Two-Stage Prosody Prediction for Emotional Text-to-Speech Synthesis

Hao Tang, Xi Zhou, Matthias Odisio, Mark Hasegawa-Johnson, and Thomas S. Huang

University of Illinois at Urbana-Champaign, Urbana, IL, U.S.A.
{haotang2,xizhou2,modisio,hasegawa,huang}@ifp.uiuc.edu

Abstract

In this paper, we adopt a difference approach to prosody prediction for emotional text-to-speech synthesis, where the prosodic variations between emotional and neutral speech are decomposed into the global and local prosodic variations and predicted using a two-stage model. The global prosodic variations are modeled by the means and standard deviations of the prosodic parameters, while the local prosodic variations are modeled by the classification and regression tree (CART) and dynamic programming. The proposed two-stage prosody prediction model has been successfully implemented as a prosodic module in a Festival-MBROLA architecture based emotional text-to-speech synthesis system, which is able to synthesize highly intelligible, natural and expressive speech.

Index Terms: TTS, speech synthesis, prosody prediction, CART, dynamic programming

1. Introduction

The process of automatically converting arbitrary textual messages into intelligible and natural sounding speech is known as text-to-speech (TTS) synthesis, which has significant and widespread applications in a variety of areas including speech-enabled mobile devices, automotive connectivities, enterprise and consumer communications, voice-assisted e-commerce, voice aid for people with disabilities, call center automation, human-computer interaction, entertainment industries, etc. [1, 2, 3]. Currently, the most commercially viable TTS software can produce neutral speech that is sometimes indistinguishable from the human speech, especially in certain types of application scenarios (namely, limited domain application scenarios) such as synthesizing telephone numbers and weather forecasting. However, there has been a lack of emotional affect in the synthetic speech of the state-of-the-art TTS systems. This is largely due to the fact that the prosodic modules in these systems are unable to predict prosody from text accurately for emotional speech. Emotional speech exhibits prosodic variations from neutral speech in very different ways for different emotional states. Researchers have been trying to model such variations since the first attempt to add emotional affect to synthetic speech [4]. In the literature, the global or macro prosodic variations have been well studied thanks to the recognition that prosody is a suprasegmental feature rather than a segmental feature [5, 6, 7]. Recently, the local or micro prosodic variations have been brought to the attention of the research community [8, 9, 10]. The local prosodic variations account for the subtle changes in prosody between emotional and neutral speech which rely on the underlying linguistic structures and contexts. It is natural to think of the prosodic variations as a superposition of the global and local prosodic variations. Thus, the prediction of the prosodic variations can be decomposed into two stages, namely, the prediction of the global prosodic variations and the prediction of the local prosodic variations.

The prediction of prosody from text can be formulated as a transformation that maps a sequence of linguistic contexts to a corresponding sequence of prosodic parameters

$$c_1, c_2, ..., c_N \longrightarrow p_1, p_2, ..., p_N \tag{1}$$

where N denotes the number of basic units (e.g., phones) in the input text. Our goal is, given the emotional state E and linguistic contexts C, to construct a prosody prediction model, $P_e(E, C)$, for emotional speech, that implements the transformation 1. In the case of prosody prediction for neutral speech, this model reduces to $P_n(C)$.

2. A difference approach

This paper adopts a difference approach to prosody prediction for emotional TTS synthesis. More precisely, we aim to construct a difference model, $\Delta P_e(E, C)$, between emotional and neutral speech such that

$$P_e(E, C) = P_n(C) + \Delta P_e(E, C) \tag{2}$$

There are very good and convincing reasons that this difference model is advantageous over the direct modeling of $P_e(E, C)$, the full prosody prediction model for emotional speech. First, the prosody prediction model for neutral speech, $P_n(C)$, has been extensively studied and implemented as robust prosodic modules in current state-of-the-art TTS systems. It would be beneficial to build the prosody prediction model for emotional speech upon these existing systems. Second, it is straightforward to see that the difference model would require less data to train than would the full model. Since a large part of the dependency of prosody on the underlying linguistic contexts has been accounted for by $P_n(C)$, the dependency of $\Delta P_e(E, C)$ on the linguistic contexts is far simpler. To capture such dependency would therefore require a significantly less amount of training data. Third but not the last, the difference model allows us to render an emotion with a continuum of intensities through

$$P_e(E, C) = P_n(C) + \alpha \Delta P_e(E, C) \tag{3}$$

where α is a continuous coefficient controlling the intensity of the emotion. For example, $0 < \alpha < 1$ will make the emotion less than full-brown while $\alpha > 1$ will exaggerate the emotion to a certain extend.

