Automatic Speaker Identification Using Clinically Depressed Speech Content

SHEERAZ MEMON*, FAISAL KARIM SHAIKH**, AND JAVED ALI BALOCH*

RECEIVED ON 26.10.2011 ACCEPTED ON 01.12.2011

ABSTRACT

The environment affects largely the performance of automatic speaker recognition. This work investigates the effects of clinical environment on the task of speaker recognition. For this task we have used two sets of speakers, a clinical set which consists of speech samples from 70 clinically depressed speakers and a control set which comprises of 68 clinically non-depressed speakers. The MFCCs (Mel Frequency Cepstral Coefficients) are applied for feature extraction, and a number of modeling methods such as GMM-EM (Gaussian Mixture Models Based on Expectation Maximization), GMM based on Kmeans (GMM-Kmeans), GMM-LBG based on Linde Buzo Gray, and GMM -ITVQ based on Information Theoretic Vector Quantization are used. The different modeling methods are evaluated for the novel speech corpus. The results suggest that the speaker recognition rates for the depressed speakers are lower (60-71%) than for the non-depressed speakers (79-89%). This paper further investigate the performance of VQ (Vector Quantization) based Gaussian modeling, and proposes a novel approach called GMM-ITVQ. The results suggest that GMM-EM has the higher recognition rates however, the performance of GMM-ITVQ is comparable to GMM-EM.

Key Words: **Speaker Recognition, Depression, Clinical Environment, Gaussian Mixture Model.**

1. INTRODUCTION

The task of identifying an individual based on his or her voice/speech samples is called speaker recognition. The record environment for train/test speech could be widely different, so validating the feature extraction and modeling methods for different record environments could lead to better results. This paper proposes a system which differentiates the speakers based on the behavioral contents such as a depressed or a non-depressed speaker. The research contribution given in this paper will also improve the understanding of diverse psychological and physiological states available in the acoustic voice/speech signal. In this paper task of automatic speaker identification is performed for two sets of speakers, the first set comprises the speakers which are labeled by the psychologists as depressed and the second set contain a number of speakers which are analyzed as non-depressed speakers. The speech corpus is obtained from OREGON research institute for Psychologists, USA [1].

The speaker recognition systems which are currently in use are limited when it comes to IntraSpeaker variability. This variability degrades the recognition rates in most of the situations and thus limits the commercial use of

* Assistant Professor, Department of Computer Systems Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Assistant Professor, Department of Telecommunicatoin Engineering, Mehran University of Engineering & Technology, Jamshoro.

Automatic Speaker Recognition. Ongoing research suggests that a number of variations available within speaker can be mapped out in psychological and physiological state of a speaker class [2-5]. A database addressing such mechanisms of speaker state influences of speech can direct to improvement of identification task. This work also contributes towards the improvement of speaker recognition by providing a better understanding of how clinical and attitudinal information can degrade the performance [6].

A speaker recognition system consists of a feature extractor followed by speaker modeling technique. A number of studies suggest MFCCs [7] for feature extraction and it does produce good results in most of the situations. Feature extraction is followed by a classification algorithm to generate the speaker specific data; GMM has shown promising results in the field of speaker recognition. A number of attempts have been made to use VQ methods with the GMM to optimize the performance of a speaker recognition system [8-9]. The VQ encompasses short term spectral vectors to code vectors. In this paper we use a number of VQ techniques such as K-means, LBG [10] and ITVQ [11] with GMM. The work presented is this paper is further organized as follows, Section 2 presents an overview of the conventional and proposed approaches for feature extraction and modeling, experiments and the ORI (Oregon Research Institute) speech corpus are summarized in Section 3, and finally Section 4 demonstrates the conclusion of this contribution.

