Multidimensional meaning annotation of listener vocalizations for synthesis

Sathish Pammi, Marc Schröder, and Marcela Charfuelan

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany and
Alt-Moabit 91c, D-10559, Berlin, Germany
{firstname.lastname}@dfki.de

Abstract. Listener vocalizations convey affective and epistemic states behind the listener's intentions while the interlocutor is talking. The meaning annotation of such vocalizations is a crucial step in synthesis of listener vocalizations. This paper presents a perception study to annotate meaning of vocalizations. In this study, subjects annotate (characterize) a set of listener vocalizations using a multi-dimensional set of meaning descriptors. The set of stimulus vocalizations is selected based on intonation clustering. We investigate the typical impressions and the appropriateness of meanings conveyed by vocalizations, based on high agreement ratings provided by the participants. We also discuss the suitability of the annotation procedure to generate expressive listener vocalizations.

Keywords: listener vocalizations, perception study, meaning, speech synthesis

1 Introduction

Nowadays spoken and multimodal dialogue systems attempt to model the computer's part of the dialogue in both the speaker and the listener role [12, 16]. That means the machine must emit signs of listening while the user is speaking: backchannels [19] or expressive feedback signals [1]. In multimodal dialogue systems, some of these signals can be visual, such as head nods, smiles, or raised eyebrows [5]; in the vocal channel, backchannel and feedback signals can be realized as listener vocalizations. Listener vocalizations like *mhm, right, yeah, uh-huh* are not only produced to make the interaction more natural but also to signal affective meanings such as *anger, amusement* and epistemic meanings such as *interested, agreeing*.

Yngve [19] investigated responses such as *uh-huh*, *yes*, *okay*; he called them as "behavior in the back channel". Duncan [6] attempted to correlate meaning with segmental forms like *yeah*, *right* and *I see*; whereas Schegloff [17] and McCarthy [10] noted the multifunctioning of vocalizations. Later studies [8, 18] indicates that several behavior properties like segmental form, intonation, voice-quality have influence on the meaning conveyed by vocalizations.

Although several studies attempted to understand meanings of vocalizations, there has been not much focus on how these vocalizations can be used for synthesis. An integrative account of all these studies must be considered in a bigger picture. It requires the following sequence of steps: (i) identification of suitable meaning descriptors; (ii)

annotation of appropriateness for each meaning descriptor; (iii) identifying a typical impression of meanings for each vocalization; (iv) analyzing the impact of behavioral properties like segmental form and intonation on perceived meaning. We attempt the above steps in this paper.

In order to synthesize an appropriate listener vocalization, we require two kinds of information about each of the available vocalizations [13]: a typical impression of the meaning that the vocalization could convey; and how appropriate is the vocalization for a given meaning. In this paper, we experiment a methodology to find meanings of vocalizations that are usable for synthesis. We conduct a listening test where subjects annotate (characterize) a set of listener vocalizations using a multi-dimensional set of meaning descriptors.

Considering the possibility to improve acoustic variability using imposed intonation contours [14], we also investigate the relevance of intonation and segmental form on the perceived meaning. This motivates the procedure of stimuli selection for the experiment. The paper is organized as follows. In Section 2 the vocalizations database used in this study is described. Section 3 describes our meaning descriptors used in this study. In Section 4 our approach to select representative vocalizations is explained. In this section the perception experiment is also explained. In Section 5 main results are discussed and in Section 6 findings are summarized.

2 Vocalizations database

To collect natural listener vocalizations from dialogue speech, we recorded about half an hour of free dialogue with professional British actors. Four British actors were selected for four Sensitive Artificial Listener (SAL) voices: cheerful (Poppy), neutral (Prudence), gloomy (Obadiah), and aggressive (Spike) voices. The British actors were originally chosen for the recordings required for building new TTS voices. In addition to speech synthesis recordings, free dialogue of around 30 minutes was recorded with each of the British speakers. The recording setup and instructions given to the actors are described in [15].

