HMM-Based Emotional Speech Synthesis Using Average Emotion Model

Long Qin, Zhen-Hua Ling, Yi-Jian Wu, Bu-Fan Zhang, and Ren-Hua Wang

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei
{qinlong, zhling, jasonwu, bfzhang}@mail.ustc.edu.cn
rhw@ustc.edu.cn

Abstract. This paper presents a technique for synthesizing emotional speech based on an emotion-independent model which is called "average emotion" model. The average emotion model is trained using a multi-emotion speech database. Applying a MLLR-based model adaptation method, we can transform the average emotion model to present the target emotion which is not included in the training data. A multi-emotion speech database including four emotions, "neutral", "happiness", "sadness", and "anger", is used in our experiment. The results of subjective tests show that the average emotion model can effectively synthesize neutral speech and can be adapted to the target emotion model using very limited training data.

Keywords: average emotion model, model adaptation, affective space.

1 Introduction

With the development of speech synthesis techniques, the intelligibility and naturalness of the synthetic speech has been improved a lot in the last decades. However, it is still a difficult problem for the TTS system to synthesize speech of various speakers and speaking styles with a limited database. It is known that the HMM-based speech synthesis can model speech for different speakers and speaking styles, and voice characteristics of the synthetic speech can be converted from one speaker to another by applying a model adaptation algorithm, such as the MLLR (Maximum Likelihood Linear Regression) algorithm, with a small amount of speech uttered by the target speaker [1], [2], [3]. Furthermore, the HMM-based emotional speech synthesis systems have been successfully constructed by directly training the models with enough emotion data or adapting the source model to the target emotion model when only a limited training data is available [4], [5].

We have realized a HMM-based speech synthesis system in which the LSP (Line Spectral Pair) coefficients and the STRAIGHT analysis-synthesis algorithm are employed [6], [7]. Then, by realizing the MLLR-based model adaptation algorithm, we provide our synthesis system with the ability of synthesizing voice of various speakers with different styles [8]. As only a very limited amount of emotion training data is acquired, we use the model adaptation method to construct our emotional speech system. Commonly, the source model for emotion adaptation is trained using only neutral speech data. But in this paper, we train an emotion-independent model using a

Q. Huo et al. (Eds.): ISCSLP 2006, LNAI 4274, pp. 233 – 240, 2006.

multi-emotion speech database, which includes the neutral, happy and sad speech data of a female speaker. Compared with the neutral model, the average emotion model which considers the distributions of all emotions in the training data is a better coverage of the affective space. In fact, it takes the possible distribution of the target emotion into account, so it can achieve a better adaptation performance than the neutral model. The average emotion model is obtained using a shared decision tree clustering method which assures all nodes of the decision tree always have training data of all emotions [9]. Then we adapt the average emotion model to the target emotion model using a small amount of target speech data and generate the target synthetic speech.

In the following part of this paper, a description of our HMM-based emotional speech synthesis system is presented in section 2. Section 3 presents the speech database information, the training set design and the results of subjective experiments, while section 4 provides a final conclusion.

2 System Description

The framework of our HMM-based emotional speech synthesis system, shown in Figure 1, is the same as the conventional HMM-based synthesis system except that an average emotion model is used as the source model and a MLLR-based model adaptation stage, using context clustering decision tree and appropriate regression matrix, is added between the training stage and the synthesis stage.

In the training stage, the LSP coefficients and the logarithm of fundamental frequency are extracted by the STRAIGHT analysis. Afterwards, their dynamic features including delta and delta-delta coefficients are calculated. The MSD (multi-space probability distribution) HMMs are introduced to model spectrum and pitch patterns because of the discontinuity of pitch observations [10]. And state durations are modeled by the multi-dimensional Gaussian distributions [11]. To obtain the average emotion model, firstly, the context-dependent models without context clustering are separately trained for each emotion. Then all these context-dependent emotion models are clustered using a shared decision tree and the Gaussian pdfs of the average emotion model is calculated by tying all emotions' Gaussian pdfs at every node of the tree. Finally, state duration distributions of the average emotion model are obtained under the same clustering procedure.

