RULES BASED FEATURE MODIFICATION FOR AFFECTIVE SPEAKER RECOGNITION[1]

Zhaohui Wu, Dongdong Li and Yingchun Yang

College of Computer Science and Technology,
Zhejiang University, Hangzhou, P.R.China, 310027

ABSTRACT

One of the largest challenges in speaker recognition applications is dealing with speaker-emotion variability. In this paper, we further investigate the rules based feature modification for robust speaker recognition with emotional speech. Specifically, we learn the rules of prosodic features modification from a small amount of the content matched source-target pairs. Features with emotion information are adapted from the prevalent neutral features by applying the modification rules. The converted features are trained together with the neutral features to build the speaker models. The effects of individual and combined modifications of duration, pitch and amplitude are also studied using EPST dataset recorded by 8 professional actors with 14 kinds of emotion expressiveness. It demonstrates that duration modifications play the most important role; and that, pitch modifications are more effective than amplitude modifications. Promising result with an improved identification rate by 7.83% is achieved compared to the traditional speaker recognition.

1. INTRODUCTION

Different emotional states affect the speech production mechanism of a speaker in different ways, and lead to acoustical changes in his/her speech [1]. These changes are a major cause of errors in speaker recognition application.

One of the most famous compensation techniques for emotion influences attempts to elicit different manners of speaking during the enrollment. This structured training approach aims at making the system become familiar with the variation likely to be encountered in that person's voice [2]. The registered users are asked to provide specially reading (emotional) speech which would lead to the unfriendliness of the system.

Analysis of emotional speech and its synthesis rules have been researched for many years [3]. What' more, speaker information is remained when emotion state changes, as shown in our previous study [4]. Based on these researches, in this paper we investigate the applicability of feature modifications of duration, pitch and amplitude parameters for the robustness of speaker recognition over affective speech. The features of the input speech are adapted to the target emotion with the modification rules derived from the same content source-target pairs and then output the new features. The speaker models are trained with both the neutral speech provided by the users and generated output speech perceived as conveying emotions. We also study the effect of each parameter on transforming the emotional information in speech utterances for affective speaker authentication.

This paper is organized as follows. In the next section we give an introduction of the emotional corpus. Section 3 is committed to presenting the system architecture. In the following section the process steps of feature extraction, statistic and modification are described in detail. The different experiments comparison and result discussion are presented in Section 5. We close with a conclusion section.

2. DATASET

For experiments, the Emotional Prosody Speech and Transcripts (EPST) corpus with the Linguistic Data Consortium (LDC) catalog number LDC2002S28 was used [5]. This database is the only one containing emotional speech provided by LDC up to now. The main objective of the corpus is to support research in emotional prosody. Speech narrated by 8 professional actors (3 male and 5 female) is sampled at 22.05 kHz with 2-channel interleaved 16-bit PCM format.

The corpus is divided into two parts, Distance Continuum and Emotion Continuum. The first one expresses a dimension of dominance and a dimension of distance. The other part contains a series of semantically neutral utterances (dates and numbers) spanning fourteen distinct emotional categories, chosen based on Banse & Scherer's selection criteria [6]. The fourteen types of expression are: 'Hot Anger', 'Sadness', 'Panic', 'Elation',

[1] This work is supported by National Natural Science Foundation of P.R.China (60273059), National Science Fund for Distinguished Young Scholars 60525202, Program for New Century Excellent Talents in University NCET-04-0545 and Key Program of Natural Science Foundation of China 60533040.

'Shame', 'Pride', 'Boredom', 'Disgust', 'Cold Anger', 'Anxiety', 'Despair', 'Neutral', 'Interest', 'Contempt'. For each speaker, the database contains 5-minute emotional speech and 0.5-to-1-minute neutral speech.

3. SYSTEM DESCRIPTION

Typically a speaker will likely enroll his/her voice with neutral speech and then wish to be verified with discretionary affective speech up to his/her mood at that time.

The system learns the conversion rules from a small speech set with the same content pairs of standard reading (neutral) speech and emotional speech. The conversion rules for feature modifications are speaker-independent but associated with corresponding target emotions. The neutral speech provided by registered users is divided into two parts. The minority of neutral speech is used to generate the target emotional speech of all types with the conversion rules, e.g. one neutral utterance to 14 types of emotional utterances in our case. The speaker model is trained using an aggregation of data with all of the converted affective speech and the rest majority of neutral speech.

