GSJ: Volume 8,Issue 3, March 2020, Online: ISSN 2320-9186

www.globalscientificjournal.com

Speaker Recognition with Emotional Speech

Ahmad Faraz Hussain, He Qianhua
School of Electronics and Information Engineering, SCUT, China

***Abstract*— In recent researches, emotional speaker recognition has emanated as an important challenging topic. Despite the fact that speaker recognition research has been ongoing for extra than four decades, the speaker recognition performance is effected by background noise, age, person health and emotional state of a speaker. I-vector is used in this study because it has been proved to be very efficient for its fixed length and low dimensions. Features are extracted using three techniques (MFCC, MFCC SDC, MFCC SDC + PNCC) and for channel/session compensation, Linear Discriminant Analysis (LDA) and Probabilistic Linear Discriminant Analysis (PLDA) are used. In the experiments, "CREMA-D" (Crowd-sourced Emotional Multimodal Actors Dataset) is used. Satisfying results are achieved using six different emotions.**

***Index Terms*— Speaker recognition, MFCC SDC, I-Vector, PLDA**

I. Introduction

THIS The importance of emotional speaker recognition is growing in many fields in Human-Computer-Interaction (HCI) [1]. The main goal is to make the computer able to know a person in real-life conditions. Many real applications can used emotional speaker recognition in noisy environments such as forensic or criminal investigation to identify the accused person who produces emotional utterances in different environments like subway, restaurant, airport, busy street etc.

Emotion is an inherent nature of human and remarkably change the speech forms [2]. Intra-speaker variation occurs in speech signal due to speaking rate and speech tones but emotion causes significant changes in speech properties like temperal-spectral patterns, formant structures and harmonic forms. GMM-UBM architecture for speech verification [3], the neutral speech used in training and other emotions can hardly represent the test utterances, thus leading severe performance degradation in verification.

I. Shahin [4] reported text-independent speaker verification on emotional dataset, including five emotions: sad, happy, angry, disgust and fear. HMMs and the cepstral mean subtraction technique used in training and testing sessions, and got better performance compared to system that is only based on HMMs.

The speaker identification is performed in emotional environment on text-independent dataset [5]. MFCC is used to extract the spectral features, while GMM is used for training and testing of the system. The performance is checked on Berlin emotional speech database that contains five emotions and neutral. The results show that the emotional state effect the speaker identification. The accuracy rate is about 60% i.e. total fail in real applications. Most difficult situations are for angry and happy; for both the accuracy rate is between 16% and 36%.

Emotional variability in speech degrades the speaker recognition performance [6]. Training is done with neutral and expressive speech is used for testing. Mismatching causes error that lead to the speaker recognition degradation. Could emotional regions be defined in which the speaker recognition performance is reliable? So, they predict the reliable regions for speaker recognition by analyzing and predicting the emotional content. The emotional database consists of 80 speakers. They evaluate speaker recognition performance as a function of arousal and valence, forming regions where they can reliably recognize a speaker. The experimental results show that the sentences classified as reliable for speaker recognition tasks have lower equal error rate (EER) as compared to sentences that are classified as unreliable.

The traditional LDA finds the transformation that minimizes the ratio of the within to between class scatters. LDA assumes speaker classes have a Gaussian distribution and share the same covariance matrix. Many variations of discriminant analysis have been proposed to partly relax the LDA assumptions. Kernel discriminant analysis or generalized discriminant analysis (GDA) [7, 8] finds a non-linear transformation, heterocedastic LDA (HLDA) [9] employs different covariance matrices for different classes, mixture discriminant analysis (MDA) [10] assumes the distribution of each class is a mixture of Gaussians.

In this work, speakers are recognized from emotional speech signals that is distorted by real subway noise. The performance of speaker recognition system is highly affected by noise. So, in the field of speaker recognition, this is a challenging topic that is not yet been studied in emotional context and the implementation of emotional speaker recognition in more practical setting. MFCC is used for feature extraction, i-vector is used for classification and some compensation techniques like LDA and PLDA are used in clean and noisy environments.

The organization of this paper is as follow: Section II illustrates the proposed emotional speaker recognition system. Cepstral mean and variance normalization are described in section III. I-vector is described in section IV. PLDA is described in section V. Experimental setup is described in section VI and finally Section VII gives the conclusion.

II. SYSTEM FRAMEWORK

The framework of proposed emotional speaker recognition system is instantiated in figure 1. CREMA-D database that consist of six emotions is used for our experiments. Three techniques are used to extract the most relevant information from the emotional speech signals to represent them in feature vectors. Gaussian Mixture Models (GMM) is used as a tool for i-vector. I-vector is very popular in speaker recognition field. A low dimensional of 400 dimensions is used. For channel compensation, PLDA techniques are used.

