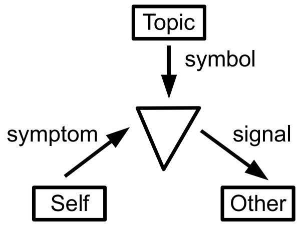
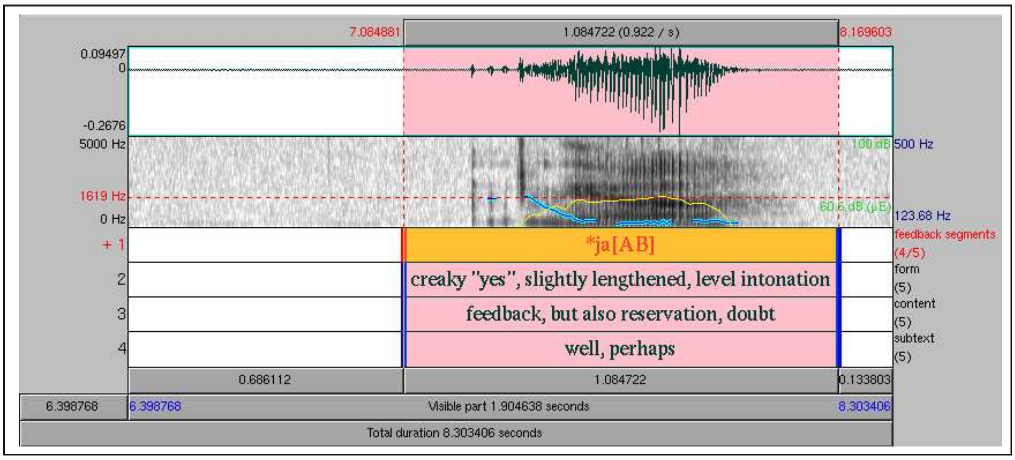


Figure 2. Bühler's Organon model of speech, adapted from [6].



*ja[AB]

creaky "yes", slightly lengthened, level intonation

feedback, but also reservation, doubt

well, perhaps

Figure 1. Example of an informal description for a listener vocalization, where the first tier represents annotation according to the ABL scheme, the second tier represents form, the third tier content and the fourth tier subtext.

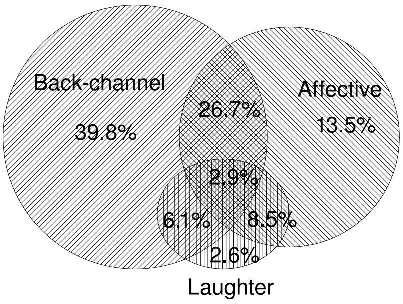


Figure 3. The distribution of listener vocalizations according to ABL annotation scheme

| | |
|---|---|
| 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.

categories were used on at least 10% of the vocalizations, and eleven categories were used on at least 5% of the data.

Annotators made frequent use of the possibility to give more than one category. 17.7% of the vocalizations were labelled with a single category; 52.9% were labelled with two categories, and 29.4% with three categories.

The characters clearly differed with respect to the categories of meaning conveyed by their listener vocalizations. In his "natural" interaction mode, the actor is friendly, interested and amused; as Spike, he is scornful, irritated, amused and ironic; as Obadiah, he is despondent and friendly; and as Poppy, he is interested and friendly (see Figure 4). This seems partly but not fully consistent with the intended personalities. A more fine-grained analysis taking into account reference annotation in addition to these meaning categories seems to show a clearer picture (see below).

### 4.3. Reference types

Annotators made very frequent use of the reference types in annotation. In 31% of the cases, they actually used all three references, which means that they considered self-related, other-related and topic-related meaning to be present in a single vocalization. In 48% of the cases, two reference types were indicated (i.e., S+O, O+T or S+T). In 14.3% of the cases, only one reference was given, and in 6.7%, no reference was specified.

The Self, Other and Topic reference based distinction seems to provide insights in the characters' expressive behavior, as shown in Figure 4. For example, the optimistic character (Poppy) shows mostly happy self expression, he is interested in the Topic, while being friendly and compassionate towards the Other.

Indeed, self-expression seems to describe very well the intended personality: despondent, irritated, uneasy and thoughtful for Obadiah, the gloomy character; happy, interested, surprised, thoughtful, excited and amused for the cheerful character Poppy; and for the aggressive character, Spike, self-expression is amused, irritated, ironic, scornful, and confident. In the same way, we can now characterize the "natural" speaking mode of our actor as amused, sometimes decisive and sometimes tentative, and thoughtful.

The only category that does not quite seem to fit the picture is the observation that Spike is predominantly

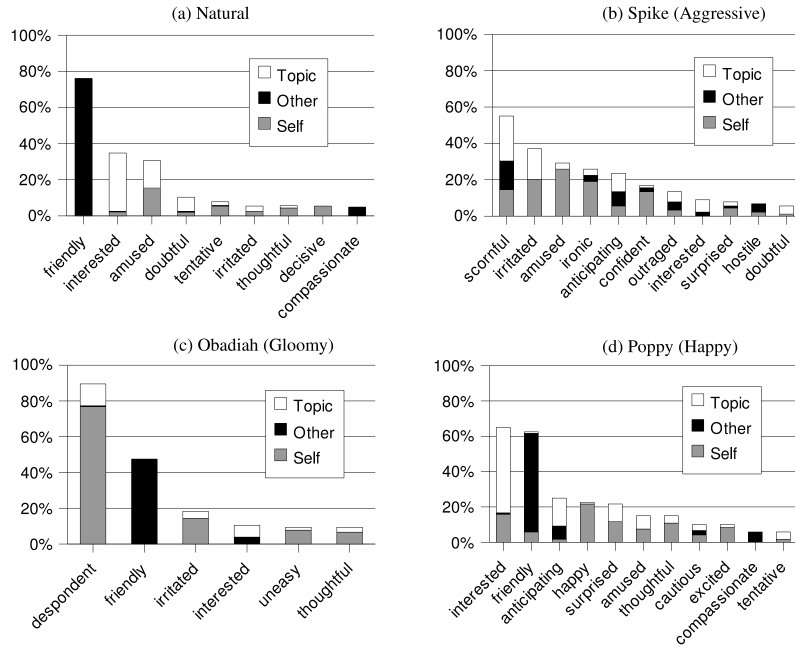


Figure 4. Most frequent affective-epistemic categories and associated reference types, per character.

