

Intonation Modelling and Adaptation for Emotional Prosody Generation

Zeynep Inanoglu and Steve Young

Cambridge University Engineering Department
Machine Intelligence Laboratory
Cambridge, UK
zi201@cam.ac.uk, sjy@eng.cam.ac.uk

Abstract. This paper proposes an HMM-based approach to generating emotional intonation patterns. A set of models were built to represent syllable-length intonation units. In a classification framework, the models were able to detect a sequence of intonation units from raw fundamental frequency values. Using the models in a generative framework, we were able to synthesize smooth and natural sounding pitch contours. As a case study for emotional intonation generation, Maximum Likelihood Linear Regression (MLLR) adaptation was used to transform the neutral model parameters with a small amount of happy and sad speech data. Perceptual tests showed that listeners could identify the speech with the sad intonation 80% of the time. On the other hand, listeners formed a bimodal distribution in their ability to detect the system generated happy intonation and on average listeners were able to detect happy intonation only 46% of the time.

1 Introduction

Emotional speech synthesis requires the appropriate manipulation of a wide range of parameters related to voice quality and prosody. In this paper, we focus on the generation of pitch contours which constitute the building blocks of human intonation. Such prosodic generation schemes can plug into a variety of speech synthesis or voice conversion frameworks. We argue that a robust model-based approach can provide an adaptive framework which allows new emotional intonations to be generated with little training data using model adaptation algorithms.

The popularity of concatenative synthesis and unit-selection schemes have made the prospect of emotional speech synthesis more viable, given the improvement in quality over formant-based approaches. One approach that has been explored is to record neutral as well as emotional speech databases and select the units with the appropriate emotional qualities [1][2][3]. However this gives only limited capability for generating new emotions, since each new emotion requires an extensive data collection effort. At the other end of the spectrum, exploring dimensional approaches has proved valuable, particularly for generating non-extreme emotions and emotion build-up over time [4]. Due to lack of a

Table 1. List of Intonation Units

a	accent	c	unstressed
rb	rising boundary	fb	falling boundary
arb	accent + rising boundary	afb	accent + falling boundary
sil	silence		

comprehensive theory of emotion, however, it is unclear how all emotions can be represented within this dimensional space.

We propose an HMM-based representation of intonation, where each HMM represents a syllable-length intonational unit. Previous work on intonation modelling based on HMM has been carried out by [6][5][7]. We revisit this approach with the purpose of automatic contour synthesis and adaptation. Our motivation is twofold: to create a set of unified models that can be used to both recognize and synthesize natural sounding intonation patterns and to generate emotional intonation by adapting model parameters with a small amount of emotional speech data. Our results show that syllable-based HMMs can generate very natural sounding intonation contours. Adaptation to a very small amount of sad speech data also resulted in considerably sadder prosody, as confirmed by preference tests. Adaptation to happy data was not as effective, suggesting that intonation may not be the only key element in producing happy sounding speech.

Section 2 describes the models in detail. Section 3 presents the performance of our models in a recognition framework, where two levels of context-sensitivity are explored. Section 4 reviews the HTS(HMM-based speech synthesis) system which has been adapted to generate continuous pitch contours from HMMs and illustrates the contours that result from our models. Section 5 describes the MLLR adaptation method and demonstrates the results of adaptation to happy and sad speech data. Section 6 summarizes conclusions and future work.

2 Intonation Models

A set of seven basic units, which constitute the parts of an intonational phrase, were chosen (Table 1). Each syllable is assumed to belong to one of the seven units of intonation. This label set is based on the work done on the Tilt intonation model [7]. An accent is represented by **a** and any intonation movement that is not an accent is either a falling or rising boundary tone (**fb,rb**) or an unstressed segment of speech(**c**). For the cases when an accent and a boundary tone coincide, the combination units (**arb, afb**) are used. Silence is also included as a marker for intonational boundaries.