The test speech has no restriction on the emotion types, which makes the system more friendly and intelligent in the interaction with users. The utterances could be either neutral ones or affective ones. The test speech is matched with all of the speaker models. The final decision is made as the highest scoring decision procedure applied to the models matcher module outputs.

4. FEATURE MODIFICATION

In this section we describe the modification rules in detail. Figure 1 shows the flow of feature modification. The rules are based on the statistics of features variation between content matched pairs of neutral and emotional speech. Duration, pitch, and amplitude are selected because of their sensitivity to the change of emotion states. In each emotional state, the rules are applied to each speaker with the same parameters. An averaging process is adopted.

4.1. Feature Extraction

Linear Predictive Coding (LPC) analysis is applied to process the speech. LPC coefficients and residual signal are extracted. The LPC coefficients represent the phoneme character and are closely related to the attributes of the voice quality like tenseness, creakiness, laxness, breathiness which could reflect the emotional content of speech [7]. Residual signal carries significant speaker specific information and shows the correlation with a subjective evaluation of voice properties. Both of them are used to build the speaker models for their importance in charactering speakers.

The result of LPC analysis is a new representation of the signal.

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \qquad (1)$$

Where $s(n)$ is the original speech signal. a_i and $u(n)$ are the outputs of the LPC analysis with a_i representing the LPC coefficients. The $u(n)$ term represents the normalized excitation source, or the residual. The G factor is a gain term.

To residual signal, we use a reduction method, and apply LPC analysis to get the main information.

$$u(n) = \sum_{i=1}^{p} a_i' u(n-i) + G' u'(n) \qquad (2)$$

The concatenation of LPC coefficients a_i and a_i' derived from speech signal and residual signal are used to train speaker models.

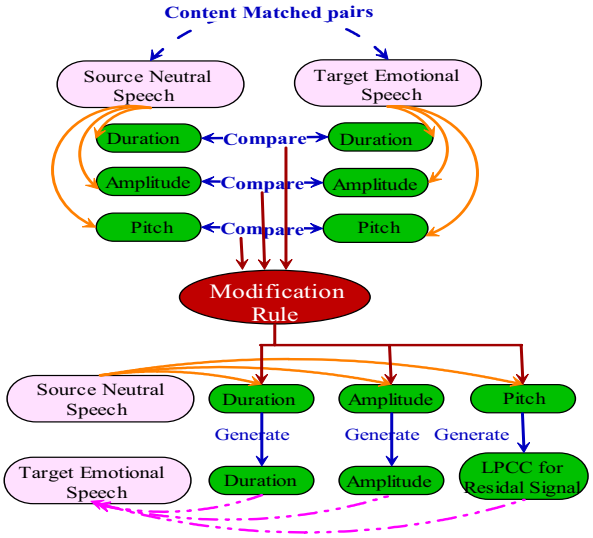

Figure1. Feature Modifications

4.2. Statistics of Feature Variation

For each feature X , X_e is the emotional value, as given by the acoustic analysis of the emotional speech, and X_n is the neutral one. The same content is shared by the emotional speech and the neutral speech.

Duration is determined by the frame number. The change of Duration can be calculated as:

$$\Delta D = D_n / (D_e - D_n) \qquad (3)$$

Here D stands for frame number of each utterance.

Pitch is determined by Subharmonic-TO-Harmonic Ratio (SHR) method provided by Sun [8].The change of average pitch (AP) can be calculated as:

$$\Delta AP = \frac{\sum_{i=1}^{D_e} P_{ie}}{D_e} - \frac{\sum_{i=1}^{D_n} P_{in}}{D_n} \quad (4)$$

Where P_i is the pitch value of frame i, and D is the frame number of the utterance.

The change of pitch range (PR) can be calculated as:

$$\Delta PR = (P_{\max e} - P_{\min e}) / (P_{\max n} - P_{\min n}) \quad (5)$$

Where $P_{\max}$ and $P_{\min}$ are the maximal and minimal value in a set of pitch an utterance.

Amplitude is referred to the short time mean amplitude. The change of average amplitude (AA) and amplitude range (AR) can be calculated in the same way as Equation (4) and (5) with replacing corresponding pitch parameters with amplitude. More details could refer to [4].

4.3. Feature modifications

Amplitude is modified by average amplitude and amplitude range, and can be calculated as:

$$A'_e = (A'_n + \Delta AA) * \Delta AR \quad (6)$$

Where A' is the value of amplitude.

