We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Frequency Shifting for Emotional Speaker Recognition

Yingchun Yang, Zhenyu Shan and Zhaohui Wu
*College of Computer Science and Technology, Zhejiang University
 P.R.China*

1. Introduction

The task of automatic speaker recognition is to determine the identity of a speaker by machine (Herbert & Michael, 1994 and Bimbot et al., 2004), which is the one of the main applications in pattern recognition. According to different application fields, it encompasses speaker verification and speaker identification. Speaker verification is to verify a person's claimed identity from his/her voice, and gives a result "Yes" or "No" (Joseph & Campbell, 1997). In speaker identification, there is no a priori identity claim, and the system decides who the person is, what group the person is a member of, or that the person is unknown.

In general, the process of speaker recognition has two steps: feature extraction and classification. The process of feature extraction is to extract features from the utterance for training or testing. Many effective features are found in researches, such as MFCC (Rivara et al., 1996 & 1999), LPC (Philippe &, 1995), PLP (Hermansky, 1990), Pitch (e.g. Atal, 1969) and Energy. The classification includes training and testing. Before testing, the recognition system must get familiar with the voices. Thus, in training, the system needs to collect speech of the registered person. Then in testing, the system compares an unidentified utterance to the trained model and makes identification or verification. The popular classification methods (or models) include GMM (Reynolds & Richard, 1995), GMM-UBM (Reynolds et al., 1999), SVM (e.g., Wan & Campbell, 2000), HMM (Tishby, 1991), and DBN (Sang et al., 2003).

Although great progress has been made in last 20 years, there are still many problems for speaker recognition system in real applications. The recognition performance is reduced by many factors (Sadaoki 1997), such as background noise (e.g. Ming et al., 2008), channel effects (e.g. Reynolds, 2004) and emotion variability. In the paper (Shan et al., 2006), a speaker check-in system is evaluated in an office with 38 users (25 males, 13 female, aged from 20 to 38) for 13 months. In the system, MFCC and GMM are applied as feature extraction method and modelling method. About 34% of the recognition errors are caused by noise and 35% by un-described reasons, in which emotion variability is the main factor.

Speaker emotion variability means the emotion states mismatch between the training and testing speech, and such kind of recognition is called emotional speaker recognition or affective speaker recognition. Different emotion states will affect speech production mechanism of a speaker in different ways, thus lead to acoustical changes in his/her speech

(Scherer, 1998 and 2000). As concluded in (Louis, 2003 and Scherer, 2003), the emotion factor will change more than 20 types of speech features, such as pitch, Intensity, formants, speech rate, energy and duration. These changes can induce intra-speaker vocal variability, which will aggravate the recognition performance.

The remainder of this chapter is organized as follows. In Section 2, the problem of emotional speaker recognition is analyzed and some methods are introduced. In Section 3, we proposed the frequency shifting method for emotional speaker recognition, including modelling and feature extraction method. In Section 4, an emotional speech corpus MASC is described briefly. In Section 5, the performance evaluation for the shifting method is given. Discussions and future works are drawn in Section 6.

2. Emotional Speaker Recognition

This section will give a review of some methods and related work in emotional speaker recognition. According to the difference of emotion states of the training and testing utterances, the emotional speaker recognition system often works in two situations. The first one is that both the training and testing data include the utterances with the same emotion states, though its performance is still worse than the situation when the emotion state is neutral (Scherer et al., 2000 and Wu W. et al., 2006). The other is that the emotion state of testing utterance is excluded in the training utterance. It usually occurs in the real application, where users often provide neutral speech in training yet various emotional utterances in testing. The system in this situation performs even worse than in the first one (Wu et al., 2006; Shan et al., 2007 and Huang et al., 2008), and it attracts our focal attention. The reason for poor performance is the mismatching between training and testing conditions. It also exists in other recognition applications, such as noise and channel effect. Some methods have been proposed to overcome this limitation. For example, for solving noise or channel effect, the compensation method, based on the relationship between different channels (Wu et al., 2006) or noise types, is applied to eliminate the negative effect. Under noisy surroundings, some rules are presented to see whether one frame is contaminated by noise and the clean part is used for testing. The idea of these two methods is helpful to emotional speaker recognition. However, emotion has its particularities. For instance, the deformation of voice by the same channel is fixed thus has its own rule, and the effects of emotion on voice deformation are more complex, for different person and text will affect the degree of deformation.