The basic models were trained on a female speaker from the Boston Radio Corpus [12]. This is a corpus of news stories read by a professional news reader. About 48 minutes of the female speaker’s speech was labelled with the intonation units. Out of the thirty-four news clips, five were set aside for testing. To achieve speaker-independence, raw fundamental frequency values were normalized by speaker mean and two orders of dynamic coefficients were computed

(delta and delta delta). The models have three states, each with a single Gaussian distribution. Training was carried out using HTK (Hidden Markov Model Tool Kit)[11]. A number of iterations of embedded Baum-Welch algorithm were performed. Context-sensitive models were also explored using incremental levels of context around each intonation unit. A **tri-unit** model set was built by replicating models based on their left and right neighbours and performing further iterations of embedded training. The use of tri-unit models is analogous to the use triphones in speech recognition. In order to overcome the problems of data sparsity and unseen contexts, decision-tree based parameter tying was used, where states of the models are clustered based on a log-likelihood criterion[10]. A **full-context** model set was also built incorporating sixteen contextual factors. The goal here was to capture more long distance intonation dependencies and take advantage of more detailed information such as the relative location of the vowel in the syllable. An intonational phrase (IP) is defined as the sequence of labels between two silences. Based on this definition, the contextual factors incorporated in the full-context models are as follows:

- identity of the unit to the left
- identity of the unit to the right
- relative position of current syllable in current IP
- the total number of each of **a/c/rb/fb/arb/afbs** in current IP
- the number each of **a/arb/afb/rb/fbs** before current syllable in current IP
- relative position of the vowel in current syllable
- total number of phones in current syllable

Decisions on how much context to incorporate in the models can depend on the framework within which one wishes to use the system. In a speech synthesis framework, for instance, access to phonetic information is straightforward and further context can be incorporated. In a voice conversion framework, it may be preferable to work with intonation units only.

3 Recognition Accuracy of Intonation Models

Evaluating the models in a recognition framework is important for two reasons: to understand how effective our models are as a means to capture and analyze intonation patterns and to ensure that they reliably represent the individual intonation units since failing to do so would also degrade the quality of synthesis. Recognition is performed using the Viterbi algorithm in HTK. Given consecutive raw f0 values extracted from an utterance of any length, the Viterbi algorithm finds the sequence of intonation units that maximizes the likelihood of the observations. In addition to the label sequence, the syllable boundaries are also determined by this process. The test data consisted of five news stories each lasting for about two minutes, with an average syllable count of 200. Table 2 shows the percent accuracy and percent correct rates for models of varying contextual complexity. The difference between percent correct and percent accuracy is that the latter also takes into consideration insertion errors (i.e. errors where a

redundant intonation unit was inserted between two correctly recognized units) and is therefore a more informative evaluation metric.

To illustrate the power of context we have also included the best results that can be achieved by increasing the number of Gaussian mixture components when context is ignored. Incorporating full context significantly improves the performance even when compared with the best performance achieved by the optimal configuration of mixtures components(N=10).

Table 2. Recognition results of varying mixture components in the no context case as well as single mixture component with tri-unit and full context. Ten mixtures was the optimal rate for the no-context case.

Model Set	Number of Mixtures	%Correct	%Accuracy
Basic	N=1	53.26	44.52
	N=2	53.36	45.48
	N=4	54.65	46.31
	N=10	59.58	50.40
Tri-Unit	N=1	61.50	49.75
Full Context	N=1	64.02	55.88

4 Intonation Synthesis from HMM

We have adapted the HMM-based speech synthesis(HTS)[9] system to generate continuous pitch values. HTS is a stand-alone speech synthesizer, which can generate speech waveforms from a set of context-sensitive phone HMMs. We have applied its cepstral parameter generation framework[8] to produce interpolated pitch values from our continuous density intonation HMMs. The key idea behind the parameter generation algorithm is the fact that in addition to f0 values, dynamic features (delta f0 and delta delta f0) are also used to optimize the likelihood of the generated parameter sequence. Without the incorporation of these dynamic features, the generated sequence would consist of the state mean vectors regardless of immediate context.