2. EXPERIMENTAL FRAMEWORK

2.1 Feature Extraction

The optimization and improvement of classification scores is directly dependent on the selection of features. The better the feature selection is achieved the improved recognition rates will be obtained. Several feature extraction techniques have been used for speaker recognition task, of which widely adapted is MFCC. MFCCs best describe the speaker model [7] because psychophysical studies have established the fact that humans perceives frequency content sound by following a subjectively defined non linear scale, which is called Mel scale [12]. It is given by:

$$f_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Where f is actual frequency in hertz, f_{mel} denotes the subjective pitch in Mels. We derive 20 cepstral coefficient vectors to define a speaker specific data matrix.

2.2 Speaker Modelling

2.2.1 GMM-EM

The GMM [13] is a feature modelling and classification algorithm widely used in the speech-based pattern recognition, since it can smoothly approximate a wide variety of density distributions. The adapted GMM [14] which consists of UBM (Universal Background Model) based on MAP (Maximum a Posteriori) estimation have turned GMMs into reality. The GMMs use EM algorithm for the optimization of GMM parameters such as means, covariances and weights. However, in this paper a number of VQ techniques are used to optimize GMM parameters and their performance is also compared with EM.

The probability density function (pdf) is given as:

$$p\left(x|\lambda\right) = \sum_{t=1}^{M} p_i\, b_i(x) \quad (2)$$

Here x denotes D-dimensional random vector, the component densities are denoted by $b_i(x)$, where i=1,2,3,…,M, and the component weights are denoted by p_i, for i=1,2,3,…,M. The component densities are given by:

$$b_i(x) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_i\right)'\Sigma_i^{-1}\left(x-\mu_i\right)\right\} \quad (3)$$

Here μ_i denotes mean vector and covariance is denoted by Σ. The GMM speaker model is the collection of weights, means and covariances, from all component densities, The speaker model can therefore be represented as class model λ, given as:

$$\lambda = \{p_i, \mu_i, \Sigma_i\}\ i=1,.....M \quad (4)$$

The EM algorithm is most commonly used to iteratively derive optimal class models, however as stated above in this paper number of VQ techniques such as Kmeans, LBG and ITVQ are used to derive optimal class models, beside EM. GMM based on VQ is further detailed in the following section.

2.2.2 GMM-VQ

A number of attempts have been made to use vector quantization methods with GMM to improve the performance [8-9] of speaker recognition system. In VQ a set of code vectors is derived and a codebook is generated from speech feature space by dividing it into K regions. In order to achieve K code vectors the speech feature space is divided into k=1,2,3,.....,K regions. The K-means, LBG and ITVQ clustering methods are used to achieve this goal [15-17].

In GMM based on EM optimization procedure, the values for weights, means and variances are estimated every time using the iterative procedure of EM algorithm; the updates are shown in Equations (5-7).

$$\pi_i(t+1) = \frac{1}{N}\sum_{p=1}^{N_C}\tau_{i_p}(t) \quad (5)$$

$$\mu i(t+1) = \frac{1}{N_C \pi_i(t)}\sum_{p=1}^{N_C}\tau_{i_p}(t) X_p \quad (6)$$

$$V_i(t+1) = \frac{1}{N_C \pi_i(t)}\sum_{p=1}^{N_C}\tau_{i_p}(t)\left((X_p - \mu_i(t))(X_p - \mu_i(t))^T\right) \quad (7)$$

Equation (5) indicates the Weight updates, which satisfies the constraint that sum of all weights should be equal to 1. Equation (6) indicates the mean updates, N_c denotes the number of Gaussian components, π_i are the mixture weights calculated in Equation (5) and τ_{ip} and X_p denote feature vector sets. Equation (7) is the variance update formula, in which μ_i is the mean vector evaluated in Equation (6). We have proposed that code vectors can be generated by using principal approach of VQ methods in order to optimize weight, mean and covariance estimates. GMM-VQ approach establishes the comparable results with GMM-EM. The distortion minimization functions of VQ methods such as K-means, LBG and ITVQ are listed in Equations (8-10).