The mismatch in emotional speaker recognition contains three conditions. In the first condition, the pattern of testing data is not contained in training speech. For example, the testing utterances contain sad speech, and the training utterances not, and the recognition system can't match the pattern of the utterance in testing. In the second, the training data contains all the pattern in the testing utterance, but the pattern is not matched extract. Owing to the difference of emotional intensity and depth, the pattern may not be similar even if the emotion state is the same. In some cases, the mismatch happens, where the testing data is mixture of happy and neutral utterance. In the third condition, only neutral speech can be gathered in training and various emotion utterances exist in testing. In the ordinary, the emotion state of testing data is also unknown. It may need to identify the emotion state in testing and the recognition performance will be affected by the emotion

identification rate. It is the most common condition in real application. And in the following, we primarily review the methods for solving this type problem.

In our conclusion, the method of solving the mismatch problem has three ideas. The first idea is to extract the feature which is not varied by the emotion variation, which is the most effective way. Nevertheless it is the hardest, because recent experiment result shows that all types of features for speaker recognition (as we known) will be affected by emotion variation. For a brief explanation, the emotion varies the frequency spectrum in a complex way and the common features for recognition are extracted from the frequency spectrum. The second method tries to find a relation between different emotion states, including its feature, model and scores. The aim is to build a relationship between neutral utterance and other emotional utterance. Then, this relationship is adopted to transfer neutral speech to emotion speech. We named this method emotion enriching. The third method is to collect the neutral part from the emotion utterance for testing. It often needs to compute the emotional factor to define the emotion part, which is named emotion regulation. In the following of this section, we will introduce the recent presented methods from the second or third idea.

Zheng(Bao et al., 2007 & Wu et al., 2006) considered that emotion effect is somewhat similar to that of channel effect on speaker recognition. They realized that when training and testing speech are in different emotions, the discrepancy between the speaker models and the test utterances will induce biases in verification scores. Thus, the ideas of handling the channel effect was borrowed to alleviate the negative effect of emotion(Bao et al., 2006). In their former work, an emotion-dependent score normalization method(E-norm) (Wu et al., 2006)derived from H-norm(Reynolds et al., 2000) was proposed, which was originally designed to alleviate channel effect in cross-channel speaker verification. It is designed to estimate these emotion-dependent biases from development data. Then, the biases are removed from verification scores. Soon afterwards, they furthered their approach by proposing emotion attribute projection(EAP) method(Bao et al., 2007), which removes emotion variability from SVM expansion dimensions. The idea is borrowed from the nuisance attribute projection (NAP), which has been proven to be a successful channel compensation method(Campbell et al., 2006 and Solomonoff et al., 2005). It can achieve a better result because it removes the subspace that may cause the emotion variability in the kernel of an SVM system.

Huang (Huang et al., 2008) studied the difference between various emotions and propose a method based on the Pitch-dependent difference detection and modification (PDDM). The basic idea is to choose the neutral part of testing utterance for final scoring. The pitch is used as the emotion factor to distinguish "neutral" part from emotion utterance. First, it classifies the test utterances into HD (High Different from neutral, and with high identification error rate) and LD (Low Different from neutral, and with low identification error rate) group. Then, it modifies the segments with intense emotional information in the HD utterance to reduce the impact of unmatched emotion states in the training and testing.

In most circumstances, the emotional element, if there is any, always lasts ephemeral when users provide testing utterance, since they tend to provide neutral speech for testing in real application(as discussed in (Shan et al., 2006). The testing utterance is a mixture of neutral and emotion speech, and it is common in emotional speaker recognition. The paper(Shan et al., 2009) experimentally analyzed the performance of the GMM-based verification system with the utterances in this situation and two results are concluded. One is applying the test

utterances with low emotion ratio, which suggests the user to provide pure neutral speech for testing or eliminating the non-neutral part from the testing speech. The other is increasing the testing utterances length, yet it can't be satisfied in most applications. Thus, a simple is proposed method to distinguish the non-neutral features from the neutral ones in the scoring processing of testing with the purpose to reduce the emotion ratio of testing utterances. First, all features' scores and the average of these scores are calculated. Second, the scores higher than the average are selected for the final score computation. In other words, not all features are effective in the score computation thus some of them are treated as non-neutral features and eliminated to decrease the emotion ratio.

In about ten years ago, Scherer(Scherer, 2000) presented a structured training approach which aims at making the system get familiar with the emotion variation of the user's voice in training. In this method, the registered speakers are asked to provide speech on various emotional states. It doesn't fit the common situation as we described above. However, it gives a direction for eliminating the emotion effect: the question is that how to synthesis emotion speech, or how to obtain the emotion feature, or how to train the emotion model, or how to calculate the scores of the feature vectors against his/her emotion model, when only the natural utterance is obtained in training.