The difference model is further decomposed by

$$\Delta P_e(E, C) = \Delta P_g(E) + \Delta P_l(E, C) \tag{4}$$

where $\Delta P_g(E)$ is the model of the global prosodic variations and $\Delta P_l(E, C)$ is the model of the local prosodic variations.

This decomposition also separates the influence of the emotional state E and linguistic contexts C on the prosodic variations, which turns out to be necessary as without such separation the global prosodic variations would become dominant and prevent the subtleties of the local prosodic variations from being precisely learned based on the underlying linguistic contexts. In this paper, a data-driven two-stage model is utilized to predict the prosodic variations, which, combined with the prosody predicted for neutral speech, yield accurate prosody predictions for emotional speech.

3. Databases

We have specifically constructed several emotional speech databases for prosody modeling and prediction. We designed a corpus consisting of a selection of 300 sentences from the IEEE Harvard Sentences [11]. The selection was performed using a greedy algorithm so that the final corpus is phonetically balanced and prosodically rich. A professional actress whose first language is American English was hired to speak the script in the neutral, happy, sad, and angry manners, respectively. The speech was recorded at 44.1 kHz using a MOTU 8pre firewire audio interface and a Studio Projects B1 condenser microphone in a soundproof booth, and was downsampled to 16 kHz prior to further processing. The average length of the recorded sentences in the resulting databases is about 3-4 seconds, depending on the emotional state. Thus, each of the databases contains speech data about 15 to 20 minutes long.

The speech signals in the databases were first aligned with the script using the HTK toolkit [12]. The forced alignment was performed at the phrase, word, syllable, and phone levels, respectively. Then, the f_0 and intensity contours of the speech signals were extracted using the Praat software [13] with a 25ms hamming window and a 10ms frame rate. The f_0 contours were first median-filtered and then linearly interpolated. Finally, the values of the prosodic parameters (f_0, duration, and intensity) were stored on a per phone basis. Note that the values of f_0 and intensity for a phone are taken to be the values of f_0 and intensity computed for the center frame over the duration of the phone, respectively.

It should be noted that these databases can be readily extended to include multiple actors/actresses and to include more emotional states. In this paper, however, we confine our model to be person-dependent and to the three basic emotional states, namely happiness, sadness, and anger.

4. Model of global prosodic variations

The model of the global prosodic variations, $\Delta P_g(E)$, determines the part of the prosodic variations between emotional and neutral speech that depends only on the emotional state, regardless of the underlying linguistic contexts. Such variations are described by the mean μ_e and standard deviation σ_e of the prosodic parameters for emotional speech as well as the mean μ_n and standard deviation σ_n of the prosodic parameters for neutral speech. The local prosodic variations can thus be isolated from the global prosodic variations by the following transformations

$$p'_e = \frac{p_e - \mu_e}{\sigma_e} \quad (5)$$

$$p'_n = \frac{p_n - \mu_n}{\sigma_n} \quad (6)$$

$$\delta p = p'_e - p'_n \quad (7)$$

where p_e, p'_e denote the total and local prosody for emotional speech, respectively, p_n, p'_n denote the total and local prosody for neutral speech, respectively, and δp denotes the local prosodic variations which capture the part of subtle prosodic variations that depends on the underlying linguistic contexts.

The mean μ and standard deviation σ of the prosodic parameters of the speech for the neutral and each emotional state can be estimated directly from the corresponding speech database. Table 1 gives the numerical values of μ and σ for the neutral as well as three emotional states, namely happiness, sadness, and anger, as a result of database analysis that we performed.

Table 1: *Means and standard deviations of prosodic parameters for neutral and three emotional states, namely happiness, sadness, and anger. The units of f_0, duration (dur.), and intensity (int.) are Hz, ms, and dB, respectively.*

emotion	mean (μ)			standard deviation (σ)		
	f_0	dur.	int.	f_0	dur.	int.
neutral	196.4	62.6	53.4	49.4	50.8	8.1
happy	265.3	63.7	59.4	78.2	55.9	9.6
sad	192.8	90.8	50.7	58.0	82.0	8.3
angry	238.2	64.4	59.5	72.0	56.2	10.2

5. Model of local prosodic variations

5.1. CART model

The model of the local prosodic variations, $\Delta P_l(E, C)$, is trained on δp for each emotional state E using the classification and regression tree (CART) algorithm [14]. A CART model is constructed based on the linguistic contexts of the individual phones in an input sentence as produced by a text analysis frontend of a TTS system. The linguistic context of a current phone is represented by a set of contextual factors of the current phone, augmented by the same contextual factors of the two phones to the left and two phones to the right of the current phone. These contextual factors include

