Evaluating the meaning of synthesized listener vocalizations

Sathish Pammi and Marc Schröder

DFKI GmbH, Saarbruecken
`firstname.lastname@dfki.de`

Abstract

Spoken and multimodal dialogue systems start to use listener vocalizations for more natural interaction. In a unit selection framework, using a finite set of recorded listener vocalizations, synthesis quality is high but the acoustic variability is limited. As a result, many combinations of segmental form and intended meaning cannot be synthesized.

This paper presents an algorithm in the unit selection domain for increasing the range of vocalizations that can be synthesized with a given set of recordings. We investigate whether the approach makes the synthesized vocalizations convey a meaning closer to the intended meaning, using a pairwise comparison perception test. The results partially confirm the hypothesis, indicating that in many cases, the algorithm makes available more appropriate alternatives to the available set of recorded listener vocalizations.

Index Terms: interactive speech synthesis, unit selection, listener vocalizations, PSOLA

1. Introduction

Speech synthesis is used in increasingly interactive dialogue settings. Whereas early spoken dialogue systems adopted a ping-pong strategy for turn taking, newer spoken and multimodal dialogue systems attempt to model the computer's part of the dialogue in both the speaker and the listener role [1, 2]. That means the machine must emit signs of listening while the user is speaking: backchannels [3] or expressive feedback signals [4]. 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.

In earlier work we have described how we have collected a corpus of listener vocalizations [6] as well as a simple implementation of listener vocalization synthesis in the unit selection framework [5]. An important limitation with the initial unit selection approach to the synthesis of listener vocalizations is the fact that we can only generate the vocalizations that have been recorded. If we require additional vocalizations, such as an existing segmental form but with a meaning that had not been produced by the original speaker during the recording session, then the simple selection algorithm can only produce the vocalization most similar to the target – which may not be acceptable.

In order to extend the space of options, we have experimented with *cross-combinations* of segmental form and intonation contour [7]: using signal processing, we impose one vocalization's intonation contour onto another vocalization.

The present paper develops this work further. We describe an extended algorithm for selecting both candidate units and intonation contours, and for combining them. In an evaluation experiment, we assess the meaning of combined stimuli compared to the original vocalizations from which they were combined, to test the hypothesis that this approach can bring non-matching vocalizations closer to the intended meaning and thus reduce the sparse data problem in the synthesis of vocalizations.

The paper is structured as follows. We start by describing the data and annotations we used as the basis for the current work. We then describe the extended selection algorithm used in the run-time synthesis system. We describe the experiment including the research question, stimuli, and test procedure. We then present the results, discuss their meaning for the research question, and conclude with ideas for future work.

2. Data collection and annotation

To collect natural listener vocalizations from dialogue speech, we recorded about half an hour of free dialogue [6] with a professional female British actor with whom we had previously recorded a cheerful expressive speech synthesis database. The actor was instructed to participate in a free dialogue, but to take predominantly a listener role. We encouraged her to use "small sounds that are not words", such as *mm-hm*, where it felt natural, in order to keep her interlocutor talking. However, she was also allowed to "say something" in the conversation where this "felt natural" to keep the dialogue going.

Listener vocalizations were marked on the time axis. Their segmental form was transcribed as a single (pseudo-)word, such as *myeah* or *(laughter)*. The dialogue speech contains 174 spontaneous listener vocalizations from the actor; the most frequent segmental forms are *yeah*, *(sigh)*, *(laughter)*, *mhmh*, *(gasp)*, *oh*.

Annotation of the meaning of these listener vocalizations proceeded as follows. 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 vocalizations [6]; the most frequently used annotations of the SEMAINE corpus [8]; and a set of affective-epistemic descriptors used to describe visual listener behavior [9]. The three sources were consolidated into a list of 11 descriptors: *anger, sadness, amusement, happiness, contempt, solidarity, antagonism, certain, agreeing, interested*, and *anticipation*.

We obtained annotation of meaning for a subset of 23 vocalizations as part of a study investigating the effect of segmental form and of intonation on the perceived meaning of listener vocalizations [10]. About half of the vocalizations annotated differed in segmental form, but had approximately the same intonation contour (low and slightly falling); the other half had approximately the same segmental form (*yeah*) but varied in intonation contour. In a listening test, 20 subjects characterized each vocalization using the 11 descriptors. In order to account for the expected inherent ambiguity of the listener vocalizations, descriptors were presented as scales, and subjects were asked to rate each vocalization on each of the scales.