In the adaptation stage, the spectrum, pitch and duration HMMs of the average emotion model are all adapted to those of the target emotion. To achieve supersegmental feature adaptation, the context decision tree constructed in the training stage is used to tie regression matrices. And because of the correlations between the LSP coefficients of adjacent orders, the appropriate regression matrix format is adopted according to the different amount of training data. At first, the spectrum and pitch HMMs are adapted to the target emotion HMMs. Then, on the basis of the converted spectrum and pitch HMMs, the target emotional utterances are segmented to get the duration adaptation data. So that the duration model adaptation can be achieved.

In the synthesis stage, according to the given text to be synthesized, a sentence HMM is constructed by concatenating the converted phoneme HMMs. From the sentence HMM, the LSP and pitch parameter sequences are obtained using the speech

parameter generation algorithm, where phoneme durations are determined based on the state duration distributions. Finally, the generated parameter sequences of spectrum, converted from the LSP coefficients, and F0 are put into the STRAIGHT decoder to synthesize the target emotion speech.

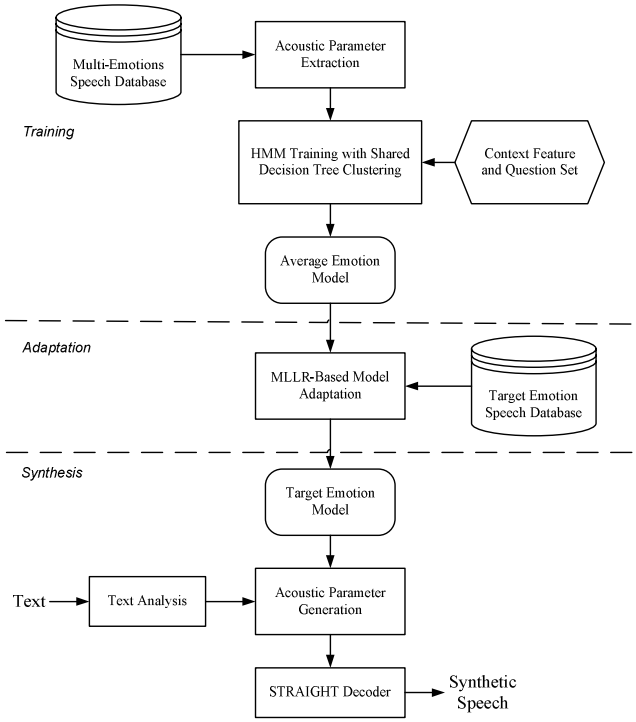

Fig. 1. HMM-based emotional speech synthesis system

3 Experiment and Evaluation

3.1 Speech Database

We constructed a multi-emotion Chinese speech database of a female speaker including four emotions, “neutral”, “happiness”, “sadness” and “anger”. There are phonetically balanced 1200 sentences for “neutral” and 400 sentences for each of the other emotions. Contexts of all the emotion samples are different from each other. Firstly, we evaluated whether the recorded speech samples were uttered in the intended emotions. All the speech samples were randomly presented to ten listeners, and the listeners were asked to select an emotion from the four emotions. The listeners were asked to recognize the emotion of speech samples not by contexts but by acoustic presentations. Table 1 shows the classification rates for each emotion of the recorded speech. We can find that most of the recorded speech can successfully represent the intended emotions.

Table 1. Classification results of the recorded natural speech

	Classification (%)			
	Neutral	Happy	Sad	Angry
Neutral	96.0	2.0	1.0	1.0
Happy	7.0	85.5	0.5	7.0
Sad	5.0	0	91.0	4.0
Angry	1.5	6.0	1.0	91.5

3.2 Training Set Design

In order to realize an average emotion model, a good coverage for the affective space of the training data is expected. The affective space can be described with Russell's circumplex model [12], [13]. As illustrated in Figure 2, Russell has developed a