	Prudence	Poppy	Spike	Obadiah
Corpus duration (in minutes)	25	30	32	26
number of vocalizations	128	174	94	45

Table 1: British English listener vocalizations recorded for the four SAL characters

Once the dialogue was recorded for all four characters, listener vocalizations were marked on the time axis and transcribed as a single (pseudo-)word, such as *myeah* or *(laughter)*. With respect to the number of listener vocalizations they produced the speakers varied enormously. Whereas Obadiah produced only 45 vocalizations, Poppy produced 174 (see Table 1).

3 Meaning descriptors

We started by establishing a list of meaning dimensions, based on three sources: the most frequent categories in an exploratory annotation study on German listener vocal-

izations [15]; the most frequently used annotations of the SEMAINE corpus [11] – a large and annotated collection of dialogue of the SAL domain; and a set of affective-epistemic descriptors used to describe visual listener behavior [4].

Descriptors	**Scale type**	**Source**
anger	unipolar	Emotional categories
sadness	unipolar	
amusement	unipolar	
happiness	unipolar	
contempt	unipolar	
solidarity	unipolar	IPA categories
antagonism	unipolar	
(un)certain	bipolar	Baron-Cohen's categories
(dis)agreeing	bipolar	
(un)interested	bipolar	
(high/low)anticipation	bipolar	

Table 2: Consolidated list of meaning descriptors used in this study

The three sources were consolidated into a list of 11 descriptors as shown in Table 2. The table shows the scale type (unipolar/bipolar) of meaning descriptors. We made sure that these categories are derived from three different backgrounds, emotional categories [7], Baron-Cohen's epistemic mental states [3] and Bales Interaction Process Analysis (IPA) [2]. Whereas epistemic states can be used to transmit attitudinal mental states of listener, IPA labels can be used to convey social meanings in dialogue.

4 Approach

This section describes our approach to annotate meanings of listener vocalizations. Annotation of meaning for all listener vocalizations is a tedious and time consuming process. Instead, annotation of selective vocalizations would be more cost effective. As literature [8, 18] suggests that the meaning of vocalization highly correlates with segmental form and intonation, we propose a semi-automatic procedure to select representative vocalizations of segmental forms and intonation contours in the corpus. This also facilitates us to investigate the relevance of segmental form and intonation on the perceived meaning.

4.1 Stimuli selection

The stimuli are selected based on a semi-automatic clustering of intonation contours. For clustering vocalizations according to intonation, a contour was automatically computed for each vocalization by fitting a 3rd-order polynomial to f0 values extracted using the Snack pitch tracker [9]. Polynomials can approximate intonation contours of speech signal in unvoiced regions. Separately for each speaker, we used K-means clustering of intonation contours to identify the vocalizations with a similar intonation.

Two sets of stimuli were manually extracted from the clustered data for the purpose of selecting representative vocalizations that cover the maximum number of possible segmental forms and intonation contours. We aimed for two sets that contain, on one hand, stimuli with the same segmental form (as determined from the single-word description) varying in intonation (identified in the following as *fixed segmental form*); and on the other hand, stimuli with the same intonation (flat intonation contour) and varying in segmental form (henceforth, *fixed intonation contour*). Thus we manually selected samples from clusters as follows: (i) in order to get wide range of contour shapes, we selected one or two representative samples from each cluster with same segmental form (i.e. *yeah*); (ii) we selected samples with different segmental forms from a single cluster where contour shape is constant. Table 3 shows the number of selected stimuli for the experiment.

