Effect of Clinical Depression on Automatic Speaker Verification

Sheeraz Memon[1], *Mukhtiar Ali Unar*[1] *and Bhawani Shankar Chowdhry*[2]

[1]Department of Computer System Engineering, Mehran UET, Jamshoro, Pakistan
[2]Department of Electronic Engineering, Mehran UET, Jamshoro, Pakistan
{sheeraz.memon, mukhtiar.unar, bhawani.chowdhry}@faculty.muet.edu.pk

Abstract

The effect of a clinical environment on the accuracy of the speaker verification was tested. The speaker verification tests were performed within homogeneous environments containing clinically depressed speakers only, and non-depressed speakers only, as well as within mixed environments containing different mixtures of both clinically depressed and non-depressed speakers. The speaker verification framework included the MFCCs features and the GMM modeling and classification method. The speaker verification experiments within homogeneous environments showed 5.1% increase of the EER within the clinically depressed environment when compared to the non-depressed environment. It indicated that the clinical depression increases the intra-speaker variability and makes the speaker verification task more challenging. Experiments with mixed environments indicated that the increase of the percentage of the depressed individuals within a mixed environment increases the speaker verification equal error rates.

Index Terms: *Speaker verification, Clinical environment, Clinical depression.*

1. Introduction

The performance of speaker recognition systems degrades due to both, the intra-speaker variability and the background noise. One of the factors affecting the intra-speaker variability is the clinical depression. Speech contents of clinically depressed speakers consist of more abstractive flow of conversations, higher frequency of pauses and more nonverbal sounds than speech of normal speakers [19]. It has been also previously demonstrated that clinical depression changes acoustic characteristics of speech [1-3, 13-15] and therefore, it can be hypothesized that the speaker verification accuracy can be affected in an environment consisting fully or partially of clinically depressed people.

This paper aims to determine the effects of clinical environments consisting of clinically depressed people on the speaker verification rates when using the state of the art speaker recognition techniques. The importance of this study is given by the fact that robust speaker recognition systems have potential applications in the clinical environments and the health care sectors such as telemedicine, biometrics and surveillance systems. It is recently reported that nearly 20 percent of military service members who have returned from Iraq and Afghanistan report symptoms of post-traumatic stress disorder or major depression, according to a new RAND Corporation study [22].

The state of the art speaker recognition systems extracts acoustic features which capture the characteristics of the speech production system such as pitch or energy contours [7], glottal waveforms [6], or formant amplitude and frequency modulation [5] and model them using statistical learning techniques [20,21]. The *Mel frequency cepstral coefficients* (MFCCs) have been commonly used to characterize acoustic properties of speech [8] often in conjunction with the *Gaussian mixtures model* (GMM), which is regarded to be one of the best statistical modeling techniques used in speaker recognition systems [8,16,17,18]. The characteristic features used in this study include MFCCs, their velocity and acceleration, short time energy and zero crossing rates. The pre-processing stage was used to separate the silence/noise intervals and to perform the pre-emphasis filtering. The speaker models were built using the Gaussian mixture modeling based on the expectation maximization (EM) procedure [16,17].

The remaining part of this paper is organized as follows. Section 2 describes the speaker verification system. In Section 3, the experimental setup and results are presented, and finally, Section 3 contains the conclusions.

2. Speaker Verification System

2.1 General Framework

The general framework of the speaker verification system used to conduct our experiments is shown in Figure.1. The system can operate in one of the three possible modes: universal background model (UBM) training mode, target speaker enrollment mode and testing mode. In each case identical speech detection and feature extraction methods are used. An energy based silence detector was used to discard the low energy intervals of the signal [16]. Previous research has shown that the MFCC based systems are not very sensitive to changes in frame size (in the range 20-50ms) and frame step (in the range 1/6 to 1/3 of the frame size). Frames whose energy is too low to be considered speaker-discriminative were therefore excluded from subsequent processing. From each remaining frame, the first 12 MFCC were computed and normalized using the cepstral mean subtraction (CMS) method.

The sequences of feature vectors were then modeled with the GMM. For each target speaker 1024 Gaussian mixtures were generated. Each model was defined by a set of parameters including its *a priori* probability, mean vector, and the diagonal covariance matrix. The speaker models were trained with around 5 minutes of data length for each speaker. After the enrollment (training) stage, the universal background model (UBM) parameters [18] were derived using the expectation maximization (EM) algorithm trained on a large speech corpus including the non-target speakers obtained from the NIST 2001 and NIST 2002 SRE corpora. The target speaker's model means were then adapted using the maximum a posteriori (MAP) estimation method, the UBM and the target speaker's data. During the testing stage, the same pre-processing and feature extraction methods were applied to the test data as in the training stage. The testing sequences of feature vectors were then scored by each speaker's model, and the verification decision was made based on the identity of the highest scoring model. The general system performance was assessed using the equal error rate (EER) measure and by plotting the detection error trade-off (DET) curves.

