Synthesis of listener vocalisations with imposed intonation contours

Sathish Pammi[1], *Marc Schröder*[1], *Marcela Charfuelan*[1], *Oytun Türk*[2], *Ingmar Steiner*[1]

[1]DFKI GmbH, Saarbrücken, Germany
firstname.lastname@dfki.de
[2]Sensory Inc., Portland, OR, USA
oturk@sensoryinc.com

Abstract

Synthesis of listener vocalisations is one of the focused research areas to improve emotionally coloured conversational speech synthesis. To communicate different intentions, a synthesiser should be capable of generating a broad range of vocalisations with different kinds of acoustic properties. However, the data collection for corpus based methods is necessarily limited in acoustic variability. This paper describes our approach to increase the acoustic variability of vocalisations in terms of intonation. After selecting the best candidate for a given target from among the available vocalisations, we use prosody modification techniques to impose a target intonation contour. In an experiment, we combine markedly distinct intonation contours with vocalisations differing in segmental form, using the prosody modification techniques MLSA vocoding, FD-PSOLA, and HNM. In a listening test, we evaluate the perceived naturalness of the resulting synthesised vocalisations, and assess the effect of segmental form, intonation contour and modification technique on perceived meaning.

Index Terms: listener vocalisations, pitch modification, FD-PSOLA, HNM, MLSA Vocoding

1. Introduction

Listener vocalisations play an important role in communicating listener intentions while the interlocutor is talking. They include non-linguistic vocalisations like *uh-huh*, *mhm*, *(laughter)*, and *(sigh)* as well as verbal response tokens such as *yes*, *right*, *really*, and *absolutely*.

In multimodal human-computer interaction, the ability of systems to generate vocal listener behaviour [5] is an important requirement for generating affective interaction. For example, embodied conversational agents (ECA) are one kind of interactive agents used to speak utterances generated by a Text-to-Speech (TTS) system. For the generation of vocal backchannels, an ECA should be able to use the same voice with which it speaks. Nowadays many good TTS systems are corpus based systems. So, the major challenge is not only to collect listener vocalisations from the same speaker with whom we recorded data for synthetic voice, but also to obtain vocalisations with good naturalness in quality and sufficient variability in quantity.

We have described a method [11] to collect natural listener vocalisations from dialogue speech. While the method seems successful in providing recordings of natural vocalisations, the variability of those vocalisations is limited in view of the prosody and segmental form of vocal behaviour. The vocalisations, most of the time, are not only conversation and situation specific, but also depend on the listener's personality. It seems difficult to obtain more varied material regarding the listener vocalisation of a single person, which would be required to cover a broad range of meanings.

Both the segmental form and the various aspects of prosody (intonation, rhythm, and voice quality) convey meaning in listener vocalisations (e.g., [1]). The present paper focuses on intonation, which is known to carry meaning on multiple levels, including linguistic accentuation and phrasing [13], emotion [9], and organisation of the discourse [3]. While the linguistic role of intonation can be expected to be minor on single-word or non-verbal utterances, the other types of meaning are expected to be present. It is an open question whether this meaning is independent of the segmental form of the vocalisation; possibly, non-trivial interactions between the segmental form and the intonation contour might occur.

In order to generate a greater variety of listener vocalisations, we impose intonation contours on recorded listener vocalisations using prosody modification techniques. In the present experiment, we cross-combine naturally observed intonation contours and segmental forms from the same speaker; the approach could also be used with other sources of intonation contours such as automatically generated contours, or contours extracted from another speaker's recordings. As prosody modification techniques, we use Mel-Log Spectral Approximation (MLSA) vocoding, Frequency-Domain Pitch-Synchronous Overlap-Add (FD-PSOLA) and Harmonics-plus-Noise Modelling (HNM). In a listener experiment, we assess the quality of the synthesised stimuli resulting from the application of different prosody modification techniques; we also investigate the effects of segmental form, intonation contour and modification technique on the perceived meaning of the vocalisation.

