# A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality

*Zeynep Inanoglu, Steve Young*

Department of Engineering, Cambridge University, Cambridge, UK
zi201@cam.ac.uk, sjy@eng.cam.ac.uk

## Abstract

This paper describes a system that combines independent transformation techniques to endow a neutral utterance with some required target emotion. The system consists of three modules that are each trained on a limited amount of speech data and act on differing temporal layers. F0 contours are modelled and generated using context-sensitive syllable HMMs, while durations are transformed using phone-based relative decision trees. For spectral conversion which is applied at the segmental level, two methods were investigated: a GMM-based voice conversion approach and a codebook selection approach. Converted test data were evaluated for three emotions using an independent emotion classifier as well as perceptual listening tests. The listening test results show that perception of sadness output by our system was comparable with the perception of human sad speech while the perception of surprise and anger was around 5% worse than that of a human speaker.

**Index Terms**: expressive speech synthesis, emotion conversion, voice conversion

## 1. Introduction

The ability to add emotions to synthesised neutral speech has become a high priority since the emergence of unit selection methods which require very large training corpora in order to generate highly intelligible and natural outputs. Because replicating a unit-selection corpus for each expressive style is a costly process, signal-processing frameworks are needed for emotion conversion that require only limited training data for each target emotion. Early work in this area has focused on rule-based modifications of acoustic features (see [1] for a review). However designing good rules for each expressive style requires manual analysis and this can only capture a limited set of phenomena. In recent years, stochastic methods have been explored for modelling and transforming both short-term spectra and prosody [2-6]. In [2], GMM-based spectral conversion techniques were evaluated but it was found that spectral conversion alone is not sufficient for conveying the required target emotion. In [4], the use of GMM and CART-based pitch conversion methods were evaluated for mapping neutral prosody to emotional prosody in Mandarin. In [3], a unified conversion system was proposed using duration embedded Bi-HMMs to convert spectra and decision trees to transform syllable F0 segments. [5] and [6] are methods specific to HMM-based speech synthesis, where emotional prosody and spectra were modelled jointly using phone HMMs. In this paper, however, it is argued that prosody and voice quality should be converted separately since we believe they operate at different timescales. We therefore present a set of conversion methods that transform each feature independently in the appropriate temporal layer: syllables for F0, phones for duration and short-term segments for spectra. These independent conversion modules are then combined in the order of most granular(segment) to least granular(syllable) to form a system for emotion transformation.

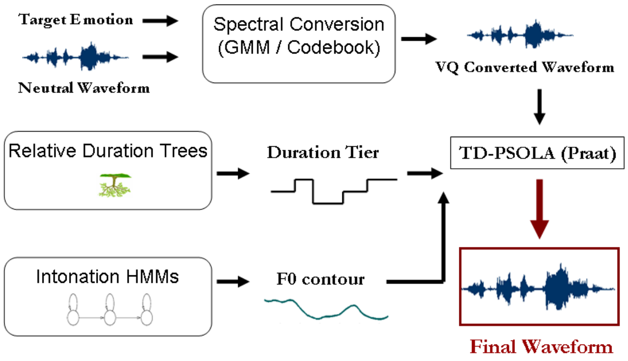

Figure 1: *System Overview*

Our previous work on using syllable HMMs to represent ToBI-based intonation units resulted in the generation of very natural intonation contours with varying accent patterns [7]. In this paper we apply the same HMM-based modelling and generation framework to model context-sensitive syllable F0 segments for each emotion. The goal is to explore the novel use of context-sensitive syllable models to represent both linguistic and emotional F0 variations in an utterance by using a set of perceptually optimized contextual factors. Conversion of phone durations were carried out as a separate module using relative decision trees given the neutral duration and phonetic context. For spectral conversion, we have explored and contrasted two alternative methods: a GMM-based linear transformation method similar to [2] and a codebook selection method. The latter was implemented to evaluate whether a framework for directly transplanting target spectra would preserve the emotional strength better than the former, which simply transforms the source spectra.

