interspeech 2008 incorporating SST 08

Emotion Conversion using F0 Segment Selection

Zeynep Inanoglu, Steve Young

Department of Engineering, University of Cambridge, Cambridge, UK

zeynepinan@post.harvard.edu, sjy@eng.cam.ac.uk

Abstract

This paper describes F0 segment selection, a novel syllable-based F0 conversion method, which provides a concatenative framework to search for F0 segments in a modest corpus of emotional speech (∼15 minutes of data). The method is compared with our earlier work on F0 generation using context-sensitive syllable HMMs. Both methods are complemented with a duration conversion module as well as GMM-based spectral conversion to form a unified emotion conversion framework in English. The system was evaluated using three target styles: surprise, anger and sadness. The results of an extensive perceptual test show that segment selection significantly outperforms the HMM-based method in terms of both emotion recognition rates and intonation quality ratings for surprise and anger. For conveying sadness both methods were effective.

Index Terms: emotion conversion, expressive prosody

1. Introduction

The problem of emotion transformation in speech and its possible solutions have been attracting much attention in the field of text-to-speech synthesis (TTS) for rapid generation of target expressive styles. A rudimentary emotion conversion framework can be implemented using rules which modify an input utterance in a deterministic way. Various rule-based transformation attempts exist in the literature (see [1] for a review). However, designing good rules for each expressive style requires manual analysis and can only capture a very limited set of acoustic-prosodic divergences. In recent years, data-driven voice conversion methods have been explored for modeling and transforming both short-term spectra and prosody [2-6]. In [2], GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. In [4], the use of GMM and CART-based F0 conversion methods were evaluated for mapping neutral prosody to emotional prosody in Mandarin speech. In [3], a unified conversion system was proposed using duration embedded Bi-HMMs to convert neutral spectra and decision trees to transform syllable F0 segments. [5] and [6] report methods specific to an HMM-based speech synthesis framework, where emotional prosody and spectra were modeled and adapted jointly using phone HMMs. In [7], we described an emotion conversion system for English which is independent of the underlying synthesis system. This paper also introduced syllable HMMs for generating F0 contours.

In this paper, we present F0 segment selection as an alternative to using syllable HMMs. Both methods adopt a linguistically motivated approach to F0 transformation in an effort to capture the interactions between the affective and linguistic layers of prosody within a single framework. We compare the methods in a unified conversion framework, where transformation of spectra and phone durations are performed to complement the F0 modules. An extensive perceptual study was performed to evaluate the effectiveness of the conversion system as a whole and the F0 conversion modules individually.

The rest of the paper is outlined as follows: In Section 2 we review the HMM-based F0 generation technique. In Section 3 we introduce F0 segment selection as an alternative. In section 4 the experimental setup of the conversion system is outlined and the duration and spectral conversion modules are summarized. Finally, in section 5, we report the results of a perceptual study evaluating the conversion outputs.

2. F0 Generation From Syllable HMMs

The HTS HMM-based speech synthesis framework was used to train syllable F0 models. Note that unlike conventional full-spectrum HMM synthesis, we model F0 only and we do so at the syllable rather than the phone level. Furthermore, because we are operating in a conversion framework, we assume that syllable durations are converted in a prior step and are therefore available as an input to the F0 generation process. Three state context-sensitive Multi-Space Probability Distribution HMMs (MSD-HMM) were used to model each syllable [8]. A combination of seven linguistic features identify each syllable: these are lexical stress (lex), position in word (wpos), position in sentence (spos), part of speech of current word (pofs), part of speech of previous word (ppofs), onset type (onset) and coda type (coda), where the onset and coda are either voiced, unvoiced or sonorant. For training, the voiced segment within each syllable was aligned with the context-dependent syllable models determined by the corresponding linguistic features. The unvoiced regions in the training utterances were modeled using a separate *uv* model which was always aligned with a zero-dimensional unvoiced symbol as defined in an MSD-HMM framework [8]. Decision tree-based parameter tying was performed to compensate for rare and unseen contexts. The parameter generation framework of HTS.1 alpha was used to generate utterance F0 contours for unseen test utterances by concatenating syllable HMMs [9].

