SPEAKER IDENTIFICATION IN ROOM REVERBERATION USING GMM-UBM

Aditi Akula, Vijendra Raj Apsingekar and Phillip L. De Leon

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico USA 88003
Phone: +1 (575) 646-3771, {aditi, vijendra, pdeleon}@nmsu.edu

ABSTRACT

Speaker recognition systems tend to degrade if the training and testing conditions differ significantly. Such situations may arise due to the use of different microphones, telephone and mobile handsets or different acoustic conditions. Recently, the effect of the room acoustics on speaker identification (SI) has been investigated and it has been shown that a loss in accuracy results when using clean training and reverberated testing signals. Various techniques like dereverberation, use of multiple microphones, compensations have been proposed to minimize/alleviate the mismatch thereby increasing the SI accuracies. In this paper, we propose to use a Gaussian mixture model-Universal background model (GMM-UBM), with the multiple speaker model approach previously proposed, to compensate for the acoustical mismatch. By using this approach, the SI accuracies have improved over the conventional GMM based SI systems in the presence of room reverberation.

Index Terms— Speaker recognition, Identification

1. INTRODUCTION

Speech is meant to carry the linguistic messages between humans for communication. In addition, it also contains unique characteristics of the speaker which are useful in speaker recognition. Speaker recognition can be broadly divided into two categories : speaker identification (SI) and speaker verification (SV). The objective of speaker *identification* (SI) is to determine which voice sample from a set of known voice samples best matches the characteristics of an unknown input voice sample [1]. The objective of SV is to verify the identity claim of a speaker.

SI is a two-stage procedure consisting of training and testing. In the training stage, speaker-dependent feature vectors are extracted from a training speech signal and a speaker model is built for each speaker's feature set. In the testing stage, feature vectors are extracted from a test signal (speaker unknown) and are scored against all speaker models and the most likely speaker identity is decided. The performance of an SI system is measured by the identification accuracy defined as the ratio of number of correctly identified tests to the total number of accuracy tests.

The accuracies for SI systems have been shown to degrade in mismatched conditions when the training signals are acquired in clean environment and testing signals in reverberant and noisy environment [2]-[7]. Such conditions occur in applications like an access control systems, security access to buildings, vehicles etc as they require hands-free sound capture [8], [9]. In such surroundings, there is some distance between the source (user) and the microphone possibly resulting in reverberation. In general, there are at least two approaches to deal with the reverberant conditions: 1) apply compensation techniques on the test signal 2) modify the training signals by distorting them in order to match them to the test signal [6], [3]. Many researchers have used the first approach to compensate the test signal with techniques like cepstral mean subtraction (CMS) [10]-[12]; Relative spectral transform (RASTA) [13] and Root MFCC [14].

In [3], the authors addressed the problem of reverberant environment for SV. To combat the effects of reverberation, they propose to train the system with reverberant speech originating from rooms different than those of the test speech. In this work, each speaker generates several training models using an auto-regressive (AR) vector method. The authors build reverberation classification models (RCMs) for a random speaker and use the Itakura distance between the RCM and the test utterance to find the training room that best matches the test reverberation; speaker models using this training room are then used in the test stage. A classification accuracy of 96.5% is reported using KING corpus.

In [4], the authors considered speaker recognition in the far-field microphone situation. They considered teleconferencing rooms where several distant microphones were used to create multichannel signals. Speech recorded with distant microphones is prone to reverberation and additive background noise. They proposed new methods for reverberation compensation and feature warping of both training and test signals in order to improve SI accuracy in the reverberant environment. For reverberation compensation, reverberation is modeled as an additive noise and noise reduction techniques like spectral subtraction are applied followed by empirical estimation of noise parameters. CMS is applied after spectral subtraction.

Three distant microphone databases which differed in the microphone positioning, room characteristics and speaking style were used [4]. The authors used the data from the multiple microphones to do multiple channel combination experiments in order to compensate for the mismatch and reported up to 87.1% *relative* improvement when using the Distant Microphone database [4]. One drawback with this work is the requirement of multiple training signals acquired in reverberant environments.