The paper is structured as follows. Our general approach is outlined in Section 2. In Section 3, we describe an experiment investigating the proposed approach. Section 4 discusses the results. In the Conclusion, we also outline future work on the synthesis of listener vocalisations.

2. 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 vocalisation. Given a request formulated using speech synthesis markup, we construct a target unit representing the ideal vocalisation. 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).

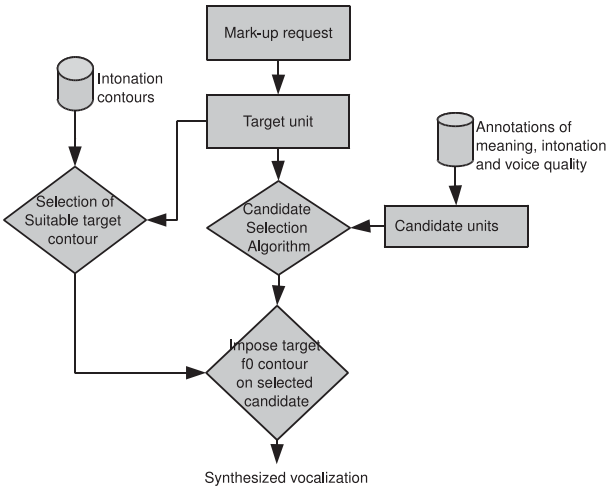

Figure 1: Overview of the approach

2.1. Markup

To support the generation of non-verbal and quasi non-verbal vocalisations such as backchannels, a new element <vocalization> is introduced into the MARY-specific markup format MaryXML. It allows a user to request a vocalisation based on the following criteria:

- meaning: the intended meaning of the vocalisation;
- intonation: the type of intonation contour used on the vocalisation;
- voice quality: the voice quality used with the vocalisation;
- name: a description of the segmental form of the vocalisation.

An example of the markup request is shown in Figure 2. All of the attributes of the <vocalization> tag are optional; if an attribute is not given, this means that the search is not constrained on that level.

```
<maryxml>
  <voice name="dfki-poppy">
    <vocalization
      name="right"
      meaning="accept"
      intonation="falling"
      voicequality="modal"/>
  </voice>
</maryxml>
```

Figure 2: Example of MaryXML markup requesting the generation of a vocalisation.

Table 1 lists the possible values currently supported for each of the criteria; note that not all values are available with every voice. The list of values results from a provisional annotation of the vocalisations of four British English speakers [12]. Clearly, the list of values needs to be broadened; notably, the values for describing intonation contours are insufficient. Given the fact that linguistically motivated descriptions of intonation, such as ToBI [13], are probably inadequate for the emotional and discourse-oriented meanings found in listener vocalisations, it is not straightforward to select an appropriate descriptive scheme for intonation contours. A thorough investigation of the matter is needed, but is beyond the scope of the present paper.

Attribute	Possible values
meaning	anger, sadness, amusement, happiness, contempt, certain, uncertain, agreeing, disagreeing, interested, uninterested, low-anticipation, high-anticipation, low-solidarity, high-solidarity, low-antagonism, high-antagonism
intonation	rising, falling, high, low
voicequality	modal, creaky, whispery, breathy, tense, lax
name	yeah, yes, mhmh, mhm, right, tsright, tsyeah, aha, (snort), (sigh), definitely, really, gosh, ah I see, oh god (gasp), yeah absolutely

Table 1: Values currently supported for each of the attributes of the <vocalization> element in MaryXML.

2.2. Selecting the best candidate

Unit selection principles are used to select the best candidate vocalisation for a given request. A unit in this case represents the entire vocalisation; 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 vocalisation; the unit features stem from manual annotation of the recordings [12], which is currently preliminary.

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 in a similarity matrix.