3. F0 Segment Selection

Segment selection makes use of a concatenative framework similar to unit selection. A sequence of syllable F0 segments are selected directly from a small expressive corpus, using target and concatenation costs. A syllable F0 segment consists of the voiced part of each syllable. A similar idea has been explored to predict F0 contours in a non-expressive TTS framework from a large corpus of Mandarin speech [10]. The goal of the method described here, however, is to generate expressive prosody from limited data in a conversion framework. Parallel neutral and emotional syllable F0 segments are stored as part of the unit definition as well as their common linguistic context.

Accepted after peer review of full paper

During segment selection, the neutral part of the unit definition is used to calculate how similar an unseen input segment is to previously observed neutral segments in the corpus. The same set of linguistic identifiers were used as in section 2. We define a syllable target cost T and an inter-syllable concatenation cost J such that the total cost over S syllables for a given unit sequence U and input specification sequence I is defined as

$$C(U,I) = \sum_{s=1}^{S} T(u_s, i_s) + \sum_{s=2}^{S} J(u_{s-1}, u_s) \quad (1)$$

The target cost T is a weighted Manhattan distance consisting of P subcosts

$$T(u_j, i_s) = \sum_{p=1}^{P} w_p T_p(i_s[p], u_j[p]) \quad (2)$$

Eight target subcosts (P=8) are used. The first seven are binary subcosts to indicate whether individual context features (e.g. lexical stress) in the specification match the corresponding syllable context in the unit. A matching feature results in zero cost whereas a mismatch results in a unit cost of 1. The final subcost, T_{f0}, is defined using the Root Mean Squared(RMS) distance between the input syllable contour $F0^i$ and the neutral contour which is part of the unit definition, $F0^n$, after the two segments are interpolated to the same length L:

$$T_{f0} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(F0^i(l) - F0^n(l))^2} \quad (3)$$

The weights for each subcost serve two functions: firstly they normalize the different ranges of categorical and continuous subcosts and secondly they rank features according to their importance for each target emotion.

The concatenation cost, J, is nonzero if and only if consecutive syllables in the input specification are "attached", i.e. within the same continuous voiced region. If the voiced syllable segment for the input specification i_{s-1} ends at time t_1 and the input specification i_s begins at time t_2, the concatenation cost for two candidate segments in the target corpus with lengths, L_{s-1} and L_s, is defined as the difference between the last F0 point in segment $F0_{s-1}$ and first F0 point in segment $F0_s$:

$$J(u_{s-1}, u_s) = \begin{cases} w_j \left(F0_{s-1}[L_{s-1}] - F0_s[1]\right) & \text{if } t_1 = t_2 \\ 0 & \text{otherwise} \end{cases}$$

The concatenation cost is included to avoid sudden segment discontinuities within voiced regions. A concatenation weight, w_j is used to prioritize this cost relative to the target subcosts when selecting segments.

Once all the costs are defined, segment selection becomes a problem of finding the path, $\hat{u}$, with the smallest cost through a trellis of possible F0 segments given an utterance specification. Viterbi search is used to find the minimum-cost path, by tracing back locally optimal candidates. Note that the concatenation cost is zero for all syllable voiced segments that are detached from the preceding voiced segments due an intervening unvoiced region or a pause. Therefore if an input utterance consists of only detached syllables, the concatenation cost plays no role in segment selection and the optimal path will simply be the sequence of units which minimize target costs locally at each syllable time step. In fact, we apply a different set of weights for detached and attached syllables. This distinction is motivated by the fact that all weights are likely to change with the introduction of a concatenation cost. The two sets of weights are estimated using a least squares linear regression method on held-out utterances.

For the detached case, a set of P weights, w_p^T, are estimated for each target subcost. For each held out syllable F0 segment in a target emotion, the N-best and N-worst candidates in the corpus are identified in terms of their RMS distance to the held-out segment. This highlights the units we most want our cost function to select and the units we most want it to avoid. The cost functions for these syllable segments are then set equal to their RMS distances, which results in a system of linear equations. Combining the equations for each of the M held-out syllables and 2N candidates yields the following system of 2NM equations which can be solved using least squares:

$$CW = D \quad (4)$$

where C is a $2NM \times P$ matrix of subcosts, W is the $P \times 1$ vector of unknown weights and D is the $2NM \times 1$ vector of distances. In our system N was set to 5 and leave-one out cross-validation was performed on all training utterances to obtain the final system of equations. The weights obtained for detached syllables are listed in Table 1. The different contextual weights indicate which features are most relevant for each target emotion (e.g. position in sentence seems to be an important subcost for surprise yet not as relevant for anger or sadness). Note that the low values for the final weights w_{f0} is due to the fact that subcost T_{f0} inherently has higher values which need to be normalized to match the categorical subcosts.