In the paper(Wu Z.H. et al., 2006), it was found the main reason that causes the performance degradation is the absence of emotion speech. Analysis of emotional speech and its synthesis rules have been researched for many years. It is a way to synthesize emotion speech based on these rules. They investigate the rules based feature modification for robust speaker recognition with emotional speech. A feature modification rule is developed to convert neutral speech to emotion speech. First, the rules of prosodic features modification are learned from a small amount of the content matched source-target pairs. Then, features with emotion information are adapted from the prevalent neutral features by applying the modification rules. Finally, the converted features are trained together with the neutral features to build the speaker models. The speaker models are trained with both the neutral speech provided by the users and generated output speech perceived as conveying emotions.

In order to train the emotion model from his/her neutral model, Shan(Shan et al., 2007) presented a neutral-emotion GMM transformation algorithm, which is based on the assumption: if two speakers' natural speech space satisfies the similar distribution, so does their emotion speech space, especially when they share the same culture. First, the KL-distance between neutral GMMs (order n) is calculated to find out k speakers from the emotion database, who have the similar feature space with the registered speaker. Then, these speakers' emotion GMM is transformed to the registered speaker's emotion GMM. However, the order of the transformed GMM is n^k times of the original one and leads to an exponential increase in computation cost. To overcome this limitation, a neutral-emotion GMM transformation algorithm was presented based on the same assumption as above. In this method, the transformation function is defined by a polynomial function(Shan et al., 2008). It establishes a relationship between neutral and emotion GMM. Applying this relationship, the emotion GMM is obtained from the neutral GMM with the same order. The model is adopted in the emotional speaker recognition with increasing less cost of training and testing. However, these methods demand that the system is aware of all the emotion states of the testing utterances in training. And only the emotion states in testing dataset are

considered. If the utterance with new emotion state is added into the testing dataset, the model needs to be retrained for this emotion state.

In following sections, we present the frequency shifting method for emotional speaker recognition. In this new method, only neutral speech is required in training, and the testing speech is in multi-emotion conditions and the emotion state is not available. To model the speech of multi-emotion conditions, we introduce the multi-condition model, which has been successfully used in speech recognition to account for varying noise sources or speaking styles. The researches (e.g., Lippmann et al., 1987 and Scherer, 2000) indicate that it can improve the system robustness when the training data sets are in multi-conditions, thus we believe it is suitable for the multi-emotion conditions. In further, the maximum scoring method is proposed which takes the emotion occurrences of testing utterances into consideration. Thus the adversely effect of the mismatch of emotion state in training and testing can be weakened. In order to generate training speech with multi-emotions, we proposed a method to synthesize emotion utterances by shifting the spectrum of neutral speech. The energy distribution of neutral utterance is changed to create utterances with different energy distribution, which are used as emotion utterances for training. Comparative experiments carried on the MASC database(Wu, T., 2006) show that the new method is superior to the baseline system. The frequency shifting method combining with the maximum scoring method improve the system robustness.

3. Frequency Shifting Method

In this section, the frequency shifting method is introduced in two parts: model and emotion speech synthesis.

3.1 Model

In general speaker recognition, the training data set (Φ_0) only contains neutral utterance of speaker S , which is represented by the likelihood function $p(X|S, \Phi_0)$ of feature vector X associated with speaker S trained on data set Φ_0 . The Gaussian Mixture Model (GMM) is a preferable likelihood function(Reynolds et al., 1995), which is a weighted sum of Gaussian components. Each speaker S is indicated by a GMM λ . Generally, Expectation-Maximization (EM) algorithm is applied in the training process to find out the parameters of λ to maximize $p(X|S, \Phi_0)$ with respect to Φ_0 .