- Identity of the current phone
- Voiced/unvoiced flag of the current phone
- Sonority class of the current phone (e.g., vowel, semivowel, nasal, fricative, stop, etc.)
- Position of the current phone in the current syllable
- Lexical stress of the current syllable
- Number of syllables in the current word
- Position of the current syllable in the current word
- Phrase-level stress of the current word
- Part of speech of the current word
- Number of words in the current phrase
- Position of the current word in the current phrase

The CART algorithm is a non-parametric statistical learning technique that builds a decision tree classifier on the data represented by attribute-value pairs. Each interior node of the decision tree splits the training examples associated with it into two separate sets based on a question test against a certain attribute in a way that each dichotomy of the training examples minimizes the Gini impurity of the overall tree [14]. Each leaf node of the tree stores a set of training examples which are assumed to belong to a single class. It is not uncommon that many examples may fall in the same leaf node. We further cluster the examples in each leaf node into K groups using the K-means algorithm [15].

Once the CART model has been trained, it can be used to map the linguistic context of a phone, c_i, to K candidates of the predicted local prosodic variations

$$c_i \longrightarrow \delta p_i^{(1)}, \delta p_i^{(2)}, ..., \delta p_i^{(K)} \tag{8}$$

where K is the number of clusters in a leaf node, $\delta p_i^{(k)}$ is the center of the k^{th} cluster, and i denotes the index of the phone.

5.2. Dynamic programming

Given a sequence of linguistic contexts $c_1, c_2, ..., c_N$, the ultimate goal of the model of the local prosodic variations, $\Delta P_l(E, C)$, is to predict a corresponding sequence of local prosodic variations $\delta p_1, \delta p_2, ..., \delta p_N$. Since the CART model described in the previous subsection outputs for each c_i a set of K candidates for δp_i instead of a single best δp_i, we may form a trellis in the following way. The $K \times N$ trellis consists of N columns with the i^{th} column corresponding to c_i and composed of the K candidates $\delta p_i^{(k)}$, $k = 1, 2, ..., K$. Each of these K candidates is associated with a target cost, S_t, which measures the unlikeliness of the candidate to be selected.

$$S_t(\delta p_i^{(k)}) = 1/n_{ik} \tag{9}$$

where n_{ik} is the number of examples in the cluster k in the leaf node that is reached by the traversal by c_i.

We define a concatenation cost, S_c, between two candidates in adjacent columns of the trellis by

$$S_c(\delta p_{i-1}^{(k)}, \delta p_i^{(j)}) = \parallel \delta p_i^{(k)} - \delta p_{i-1}^{(j)} \parallel_2 \tag{10}$$

where $\parallel \cdot \parallel_2$ denotes the L_2 norm. The concatenation cost is a measure of the jerkiness of the concatenation of the two candidates $\delta p_{i-1}^{(k)}$ and $\delta p_i^{(j)}$.

The dynamic programming technique [16] is used to find a path, $\delta p_1, \delta p_2, ..., \delta p_N$, in the trellis which is the best in the sense that the total cost, S_{tot}, of the path is minimum among all possible paths. The total cost is given by

$$S_{tot}(\delta p_1^{(k_1)}, \delta p_2^{(k_2)}, ..., \delta p_N^{(k_N)}) = \\ \sum_{i=1}^{N} \alpha_t S_t(\delta p_i^{(k_i)}) + \sum_{i=2}^{N} \alpha_c S_c(\delta p_{i-1}^{(k_{i-1})}, \delta p_i^{(k_i)}) \tag{11}$$

where α_t and α_c are the weights of the target cost and concatenation cost, respectively. These two numbers are often empirically set to balance the respective two costs.

Once the local prosodic variations, $\delta p_1, \delta p_2, ..., \delta p_N$, are determined, they are combined with the global prosodic variations and the prosody for neutral speech predicted by the model $P_n(C)$. The following transformations will yield the predicted prosody for emotional speech.

$$\widehat{p'_n} = \frac{\widehat{p_n} - \mu_n}{\sigma_n} \tag{12}$$

$$\widehat{p'_e} = \widehat{p'_n} + \widehat{\delta p} \tag{13}$$

$$\widehat{p_e} = \widehat{p'_e} \times \sigma_e + \mu_e \tag{14}$$

where the hat $\widehat{(\cdot)}$ denotes a predicted value.