The rest of the paper is outlined as follows: Section 2 briefly introduces the overall system. Details of the speech data collection are given in Section 3. Section 4 explains the use of two alternative spectral conversion methods and reports objective results based on root mean squared error (RMSE). Section 5 describes the relative phone duration trees as well as syllable HMMs for F0 generation. Finally in section 6, the system is evaluated both objectively using an emotion classifier as well as subjectively using a traditional forced-choice listening test.

## 2. System Overview

Figure 1 illustrates the three modules of our emotion transformation framework. The spectral conversion step outputs a waveform with the source prosody and converted spectra. The second step is to change the phone durations using the relative CART trees. Finally the converted durations are used to generate a pitch contour for the entire utterance using a sequence of syllable HMMs. The F0 and duration modifications are carried out using the TD-PSOLA algorithm implemented by Praat[8]. Conversion samples can be found at http://mi.eng.cam.ac.uk/∼zi201/interspeech2007.html

## 3. Data Collection

A corpus of emotional data was collected in collaboration with the Speech Technology Group, Toshiba Research Europe. A professional female voice talent recorded parallel speech data in four expressive styles (angry, surprised, happy, sad) as well as a neutral style. In expressing the emotions, she was asked to assume a natural, conversational style rather than a dramatic intensity. Utterances had a sampling rate of 22.05Khz. For the experiments described below, 272 utterances per emotion were used for training and 28 were set aside for testing. All utterances were automatically force-aligned using context-sensitive phone models in HTK [9]. Lexical stress labels and part of speech tags were also extracted automatically. A preliminary perceptual study was conducted on the speech data to evaluate the degree to which the speaker was able to convey the desired emotions. Ten random utterances were chosen for each emotion and presented to 15 listeners in mixed order. Table 1 illustrates the confusion matrix for the listener responses. Happy was highly confusable with neutral. The listeners' informal consensus was that the speaker tried to mimic happiness using a stylized version of neutral speech which did not come across as happy. For this reason, we have excluded happy speech from the experiments reported in this paper.

Table 1: *Perceptual Tests On Emotional Speech Data*

| | Sad | Surp | Hap | Ang | Neu |
|---|---|---|---|---|---|
| Sad | 83.3 | 0 | 0 | 2.7 | 14 |
| Surp | 0 | 65.3 | 22.7 | 11.3 | 0.7 |
| Hap | 4 | 8 | 33.3 | 12.7 | 42 |
| Ang | 0.7 | 2 | 2 | 93.3 | 2 |

## 4. Spectral Conversion

Our analysis of long-term average spectra (LTAS) for vowels revealed emotion-specific patterns. Figure 2 illustrates the LTAS for vowel /ae/ for four speaking styles. In order to capture such differences, we have contrasted two spectral conversion techniques. For both methods, line spectral frequencies (LSF) were used as the acoustic features to be converted. LSF parameter vectors of order 30 were computed from time-domain LPC analysis using a frame size of three pitch-periods and a step size of one pitch period. Three pitch period residuals were stored at every frame step. During synthesis, the residuals were filtered with the converted LSF vectors, a Hamming window was applied at each synthesis frame and overlapping frames were added to give a smooth envelope transition.

### 4.1. GMM-based Spectral Conversion

A GMM-based linear transformation method was adopted for spectral conversion[10]. This method represents the neutral acoustic space with a GMM and trains a linear transformation function for each mixture component. In our case, training data from neutral and emotional speech was aligned using the state-based phonetic alignments returned by HTK. The number of mixtures was a set to 16.

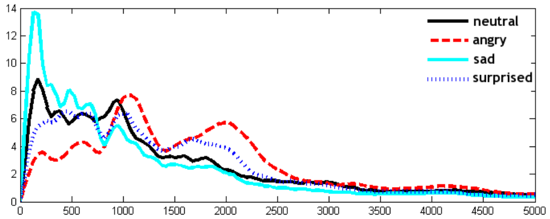

Figure 2: *Long Term Average Spectra for Phone /ae/ in the training data for neutral, angry, sad and surprised utterances.*