In emotional speaker recognition, training speech with various emotion states should be created from Φ_0 at first. This leads to augment training sets $\Phi_1, \Phi_2, \dots, \Phi_L$, where Φ_l denotes the l^{th} emotion speech and L is the number of emotion states. The direct modeling method is to apply the structured training method to employ all the emotional data sets to train a likelihood function $p(X|S, \Phi_0, \Phi_1, \dots, \Phi_L)$. It demands that the length of the training utterance of each emotion state is almost the same; otherwise its performance will be affected, when the specific emotion state of the testing utterance takes small proportion in training. According to overcome this limitation, we introduce the multi-condition model(Atal, 1974), which is formed by combining the likelihood functions trained on the individual training set:

$$p(X|S) = \sum_{l=0}^L p(X|S, \Phi_l) P(\Phi_l|S) \quad (1)$$

where $p(X|S, \Phi_l)$ is the likelihood function of frame vector X trained on set Φ_l , and $P(\Phi_l|S)$ is the prior probability for the occurrence of l^{th} emotion speech of speaker S . As $P(\Phi_l|S)$ is fixed, maximizing $p(X|S)$ in training is to maximize $p(X|S, \Phi_l)$ individually. In testing, the first is to calculate the posterior probability of speech given feature vector X with respect to speaker S :

$$p(S|X) = \frac{p(X|S)P(S)}{P(X)} \quad (2)$$

where the equation is based on Bayes' rule. By replacing $p(X|S)$ in Equation (2) with Equation (1) and assuming an equal prior $P(S)$ for all the speakers, we obtain an operational version of Equation (2) for recognition.

$$p(S|X) = \sum_{l=0}^L p(X|S, \Phi_l) P(\Phi_l|S) \quad (3)$$

And the posterior likelihood of a speaker given an utterance $X^T = \{X_1, X_2, \dots, X_T\}$ with T frames is defined as:

$$p(S|X^T) = \frac{1}{T} \sum_{t=1}^T \log p(S|X_t^T) \quad (4)$$

where $p(S|X_t)$ is defined by Equation (3).

In our assumption, the emotion state of feature vector X is not available, so the value of $P(\Phi_l|S)$ can't be decided. If each emotion state occurrences in testing is the same, $P(\Phi_l|S)$ has the equal value for each emotion state. Equation (3) can be simplified:

$$p(S|X) = \text{mean}_{l=0}^L p(X|S, \Phi_l) \quad (5)$$

In this equation, the average of $p(X|S, \Phi_l)$ is used for scoring, so it is named the mean scoring method. Since all of $P(\Phi_l|S)$ is the same, it neglects the effect of the emotion state of feature vector X . In our opinion, the emotion information of testing speech may overcome the degradation of the performance caused by different emotion states of testing and training speech. In testing, it needs to find the most likely emotion state for a given feature vector X . One feature vector only belongs to one emotion condition. Thus, the value of $P(\Phi_l|S)$ for a feature vector X is defined as:

$$P(\Phi_l|S) = \begin{cases} 1, & \text{if the emotion state of } X \text{ is } l \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

As $p(X|S, \Phi_l)$ in Equation (1) describes the distribution of both the speaker and emotion feature space, a simple method based on $p(X|S, \Phi_l)$ is proposed. The object is to find which emotion state has the maximum posteriori probability to feature X . Formally,

$$l = \arg \max p(S, \Phi_l | X) = \arg \max \frac{p(X | \Phi_l, S)p(X)}{p(\Phi_l, S)} \approx \arg \max p(X | \Phi_l, S) \quad (7)$$

where the second equation is due to Bayes' rule. Combining Equation (6) and Equation (7), Equation (3) can be reduced to:

$$p(S | X) = \max p(X | S, \Phi_l) \quad (8)$$

According to the deduction, the final result relates to the maximum of $p(X | S, \Phi_l)$, so it is named the maximum scoring method.

3.2 Emotion Speech Synthesis

In the last subsection, we assume only neutral speech Φ_0 can be obtained in training. To adopt the multi-condition model in multi-emotion conditions, the training sets $\Phi_1, \Phi_2, \dots, \Phi_L$ should be generated from neutral speech Φ_0 . The method of converting neutral speech to emotion speech will be described in this section.

Much effort has been devoted to analyze the relationship between neutral and emotion speech, and many qualitative relations have been published. But it is troublesome to quantitatively analyze the relationship, because it is not only speaker dependent but also text dependent. Thus, it is hard to convert neutral speech to emotion speech exactly in the text independent speaker recognition. However, the research on acoustic changes benefits us to transform neutral speech to emotion speech.

Fig. 1. The spectrum of certain frame vector of anger and neutral speech (voicing [e]). The dashed and solid line represents the anger and neutral spectrum individually.

Nowadays, the research on the acoustic changes is mainly in the frequency domain. The variation of formats represents the difference between neutral and emotion speech. Take anger speech for example, the average of F0, F1 and F2 of anger speech is smaller than neutral speech. Figure 1 shows the spectrum of certain feature vector extracted from anger and neutral utterance (voicing [e]). It indicates the anger speech can be shifted from neutral spectrum to change the formats. Hence, the emotion spectrum can be converted from the neutral spectrum by shifting its spectrum, though it is not precise.