6. System implementation

We have successfully implemented the proposed two-stage prosody prediction model as a prosodic module in an emotional TTS synthesis system based on a Festival-MBROLA architecture. Given an input sentence, the built-in prosody prediction model of Festival [17] is first employed to predict the prosody for neutral speech. Then, the text analysis frontend of Festival is used to generate the linguistic context of each phone in the sentence, which is fed into our two-stage difference model to predict the prosodic variations. Finally, the predicted prosody for neutral speech and prosodic variations are combined to yield the prosody for emotional speech, which is then sent to MBROLA [18], a diphone-based speech synthesizer, to produce the desired emotional synthetic speech. This emotional TTS system has demonstrated to be able to synthesize highly intelligible, natural, and expressive speech. A functional diagram of the system is shown in Figure 1.

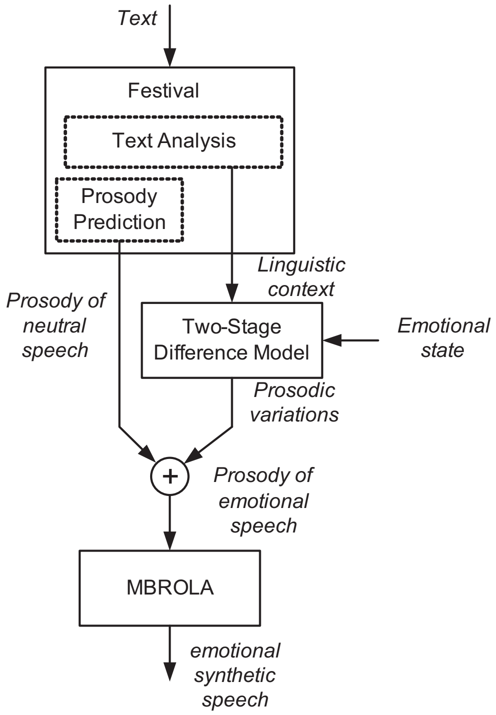
Text
Festival
Text Analysis
Prosody Prediction
Linguistic context
Prosody of neutral speech
Two-Stage Difference Model
Emotional state
Prosodic variations
+
Prosody of emotional speech
MBROLA
emotional synthetic speech

Figure 1: *Functional diagram of an emotional TTS system based on a Festival-MBROLA architecture. The proposed two-stage difference model for prosody prediction has been successfully implemented as a prosodic module in the system.*

Figure 2 illustrates the sample speech synthesis results of the above system. The input sentence to the system was "the boy was there when the sun rose." The neutral as well as happy, sad, and angry speech was synthesized. The waveforms of the synthetic speech were plotted along with the respective f_0 and intensity contours. Our initial informal listening experiments indicate that the synthesized speech, for the neutral as well as the three emotional states (happiness, sadness, and anger), is highly intelligible, natural, and expressive. These informal listening experiments represent a sample of our ongoing work on the more formal subjective and objective evaluations of the model and system, which are needed in assessing the effectiveness and efficiency of the algorithm and implementation.