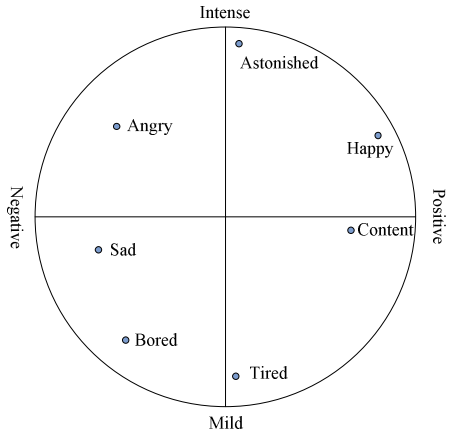

Fig. 2. Circumplex model of affect as described by Russell (1980)

two dimensional circumplex model of affection that makes it straightforward to classify an emotion as close or distant from another one. He called the two dimensions "valence" and "arousal". These terms correspond to a positive/negative dimension and an activity dimension respectively. As the multi-emotion database can only contain several kinds of emotions sampled from the affective space, it is important to choose the most representative emotions for training. In our experiment, the multi-emotion database has four emotions, neutral, happiness, sadness, and anger. We decide to use the speech data of neutral, happiness and sadness as the training data for the average emotion model, because happiness that is a very positive emotion with high arousal and sadness that is a very negative emotion with low arousal almost are two corresponding emotions and can be a rational representation of the affective space. Meanwhile, the angry speech data is left for model adaptation and evaluation.

3.3 Experimental Conditions

The average emotion model is trained by 300 sentences of each emotion, including neutral, happy and sad, selected from the multi-emotion database. A neutral model is trained by 1000 neutral sentences selected from the multi-emotion database for comparison. And 100 angry sentences are used for the model adaptation and evaluation. The speech is sampled at a rate of 16KHz. Spectrum and pitch is obtained by the STRAIGHT analysis. Then they are converted to the LSP coefficients and the logarithm F_0 respectively, and their dynamic parameters are calculated. Finally, the feature vector of spectrum and pitch is composed of the 25-order LSP coefficients including the zeroth coefficient, the logarithm F_0, as well as their delta and delta-delta coefficients. We use the 5-state left-to-right no-skip HMMs in which the spectral part of each state is modeled by the single diagonal Gaussian output distributions. The duration feature vector is a 5 dimensional vector, corresponding to the 5-state HMMs, and the state durations are modeled by the multi-dimensional Gaussian distributions.

3.4 Experiments on the Average Emotion Model and the Neutral Model

Table 2 shows the number of distributions of the average emotion model and the neutral model after decision tree context clustering. Here, we set the weight for adjusting the number of parameters of the model during the shared decision tree context clustering as 0.6. From the table, it can be seen that the two models have comparable distributions.

Table 2. The number of distributions after context clustering

	Neutral Model	Average Emotion Model
Spectrum	3247	3115
F0	4541	5020
Duration	599	589

50 sentences of the synthetic speech generated by each model were also presented to 10 listeners to choose the emotion from the four emotions and the result is illustrated in Table 3. It can be found that both the two models can effectively synthesize neutral speech. However, the result of the neutral model is a little better than that of

Table 3. Classification results of the synthetic speech generated by the neutral model and the average emotion model

	Classification (%)			
	Neutral	Happy	Sad	Angry
Neutral Model	92.2	5.7	2.1	0
Average Emotion Model	84.2	5.0	10.1	0.7

the average emotion model. Some of the synthetic speech generated by the average emotion was misrecognized as sad. That may be because sadness has a better expression than happiness in the training data, as shown in Table 1, so that the average emotion model has a slight bias towards sadness.

3.5 Experiments on the Emotion Adaptation

In the model adaptation stage, the neutral model or the average emotion model is adapted to the target emotion model with 50 angry sentences which are not included in the adaptation training data. The 3-block regression matrix is adopted and the regression matrices are grouped using a context decision tree clustering method. First, 10 listeners were asked to recognize the emotion of 50 synthetic speech samples generated by the two methods from the four emotions. The classification results are presented in Table 4. It can be found that about 70% of the synthetic speech can by successfully recognized by the listeners and the average emotion model has a better adaptation performance.

