COMPENSATION FOR ROOM REVERBERATION IN SPEAKER IDENTIFICATION

Aditi Akula 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, pdeleon}@nmsu.edu

ABSTRACT

Accuracy in speaker recognition systems may degrade if channel conditions during the training and testing stages are significantly different. Such channels may include different microphones, telephone and mobile handsets, speech coders, and VoIP. Many compensation techniques have been proposed which seek to minimize the channel mismatch condition thereby improving accuracy rates in these systems. More recently, the acoustic channel and its effect on speaker identification (SI) have been investigated and it has been shown that when using clean training signals and reverberated test signals, a loss in accuracy results. In this paper, we improve upon a proposed method to compensate for this acoustic channel mismatch by utilizing a more accurate room reverberation model during the training stage. This model allows us to pre-distort (reverberate) clean training signals in order to approximate the expected reverberation present in test signals. By utilizing a set of reverberated training models for each speaker, SI accuracies can be improved.

1. INTRODUCTION

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]. 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) and are scored against all S speaker models and the most likely speaker identity decided.

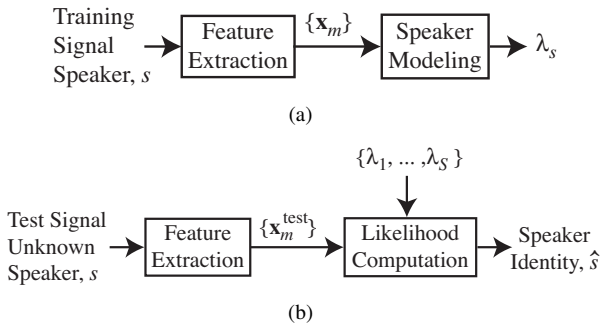

Figure 1: (a) Training and (b) testing stages in SI.

In the feature extraction blocks in Fig. 1, L-dimensional feature vectors are constructed using mel-frequency cepstral coefficients (MFCCs) as elements. In the speaker modeling block of the training stage, a Gaussian Mixture Model (GMM) is constructed and its parameters (weights, mean vectors, and covariance matrices) are estimated using the Expectation Maximization (EM) algorithm [1]. After computing all speaker models, the system is trained and ready for the test stage.

For the SI test stage, the likelihood computation block in Fig. 1(b) scores 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

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

where M' is the number of test feature vectors [1]. SI accuracy is computed as the number of correct identification tests divided by the total number of tests.

Accuracy in SI systems is known to degrade if channel conditions during the training and testing stages are significantly different. Such channel mismatches may occur when using different microphones or when different telephone, mobile, GSM coders, or VoIP channels are used for training and testing [2], [3], [4], [5], [6]. In general, there are at least two approaches one could take in dealing with the non-ideal test environment: 1) inverse filter the test signal in order to undo channel distortions or at least partially compensate or 2) modify the training signal or speaker model in order to minimize the mismatch with the test signal. Many researchers have taken the first approach and compensated or equalized the test speech signal in order to improve SI accuracy. Some of the channel compensation techniques include Cepstral Mean Subtraction (CMS) [2], [7], [8], [5]; Relative Spectral Transform (RASTA) [9]; Root MFCC (RMFCC) [10]; and Running Spectrum Filtering (RSF) and Dynamic Range Adjustment (DRA) [11], [12], [13].

The first approach has also been considered with far-field microphones in SI applications [14]. In this work, the authors consider several teleconferencing rooms where multiple, distant microphones are used to create multichannel training signals. Speech recorded with distant microphones is prone to reverberation and additive background noise. In their SI system, the authors use traditional enhancement methods and propose 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, the authors model reverberation as an additive noise and apply noise reduction techniques

like spectral subtraction followed by empirical estimation of noise paramters in [14]. Three distant microphone databases which differed in the microphone positioning, room characteristics and speaking style were used. 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 [14]. One drawback with this work is the requirement of multiple training signals acquired in reverberant environments.