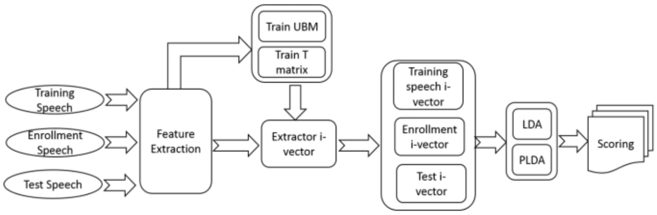

Fig. 1. Emotional speaker recognition system

III. FEATURE EXTRACTION

A. MFCC

MFCCs is used for feature extraction and the steps are shown in fig. 2.

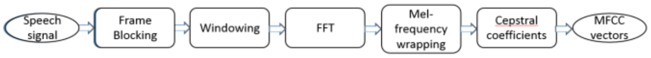

Fig. 2. MFCC features extraction

The aim of robust feature extraction is features that are less distorted by noise vitiate. Cepstral mean and variance normalization (CMVN) is a good noise normalization technique for speaker recognition. Due to insufficient data for parameter estimation and loss of discriminable information, the performance of CMVN is known to degrade for short utterances as all utterances are forced to have zero mean and unit variances. Instead of maximum likelihood projections, we suggest using posterior estimates of mean and variance in CMVN. In addition to providing a reliable estimation of parameters, this Bayesian method also shows that discriminable information is retained without an increase in computational.

B. MFCC SDC

Shifted Delta Cepstra (SDC) coefficients consists of four parameters called N, d, P and k. For each data frame, first MFCCs are calculated based on N; (i.e $c_0,c_1,c_2,c_3….c_{N-1}$). Parameter d specifies the spread over which deltas are calculated. Parameter P establishes the distances between successive delta calculations and K specifies number of blocks. The SDC coefficients are extracted as shown in fig 3.
Therefore, SDC coefficients revealed in 4 are the stacked version of MFCC coefficients given in 1, and k×N parameters are used for each SDC feature vector.
For a given time t, we get

$$\Delta c(t,i) = c(t,iP+d) - c(t+iP-d) \qquad (1)$$

The stacked version of SDC coefficients are given by

$$SDC(t) = [\Delta_c(t,0)^t \ldots\ldots \Delta_c(t,k-1)^t]^t \qquad (2)$$

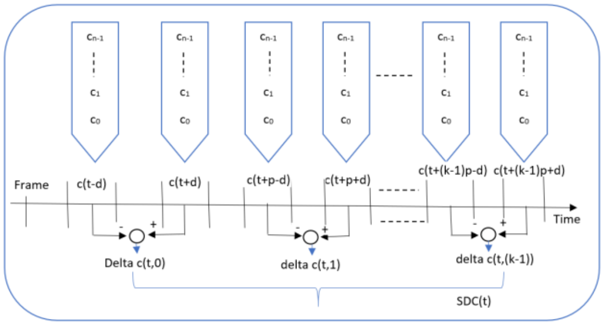

Fig. 3. SDC coefficients extraction

C. MFCC SDC+PNCC

The front-end technique Power Normalized Cepstral Coefficients, similar to MFCCs but Mel -scale is replaced by Gammatone filters. Bias vectors provide robustness. Combination with SDC, it provides high recognition in different emotional environment. As we used a dataset that consists of six different emotions, so this feature extraction technique performs very well.

IV. I-VECTOR MODELING

I-vector means identity vector which is firstly introduced by Dehak et al. [11]. An i-vector is represented by a low dimensional space having fixed length known as Total Variability Space (TVS). The i-vector algorithm is derived from the traditional JFA (Joint Factor Analysis). JFA consists of two distinct subspace called channel variability and speaker variability. During calculation, the channel dependent components are rejected and speaker dependent super vector is calculated. Dehak proved that channel dependent components contain some useful information that can be used distinguish speakers.

According to Dehak, the GMM super-vector M can be decomposed as follow:

$$M = m + T\omega \qquad (3)$$

where m is the speaker and channel independent super-vector which are estimated using the Universal Background Model (UBM). $\omega N(0;1)$ is the i-vector with standard prior distribution. It represents speaker coordinates in reduced total variability space. T represents low rank rectangular total variability matrix.