2.3. Imposing a target intonation contour

Using a selection process similar to the selection of the actual unit, we select an intonation contour from a collection of intonation contours using a target cost mechanism. Prosody modification techniques are used to enforce this intonation contour on the selected unit.

The present paper investigates the properties of this last aspect in the work flow: the effect of imposing a markedly different intonation contour onto a listener vocalisation.

3. Experimentation

The present experiment aims to identify the effects of applying different signal modification technologies when imposing intonation contours on vocalisations. The experiment is designed to address the following two research questions:

1. How good is the perceived naturalness of the resulting listener vocalisations after imposing an intonation contour (depending on the signal modification technology used)?

2. How does the meaning of the listener vocalisations change when cross-combining segmental form and intonation contour?

In the following subsections, we first describe the database of listener vocalisations used, and give an account of the signal modification technologies used to impose the intonation contours. We then describe how we created stimuli and carried out the listening test.

3.1. Database

As listener vocalisations appear natural only in conversation, free dialogue of around 30 minutes was recorded with a professional female British actor with whom we had previously recorded a cheerful expressive speech synthesis database: the voice *dfki-poppy*, available with MARY TTS 4.0. 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 for as long as possible. However, she was also allowed to "say something" and therefore to become the speaker in the conversation where this "felt natural" to keep the dialogue going.

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

3.2. Signal modification techniques

We used three state-of-the-art signal modification techniques to impose the target intonation contours onto the vocalisations.

3.2.1. MLSA vocoding

The MLSA or MGLSA (Mel-Generalised Log Spectral Approximation) vocoder is a digital filter for speech synthesis included in the HTS HMM-based synthesis engine [17]. In the MARY framework this engine has been ported to Java, and the MLSA vocoder has been enhanced to use mixed excitation as in [19]. The mel-generalised cepstral coefficients used in this vocoder are extracted with SPTK [6] and the pitch contour with Snack [14]; pitch modification for the different vocalisations is realised resizing the target prosody to a candidate number of frames. Mixed excitation is realised with ten Fourier magnitudes for pulse excitation generation and five bandpass voicing strengths for better pulse/noise spectral shaping. Fourier magnitudes are calculated on the residual signal, obtained by inverse filtering, by detecting the first ten pitch harmonic peaks in the residual spectrum. Bandpass voicing strengths are estimated by filtering the signal into five frequency bands and calculating peak normalised cross correlation in each band. Voicing strengths and Fourier magnitudes were calculated with SPTK and Snack. Mixed excitation is calculated as follows: a pulse train is generated by inverse Fourier transform of the Fourier magnitudes for one pitch period. The pulse train and noise are passed through the five spectral shaping filters and then added together to give a full band excitation. For each frame, the frequency shaping filter coefficients are generated by a weighted sum of fixed bandpass filters. The pulse filter is calculated as the sum of each of the bandpass filters weighted by the voicing strength in that band. The noise filter is generated by a similar weighted sum, with weights set to keep the total pulse and noise power constant in each frequency band [7].

3.2.2. Frequency domain pitch synchronous overlap-add

FD-PSOLA employs linear prediction to compute the spectral envelope and the excitation spectrum using pitch synchronous speech frames [8]. Pitch modification is achieved by linear interpolation of the spectral envelope. The residual spectrum is either shortened or expanded to match the new size of the spectral envelope. The modified spectral envelope and residual spectrum is then multiplied and the time-domain waveform is obtained by an inverse Fourier transform.

The prosodically modified speech signal is generated using time domain overlap-add operations. The major advantage of FD-PSOLA is its ease of implementation. Frequency domain operation makes it straightforward to perform spectral envelope modifications such as speaker identity transformation or normalisation. Similar to other PSOLA variants, FD-PSOLA lacks the functionality to provide explicit control of phase continuity. Therefore, when used in the context of concatenative synthesis, it may lead to discontinuities at concatenation boundaries. TD-PSOLA, the time domain equivalent of FD-PSOLA, results in degraded quality with increasing amounts of pitch modification [10]. We have obtained similar results in informal evaluations using FD-PSOLA.

