SPEAKER RECOGNITION IN REVERBERANT ENVIRONMENTS

Joseph Gammal, Rafik Goubran
Dept. of Systems and Computer Engineering, Carleton University, Colonel By Dr., Ottawa, Canada, K1S 5B6
{jgammal,goubran}@sce.carleton.ca

1. INTRODUCTION

Closed set speaker recognition is the task of, given a speech utterance, correctly selecting the identity of an unknown speaker from a limited set of speakers. Different methods are used for this purpose and each can employ a variety of feature vectors extracted from the speech. The objective of this paper is to objectively compare these methods and parameterizations when there is a mismatch between training and test conditions caused by the existence of reverberation.

Three classes of speaker recognition algorithms were used: Gaussian mixture models (GMM), covariance models and AR-vector models. Each of the last two classes employs two different speaker recognition measures.

Two different feature vectors were extracted from the speech, these are LPC cepstral (LPCC) and Mel-warped cepstral (MFCC) vectors [1]. The effect of the addition of delta-cepstral vectors was investigated.

2. METHOD

2.1 Reverberation Models

Reverberation was simulated using the image-method [2]. Impulse responses were generated for two rooms. The first had dimensions 3.6x4.2x3m with reflection coefficients for the walls as 0.9 and floors as 0.7, and the second had dimensions 3x6x2.5m with reflection coefficients of 0.93 and 0.8. Microphone to speaker separation in the first room was 0.75m and in the second was 0.54m. The first configuration is characterized as minor reverb., the second as major reverb. The impulse responses were generated using a sampling frequency of 8Khz. The lengths of the impulse responses were 1 and 2 seconds respectively.

2.2 Database and Signal Processing

The KING speaker recognition database was used. The speech files in the 51-speaker database were band limited to the 300-3400Hz telephone band and μ-law coded. A sampling frequency of 8Khz was used. Reverberated versions of the database were produced by convolving the original database with either of the impulse responses before further processing. 20ms frames were extracted every 10ms after silence removal. Either MFCC or LPCC vectors were extracted. For MFCC vectors the filter bank contained 19 filters. When delta-cepstral (Δ) vectors were used, they were produced using a 5-frame first order orthogonal polynomial fit. Cepstral mean subtraction (CMS) was applied to all vectors.

2.3 Recognition Methods

The GMM was produced using the method outlined in [3] and using diagonal covariance matrices. Both the sphericity (SM) and divergence shape (DS) [4] were used for the covariance-based methods. The DS and SM, which are distances between a claimant model and test utterance are calculated as follows [4]:

$$DS_{1,2} = DS(C_1, C_2) = \frac{1}{2} trace[(C_1 - C_2)(C_2^{-1} - C_1^{-1})] \quad (1)$$

$$SM_{1,2} = SM(C_1, C_2) = \frac{1}{2} trace(C_1 C_2^{-1}) trace(C_2 C_1^{-1}) \quad (2)$$

Here C_1 is the covariance matrix of the training speech and C_2 is that of the test utterance. For both the second-order AR-vector methods the training method specified in [6] was used to train the models. For a set of p-dimensional training vectors $\{\vec{x}_t\}_{1 \le t \le N}$ with mean $\vec{\mu}$, the 2nd order AR-vector model whose solution as follows [6]:

$$\sum_{i=0}^{2} A_i(\vec{x}_{t-i} - \vec{\mu}) = \vec{e}_t \text{ with } A_0 = I_p \quad (3)$$

The lagged covariance matrices are defined as follows [6]:

$$\chi_k = \frac{1}{N} \sum_{t=k+1}^{N} (\vec{x}_t - \vec{\mu})(\vec{x}_{t-k} - \vec{\mu})^T \text{ with } k = 0..2 \quad (4)$$

The Toeplitz matrix X is defined as follows:

$$X = \begin{bmatrix} \chi_0 & \chi_1^T & \chi_2^T \\ \chi_1 & \chi_0 & \chi_1^T \\ \chi_2 & \chi_1 & \chi_0 \end{bmatrix} \quad (5)$$

