REVERBERATION MATCHING FOR SPEAKER RECOGNITION

Itai Peer[1], *Boaz Rafaely*[2] *and Yaniv Zigel*[3]

[1,2]Department of Electrical and Computer Engineering, [3]Department of Biomedical Engineering
Ben-Gurion University of the Negev, Beer-Sheva, 84105, Israel
[1] itaipee@bgu.ac.il, [2] br@ee.bgu.ac.il, [3] yaniv@bgu.ac.il

ABSTRACT

Speech recorded by a distant microphone in a room may be subject to reverberation. Performance of a speaker verification system may degrade significantly for reverberant speech, with severe consequences in a wide range of real applications. This paper presents a comprehensive study of the effect of reverberation on speaker verification, and investigates approaches to reduce the effect of reverberation: training target models with reverberant speech signals and using acoustically matched models for the reverberant speech under test, score normalization methods to improve the reverberation robustness, and also reverberation classification via the background model scores. Experimental investigation is performed, using simulated and measured room impulse responses, NIST-based speech database, and AGMM based speaker verification system, showing significant improvement in performance.

Index Terms— Model Matching, Reverberation, Robust Recognition, Speaker Recognition.

1. INTRODUCTION

A wide range of speech communication systems are in use today. This variety introduces acoustic mismatch to speaker recognition systems i.e. the acoustic environments in the training stage and testing stage might be different. Performance of current speaker recognition systems can degrade significantly under mismatched conditions [1], [2] such as different room reverberation [3]-[5]. Speech communication systems operating inside rooms, in which the microphone is distant from the speaker, will produce reverberant speech. Therefore, speaker recognition systems utilizing current speech communication systems will produce poor performance under reverberation mismatch.

Several approaches for overcoming the effect of reverberation have been reported. Dereverberation methods attempt to reconstruct the clean speech signal from the reverberant signal [6], however, no successful method for single channel dereverberation seems to exist. Another approach is to use a microphone array [7], to reduce room reflections and enhance the direct sound. However, in most speaker recognition applications, microphone arrays are not readily available.

For speaker recognition systems it is common to divide solutions for general channel mismatch into three domains [2], [8]: feature domain compensation is aimed at removing the channel effects from the feature vectors prior to model training or verification. It includes cepstral mean subtraction (CMS) [2], feature mapping [8] and joint factor analysis [9]. Model domain techniques such as speaker model synthesis (SMS) [10] and eigenchannel modeling [9] attempt to modify the models to minimize the effects of varying channel. Score domain solutions aim to remove score shifts and scaling caused by the varying channel conditions. Among these methods are Z-norm [1], H-norm [11], T-norm [12], and Top-norm [13]. Although these methods have been proposed for general channel mismatch, most of them were not examined for reverberation mismatch.

In this paper we study the methods of acoustic model matching and score normalization. In [4] the use of model matching showed to improve speaker recognition when training with the autoregressive (AR) vector method. Another work suggests feature domain reverberation compensation with model matching [5]; however, this method was tested with both the training data and test segments recorded in the same rooms, and might therefore suffer from over-fitting.

The aim of the research presented in this paper is to improve speaker recognition performance for reverberant speech by extracting and utilizing information from the acoustic environment. The paper presents several contributions. Room acoustic model matching is employed in speaker recognition, supporting previous results [4], [5] and extending these results to a speaker verification system employing the widely used adaptive Gaussian mixture model (AGMM) [11]. Score normalization is also employed showing significant performance improvement. Finally, the model matching method leads to another problem – how to select the right target model? Reverberation classification is introduced which uses the background models scores, and in addition to the widely used reverberation time for reverberation parameterization, a new parameter, the frame definition is introduced showing improved performance.

2. THE EFFECT OF REVERBERATION ON SPEECH

Speech measured by a microphone in a room can be modeled by convolution between the speech signal and the room impulse response [14], the latter composed of the

This work was supported in part by the Ministry of Industry and Trade, grant no. 36294

direct sound and reflections from the room walls and objects in the room. When contributions from room reflections are significant compared to the direct sound, the speech is said to be reverberant. The most common measure of reverberation is the reverberation time (T_{60} or *RT*) [14], [15] which is the time it takes the acoustic energy in the room to decay by 60dB after the source is switched off. T_{60} can be calculated from the room impulse response $h(t)$ using Schroeder method [15]:

$$T_{60} = \left\{ t : P(t) = P(0)/10^6 ; P(t) = \int_t^{\infty} h^2(\tau) d\tau \right\} \quad (1)$$

Although T_{60} has been used in a previous speaker recognition study [4], another parameter, called frame-definition in this paper, seem to better represent the signal leakage from adjacent frames due to reverberation. It is based on the intelligibility parameter, the definition [14], defined as the ratio between the energy in the early part of the room impulse response and the total energy. With the early part set to 20ms equal to a frame length in the speaker recognition system, frame-definition, *FD*, or D_{20}, is given by

$$D_{20} = \int_0^{20ms} h^2(t)dt \Big/ \int_0^{\infty} h^2(t)dt \,. \quad (2)$$

In this paper both T_{60} and D_{20} are used for reverberation classification.

