Open Access

Study on the Method of Emotional Speech Recognition Modeling

Lianyi Fan*

Foreign Language Department, Shanghai Lixin University of Commerce, Shanghai 201620, P.R. China

Abstract: Emotions play a very significant role in speech recognition. The model built on neutral speech degrades dramatically the recognition of emotional speech. How to deal with emotional issues properly is crucial to achieve good performance in recognition. Most widely used approaches include robust feature extraction, speaker normalization and model retraining. In this paper, a novel method is proposed, which is an adaptive method to transform a neutral-based model into emotion-specific one with a small amount of emotional speech data. It is shown experimentally that the new model achieves higher accuracy in overall performance.

Keywords: Adaptation, Affective computing, emotional, recognition, speech.

1. INTRODUCTION

The research of speech recognition began in 1950s and its symbol is the invention of Audrey, the first speech recognition system that could recognize ten English Alphanumeric in AT&T BELL labs [1]. And since then, speech recognition research has made great achievements in many different fields from the initial stage of isolated words or specific person's speech recognition to today's non-specific person or consistent speech recognition. But due to the fact that all the data used in the lab were collected under ideal condition, the effect of speech recognition in reality is far from being satisfactory because the voice channel, noise, pronunciation or emotions, any of them can influence the effect of speech recognition. Later, researchers focused on how to solve these problems and have made some progress in the fields such as cross-channel, noise-removal speech pattern, *etc* [2-4]. But so far very few papers have been recognized on how to solve emotion-related problems.

With the rapid development in human-machine interaction system, researchers lay more emphasis on the study of "emotion" or "affective computing" and have made some progress regarding facial expression and gesture analysis [5-7]. As one of the important communication means, speech is the most convenient and direct way for humans to communicate with each other. Just like facial expression, speech can also convey rich emotional information. Thus it can be said that the ultimate task of the research of speech recognition is to be able to recognize written as well as emotional information. Lately, emotional speech recognition research has just started all across the world [6, 8]. Considering the great impact that emotion and attitude have on the speech synthesis and recognition, emotional speech research has attracted more attention. Researchers generally focus on the analysis, recognition and synthesis of emotions [9-12], but very few researches have been conducted so far on the recognition of emotional speech.

J.H.L Hansen and S.E. BouGhazale thought that among the affecting factors such as background noise, transmission channel, psychological stress, working pressure and mood changes *etc*, mood changes have the greatest impact on speech recognition [2]. As for the voice variation problem, many researchers have done some relevant studies since 1970s and the solutions can be concluded in terms of levels from the bottom to the top as follows: 1) **Feature level**. The main idea of feature level is to extract some robust features that can represent the voice content without being affected by various variability factors, or can add a variation-regulation process at the recognition stage to reduce its impact on voice features to make sure that the regulated voice and natural voice are as close as possible. In this case, speech recognition device used in regular speech training can obtain a good effect. Hansen assumed that the impact of variability factors can be reduced *via* compensating formant bandwidth and formant position [4]. 2) At acoustic model level some adjustments are made in the features and model training methods according to the characteristics of different voices. For example, Lippm put forward Multi-style training method [13], and the acoustic adaptive method adopted in the paper can be classified into this category. 3) At language model level, some adjustments are made in the language model by using high-level knowledge, for example, Athanaselis T. improved emotional speech recognition rate by increasing the proportion of emotional sentences in the language model [14].

While modeling each different emotional speech, its recognition rate can certainly be improved to a great extent but it is impractical in reality because it is very difficult to collect neutral voice samples to obtain the related data, as it is very demanding for the readers. This paper selects several basic speech patterns to study their impact on speech recognition and uses a small amount of the related emotion-

*Address correspondence to this author at the Foreign Language Department, Shanghai Lixin University of Commerce, Shanghai, P.R. China, 201620; Tel: 021-67705162; E-mail: fanlianyi2009@163.com

al speech data, which are based on neutral voices *via* adaptive method, to improve their recognition rate.

The structure of this paper is as follows: the second part involves the instruction of emotion database and its application; the third part introduces the baseline system and adaptive system; the fourth part shows the analysis of the experimental results and the last part provides the conclusion of the paper and the future research orientation.

2. EMOTIONAL SPEECH DATABASE

2.1. Speech Data Category

Many researchers have performed researches on how to categorize the emotions but so far they have not reached a consistent standard as it is a very complicated issue [15, 16]. At the present stage, researchers usually define several basic emotions based on the actual situation and their own understanding. In this research, the most common five emotions are discussed, namely, neutral, happiness, anger, fear, and sadness. In order to facilitate drafting, the symbol Nis used to represents neutral, H for happiness, A for anger, F for fear and S for sadness.