In [6], the authors used the second approach for test signals acquired in lossy, packet channels (VoIP) but clean training signals. They found that SI accuracy can be significantly improved when a packet loss model with a similar loss rate is applied to the training data. Because it is unrealistic to know packet loss rates in advance, the authors applied a *set* of packet loss models each with different loss rates to the training signals thereby creating multiple models for each speaker. The log-likelihood detection is the same as in (1) except that now scoring is computed over all speakers' multiple models. The authors demonstrated that SI accuracies can improved from 30–60% baseline levels (clean training signals, VoIP test signals) to over 95% (YOHO corpus) [6]. It is worth noting that a similar approach to multiple training models was proposed over a decade ago in order to address the problem of intersession variability [15], [16]. 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 [16].

In [17], the authors addressed the problem of reverberant environment for speaker *verification* (SV). They proposed to combat the effects of test speech acquired in a reverberant environment by training 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. The authors report a classification accuracy of 96.5% on KING corpus.

In [18], the second approach (modify the training signal in order to minimize the mismatch with the test signal) was also investigated for test signals acquired in reverberated rooms. Similar to the work in [6], a *set* of simulated rooms similar to but not identical to the test room were used to generate *reverberation filters*. Each speaker's clean training signal was filtered with the reverberation filters creating a set of reverberated training signals for each speaker. The training signals were then used to create a set of speaker models for each speaker and the test signal was then scored against all of the speakers' multiple models. The advantage of this approach over the one in [14] is that we avoid the problem of having to acquire multiple training signals in reverberant environments–ours requires only a clean training signal. Using this approach, the authors demonstrated that SI accuracies can improved by 20% over baseline levels (clean training signals, reverberated test signals) [18]. In generating the impulse responses for the simulated training rooms, the image method was used [19]. 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 [18].

In this paper, we extend the previous work and utilize a more advanced algorithm to generate training room impulse responses which does properly consider phase and the echo arrival time. We also consider a wider range of reflection coefficients in our simulations and measure the reverberation time R_t, to better understand its impact on SI accuracy. Finally, we use a real test room impulse response in order to validate the method. The paper is organized as follows. In Section 2, we describe the method whereby training signals are first filtered with a family of reverberation filters prior to construction of speaker models. In Section 3, we describe the experimental evaluation and provide results using the TIMIT corpus and in Section 4 we conclude the article.

2. REVERBERATION FILTERING OF TRAINING SIGNALS

In the SI problem under investigation, we assume 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 under controlled conditions or covertly but the test signals have been acquired in an environment which cannot be controlled. Furthermore, we do not assume knowledge of the impulse response of the test room but only general information such as its approximate size and approximate positioning of the speaker and microphone. We use this general information to artificially create reverberation filters which approximate the expected test room impulse response. The reverberation filters or "training rooms" are applied to the clean training signals as 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).

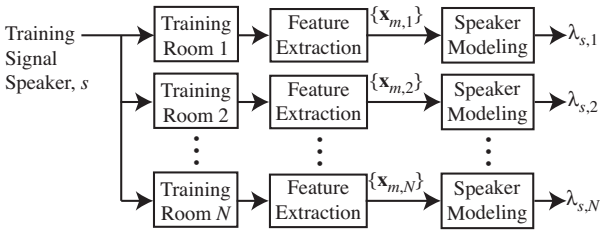

Figure 2: Training stage where signals are filtered using impulse responses of various training rooms.

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 (1) over all speaker models

$$\hat{s} = \arg \max_{1 \le s \le S, 1 \le n \le N} \sum_{m=1}^{M'} \log p(\mathbf{x}_m^{\text{test}} | \lambda_{s,n}) \qquad (2)$$

where $\lambda_{s,n}$ denotes the GMM parameters for speaker s using training room n where N is the number of training rooms.

Although in our proposed approach we use multiple training models, we do not assume the availability of multiple training sessions each acquired in a different training room to generate these models as in [14]. 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.

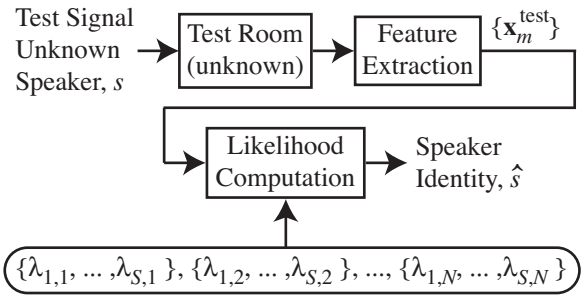

Figure 3: Testing stage using multiple speaker models for each speaker.