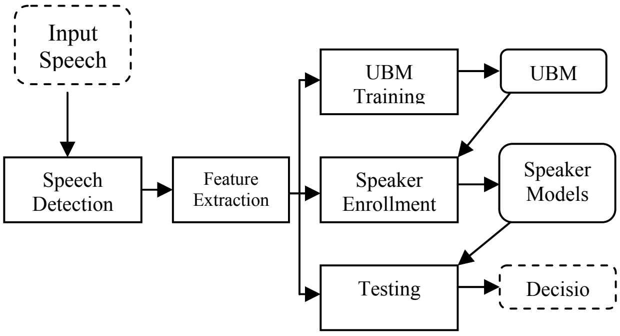

Figure.1. *General framework of the speaker verification system.*

2.2. Clinical Speech Corpus

The clinical data used in this study (CC-ORI) was obtained from the Oregon Research Institute, USA. The non-clinical data was obtained from the NIST 2004 corpus.

The CC-ORI consists of audio recordings of 139 adolescents (93 females and 46 male) aged 13-19, participating in typical discussions between family members. Each subject was represented by around one hour of recording. Details of the data acquisition sessions can be found in [9-11]. Through the self-report and interview measures of the adolescence's depression evaluated by psychologists from ORI [9], 68 (49 females and 19 males) were diagnosed as suffering from major depressive disorder (MDD), and the remainder (44 females and 27 males) were healthy controls (i.e., no current or lifetime history of MDD). The clinically depressed and healthy groups were matched on their demographic data which included their sex, race and age. The CC-ORI recordings were sampled with the 8 kHz rate and 16 bits/sample.

The NIST 2004 samples were derived from the Switchboard 2, and the Mixer projects. The Switchboard-2 Corpora included mostly college or early post-college age students [4] from a specific area of the United States. The NIST 2004 data was also sampled at 8 kHz rate with 16bits/s.

2.3. Speech Segmentation and Feature Extraction

The speech signal was segmented using the Hamming window into short frames of length 20 ms within which the spectral and temporal properties of speech such as signal energy and pitch can be assumed stationary [8,12]. There was 50% overlap between frames. The feature extraction was performed on the frame-by-frame basis. Each frame was used to derive a feature vector consisting of: 12 MFCCs coefficients, 12 Δ-MFCC (first derivative of MFCCs), 12ΔΔ-MFCC (second derivative of MFCCs), 1-short time energy coefficient and 1-zero-crossing coefficients. The resulting arrays of 38-dimensional feature vectors were used to test the speaker verification rates in different environments

3. Experiments and Results

3.1. Individual Class Speaker Verification (ICSV) within Homogeneous Environments using CC-ORI

In this experiment, the speaker verification was performed within two homogeneous environments. The first environment contained 100% of depressed speakers and the second environment contained 100% non-depressed speakers. The depressed environment contained 68 speakers (49 females and 19 males), and the non-depressed environment contained 71 speakers (44 females and 27 males). The experimental results for the intra-class speaker verification tests (ICSV) test within homogeneous environments are presented in Figure.2. It can be clearly observed that the speaker verification task within the depressed environment is more challenging than within the non-depressed environment. The speaker verification equal error rate (EER) for the depressed speakers is 5.1% higher than for the non-depressed speakers. Since the numbers of speakers in both clinically depressed and non-depressed classes were almost the same, and the utterances were recorded under the same background noise conditions, it can be concluded that the clinical depression was the main factor causing the degradation of speaker verification accuracy within the depressed environment when compared with the non-depressed environment. It also indicates that the intra-speaker variability within the depressed environment is higher than within the non-depressed environment.

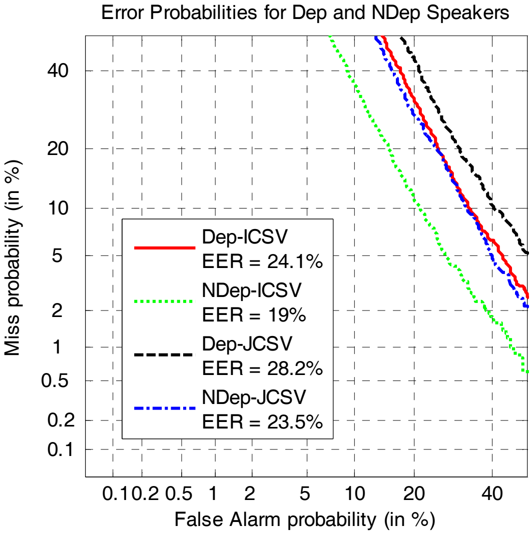

Figure. 2. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the ICSV test within homogeneous environments.*