In the multi-conditions model, the aim is to synthesize all kinds of emotion speeches while the emotion state of synthesized speech is not concerned. Thus, we are even unaware of that in which direction and how long to shift the neutral spectrum. The training sets $\Phi_1, \Phi_2, \dots, \Phi_L$ can be synthesized by changing the spectrum of neutral speech Φ_0 . The spectrum shifting method is proposed:

$$E_e(f) = E_n(f + h) \quad (9)$$

where E_e and E_n is the energy in frequency f of emotion and neutral speech individually, and h is the length of shifting. According to this, the “emotion” speech can be obtained from the neutral speech.

The synthesized speech can't be categorized as a specific emotion state, only indicating the formats variation of neutral speech. In fact, it also changes the pitch and the energy distribution of the frame. In the structured training method, the aim is to synthesize all kinds of emotion utterances while the emotion state of synthesized speech is not concerned. Thus, the shifting method is suitable for obtaining the training dataset with different emotion states.

The MFCC is most widely used frequency domain feature in speaker recognition. Adding the spectrum shifting method to MFCC extraction, it can extract emotion features from neutral speech with the different value of h . As shown in Figure 2, the process in the dashed rectangle is added and other steps are the same with the original.

Fig. 2. The process of MFCC and the crewel frame is added to extract emotion features from neutral speech.

4. Emotional Database

An emotional speech database MASC (Mandarin Affective Speech Corpus) is used in our experiments (Wu, T. et al., 2006), which is available publicly. It is constructed with two major objectives. On one hand, it is used for prosodic and linguistic investigation of emotion expression in Mandarin. On the other hand, it supplies a training set as well as a test data set for speaker recognition system affected by emotional factors. Compared with other emotional speech database, MASC concentrates on showing both the characteristics of different emotional states and the intra-speaker variability caused by state changes of speakers. In particular, these records are spoken in Mandarin which is a fairly fresh area of emotional speech studies.

The selection of emotional states is expected to put the speech on the emotion wheel which has been derived from the Plutchik's work (Plutchik, 1980). It is a model to describe the activation-evaluation-power space of emotion. According to this emotion model, emotional states distribute on a circle which is named Emotion Wheel. The centre of this circle stands for the natural origin, a state which gathers all kinds of emotional factors. However, the effects from these emotional factors are so weak that they cannot emerge at the origin. Each emotional state is defined with a unique planar vector $\vec{E}$ that has two parameters, emotional intensity and emotional orientation. Emotional intensity indicates the range of $\vec{E}$ and emotional orientation renders the angle of $\vec{E}$. In terms of the emotion wheel, four emotion types are selected in the corpus: anger, elation, panic and sadness, whose descriptions are consulted by Banse & Scherer (1996).

- ◆ Neutral - Simple statements without any emotion.
- ◆ Anger - A strong feeling of displeasure or hostility.
- ◆ Elation - Be glad or happy because of praise.
- ◆ Panic - A sudden, overpowering terror, often affecting many people at once.
- ◆ Sadness - Affected or characterized by sorrow or unhappiness.

The corpus contains recordings of 68 native speakers (23 female and 45 male) and five kinds of emotion states: neutral, anger, elation, panic and sadness. Each speaker reads 5 phrases, 20 sentences three times for each emotion state and 2 paragraphs only for neutral. These materials cover all the phonemes in Chinese. The sentences include all the phonemes and most common consonant clusters in Mandarin. The types of sentences are: simple statements, a declarative sentence with an enumeration, general questions (yes/no question), alternative questions, imperative sentences, exclamatory sentences, special questions (wh-questions). The sampling rate of the utterances used in the experiments is 8000Hz. More details can be found in (Wu, T. et al., 2006).

5. Experiments

5.1 Evaluation of the Shifting Method

In this subsection, an experiment is designed to compare the speakers' emotion speech with the ones synthesized by the spectrum shifting method. The object is to see their similarity. The experiment employs five phrases read for three times of all speakers from the MASC. The neutral and emotion phrase with the same text and same speaker is compared, so the comparison is text dependent and speaker dependent.

