SPEAKER IDENTIFICATION IN THE PRESENCE OF ROOM REVERBERATION

Phillip L. De Leon and Audrey L. Trevizo

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico USA 88003
pdeleon@nmsu.edu, atrevizo.ee@gmail.com

ABSTRACT

Speaker identification (SI) systems based on Gaussian Mixture Models (GMMs) have demonstrated high levels of accuracy when both training and testing signals are acquired in near ideal conditions. These same systems when trained and tested with signals acquired under non-ideal channels such as telephone have been shown to have markedly lower accuracy levels. In this paper, we consider a reverberant test environment and its impact on SI. We measure the degradation in SI accuracy when the system is trained with clean signals but tested with reverberant signals. Next, we propose a method whereby training signals are first filtered with a family of reverberation filters prior to construction of speaker models; the reverberation filters are designed to approximate expected test room reverberation. Reverberant test signals are then scored against the family of speaker models and identification is made. Our research demonstrates that by approximating test room reverberation in the training signals, the channel mismatch problem can be reduced and SI accuracy increased.

1. INTRODUCTION

In speaker identification (SI) the goal is to identify the most likely speaker of an unknown voice sample while in speaker verification (SV) the goal is to validate an identity claim based on a voice sample [1]. Our research focusses on the former. SI is a two-stage procedure consisting of training and testing. In the training stage shown in Fig. 1(a), speaker-dependent feature vectors, $\mathbf{x}_m$ are extracted from a training speech signal and a speaker model, λ_s is built for each speaker's feature set. In the testing stage shown in Fig. 1(b), feature vectors $\mathbf{x}_m^{\text{test}}$ are extracted from a test signal (speaker unknown). The test feature set is compared and scored against all S speaker models and the most likely speaker identity, $\hat{s}$ is decided.

Specific details regarding the SI system we use in this work are as follows. In the feature extraction blocks in Fig. 1, L-dimensional feature vectors are constructed using mel-frequency cepstral coefficients (MFCCs) as described in [1]. In the speaker modeling block of the training stage for a SI system, a Gaussian Mixture Model (GMM) is constructed so as to probabilistically model the distribution of the speaker's training feature vectors. The GMM-based approach has shown to be very successful in accurately identifying speakers from a large population [2]. In utilizing a GMM, we assume the probability density function (pdf) for feature vector $\mathbf{x}$ given speaker model λ_s can be modeled as a weighted mixture of W Gaussian pdfs

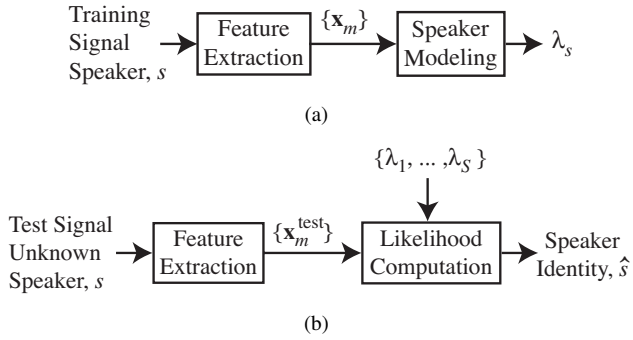

Fig. 1. (a) Training and (b) testing stages in SI

$$p(\mathbf{x}|\lambda_s) = \sum_{i=1}^{W} w_i p_i(\mathbf{x}) \quad (1)$$

where $p_i(\mathbf{x})$ is the ith component density and w_i is the associated weight. In the GMM, the weights are constrained to sum to one. The component density in (1) is an L-dimension Gaussian pdf of the form

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{L/2}|\mathbf{\Sigma}_i|^{1/2}} \times \quad (2)$$

$$\exp\left\{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right\} \quad (3)$$

where $\boldsymbol{\mu}_i$ is the mean vector and $\mathbf{\Sigma}_i$ is the covariance matrix (assumed to be diagonal) corresponding to the ith mixture component. The weights, mean vectors, and covariance matrices collectively form the speaker model, λ_s as in Fig. 1(a). A standard approach in estimating the parameters

of the GMM (weights, mean vectors, and covariance matrices) is to use the Expectation Maximization (EM) algorithm [1]. After computing all speaker models, the SI system is trained and ready for the test stage.