For attached syllables, a different set of P+1 weights are estimated which includes all the target weights plus the weight for the concatenation cost. A similar estimation process is performed, this time on pairs of held-out segments in order to incorporate the concatenation cost in the linear system of equations. The weights for attached syllables are listed in Table 2. Note that most contextual weights apart from lexical stress are set to zero and the concatenation cost dominates segment selection in the case of surprise and sadness. For anger, the subcost T_{f0} still plays an important role, as evidenced by its higher weight relative to the other emotions (0.68).

Table 1: *Estimated weights for detached syllables across three target emotions*

	Surprised	**Sad**	**Angry**
w_{lex}	13.67	12.30	18.74
w_{wpos}	24.52	11.29	18.47
w_{spos}	11.33	4.91	3.31
w_{pofs}	1.13	4.82	8.82
w_{ppofs}	24.27	6.49	10.54
w_{onset}	15.08	0.33	5.54
w_{coda}	8.23	6.09	6.36
w_{F0}	0.47	0.69	1.00

4. Experimental Setup

The output of the F0 prediction component is combined with a duration conversion module and a spectral conversion technique to form a system of emotion conversion. Spectral conversion is performed first using a pitch-synchrounous LPC analysis/synthesis framework. Durations and F0 contours of the input utterance are modified using the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) implementation provided by the Praat software [11]. All modules use 272 utterances ($\sim$15 minutes) of emotional speech data for training and 28

Table 2: *Estimated weights for attached syllables across three target emotions*

	Surprised	**Sad**	**Angry**
w_{lex}	17.89	6.43	15.98
w_{wpos}	0.0	0.0	0.0
w_{spos}	0.0	0.0	0.0
w_{pofs}	0.0	0.0	0.0
w_{ppofs}	0.0	0.0	0.0
w_{onset}	3.23	0.0	0.0
w_{coda}	0.0	0.0	8.74
w_{F0}	0.27	0.37	0.68
w_{concat}	0.74	0.70	0.48

utterances were set aside for testing [1]. A professional female voice talent recorded parallel speech data in three expressive styles (angry, surprised, sad) as well as a neutral style. All utterances were automatically force-aligned using context-sensitive phone models in HTK [12]. Lexical stress labels and part of speech tags were also extracted automatically.

4.1. Duration Conversion

To convert phone durations, relative regression trees were built for four broad classes: vowels, nasals, glides and fricatives [7]. Rather than absolute durations, the relative trees predict scaling factors which modify the neutral phone durations. For each broad class and emotion, we have identified the set of predictors from a feature pool which minimize RMS error on test data. The feature pool consisted of input neutral durations, phone identity, broad class of left phone, broad class of right phone, lexical stress, position in word, position in sentence and part of speech. Details of duration conversion can be found in [13].

4.2. Spectral Conversion

A GMM-based spectral conversion method is used to map each neutral spectrum to that of a desired target emotion [14][15]. Line spectral frequencies (LSF) were used as the acoustic features to be converted. LSF parameter vectors of order 30 were computed for parallel pairs of neutral-emotional utterances. These were then time-aligned using the state-based phonetic alignments computed using HTK. The number of mixture components was set to 16. An Overlap and Add (OLA) synthesis scheme was used to combine the converted spectral envelope with the neutral (unmodified) residual.

5. Perceptual Evaluation

A direct comparison between F0 conversion methods was facilitated using a three-way preference test. Subjects were asked to compare three utterances which were identical except for the method used to convert the F0 contours: segment selection, syllable HMMs or a simple Gaussian normalization scheme which served as a baseline. This baseline method scales every pitch point, s, in the neutral source contour to match the mean, μ_t and variance σ_t of the target emotion:

$$F(s) = \frac{\sigma_t}{\sigma_s} s + \mu_t - \frac{\sigma_t \mu_s}{\sigma_s} \quad (5)$$

Spectral conversion was applied to all utterances but neutral durations were left unmodified. 30 subjects participated in the test

[1]The data was collected in collaboration with the Speech Technology Group, Toshiba Research Europe.