First, the speech is segmented into frames by a 20-ms window progressing at a 10-ms frame rate. Secondly, FFT of the size 512 is used to transform the speech frames into frequency domain. Then, spectrogram shifting method is applied to obtain synthesized spectrums. The length unit of shifting is $8000/512=15.625\text{Hz}$. And in the experiment, the shifting length h is $\pm 1, \pm 2, \pm 3, \pm 4, \pm 5$ unit. Thus, 10 phrases are created from one neutral phrase. Finally, reverse FFT transforms them into time domain to get emotion utterance.

Here, the similarity of two phrases is indicated by the DTW(Sakoe, 1978 and Joseph 1997) distance. The distances from speaker's emotion phrase to the neutral one and to the synthesized phrases are calculated. For each emotion state to neutral, there is $5 \times 3 \times 68 = 1020$ times of distance calculation. Because the number of synthesized phrases of one neutral phrase is 10, so its distance is based on DTW:

$$D = \min_h (DTW(X_E || X_N^h)) \quad (10)$$

where X_E and X_N^h represent emotion speech and synthesized speech by shifting h HZ individually.

The result is shown in Figure 3. The distance from synthesized phrase to emotion phrase is about half the distance to neutral one. The synthesized is more similar to speaker's emotion speech than neutral. It indicates the synthesized speech can make the system familiar with the emotion speech in a certain way.

Fig. 3. The comparison of the distances from speaker's emotion phrase to the neutral one and to the synthesized phrases. Note: A : Anger, S : Sadness, E : Elation, P : Panic.

5.2 Experiment Strategies

In the following experiments, the speech is segmented into frames by a 20-ms window progressing at a 10-ms frame rate. An energy-based voice activity detector is used to remove silence. Then, the 13-dimension MFCCs are extracted from the speech frames. The final performance is evaluated by identification rate (IR), which is computed as the percent of correctly identified sentence over all testing sentences.

Only the sentences(2-10s) of the corpus are used in the experiments. All sentences in MASC are divided into two parts. The first 5 sentences read for three times for each speaker are used for training, named training data set. And the left 5*3*15*68=15300 are for testing. Six speaker identification experiments are designed to evaluate the performance. In the first experiment, only the neutral sentences are used to train a GMM with 32 order and the scores is calculated as Equation(2). Because the emotion mismatch between the training and testing speech, the result is the lower limit of the speaker identification performance. In the second, all the sentences in training data set are used to train a GMM with 160 orders and the scoring method is the same as the first experiment. It aims to evaluate the structured training approach. The training and testing speech in the third and fourth is the same with the second experiment. The difference is that they apply the multi-condition model with 32 order GMM for each. The mean scoring method and maximum scoring method are adopted individually. These three methods are design to evaluate the performance of multi-condition model and maximum scoring method. The training and testing sentences in the fifth and sixth experiment are the same as the first one. The multi-emotion training data is generated by the frequency shifting method to evaluate its performance. Their training and testing methods are the same as the third and forth one individually. The experiment strategies are summarized in Table 1.

5.3 Results

	Experiment Strategies				Results
	Train dataset	Test dataset	Scoring Method	Frequency Shifting	IR (%)
1	Neutral	ALL	Eq.2	No	45.17
2	ALL	ALL	Eq.2	No	63.54
3	ALL	ALL	Eq.4	No	58.71
4	ALL	ALL	Eq.7	No	77.76
5	Neutral	ALL	Eq.4	Yes	47.49
6	Neutral	ALL	Eq.7	Yes	54.43

Table 1. The strategies and results of the experiments.

The details of identification rate are shown in Table 1. Compared with the first experiment, the IR of the second, third and forth one is increased 12% at least. It indicates the system performance can be improved if the training data contains all kinds of emotion states. And it is important to make the system get familiar with the emotion speech in training. The multi-condition model combining with the maximum scoring method (77.76%) exceeds the structured training method (63.54%), with an increasing of 14.22%. The multi-condition model serves to model the speech of multi-emotions. And the maximum scoring method can alleviate the negative effect of unavailable of the emotion state of testing speech. However, it is surprising that the structured training method outperforms the multi-condition model combining with mean scoring method (58.71%). The possible reason is that the length of speech with each emotion state is almost same in the experiment which is suitable for the structured training method. The use of shifting method improves the IR increased from 45.17% to 54.43%. It shows the shifting method indeed improves the performance when there is only neutral speech in training. The synthesized speech is helpful to make the system get familiar with the emotion

speech. However, it can't achieve the result of forth experiment, in which all kinds of emotion states can be obtained. In summary, our emotional speaker recognition method based on the multi-condition model is more robust, when there is only neutral speech in training and various emotion utterances in testing.