3.2.3. Harmonics Plus Noise Model (HNM)

In order to provide better control over phase continuity, we have implemented the harmonics plus noise framework as described in [15]. HNM models the lower frequency portion of the speech signal using a set of harmonically related sinusoids. The difference between the original signal and the signal re-synthesised from the harmonic part is modelled as bandpass filtered random noise. The frequency boundary between the two bands is dynamically computed by analysing and separating harmonic peaks from noisy peaks and then smoothing the result over consecutive speech frames.

Pitch modification is performed by computing a new set of harmonics according to the pitch scaling ratio while preserving the spectral envelope shape. The modified speech signal is obtained by interpolating phases and amplitudes across successive synthesis frames. Explicit phase interpolation reduces discontinuities at concatenation boundaries. As a variation of the original algorithm, we have used the waveform corresponding to the noise part instead of employing the bandpass filtered noise model. This approach enables perfect reconstruction when no pitch modification is performed. The modified noise part generation uses simple overlap-add since no pitch modification is required for the noise part. It appears that less distortions can be expected from HNM based signal modification compared to PSOLA [16].

3.3. Creation of stimuli

To create stimulus material, three vocalisations were chosen through a semi-automatic process. We first applied a K-means algorithm to cluster the 174 vocalisations based on their intonation contours, using as criterion the second-order polynomial distance proposed by [4]. Out of the resulting clusters, we selected a set of 17 vocalisations differing in segmental form. As a final step, we chose three vocalisations that were as different from one another as possible, in order to cover a reasonable range of segmental form as well as markedly different intonation contours.

As shown in Figure 3, the vocalisations have approximately the same length, and are voiced throughout. They are described

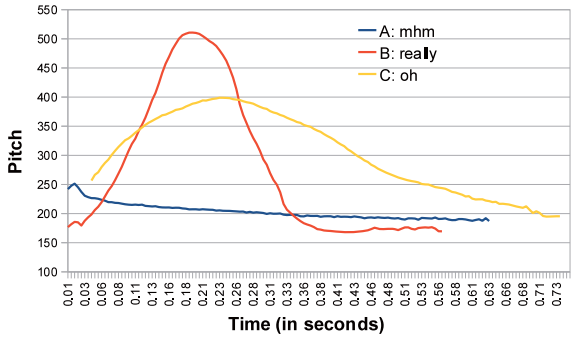

Figure 3: Intonations of the listener vocalisations

as follows:

- mhm (A): low-falling contour with very narrow range;
- really (B): high-low jump with a very large F_0 range;
- oh (C): slow melodious fall from high to mid-range.

From these three original vocalisations, the synthesised stimuli were created as follows. Each of the three vocalisations was synthesised with each of the three intonation contours, using each of the three signal processing techniques (MLSA, FD-PSOLA and HNM). Although the vocalisations are mostly voiced, F0 interpolation is used in all those cases where there is a voiced-unvoiced mismatch between candidate and target contour regions; unvoiced regions in candidate contour were always retained. In total 27 synthetic stimuli are generated, out of these, nine are re-synthesised using the original intonation contour of the respective vocalisation, and 18 are cross-synthesised using the other two intonation contours. We used these 27 synthetic plus the three original vocalisations as stimuli in the listening test. The original stimuli are included to provide reference data regarding meaning and naturalness ratings. The re-synthesised stimuli are included to provide insights in the effect of the signal modification algorithms as such, irrespective of a change in intonation contour. The cross-synthesised stimuli, finally, measure the effect of segmental form and intonation contour on ratings of meaning, and show the amount of degradation due to large modifications in intonation.