With A= [A_0 A_1 A_2] let $E_X^{(A)} = AXA^T$. A set of vectors from the test utterance $\{\vec{y}_t\}_{1\le t\le M}$ has model B. Let $E_Y^{(B)} = BYB^T$, $E_X^{(B)} = BXB^T$ and $E_Y^{(A)} = AYA^T$ [6]. The distance measure referred to in this paper as AR1 is as follows [5]:

$$AR1_{1,2} = \frac{1}{2}\log[trace(\frac{E_Y^{(A)}}{E_Y^{(B)}})\ x\ trace(\frac{E_X^{(B)}}{E_X^{(A)}})] \qquad (6)$$

AR2 requires that the vectors be sorted randomly prior to training and testing. The distance measure is as follows [6]:

$$AR2_{1,2} = \log\left[\frac{\frac{1}{p}trace(E_X^{(A)^{-\frac{1}{2}}}E_Y^{(A)}E_X^{(A)^{-\frac{1}{2}}})}{[\det(E_X^{(A)^{-\frac{1}{2}}}E_Y^{(A)}E_X^{(A)^{-\frac{1}{2}}})]^{\frac{1}{p}}}\right] \qquad (7)$$

3. RESULTS

Table 1. Recognition accuracy.

Method & feature	Recognition accuracy (%)		
	No reverb	Minor reverb	Major reverb
GMM64 LPCC	93.4	79.0	70.6
GMM64 LPCC+Δ	96.3	72.0	62.0
GMM64 MFCC	93.1	77.5	67.1
GMM64 MFCC+Δ	94.2	76.4	66.3
AR1 LPCC	91.9	40.3	30.8
AR1 LPCC+Δ	92.2	50.1	38.9
AR1 MFCC	90.5	33.7	29.1
AR1 MFCC+Δ	85.9	44.4	40.9
AR2 LPCC	95.1	80.7	75.8
AR2 LPCC+Δ	96.3	68.6	63.4
AR2 MFCC	91.4	53.3	46.4
AR2 MFCC+Δ	94.2	35.2	29.4
SM LPCC	94.8	73.2	67.4
SM LPCC+Δ	94.2	57.6	52.2
SM MFCC	89.0	58.5	50.7
SM MFCC+Δ	93.9	43.5	40.9
DS LPCC	94.2	76.1	69.5
DS LPCC+Δ	94.2	44.4	33.4
DS MFCC	91.9	70.3	67.1
DS MFCC+Δ	93.9	48.4	41.8

Each recognition method was trained using sessions 1- 3, and tested using 30s segments from sessions 4-10.

4. DISCUSSION

The results reveal that performance for all methods degrades under reverberation. Delta-cepstral coefficients degrade recognition performance considerably under conditions of reverberation for all methods except AR1, where in the case of AR1 they enhance recognition performance. When delta-cepstral coefficients are used the GMM is the most robust to reverberation. When delta-cepstral coefficients are not used AR2 using LPCC vectors is the most robust to reverberation followed by the GMM. LPCC and MFCC vectors are affected differently by reverberation. When delta-cepstral features are not used, LPCC vectors are more robust to reverberation than MFCC vectors for all methods except the GMM.

It was found that when training was performed with reverberant speech prior to testing with minor reverb or major reverb that recognition accuracy improved for all methods. This was found regardless of whether the impulse response used during training was the same as that used during testing. It was also found that training with reverberant speech and testing with non-reverberant speech gave better results than when training was performed with non-reverberant speech and testing was performed with reverberant speech.

REFERENCES

[1] S. Davis, P. Mermelstein (1980) "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Signal Processing, vol. 28, pp. 357-366

[2] J. B. Allen, D. A. Berkley, (1979) "Image method for efficiently simulating small-room acoustics", Journal of the Acoustical Society of America, vol. 65. pp. 943-950

[3] D. Reynolds, R. Rose, (1995) "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, vol. 3, Issue: 1, pp. 72-83

[4] R. Zilca, (2001): "Using second order statistics for text independent speaker verification", in ODYSSEY-2001, pp. 45-49.

[5] C. de Lima, D. da Silva, A. Alcaim, J. Apolinario Jr., (2002) "AR-Vector using CMS for robust text independent speaker verification", in Proc. 14th International Conference on DSP, vol 2, pp 1073-1076

[6] I. Magrin-Chagnolleau, J. Wilke, F. Bimbot (1996), "A further investigation on AR-vector models for text-independent speaker identification", in Proc. ICASSP'96 pp. 401-404