The effect of room reverberation on SI was also addressed in [5]. The authors proposed to modify the training signals in order to mitigate the mismatch with the reverberant test signals. Similar to the work in [15], a *family* of reverberation filters were generated using a set of training rooms that mimic the test room. Each speaker's clean training signals were filtered with the family of reverberation filters to generate reverberated training signals. These reverberated training signals were used to create a set of speaker models for each speaker, one for each room. The test signal was then scored against all the speaker models from all the rooms.

Using this approach, the authors demonstrated that SI accuracies can improved by 20% over baseline levels (clean training signals, reverberated test signals) [5]. Impulse responses for the simulated training rooms is generated using the image method[16]. Unfortunately, with this method, phase is not properly considered in that all the reflections are assumed to arrive in-phase to the microphone which is not the case in real rooms. In addition, the reflection coefficients used in the image method for generating the training rooms were in a narrow range which does not allow for investigation of higher levels of reverberation.

Another paper on reverberation matching is presented in [7] and the authors investigated the effect of reverberation on SV. Particularly the authors study the methods of acoustic model matching and score normalization. They have extended the results of [4], [3] to an SV system employing the widely used adaptive Gaussian mixture model (AGMM) or commonly called as Gaussian mixture model-Universal background model (GMM-UBM). For acoustic matching of speaker models, several models were generated for each speaker under various reverberation conditions/reverberation times (RT's) in the training stage. A reverberation background model (RBM) is built for each RT by training the speech segments from various speakers filtered through a simulated impulse response. These RBM's are used for reverberation classification.

The speaker models are then adapted from the RBM and speaker reverberant speech signal. There are several target models for each speaker, each for a different reverberation and the correct target model for classification is selected by scoring the unknown test segment against all the RBM's and then selecting the RBM that gives the highest score. This is referred to as RBM classification. In [7], the authors used the NIST-99 speaker recognition evaluation corpus. The authors reported an EER of 6.79% for the clean speech segments and an EER of 18.32% for reverberant test segments. Various normalizations were applied on these scores and the EER of reverberant segments reduced to 8.97%.

In this paper, we propose a method based on Gaussian mixture model-Universal background model (GMM-UBM), which utilizes reverberation filters in the training stage, to reduce the mismatch between the training and testing signals by modifying the training signals. By building a GMM-UBM, there is a tighter coupling between the UBM and the speaker model which helps in identifying the speaker correctly.

2. MULTIPLE SPEAKER MODEL APPROACH

Multiple speaker model approach has been proposed in [5], [6]. In [6], the authors used the similar approach to that described in [5]. They differed in the way of generating the reverberation filters by using the modified image method [17]. Also they used higher levels of reverberation than those used in [5] to generate the training room impulse responses. Simulations were performed on GMM based system and the authors reported a 13% increase in accuracy over baseline (when clean training signals are used) using simulated rooms and a 12% increase in accuracy over baseline using a real room. Though for higher levels of reverberation i.e. for reflection coefficients over 0.8 which is the case in real rooms, this method does not perform as expected.

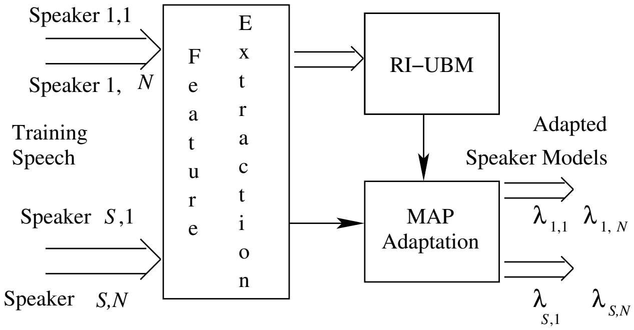
In this paper we extend the research done in [6] by implementing the multiple speaker model approach on the GMM-UBM based SI system and by using the reflection coefficients greater than 0.8 to model more realistic levels of reverberation. We generated the impulse responses of the training rooms using the modified image method [17] and convolved the clean training signals with these synthetic room impulse responses to simulate reverberation. We propose two configurations to train our SI system and associated two test stage techniques for testing the system.

2.1. Training the SI system: Configuration #1

In the first configuration, the clean training signals are filtered through these set of impulse responses, as shown in Fig. 1. A speaker-independent and room-independent UBM (RI-UBM) is built by concatenating the feature vectors of all the speakers from all the rooms. Once the RI-UBM is built, each speaker model is adapted by using the feature vectors associated with that speaker from each room. Thus each speaker has a speaker model associated with a particular training room. Thus if there are S speakers and N rooms then there are $S \cdot N$ speaker models.