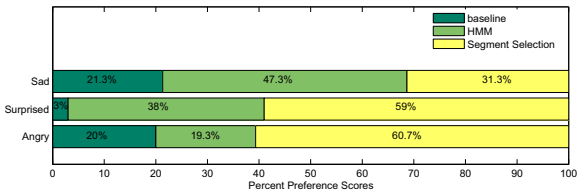

Figure 1: *Preference scores for each F0 prediction method and emotion.*

and each subject performed 10 comparisons per emotion. The preference scores for each emotion are displayed in Figure 2. For anger and surprise segment selection was preferred significantly more often than the other two methods reaching 61% and 59% preference rates (t-tests, $p << 0.01$). For surprise the naive baseline was preferred very infrequently, while for anger there was no significant difference between the baseline and HMM-based contours. In the case of sadness, HMM-based contours were the most popular followed by segment selection ($p << 0.01$).

A further evaluation of the full conversion system was performed using a multiple-choice emotion classification test, where subjects were asked to guess the emotion in an utterance[2]. The test was first conducted using the original emotional utterances of the speaker. A "Can't decide" option was included in the choices to avoid forcing the subjects to choose an emotion. 5 utterances per emotion were presented to 30 subjects. The confusion matrix for this test is summarized in Table 3.

Table 3: *Confusion scores for the emotion classification task for original emotional utterances spoken by the actor*

	Angry	**Surprised**	**Sad**	**Can't decide**
Angry	99.3%	0.7%	0%	0%
Surprised	20.0%	66.0%	0%	14.0%
Sad	0.7%	0%	96.0%	3.3%

The same test was then conducted using converted neutral utterances generated by our conversion system. 10 utterances per emotion were classified by 30 subjects in random order. Duration conversion and spectral conversion were applied to all outputs. Additionally, there were two hidden groups within each emotion: five of the conversions were synthesized using HMM-based contours and the other five were synthesized using segment selection. Confusions between emotions were analyzed separately for the two F0 conversion methods (Table 4 and Table 5). The conversion outputs using HMM-based F0

Table 4: *Confusion scores for the emotion classification task for utterances where HMM-based contours are used*

	Angry	**Surprised**	**Sad**	**Can't decide**
Angry	64.7%	8.0%	4.7%	22.6%
Surprised	10.0%	60.7%	0%	29.3%
Sad	0.7%	0.7%	96.0%	2.6%

contours conveyed sadness as well as the original sad speech, while the recognition rate for surprise (60.7%) was slightly lower than that of the original surprised speech (66%) and the rate for anger (64.7%) was much lower than that of original

[2]Speech samples are available online at http://mi.eng.cam.ac.uk/∼zi201/conversions.html

Table 5: *Confusion scores for the emotion classification task for utterances where F0 segment selection is used*

	Angry	**Surprised**	**Sad**	**Can't decide**
Angry	86.7%	0.7%	0%	12.6%
Surprised	8.7%	76.7%	0	14.7%
Sad	0.7%	0%	87.3%	12%

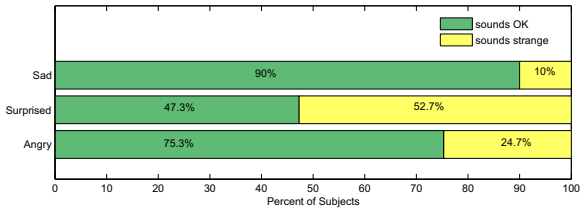

Figure 2: *Categorical quality ratings for spectral conversion + duration conversion + HMM-based contour generation*