Both measured and simulated room impulse responses have been employed in this paper. The simulated room impulse responses have been generated using the widely used image method [16]. Measured room impulse responses were generated by measuring the response between a loudspeaker (KRK model RP-6) to a microphone (Bruel & Kjaer type 4133) in various lecture and laboratory rooms at Ben-Gurion University.

3. THE SPEAKER VERIFICATION SYSTEM

3.1. The baseline system

Figure 1 shows a block diagram of the speaker verification system. 12 Mel frequency cepstral coefficients (MFCCs) and 12 delta-MFCCs (ΔMFCCs) [1] are extracted from each frame. In addition, CMS was performed on the features to improve robustness [2]. In the training phase, the background models were trained using GMM with 1024 Gaussians and diagonal covariance matrixes. AGMM training was performed in order to adapt the target models from the background models. In the testing phase, after feature extraction, a log-likelihood ratio test was employed to compute target and impostors scores, involving the test segments, target models and background models.

System expansions beyond the baseline system are the reverberation classification, the model matching and the designing of the score normalization which are detailed in the following sections.

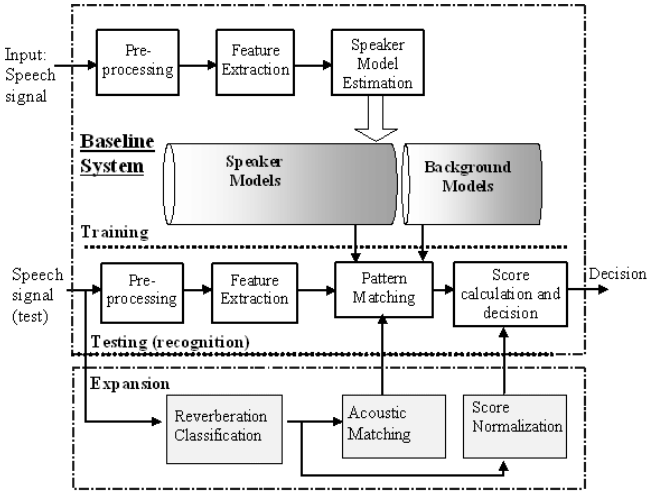

Fig. 1. The speaker recognition system

3.2. Acoustic model matching

Acoustic matching of speaker models involves training and testing under the same room acoustic conditions, e.g. same *RT*. In the training phase, several models are generated for each speaker under various reverberation conditions. First, reverberant background model (RBM) for each *RT* is produced: Clean[1] speech segments from various speakers are filtered by simulated room impulse response and a reverberant background model is trained using these segments. These RBMs are used also for reverberation classification (section 3.4). Then, using the AGMM training, the reverberant speaker model is adapted from the RBM and the speaker reverberant speech signal. Although measured room responses in the training data might better match the reverberant test segments, simulated responses [16] were used for convolving with the training speech signals due to their availability and simplicity.

3.3. Score normalization

This section details the score normalization methods which were used in the experiments. Model-dependant normalizations aim to remove the model bias. Among them is the Z-Norm [1] which uses a development group with several room responses for estimating the normalization parameters of each target model; Top-Norm [13] which uses the top 10% highest scores, and a modified H-Norm [11] designed for reverberation mismatch instead of handset mismatch. The modified H-Norm uses reverberant test segments with similar reverberation parameters as the target model. Also, a test-dependant normalization which aims to remove the test bias, T-Norm [12], was used. In the experiments, combined normalizations were

[1] The speech signals were recorded via landline telephone channels, but are referred to as "clean" for not including reverberation

employed: ZT-Norm, Top-T-Norm and HT-Norm.

3.4. Reverberation classification

Having several target models for each speaker, each for a different reverberation, leads to another problem – how to select the target model for the score calculation? We suggest the use of RBMs likelihood scores for classification. Each RBM is trained with a large number of speakers under specific reverberation conditions. Then, each test segment with an unknown acoustic condition is tested against all RBMs and is classified by selecting the highest score. The assumption at the base of this method is that the highest score will correspond to the most similar reverberation condition, regardless of the speaker identity. This method is referred to as RBM classification.

4. EXPERIMENTAL SETUP

This section describes a set of experiments aimed to investigate speaker verification using model matching method and the RBM reverberation classification. A database comprising one-minute long (clean) speech segments, taken from NIST-99 speaker recognition evaluation corpus, was used for training target models for 198 male speakers. An additional set of 50 one-minute speech segments of different speakers, taken from NIST-98, was used to train the background models. NIST-99 and NIST-98 speaker recognition evaluation corpora include conversational speech over land-line telephone, sampled at 8 kHz, with 8 bits μ-low.