Character	*Fixed segmental form*	*Fixed intonation contour*
Poppy	15	8
Spike	10	9
Obadiah	5	8
Prudence	8	9
Total	38	34

Table 3: Character wise number of vocalizations selected for meaning annotation

4.2 Perception experiment

Scale-based ratings capture inherent ambiguity more than forced-choice test. We designed a web-based perception study for participants. The first page provided instructions, the second page collected demographic information and the following pages present the audio and rating scales one at a time, as shown in Figure 1. The stimuli were presented to the participants in a random order for eliminating order and fatigue effects. Participants could play the audio as many times as they liked before providing meaning ratings. A 5-points Likert scale for each meaning was used: from 1 (absolutely no attribution) to 5 (extremely high attribution) for unipolar meaning categories; from -2 (extremely negative attribution) to +2 (extremely positive attribution) for bipolar meaning categories. “No Real Impression” option was provided for each meaning scale in case the participant is unsure.

44 participants (20 women, 24 men) took part in the annotation study. 22 participants provided ratings for the vocalizations in test set *fixed segmental form* (9 women, 13 men) and 22 participants rated vocalizations in test set *fixed intonation contour* (11 women, 11 men).

5 Results and discussion

In order to study each of the vocalizations per meaning, we first introduce the term *meaning-vocalization* combination that is used in the rest of this paper. Each vocalization can convey maximally 11 meanings used in the corpus annotation. One stimulus

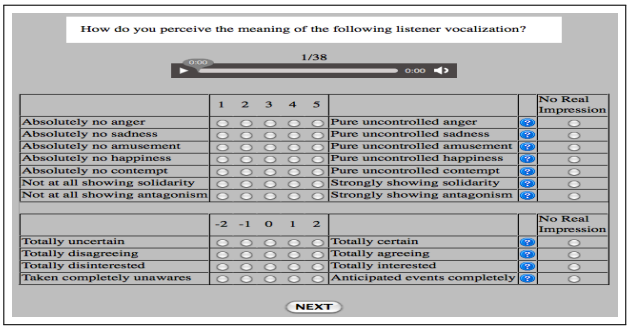

Fig. 1: A screenshot of the web page for the perception study

indicates 11 *meaning-vocalization* combinations. For example, in the case of Prudence (see Table 4), 187 *meaning-vocalization* combinations (17 stimuli * 11 meaning categories) were available for analysis.

5.1 High versus Low agreement

Table 4 shows the high variability on agreement of *meaning-vocalization* combinations for Prudence. In this table high agreement is identified with circles or arrows and low agreement is identified with a dot (·). In order to identify high agreement versus low agreement of *meaning-vocalization* combinations, we computed the interquartile range (IQR) of ratings provided for each combination. We considered that a combination has high agreement if the IQR of the combination is less than one third of the meaning scale range. In other words, a combination has high agreement if more than 50% of the raters agree within one third of the meaning scale range. The high agreement combinations indicates typical impression of the meaning on the vocalization.

Table 4 shows that the number of low agreement annotations (identified as ·) are higher in the *fixed intonation contour* set when compared to the *fixed segmental form* set for Prudence. The same tendency was observed when taking into account all the vocalizations in our corpus, that is 792 (72 stimuli * 11 categories) *meaning-vocalization* combinations, from which 418 combinations belong to the *fixed segmental form* set and 374 belong to the *fixed intonation contour* set. Figure 2 shows a global picture of high agreement versus low agreement combinations for all the corpus. While around 60% of the *fixed segmental form* combinations show high-agreement, only 40% of the *fixed intonation contour* combinations show high-agreement. This seems to indicate that the participants perceived more distinguishable information from intonation when compared to segmental form. In other words, this evidence indicates that the intonation contour is highly relevant for signaling meaning when compared to phonetic segmental form.

