Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 3323-3328

Multi-Dimensional Meaning Annotation in Synthesis of Listener Vocalizations

Viswanatha Reddy Allugunti, Prof. Dr. Biplab Kumar Sarkar
Research Scholar, Professor
Glocal University, Uttar Pradesh 247121, India

Abstract:

The meaning annotation of listener vocalizations is a crucial step towards the synthesis of these vocalizations.This chapter describes a systematic study of vocalizations' meanings. We propose a multi-dimensional annotation approach aimed at obtaining appropriateness ratings of each vocalization for each of the meanings. We conduct a listening test where multiple subjects annotate (characterize) a set of listener vocalizations using a multidimensional set of meaning descriptors. Typical impressions on context-independent meaning of listener vocalizations are being investigated. We also analyse the relevance of behaviour properties for the meaning perception of listener vocalizations.

Keywords:open-endedness,segmental form, intonation, appropriateness.

1. Introduction:

The meaning annotation of listener vocalizationsis a critical step in the synthesis of these vocalisations. In the preceding chapter, an open-ended exploratory investigation was given in order to determine the list of probable interpretations accessible in a database of German listener vocalisations. Despite the fact that such research assisted us in identifying a plausible list of meanings in the corpus, we made a few findings as a result of it.

2. Multi-Dimensional Meaning Annotation

In addition, various additional research was conducted in order to better comprehend the meanings of vocalisations. None of them, however, were concerned with determining the appropriateness of meanings. All of these research must be examined in the context of a larger picture. It necessitates the following procedure: Identification of relevant meaning descriptors; annotation of appropriateness for each meaning descriptor; identification of a typical perception of meanings for each vocalisation; analysis of the influence of behavioural features such as segmental shape and intonation on perceived meaning. We attempt the above steps in this chapter. In order to synthesize an appropriate listener vocalization, we require two kinds ofinformation about each of the available vocalizations:

A typical impression of the meaning that the vocalization could convey. How appropriate is the vocalization for a given meaning.

3. Experimental corpus:

Table 1.1 shows the database of vocalizations, which is recorded by four professionalBritish actors, used for multi-dimensional meaning annotation.

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

Table 1.1: British English listener vocalizations recorded for the four SAL characters

4, Approach

Consolidating meaning descriptors

As described, We want to combine the list of meaning descriptors produced in the previous chapter in order to strike a balance between the time and effort required to annotate the vocalisations and the proportion of vocalisations covered by the consolidated list.

The most frequently used annotations of the SEMAINE corpus (McKeown et al. 2010) – a large and annotated collection of dialogue in the SAL domain; and a set of affective-epistemic descriptors used to describe visual listener behaviour (Bevacqua et al. 2007) – were used to create a list of meaning dimensions.at ease, thoughtful, concentrating, shows solidarity, shows antagonism and so on.

We also ensured that the consolidated list of categories comes from three sources: emotional categories (Ekman 1999), Baron-epistemic Cohen's mental states (Baron-Cohen et al. 2004), and Bales Interaction Process Analysis (IPA) (Bales 1950).

The rationale is simple: as explained in Chapter 2, listener vocalisations communicate emotional, epistemic, and turn-taking cues, as well as cognitive, social, and discourse regulation functions. The three backgrounds, in this author's opinion, are the greatest sources accessible to cover these states and functions. Emotional categories may be used to transmit affective meanings; epistemic states can be used to reflect the listener's attitudinal mental states; and IPA labels can be used to indicate social meanings in discourse.

5. Stimuli selection

The stimuli are chosen using a semi-automated intonation contour clustering method. A contour was automatically generated for each vocalisation for grouping vocalisations according to intonation by fitting a 3rd-order polynomial to f0 values retrieved using the Snack pitch tracker (Sjölander 2006). In unvoiced regions, polynomials can approximate intonation contours of speech signals. We utilised K-means clustering of intonation contours to find vocalisations with comparable intonation for each speaker separately.

Two sets of stimuli were manually picked from the clustered data in order to identify sample vocalisations that covered the widest range of segmental shapes and intonation contours conceivable. We aimed for two sets of stimuli: one with the same segmental form (as determined from the single-word description) but varying in intonation (referred to as fixed segmental form); and the other with the same intonation (flat intonation contour) but varying in segmental form (referred to as variable segmental form) (henceforth, fixed intonation contour). As a result, we manually chose samples from clusters in the following manner: I we selected samples with different segmental forms from a single cluster where contour shape is constant in order to get a wide range of contour shapes; (ii) we

selected samples with different segmental forms from a single cluster where contour shape is constant in order to get a wide range of contour shapes.Table 1.2shows the number of selected stimuli for the experiment.

