PROSODY ANALYSIS AND MODELING FOR EMOTIONAL SPEECH SYNTHESIS*

Dan-ning Jiang[1], Wei Zhang[2], Li-qin Shen[2], Lian-hong Cai[1]

[1]Department of Computer Science & Technology, Tsinghua University, China
[2]IBM China Research Lab
jdn00@mails.tsinghua.edu.cn, {zhangzw,shenlq}@cn.ibm.com, clh-dcs@tsinghua.edu.cn

ABSTRACT

Current concatenative Text-to-Speech systems can synthesize varied emotions, but the subtle and range of the results are limited because large amount of emotional speech data are required. This paper studies a more flexible approach based on analyzing and modeling the emotional prosody features. Perceptual tests are first performed to investigate whether just manipulating prosody features can attain the communication purposes of emotions. Then, based on the positive results, the same corpus with sufficient prosody coverage is shared by different emotions in unit selection. Finally, an adaptation algorithm is proposed to predict the emotional prosody features. It models the prosodic variations by linguistic cues and emotion cues separately, and requires only a small amount of data. Experiments on Mandarin show that the adaptation algorithm can obtain appropriate emotional prosody features, and at least several emotions can be synthesized without the use of special emotional corpus.

1. INTRODUCTION

Current concatenative Text-to-Speech (TTS) systems can synthesize high-quality speech [1]. Therefore, how to incorporate emotions in the neutral systems becomes recent research focus. Previous works [2][3][4] have proved the synthesizers' capability to produce varied emotions. In these systems, a target emotion is synthesized by selecting appropriate units from its special large corpus. Emotional prosody features are controlled by Rule-based methods [5] or data-driven methods. Since it is not easy to describe the prosody characteristics comprehensively by rules, data-driven methods are preferred in these years, but they also require large amount of training data. Unfortunately, it is quite difficult to record so much emotional speech data and guarantee that the same emotion is recorded consistently. The approach is very costly and unable to synthesize more subtle and mixed emotions flexibly.

Considering that it is relatively easy to control prosody features in TTS systems, the paper studies a more flexible synthesis approach based on analyzing and modeling the emotional prosody features. Although relevant analysis suggested that just manipulating prosody features is not sufficient to present intense emotions [6][7], it still may be feasible to attain the communication purposes of emotions. Perceptual tests are first performed to validate this notion. Then, based on the positive results, the same corpus with sufficient prosody coverage is shared by different emotions in unit selection. Finally, an adaptation algorithm is proposed to predict the emotional prosody features. It models the prosodic variations by linguistic cues and emotion cues separately, and requires only a small amount of data. Experiments on Mandarin show that the adaptation algorithm can obtain appropriate emotional prosody features, and several emotions (at least sadness and happiness) can be synthesized without the use of special emotional corpus.

The paper is organized as follows. Section 2 presents perceptual tests to investigate the prosody's contribution in emotional communications, and discusses the results. Section 3 illustrates the prosody adaptation algorithm. In section 4, experiments on Mandarin are performed and evaluation results are shown. Conclusions are finally given in section 5.

2. PROSODY'S CONTRIBUTION IN EMOTIONAL COMMUNICATIONS

We presume that what more important to emotional synthesis is to attain the communication purposes of emotions, rather than obtain intense emotions. As suggested in [8], there are two communication functions in emotions. One is appropriateness, and the other is efficiency. The former means that the emotional expression should be appropriate for the verbal content and condition, for example, bad news should be presented unhappily. The latter means that the expression should deliver information about the speaker's attitude when the content does not show it. Thus, perceptual tests are performed to investigate whether prosody features alone can attain the appropriateness and efficiency in communications.