anger (99%). There was considerable indecision amongst subjects while classifying surprise and anger. With segment selection, however, the classification rate for anger increased significantly to 86.7%. This indicates that appropriate F0 prediction is a critical component of anger, which is considered to be a voice-quality dominated emotion. Surprise is also recognized significantly better with segment selection (76.7%), in fact, even better than the perception of the original surprised utterances. This may be explained by the imperfection of our spectral conversion module: we believe that the GMM-based spectral conversion smoothed out the spectra slightly to reduce the tension in surprised speech, which may have created confusion between anger and surprise in the original utterances. Overall, the effect of F0 prediction method on emotion recognition rates was significant for all emotions. Segment selection resulted in better recognition in the case of anger ($p = 0.0006$) and surprise ($p = 0.004$), while HMM-based contours resulted in higher recognition scores for sadness ($p = 0.018$) Finally, as part of the emotion classification test, we also asked subjects to categorize each utterance in terms of intonation quality using the options "Sounds OK" or "Sounds Strange." The intonation quality ratings are illustrated in bar charts for each method (Figures 3 and 4). For both methods, the percent quality ratings for sadness are identical and generally very high (90%"sounds OK"). Subjects also thought that both methods attempted to convey anger naturally most of the time, even though the actual emotion recognition rates are very different between the methods. The effect of F0 prediction method on quality perception was, however, significant in the case of surprise ($p = 0.0006$). In the surveys, a number of subjects noted that in some of the utterances, "the surprise element was there" but it was "slightly misplaced", which made them choose the "can't decide" option. Therefore, unlike anger, the recognition rates and quality ratings for surprise were somewhat correlated.

6. Conclusions

This paper has described two novel syllable-based F0 prediction techniques for emotion conversion in English. The techniques were evaluated within a full conversion framework with complementary duration and spectral transformation modules. Both methods of F0 prediction were successful in conveying the desired emotions well above chance level. However, in the case of

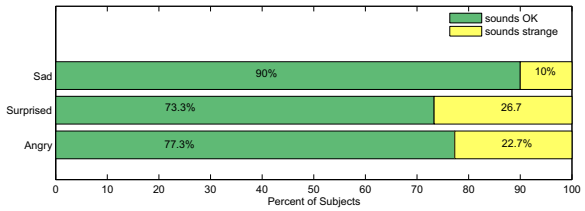

Figure 3: *Categorical quality ratings for spectral conversion + duration conversion + F0 Segment Selection*

anger and surprise, F0 segment selection produced significantly more natural and convincing expressive intonation than our earlier HMM-based technique as evidenced by preference and classification tests. Overall, the methods achieved comparable performance to a human actor. Incorporating more advanced spectral conversion techniques within the conversion framework is likely to boost recognition rates even further particularly in the case of anger.

7. References

[1] Schroder, M., "Emotional Speech Synthesis - A Review", Proc. of EUROSPEECH, vol.1:561–564, 1999.

[2] Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., and Shikamo, K. "GMM-based Voice Conversion Applied to Emotional Speech Synthesis", IEEE Trans. Speech and Audio Proc., 7(6):697–708, 1999.

[3] Wu, C.H., Hsia, C.-C., Liu, T.-E., and Wang, J.-F., "Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis", IEEE Trans. Audio, Speech and Language Proc., vol.14(4):1109–1116, 2006.

[4] Tao, J., Yongguo, K., and Li, A. "Prosody Conversion from Neutral Speech to Emotional Speech", IEEE Trans. Audio, Speech and Lang Proc., vol.14:1145–1153, 2006.

[5] Tsuzuki, H., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S. "Constructing emotional speech synthesizers with limited speech database", Proc. of ICSLP vol.2:1185-1188, 2004

[6] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., "Modeling of various speaking styles and emotions for HMM-Based Speech Synthesis", Proc. EUROSPEECH,vol.3:2461-2464, 2003.

[7] Inanoglu, Z., Young, S., "A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality", Proc. of Interspeech, 2007.

[8] Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., "Multi-Space Probability Distribution HMM", IEICE Trans. Inf. and Systems,vol.3:455-463, 2002.

[9] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T, Kitamura, T., "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis", Proc. of ICASSP,vol.3:1315-1318, 2000.

[10] Tian, J., Nurminen, J., Kiss, I., "Novel Eigenpitch-based Prosody Model for Text-to-Speech Synthesis", Proc. of Interspeech, 2007.

[11] http://www.fon.hum.uva.nl/praat/

[12] http://htk.eng.cam.ac.uk

[13] Inanoglu, Z., "Data-driven Parameter Generation for Emotional Speech Synthesis", PhD Thesis, University of Cambridge, 2008.

[14] Stylianou et al.,"Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. Speech and Audio Proc. vol.6:131-142, 1998.

[15] Ye, H.,"High-Quality Voice Morphing", PhD Thesis, Cambridge University, 2005.