The input to our system is an intonation label sequence with corresponding syllable boundaries as well as a speaker mean value. The intonation models are then concatenated according to the label sequence to form a larger HMM network that represents the entire contour. The parameter generation algorithm is then applied to generate an interpolated f0 contour. Since training includes a mean normalization for the speaker, the input mean value is added to the generated f0 values. Currently the system also assumes that syllable boundaries are given since our focus is on intonation generation and a precise syllable-based duration model will require a separate investigation of its own. We have generated contours from both tri-unit and full-context model sets. Figure 1 illustrates the interpolated contours generated by our system given two different sequences of

intonation units for an utterance with six syllables. The introduction of accents and boundary tones in the contour is clearly observable in these plots. Figure 2 is a comparison of a sample contour generated by the tri-unit and full context models for the same label sequence. We were able to observe that, compared with the tri-unit model, the full context model may shift accents slightly to the right or to the left, based on the location of the vowel in the syllable. The full context model was also found to frequently vary the amplitude of a pitch accent based on the number of preceding accents. For instance, in an utterance with many accents, some of the later ones may not be as pronounced as the earlier ones. A number of these contextual modifications are observable in Figure 2.

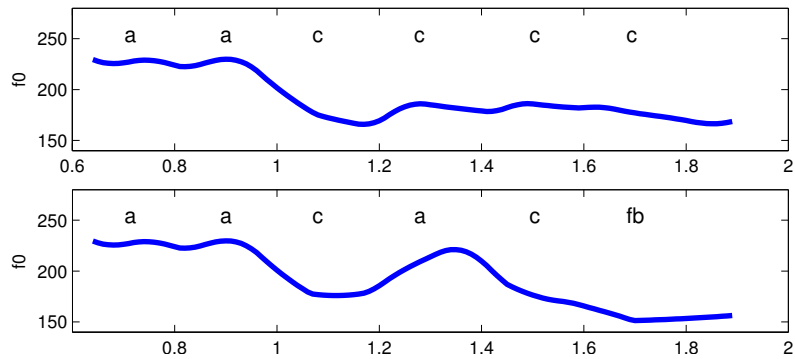


Fig. 1. Contours generated from two different label sequences using tri-unit model set with mean $f_0=200\text{Hz}$

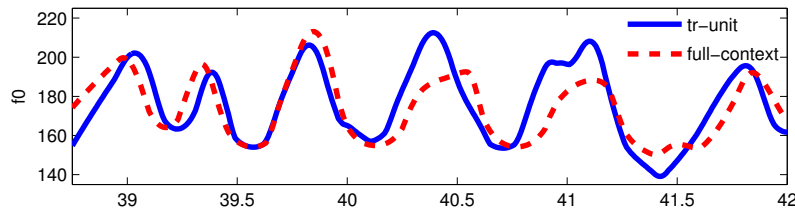


Fig. 2. Contours generated by tri-unit and full context models for the same label sequence, mean $f_0=180\text{Hz}$

The perceptual evaluation framework was based on transplanting the contours onto naturally spoken utterances. In order to focus on the effects of the generated contours, real utterances were modified using the PSOLA algorithm [14] to replace their existing f_0 contours with those generated by the intonation models. We were able to transplant our contours onto a wide range of recordings from different speakers and still obtain very natural sounding results. A brief perceptual test was conducted to quantify the naturalness of generated contours. Listeners were asked to rate eight utterances on a scale of 1 to 5 for their naturalness, 1 corresponding to least natural and 5 to very natural. Four of the utterances were presented with their original, unmodified pitch contours

and the other four had the synthetic contours generated by our tri-unit models. The utterances were presented in random order. The objective of the test was to ensure that the utterances with synthetic contours had ratings that overlapped sufficiently with those of the original recordings. The mean rating was 4 out of 5 for the unmodified utterances and 3.55 for the modified utterances. A t-test ($p < 0.05$) on the two samples confirmed that the samples are not statistically distinguishable.