The meaning of the unmodified vocalizations used in the present study is based on the median of the 20 ratings.

3. Overview of the approach

The basic idea of our approach, as shown in Figure 1, is to combine unit selection principles with signal post-processing to impose a suitable intonation contour onto an approximately suitable vocalization. Given a request formulated using speech synthesis markup, we construct a target unit representing the ideal vocalization. A target cost function is used to select the best candidate from among the available recordings in the given voice. The target unit is also used to select a suitable intonation contour, which is then imposed onto the selected unit. The approach is implemented in our unit selection synthesis framework MARY (http://mary.dfki.de).
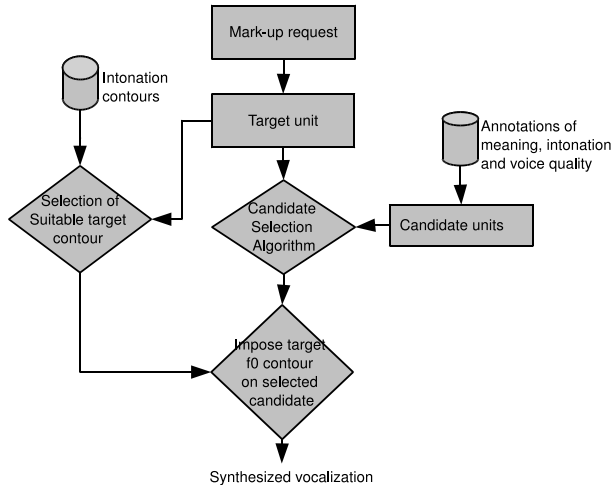

Figure 1: Overview of the approach

3.1. Markup

To support the generation of vocalizations, a new element `<vocalization>` is introduced into the MARY-specific markup format MaryXML [7]. It allows a user to request a vocalization by describing the intended *meaning*, *intonation*, *voice quality*, and *name* (i.e., segmental form) of the vocalization.

3.2. Selecting the best unit candidates

Unit selection principles are used to select the best candidate vocalization for a given request. A unit in this case represents the entire vocalization; therefore, our cost function uses only target costs, no join costs. A target unit is created from the markup request, containing as features the values given in the markup attributes, or "unspecified" if the respective attribute is omitted. Each candidate unit represents one recorded vocalization.

The target cost is a weighted sum of feature costs. In the case of unit candidate selection, we give more weight to the segmental form rather than meaning of the vocalization. This reflects the choice that, if at all possible, the segmental form of the selected vocalization should be realized as requested.

The cost function uses a manually created similarity matrix for each feature. Compared to the classical evaluation function, which assigns cost 0 for equal values and cost 1 when values are different, the similarity matrix has the advantage that it can capture the degree of similarity between feature values. Where a unit exactly matching the target is not available, it is preferable (i.e., less costly) to use a similar unit rather than a very different one. For example, the similarity between the segmental forms

'yeah' and 'myeah' is high (resulting in low cost), whereas the similarity between 'yes' and 'no' is low, and thus results in high cost for that feature. We manually fill the similarity matrices and assign the weights to the different features. The special value "unspecified" has cost 0 for all feature values.

The candidate selection algorithm selects a configurable number of best candidates.

3.3. Selecting the best contour candidates

The contour selection algorithm selects a number of best contour candidates among the available vocalizations using the same algorithm as for the unit candidates, except that different weights are used. Whereas the unit candidate selection gives more weight to the segmental form, the contour candidate selection gives more weight to the meaning features and zero weight to the segmental form and voice quality. Therefore, contour candidates are selected irrespective of their segmental forms.

In order to be included in the list of contour candidates, an intonation contour should bring the combined vocalization closer to the target than any of the candidate units by themselves, i.e., it should actually reduce the resulting cost. In order to decide this, we compute the smallest *contour* cost for the *unit* candidates, CC_{min}; only intonation contours that have a cost not greater than CC_{min} are considered.

If there are no suitable contour candidates, the best unit candidate is synthesized without imposing an intonation contour.