Let d is the acoustic feature dimensionality and uv is the mixture number in GMM. So, the size of Gaussian mean super-vector M is (uv*d). The Expectation Maximization (EM) algorithm is used to estimate the total variability matrix T [12].

In reality the feature vector of each input utterance is an estimation of ω using Maximum a Posteriori (MAP) adaptation. The new feature vector is given by:

$$\omega = B^{-1}T^{-1}\Sigma^{-1}F_c \qquad (4)$$

where

$$B = I + T^t\Sigma NT \qquad (5)$$

where Σ is the block diagonal covariance matrix obtained by the covariance matrix of UBM and T is the total subspace matrix. F_c and N_c in eq. 4 and eq. 5 represent the zero and first order Baum-Welch statistics. UBM is used to calculate these sufficient statistics [13].

Given c is the UBM component, L is sequence of feature frames, then Baum-Welch statistics of utterance can be calculated as follow:

$$N_c = \sum_{t=1}^{L} \gamma_t(c) \qquad (0^{th}\ order\ statistics) \qquad (6)$$

$$F_c = \sum_{t=1}^{L} \gamma_t(c) Y_t \qquad (1^{th}\ order\ statistics) \qquad (7)$$

where, Y is the feature vector at frame t and $\gamma_t(c)$ is the posterior probability of Gaussian component c for frame t. Removing mean these statistics are then centered. The dimension of a single N_c is 1 for an UBM component c; F_c has (d*1) dimensions for each c where d is the dimension of the feature vector [14]. Finally, in the recognition phase given two i-vectors $\widehat{\omega}_1$ and $\widehat{\omega}_2$ we need to confirm that these i-vectors are produced by the target or non-target, that can by identified by the following log-likelihood ratio,

$$log - likelihood = log \frac{p(\widehat{\omega}_1, \widehat{\omega}_2 | target)}{p(\widehat{\omega}_1, \widehat{\omega}_2 | non - target)} \qquad (8)$$

V. LINEAR DISCRIMINANT ANALYSIS (LDA) & PLDA

For dimensionality reduction in pattern recognition problems and classification, LDA is widely used [15]. When each class has a Gaussian distribution with a common covariance matrix, it finds the exact optimal linear transformation. The traditional LDA is given by the following eq.

$$\lambda = \frac{A^T S_b A}{A^T S_w A} \qquad (9)$$

where S_w and S_b indicates within and between class covariance matrices. The projection matrix A that contains the k eigenvectors corresponding to the k largest eigenvalues of $S_w^{-1} S_b$ is the solution for LDA optimization problem. For feature vectors x, the within and between class scatters are calculated by,

$$S_w = \textstyle\sum_{c=1}^{C} n_c (\mu_c - \mu)(\mu_c - \mu)^T \qquad (10)$$

$$S_w = \textstyle\sum_{c=1}^{C} \sum_{k \epsilon c} (x_k - \mu_c)(x_k - \mu_c)^T \qquad (11)$$

Where, C is the total number of speaker classes, n_c is the number of samples in the class , μ is the total mean of all samples, μ_c is the mean of samples in class c. The PLDA representation for an i-vector ω corresponding to a speaker utterance is given by;

$$S_w = \textstyle\sum_{c=1}^{C} \sum_{k \epsilon c} (x_k - \mu_c)(x_k - \mu_c)^T \qquad (12)$$

Where, H and G are the matrices representing the speaker and channel subspaces respectively, ρ is the global mean of i-vector population, g and h are the channel and speaker factors, having standard normal prior distribution, and ϵ is the residual factor having standard normal prior with diagonal covariance. For handling the effect of outliers in the data and the model, heavy-tailed priors are assumed for the latent variables is called as heavy –tailed PLDA (HPLDA) [16]. A simplified version of PLDA has been projected [61] that ignore the term Gg and appoint standard normal prior to the latent variables. The non-diagonal covariance denoted by S and the residual term is modeled with Gaussian distribution with zero mean. This approach is known as Gaussian PLDA (GPLDA) and its performance is similar to HPLDA with very less complexity.

VI. EXPERIMENTAL SETUP

The target of this paper is to design an emotional speaker recognition system based on i-vector that perform on emotional speech. The impact of i-vector on the emotional speaker recognition rate showed and for channel compensation PLDA is used. MFCC, MFCC SDC and MFCC SDC + PNCC are used for features extraction.