3.4. Listening test

A web-based listening test was conducted. Participants were presented with a task description, which included an explanation and examples of listener vocalisations, and made it explicit that synthetic vocalisations would be presented. Subjects were encouraged to use headphones and to adjust the playback volume before starting the test. The 30 stimuli were presented in an individually randomised order. Each stimulus, which could be re-played as often as the subject wished, had to be characterised using twelve five-point scales. The first scale measured the perceived naturalness (from 1 = completely artificial to 5 = completely natural). The remaining eleven scales were used to assess various aspects of meaning. This set of scales was consolidated from three different sources. It includes the most frequent categories used in a previous study on annotating listener vocalisations [11]; the most frequently used annotations of the SEMAINE database [18]; and the affective-epistemic descriptors used to describe visual listener behaviour [2]. The meaning descriptors include seven unipolar scales: degree of *anger*, *sadness*, *amusement*, *happiness*, *contempt*, *solidarity* and *antagonism* (from "absolutely no X" to "pure uncontrolled X"), as well as four bipolar scales: *certain/uncertain*, *agreeing/disagreeing*, *interested/uninterested*, and *unexpected/anticipated*. Each of the eleven meaning descriptors was presented as a five-point scale. For each of the meaning scales (but not for the naturalness scale), subjects could tick a field "no real impression" if they felt it inappropriate to provide any scale value for a given meaning scale.

Nine subjects (five male, four female) participated in the test, most of whom were university staff from different language backgrounds. Given this heterogeneous pool of raters, any patterns with respect to meaning categorisation are likely to be rather robust and not likely to be strongly culture-specific; however, it can only show a first trend. A larger-scale listening test in collaboration with a Psychology department is ongoing; it uses the same eleven meaning descriptors, but a much broader range of naturally occurring listener vocalisations, and a large pool of monolingual British-English speaking subjects. Given this background, the annotation of meaning in the present test should be considered only as a first peek into the relative effects of segmental form and intonation in the perception of meaning.

4. Results and discussion

4.1. Naturalness

The naturalness ratings of the stimuli are shown in Figure 4. A clear pattern can be observed. First, the original stimuli are rated as most natural. Second, the stimuli which were re-synthesised with their own original intonation contour are slightly less natural. The third group of cross-synthesised stimuli, which are synthesised with a different vocalisation's intonation contour, are substantially less natural. Within each group, HNM synthesis scores best, closely followed by FD-PSOLA, whereas MLSA scores clearly worse.

These findings confirm that the re-synthesis using FD-PSOLA and HNM introduce very few artifacts, whereas the quality already drops somewhat with re-synthesis using MLSA vocoding.

The fact that cross-synthesis is rated less natural than re-synthesis confirms the expectation that larger intonation modifications lead to more distortions. While this is established knowledge for the signal *modification* techniques FD-PSOLA and HNM, it might have been different in the case of MLSA vocoding. Given the fact that the signal is decomposed into a spectral envelope and an excitation and then vocoded from these representations, it would have been conceivable that this technology is more robust to larger F_0 changes. Our findings suggest that this is not the case.

4.2. Meaning

In analysing the ratings of meaning, we first looked at the "no real impression" ratings. Any scales for which more than half of the subjects indicated "no real impression" would be discarded; however, this criterion was never reached, so that all stimuli can be located on every scale.

The ratings of the meaning conveyed by the three original vocalisations *mhm_A*, *really_B* and *oh_C* can be seen in Figure 5 (a). First, it can be seen that all three vocalisations have received only moderate ratings on all scales, indicating that none of them was perceived as "pure uncontrolled" expression of any emotion. *mhm_A* was rated as somewhat sad, showing solidarity, uncertain and disagreeing. *really_B* was slightly amused and happy, showing solidarity, antagonistic, uncertain, clearly interested, and taken unawares. *oh_C*, finally, seems to have a rather