3.2. Speaker Verification within Mixed Environments using CC-ORI

This set of experiment tested the speaker verification accuracy within mixed environments i.e., environments consisting of different mixtures of both, depressed and non-depressed speakers. The detection error trade-off (DET) curves and the equal error rates (EERs) for the mixed environments constructed out of the CC_ORI data are presented in Figure.3, denoted as Joint-class speaker verification (JCSV). Four different environmental mixtures were used. Each mixture contained a fixed number of 68 non-depressed speakers. The first mixture had no depressed speakers, the second mixture contained 17 depressed speakers, the third mixture contained 34 depressed speakers and the fourth mixture contained 68 depressed speakers. Since, the mixed environments were composed of speakers from the CI-ORI corpus only; the conditions of recordings and the background noise were the same for all speakers. Also, each environment contained the same number of non-depressed speakers; only the number of depressed speakers was changing. It can be therefore assumed that the observed effects on speaker verification were mostly due to different amounts of depressed individuals within a given environment. The results in Figure.3 show that the increasing percentage of the depressed speakers within a mixed environment leads to an increase of the EER values.

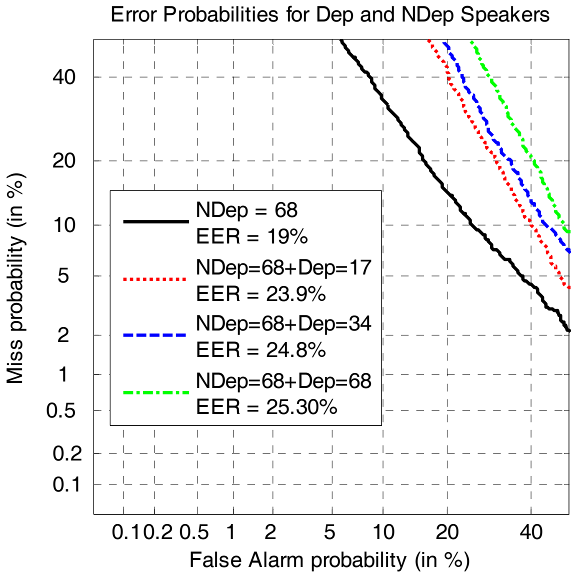

Figure.3. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the JCSV test within mixed environments using CC-ORI.*

3.3. Speaker Verification within Mixed Environments using CC-ORI and NIST2004

This set of experiment again tested the speaker verification accuracy within mixed environments however this time the depressed speakers were sourced from the CC-ORI and the non-depressed speakers were taken from the NIST 2004 corpora. The resulting detection error trade-off (DET) curves and the equal error rates (EERs) for the mixed environments constructed out of the CC_ORI data are presented in Figure.4. Three different environmental mixtures were used. The first mixture contained 68 depressed speakers from CC-ORI and no non-depressed speakers. The second mixture contained 616 non-depressed speakers from NIST 2004 and the third mixture contained 616 non-depressed speakers from NIST 2004 and 68 depressed speakers from CC-ORI.

Since, the mixed environments were composed of speakers from two different data bases (CI-ORI and NIST 2004), the recording conditions and the background noise were different. It is therefore difficult to draw any definite conclusions. Figure.4 shows that the addition of 68 depressed CC-ORI speakers to the 616 NIST 2004 non-depressed speakers increases the EER values compare to the environment containing only the 616 NIST 2004 speakers. This could be the result of both, the clinical depression and the different noise level in the CC-ORI recordings. To be able to draw more definite conclusions, further research on equalization methods compensating for the differences in the recording conditions is needed.

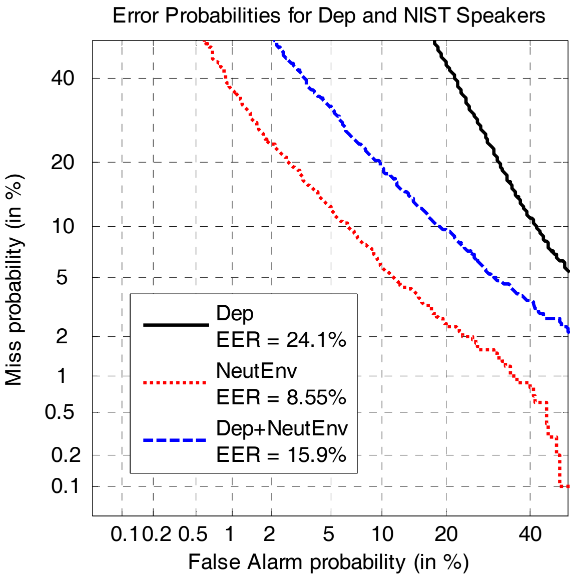

Figure.4. *Detection Error Trade-Off (DET) curves and Equal Error Rates (EERs) for the JCSV test within mixed environments using CC-ORI+NIST200*