Character	Fixed segmental form	Fixed intonation contour
Poppy	25	12
Spike	20	10
Obadiah	15	8
Prudence	10	9
Total	70	39

Table 1.2: Character wise number of vocalizations selected for meaning annotation

6. Perception experiment

More than a forced-choice exam, scale-based assessments convey underlying ambiguity. For participants, we created a web-based perception research (see Appendix A). As illustrated in Figure 8.1, the first page offered instructions, the second page gathered demographic information, and the succeeding pages presented the audio and rating scales one by one. To eliminate order and tiredness effects, the stimuli were given to the subjects in a random order. Before submitting meaning evaluations, participants may listen to the recording as many times as they wanted. For each meaning, a 5-point Likert scale was used: for unipolar meaning categories, from 1 (absolutely no attribution) to 5 (very high attribution); for bipolar meaning categories, from -2 (highly negative attribution) to +2 (extremely positive attribution). For each meaning scale, a "No Real Impression" option was supplied in case the participant was unsure. 44 participants (20 women, 24 men) took part in the annotation study. 22 participantsprovided 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).

7. Results and discussion

In order to study each of the vocalizations per meaning, we first introduce the termmeaning-vocalization combination that is used in the rest of this chapter. Each vocalizationcan convey maximally 11 meanings used in the corpus annotation. Onestimulus indicates 11 meaning-vocalization combinations.

7.1. High versus Low agreement

Table 1.3 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[17]
low agreement of meaning-vocalization combinations, we computed the interquartilerange (IQR) of ratings provided for each combination. We considered that a combinationhas high agreement if the IQR of the combination is less than one third of themeaning scale range. In other words, a combination has high agreement if more than50% of the raters agree within one third of the meaning scale range. The high agreementcombinations indicates typical impression of the meaning on the vocalization. Table 1.3shows that the number of low agreement annotations (identified as _)

arehigher in the fixed intonation contour set when compared to the fixed segmental formset for Prudence[16]. The same tendency was observed when taking into account all the vocalizations in our corpus[15], that is 792 (72 stimuli * 11 categories) meaning-vocalizationcombinations, from which 418 combinations belong to the fixed segmental form setand 374 belong to the fixed intonation contour set. Figure 1.1shows a global pictureof high agreement versus low agreement combinations for all the corpus. While around60% of the fixed segmental form combinations show high agreement, only 40% of thefixed intonation contour combinations show high agreement. This seems to indicatethat the participants perceived more distinguishable information from intonation whencompared to segmental form. In other words, this evidence indicates that the intonationcontour is highly relevant for signalling meaning when compared to phonetic segmentalform[14].

Fixed segmental form												
segmental form	intonation-contour	voicequality	anger	sadness	amusement	happiness	contempt	solidarity	antagonism	certain	agreeing	interested	anticipation
yeah	⁀	modal	o	·	o	o	·	↑	o	↑	↑	o	o
yeah	⁀	modal	o	⇑	o	o	o	·	o	o	o	·	o
yeah	⁀	creaky	o	·	o	o	·	·	o	↑	↑	o	↑
yeah	⁀	modal	o	o	·	·	o	⇑	o	↑	↑	↑	·
yeah	⁀	modal	o	·	o	o	o	·	·	o	↑	·	·
yeah	⁀	modal	o	o	⇑	⇑	o	⇑	o	↑	↑	↑	·
yeah	⁀	creaky	o	·	o	o	o	·	o	o	o	↓	o
yeah	⁀	modal	o	·	o	o	·	↑	·	·	·	↓	o

Table 1.3: Fixed segmental form set: 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.