For the SI test stage, the likelihood computation block in Fig. 1(b) compares test feature vectors from the unknown speaker against all speaker models. Assuming equally-likely speakers and independent feature vectors, the maximum likelihood (ML) (log-likelihood) detection for identification of the unknown speaker is given by

$$\hat{s} \quad = \quad \arg \max_{1 \leq s \leq S} \sum_{m=1}^{M'} \log p(\mathbf{x}_m^{\text{test}} | \lambda_s) \qquad (4)$$

where M' is the number of test feature vectors [1]. In assessing a SI system, we measure the identification accuracy, computed as the number of correct identification tests divided by the total number of tests. Using the TIMIT corpus (630 speakers, clean speech), approximately 24 s training signals/6 s test signals, our SI system (29×1 feature vector and 32 component GMM) has 99.68% accuracy which agrees closely with that published in recent literature [3].

One well-known problem in both SI and SV is the loss of accuracy when channel distortions such as those from the telephone are present in the speech signals. For SI systems which use the NTIMIT corpus (630 speakers, telephone-quality speech), accuracy of approximately 70% (30% lower than with TIMIT) has been reported [4]. When using the NIST 1999 speaker recognition evaluation corpus (230 speakers, telephone-quality speech), SI accuracy of approximately 83% has been reported [3]. For signals which come from cellular telephones, two distortions can be present: distortions due to speech coding and distortions due to packet loss. In [5], the authors passed TIMIT signals through Global System for Mobile (GSM) speech coders and measured SI accuracy of approximately 60% (40% lower compared to TIMIT). In [6], the authors considered the problem whereby a SI system is trained with clean speech but tested with speech acquired over lossy, packet channels, such as cellular and VoIP. Accuracy levels for SI using the YOHO speech corpus were reported as 30–60% depending on the packet loss rates.

In addition to evaluating SI performance under various channel conditions, many methods have also been proposed to compensate or equalize speech signals in order to improve SI accuracy with varying degrees of success [2], [7], [8], [9], [10]. Likewise, in SV applications many methods have also been proposed including factor analysis [11]. In the majority of this work, the channels under consideration have been various microphone types, telephone and mobile channels, or VoIP.

In this paper, we consider the impact of acoustic room reverberation on SI (not SV). We specifically investigate the problem where we have access to clean training signals but have acquired the test signals in a reverberant environment. Such a mismatch between training and test signals can easily occur in audio surveillance applications where higher-quality training signals have been acquired covertly or under controlled conditions, but the test signals were acquired in an uncontrollable environment.

The paper is organized as follows. In Section 2, we propose a method whereby training signals are first filtered with a family of reverberation filters prior to construction of speaker models in order to indirectly transform speaker models; the reverberation filters are designed to approximate expected test room reverberation. In Section 3, we describe the experimental evaluation and provide results using the TIMIT corpus. In Section 4 we discuss both practical and computational aspects of the solution and in Section 5 we conclude the article.

2. INDIRECT SPEAKER MODEL TRANSFORMATIONS

Under the assumption of clean training signals but reverberant test signals there are at least two approaches one could take in dealing with the non-ideal test environment: 1) inverse filter (dereverberate) the test signal in order to undo distortions or 2) modify the training signal or speaker model in order to minimize the mismatch with the test signal. In previous SI work regarding test signals acquired in lossy, packet channels, we utilized the second approach. We found that when a test signal undergoes packet losses and we apply a packet loss model with a similar loss rate to the training data, SI accuracy can be improved from 30–60% to over 95% (YOHO corpus) [6]. This suggested that when packet loss rates of training and test signals can be matched, SI accuracy can be significantly improved. Because it is unrealistic to know in advance packet loss rates and somewhat difficult to accurately measure loss rates, training channels can only approximate test channels. Nevertheless, it was found that if a *set* of packet loss models with different loss rates was applied to the training signals in advance, thereby creating a *set* of speaker models, and the likelihood was computed over the set of models, SI accuracy could be improved even if the loss rate for the test signal was unknown [6].

Prior to publication of [6], we were not aware of any method which used *multiple* training models for each speaker in order to improve SI with non-ideal test signals. As it turns out a similar approach was proposed over a decade ago in order to address the problem of intersession variability [12], [13]. In this work, the authors assume the availability of multiple training signals for each speaker acquired in different sessions in a variety of conditions and channels. Separate models (not GMMs) are constructed for each speaker's sessions [13]. Since the work was developed prior to [1], it does not employ GMM speaker models but rather used three separate statistical models for modeling cepstra mean (Gaussian) and covariance (Wishart), and sample covariance of the differential cepstra (Wishart) [13].