3.4. Selecting the best unit-contour pair

In this step, it is decided which of the contour candidates to impose on one of the unit candidates. Given a unit candidate u_i with original contour c_i and a contour candidate c_k, we define the Unit-Contour Cost UCC as the weighted sum of a merged target cost MC, reflecting an estimate of the similarity of the merged vocalization to the target, and an intonation cost IC, with weight factor α, which attempts to reflect the distortions caused by imposing the intonation contour:

$$UCC(u_i, c_k) = MC(u_i, c_k) + \alpha\, IC(c_i, c_k)$$

We compute the merged target cost using the formula,

$$MC(u_i, c_k) = segCost(u_i) + f0Cost(c_k) + vqCost(u_i)$$
$$+ \frac{1}{2}(meaningCost(u_i) + meaningCost(c_k)),$$

where $segCost$ is segmental form cost, $f0Cost$ is intonation cost, $vqCost$ is voice quality cost and $meaningCost$ is meaning cost.

$IC(c_i, c_k)$ is computed as a distance between third-order polynomial approximations of the respective contours [11]. The pair of unit and contour candidates that minimizes the Unit-Contour Cost is selected.

3.5. Imposing a target intonation contour

The selected unit and contour are combined using the Frequency-Domain Pitch-Synchronous Overlap Add (FD-PSOLA) algorithm [12]. In order to make the synthesized vocalization insensitive to unvoiced regions and large pitch excursions, a third order polynomial approximation of the source contour is used as the target contour for imposition. A reduced copy of the original intonation is used in case of extreme pitch ranges to reduce the effect of distortions.

4. Evaluation

The approach described in Section 3 allows us to synthesize arbitrary combinations of segmental forms and intonation contours. In this experiment we investigate whether the meaning of a synthesized vocalization can be modified towards an intended target meaning by imposing a suitable intonation contour onto an original listener vocalization. Since the signal modification has been shown to produce noticeable distortions, especially for large modifications [7], we avoid large modifications.

We hypothesize that the approach described above makes the synthesized vocalizations convey a meaning closer to the intended meaning than an unmodified original in cases where no suitable match for the requested vocalization exists in the corpus. In order to test the hypothesis, we use a pairwise comparison test where participants are requested to indicate which stimulus in the pair seems more appropriate for a given meaning. We chose this approach because we expect the pairwise presentation to make apparent more fine-grained distinctions than separate scale ratings.

We prepared three types of stimuli – the original vocalizations of unit candidates (*A*), original vocalizations of contour candidates (*B*), and the synthesized vocalizations resulting from imposing B's contour onto A (*AB*).

We selected 11 combinations of meanings and segmental forms to create stimuli, making sure that the segmental form with the intended meaning is not available in the corpus. Therefore, the unmodified vocalizations were expected not to convey the intended meaning. We tried to cover a reasonable range of segmental forms and meanings in the stimuli.

To evaluate the new approach, the three types (*A*, *B*, *AB*) of stimuli were generated for the selected target combinations of segmental form and meaning. Table 1 shows the 33 stimuli with corresponding segmental forms, stylized intonation shapes, and meanings as previously annotated. It can be seen from Table 1 that the stimuli of type A were annotated as "not appropriate" or "somewhat appropriate" for the intended meaning, whereas most stimuli of type B were rated as "very appropriate". To the extent that the meaning of B is conveyed by the intonation contour, we would expect that imposing an approximation of B's contour on A (=AB) should increase the appropriateness.

The pairwise comparison test is divided into three parts: *AB-A*, *AB-B*, and *A-B*. The comparison *AB-A* is aimed to test the hypothesis, whereas *A-B* is primarily a sanity check, verifying the expectation that *B* is generally rated as better for the intended meaning than *A*. In addition, we carried out an *AB-B* comparison to see whether the appropriateness of *AB* reaches that of the original *B*.

The three parts of the evaluation experiment were carried out through a web-based online perception test. Participants were presented with a task description, which included an explanation and examples of listener vocalizations, and made it explicit that synthetic vocalizations would be presented. Subjects were encouraged to use headphones and adjust the playback volume before starting the test.

Participants were asked which one among the two stimuli sounds more appropriate for a given meaning. For example, one question being asked in a comparison test was: *Which one of the following audio examples sounds more like "amusement"?* In total, 21 subjects participated in the online perception test.

5. Results and Discussion

The results of the listening test are shown in Figure 2.