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

Table 1.4: Fixed intonation contour set: Segmental form, intonation contour and meaning

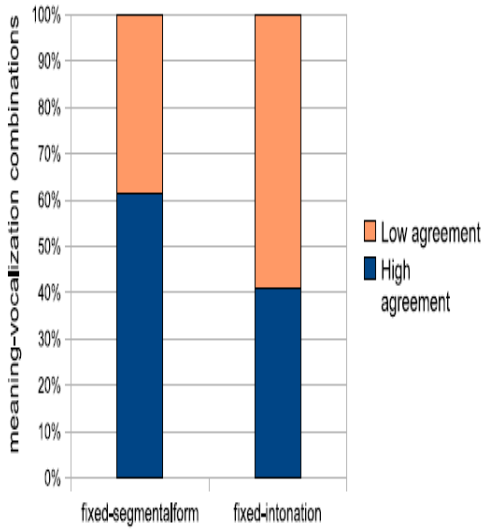

Figure 1.1: Percentage of high vs. low agreement meaning-vocalization combinations

8. Conclusion

We looked at a multi-dimensional annotation approach for annotating listener vocalisations in the context of conversational speech synthesis in this chapter. We conclude the following from this analysis: I unit-selection algorithms can benefit from the annotation of meaning on scales: it captures the appropriateness of listener vocalisations for a given meaning; (ii) one vocalisation can convey multiple meanings, which is useful for the use of the same vocalisation in multiple instances; (iv) the evidence seems to indicate that the intonation contour is highly relevant for the use of the same vocalisation in multiple instances; (v) the evidence seems to indicate that the intonation contour is highly relevant

9. REFERENCES

[1].Thórisson, K.R. (1996). "Communicative humanoids: a computational model of psychosocialdialogue skills". PhD thesis. Massachusetts Institute of Technology.

— (2002). "Natural turn-taking needs no manual: Computational theory and model, from perception to action". In: *Multimodality in language and speech systems*,pp. 173–207.

[2]. Thórisson, K.R. et al. (2005). "Whiteboards: Scheduling blackboards for semanticrouting of messages & streams". In: *AAAI Workshop on Modular Constructionof Human-Like Intelligence*, pp. 8–15.

[3]. Toda, T. and K. Tokuda (2005). "Speech parameter generation algorithm consideringglobal variance for HMM-based speech synthesis". In: *Proc. Ninth European Conferenceon Speech Communication and Technology*.

[4]. Tokuda, K., H. Zen, and A.W. Black (2002). "An HMM-based speech synthesis systemapplied to English". In: *Proc. IEEE Workshop on Speech Synthesis. IEEE*, pp. 227–230.

[5]. Tokuda, K. et al. (2000). "Speech parameter generation algorithms for HMM-basedspeech synthesis". In: *Proc. International Conference on Acoustics, Speech andSignal Processing (ICASSP). IEEE*, pp. 1315–1318.

[6]. Tokuda, K. et al. (2008). The HMM-based speech synthesis system (HTS) Version 2.1.Online: http://hts.sp.nitech.ac.jp/ (accessed on 25th June, 2011).

[7]. Tottie, G. (1991). "Conversational style in British and American English: The case of backchannels". In: *English corpus linguistics*, pp. 254–271.

[8]. Versloot, CA (2005). What do Listeners do? A Simple Annotation Schema. Tech. rep.Enschede: University of Twente (HMI).

[9]. Vilhjálmsson, H. et al. (2007). "The behavior markup language: Recent developmentsand challenges". In: *Proc. Intelligent Virtual Agents. Springer*, pp. 99–111.

[10]. Ward, N. and W. Tsukahara (2000). "Prosodic features which cue back-channel responsesin English and Japanese". In*: Journal of Pragmatics 32.8*, pp. 1177–1207.

[11]. Ward, Nigel (2006). "Non-lexical conversational sounds in American English". In:*Pragmatics & Cognition 14.1*, pp. 131–184.

[12]. Xudong, D. (2009). "Listener response". In*: The pragmatics of interaction 4*, pp. 104–124.

[13]. Yamagishi, J. et al. (2003). "Modeling of various speaking styles and emotions for HMM-based speech synthesis". In: Proc. Eighth European Conference on Speech Communication and Technology.

[14]. VR Allugunti, CKK Reddy, NM Elango, PR Anisha - Intelligent Data Engineering and Analytics, 2021

[15]. VR Allugunti, NM Elango "Development of a Generic Secure Framework for Universal Device Interactions in IoT of Fifth Generation Networks"

[16]. Viswanatha Reddy Allugunti, D Jayaramaiah, Prasanth A and Dr.AnirbanBasu "Agent Based Performance Analysis of Next Generation Mobile Networks(LTE)" International Journal of Computer Science and Information Technology & Security (IJCSITS).

[17]. Viswanatha Reddy Allugunti, Dhana Naga Kalyani Tummala "Designing and Development of Framework for Supply Chain AI Bots" Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020 .