For reverberant test signals (the problem under consideration), we also employ multiple speaker models. Similar to the previous work in [6], we replace a set of packet loss models with a set of reverberation filters applied to clean training signals prior to speaker modeling. With this approach, rather than compensating test signals to match training signals (speaker models), we artificially distort *training* signals to match expected distortions of test signals and thus *indirectly* transform the speaker models. As would likely be the case in an uncontrolled acoustic environment, the room impulse response that the test signal is acquired in, is unknown and therefore cannot be used with the training signals. However, we propose to use a family or set of reverberation filters which *approximate* the expected reverberation of the test room or simply span a reasonable number of typical acoustic settings. The idea is illustrated in Fig. 2. Feature extraction and speaker modeling proceed as usual except that each speaker will now have a set of speaker models (one for each room).

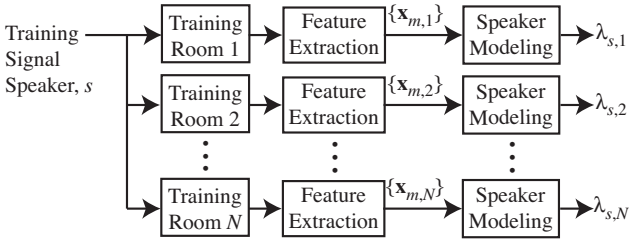

Fig. 2. Training stage whereby signals are filtered using reverberation filters which mimic various training rooms.

Under the scenario in Fig. 2, the proposed SI test stage is illustrated in Fig. 3 and shows that testing proceeds as usual except that we now conduct likelihood calculations in (4) over all speaker models

$$\hat{s} = \arg \max_{1 \le s \le S, 1 \le n \le N} \sum_{m=1}^{M'} \log p(\mathbf{x}_m^{\text{test}} | \lambda_{s,n}) \quad (5)$$

where $\lambda_{s,n}$ denotes the GMM parameters for speaker s using training room n where N is the number of training rooms. In our proposed approach we use multiple training models all derived from clean speech. We do not assume availability of multiple training sessions each acquired in a different training room to generate these models—this would impose an unrealistic burden for training the system. Rather, our multiple speaker models in Fig. 2 arise from taking a single clean training signal and filtering it with a set of filters each approximating test room acoustic conditions.

3. EXPERIMENTS AND RESULTS

We conducted three sets of experiments using the first 100 speakers from the TIMIT corpus. In order to generate reverberation filters, room impulse responses were simulated for different room sizes, source/microphone locations, and reflection coefficients using the image method [14]. This method gives good approximations to actual room impulse responses and has been used extensively in acoustic echo cancelation research [15]. The parameters for the various rooms are detailed below based on the diagram in Fig. 4. In the first set of experiments, we establish baseline results for an SI system which uses clean training signals and reverberant test signals. In the second set of experiments, we evaluate the proposed method using both reverberant training and test signals. We note that the parameters used to generate the test room reverberation filter are approximated in the training room reverberation filters as described in Section 2. In the third set of experiments, we evaluate our proposed method using actual collected data. We use test signals filtered with an impulse response measured from an actual room (office); training room reverberation filter parameters were based on actual dimensions and source/microphone locations of the test room.

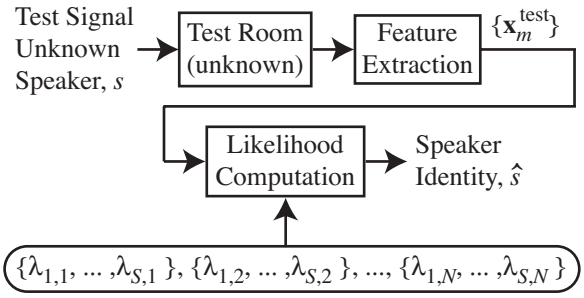

Fig. 3. Testing stage using multiple speaker models for each speaker.