A. *Emotional Corpus*

The text-dependent dataset that is used in our experiments is “CREMA-D” (Crowd-sourced Emotional Multimodal Actors Dataset) [17]. Out of 91, 48 are male and 43 are female. Actors were between the ages 20-74 years coming from a variety of races and ethnicities (Asian, African, American, Caucasian, Hispanic and Unspecified). Actors spoke 12 sentences in six different emotions (Happy, Disgust, Fear, Anger, Neutral and Sad). Total of 6,552 utterances should be produced if 12 sentences were spoken by 91 actors. Some actors didn’t have 72 utterances. The duration of each utterance is 2 seconds to 3 seconds. All audio files are sampled with 16bit resolution at a rate of 16 KHz.

B. *Experimental Setting*

Kaldi toolkits D. Povey, 2011[18] used for performing experiments. MFCC features extraction (20 ms hamming window, every 10ms), 19 Mel-frequency cepstral coefficient together with log energy were used. Delta and delta-delta coefficient were evaluated to generate 60-dimensional feature vector.

256 Gaussian Mixtures, 400-dimensional i-vector and 150-dimensional LDA/PLDA.GPU: GTX 1080T used for our experiments. As mentioned earlier, a total of 6,539 utterances from 91 speakers. 42 utterances from each speaker i.e. total of 3,816 utterances (8 utterances are missing in the dataset) were put in training to train the GMM-UBM. 24 utterances from each speaker i.e. total of 2,177 utterance (5 utterances are missing in the dataset) in testing to get the speaker models and 6 utterances from each speaker i.e. 546 utterances used in development. 3 utterances are randomly selected for each speaker from 2,177 utterances as enroll samples and others are used as eval samples for evaluation. It means that in testing there are 24 utterance of each speakers, so 3 utterances are used for enrollment and 21 utterances are used for evaluation. A development set was used to optimize our model against during the development process. It is used to tuned the parameters of our training algorithm and prevent overfitting. Test is used to evaluate the performance on unseen data but it is not used for tuning. In this experiment, GMM: 256 and channel compensation techniques (LDA and PLDA) are used. Table 1 and 2 shows the results.

Table 1: The EERs (%) of Emotional Speaker Verification using GMM: 256 & LDA

EMOTION	ANG	DIS	FEA	HAP	NEU	SAD
EER (%)	5.304	4.751	5.635	4.309	3.862	5.083

Table 2: The EERs (%) of Emotional Speaker Verification using GMM: 256 & PLDA

EMOTION	ANG	DIS	FEA	HAP	NEU	SAD
EER (%)	2.857	2.418	3.58	2.198	1.758	3.297

Now, we used MFCC SDC and MFCC SCD+PNCC as features extraction techniques. They have better performance as compared to MFCC. The results are shown in Table 3 and 4.

Table 3: The EERs (%) of Emotional Speaker Verification using GMM: 256 & PLDA & MFCC SDC

EMOTION	ANG	DIS	FEA	HAP	NEU	SAD
EER (%)	1.557	1.448	2.38	1.157	0.848	2.47

Table 4: The EERs (%) of Emotional Speaker Verification using GMM: 256 & PLDA & MFCC SDC+PNCC

EMOTION	ANG	DIS	FEA	HAP	NEU	SAD
EER (%)	0.963	0.589	1.198	0.543	0.334	1.350

The main idea of this work was the anticipation of the speaker verification enhancement when using emotional speech. The first two experiments were performed using MFCC features extraction method. Using MFCC SDC features extraction technique in 3rd experiment showed better results. In the last experiment the combination of MFCC SDC and PNCC feature extraction technique showed best results as shown in fig 4. Fear and Sad emotions were not showing good results because the utterances for these emotions were spoken very slowly, so the system has difficulty in recognizing speakers of these emotions.

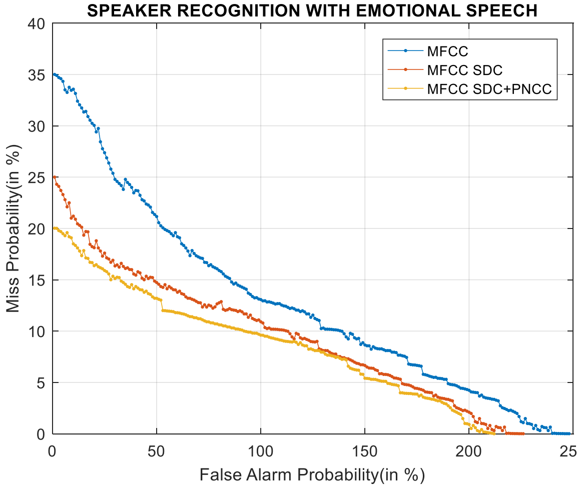

Fig. 4. DET curves of Emotional Speaker Recognition using different feature extraction techniques.