To isolate the influence of prosody features, the copy synthesis technology is used. Pairs of utterances with the same content and different expressions (neutral and emotional) are recorded. Then pitch, duration, and energy features of the emotional utterance are copied to the neutral counterpart through FD-PSOLA algorithm. So there are neutral segments and emotional prosody in the synthesis samples. All these three types of expressions are used as stimulus in perceptual tests. To test appropriateness of the above expressions, the text materials are emotional, and the subjects are asked whether the expression is appropriate for the content. In the efficiency test, the contents are

* This work was done in IBM China Research Lab.
It was also supported by China National Natural Science Foundation (60433030).

neutral, and the subjects are asked to guess the most possible condition according to the expression, for example, whether it delivers good news or bad news, or there is no emotional bias. Examples of the tests are shown in figure 1. Here, the examined emotion is sadness.

(a) Appropriateness Test Example
Text: 如今我的生活变得十分艰难。(Now my life is made extremely difficult.)
Question: Is the expression appropriate for the content?
Selections: A.appropriate; B.some inappropriate; C.very inappropriate.

(b) Efficiency Test Example
Text: 我已经知道自己的考试结果了。(I have known my exam result.)
Question: Which is the most possible exam result?
Selections: A.bad; B.I don't know; C.good.

Figure 1. Two examples of the perceptual tests. One is for the appropriateness test, and the other is for the efficiency test.

The text materials are 8 sentences, 4 with negative contents for the appropriateness test, and 4 with neutral contents for the efficiency test. The subjects are 24 university students unfamiliar with TTS. All utterances are randomly ordered.

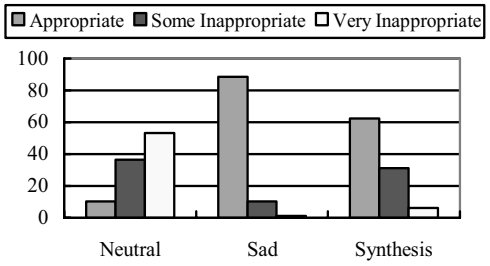

Figure 2. Results of the appropriateness test.

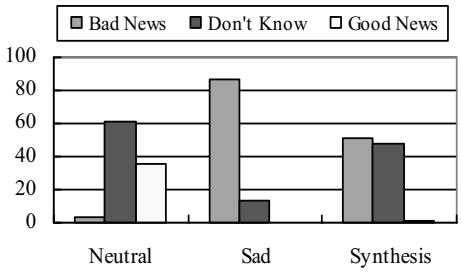

Figure 3. Results of the efficiency test.

Results of the two tests are shown in figure 2 and figure 3 respectively. From figure 2, it can be seen that neutral expressions are perceived as appropriate for the negative contents only at a rate about 10%, and as very inappropriate at most time. On the contrary, the sad and synthesis expressions are perceived as appropriate at most time, and very occasionally to be judged as very inappropriate. Figure 3 shows that neutral expressions nearly deliver no negative information, and sometimes even give subjects positive information. Sad expressions are associated with the bad news at a high rate over 80%. Although synthesis expressions deliver negative information only slightly over a half of time, they do not mislead the subjects with wrong information. The results indicate that although prosody features alone seem not enough to present intense emotions, they are still very meaningful in emotional communications.

Other works [4][6][7] suggested that some emotions (e.g. anger) are more heavily influenced by segmental features than others, but they also agreed that prosody features are more important for a number of emotions. So it is reasonable to generalize the analysis results to several other emotions.

3. PROSODY ADAPTATION

Emotional prosody features encode information from at least two sources. One source is linguistics, and the other is emotion. Variations caused by the former are determined by linguistic contexts, such as position in the prosodic structure, syllable tone type (as Chinese is a well-known tonal language), accent level, and so on. A large amount of training data is necessary to obtain a stable prediction of the linguistic component, typically several thousands of sentences [1]. This drives up the required data amount to train an emotional prosody model. However, the linguistic component is unrelated to emotions, so it does not have to be contained repeatedly in each emotional prosody model.

The paper proposes an adaptation algorithm. In the algorithm, the linguistic and emotion component of prosody features are modeled separately. The former is modeled only once by using a huge amount of neutral speech data. The latter is estimated as the differences between the emotional and neutral prosody features, and modeled with only a small amount of data for each target emotion. Then these two parts are combined together to predict the emotional prosody features.