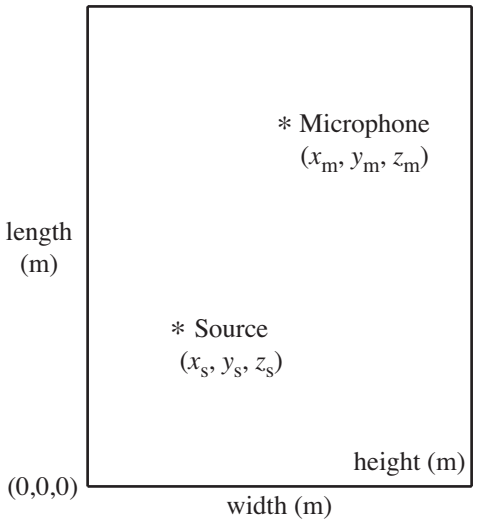

Fig. 4. Room diagram.

3.1. Experiment 1: Impact of Room Reverberation on SI Accuracy

In order to evaluate the impact of room reverberation on SI accuracy, we trained the SI system using clean speech signals and tested it using reverberant test signals. Test room dimensions used in generating the room impulse responses are given in Table 1. Test signals were filtered prior to SI in order to generate the reverberant test signals. The SI accuracy results are shown in Fig. 5.

Table 1. Experiment 1 test room parameters

Test Room	Size (w, l, h)	Source $(x_{\rm s}, y_{\rm s}, z_{\rm s})$	Microphone $(x_{\rm m}, y_{\rm m}, z_{\rm m})$
1	(4.0, 5.0, 3.0)	(2.0, 1.5, 1.5)	(2.0, 3.5, 1.5)
2	(4.0, 5.0, 3.0)	(2.0, 0.5, 1.5)	(2.0, 4.5, 1.5)
3	(4.0, 5.0, 3.0)	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)

The results demonstrate that test room geometry and acoustics can affect SI accuracy rates. Two factors—source/microphone distance and reflection coefficients—appear to degrade SI accuracy levels consistent with how increases in these factors can increase reverberation levels. The combination of these two factors is most predominant in Room 2, which has the greatest source/microphone distance and also has the largest decrease in SI accuracy as the reflection coefficient increases. Room 3 has the second greatest source/microphone distance and also has significant decreases in SI accuracy as the reflection coefficient increases. Due to the small source/microphone distance in Room 1 there are not significant levels of room reverberation and hence SI accuracy is fairly immune to reverberation even with high reflection coefficients.

3.2. Experiment 2: Evaluation of Proposed Method using Speaker Models based on Reverberant Training Rooms

For experiment 2, we conducted simulations using three different sets of reverberated test signals. We constructed training room impulse responses based on geometries similar, but not identical, to the test room in order to approximate test room acoustics. Test and training room parameters for each of the simulations are listed in Tables 2–4 where we note that all room sizes are $4.0 \times 5.0 \times 3.0$ (m) with the exception of training rooms 5 and 6 used in simulations 1 and 2 where the room sizes were $4.4 \times 5.4 \times 3.3$ (m). Clean TIMIT training signals were filtered with the various training room impulse responses and used to create a set of speaker models as in Fig. 2. Using the reverberated test signals and speaker model families as in Fig. 3, we measured SI accuracy. The results are given in the first three rows of Table 5 where we include the baseline accuracy when speaker models based on clean training signals are used (Experiment 1). Our results indicate

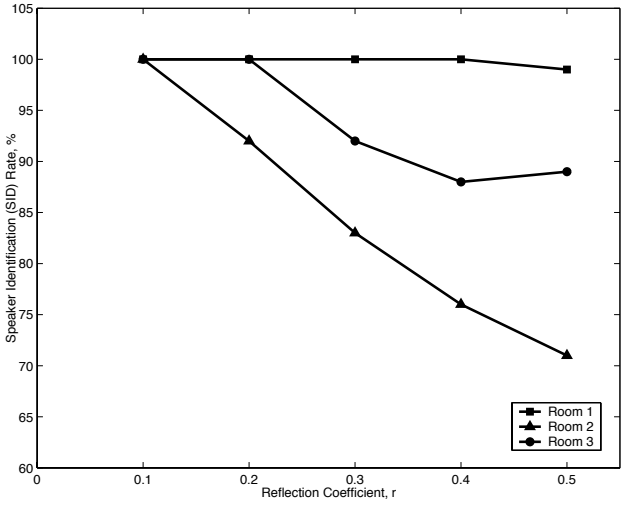

Fig. 5. SI accuracy versus reflection coefficient for test rooms in Table 1. Speaker models are based on clean training signals, test signals are reverberant.

that using the proposed method, we are able to significantly increase SI accuracy rates compared to the baseline case when speaker models based on clean training signals are used. The increase was most significant in Experiment 2, Simulation 3 (most reverberant test room) where the proposed method resulted in a 28% increase in SI accuracy.