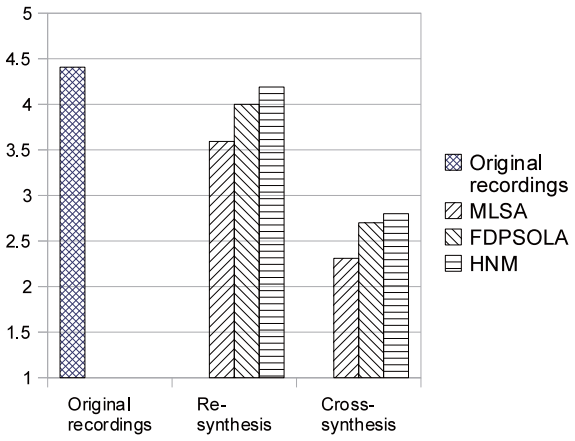

Figure 4: Naturalness ratings, with 1 = *completely artificial* and 5 = *completely natural*. Re-synthesis: the vocalisation is synthesised with its own original intonation contour; Cross-synthesis: the vocalisation is synthesised with the intonation contour taken from a different vocalisation.

diffuse meaning. According to the raters, it could express some sadness and contempt, but also amusement and happiness; it is high on solidarity but also shows some antagonism. In other words, *mhm_A* is a passive expression with negative valence but not directed against the interlocutor. *really_B* is a positive sign of interest and unexpectedness. *oh_C* seems to be an unspecific sign of solidarity with the interlocutor.

Figures 5 (b) and (c) show the extent to which these meanings are stable with the segmental form and with the intonation contour, respectively, when the other element is varied. In fact, it seems that the meaning differences due to segmental form (Figure 5 (b)) are rather small. *oh* is rated higher on solidarity than the other vocalisations, slightly lower on anger and antagonism, and higher on interest; *really* seems to express some antagonism, uncertainty, disagreement, and unexpectedness; *mhm* seems to have an element of disagreement but seems otherwise unspecific.

The rating patterns associated with the intonation contours, across vocalisations, are more conclusive (Figure 5 (c)). Contour *A*, the low and flat contour, is rated consistently high on sadness, low on amusement and happiness, and shows some disagreement and lack of interest. In contrast, contour *B*, the high-low jump, is low on sadness but rather high on amusement, happiness and interest, and has an element of unexpectedness. The ratings for contour *C*, the high melodious fall to a mid range, show no clear pattern.

There were no systematic effects of signal modification method on meaning.

A detailed analysis of the interactions of segmental form and intonation (not displayed here due to space limitations) shows interesting and partially unexpected interactions. *really* is rated as somewhat angry with contours *A* and *C* but not with contour *B*; contours *B* and *C* are rated as more amused and happier with *mhm* and *oh* than with *really*; *really* is rated as quite contemptuous only when combined with contour *A*. Solidarity ratings for contours *A* and *C* are rather low with *really* but high with *oh*. *mhm* is rated as uncertain only with its original contour *A*; it is somewhat disagreeing with contours *A* and *C*, but is neutral or slightly agreeing with contour *B*. *really* is rated as highly interested with its original contour *B* but as quite uninterested with contour *A*.

These findings, even though the details may be questioned due to the small and heterogeneous set of listeners, seem to point out two important trends regarding the relative role of segmental form and intonation contour in determining the meaning of listener vocalisations. First, some but not all intonation contours seem to carry a specific meaning, which survives the combination of the contour with different segmental forms; similarly, some segmental forms seem to carry more specific meaning than others. Secondly, the meaning may change in unexpected ways when cross-combining segmental forms and meaning. For example, none of the ratings of the original vocalisations (Figure 5 (a)) allowed us to predict that *really* with the low and flat intonation contour *A* would convey anger and contempt.