For the duration modification, the frames are reconstructed. Excessive frames are cut to reduce its duration; and appropriate frames are repeated to expand its duration.

The whole residue is often summarized in one number representing F0. The change of average pitch and pitch range are used to direct the modification of LPCC of residual signal:

$$R'_e = (R'_n + \Delta AP) * \Delta PR \quad (6)$$

Here R' stands for the LPCC of residual signal.

5. EXPERIMENT

5.1. Experiment Strategies

Three experiments are designed in this work to evaluate the performance of the proposed approach:

- ***Baseline***: In the first set of experiments, we conducted a baseline strategy. The user's own speech is used to build the speaker models. Two systems are evaluated here: traditional GMM system trained with neutral features (Model Type: A1) and GMM system trained with both neutral features and emotional features (Model Type: A2).
- ***Trained with Converted Emotional features***: In this series of experiments, the effects of individual and combined modifications of duration, pitch and amplitude on the performance of the system is studied. The state of the affective speech is assumed to be classified before feature extraction. The utterance is matched with GMM models trained with the corresponding type of converted emotional features (Model Type: B) either single parameter modification or combination parameters modification.
- ***Trained with Neutral and Converted Emotional Features***: The performance of automatic emotion recognition is still far from satisfactory, which makes Model Type B trained with the converted emotional feature unpractical. Compared with the above sets of experiments, speaker models are built with both neutral and converted emotional features (Model Type: C). This strategy makes the advance emotion recognition become inessential.

In each strategy, speech is cut in overlapping frames of 30ms duration stepped each 20ms. After pre-emphasis with $u = 0.9378$, each frame is fed to LPC with p=14 as analysis order. The 14-dimensional LPCC of residual signal are then appended to the 14-dimensional LPCC of speech signal directly resulting in a 28-dimensional feature vector. AuditoryToolbox [9] is used as the interface in our source code.

5.2. Results and Discussions

The details of identification rate (IR) for the fourteen kinds of affective speech tested independently with traditional GMM system are shown in the Figure 2. The traditional system yields the average IR of 64.19%. While the neutral state of speech gives the highest one of 78.75% among all the affective states. This technique demonstrates that the performance of speaker recognition system drops sharply on emotional speech and the consistence speech state of the training and test utterances is important.

When neutral features and emotional features trained together, we selected both different numbers of utterances (1, 3, 5 seconds) and different numbers of emotion states (first 4, 8 and 14 types). Figure 3 shows the accuracy with diverse features for training speaker models. As expected, the accuracy of the system increases with the number of utterances and the emotion states.

Results of strategy with Model Type B are shown in Figure 2. The results indicate following trend, in terms of producing successful speaker recognition with converted speech, among proposed features modification methods: Duration > pitch> amplitude, with the average IR improvement of 3.43%, 1.87% and 0.84% respectively. When combined together, a profit 7.23% is gained compared to the traditional models.

The detail comparison results for the fourteen kinds of affective speech between Model Type C and Model Type A2 are shown in Figure 4. And the average performance is list in Table 1. As we can see, the accuracy increased with the number of utterances. When 1 second of 14 types emotional utterances added, the proposed model with neutral and converted speech shows the comparative identification rate with the model built on additional original

emotional speech. The profit in performance increases by 7.83% for all emotional testing utterances with 5 seconds utterances added compared with the baseline technique. However the enhancement of the proposed model is not as distinct as in the case of Model Type A2 when utterances increase. The enhancement for different emotional type is also different. The main reason lies on the diverse expression of emotion even in one emotion state. The original emotional utterances fill the speaker models with more emotion information. The converted speech could only express one rule for one emotion state. What's more, the converted rules are based on the content matched speech and couldn't exert all strength to the text-independent speech. Further experiments should be paid attention to the improvement of feature modification.