Table 4. Classification results of the synthetic speech generated by the angry model adapted from the neutral model and the average emotion model

	Classification (%)			
	Neutral	Happy	Sad	Angry
Neutral Model	16.7	2.3	10.4	70.6
Average Emotion Model	13.1	3.4	10.0	73.5

Compared to the speech synthesized by the adapted average emotion model, some speech samples generated by the adapted neutral model sound to be not natural especially in prosody. Figure 3 demonstrates the F0 contours of the synthetic speech generated from the adapted neutral model and the adapted average emotion model respectively, meanwhile the F0 contour of the target speech is also presented. The dotted red line presents the F0 contour generated from the adapted neutral model while the solid

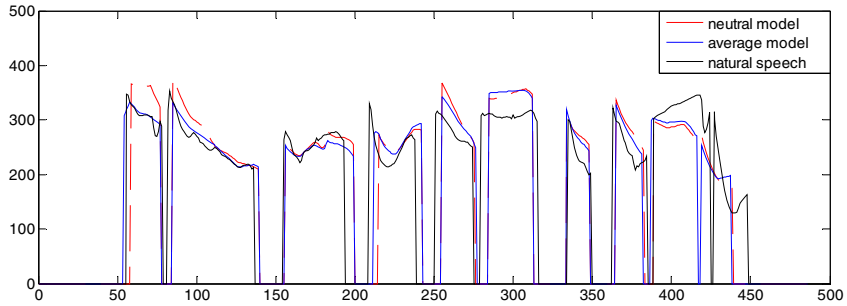

Fig. 3. Comparison of F0 contours generated by the angry model adapted from the neutral model and the average emotion model

blue line is the result of the adapted average emotion model and the solid black line is the F0 contour of target speech. We can see that the values of F0 generated from the adapted average emotion model are more similar to those of the target speech.

4 Conclusion

A HMM-based emotional speech synthesis system is realized using a model adaptation method. At first, an average emotion model is trained using a multi-emotion speech database. Then, the average emotion model is adapted to the target emotion model with a small amount of training data using a MLLR-based model adaptation technique in which a context decision tree is built to group HMMs of the average emotion model. To compare the performance of the proposed method, a neutral model is also trained and adapted. From the results of the subjective tests, it can be seen that both methods can effectively synthesize the intended emotion speech. In addition, the adaptation performance of the average emotion model is slightly better than that of the neutral model.

If having more emotional speech data, there will be a better coverage of the affective space, so we can train a more reasonable average emotion model. Our future work will focus on increasing the number of emotion categories in the multi-emotion database and improving the performance of the average emotion model. At the same time, various emotions will be selected as the target emotion to evaluate the effectiveness of the average emotion model.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under grant number 60475015.

References

1. T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” Proc. ICASSP-1996, pp. 389-392, 1996.
2. C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, no.2, pp. 171-185, 1995.
3. T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speaker adaptation for HMM-based speech synthesis system using MLLR,” The Third ESCA/COCOSDA Workshop on Speech Synthesis, pp. 273-276, Nov. 1998.
4. J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Information and Systems, vol. E88-D, no.3, pp.502-509, March 2005.
5. J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, “Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis,” Proc. ICASSP-2004, vol.1, pp. 5-8, May 2004.

6. H. Kawahara, “Restructuring speech representations using a pitch-adaptive time frequency smoothing and a instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sound”, Speech Communication 27, pp. 187-207, 1999.
7. Y.J. Wu and R.H. Wang, “HMM-based trainable speech synthesis for Chinese,” to appear in Journal of Chinese Information Processing.
8. Long Qin, Yi-Jian Wu, Zhen-Hua Ling, and Ren-Hua Wang, “Improving the performance of HMM-base voice conversion using context clustering decision tree and appropriate regression matrix,” to appear in Proc. ICSLP-2006.
9. J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A context clustering technique for average voice models,” IEICE Trans. Information and Systems, vol. E86-D, no. 3, pp. 534-542, March 2003.
10. K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling,” Proc. ICASSP-1999, pp. 229-232, Mar. 1999.
11. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for HMM-based speech synthesis,” Proc. ICSLP-1998, vol.2, pp. 29-32, Nov. 1998.
12. J.A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, pp. 1161-1178, 1980.
13. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, “ Emotion recognition in human-computer interaction,” IEEE Signal Processing Magazine, Vol. 18, Issue 1, pp. 32-80, Jan. 2001.