$$D\sum_{j=1}^{S}\min_{1\le k\le K} d\left(x_i, c_k\right) \quad (8)$$

$$MQE \equiv D(\{Y,S\}) = \frac{1}{N_p}\sum_{p=1}^{N_p} d\left(x_p, q\left(x_p\right)\right) = \frac{1}{N_p}\sum_{i=1}^{N_C} D_i \quad (9)$$

$$w_k(n+1) = w_k(n) - \eta\left(\frac{\Delta V}{V} - 2\frac{\Delta V}{C}\right) \quad (10)$$

In Equation (8) x_j denotes the feature vectors and c_k denotes the centroids calculated based on those feature vectors. In Equation (9) N_p is the number of centroids and x_p and $q(x_p)$ denote the distance between two points in feature set. In Equation (10) the information theoretic weight update formula is shown, in which η is a constant, for our set of experiments η=0.03, provided satisfactory results. The detail description of Equation (10) can be found in [10-11].

3. EXPERIMENTS

This section first describes the speech corpus which has been used to evaluate the performance of a speaker recognition system, followed by experimental results achieved for GMM based modelling methods.

3.1 Clinical Speech Corpus

The experiments have been conducted on the clinical speech corpus and are based on the idea that the recognition rates for voice based biometric systems degrades if the speaker speech samples contain behavioural contents. The video recordings obtained from ORI [18] were used to select speech samples for processing. The data included 138% speaker's (including both depressed and non-depressed) recordings while being occupied in discussion with the children. The family discussion is based on given tasks by psychologists. The videotapes were interpreted by a professional psychologist following the coding system called LIFE (Living in Family Environments) [19]. The video files were converted to audio files with a sampling frequency of 8kHz. The clinical speech corpus which we have used to conduct the initial experiments contains various sets of behavioural contents which are classified based on affect codes such as Aversive affect like contempt, anger and belligerence, the positive and neutral affect such as neutral, pleasant, happy and caring and distressed and depressed affect such as anxious, dysphonic and whine [19].

3.2 Experimental Results

In this section experimental evaluation of the Gaussian mixture speaker model for clinical speech corpus is presented. The experiments are conducted to validate the hypothesis that performing training on clinical speech, as defined in the introduction, is useful in making a speaker verification system robust and is also useful in reducing the false rejection rate for a given false acceptance rate. Two enrolment sets are defined to conduct the experiments. First set represents the speakers who are recorded in the clinical environment and are labelled as depressed speakers and the second set represents the speakers who are labelled as Non-Depressed speakers recorded in the same environment. The experiments are performed to compare the two sets. For first set around 70 speakers were first enrolled as clients in the system, and each speaker was assigned a MFCC data matrix of 20 Cepstral coefficient vectors and then speaker specific model values were derived using GMM modelling. We are using around 4 min length of speech files for training and we are not taking into account which affect-code or content-code is used. As for this initial set of experiments our principal goal is to obtain the recognition rates of speakers who are either depressed as in first set or non-depressed as in the second set. When analyzing results from the experiments four different modelling techniques were used to classify the speakers in both the sets as shown in Table 1. The results in Table 1 show that the speakers who are showing the behavioural speaking styles or in other words the speakers who are identified by the psychologists as depressed show degraded performance when matched with the models to perform classification. A degradation of about 18% in recognition rates can be observed for GMM-EM training, and similarly the degradation continues for other modelling methods too. This leads to the conclusion that speakers with changed behavioural speech contents can cause the degradation in speaker recognition rates.

The DET plot shown in Fig. 1 depicts the false alarm probability and miss probability for the depressed speakers under various modeling methods. The plot shows that a true speaker which is identified as imposter has a highest probability with GMM-Kmeans modeling while the minimum probability rate appears with the GMM-EM modeling method. Similarly the DET plot in Fig. 2 represents the error probability rates for Non-Depressed Speakers. The comparison of the two plots shows that the performance rate is degraded when the speech contains behavioral contents.