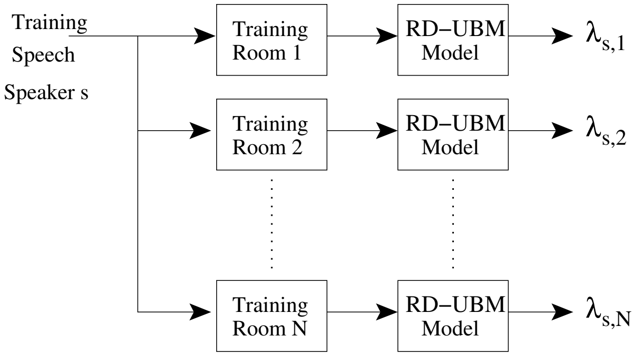
2.2. Training the SI system: Configuration #2

In the second configuration, similar to the configuration #1 a RI-UBM is built, then a room-dependent UBM (RD-UBM) is adapted from the RI-UBM by concatenating the feature vectors of all the speakers from each room. Individual speaker models are then built by adapting these RD-UBMs as illustrated in Fig. 2.



(a)

Fig. 1. Training stage using Multiple Speaker model approach, Configuration #1



(a)

Fig. 2. Speaker adaptation in Multiple speaker model approach used in training stage, Configuration #2

2.3. Testing the SI system: Configuration #1

Fig. 3 illustrates the test stage for Configuration #1. Here when a test utterance is acquired from an unknown room and unknown speaker, feature vectors are extracted and scored against all the $S \cdot N$ speaker models. The speaker identity is then identified as

$$\hat{s} = \arg \max_{1\leq s\leq S, 1\leq n\leq N} \sum_{m=1}^{M'} \log p(\mathbf{x}_m^{\text{test}}|\lambda_{s,n}) \qquad (1)$$

where $\lambda_{s,n}$ denotes the adapted GMM-UBM parameters for speaker s using training room n, where N is the number of training rooms and S is the number of speakers in the database.

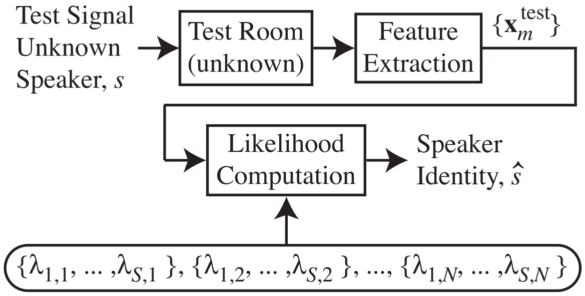

Test Signal Unknown Speaker, s
Test Room (unknown)
Feature Extraction
$\{\mathbf{x}_m^{\text{test}}\}$
Likelihood Computation
Speaker Identity, $\hat{s}$
$\{\lambda_{1,1}, \ldots, \lambda_{S,1}\}, \{\lambda_{1,2}, \ldots, \lambda_{S,2}\}, \ldots, \{\lambda_{1,N}, \ldots, \lambda_{S,N}\}$

Fig. 3. Testing stage using multiple speaker models for each speaker in Configuration #1.

2.4. Testing the SI system: Configuration #2

In the test stage of Configuration #2, the testing is carried in two steps as illustrated in Fig. 4. First, the test feature vectors are scored against the RD-UBMs to identify the closest room to the test signal with highest likelihood score. In the second step, only the speaker models associated with the highest scoring RD-UBM's are searched to identify the speaker. The advantage of this technique is that we need to score against the speaker models from only one room instead of N rooms. This can be seen as a gain in terms of computational speedups. To trade-off any loss in accuracy additional rooms can be searched, sorted according to the scores against the RD-UBMs. In our experiments we searched the two highest scoring rooms.