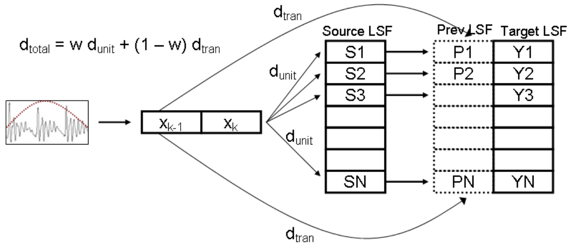

Figure 3: *Codebook Selection*

### 4.2. Codebook-based Spectral Conversion

As an alternative to linear transformation, the codebook approach encourages piecewise transplantation of LSF values directly from the target training data. The aligned neutral and target frames are stored as codebooks along with the left context for each target frame as shown in Figure 3. If the left context of the input frame is unvoiced, than the distance measure is simply the Euclidian distance between the input LSF vector and the source LSF entries. If the left context is voiced, a transition cost is also introduced which takes into account the difference between the left context of the input frame and the left context of all target entries. To be consistent with unit-selection terminology, we call the former, unit distance ($d_{unit}$) and the latter, transition distance ($d_{tran}$). We have used equal weights for both distance measures. The target codebook entry that minimizes the overall cost function is transplanted onto the input. The transition cost encourages selection of consecutive frames from the training data while the unit cost ensures that the spectral identity of each frame is preserved.

### 4.3. Objective Results

Both spectral conversion methods reduce RMS error compared with the no conversion case (Table 2). However, GMM-based transformation outperforms the codebook-based transformation for all three emotions.

Table 2: *RMS Error between source/converted and target LSF vectors*

| RMSE | Angry | Sad | Surprised |
|---|---|---|---|
| Leave Spectra Unchanged | 0.055 | 0.046 | 0.052 |
| **GMM Conversion** | **0.041** | **0.037** | **0.042** |
| Codebook Conversion | 0.048 | 0.042 | 0.047 |

# 5. Prosody Conversion

## 5.1. Duration Conversion

To convert phone durations, regression trees were built for four broad classes: vowels, nasals, glides and fricatives. We have investigated the use of both absolute trees which directly predict a duration in milliseconds as well as relative trees which predict a scaling factor to be applied to the neutral duration. Relative trees strongly outperform absolute trees on the basis of RMS error regardless of the contextual features used to build the tree. Therefore in the following experiments only relative trees are used. We have investigated four feature groups:

- F0: original(neutral) duration (continuous)
- F1: F0 + phone identity, previous phone, next phone,
- F2: F1 + lexical stress, sentence position
- F3: F2 + word position, word length, part of speech

Once all the trees were built, cross-validation across the training data was used to find the optimal pruning level. Cross-validation was performed by holding out data in groups of ten. The final pruned trees were used to predict phone durations on the test data. Table 3 outlines the RMSE per feature group. While all trees did better than leaving durations unchanged, the best results were acquired by group F0, i.e. simply using original duration to predict the scaling factor. Introducing contextual factors seemed to cause confusion. This may be explained by the fact that the speaker used varying strategies related to specific contexts. For the purposes of evaluation, we chose to work with the feature group F0 which minimized RMS error.

Table 3: *RMSE for duration.*

| RMSE(ms) | Angry | Sad | Surprised |
|---|---|---|---|
| Leave Duration Unchanged | 33.51 | 30.19 | 32.27 |
| **Duration Tree (F0)** | **18.72** | **16.49** | **13.25** |
| Duration Tree (F1) | 22.12 | 19.86 | 17.84 |
| Duration Tree (F2) | 22.33 | 19.26 | 28.10 |
| Duration Tree (F3) | 21.92 | 19.30 | 25.94 |