TABLE 1. CLASSIFICATION RATE OF DEPRESSED/NON-DEPRESSED SPEAKERS

Modelling Method	Speaker Recognition Rate	
	Depressed (%)	Non-Depressed (%)
GMM-EM	71	89
GMM-K-means	60	79
GMM-LBG	61	80
GMM-ITVQ	69	84

4. CONCLUSION

This paper summarizes experiments on speech containing behavioural contents, and the subject of concern here is to see the performance rates for speaker recognition, However the extensive acoustic analysis and consideration of the various classes of behavioural contents is ongoing. The results address this issue that the behavioural speech or attitudinal speech should be taken as a serious issue for speaker recognition. It is evident from the experiments that the training of speaker verification algorithms on speech samples including behavioural contents instead of neutral speech samples can lead to improvement of speaker recognition rate. The conclusion from these set of experiments also address that study of the role of speaker disparities in vocal reactivity and recognition is useful. The results suggest that classification of speakers using GMM-EM method has established better results while the results achieved with GMM-ITVQ are also comparable. A degradation rate of 18% was observed when depressed/ Non-depressed speakers were classified using GMM-EM method.

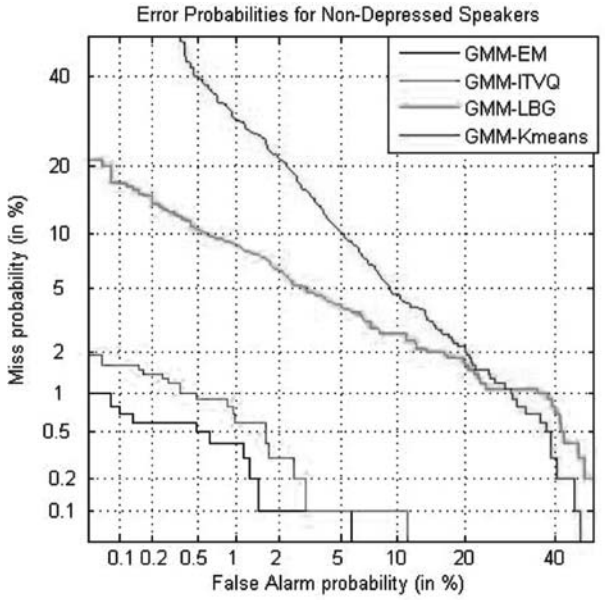

FIG. 1. DET PLOT FOR THE NON DEPRESSED SPEAKERS

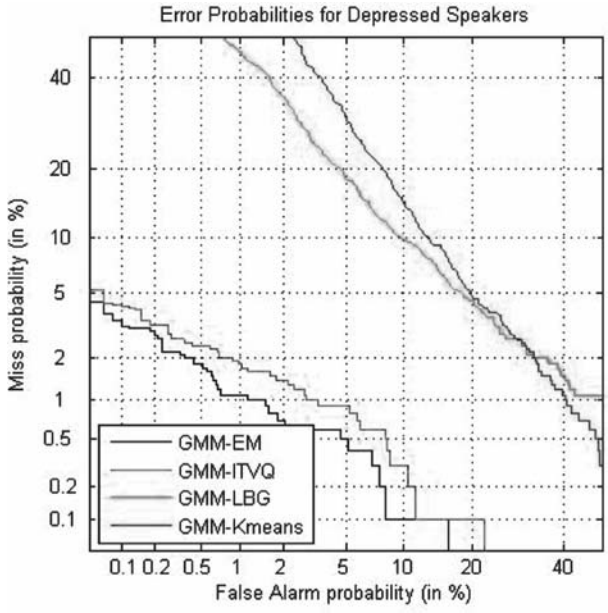

FIG. 2. DET PLOT FOR THE DEPRESSED SPEAKERS

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr. Abdul Qadeer Khan Rajput, Vice-Chancellor, and Prof. Dr. Mukhtiar Ali Unar, Director, Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan, for their support and guidance in order to complete this research.