4. Conclusions

The effects of the clinical depression on the speaker verification accuracy were investigated. The tests conducted within homogeneous environments using the same data base clearly indicated that the speaker verification within the clinically depressed environment is more challenging than within the non-depressed environment and the EER values obtained within the depressed environment are higher than within the non-depressed environment. The tests conducted within mixed environments composed of the same data base indicated that the higher is the percentage of the depressed speakers, the larger are the speaker verification EER values. Finally, the tests conducted within mixed environments constructed out of two different data bases were not conclusive due to the lack of equalization methods allowing to directly merge data recorded under different conditions.

5. Acknowledgements

The authors would like to thank the Oregon Research Institute, U.S.A. and Dr. Lisa Sheeber for providing the clinical data base. This study supported in part by the Australian Research Council Linkage Project LP0776235.

6. References

[1] Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K. "Speaker verification with elicited speaking styles in the VeriVox project," Speech Communication 31(2-3): 121-129, 2000.

[2] K. R. Scherer, T. Johnstone, G. Klasmeyer, & T. Bänziger "Can automatic speaker verification be improved by training the algorithms on emotional speech" University of Geneva, Switzerland.

[3] Murray, I. R., & Arnott, J. L. “Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” JASA, 93: 1097-1108, 1993.

[4] Alvin F. Martin, “Encyclopedia of Biometrics” National Institute of Standards and Technology Gaithersburg, Maryland, USA.

[5] C. R. Jankowski jr. et al., “Fine structure features for speaker identification,” in Proc. ICASSP, 1996, pp. 689–692.

[6] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification,” IEEE Trans. Speech Audio Process., vol. 7, no. 5, pp. 569–586, Sep. 1999.

[7] B. Peskin et al., “Using prosodic and conversational features for high performance speaker recognition: Report from JHU WS02,” in Proc. ICASSP, vol. 4, 2003, pp. 792–795.

[8] D. A. Reynolds, “Experimental evaluation of features for robust speaker identification,” IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 639–643, Oct. 1994.

[9] L. Sheeber, H. Hops, J. Andrews, T. Alpert, and B. Davis, “Interactional processes in families with depressed and non-depressed adolescents: reinforcement of depressive behaviour,” Behaviour Research and Therapy, vol. 36, pp. 417-427, 1998.

[10] H. Hops, A. Biglan, A. Tolman, L. Sherman, J. Arthur, and N. Longoria, “Living in family environments (LIFE) coding system: Reference manual for coders,” Oregon Research Institute, Eugene, OR, Unpublished manuscript, 2003.

[11] H. Hops, B. Davis, and N. Longoria, “Methodological issues in direct observation-illustrations with the living in familial environments (LIFE) coding system,” Journal of Clinical Child Psychology, vol. 24, pp. 193-203, 1995.

[12] L.R. Rabiner, R.W. Schafer, “Digital Processing of Speech Signals”, Prentice-Hall Inc.

[13] D. J. France, et al. “Acoustical properties of speech as indicators of depression and suicidal risk,” IEEE Transactions, Biomedical Engineering, vol. 47, pp. 829-837, 2000.

[14] E. Moore, et al. “Critical analysis of the impact of glottal features in the classification of clinical depression in speech,” IEEE Trans, Biomed Eng, vol. 55, pp. 96-107, 2008.

[15] A. Ozdas, et al. “Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk,” IEEE Trans, Biomed Eng, vol. 51, pp. 1530-1540, 2004.

[16] Reynolds, D. A., Rose, R. C., and Smith, M. J. T., PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system. In Proceedings of the International Conference on Signal Processing Applications and Technology, November 1992, pp. 967–973.

[17] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” in *IEEE Trans. Speech and Audio Processing*, 1995, vol. 3, pp. 72–83.

[18] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[19] Christopher. D and Marybeth. S,” Intra-speaker variability in palatometric measures of consonant articulation”, Journal of Communication Disorders, Volume 42, Issue 6, Dec 2009, Pages 397-407.

[20] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li and Li Rong Dai, "Local variability vector for text-independent speaker verification", *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 54-58, 2014.

[21] Patrick Kenny; Gilles Boulianne; Pierre Ouellet; Pierre Dumouchel, “Speaker and Session Variability in GMM-Based Speaker Verification “ , IEEE Transactions on Audio, Speech, and Language Processing,, Volume: 15, Issue: 4 Pages: 1448 - 1460, 2007.

[22] The masks of war: American military styles in strategy and analysis: A RAND Corporation research study, C Builder, Johns Hopkins University Press.