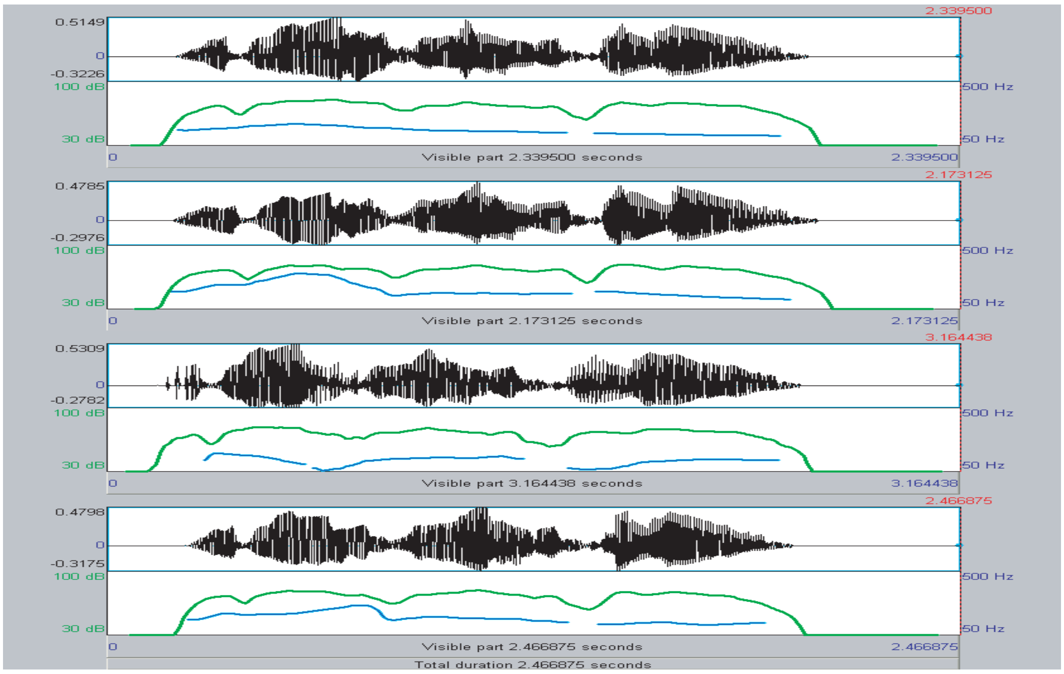

Figure 2: *Sample speech synthesis results. 1st row: neutral; 2nd row: happiness; 3rd row: sadness; 4th row: anger. The blue curves are the f_0 contours and the green curves are the intensity contours. All the plots were made by Praat [13].*

7. Conclusions

In this paper, we adopt a difference approach to prosody prediction for emotional TTS synthesis, where the prosodic variations between emotional and neutral speech are decomposed into the global and local prosodic variations and predicted using a two-stage model. The global prosodic variations are modeled by the means and standard deviations of the prosodic parameters, while the local prosodic variations are modeled by the CART model and dynamic programming. This two-stage prosody prediction model has been successfully implemented as a prosodic module in an emotional TTS synthesis system based on a Festival-MBROLA architecture. Our initial informal listening experiments show that the synthesized speech, for the neutral as well as happy, sad, and angry states, is highly intelligible, natural, and expressive. We are in the process of formalizing the subjective and objective evaluations of the model and system and making continuous improvements to the model and system.

8. Acknowledgements

This research was funded in part by the U.S. Government VACE Program and in part by the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies. The authors wish to thank Eleni Pappageorge who gave permission for her voice to be used.

9. References

[1] Thierry Dutoit, An introduction to text-to-speech synthesis, Kluwer Academic Publishers, Norwell, MA, 1997

[2] Richard W. Sproat, Multilingual Text-to-Speech Synthesis, Kluwer Academic Publishers, Norwell, MA, 1997

[3] Shrikanth Narayanan, Abeer Alwan, Text to speech synthesis: new paradigms and advances, Prentice Hall, 2005.

[4] M. Schroder, "Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (Ph.D thesis)," Saarland University, 2004.

[5] J.E. Cahn, "Generating expression in synthesized speech (Master thesis)," MIT Media Lab, MIT, 1989.

[6] I.R. Murray, Simulating emotion in synthetic speech (Ph.D thesis), University of Dundee, UK, 1989.

[7] F. Burkhardt, "Emofilt: the Simulation of Emotional Speech by Prosody-Transformation," In Proc. of INTERSPEECH-2005, pp. 509-512, Lisbon, Portugal, 2005.

[8] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J., "A Corpus-Based Approach to Expressive Speech Synthesis," in Proc. of the 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA, June 14-16, 2004.

[9] Fabio Tesser, Piero Cosi, Carlo Drioli, Graziano Tisato, "Emotional FESTIVAL-MBROLA TTS synthesis," In Proc. of INTERSPEECH-2005, pp. 505-508, 2005.

[10] J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M.A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," IEEE Transactions on Audio, Speech and Language Processing, vol.14, no. 4, pp. 1099-1108, 2006.

[11] IEEE Harvard Sentences. From the appendix of IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. IEEE Transactions on Audio and Electroacoustics. vol 17, 227-46, 1969.

[12] The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/.

[13] Boersma, Paul & Weenink, David (2008). Praat: doing phonetics by computer (Version 5.0.18) [Computer program]. Retrieved March 31, 2008, from http://www.praat.org/.

[14] L. Breiman, J. Friedman, R. A. Olshen and C. J. Stone, "Classification and regression trees," Wadsworth, 1984.

[15] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification (2nd ed.), John Wiley and Sons, 2001.

[16] Richard Ernest Bellman, Dynamic Programming, Dover Publications, Incorporated, 2003.

[17] The Festival Project, http://www.cstr.ed.ac.uk/projects/festival/.

[18] The MBROLA Project, http://mambo.ucsc.edu/psl/mbrola/.