5.2 Appropriateness of high agreement annotations

Not all vocalizations with high agreement may be suitable to convey a specific meaning for synthesis. In this work the suitability of a *meaning-vocalization* combination is cal-

Fixed segmental form													Fixed intonation contour											
segmental form	intonation-contour	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation	segmental form	intonation-contour	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	⁀	○	·	○	○	·	↑	○	↑	↑	○	○	tsyes	‿	·	·	·	·	·	·	·	↑	○	·	·
yeah	⁀	○	⇑	○	○	○	·	○	○	○	·	○	tsyeah	‿	·	·	·	○	·	·	·	·	·	·	○
yeah	‿	○	·	○	○	·	·	○	↑	↑	○	↑	mhm	‿	·	·	○	○	·	↑	·	·	○	·	·
yeah	⁀	○	○	·	·	○	⇑	○	↑	↑	↑	·	yeah	‿	·	·	○	○	·	·	·	↑	·	·	·
yeah	‿	○	·	○	○	○	·	·	○	↑	·	·	yes	⁀	·	·	○	○	·	·	·	○	○	○	↑
yeah	‿	○	○	⇑	⇑	○	⇑	○	↑	↑	↑	·	right	‿	·	○	○	○	·	·	·	·	·	·	·
yeah	‿	○	·	○	○	○	·	○	○	○	↓	○	tsright	‿	·	·	○	○	·	·	·	↑	·	·	○
yeah	‿	○	·	○	○	·	↑	·	·	·	↓	○	aha	⁀	○	○	·	·	·	⇑	○	↑	↑	↑	↑
													tsgosh	‿	○	○	·	·	·	·	○	○	○	○	○

Table 4: Segmental form, intonation contour and meaning of Prudence's stimuli. *Meaning-vocalization* combination is represented using the following symbols.
○: vocalization is not appropriate for the meaning;
↑ or ↓ : vocalization is somewhat appropriate;
⇑ or ⇓ : vocalization is very appropriate for the meaning;
·: the annotation has low agreement (we can not conclude on appropriateness);
↓ and ⇓ : negative sides of bipolar scales

culated by computing the median of ratings provided for that combination. However, we can not conclude about suitability of low agreement ratings.

We distinguish three levels of appropriateness based on where the participants tend to agree on the meaning scale. A *meaning-vocalization* combination is very appropriate if the participants tend to agree on positive (in case of unipolar and bipolar scales) or negative (in case of bipolar scale) end of meaning scale. The combination is not appropriate to convey the meaning if they tend to agree on '0'. In other words, we can say that the combinations are "very appropriate", "somewhat appropriate", and "not appropriate" when the median is greater than two third of meaning scale, between one third and two third, and less than one third respectively. Among high-agreement *meaning-vocalization* combinations available in our corpus, it was found that, 7.2% (30) are very appropriate, 22.4% (93) are somewhat appropriate, and 70.4% (293) are not appropriate combinations. This result is highly relevant in speech synthesis, that is, one vocalization can be "not appropriate", "somewhat appropriate" or "very appropriate" for several different meanings at the same time. These three categories can be used, for example, in an algorithm for unit-selection synthesis (i.e. vocalization selection) that considers

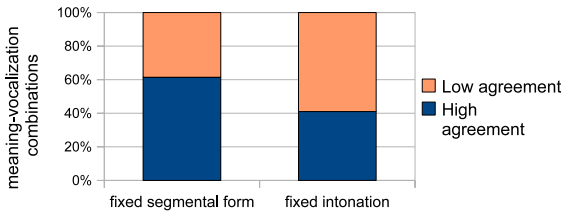

Fig. 2: Percentage of high and low agreement *meaning-vocalization* combinations

appropriateness to realize a particular intended (target) meaning. The evaluation of such unit-selection algorithm has been presented in [13].

5.3 Inherent ambiguity of listener vocalizations

According to Table 4, the vocalization *aha* can convey 5 meanings (*solidarity, certain, agreeing, interested, anticipation*), whereas the vocalization *right* does not convey any meaning available in our descriptors. Figure 3 shows the histogram of possible meanings for the listener vocalizations in our corpus. Among 72 stimuli, 14 vocalizations (19.5%) convey no meaning, 27 (37.5%) convey single meaning, and the remaining 31 (43%) convey multiple meanings. On average, a single vocalization in this corpus can convey 1.68 meanings, this confirms the argumentations already made in the literature [10,17]. Indeed the inherent ambiguity of listener vocalizations is a very interesting feature to exploit in speech synthesis, because a single vocalization can be used in multiple instances.