3. SIMULATIONS AND RESULTS

We conducted three sets of experiments using the first 100 speakers from the TIMIT corpus. We use a relatively simple GMM-based SI system in order to evaluate the proposed method. Our SI system removes silence from the speech signals using voice activity detector and uses 29 dimensional MFCC feature vectors computed every 25 ms with 10 ms overlap. We use a 32 component GMM for speaker modeling. Using clean TIMIT training and testing signals, our SI system has an accuracy of 100%.

Training room impulse responses were computed for a fixed room size, three different source/microphone locations, and several reflection coefficients using a room impulse response generator [20]. In [20], the author uses the image method described in [21] to find the room impulse response and modifies it using [22] to simulate received echo arrival time accurately. Each echo is then lowpass filtered using a Hamming window so as to add phase to the impulse response [22]. The parameters for the various rooms are detailed below and based on the diagram in Fig. 4.

3.1 SI Accuracy for Clean Training Signals and Reverberant Test Signals

We first establish baseline results for an SI system which uses clean training signals and reverberant test signals. For the baseline case, no modification of the training signal or test signal is made. To simulate test room reverberation, we filtered TIMIT test signals using a test room impulse response. The various test room impulse responses were generated using [20] for a room measuring $3.35 \times 4.27 \times 3.5$ (m), a three different source/microphone locations (see Table 1), and a range of reflection coefficients. Test room reverberation times were calculated to be range from $R_t = 44$ ms for $r = 0.3$ to $R_t = 155$ ms for $r = 0.7$ which are representative of a typical small office with furnishings.

SI accuracy results for the simulation of SI using clean training signals and reverberant test signals are shown in Fig. 5. The results confirm that the geometry and acoustics of the test rooms can affect SI accuracy rates when

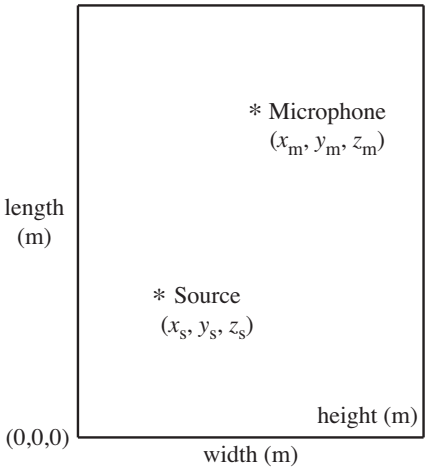

Figure 4: Diagram of the room.

Table 1: Reverberant Test room parameters

Test Room	Source (x_s, y_s, z_s)	Microphone (x_m, y_m, z_m)
1	(2.0,1.5,1.5)	(1.0,3.5,1.5)
2	(0.95,1.25,1.0)	(2.25,3.75,1.5)
3	(0.53,0.79,0.53)	(2.99,3.80,2.13)

training signals are clean, i.e. channel mismatch. Two factors, source/microphone distance and reflection coefficient, 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 3, which has the greatest source/microphone distance (and highest R_t) and also has the largest decrease in SI accuracy as the reflection coefficient increases. Room 2 has the second greatest source/microphone distance and also has significant decreases in SI accuracy as the reflection coefficient increases.

3.2 SI Accuracy using Proposed Method with Synthetic Test Room

In the second set of simulations, we evaluate the proposed method using both reverberant training and test signals where now the reverberant training signals are constructed by filtering clean training signals with reverberation filters as in Fig. 2. We constructed training room impulse responses based on geometries similar, but not identical, to the test rooms 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 training room sizes are $3.7 \times 4.7 \times 3.8$ (m) except Rooms 5 and 6 where the room sizes were $4.0 \times 5.0 \times 3.3$ (m) and the test room sizes are $3.35 \times 4.27 \times 3.5$ (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 signal and speaker model families as in Fig. 3, we measured SI accuracy.