5. Conclusion

We have presented a framework for generating synthetic listener vocalisations in unit selection speech synthesis from markup, using a combination of unit selection and signal modification techniques to generate synthetic vocalisations with more prosodic variety than what is contained in the recorded speech material. We have experimentally investigated the perceptual effects of imposing intonation contours onto a small selection of different vocalisations, using three state-of-the-art signal modification techniques: MLSA vocoding, FD-PSOLA and HNM. Our findings indicate that the drop in naturalness seems strongest for MLSA and smallest for HNM and FD-PSOLA; naturalness degrades substantially when imposing intonation contours that are very different from the original contour, but at least for HNM and FD-PSOLA stays high when re-synthesising the original contour. In line with the literature, we expect this to be a continuous effect, in the sense that smaller changes to the intonation contour should also lead to smaller degradations.

Regarding the meaning of listener vocalisations, we have found distinguishable meanings of some, but not all, segmental forms and intonation contours. Unexpected interactions were observed, where certain configurations of segmental form and intonation caused a perceptual impression that was not predictable from the individual meanings of segmental form and intonation separately. This means that, when synthesising from meaning-level markup, caution seems to be of order when combining segmental forms and intonation contours.

Future work will address both technical and conceptual aspects. Conceptually, an appropriate abstraction needs to be found for representing intonation in markup. Also, a broad and systematic investigation of a continuum of intonation contours would be needed to properly assess the extent to which intonation contours can be changed gradually without causing unexpected effects on meaning. This would allow us to change nuances in meaning by gradually changing properties of intonation contours.

On the technical level, one area for improvement is the automatic selection of a suitable intonation contour to impose onto the selected vocalisation. The mechanism currently selects a single best contour, but this may be changed in the future to provide a set of n-best contours; a distance measure comparing the candidate contours to the original intonation contour of the selected unit could then be used to choose the actual intonation contour to impose. Providing more freedom to select a contour requiring less signal modifications should result in less distortions.