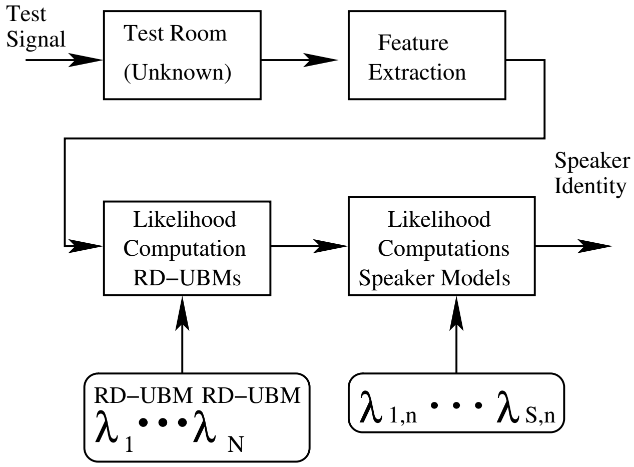

Test Signal
Test Room (Unknown)
Feature Extraction
Likelihood Computation RD–UBMs
Likelihood Computations Speaker Models
Speaker Identity
RD–UBM RD–UBM
$\lambda_1 \cdots \lambda_N$
$\lambda_{1,n} \cdots \lambda_{S,n}$

Fig. 4. Testing stage using multiple speaker models for Configuration #2.

Table 1. Results on TIMIT.

Test Room	Baseline	Configuration #1	Configuration #2
Small Office	74%	93%	78%
Student Lounge	74%	81%	78%
Conference Room	62%	86%	80%

Table 2. Results on NIST-2002.

Test Room	Baseline	Configuration #1	Configuration #2
Small Office	24%	40%	54%
Student Lounge	22%	40%	44%
Conference Room	19%	32%	35%

3. EXPERIMENTS AND RESULTS

We performed our experiments on first 100 speakers of TIMIT and NIST-2002 speech corpora. The TIMIT speech signals are recorded in a laboratory environment. NIST-2002 speech signals were recorded via cellular channels but are referred to as *clean* as they do not include reverberation. To demonstrate the applicability of the methods proposed in Section 2 to a wide variety of GMM-UBM based SI systems, we have added to this system some additional elements such as delta MFCCs, cepstral mean subtraction (CMS) and RASTA processing depending on the corpus being used. Specifically, our baseline system uses an energy-based voice activity detector to remove silence; feature vectors composed of 29 MFCCs for TIMIT and 13 MFCCs + 13 delta MFCCs for NIST-2002 extracted every 10 ms using a 25 ms window; CMS and RASTA processing on NIST-2002; and $W = 1024$ component densities for the GMM-UBMs. For TIMIT, we use approximately 24s training signals and 6s test signals and for NIST-2002 (one speaker detection cellular task) we use approximately 90s training signals and 30s test signals. Our results with clean training and clean testing on TIMIT corpus is 100% and on NIST-2002 corpus is XXX%.

We used three real room impulse responses to filter our test speech and obtain reverberated test signals: a small office, student lounge and conference room, all the rooms are on the New Mexico State University campus. Our baseline results with clean training and reverberant testing with TIMIT corpus are 74%, 74% and 62% and with NIST-2002 corpus are 24%, 22% and 19% on small office, student lounge and conference room respectively.

To each associated real room, six synthetic rooms are chosen, whose dimensions are similar to that of the real rooms but not same with varying reflection coefficients. We chose six rooms in order to balance between the accuracy and the computational overhead. The synthetic room impulse responses are generated using the modified image method described in [17] with reflection coefficients ranging from 0.80 to 0.92. The reflection coefficients were chosen in order to match the reverberation time (RT) of the synthetic room to that of the test room. For example, the RT of the student lounge is 0.65s and that of the synthetic rooms is 0.51s for a reflection coefficient of 0.85 and 0.69s for a reflection coefficient of 0.9. The real room impulse responses are measured using Cool Edit pro 2.0, a digital audio editor, with Aurora plug-in. The clean training signals are filtered through the various training room impulse responses to create a *family* of reverberant training signals and clean test signals are filtered through the real room impulse responses to get the reverberant test signals. These reverberated training and test signals are then used with the Configuration #1 and Configuration #2 as described in Section 2. The real room impulse response for a study lounge and a simulated impulse response for the synthetic room used in its training stage are shown in Fig 5.