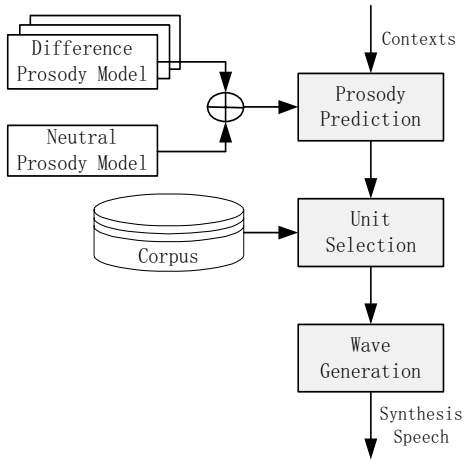

Figure 4. Paradigm of the emotional synthesis system using prosody adaptation.

Figure 4 shows paradigm of the emotional synthesis system using prosody adaptation. Here, the "difference prosody model" and "neutral prosody model" represent models of the emotional

component and the linguistic component respectively. A difference prosody model concerned with the target emotion is first selected from several ones. Then it is combined with the neutral prosody model to predict emotional prosody features. In unit selection, based on the analysis in section 2, different target emotions share the same corpus with sufficient prosody coverage. Note that the natural relationship between prosody and segmental features is maintained in the corpus, which is important to the naturalness of the synthesis results [7].

3.1. Probability Based Prosody Model

In prosody adaptation, the difference and neutral prosody models are all built through the probability based prosody model [1]. It calculates the target and transition costs for unit selection in a probabilistic way. In the training stage, linguistic contexts are first clustered by decision trees based on the distribution of the concerned prosody features; those with similar prosody features are clustered into the same terminal leaf. Then for each leaf, GMM is trained to model the probabilistic distribution of the prosody features. In unit selection, the target and transition decision trees are first traversed according to the input contexts, and the associated GMMs are obtained. Then the target and transition costs are both given by the corresponding GMM probabilities. Finally, a dynamic program algorithm is performed to find out a candidate sequence with the maximal combination of the target and transition probabilities as the prediction result.

It can be seen that when there are sufficient candidates in corpus, the predicted prosody features are determined by the input contexts. Thus the result is quite stable.

3.2. Adaptation Algorithm

Suppose PF is the prosody feature vector of a synthesis unit, then it contains both the linguistic component PF^l and the emotion component PF^e,

$$PF = PF^l + PF^e \tag{1}$$

The adaptation algorithm is divided into three steps: predict PF^l using a huge amount of neutral speech data, model PF^e for each target emotion, and finally combine them to predict PF.

(1) Predict PF^l using a huge amount of neutral speech data.

The linguistic component PF^l is predicted by using the model described in section 3.1. The neutral speech corpus contains 5,000 sentences, which can guarantee the stability of prediction results at most time. So in the following two steps, PF^l is regarded as a constant rather than a distribution.

(2) Model PF^e for each target emotion.

For each target emotion, PF^e is estimated as the difference between the emotional prosody feature PF and the predicted linguistic component PF^l. The modeling method is nearly the same with that used to model PF^l, except that the prosody feature is PF^e. So, given the input contexts CT, the corresponding GMM can be obtained, as formula (2) shows,

$$P(PF^e \mid CT) = \sum_{k=1}^{M} w_k N(\mu_k, \sigma_k) \tag{2}$$

Where M is the number of components in the GMM, w_k, μ_k and σ_k are the weight, mean and variation of the k-th component respectively.

(3) Combine the two components to predict PF.

Having the prediction of PF^l and the probabilistic description of PF^e, PF is modeled as,

$$P(PF \mid CT) = \sum_{k=1}^{M} w_k N(\mu_k + PF^l, \sigma_k) \tag{3}$$

4. EXPERIMENTS

4.1. Speech Materials

In experiments, two emotions in Mandarin are implemented. One is sadness, and the other is happiness. All emotional speech data are recorded in the lab, by asking a female speaker to read emotional paragraphs with a proper degree of emotion. Here, one example of the "proper degree" is the way that a newscaster reports someone's death with a pity. The segmental features are not much different with those of neutral speech.