2.2. Description of Database

In the experiment, 200 lexical items were recorded in the emotional speech database with almost all the most frequently used initial and final sounds [17]. The pronunciation personnel were asked to pronounce each lexical item at a time with five different basic emotions as mentioned above. In order for the recording to be authentic, all the pronunciation samples were well selected from the university students, making a total of 50 students, including 25 boys and 25 girls. And the recording was done in a quiet office with minimum noise. In the study, 10 boys and 10 girls' recordings were randomly selected as the testing voice and the rest of the students' voices were left as retraining voices. The emotional voice database composition is shown in Table **1**.

3. BASELINE SYSTEM

The research adopted extended context-related initial and final sounds to design the acoustic model. (Tri-XI F) [17]. Considering the limitation of the recorded data and the nonexistence of some specific acoustic primitives, some adjustments were made in the acoustic primitives. The acoustic model is based on HMM with each primitive comprising three states and each mixed state being described with four Gauss model. Model training was carried out in a mixed-and-split way. The number of states of each model was controlled at about 200 by using state sharing strategy based on the policy- decision-tree. The characteristic parameter of the model is the 39-dimensional MFCC including energy parameters, as well as the first-order and second-order differentials. In the experiment, neutral, fear, sadness and happiness voices were used respectively to retrain five different acoustic models, including one neutral acoustic model and four emotion acoustic models and each model's performance was evaluated in terms of its related speech recognition rates. Fig. (**1**) shows the comparison of the testing results' statistics between neutral acoustic model and the remaining four emotion acoustic models.

From Fig. (**1**) it can be seen that : 1) voice variations caused by emotions have great impact on speech recognition rate. When neutral voice model was used to test various different voices' data, it was observed that the neutral voice accuracy rate reached 90.83% but the recognition of the other four emotional voices had different degrees of decline and the recognition of the anger voice had the lowest accuracy rate. These results indicate that emotions indeed have great impact on speech recognition rate and the neutral acoustic model has poor performance on the recognition of emotion speeches. The research on the recognition of emotion speeches is of great significance because they are often encountered in practice. 2) The acoustic model of emotional data retraining can effectively improve emotional voices' recognition rate. Compared with neutral acoustic model, the acoustic model of emotional voice retraining can tremendously improve the recognition rate of the same type of emotional data to be tested. It needs to collect a great amount of emotional voice data for modeling each specific emotional voice and with the increase in the number of emotional categories, the amount of the related data will dramatically increase. But if the emotional voice is already known and the relevant compatible acoustic model is selected, the whole recognition rate can reach as high as 82.56%, though the accuracy rate is not as high as 90.83% for the neutral voice as shown in Table **2**. Therefore, it does not necessarily imply high cost to collect a great amount of emotional data for specific training as well as guarantee a high recognition rate.

4. EMOTIONAL ACOUSTIC MODEL ADAPATATION

Adaptive technology of acoustic model has been widely used in the field of speech recognition [1]. If a small amount of adaptation data is used to regulate the acoustic model, the recognition system can better match the variations caused by microphone, channel, environmental noise, speakers *etc*. [8, 18]. In order to solve the problem of voice changes caused

Table 1. The emotional voice database composition.

Emotions Speeches	N	F	S	A	H
Training (15boys+15girls)	6000 sentences	6000 sentences	6000 sentences	6000 sentences	6000 sentences
Test（10boys+10girls）	4000 sentences	4000 sentences	4000 sentences	4000 sentences	4000 sentences

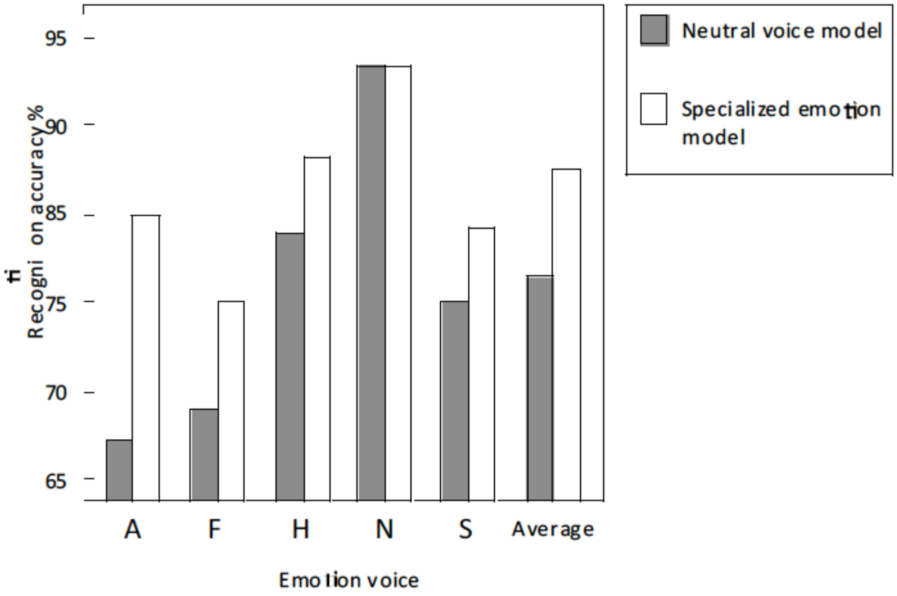

Fig. (1). Performance comparison between neutral voice model and specialized emotion model.