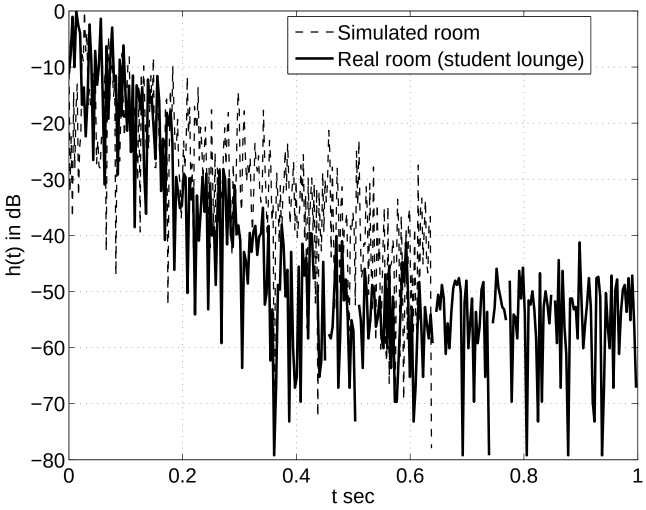


Fig. 5. Impulse response of Student lounge and a simulated room used in the training stage.

Tables 1 and 2 compare the accuracies of two configurations against the baseline. Baseline accuracies are calculated when the training signals are unreverberated and test signals are reverberated. Table 1 shows the improvement in accuracies using the multiple speaker model approach. Here configuration #1 is working better than configuration #2. In small office there is 19% improvement in accuracy over baseline and an improvement of 7% and 24% over baseline on student lounge and conference room respectively.

Table 2 shows the improvement in accuracies using the multiple speaker model approach over baseline on NIST-2002 speech corpora. There is an improvement of 30% in accuracies over baseline in small office and an improvement of 22% and 16% in student lounge and conference room respectively.

4. CONCLUSIONS

We have extended multiple speaker model approach to GMM-UBM based SI systems and illustrated the potential advantage of using the same in improving the accuracies of the SI system. We introduced two training configurations and associated test configurations. We demonstrated our approach on two different speech corpora and on three real rooms, we could achieve a gain in accuracy as high as 24% over baseline on TIMIT corpus and 30% improvement in accuracy over baseline on NIST-2002 corpus.

5. REFERENCES

[1] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Signal Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[2] P. J. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverebrant enclosures," vol. 1, pp. 117–120, May. 1996.

[3] J. S. Gammal and R. A. Goubran, "Combating reverberation in speaker verification," in *Proc. Instrumentation and Measurement Tech. Conf*, 2005.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 7, pp. 2023–2032, Sept. 2007.

[5] P. De Leon and A. Trevizo, "Speaker identification in the presence of room reverberation," *Biometric Symposium*, 2007.

[6] A. Akula and P. DeLeon, "Compensation for rooom reverberation in speaker identification," in *European. Sig. Proc. Conf (EUSIPCO)*, Aug. 2008.

[7] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. IEEE ICASSP*, 2008.

[8] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in gmm speaker recognition systems for noisyand reverberant speech with low complexity microphone arrays," in *Proc. IEEE ICSLP*, Oct 1996, vol. 3, pp. 1333–1336.

[9] I. A. Mccowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *In Proceedings of 2001: A speaker odyssey*, 2001.

[10] K. Leung, M. Mak, and S. Kung, "Applying articulatory features to telephone-based speaker verification," in *Proc. IEEE ICASSP*, 2004.

[11] M. Weintraub F. Beaufays, "Model transformation for robust speaker recognition from telephone data," in *Proc. IEEE ICASSP*, 1997.

[12] H.A. Murthy, F. Beaufays, L.P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 5, pp. 554 – 568, Sep. 1999.

[13] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, pp. 578–579, Oct. 1994.

[14] J. Han, M. Han, and W. Gao, "Channel compensation for robust telephone speech recognition," in *Proc. IEEE Region 10 Conf. Speech and Image Tech. for Computing and Telecomm (TENCON)*, 1997.

[15] D. Borah and P. DeLeon, "Speaker identification in the presence of packet losses," in *Proc. IEEE DSP Workshop*, 2004.

[16] S. G. McGovern, "A model for room acoustics," 2003.

[17] E.A.P. Habets, "Room impulse response generator," 2006.