SI accuracy results are given in the first three rows of Table 5 where we include the baseline accuracy when speaker

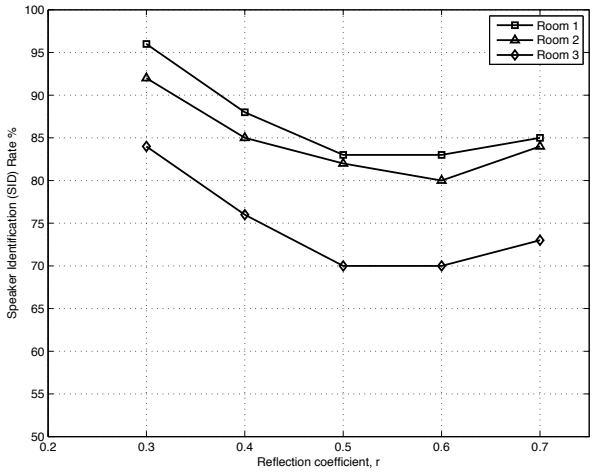

Figure 5: SI accuracy (clean training signals) versus reflection coefficient for test rooms. Baseline (no test room reverberation) accuracy is 100%.

models based on clean training signals are used. Our results indicate that using the proposed method, we are able to significantly increase SI accuracy rates compared to the baseline case where speaker models based only on clean training signals are used. The accuracy increases are approximately 13%.

Table 2: Simulation 1 testing and training room parameters

Room	(x_s, y_s, z_s)	(x_m, y_m, z_m)	r
Test	(0.75, 0.75, 0.9)	(2.85, 2,75, 1.95)	0.5
Training 1	(1.05, 1.38, 1.1)	(2.48, 4.13, 1.65)	0.4
Training 2	(1.05, 1.38, 1.1)	(2.48, 4.13, 1.65)	0.6
Training 3	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.4
Training 4	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.6
Training 5	(0.95, 1.25, 1.0)	(2.25, 3.75, 1.5)	0.4
Training 6	(0.95, 1.25, 1.0)	(2.25, 3.75, 1.5)	0.6

Table 3: Simulation 2 testing and training room parameters

Room	(x_s, y_s, z_s)	(x_m, y_m, z_m)	r
Test	(0.95, 1.25, 1.0)	(2.25, 3.75, 1.5)	0.5
Training 1	(2.0, 1.5, 1.5)	(1.0, 3.5, 1.5)	0.4
Training 2	(2.0, 1.5, 1.5)	(1.0, 3.5, 1.5)	0.6
Training 3	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.4
Training 4	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.6
Training 5	(0.75, 2.0, 0.9)	(2.85, 2.75, 1.35)	0.4
Training 6	(0.75, 2.0, 0.9)	(2.85, 2.75, 1.35)	0.6

3.3 SI Accuracy using Proposed Method with Real Test Room

In the third set of experiments, we use test signals filtered with an impulse response measured from an actual room. The test room is a study lounge on the New Mexico State University campus (Thomas and Brown, Room 102) and had

Table 4: Simulation 3 testing and training room parameters

Room	(x_s, y_s, z_s)	(x_m, y_m, z_m)	r
Test	(0.53, 0.79, 0.53)	(2.99, 3.8, 2.13)	0.5
Training 1	(1.05, 1.38, 1.1)	(2.48, 4.13, 1.65)	0.4
Training 2	(1.05, 1.38, 1.1)	(2.48, 4.13, 1.65)	0.6
Training 3	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.4
Training 4	(1.5, 1.15, 0.95)	(1.5, 3.5, 1.25)	0.6
Training 5	(0.83, 2.2, 1.0)	(3.14, 3.03, 2.15)	0.4
Training 6	(0.83, 2.2, 1.0)	(3.14, 3.03, 2.15)	0.6

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

	Baseline Accuracy	Final Accuracy
Simulation 1	83%	97%
Simulation 2	83%	96%
Simulation 3	72%	84%
Actual Room	82%	94%

a carpeted floor, painted walls, some chairs and tables, a couch and measures $10 \times 12.47 \times 2.8$ (m). We measured $R_t = 1.5$ s and the impulse response of the actual test room is shown in Fig. 6. We used [20] to construct training room impulse responses using dimensions and source/microphone locations which approximate those of the test room–all training room sizes are $9.0 \times 11.0 \times 3.5$ (m) except Rooms 5 and 6 where the room sizes were $11 \times 13.7 \times 4.0$ (m). We used reflection coefficients, $r = 0.6$ or $r = 0.8$ since the actual value from the test room was not known. The training rooms have $R_t = 0.26$ s when $r = 0.6$ and $R_t = 0.52$ s when $r = 0.8$. Finally, source and microphone locations are given in Table 6.