## 5.2. Intonation Modelling and Generation

Three state context-sensitive left-to-right HMMs were used to model each syllable. The relevant contextual factors were chosen as a result of informal perceptual tests. In this experiment the factors found to be perceptually most significant were lexical stress, word position, sentence position, part of speech and part of speech of previous word. Models were trained with HTK using interpolated sentence F0 contours from the training data as well as first and second order differentials of F0. Decision tree-based parameter tying was performed to compensate for rare and unseen contexts. For synthesis, the goal is to generate the sequence of F0 values that maximizes the probability of observations given the models $\lambda$ and the state sequence $Q_{max}$

$$P(O|\lambda) = \max_{Q} P(O|Q,\lambda)P(Q|\lambda)\ . \qquad (1)$$

The cepstral parameter generation algorithm used in the HTS speech synthesis system[11] was used to generate smoothly varying sentence F0 contours by concatenating syllable HMMs. The generation algorithm takes advantage of the relationship between static parameters (F0) and dynamic parameters($\Delta F0$, $\Delta\Delta$F0) to optimize the above probability with respect to F0 when dynamic parameters are used as constraints. The predicted phone durations were used to determine the duration of each syllable F0 segment. Figure 4 illustrates three F0 contours generated by the different emotion models for the same utterance when durations are left unchanged.

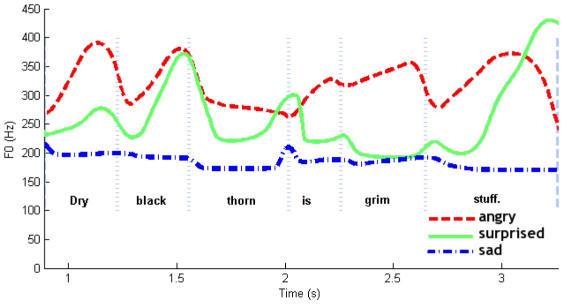

Figure 4: *F0 Contours generated by syllable HMMs*

# 6. Evaluation

A perceptual listening test was conducted to evaluate the combined conversions. In order to evaluate the contribution of individual modules, we have investigated the use of an independent emotion classifier. The classifier provides useful clues about how our modules interact while allowing us to keep the listening test focused on overall system output.

## 6.1. Emotion Classifier

The classifier used in the evaluation was originally developed for a speaker independent emotion classification task involving voicemail messages [12] and uses features acquired from a larger temporal stretch (the voiced segment) than any of our conversion modules. Each emotion is represented by a fully connected 5 state HMM. 21 features are extracted from each voiced segment and a sequence of feature vectors are formed for each utterance. The features used are:

- Prosodic Features: F0 (min, max, range, mean, tilt, number of optima), duration of voiced segment
- Spectral Energy: 25th percentile RMS, 75th percentile RMS, Mean spectral loudness 4-14 barks (10 features)

When evaluated on the original test utterances spoken by the speaker, the classifier performed quite strongly with slight confusion between surprise and anger (Table 4). Table 5 illustrates the classification results for individual and combined conversions. When only the spectra were converted with the GMM method, all but one utterance were classified as neutral. This is consistent with the findings of [2]. When only prosody was converted, sad utterances were identified perfectly, while anger and surprise reached a high recognition rate. The combination of spectral and prosody conversions resulted in the highest scores for all emotions. When applied with the prosody conversion module, spectral conversions made a notable difference in the recognition rate for anger.

Table 4: *Classifier Recognition Scores on Original Test Data*

| | Sad | Surp | Ang | Neu |
|---|---|---|---|---|
| Sad | 96.4 | 0 | 0 | 3.57 |
| Surp | 0 | 92.86 | 7.14 | 0 |
| Ang | 0 | 10.71 | 89.29 | 0 |

## 6.2. Perceptual Listening Test

Ten conversions were randomly chosen for each emotion (sad, angry, surprised) and presented to 15 listeners in a mixed order.