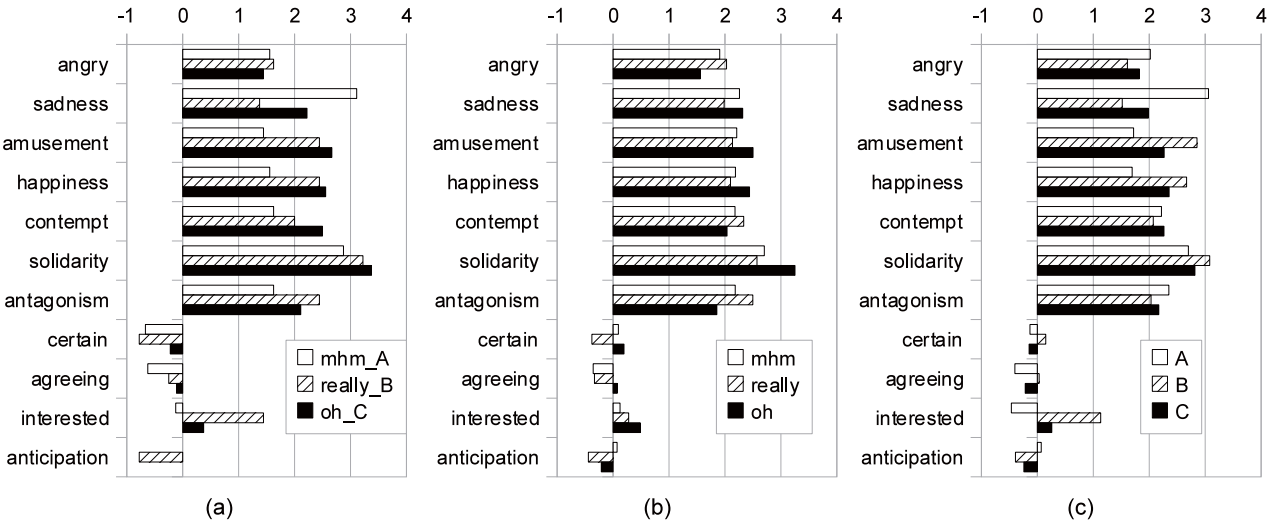

Figure 5: Average ratings of meaning, for (a) the three original vocalisations, (b) the three segmental forms (averaged over different intonation contours), and (c) the three intonation contours (averaged over different segmental forms). Scale values range from 1 (absolutely no X) to 5 (pure uncontrolled X) for the unipolar scales *angry*, *sadness*, *amusement*, *happiness*, *contempt*, *solidarity*, and *antagonism*, and from -2 to 2 for the bipolar scales, where -2 = *totally uncertain*, *totally disagreeing*, *totally disinterested* and *totally taken unawares*, and 2 = *totally certain*, *totally agreeing*, *totally interested*, and *anticipated events completely*.

Another technical challenge worth addressing is the possibility to generate vocalisations from normal diphone speech units. This would reduce the need to record vocalisations in dialogue speech. Given the lack of control for phase continuity in FD-PSOLA, we would expect larger distortions with FD-PSOLA than with HNM.

6. Acknowledgements

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) and from the DFG project PAVOQUE.

7. References

[1] J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26, 1992.

[2] E. Bevacqua, D. Heylen, C. Pelachaud, and M. Tellier. Facial feedback signals for ECAs. In *AISB 2007 Annual convention, workshop "Mindful Environments"*, pages 147–153, Newcastle, UK, 2007.

[3] D. L. Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[4] K. Fujii, H. Kashioka, and N. Campbell. Target cost of F_0 based on polynomial regression in concatenative speech synthesis. In *Proceedings of the 15th International Conference of Phonetic Sciences*, pages 2577–2580, Barcelona, Spain, 2003.

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

[6] T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, et al. Speech signal processing toolkit (SPTK), version 3.3. http://sp-tk.sourceforge.net/, 2009.

[7] A. V. McCree and T. P. Barnwell. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing*, 3(4):242–249, 1995.

[8] E. Moulines and W. Verhelst. Time-domain and frequency-domain techniques for prosodic modification of speech. In W. Kleijn and K. Paliwal, editors, *Speech coding and synthesis*, page 519–555. Elsevier, 1995.

[9] S. J. L. Mozziconacci and D. J. Hermes. Role of intonation patterns in conveying emotion in speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2001–2004, San Francisco, USA, 1999.

[10] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha. Modification of pitch using DCT in the source domain. *Speech Communication*, 42(2):143–154, 2004.

[11] S. Pammi and M. Schröder. Annotating meaning of listener vocalizations for speech synthesis. In *Proc. International Conference on Affective Computing & Intelligent Interaction*, Amsterdam, The Netherlands, 2009. IEEE.

[12] C. Pelachaud, E. Bevacqua, M. McRorie, S. Pammi, M. Schröder, I. Sneddon, and E. de Sevin. SEMAINE deliverable D5a: SAL multimodal generation component with customised SAL characters and visual mimicking behaviour. SEMAINE project deliverable, 2009.

[13] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labeling English prosody. In *Proceedings of the 2nd International Conference of Spoken Language Processing*, pages 867–870, Banff, Canada, 1992.

[14] K. Sjölander. The Snack sound toolkit, version 2.2.8. http://www.speech.kth.se/snack/, 2004.

[15] Y. Stylianou. *Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modification*. PhD thesis, École nationale supérieure des télécommunication, 1996.

[16] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter. TD-PSOLA versus harmonic plus noise model in diphone based speechsynthesis. In *Proc. ICASSP*, volume 1, Seattle, WA, USA, 1998.

[17] K. Tokuda, K. Oura, K. Hashimoto, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, A. W. Black, and others. HMM-based Speech Synthesis System (HTS). http://hts.sp.nitech.ac.jp/, 2010.

[18] M. F. Valstar, G. McKeown, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *Proc. IEEE International Conference on Multimedia & Expo*, Singapore, 2010.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Mixed excitation for HMM-based speech synthesis. In *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001.