Using the Image method [16], four room impulse responses were created with reverberation times of 0.238, 0.513, 0.863 and 1.111 seconds. Five background models (one clean and 4 reverberant) and 5 target models for each speaker were trained as detailed in section 3.2.

The database for testing was also taken from NIST-99, incorporating 589 speech segments, with durations ranging from 20 to 60 seconds. The results in the following sections are based on 589 target scores and about 116,000 impostor scores under given acoustic conditions and a total number of 3534 target scores and about 696,000 impostor scores. The classification numbers are similar to the target scores numbers.

Since the image method presents a simplistic model of an empty shoe-box room, the room impulse responses for the test segments were measured from real rooms with *RT* of 0.138, 0.446, 0.491, 0.808 and 1.094 seconds. These were used for filtering the test speech segments, while the target models employed the simulated room impulse responses.

5. RESULTS AND DISCUSSION

Using the database described above, initial results show degradation with the baseline system (models trained on clean speech only and no score normalization) from equal error rate (EER) of 6.79% with the clean test segments, up to 18.32% for reverberant test segments (table II) which verifies the need for a more robust system. The classification results and then the complete speaker recognition results are presented in the reminder of this section.

Table I: Confusion matrix for reverberation classification using RBMs. In A, the reverberation is sorted by RT. B is the same table as A only sorted by frame definition.

A

		Background model				
test	**Rev time[s]**	**clean**	**0.238**	**0.513**	**0.863**	**1.111**
	clean	99.56	0.15	0.29	0	0
	0.138	98.69	0.58	0.73	0	0
	0.446	0	4.52	91.55	0.58	3.35
	0.491	7	79.45	12.24	1.17	0.15
	0.808	0	11.95	69.1	18.08	0.87
	1.094	6.27	63.41	28.86	1.17	0.29
Frame definition[%]		100	75.702	59.333	35.329	23.26

B

	Rev time[s]	**clean**	**0.238**	**0.513**	**0.863**	**1.111**
100	**clean**	99.56	0.15	0.29	0	0
98.96	**0.138**	98.69	0.58	0.73	0	0
82.81	**0.491**	7	79.45	12.24	1.17	0.15
81.96	**1.094**	6.27	63.41	28.86	1.17	0.29
63.18	**0.446**	0	4.52	91.55	0.58	3.35
59.77	**0.808**	0	11.95	69.1	18.08	0.87

test

5.1. Reverberation classification

Table I.A presents a confusion matrix for the reverberation classification according to *RT* with highest scores marked with gray background. Scores are less than 100% for two possible reasons. First, *RT* values are different for the test and model data. Second, test data used measured room responses while the model data used simulated ones. Furthermore, the classification does not seem to match *RT* values between test and model response. However, it does seem to better match *FD* values. This is evident by the diagonal behavior of the maximum score values in Table I.B, compared with the less diagonal behavior in Table I.A. This gives a motivation for categorizing the reverberation with *FD* rather than with *RT*. Also, a major advantage of this method is the very low rate of clean test segments which were classified as reverberant (0.44%).

5.2. Complete speaker recognition system

In this section the method of acoustic model matching is combined with the RBM classification and score normalization. In addition to the background models scores, the classification was also performed manually by matching test and model data according to best fits *RT* and *FD* values. Table II summarizes the results for all experiments. Rev in the test column refers to test segments with various reverberation conditions. Rev on the train column refers to

Table II. EER [%] results for various experiments

test	train	matcing	No-Norm	ZT-Norm	HT-Norm	TopT-Norm
Clean	Clean	-	6.79	4.07	-	3.74
Rev	Clean	-	18.32	16.44	-	24.23
Rev	Rev	*RT*	18.96	9.9	9.79	10.45
Rev	Rev	*FD*	19.27	9.08	9.76	9.68
Rev	Rev	RBM	19.36	8.97	9.93	9.6