The sad paragraphs are reports of illness and death, while the happy paragraphs are descriptions of the spring's beautiful views and a girl's imagination about her wedding. Note that the "happiness" is not a pure one; it mixes the love and enthusiasm about life. We totally use 216 sad sentences and 143 happy sentences, from which 10 sentences of each emotion are testing data, and the remaining sentences are used to train the emotional prosody models. These emotional data, together with other 500 neutral sentences, are put into the corpus for unit selection.

4.2. Emotional Prosody Characteristics

The neutral, emotional, and synthesis prosody characteristics are compared in figure 5 and figure 6. From figure 5, it can be seen that sadness is mainly characterized by low pitch mean, narrow pitch range, and very flat pitch contour. The limited pitch variations merely realize the syllable tone shape, and there is little pitch declination in the intonation phrase. The synthesis sad pitch contour does have above characteristics. Figure 6 shows that the happy prosody has a faster speak rate and less pitch declination. The pitch range is also slightly narrower. The synthesis prosody also shows similar characteristics.

4.3. ABX Tests

To evaluate whether the synthesis speech conveys the target emotion, we perform ABX tests, which are often used in studies on voice conversion. A, B and X represent the neutral speech, emotional speech, and synthesis speech respectively. In the test, three utterances are played with the order of A, B, X or B, A, X. Subjects are asked to judge whether the emotion implied in X is similar with that in A or B. There are 20 university students as the subjects.

Table 1. Results of the ABX tests.

	Sadness	Happiness
Ave. Correct Rate	100.0%	82.5%

Table 1 shows the average rate that the subjects judge the synthesis expression similar with the target emotions. The correct rate of sadness is very high, which may be because that the sad pitch contour is very flat and easy to predict. The correct rate of happiness is some lower, but still much higher than the random rate.

5. CONCLUSIONS

This paper studies a flexible emotional synthesis approach using the concatenative synthesizer. Perceptual tests are first performed to investigate the prosody's contribution in emotional communications. The analysis results show that just manipulating prosody features can obtain meaningful results at least for some emotions. Then, the same corpus with sufficient prosody coverage is shared by different emotions in unit selection. Finally, an adaptation algorithm is proposed to predict the emotional prosody features. In experiments, two emotions (sadness and happiness) in Mandarin are implemented. By using no more than 200 sentences as the training data, appropriate emotional prosody features can be predicted for each emotion, and the synthesis speech is judged as similar with the target emotions at a rate over 80%. The experiment results show that the adaptation algorithm is efficient when there are only small amount of training data. So, prosody features of more subtle and mixed emotions can be obtained with fewer difficulties. The results also indicate that at least several emotions can be synthesized without the use of special emotional corpus. However, since it is difficult to conclude that prosody features are important in all emotions, in further work, it may be necessary to extend the corpus by adding some emotional segments.

6. ACKNOWLEDGEMENTS

This work was done in IBM China Research Lab. I thank the work foundation and environment provided by the company, as well many valuable suggestions and help from Xijun Ma, Qin Shi, Weibin Zhu, and Yi Liu.

7. REFERENCES

[1] X.J. Ma, W. Zhang, W.B. Zhu, etc, "Probability based Prosody Model for Unit Selection," *Proc. of. ICASSP'04*, Montreal, Canada, pp. 649-652, May 2004.

[2] E.Eide, "Preservation, Identification, and Use of Emotion in a Text-to-Speech System," *Proc. of IEEE Workshop on Speech Synthesis*, Santa Monica, Sep. 2002.

[3] A.W. Black, "Unit Selection and Emotional Speech," *Proc. of EuroSpeech'03*, pp. 1649-1652, Sep. 2003.