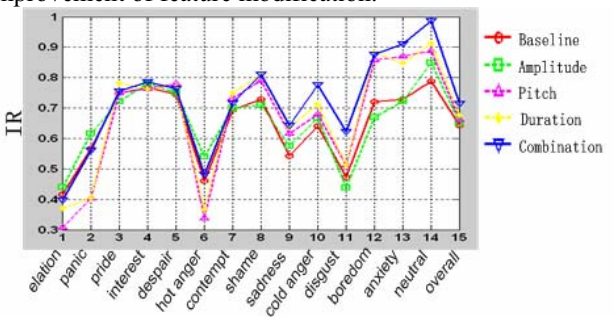

Figure 2. The performance of traditional speaker system & the effects of individual and combined modifications of duration, pitch and amplitude on speaker recognition

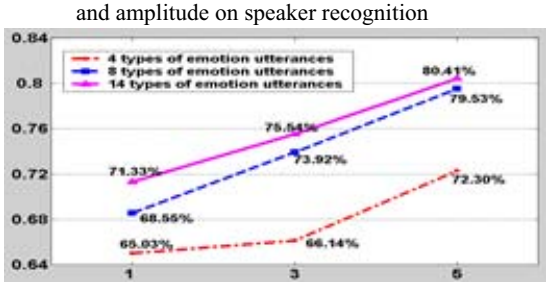

Figure 3. The accuracy obtained with diverse speaker models trained by 1, 3, or 5 seconds of utterances and different numbers of emotion states. The 4 types of emotion refer to hot anger, elation, sadness, panic. The 8 types of emotion include boredom, shame, pride, disgust, together with the four types above.

6. CONCLUSION

In this paper we studied the robust speaker recognition system with affective speech by training speaker models with neutral speech and converted emotional speech derived by modifying prosody features of neutral speech. Our results show that the combination of prosody feature modifications successfully adds new emotional coloring to the neutral speech. The proposed speaker models cover the variation of voice when emotion state changes. We also investigate the effect of different features on affective speaker recognition. Duration modifications are the most effective.

Further experiments on a larger database (68 people) will focus on robustness with respect to the universality of the approach.

Table 1. The identification rate with GMM models build on both neutral and converted emotional features (Model Type C)and GMM models build on both neutral and original emotional features (Model Type A2).

IR(%)	1s'	3s'	5s'
Model Type A2	71.33	75.54	80.40
Model Type C	70.45	71.42	72.02

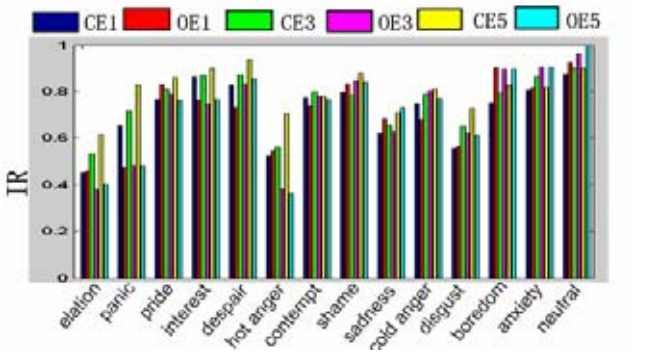

Figure 4. The detail IR of 14 type emotional speech. Model Type C with 1 second (CE1); Model Type A2 with 1 second (OE1); Model Type C with 3 second (CE3); Model Type A2 with 3 second (OE3); Model Type C with 5 second (CE5); Model Type A2 with 5 second (OE5).

7. REFERENCES

[1]I. R. Murray and J. L. Arnott, "Synthesizing emotions in speech: is it time to get excited?" Proceedings of ICSLP '96 pp.1816-1819. Invited Paper, Philadelphia, PA, USA, 3-6 October. 1996

[2]K. R. Scherer, T. Johnstone, G. Klasmeyer, "Can automatic speaker verification be improved by training the algorithms on emotional speech?" Proceedings of ICSLP'2000 vol. 2, pp807-810. Beijing, China.

[3]K.R. Scherer, "Vocal communication of emotion: A review of research paradigms", Speech Communication, vol. 40 (1-2), pg. 227-256, 2003.

[4]Dongdong Li, Yingchun Yang et al, "Emotion-State Conversion for Speaker Recognition " Proceedings of ACII 2005, LNCS 3784, pp. 403 – 410, 2005.

[5]http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

[6] R. Banse, K.R Scherer, "Acoustic Profiles in vocal emotion expression" Journal of Personality and Social Psychology. 1996 Mar; 70(3):614-36

[7] C. Gobl and A. N Chasaide, "The role of voice quality in communicating emotion, mood and attitude", Speech Communication, vol. 40. pp. 189-212, 2003.

[8] S.Sun,: Pitch Determination and Voice Quality Analysis Using Subharmonic-TO-Harmonic Ratio. IEEE International Conference on Acoustics, Speech, and Signal processing (2002)

[9]http://www.speech.cs.cmu.edu/comp.speech/Section1/HumanAudio/auditory.tlbx.html