5 MLLR Adaptation of Intonation Models with Happy and Sad Speech

MLLR is an adaptation technique frequently used in speech recognition to adapt phone models to different regional dialects[13]. The algorithm estimates a set of linear transformations for the mean and variance parameters of the Gaussian distributions for each state. Using HTK, MLLR can be applied flexibly depending on the amount of training data available. A regression tree is used to cluster the Gaussian parameters based on a similarity criteria and different transformations can be applied to different nodes of the tree based on the data available.

Thirty nine and ninety short segments of speech were acquired for sad and happy emotions, respectively. The data came from the Emotional Prosody Speech Corpus[15], which contains neutral and emotional utterances of four syllable phrases, mainly dates and numbers. Since these are extremely short utterances with limited long-distance context, tri-unit models were used for adaptation. Figure 3 illustrates sad, happy and neutral contours all generated by their respective model sets for a given label sequence. Detailed analysis of the emotion-specific changes in intonation is beyond the scope of this paper. However, it was observed that both happy and sad contours generally had a higher mean pitch, while the pitch range for happy contours was frequently wider than both sad and neutral contours, particularly in the case of accents and accents with rising boundaries. Sad contours manifested very unpronounced intonation units, particularly accents.

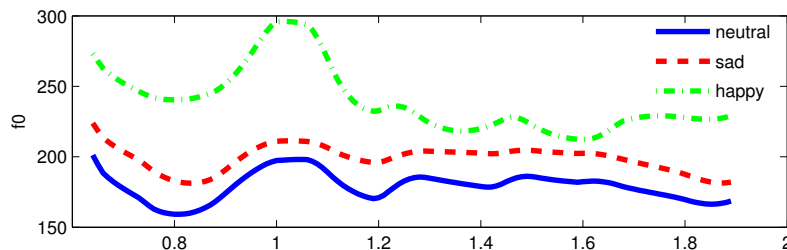


Fig. 3. Synthesized sad, happy and neutral contours by tri-unit models for the label sequence: **c arb c c c c.** mean $f_0=200\text{Hz}$

A two-part, perceptual preference test was conducted to compare the two types of generated contours: neutral and emotional. In part one, listeners were asked to listen to twenty pairs of short utterances and decide which utterance out of each pair sounds happier. In part two, the same procedure was repeated for sad speech. In order to evaluate the relative performance of emotional models over neutral ones, a forced choice was required between a generated neutral contour and a generated emotional contour. The utterances were taken from five different speakers: 3 female and 2 male. The presented contours covered all of the seven basic intonation units. Fourteen listeners took the test and a total of 280 choices were made in each part. The speech with sad contours were identified as sounding sadder 80% of the time. This was statistically significant (Chi-Square, $p < 0.01$). On the other hand, only 46% of listener’s choices associated the happy contours with happier sounding speech. While we would have hoped that our models performed better for the happy case, our results seem to confirm previous attempts in emotion synthesis[3][2].

When analyzed on a per listener basis, the distribution of preferences in part one (happy) suggest strong bimodality: eight of the fourteen listeners had a mean identification rate of 5 out of 20, while the remaining six had a mean identification rate of 15. Figure 4 illustrates the distribution of speakers in both parts. Clearly the intonational correlates of happiness are not universally sufficient on their own to express the full texture of the emotion. On the other hand, there may also be differences in imagined context for the two groups of listeners who seem to disagree on the effectiveness of our models.