REFERENCES

[1] Oregon Research Institute, "An Independent Behavioural Sciences Research Centre", USA.

[2] Karlsson, I., Banziger, T., Dankovicová , J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., and Scherer, K., "Speaker Verification with Elicited Speaking Styles in the VeriVox Project", Journal of Speech Communication, Volume 31, Nos. 2-3, pp. 121-129, 2000.

[3] Scherer, K.R., Johnstone, T., Klasmeyer, G., Bänzigerand, T., "Can Automatic Speaker Verification be Improved by Training the Algorithms on Emotional Speech", Proceedings of Interspeech, 2000.

[4] Moore, E., Clements, M.A., Peifer, J.W., and Weisser, L., "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech", IEEE Transactions on Biomedical Engineering, Volume 55, pp. 96-107, 2008.

[5] He, L., Lech, M., and Maddage, N.C., "Study of Empirical Mode Decomposition and Spectral Analysis for Stress and Emotion Classification in Natural Speech", Elsevier Journal of Biomedical Signal Processing and Control, Volume 6, No. 2, pp. 139-146, April, 2011.

[6] Murray, I.R., and Arnott, J.L. "Toward a Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion", Journal of the Acoustical Society of America, Volume 93, pp. 1097-1108, 1993.

[7] Reynolds, D.A., "Experimental Evaluation of Features for Robust Speaker Identification", IEEE Transactions on Speech Audio Processing, Volume 2, No. 4, pp. 639-643, October, 1994.

[8] Jialong, H., Liu, L., and Gunther, P., "A new Codebook Training Algorithm For VQ-Based Speaker Recognition", IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 1091-1094, 1997.

[9] Singh, G., Panda, A., Bhattacharyya, S., and Srikanthan, T., "Vector Quantization Techniques for GMM Based Speaker Verification", IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 1165-1168, 2003.

[10] Linde, Y., Buzo, A., and Gray, R.M., "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Volume 28, No. 1, January, 1980.

[11] Tue, L., Hedge, A., Deniz, E., and Principe, J.C., "Vector Quantization Using Information Theoretic Concepts", International Journal of Natural Computing, Volume 4, No. 1, pp. 39-51, January, 2005.

[12] Ben, G., and Nelson, M., "Speech and Audio Signal Processing", Part-IV, Chapter-14, pp. 189-203, John Willy & Sons, 2002.

[13] Campbell, J.P.Jr., "Speaker Recognition: A Tutorial", Proceedings of the IEEE, Volume 85, No. 9, pp. 1437-1462, September, 1997.

[14] Reynolds, D.A., Quatieri, T., and Dunn, R., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Volume 10, pp. 19-41, 2000.

[15] Memon, S., and Lech, M., "Speaker Verification Based on Information Theoretic Vector Quantization", Proceedings of IMTIC, Communication in Computer and Information Science Series, Volume 20, pp. 391-399, Springer Berlin Heidelberg, April, 2008.

[16] Memon, S., and Lech, M., "Using Information Theoretic Vector Quantization for GMM Based Speaker Verification", Proceedings of EUSIPCO, Eurasip Library, Lausanne Switzerland, August, 2008.

[17] Memon, S., Khanzada, T.J.S., and Bhatti, S., "Text-Independent Speaker Verification Based on Information Theoretic Learning", Mehran University Research Journal of Engineering and Technology, Volume 30, No. 3, Jamshoro, Pakistan, July, 2011.

[18] Davis, B., Sheeber, L, Hops, H., and Tildesley, E., "Adolescent Responses to Depressive Parental Behaviors in Problem-Solving Interactions: Implications for Depressive Symptoms", Journal of Abnormal Child Psychology, Volume 28, No. 5, pp. 451-465, 2000.

[19] Longoria, N., Sheeber, L., and Davis, B., "Living in Family Environment Coding", A Reference Model for Coders, OREGON Research Institute, 2006.