Table 1: Segmental form, intonation contour and previously annotated meaning of stimuli. A: original vocalization of unit candidate; B: original vocalization of contour candidate; AB: synthesized vocalization, with segmental form from A and intonation contour moved towards B (see Section 3.5). def'ly: definitely. Meaning is represented using the following symbols. ○: vocalization is not appropriate for the given meaning; ●: vocalization is somewhat appropriate; ●●: vocalization is very appropriate for the given meaning.

intended meaning	A		B		AB	
amusement	(sigh)	○	yeah	●●	(sigh)	
sadness	def'ly	●	yeah	●●	def'ly	
anger	mh	○	yes	●	mh	
happiness	yes	●	yeah	●●	yes	
solidarity	mhmh	●	yeah	●●	mhmh	
antagonism	def'ly	○	really	●●	def'ly	
uncertain	yeah	●	(sigh)	●●	yeah	
interested	gosh	●	yeah	●●	gosh	
agreeing	mhmh	○	yeah	●●	mhmh	
disagreeing	mhmh	○	yeah	●	mhmh	
high anticipation[1]	gosh	○	def'ly	●	gosh	

Figure 2 (c) shows the A-B comparison for the 10 usable stimulus pairs[1]. The ratings generally matched the expectation that B should be perceived as closer to the respective meaning category than A. For *interest*, no clear preference between the two original vocalizations was found. This is not necessarily in conflict with the previous ratings (Table 1), where both A and B were described as somewhat interested.

Figure 2 (a) shows the results that directly address our research question whether imposing a suitable intonation contour makes a vocalization more suitable for the intended meaning. Globally, the findings confirm the hypothesis. On average, the modified stimuli (AB) are preferred over the unmodified vocalizations (A) in 60% of the cases. This effect is statistically significant (Exact Binomial Test, two-sided $p_{binomial}(124, 207) < 0.01$). A closer look at Figure 2 (a) shows an inhomogeneous picture, however. For *amusement*, *happiness* and *interested*, the intonation contour improved the recognition as the intended meaning category significantly (Exact Binomial Test, $p < .05$ or better). A significant effect in the opposite direction is found for *disagreeing*: here, the combined stimulus is rated as consistently *less* appropriate than the unmodified original. Most of the remaining pairs showed a trend towards a preference for the AB vocalization, but the effects did

---

[1]The comparison A-B was included as a sanity check to make sure that subjects understood the terms used. For one meaning category, *high anticipation*, subjects nearly unanimously chose the *opposite* of the expected meaning. We conclude that they misunderstood the intended meaning (viz., as something that was highly anticipated and predictable), and therefore removed the data from further analysis.
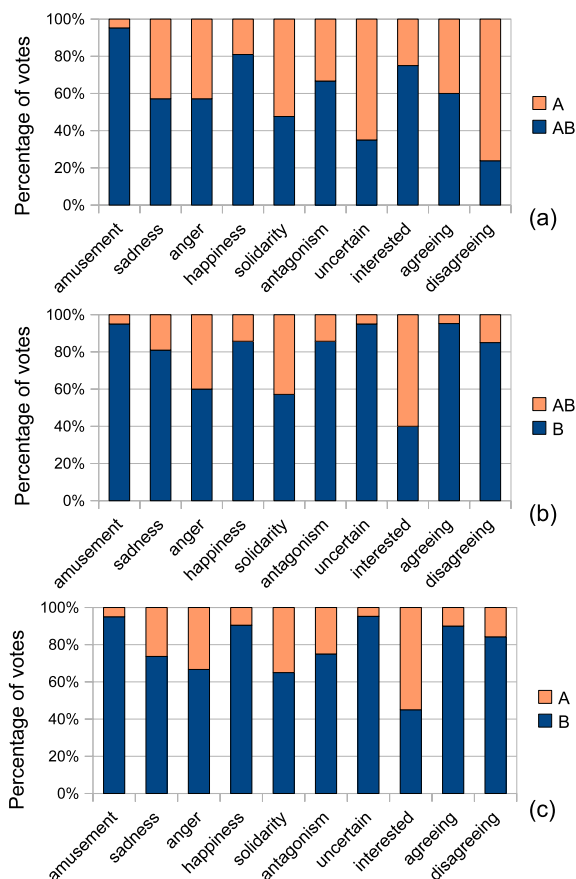
Figure 2: Percent of vocalizations rated as more appropriate for the given meaning, comparing (a) original A with synthesis AB; (b) original B with synthesis AB; and (c) originals A and B.

not reach significance individually.