Table 2. Experiment 2, Simulation 1 testing and training room parameters

Room	$(x_{\rm s}, y_{\rm s}, z_{\rm s})$	$(x_{\rm m}, y_{\rm m}, z_{\rm m})$	r
Test	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)	0.3
Training 1	(1.56, 1.75, 1.0)	(2.43, 3.25, 2.0)	0.3
Training 2	(1.56, 1.75, 1.0)	(2.43, 3.25, 2.0)	0.5
Training 3	(2.0, 0.5, 1.5)	(2.0, 4.5, 1.5)	0.3
Training 4	(2.0, 0.5, 1.5)	(2.0, 4.5, 1.5)	0.5
Training 5	(1.35, 1.25, 0.65)	(3.1, 4.25, 2.65)	0.3
Training 6	(1.35, 1.25, 0.65)	(3.1, 4.25, 2.65)	0.5

3.3. Experiment 3: Evaluation of Proposed Method using an Actual Test Room

For experiment 3, we use test signals which are filtered using an actual room impulse response. The test room was an office on the New Mexico State University campus (Goddard Hall, Room 171) and had a window, tiled floor, painted walls, and no furniture. We constructed training room impulse responses with similar geometries to the test room and used reflection coefficients, $r = 0.3$ and $r = 0.5$ since the actual value from the test room was not known. Test and training room sizes

Table 3. Experiment 2, Simulation 2 testing and training room parameters

Room	$(x_{\mathrm{s}}, y_{\mathrm{s}}, z_{\mathrm{s}})$	$(x_{\mathrm{m}}, y_{\mathrm{m}}, z_{\mathrm{m}})$	r
Test	(2.0, 0.5, 1.5)	(2.0, 4.5, 1.5)	0.3
Training 1	(2.0, 1.5, 1.5)	(2.0, 3.5, 1.5)	0.3
Training 2	(2.0, 1.5, 1.5)	(2.0, 3.5, 1.5)	0.5
Training 3	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)	0.3
Training 4	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)	0.5
Training 5	(2.2, 0.75, 1.65)	(2.2, 4.75, 1.65)	0.3
Training 6	(2.2, 0.75, 1.65)	(2.2, 4.75, 1.65)	0.5

Table 4. Experiment 2, Simulation 3 testing and training room parameters

Room	$(x_{\mathrm{s}}, y_{\mathrm{s}}, z_{\mathrm{s}})$	$(x_{\mathrm{m}}, y_{\mathrm{m}}, z_{\mathrm{m}})$	r
Test	(2.0, 0.5, 1.5)	(2.0, 4.5, 1.5)	0.5
Training 1	(2.0, 1.5, 1.5)	(2.0, 3.5, 1.5)	0.3
Training 2	(2.0, 1.5, 1.5)	(2.0, 3.5, 1.5)	0.5
Training 3	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)	0.3
Training 4	(1.15, 1.0, 0.5)	(2.9, 4.0, 2.5)	0.5
Training 5	(1.56, 1.75, 1.0)	(2.43, 3.25, 2.0)	0.3
Training 6	(1.56, 1.75, 1.0)	(2.43, 3.25, 2.0)	0.5

Table 5. SI accuracy using reverberated test signals with proposed method.

Experiment	Baseline Accuracy	Final Accuracy
Exp 2, Sim 1	92%	100%
Exp 2, Sim 2	83%	100%
Exp 2, Sim 3	71%	99%
Exp 3	63%	83%

Table 6. Experiment 3 testing and training room sizes

Room	(width, length, height)
Test (Office)	(4.05, 3.26, 3.69)
Training 1	(4.0, 3.0, 3.5)
Training 2	(4.0, 3.0, 3.5)
Training 3	(4.0, 3.0, 3.5)
Training 4	(4.0, 3.0, 3.5)
Training 5	(4.4, 3.3, 3.85)
Training 6	(4.4, 3.3, 3.85)

and parameters are listed in Tables 6 and 7. Clean TIMIT training signals were filtered with the various training room impulse responses and used to create a set of speaker models as in Fig. 2. Using the reverberated test signal and speaker model families as in Fig. 3, we measured SI accuracy. The result is given on the last row of Table 5 where we also include the baseline accuracy when speaker models based on clean training signals are used. Once again we note that with the proposed method and the use of an actual room impulse response, we are able to increase SI accuracy when using reverberant test signals by 20% over baseline results.