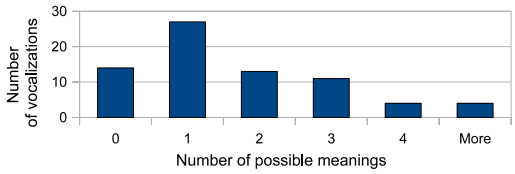

Fig. 3: Histogram of multiple meanings

6 Conclusion

In this paper, we explored a multi-dimensional annotation methodology to annotate listener vocalizations in view of conversational speech synthesis. We conclude the following issues from this study: (i) this methodology can provide a typical impression of meanings from high agreement annotations; (ii) unit-selection algorithms can benefit from the annotation of meaning on scales: it captures appropriateness of listener vocalizations for a given meaning; (iii) one vocalization can convey several meanings,

which is useful for the usage of the same vocalization in several instances; (iv) the evidence indicates that the intonation contour is highly relevant for signaling meaning when compared to the phonetic segmental form - in support for improving acoustic variability using imposed-intonation contours.

7 Acknowledgements

This research has received funding from the European Community's Seventh Framework Programme (FP7-ICT) under grant agreement no. 211486 (SEMAINE), 248116 (ALIZ-E) and 231287 (SSPNet). We would like to thank Professor Roddy Cowie and Dr. Gary McKeown for useful discussions.

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9(1), 1–26 (1992)
2. Bales, R.: Interaction process analysis. Cambridge, Mass (1950)
3. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.: Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers, London (2004)
4. Bevacqua, E., Heylen, D., Pelachaud, C., Tellier, M.: Facial feedback signals for ECAs. In: AISB 2007 Annual convention, workshop "Mindful Environments". Newcastle, UK (2007)
5. Bevacqua, E., Pammi, S., Hyniewska, S., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: IVA 2010. Philadelphia, USA (2010)
6. Duncan, S.: On the structure of speaker–auditor interaction during speaking turns. Language in society 3(02), 161–180 (1974)
7. Ekman, P., Dalgleish, T., Power, M.: Handbook of cognition and emotion. Chihester, UK: Wiley (1999)
8. Kowtko, J.: The function of intonation task-oriented dialogue (1996)
9. KTH: The snack sound toolkit. http://www.speech.kth.se/snack (2006)
10. McCarthy, M.: Talking back:" small" interactional response tokens in everyday conversation. Research on Language & Social Interaction 36(1), 33–63 (2003)
11. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Proc. IEEE ICME 2010. Singapore (2010)
12. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive ECA system. In: Proc. AAMAS. p. 1399–1400 (2009)
13. Pammi, S., Schröder, M.: Evaluating the meaning of synthesized listener vocalizations. In: INTERSPEECH 2011 (2011)
14. Pammi, S., Schröder, M., Charfuelan, M., Türk, O., Steiner, I.: Synthesis of listener vocalisations with imposed intonation contours. In: SSW7 Workshop. Kyoto, Japan (2010)
15. Pammi, S., Schröder, M.: Annotating meaning of listener vocalizations for speech synthesis. In: Proc. Affective Computing & Intelligent Interaction. Amsterdam, The Netherlands (2009)
16. Pfleger, N., Alexandersson, J.: Modeling non-verbal behavior in multimodal conversational systems. Information Technology 46(6), 341–345 (2004)
17. Schegloff, E.: Discourse as an interactional achievement: Some uses of "uh-huh" and other things that come between sentences. Analyzing discourse: Text and talk 71(93) (1982)
18. Ward, N.: Non-lexical conversational sounds in american english. Pragmatics &# 38; Cognition 14(1), 129–182 (2006)
19. Yngve, V.H.: On getting a word in edgewise. In: Chicago Linguistic Society. Papers from the 6th regional meeting. vol. 6, pp. 567–577 (1970)