VII. Conclusion

In this study an emotional speaker verification system presented in which the factor analysis is done by low-dimensional space which consists of both speaker and channel variabilities. In modelling, each recording is represented by low-dimensional vector called i-vector extracted by a simple factor analysis. The classical use of joint factor analysis addresses the channel effect in the high-dimensional GMM mean supervector space. But in i-vector approach, addressing the channel effect in low-dimensional i-vector space, so less computation is required as compared to classical joint factor analysis method. To compensate the intersession problem, two different techniques like linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (PLDA) were used. CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) emotional database is used for our experiments. MFCC, MFCC SDC and MFCC SDC + PNCC are used for features extraction. MFCC SDC+PNCC showed best performance. The emotions like sad and fear having low voice pronunciation are not showing good results as compared to other emotion because these emotions were spoken very slowly.

Deep Neural Network (DNN), will be used as a classifier in future work to prove the robustness of emotional speaker recognition in noisy environment. We will also propose some new features extraction methods rather than using the methods used in this work.

Financial and Ethical disclosures

Funding: This researched has been completed without any funding.

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

[1] M.V. Ghiurcau, C. Rusu and J. Astola, A study of the effect of emotional state upon text-independent speaker identification, IEEE Int.Conf. on Acoustics, Speech and Signal Processing, (ICASSP),pp.4944-4947, 2011.

[2] F. Bie, D. Wang, T. F. Zheng and R. Chen, "Emotional speaker verification with linear adaptation," 2013 IEEE China Summit and International Conference on Signal and Information Processing, Beijing, 2013, pp. 91-94.

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[4] I. Shahin, "Speaker Recognition Systems in the Emotional Environment," 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, 2008, pp. 1-5.

[5] M. V. Ghiurcau, C. Rusu and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011, pp. 4944-4947.

[6] S. Parthasarathy and C. Busso, "Predicting speaker recognition reliability by considering emotional content," 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, 2017, pp. 434-439.

[7] B. Scholkopft and K.R. Mullert, “Fisher discriminant analysis with kernels,” Neural Networks for Signal Processing IX, vol. 1, no. 1, pp. 1, 1999.

[8] G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” Neural Computation, vol. 12, no. 10, pp. 2385–2404, 2000.

[9] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition,” Speech Communication, vol. 26, no. 4, pp. 283–297, 1998.

[10] T. Hastie and R. Tibshirani, “Discriminant analysis by gaussian mixtures,” Journal of the Royal Statistical Society.Series B (Methodological), pp. 155–176, 1996.

[11] P. Kenny N. Dehak, R. Dehak. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification,. Interspeech, 2009.

[12] Boulianne G Kenny, P. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, 2005.

[13] Reynolds, Douglas A., Quatieri, Thomas F., and Dunn, Robert B., Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10(2000), 19–41.

[14] Mansour, Asma & Chenchah, Farah & Lachiri, Zied. (2016). Emotional speaker recognition based on i-vector space model. 1-6. 10.1109/CEIT.2016.7929127.

[15] Bahmaninezhad, Fahimeh & Hansen, John. (2017). i-Vector/PLDA speaker recognition using support vectors with discriminant analysis. 10.1109/ICASSP.2017.7953190.

[16] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in In Proc. Odyssey: The Speaker and Language Recognition Workshop, Jun 2010.

[17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," in IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377-390, 1 Oct.-Dec. 2014.

[18] Povey_Idiap-RR-04-2012: The Kaldi Speech Recognition Toolkit, Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukas, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlicek, Petr, Qian, Yanmin, Schwarz, Petr, Silovsky, Jan, Stemmer, Georg and Vesely, Karel, Idiap-

Ahmad Faraz Hussain

received the B.Eng. degree in Electrical(communication) Engineering from the University of Engineering and Technology, Peshawar, Pakistan. He received M.S. degree in Information and Communication Engineering from South China University of Technology, Guangzhou. His research interests include marine object detection and underwater communications.

He Qianhua

received the B.S. degree in physics from Hunan Normal University in 1987, the M.S. degree in medical instrument engineering from Xi’an Jiaotong University in 1990, and the Ph.D. degree in electronic engineering from South China University of Technology in 1993. Since 1993, he has been with the School of Electronic and Information Engineering (SEIE), South China University of Technology (SCUT). From 1994 to 2001, he was a Researcher with the Department of Computer Science, City University of Hong Kong. From 2007 to 2008, he was a Visitor with University of Washington, Seattle. He is currently a Professor with the SEIE, SCUT. He is interested in audio forensics.