amused. To understand better the instances in which Spike is amused, we show the most frequent categories co-occurring with "amused" for Spike and for the other character showing substantial self-amusement, the natural speaking mode of the actor (Figure 5). It is very obvious that Spike's amusement co-occurs nearly exclusively with negative emotions such as scornful, outraged and ironic, whereas the natural actor shows amusement mostly with the positive categories friendly and interested. This suggests that the two kinds of amusement are actually very different – a point that would have been difficult to make if only a single meaning category had been annotated.

The Other reference seems to show clear differences in interpersonal stance among the characters. For Spike, the aggressive character, Other-related expressions are scornful, outraged, ironic or hostile, whereas other characters are friendly or compassionate. The attitude towards the topic of discussion seems to be sensibly indicated by the Topic reference: the actor himself and Poppy show a lot of interest, whereas Spike shows a predominantly scornful and irritated attitude, and Obadiah shows little topic-related signs at all.

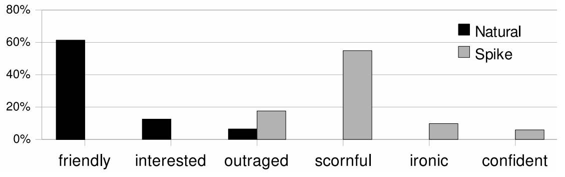


Figure 5. Most frequent ($> 5\%$) meaning categories co-occurring with the category 'amused', for natural and Spike listener modes.

These results suggest that distinguishing the reference in addition to affective-epistemic meaning categories may be 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).

## 4.4. Inter-rater agreement

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 as described above. As we allowed for more than one category per instance, we computed Cohen's Kappa separately for each category, treating annotations as a binary "present/absent" feature. On this basis, 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 S, O, and T, Kappa was close to 0, indicating 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.

## 4.5. Laughter

A behavior-level description of the listener vocalizations in our data is beyond the scope of the present paper. In a similar way as for meaning categories, it will involve the abstraction of relevant behavior elements from the informal descriptions presented in Section 3.2. As a sketch of methodology, we present first results on laughter that we have obtained from the initial ABL annotation.

Treating in the first instance laughter as a single behavioral category, we can investigate the meaning categories associated with it (Figure 6). It can be seen that laughter nearly exclusively occurs with amusement, and that much of it is friendly. However, some laughter is not friendly, and even scornful. For a synthesis system, it would be extremely important to know whether the laughter itself, in isolation, contains the "friendly" vs. "scornful" elements of meaning or if these have been derived from the context. Listening tests presenting laughter in isolation could be used to answer this question. If appropriate, then, several kinds of laughter should be distinguished in order to obtain as simple as possible a mapping between meaning and behavior.

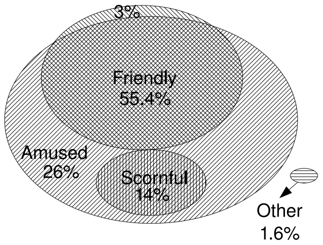


Figure 6. Distribution of meaning categories on laughter vocalizations.

## 5. Conclusion

In this paper, we have described a method for collecting listener vocalizations in view of emotionally colored conversational speech synthesis, and explored a methodology for annotating listener vocalizations. Firstly, we started with a simple affect-backchannel-laughter annotation scheme. Secondly, we continued with an open annotation through informal descriptions of behavior and meaning. Finally, we used a joint categorical description for meaning and reference, in which the meaning is described using affective-epistemic categories, and reference annotation is based on Bühler's Organon model. We found that Baron-Cohen's affective-epistemic categories were not sufficient to describe our data – it was necessary to add a number of categories from the Geneva Emotion Wheel as well as some custom categories. The annotation of reference allowed us to describe the different characters' intended personality, interpersonal stance towards the interlocutor, and attitude towards the topic of discussion.

The generally low inter-rater agreement shows that further work is needed before the meaning and reference annotation scheme can be considered a reliable tool for describing data. Improvements can be expected from a consolidation of the large set of meaning categories into a smaller set of clearly distinguishable categories, as well as improved annotation instructions.

The next steps towards synthesis consist in an annotation of behaviour as outlined in Section 4.5, definition of markup for requesting the synthesis of listener vocalisations, and the investigation of various synthesis technologies for generating the actual synthesis audio.

## 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 Annette Klepp and Juliane Schmidt for their help in recording and annotating the data.

## References

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

[2] K. Bühler. *Sprachtheorie*. Gustav Fischer Verlag, Stuttgart, Germany, 1934.

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

[4] R. Gardner. *When Listeners Talk: Response Tokens and Listener Stance*. John Benjamins Publishing Co, Feb. 2002.

[5] Geneva emotion wheel, 2005. Accessed 6 April 2009.

[6] K. R. Scherer. On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, pages 7:79–100, 1988.

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

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

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

[10] N. Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96).*, volume 3, pages 1728–1731 vol.3, 1996.

[11] N. Ward. Non-lexical conversational sounds in american english. *Pragmatics & Cognition*, 14(1):131–184, 2006.

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

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