Regarding the comparison between the modified vocalization AB and the vocalization B from which the respective intonation contour was taken (Figure 2 (b)), we find a very strong global effect of B being preferred over AB (Exact Binomial Test, two-sided $p_{binomial}(45, 205) < 0.001$). This effect is also significant for all individual pairs except for *anger*, *solidarity* and *interested* (Exact Binomial Test, $p < .01$ or better).

Figures 2 (b) and (c) show a nearly identical pattern: obviously, imposing the approximate intonation contour was not sufficient to reach the appropriateness of the original.

These findings appear to confirm, with qualifications, the hypothesis tested in this paper: that a vocalization's suitability for a certain meaning can be improved by imposing on it an intonation contour taken from a "good example" for the intended meaning. The finding is particularly strong with the stimuli using the most extreme intonation contour in the set: out of the four stimuli using the high-fall contour as a target (see Table 1), three are rated as significantly more appropriate: *amusement*, *happiness*, and interested (see Figure 2 (a)). It may be that for the remaining stimuli, the intonation contours of source A and contour target B were actually too similar, so that the perceptual effect on AB may have been too subtle.

6. Conclusion

This paper has presented an algorithm in the unit selection domain for increasing the range of vocalizations that can be synthesized with a given set of recordings. The algorithm takes a vocalization with the intended segmental form and imposes an intonation contour from another vocalization with the intended meaning onto it using FD-PSOLA. In a listening test, the modified versions were rated as significantly more appropriate for the intended meaning category than the unmodified vocalizations. This appears to confirm that the algorithm can make available for use in synthesis combinations of segmental form and meaning that have not been recorded.

The effect was clearest in the cases where an extreme intonation contour was imposed. It may be suboptimal to favor target contours that are as similar as possible to the source contour; it remains to be seen to what extent the benefits of using a markedly different contour outweigh the cost of more perceivable distortions. Alternatively, if the annotation of vocalizations included the extent to which segmental form, voice quality and intonation conveyed a certain meaning, we could use only contours that are informative. For the moment, however, we do not see how to obtain such annotation with reasonable effort.

8. References

[1] N. Pfleger and J. Alexandersson, "Modeling non-verbal behavior in multimodal conversational systems," *Information Technology*, vol. 46, no. 6, pp. 341–345, 2004.

[2] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive ECA system," in *Proc. AAMAS*, 2009, pp. 1399–1400.

[3] V. H. Yngve, "On getting a word in edgewise," in *Chicago Linguistic Society. Papers from the 6th regional meeting*, vol. 6, 1970, pp. 567–577.

[4] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, no. 1, pp. 1–26, 1992.

[5] E. Bevacqua, S. Pammi, S. Hyniewska, M. Schröder, and C. Pelachaud, "Multimodal backchannels for embodied conversational agents," in *Proc. Intelligent Virtual Agents*. Philadelphia, USA: Springer, 2010, pp. 194–200.

[6] S. Pammi and M. Schröder, "Annotating meaning of listener vocalizations for speech synthesis," in *Proc. Affective Computing & Intelligent Interaction*, Amsterdam, The Netherlands, 2009.

[7] S. Pammi, M. Schröder, M. Charfuelan, O. Türk, and I. Steiner, "Synthesis of listener vocalisations with imposed intonation contours," in *Proc. Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, 2010.

[8] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. ICME*, Singapore, 2010, pp. 1079–1084.

[9] E. Bevacqua, D. Heylen, C. Pelachaud, and M. Tellier, "Facial feedback signals for ECAs," in *AISB 2007 workshop "Mindful Environments"*, Newcastle, UK, 2007, pp. 147–153.

[10] S. Pammi and M. Schröder, "Relevance of intonation and segmental form on the meaning perception of listener vocalizations," submitted to Affective Computing and Intelligent Interaction (ACII), Memphis, USA, 2011.

[11] K. Fujii, H. Kashioka, and N. Campbell, "Target cost of f0 based on polynomial regression in concatenative speech synthesis," in *Proc. ICPhS*, Barcelona, Spain, 2003, pp. 2577–2580.

[12] E. Moulines and W. Verhelst, "Time-domain and frequency-domain techniques for prosodic modification of speech," in *Speech coding and synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, pp. 519–555.