Table 7. Experiment 3 testing and training room parameters

Room	$(x_{\mathrm{s}}, y_{\mathrm{s}}, z_{\mathrm{s}})$	$(x_{\mathrm{m}}, y_{\mathrm{m}}, z_{\mathrm{m}})$	r
Office	(1.16, 2.01, 1.01)	(0.95, 2.9, 1.37)	N/A
Training 1	(2.0, 1.5, 2.0)	(2.0, 2.0, 2.87)	0.3
Training 2	(2.0, 1.5, 2.0)	(2.0, 2.0, 2.87)	0.5
Training 3	(2.0, 2.5, 1.5)	(2.0, 0.5, 1.5)	0.3
Training 4	(2.0, 2.5, 1.5)	(2.0, 0.5, 1.5)	0.5
Training 5	(1.15, 2.0, 1.0)	(1.0, 2.9, 1.4)	0.3
Training 6	(1.15, 2.0, 1.0)	(1.0, 2.9, 1.4)	0.5

4. DISCUSSION

The proposed method could easily be employed in practical settings. For example, in audio surveillance applications, although the actual test room impulse response may not be known, room dimensions and geometry are often known or can be estimated. Room impulse responses could be generated ahead of time from the known or estimated room geometry and used in building speaker models as in Fig. 2. From a computational standpoint, the proposed method requires computation and storage of N times as many speaker models where N is the number of training rooms used in the system. However, computation is done during the training stage where speed is not a concern and storage requirements for speaker models are relatively small. The proposed method also requires N times as many likelihood calculations in the test stage [compare (5) with (4)] which will increase identification time.

According to the proposed method, the speaker model for speaker s in training room n, $\lambda_{s,n}$ with the highest log-likelihood score is identified as the speaker. During simulation, the frequency with which the training rooms were associated with the identified speaker were tabulated and are given in Table 8. We note that in general training rooms which most resemble (acoustically) the testing rooms, appear to have the highest frequency of utilization. This is no surprise since we expect the training room which can minimize test room mismatch is most likely to have the highest log-likelihood score. This result may be a useful means of indirectly estimating the

geometry and acoustics of the actual test room.

Table 8. Frequency of each training room used

Room	Exp. 2 Sim. 1	Exp. 2 Sim. 2	Exp. 2 Sim. 3	Exp. 3
Training 1	4%	10%	0%	21%
Training 2	68%	90%	28%	38%
Training 3	0%	0%	0%	8%
Training 4	0%	0%	1%	10%
Training 5	22%	0%	1%	4%
Training 6	6%	0%	69%	2%

5. CONCLUSIONS

In this paper we have considered the impact of room reverberation on SI. Using clean (no reverberation) training signals and reverberated test signals, we find that SI accuracy can be degraded by as much as 30% from baseline (clean test signals) levels depending on the level of reverberation. We proposed the use of multiple speakers models for each speaker based on a clean training signal which has been filtered using a set of reverberation filters designed to approximate expected reverberation in the test room. Using both real and simulated test room impulse responses, we are able to significantly improve accuracy levels by up to 20% from baseline levels where only clean training signals are used in speaker models.

6. REFERENCES

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

[2] D. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Process. Lett.*, vol. 2, no. 3, pp. 46–48, Mar. 1995.

[3] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[4] D. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–190, 1995.

[5] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM speech coding and speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2000.

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

[7] M. W. F. Beaufays, "Model transformation for robust speaker recognition from telephone data," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1997.

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

[9] T. Ganchev, A. Tsopanoglou, N. Fakotakis, and G. Kokkinakis, "Probabilistic neural networks combined with gmms for speaker recognition over telephone channels," in *Proc. Int. Conf. Dig. Sig. Proc.*, vol. 2, 2002, pp. 1081– 1084.

[10] K. Leung, M. Mak, and S. Kung, "Applying articulatory features to telephone-based speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2004.

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1435 – 1447, May 2007.

[12] H. Gish and M. Schmidt, "Text-indepentent speaker identification," *IEEE Signal Processing Mag.*, vol. 11, no. 4, pp. 18–32, Oct 1994.

[13] H. Gish, M. Schmidt, and A. Mielke, "A robust, segmental method for text independent speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1994.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.

[15] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 1–12, Jan 2005.