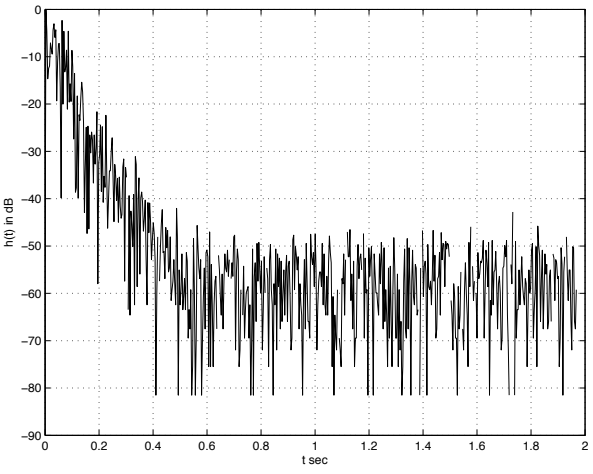

Figure 6: Impulse Response of the Actual Room.

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 mea-

Table 6: Actual Room testing and training parameters

Room	(x_s, y_s, z_s)	(x_m, y_m, z_m)	r
Office	(6.10, 4.82, 1.01)	(4.02, 6.74, 1.37)	N/A
Training 1	(6.5, 5.5, 1.0)	(3.5, 2.5, 1.5)	0.6
Training 2	(6.5, 5.5, 1.0)	(3.5, 2.5, 1.5)	0.8
Training 3	(3.0, 6.5, 1.25)	(8.0, 3.0, 1.25)	0.6
Training 4	(3.0, 6.5, 1.25)	(8.0, 3.0, 1.25)	0.8
Training 5	(2.5, 7.5, 1.35)	(9.5, 7.5, 1.75)	0.6
Training 6	(2.5, 7.5, 1.35)	(9.5, 7.5, 1.75)	0.8

sured 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. The low baseline accuracy is likely due to the high reverberation time associated with the real test room. Using the proposed method, we are able to increase SI accuracy when using reverberant test signals by 12% over baseline results. We emphasize that this accuracy improvement does not require costly acquisition of training signals in reverberant environments.

4. CONCLUSIONS

We have considered the impact of test room reverberation on SI and found that accuracy can be degraded by as much as 30% over the baseline case where test signals are clean. We have improved upon earlier work which proposed filtering clean training signals with a set of reverberation filters designed to approximate expected test room conditions. Our improvements include the use of a more refined room impulse generator for computing the training room impulse responses. The set of artificially reverberated training signals leads to a set of speaker models for each speaker and test signals are scored using the models. In our simulations, we have improved SI accuracy by 13% using simulated test rooms and 12% using an actual test room. Unlike prior work, these improvements are made without having to acquire actual training signals in reverberated environments.

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] 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.

[4] 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.

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

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

[10] 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.

[11] S. Yoshizawa and Y. Miyanaga, "Robust recognition of noisy speech and its hardware design for real time processing," *ECTI Trans. Elect. Eng., Electronics, and Communications (EEC).*, vol. 3, no. 1, pp. 36–43, Feb. 2005.

[12] N. Wada, Y. Miyanaga, N. Yoshida, and S. Yoshizawa, "A consideration about an extraction of features for isolated word speech recognition in noisy environments," in *Proc. ISPAC*, 2002.

[13] N. Hayasaka, N. Wada, and Y. Miyanaga, "Running spectrum filtering in speech recognition," *SCIS Signal Process. and Commun. with Soft Computing*, Oct. 2002.

[14] 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.

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

[16] H. Gish, M. Schmidt, and A. Mielke, "A robust, segmental method for text independent speaker identification," in *Proc. IEEE ICASSP*, 1994.

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

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

[19] S. G. McGovern. (2003) A model for room acoustics. [Online]. Available: www.2pi.us/rir.html

[20] E.A.P. Habets. (2006) Room impulse response generator. [Online]. Available: http://home.tiscali.nl/ehabets/

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

[22] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, pp. 1527–1529, Nov. 1986.