Table 2. Statistics of the results of cross-test of emotional acoustic models on different specific emotional speeches.

M / T	A		F		H		N	S		Average	
	A-mod	**A-Adapt**	**F-mod**	**F-Adapt**	**H-mod**	**H-Adapt**	**N-mod**	**S-mod**	**S-Adapt**	**X-mod**	**X-Adapt**
A	81.67	74.45	70.93	68.22	74.96	69.55	67.67	63.47	65.54	71.74	69.09
F	61.24	67.90	75.45	73.18	70.12	69.22	69.94	72.55	71.69	69.86	70.39
H	71.82	79.66	79.76	81.16	84.59	83.04	80.66	77.33	79.12	78.83	80.73
N	66.48	83.77	79.64	88.20	81.37	86.92	90.83	86.60	89.18	80.98	87.78
S	50.89	68.37	71.99	76.31	67.13	70.91	75.82	80.24	78.04	69.21	73.89
Average	66.42	74.83	75.55	77.41	75.63	75.93	76.98	76.04	76.71	74.13	76.37
E RR	25.04		7.61		1.23		--	2.80		8.69	

by speech emotions, this paper aims to discuss how to apply acoustic model adaptive to the emotional speech recognition in an attempt to discover the way to reduce emotional voice impact on the speech recognition.

At present, the most commonly used acoustic adaptive technology is model parameter adaptive conversion method. It is mainly targeted at the HMM model to make adaptive transform, for example, MAP method and MLLR method [17]. Considering the need of adaptive data and speed, this research selected MLLR method because it needs few adaptive data but can retain fast adaptive speed. This method can transform an initial model into a new adaptive model *via* a linear transformation using Baum-Welch maximum likelihood principle to re-evaluate linear transformation matrix.

In the following experiment based on adaptive acoustic model , the original acoustic model was obtained by neutral acoustic data training, namely, neutral acoustic model in baseline system. In the training process, all together 6000 neutral acoustic data were used.

4.1. Specific Emotion Adaptation

In order to improve neutral acoustic model's emotional speech recognition rate, few emotional data and MLLR method were used to adapt neutral acoustic model and four emotional acoustic models were obtained. This study used 1500 sentences for each specific emotion, respectively.

The following Table **2** shows the statistics of the cross-test results of the emotion acoustic models, which were obtained *via* different trainings, and large emotional acoustic data. From the table it can be observed that the average recognition rate of the adaptive emotional acoustic models on the tested data is 8.69% higher than that of the acoustic model simply obtained by emotional acoustic training. This is because in the training, not only adaptive data was involved but it also involved the data used in the neutral acoustic training.

But one point should be noticed that though both of the two different emotional acoustic models, whether obtained

Table 3. Comparison between the different acoustic models.

Emotion category of the tested speech	A	F	H	N	S	Average
N-mod	67.67	69.94	80.66	90.83	75.82	76.98
X-Adapt	74.45	73.18	83.04	90.83	78.04	79.91
Mix-Adapt	71.24	73.12	83.32	89.45	76.16	78.66

through emotional acoustic training or resulting from adaptive emotion, can improve the related emotional speech recognition rate, they also increased the error rate of the other different emotional speech recognition. Therefore, it is obvious that the compatibility between different emotional speech models is poor. Thus it is recommended to select appropriate emotional speech model in light of the category of emotional speech.

4.2. Mixed Data Adaptation

This paper divided emotion models into different categories for adaptation, though the test performance can be improved on the related emotional speech, but the average recognition rate cannot be improved in the whole collected data. It is because of the fact that different emotions have different impact on the voice variation. If only one emotional voice is used for adaptation and obtaining a model, the designed model will not effectively characterize other different types of speech emotions. And thus the emotional speech recognition rate would not be improved.

In order to make the adaptive acoustic model effectively characterize different emotional speeches, 4 different adaptive acoustic data were integrated and a set of new adaptive data was obtained. Table **3** shows the result of the test of the neutral acoustic model on mixed adaptive model, and the result of the emotional voice tested on the adaptive model.

As this paper used many different kinds of emotional voices, therefore, voice variations caused by different emotions could be reflected in the adaptive data. As shown in Table **3**, the recognition rate of the mixed adaptive model on neutral voice decreased a little, but as for other emotional voices, its recognition rate improved tremendously. Besides this, its average recognition rate of the whole data tested was improved in comparison to the neutral acoustic model, while, its error rate decreased by 7.3%. Its recognition rate was 79.91% not as high as the proprietary emotion model can achieve, but the latter needs to integrate the results of many emotion adaptive models. Besides, it needs to correctly categorize the voices to be tested.

CONCLUSION

The paper discussed how to model emotion acoustic model by adaptive acoustic model method and its application. The results can be concluded as follows:

1. To each emotional voice, if it is tested on the same type of emotional acoustic model, the recognition rate is higher than that of other emotion acoustic model, but it is not as high as neutral voice tested on the neutral acoustic model. It indicates that acoustic characteristics parameter (MFCC) used in the experiment cannot effectively characterize the emotional voices and it leaves much room for improvement.
2. Emotion adaptive models obtained from the integration of small amount of emotion data and neutral acoustic model can effectively improve its recognition rate of the related emotion voices. And the mixed adaptive model can improve overall recognition rate of the emotional voices.

As the research on emotional speech is only at the initial stage therefore many problems remain to be resolved. How to take advantage of emotional information and different integrated models to improve overall emotion speech? Or how to remove the impact of emotion on the speech at the feature level for extracting more robust features characterizing the emotion variation? All these questions are worth further exploring in the future.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Y. Wang, *Speech Recognition Adaptation Application Technology Research and Realization*, Tsinghua University, China, 2000, pp. 3-10.

[2] J.H.L. Hansen, and S.E. BouGhazale, “Getting started with SUS AS: A speech under simulated and actual stress database,” In: *EU-ROSPEECH-97: Eur Conference on Speech Communication Technology*, vol. 4, 1997, pp. 123-145.

[3] S. E. Bou-Ghazale, and J.H.L. Hansen, “A comparative study of traditional and newly proposed features for recognition of speech under stress,” *IEEE Transactions on Speech & Audio Processing*, pp. 76-87, 2000.

[4] J.H.L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Communication*, vol. 20, no. 1-2, pp. 151-173, 1996.

[5] Ba. Hu, and T. Tan, “Affective computing-the new development and research,” *Science Times*, vol. 31, no. 3, pp. 3-5, 2000.

[6] R. W. Picard, *Affective Computing*, Cambridge, Massachusetts, London, England, The MIT Press, 1998, pp. 145-149.

[7] J. Tao, and T. Tan, “Digitized human emotions-harmonious human machine interactive emotional computing,” *Micocomputer World*, vol. 9, no. 1, pp. 29-32, 2004.

[8] J. Han, and Y. Shao, “*New development of speech-based signal processing*,” Online Chinese Scientific Paper, 2005, pp. 57-63.

[9] J. Zhao, and X. Qian, “Emotion features analysis & recognition study in speech signal,” *Journal of Communications*, vol. 39, no. 4, pp. 41-48, 2000.

[10] J. Tao, and Y. Kang, "Features importance analysis for emotional speech classification," In: *Proceeding of ACII*, 2005, pp. 56-63.

[11] T.L. New, "Speech emotion recognition using Hidden Markov models," *Speech Communication*, vol. 41, pp. 603-623, 2003.

[12] D.N. Jang, W. Zhang, L. Shen, and L.H. Cai, "Prosody analysis and modeling for e- motional speech synthesis," *International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 453-468, 2005.

[13] R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-style training for robust isolated word speech recognition," In: *International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), IEEE Press, USA, 1987, pp. 705-708.

[14] T. Athanaselis, "ASR For emotional speech: Clarifying the issues and enhancing performance," *Neutral Networks*, vol. 18, 2005, pp. 23-28.

[15] D.N. Jiang, and L.H. Cai, "Classifying emotion in chinese speech by decom posing prosodic features," *International Conference on Spoken Language Processing INTER SPEECH 2004-ICS LP*, 2004, pp. 56-61.

[16] J. Nicholson, K. Takahash, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing and Applications*, vol. 18, no. 3, 2000, pp. 324-329.

[17] Ling Li and Fang Zheng. "Chinese consistent speech recognition of the context-related Vowel sound modelling research," *Tsinghua University Journal (Natural Sciences)*, vol. 44, no. 1, 2004, pp. 61-64.

[18] C.J. Leggetter, and P.C. Woodland, "*Speaker Adaptation of HMMs Using Linear Regression*", Technical Report, CUED / F-INFENG/TR181. Cambridge University Technical Report, 1994, pp. 87-92.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015