[4] M. Bulut, S. S. Narayanon, and A. K. Syrdal, "Expressive Speech Synthesis Using a Concatenative Synthesizer," *Proc. of ICSLP'02*, Denver, pp. 1265-1268, Sep. 2002.

[5] I.R. Murray, M.D. Edgington, D. Campion, etc. "Rule-Based Emotion Synthesis Using Concatenated Speech," *Proc. of ISCA Workshop on Speech and Emotion*, Belfast, North Ireland, pp. 173-177, 2000.

[6] K. Hirose, N. Minematsu, and H. Kawanami, "Analytical and Perceptual Study on the Role of Acoustic Features in Realizing Emotional Speech," *Proc. of ICSLP'00*, pp. 369-372, Oct. 2000.

[7] C. Gobl, E. Bennet, and A.N. Chassaide, "Expressive Synthesis: How Crucial is Voice Quality?" *Proc. of IEEE Workshop on Speech Synthesis*, Santa Monica, Sep. 2002.

[8] E. Eide, R. Bakis, W. Hamza, and J. Pitrelli, "Multilayered Extensions to the Speech Synthesis Markup Language for Describing Expressiveness," *Proc. of EuroSpeech'03*, pp. 1645-1648, Sep. 2003.

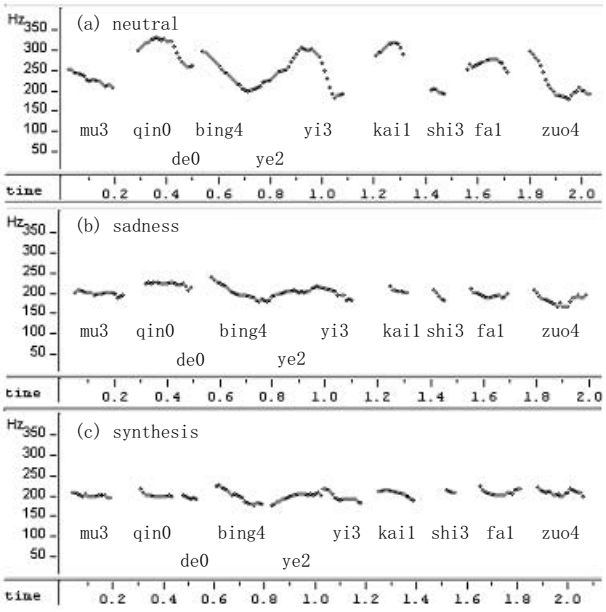

Figure 5. Neutral, sad, and synthesis pitch contour of a sad sentence, which means "Mother's illness has broken out."

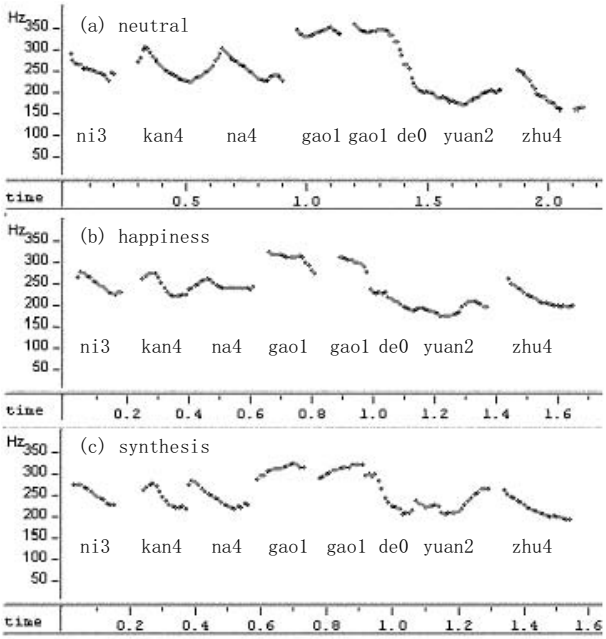

Figure 6. Neutral, happy, and synthesis pitch contour of a happy sentence, which means "You see that tall columns.