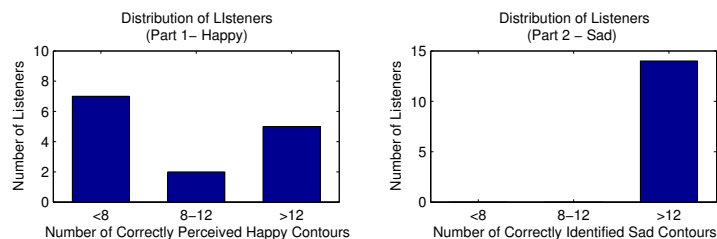


Fig. 4. Distribution of listener agreements with system.

An interesting observation was made on the possible dependence of system performance on the actual label sequence. In part one, the utterance where thirteen out of fourteen listeners agreed with the contours produced by happy models, consisted of six consecutive accents. The same label sequence was also the least successful in part two, meaning that listeners consistently preferred neutral contours over sad ones for sadder sounding speech. These facts suggest that the frequency and organization of certain constituents such as accents may be directly correlated with the proper expression of the two emotions at hand, contributing to the generative power of our models. More rigorous analysis of correlations between intonation unit sequences and emotions will be part of future work.

6 Conclusions and Future Work

A novel approach to synthesizing pitch contours based on syllable-based intonation models has been proposed. The effectiveness of these models was assessed both in a classification framework through recognition accuracy figures as well as in a generative framework through perceptual tests. Model adaptation to sad data proved successful while adaptation to happy data resulted in a division between groups of listeners. One of our immediate objectives is to analyze the dependence of emotions on the actual intonation unit sequence. Incorporation of an emotion-specific pattern of basic units may actually improve the perception of difficult emotions such as happiness. Since the poor performance of the happy synthesis may also be a consequence of the limited context in the specific adaptation data, richer data sources will be evaluated. Further work on emotion recognition using adapted intonation models will also be pursued in order to explore advantages of recognition/synthesis duality.

References

1. Akemi, I., Campbell N., Higuchi F., Yasamura M.: A Corpus-based Speech Synthesis System with Emotion. Proc. of Speech Communication (20 and 03) vol.40 161–187
2. Montero, JM., Arriola G.J, Colas, J., Enriquez, E. Pardo, J.M. : Analysis and Modelling of Emotional Speech in Spanish. Proc. of ICPhS (1999) vol.2 957–960
3. Bulut, M., Narayanan S., Syrdal A.: Expressive Speech Synthesis Using a Concatenative Synthesizer. Proc. of ICSLP (2002)
4. Schroder, M.: Dimensional Emotion Representation as a Basis for Speech Synthesis with non-extreme emotions. Workshop on Affective Dialogue Sys. (2004) 209–220
5. Jensen, U., Moore, R.K., Dalsgaard, P., Lindberg B.: Modelling Intonation Contours at the Phrase Level using Continuous Density Hidden Markov Models Computer Speech and Language 8, (1994) 247–260
6. Ljolje, A., Fallside, F.: Recognition of Isolated Prosodic Patterns using Hidden Markov Models Speech and Language (1987) vol.2 27–33
7. Taylor, P.: Anaysis and Synthesis of Intonation using the Tilt Model. Journal of the Acoustical Society of America, (2000) 107(3): 1697–1714
8. Tokuda, K., Yoshimura T., Masuko, T., Kobayashi T., Kitamura T.: Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis Proc. of ICASSP Processing (2000) vol.3 1315–1318
9. Tokuda, K., Zen H., Black A.: An HMM-Based Speech Synthesis System Applied To English. IEEE Speech Synthesis Workshop (2002)
10. Odell, J.J.: The Use of Context in Large Vocabulary Speech Recognition. PhD Dissertation, Cambridge University (1995)
11. <http://htk.eng.cam.ac.uk>
12. Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S.: The Boston University Radio Corpus. Technical Report ECS-95-001 (1995)
13. Gales, M., Woodland, P.: Mean and Variance Adaptation within the MLLR Framework. Computer Speech and Language (1996) vol.10
14. Moulines E., Charpentier, F.: Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones. Speech Communication (1990) vol.9 453–467.
15. <http://www ldc.upenn.edu/Catalog/LDC2002S28.html>