6. Discussions & Conclusions

This chapter introduces the emotional speaker recognition, and presents a frequency shifting method. The shifting method combined with MFCC is adopted for feature extraction. It converts neutral speech to emotion speech for emotional speaker recognition. The extracted emotion feature is applied with maximum scoring method for emotional speaker recognition. In our experiments, the synthesized emotion utterance by shifting frequency spectrum proves more similar to emotion speech than the neutral is. The results demonstrate the shifting method combined with maximum scoring method is effective to improve the performance of emotional speaker recognition.

All the methods we introduced above indeed outperform the baseline speaker recognition system in emotional condition. However, it can't achieve the result where neutral speech is used for both training and testing. Even the performance is still worse than the situation where all kinds of emotion states can be obtained in training. It shows that we haven't building an exact relationship between neutral and emotion utterances. More works should be carried out to find the variation rules of different emotion speeches.

7. Acknowledgements

This work is supported by the foundings: NCET-04-0545, NSFC_60525202/ 60533040, 863 Program 2006AA01Z136, PCSIRT0652, ZPNSF Y106705, National Key Technology R&D Program (No. 2006BAH02A01).

8. References

- Atal, B.S.(1969). Automatic Speaker Recognition Based on Pitch Contours, *Acoust. Soc. Am*, Vol 45, Issue 1, pp. 309-309, 1969.
- Atal, B.S.(1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *Acoust. Soc. Amer.*, vol 55, pp. 1304-1312, 1974.
- Banse, R.; Scherer, K.R.(1996). Acoustic profiles in vocal emotion expression, *Journal of Personality and Social Psychology*, vol 70, pp. 614-636, 1996.
- Bao, H.J.; Xu, M.X. & Zheng, T.F.(2007). Emotion Attribute Projection for Speaker Recognition on Emotional Speech, *InterSpeech 2007*, pp. 758-761, 2007.
- Bimbot, F.; Bonastre, J.F. & Fredouille, C.(2004). A Tutorial on Text-independent Speaker Verification, *Journal on Applied Signal Processing*, Vol.4, pp.430-451, 2004.
- Campbell, W.M.; Sturim, D.E. and Reynolds, D.A.(2006), SVM based speaker verification suing a GMM supervector kernel and NAP variability compensation, *Signal Processing Letters*, vol 13, no 5, pp.308-311, 2006.
- Hermansky, H.(1990). Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, 1990

- Herbert, G.; Michael, S.(1994). Text-Independent Speaker Identification, IEEE SIGNAL PROCESSING MAGAZINE, pp.18-32, 1994.
- Huang, T.; Yang, Y.C.(2008). Applying pitch-dependent difference detection and modification to emotional speaker recognition, InterSpeech 2008, pp. 2751-2754, 2008.
- Joseph, P.; Campbell, J.R.(1997). Speaker Recognition: A Tutorial, PROCEEDINGS OF THE IEEE, VOL. 85, NO. 9, pp.1437-1462, 1997.
- Lippmann, R. P.; Martin, E. A. and Paul, D. B.(1987). Multi-style training for robust isolated-word speech recognition," in Proc. ICASSP87, pp. 705-708, 1987.
- Louis, T.B. (2003). Emotions, speech and the ASR framework, Speech Communication, vol 40, pp.213-225, 2003.
- Ming, J.; Hazen, T. and et al.(2008) Robust Speaker Recognition in Noisy Conditions, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2008.
- Philippe, T.; Heinz, H.(1995). Usefulness of the LPC-residue in text-independent speaker verification, Speech Communication, Vol17, pp.145 - 157, 1995.
- Plutchik, R.(1980). Emotion: A Psycho evolutionary Synthesis. New York: Harper and Row, 1980.
- Reynolds, D.A.; Richard, C.(1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Transactions on Speech and Audio Processing, Vol.1, No.3, pp.72-83, 1995.
- Reynolds, D.A.; Quatieri T.F. & Dunn R.B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, Vol 10, pp.19-41, 2000.
- Reynolds, D.A.(2004). Channel robust speaker verification via feature mapping, ICASSP2004, 2004.
- Rivara, V.; Douglas, S. & Vishwa, G.(1996). Compensated Mel Frequency Cepstrum Coefficients, Proceedings of the Acoustics, Speech, and Signal Processing, pp.323-326, 1996.
- Rivara, V.; Douglas, S. & Azarshid, F.(1999). Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition, IEEE Transactions on speech and audio processing, Vol.7, No.5, pp.525-532, Sep., 1999.
- Sakoe, H.; Chiba, J.(1978). Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoustic, Speech, Signal Processing, vol. ASSP-26, no. 1, pp. 43-49, 1978.
- Sang, L.F.; Wu Z.H.; Yang Y.C.; Zhang, W.F.(2003). Automatic speaker recognition using dynamic Bayesian network, ICASPP, 2003.
- Sadaoki, F.(1997). Recent advances in speaker recognition. Pattern Recognition Letters, vol 18, pp.859-872, 1997.
- Scherer, K.R.(2000). A cross-cultural investigation of 40 E-norm emotion inferences from voice and speech: implication for speech technology, Proceedings of ICSLP, 2000.
- Scherer, K.R.; Johnstone, T. & Klasmeyer, G.(2000). Can automatic speaker verification be improved by training the algorithms on emotional speech, Proceedings of ICSLP, Vol.2, pp.807-810, Beijing, China, 2000.
- Scherer, K.R.; Johnstone, T. & Bänziger, T.(1998). Automatic verification of emotionally stressed speakers: The problem of individual differences, Proc. of SPECOM, 1998.