Table 5: *Classifier Recognition Scores on Individual and Combined Conversions*

| | **Spectral Conversion** | | | | **Prosody Conversion** | | | | **Spectral and Prosody Conversion** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sad | Surp | Ang | Neu | Sad | Surp | Ang | Neu | Sad | Surp | Ang | Neu |
| Sad | 0 | 0 | 0 | 100 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| Surp | 0 | 0 | 0 | 100 | 0 | 85.72 | 14.28 | 0 | 0 | 89.29 | 10.71 | 0 |
| Ang | 0 | 0 | 3.57 | 96.43 | 0 | 10.71 | 89.29 | 0 | 0 | 0 | 100 | 0 |

A possible realistic scenario for a TTS dialogue system is one where listeners initially hear output in a neutral voice, then at some point later in the interaction they are confronted with an expressive output. To mimic this scenario, listeners were first presented with the neutral version of each utterance followed by the expressive version. They were then asked to indicate the perceived emotion of the second utterance by selecting one of the three possible emotions. In addition, they were asked to rate the overall quality of the utterance as natural or unnatural. A sub-experiment was performed to compare the two spectral conversion techniques: for the case of anger, which is the emotion most affected by spectral changes, half of the utterances had their spectra converted with the GMM approach and half of them were converted with the codebook approach. For sad and surprised, all spectra were converted with the GMM approach.

To compare the overall recognition results (Table 6) with the perceptual test on the original emotional data (Table 1), the confusions with neutral and happy must be accounted for. In Table 1 sadness was only confused with neutral so it may be assumed that in the absence of a neutral choice, sadness would be identified correctly nearly all the time. Our conversions were also able to reach this high recognition rate (96%). In Table 1, surprise was mostly confused with happiness. Conflating these gives a recognition rate of about 88% which is comparable to the rate achieved for our conversions (84%). In evaluating anger, however, listeners seemed to do notably better on the original speech data (93.3%) than the conversions (80%). This is partially explained by the breakdown for spectral conversion methods for anger (Table 7). The GMM method performed better at conveying anger (86.67% ) than the codebook method (70.8%), which was clearly bringing the overall average down. This result was consistent with RMSE values reported in Section 4. The codebook method caused unpredictable discontinuities in voice quality, which may have resulted in the lower quality ratings (only 41.33% natural) and confused the listeners. In general, listener quality ratings and confusions were somewhat correlated: sadness was judged to be natural much more frequently than anger and surprise. It is unclear, however, whether the lower quality ratings for anger and surprise are due to the output of our conversion modules or the limitations of underlying speech modification frameworks.

Table 6: *Human Recognition and Quality Scores from the Listening Test*

| | **Emotion Classification** | | | **Quality Evaluation** | |
|---|---|---|---|---|---|
| | Sad | Surp | Ang | Natural | Not Natural |
| Sad | 96.00 | 0 | 4.00 | 80 | 20 |
| Surp | 4.00 | 84.00 | 12.00 | 63.33 | 36.67 |
| Ang | 5.33 | 14.67 | 80.00 | 49.33 | 50.67 |

Table 7: *Comparison of Spectral Conversion Methods for Anger*

| | **Emotion Classification** | | | **Quality Evaluation** | |
|---|---|---|---|---|---|
| | Sad | Surp | Ang | Natural | Not Natural |
| Codebook | 9.33 | 17.34 | 73.33 | 41.33 | 58.67 |
| GMM | 1.33 | 12 | 86.67 | 58.67 | 41.33 |

## 7. Conclusions

A system for emotion transformation has been described which consists of a cascade of modules for converting short-term spectra, durations and F0. The novel use of context-sensitive syllable HMMs for F0 contour generation resulted in convincing intonation when combined with phone durations output by the relative decision trees. Listening tests showed that the GMM-based spectral conversion method performed significantly better in conveying anger than the codebook selection approach. Overall perceptual recognition rates were encouraging being similar to those for human emotional speech in the case of sadness and being only 5-6% worse for surprise and anger. More substantial spectral and prosodic modifications (e.g. anger) resulted in more confusions and lower quality ratings. Future work will explore the underlying causes of this effect.

## 8. 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., ”Intonation Modelling and Adaptation for Emotional Prosody Generatation” Proc. of ACII, 286-293, 2005.

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

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

[10] Stylianou et al.,”Continuous Probabilistic Transform for Voice Conversion”, IEEE Trans. Speech and Audio Proc..vol.6(2):131-142, 1998.

[11] 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.

[12] Inanoglu, Z., Caneel, R., ”Emotive Alert:HMM-Based Emotion Detection in Voicemail Messages”, Proc. of IUI, 251-253, 2005.