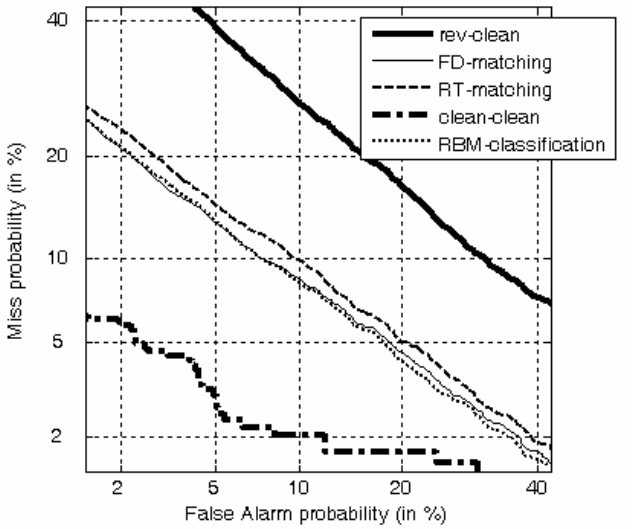

Fig. 2. DET curves for reverberant test segments on various situations

reverberant target models while the matching was performed according to the reverberation time (*RT*), frame definition (*FD*) and with the RBM classification. Various score normalization methods were employed for each experiment. The high EER results for model matching without score normalization, as shown in Table II, are due to the five different RBM biases. The score normalization methods eliminate these biases and reduce significantly the EER. Note that score normalization without model matching does not improve significantly the results either. It is the combination of model matching with score normalization which produces a significant improvement.

EER with *FD* matching is a lower then EER with *RT* matching, showing that *RT* might not be the best parameter in this case. The RBM classification which was consistent with the *FD* classification reduces the EER even further with the use of ZT-Norm. Moreover, the *RT* and *FD* matching are based on the assumption that the reverberation parameters of the test segment are known, an assumption which does not hold in many cases and is not required for the RBM classification. Note that for clean test segments, the Top-T-norm was superior.

Fig. 2. presents DET (detection error trade-off) curves for various conditions: Reverberant test segments with matching according to *RT*, according to *FD* and with the use of the RBM classification. Also, the DET curves for clean-clean (as a reference) and rev-clean (no model matching) are presented. For each condition the ZT-Norm was performed. The DET curves for matching via background models and via frame definition is similar, showing significant improvement over the rev-clean baseline system.

6. CONCLUSION

This paper investigated the effect of reverberation on speaker recognition, and proposed and evaluated methods for improving recognition robustness in the face of reverberation. We showed that significant degradation in performance arise due to reverberation. The use of acoustic model matching and score normalization was found useful in improving verification scores. The background model scores were proved to be successful for reverberation classification. Also, we found that the reverberation effects have higher correlation with definition than with the reverberation time. Another important conclusion is the useful utilization of simulated room impulse responses for target model creation and the development group. The simulated room impulse responses are easier to generate, and could be more practical than the measured ones in a real application.

REFERENCES:

[1] F. Bimbot, J-F Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. on App. Sig. Proc.*, vol. 4, pp. 430-451, 2004.

[2] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust Speaker Recognition," *IEEE Sig. Proc. Mag.* Vol. 13, No. 5, pp. 58-71, Sept. 1996

[3] P. J. Castellano, S. Sridharan and D. Cole, "Speaker Recognition in Reverberant Enclosures", *in Proc. Int. Conf. Acoust., Speech, Sig. Proc.* Vol. 1 pp. 117-120, May 1996.

[4] J. S. Gammal, R. A. Goubran, "Combating Reverberation in Speaker Verification", *Instrumentation and Measurement, Technology Con*. Ottawa, Canada, May 2005

[5] Q. Jin, T. Schultz and A. Waibel, "Far-Field Speaker Recognition", *IEEE Trans. Audio Speech Lang. Process*., vol. 15, no.7, pp. 2023-2032, Sept. 2007

[6] T. Nakatani and M. Miyoshi, "Blind Dereverberation of Single Channel Speech Signal Based on Harmonic Structure," *in Proc. IEEE Int. Conf. Acoustic. Speech, Sig. Proc*. Hong Kong, China, vol. 1, Apr. 2003, pp. 92-95

[7] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust Speaker Recognition Using Microphone Arrays," *in Proc. 2001: A Speaker Odyssey,* pp. 101-106, June 2001.

[8] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," *proc. of Int. Conf. Acoustics, Speech, and Sig. Proc*. 2003.

[9] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[10] R. Teunen, B. Shahshahani, and L. Heck, "A Model-Based Transformational Approach to Robust Speaker Recognition*," Int. Conf. Spoken Lang. Proc*. Oct. 2000.

[11] D.A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Sig. Proc.* vol. 10, No. 1-3, pp. 19-41, July 2000.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification System," *Digital Sig. Proc*. vol. 10, no. 1, 2000.

[13] Y. Zigel and M. Wasserblat, "How to Deal with Multiple Targets in Speaker Identification Systems?," *2006 IEEE Odyssey – The Speaker and Language Recognition Workshop*, Puerto-Rico, 2006.

[14] H. Kuttruff, *Room acoustics*, New York, John Wiley & Sons, 2000.

[15] M. R. Schroeder, "New Method for Measuring Reverberation Time," *J. Acoust. Soc. Am.* Vol. 37, No. 3, pp. 409-412, 1965

[16] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.* vol. 65, no. 4, pp. 943-950, 1979