- Scherer, K.R.(2003). Vocal communication of emotion: A review of research paradigms, *Speech Communication*, vol 40, pp.227-256, 2003.
- Shan, Z.Y.; Yang, Y.C. & Ye R.Z.(2007). Natural-Emotion GMM Transformation Algorithm for Emotional Speaker Recognition, *InterSpeech*, pp. 782-785, 2007.
- Shan, Z.Y.; Yang, Y.C.(2008). Learning Polynomial Function Based Neutral-Emotion GMM Transformation for Speaker Recognition, *ICPR 2008*, pp.236-240, 2008.
- Shan, Z.Y.; Yang, Y.C. & Wu Z.H.(2006). SCS: A Speech Check-in System, *The 8th International Conference on Signal Processing*, pp.752-756, 2006.
- Shan, Z.Y.; Yang, Y.C. (2009). Scores Selection for Emotional Speaker Recognition, *The 3rd edition of the International Conference on Biometrics*, 2009.
- Solomonoff, A.; Campbell, W. M.; and Boardman I.(2005). Advances in channel compensation for SVM speaker recognition, in *Proceedings of ICASSP*, 2005.
- Tishby, N. Z.(1991). On the application of mixture AR hidden Markov models to text independent speaker recognition, *IEEE Transaction on Acoustic, Speech, Signal Processing*, vol. 39, no. 3, pp. 563-570, 1991.
- Wan, V.; Campbell, W.M.(2000). Support vector machines for speaker verification and identification, *Neural Networks for Signal Processing*, 2000.
- Wu, T.; Yang, Y.C., Wu, Z.H. & Li, D.D(2006). MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition, *ODYSSEY 2006*, pp.1-5, June 2006.
- Wu, W.; Zheng, T.F., Xu, M.X. & Bao, H.J.(2006). Study on Speaker Verification on Emotional Speech, *Proceedings of ICSLP 2006*, pp.2102-2105, 2006.
- Wu, W.; Zheng T.F. and Xu, M.(2006). Cohort-based speaker model synthesis for channel robust speaker recognition, *ICASSP*, 2006.
- Wu, Z.H.; Li, D.D. & Yang, Y.C.(2006). Rules Based Feature Modification for Affective Speaker Recognition, *ICASSP 2006*, pp.661-664, May 2006.

IntechOpen

Pattern Recognition

Edited by Peng-Yeng Yin

ISBN 978-953-307-014-8

Hard cover, 568 pages

Publisher InTech

Published online 01, October, 2009

Published in print edition October, 2009

For more than 40 years, pattern recognition approaches are continually improving and have been used in an increasing number of areas with great success. This book discloses recent advances and new ideas in approaches and applications for pattern recognition. The 30 chapters selected in this book cover the major topics in pattern recognition. These chapters propose state-of-the-art approaches and cutting-edge research results. I could not thank enough to the contributions of the authors. This book would not have been possible without their support.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yingchun Yang, Zhenyu Shan and Zhaohui Wu (2009). Frequency Shifting for Emotional Speaker Recognition, Pattern Recognition, Peng-Yeng Yin (Ed.), ISBN: 978-953-307-014-8, InTech, Available from:
<http://www.intechopen.com/books/pattern-recognition/frequency-shifting-for-emotional-speaker-recognition>

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2009 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen