Delta-Spectral Cepstral Coefficients for Robust Speaker Recognition

Jonathan Deutsche, Xinhui Zhou, Carol Espy-Wilson

Abstract— **Most current speaker recognition systems use mel-frequency cepstral coefficients (MFCCs) in conjunction with delta-cepstral coefficients (DCCs) as their front-ends. However, speech signal can be altered significantly by channel effects, everyday background noise, and reverberation due to room acoustics. Such changes can greatly reduce the accuracy of speaker recognition systems [2]. For this reason, it is desirable to develop a more robust front-end processing system.**

Recently, it was shown that the accuracy of *speech* recognition systems can be improved by taking delta features in the spectral domain instead of in the cepstral domain. These delta-spectral cepstral coefficients (DSCCs), when used in conjunction with the MFCCs have been shown to be more robust to additive noise and reverberation as compared to the MFCC+DCCs [3]. In this project, the robustness of the MFCC+DSCCs was tested in a text-independent *speaker* recognition system according to the NIST speaker recognition evaluation core-task [4]. It was found that MFCC+DSCCs were more robust to white noise and reverberations than MFCC+DCCs when training and test data were recorded on the same channel type.

Index Terms— **Denoising, Dereverberation, Intersession Variability Compensation, Speaker Recognition, Speech Processing**

I. Introduction

Speaker recognition systems are a popular area of research due to their many applications in fields such as forensics, security, and telephone services. These systems are able to recognize a person from his or her voice by analyzing feature vectors, collections of data that convey information about a person's unique voice characteristics [2]. The features most commonly used in speaker recognition systems are mel-frequency cepstral coefficients (MFCCs). Delta-cepstral coefficients (DCCs) are often appended to the MFCCs for improved accuracy [2].

The speech data evaluated in speaker recognition systems can vary widely in recording quality. These data may have been recorded through different types of channels, such as a landline telephone, a cell phone, and a room microphone. Additionally, the data may contain different levels and types of additive noise, such as white noise, babble noise, and music. Finally, speech may be recorded in different acoustic environments with different impulse responses. Therefore, an ideal speaker recognition system, would need to be robust to channel effects, noise, and reverberation [2].

Development of more robust systems has been the primary focus of speaker recognition research. Some systems utilize signal enhancement in order to emphasize the speech components within a recording. Signal enhancement methods include voice activity detection, denoising, and dereverberation. However, these strategies add additional steps to the speaker recognition process and increase the computational load. Certain methods improve speaker recognition robustness in the front-end by modifying the feature vectors. Such approaches include feature normalization, feature warping, short-term Gaussianization, and relative spectral filtering. Additionally, some techniques improve robustness in the back-end; these include speaker model synthesis and feature mapping [2].

In this project, we seek to develop a feature extraction method which is itself more robust. In a paper by K. Kumar, C. Kim, and R. M. Stern [3], a novel set of features was proposed for more robust *speech* recognition. This set of features, called delta-spectral cepstral coefficients (DSCCs), was sought to improve recognition accuracy via performing the first delta operation in the spectral domain rather than the cepstral domain. It was shown that DSCCs were more robust to noise and reverberation than DCCs when both feature types were used alone and in conjunction with MFCCs [3].

In this report, we discuss whether MFCC+DSCCs are more robust than MFCC + DCCs when applied to *speaker* recognition, according to the NIST 2008 Speaker Recognition Evaluation Plan core-task. In order to test the MFCC+DSCCs against the MFCC+DCCs for robustness, we ran separate evaluations with white noise, babble noise, and reverberation added to the clean NIST test data.

The paper is organized as follows: in Sec. II.i, we discuss current speaker recognition systems and the features used most frequently in these systems. Additionally, we propose

Manuscript received August 1, 2011. This material is based upon work supported by the National Science Foundation under Grant No. 1063035.

Jonathan Deutsche is with the Department of Electrical Engineering and Computer Science, Loyola Marymount University, 1 LMU Dr., Los Angeles, CA 90045 USA (e-mail: jon.deutsche@gmail.com)

Dr. Xinhui Zhou is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: zxinhui2001@gmail.com).

Dr. Carol Espy-Wilson is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: espy@umd.edu)

delta-spectral cepstral coefficients as a more robust set of features. In Sec. II.ii, we briefly describe our back-end system. In Sec. II.iii, we list some of the details of the NIST 2008 Speaker Recognition Evaluation Plan, a benchmark for this particular experiment. In Sec. III, we display the results of the experiment. In Sec. IV we draw conclusions from our findings, and in Sec. V we propose future research.

II. Experimental Setup

i. DCC vs. DSCC

Speaker recognition systems consist of a front-end, in which speaker-specific features are represented as a finite set of data, and a back-end, in which speaker models are trained and test data is compared to said models [2].

Any front-end system entails transforming the speech signal into feature vectors, containing information unique to each speaker and useful for speaker recognition. Speaker-specific information includes: short-term spectral features, voice source features, spectro-temporal features such as rhythm, and high level features such as word usage. Different feature extraction systems highlight different voice characteristics [2].

Mel-frequency cepstral coefficients, or MFCCs, are some of the most popular features for speech processing. Computation of MFCCs involves many steps. After being sampled the input speech signal is pre-emphasized, multiplied by a smooth window function, and then a short-time Fourier transform is performed. The spectrum is multiplied by a mel-scale filter bank, which weights the features so as to mimic human perception of pitch. After mel-filter integration, the signal undergoes logarithmic compression and a discrete cosine transform which reduces the number of features and eliminates complex components of the signal [2].

In summary, MFCC's are obtained as follows:

$$c_n = \sum_{m=1}^{M} [\log Y(m)] \cos[\frac{\pi n}{M}(m - \frac{1}{2})]$$

Where the mel-scale filter bank consists of M-channels and is denoted as $Y(m)$, $m = 1, \ldots, M$, and n is the index of the cepstral coefficient [2].

It is to be noted that the logarithmic nonlinearity stage allows for a reduction of channel effects via cepstral mean subtraction, which recovers the original speech sequence from its convolution with the recording medium [5].

For improved accuracy, the first and second time derivatives of the MFCCs, delta and double-delta-cepstral coefficients (DCCs) are used as additional features. DCCs incorporate speaker-specific temporal information, such as formant transitions and energy modulations (rhythm) [2].

Though the addition of DCCs to MFCCs does improve recognition accuracy over the use of MFCCs alone, this improvement decreases with the addition of noise or reverberation [3].

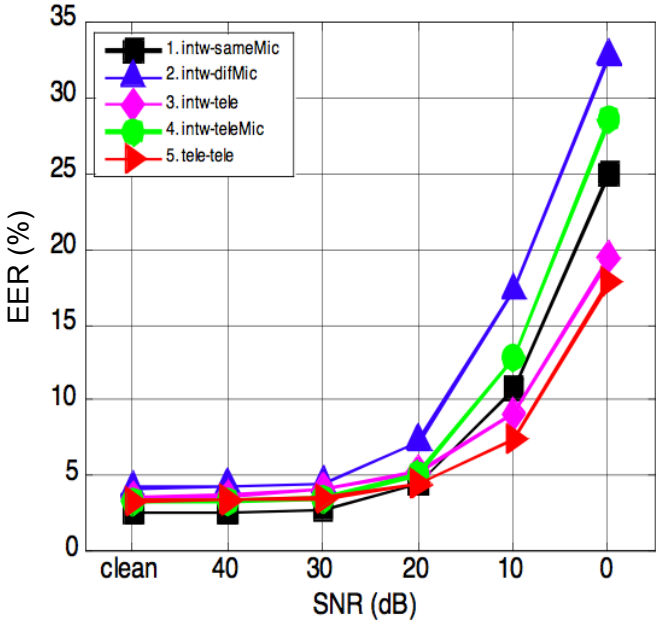
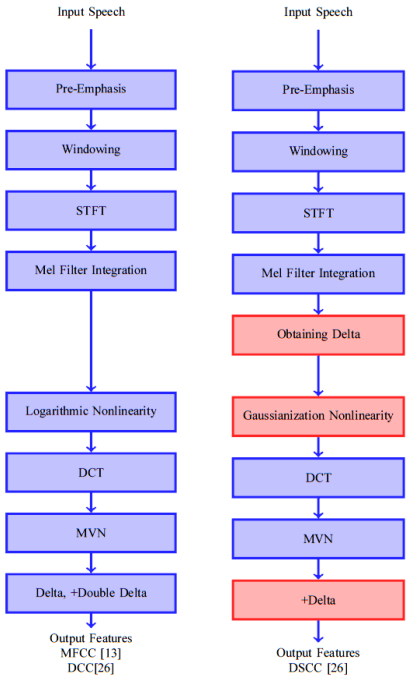


Fig. 1. Equal error rates of a speaker recognition system with an MFCC front-end in different levels of white noise [1]. Line colors indicate conditions for training and testing data according to the NIST 2008 SRE evaluation conditions: "intw-sameMic" indicates that data contained interview speech recorded on the same microphone in training and test; "tele-tele" indicates that telephone speech was used in both training and test [4].

Figure 1 displays a plot of the equal error rate of an MFCC-based speaker recognition system. As the level of white noise in the speech signal increases, the system's performance worsens (the equal error rate increases). In general, the system also suffers when there are channel mismatches in training and test data.

In our speaker recognition system, we implemented a feature extraction method developed by Kumar, et al. [3]. We appended MFCCs to delt-spectral cepstral coefficients (DSCCs) instead of DCCs. DSCCs differ from DCCs in that the first delta operation is performed immediately after mel-filter integration while the feature vector is still in the spectral domain. A second delta-operation is performed after the discrete cosine transform [3]. Figure 2 compares the two extraction processes:



(a) (b)

Fig. 2. Different extraction processes for (a) 13-dimensional MFCC features and 26-dimensional delta-cepstral coefficients (DCC), (b) 26-dimensional delta-spectral cepstral coefficients (DSCC) [3].

Note that the DSCCs use a Gaussianization nonlinearity rather than a logarithmic nonlinearity.

DSCCs were proposed over DCCs because it was shown that DSCCs were more robust to noise and reverberation than DCCs in the area of speech recognition [3]. We applied DSCCs to speaker recognition hoping to attain a more robust front-end system.

The short-time power plots in Figure 3 come from the aforementioned paper [3]:

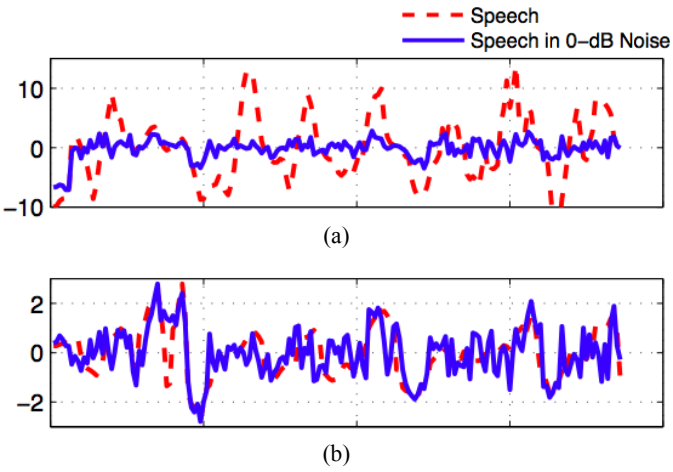


Fig. 3. Short-time power plots of a single mel-channel for (a) temporal difference over the logarithmic power of a speech signal (reflecting DCC) and for (b) the Gaussianization operation over the temporal difference of a speech signal (reflecting DSCC) [3].

Note that the lines representing the clean and noisy speech signals are more similar in the plot representing DSCCs; this similarity suggests that DSCCs should be more robust features in noise than DCCs, possibly due to the Gaussianization nonlinearity which replaces the logarithmic nonlinearity.

Figure 4 further suggests the improvement of DSCCs over DCCs, using power-plots generated from the actual feature extraction program being used in this experiment. These power-plots include all forty mel-channels.

The top plots, (a) and (b), represent the spectrogram of a clean speech signal (on the left) and a speech signal with 0 dB Gaussian white noise added (on the right). The robustness of each type of feature can be assumed by comparing the plot corresponding to the clean signal to the plot corresponding to the noisy signal. The plots for clean and noisy speech are most similar in the "Gaussianization operation over the temporal difference of speech signal," plots (g) and (h), which corresponds to the DSCC front-end. It can be expected that the DSCCs will be more robust than the DCCs, at least in Gaussian white noise, based on these power plots.

ii. Back-end System

In the back-end system, we used the standard Gaussian mixture model (GMM) approach to speaker modeling. A universal background model (UBM) was trained in each condition using the expectation-maximization algorithm. From each UBM, we adapted speaker-specific GMMs (target models) using the maximum a posteriori (MAP) method [2]. Figure 5 illustrates the MAP adaptation method.

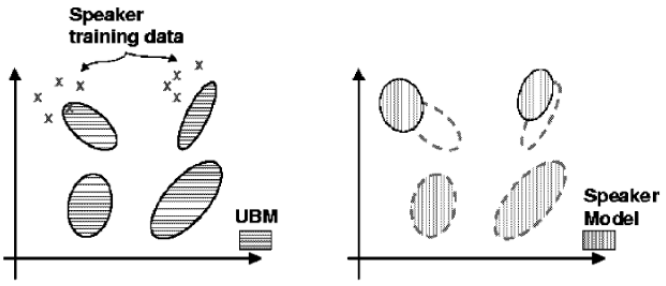
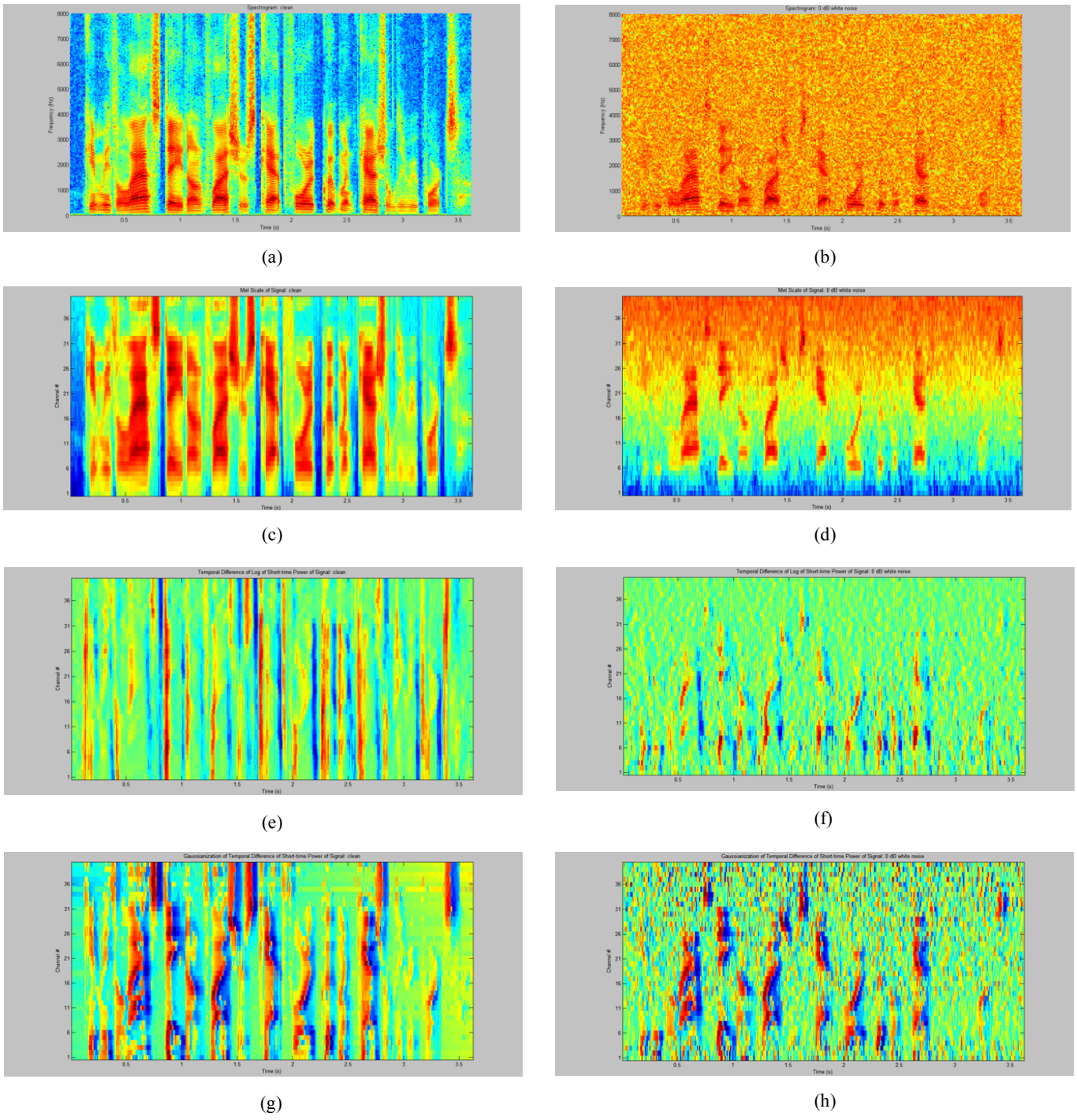


Fig. 5. GMM target model training using MAP adaptation [1].

Using the MAP method decreases the computational load; generating a UBM effectively "narrows the search" for the target model. Adapting the target model from the speaker training data and the UBM requires less time and computation than would adapting the target model from the training data alone [2].



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 4. Spectrograms and short-time power plots for clean and noisy speech signal, comparing three different types of features. (a) and (b): spectrograms for clean and noisy speech signal, respectively. (c) and (d): short-time power plots for all mel-channels of a mel-filtered speech signal, clean and noisy, respectively. (e) and (f): short-time power plots for the temporal difference operation over the logarithmic power of speech signal, clean and noisy, respectively, representing DCC's. (g) and (h): short-time power plots for the Gaussianized temporal difference of the speech signal, clean and noisy, respectively, representing DSCC's. Note that the clean and noisy plots are most similar in (g) and (h), indicating that DSCC's with Gaussianization nonlinearity should be more robust features than DCC's with logarithmic nonlinearity.

iii. NIST Speaker Recognition Evaluation Plan

All training and test data in our experiment came from the NIST 2008 Speaker Recognition Evaluation Plan. We implemented our speaker recognition system on two of the eight conditions detailed in the NIST 2008 SRE core test. These were condition 2, in which all data is interview speech and the same microphone was used in training and test, and condition three, also interview speech, in which different microphones were used in training and test [4].

We performed the speaker recognition task with two different front-end systems: using a 13-dimensional set of MFCCs appended to a 26-dimensional set of DCCs and using a 13-dimensional set of MFCCs appended to a 26-dimensional set of DSCCs. Because all test data provided by NIST were clean speech signals, we added white noise, babble noise, and reverberation in order to evaluate the robustness of each set of features. We tested both systems on clean speech as well as in different levels and types of noise and reverberation in order to see if the novel delta-spectral features showed any improvement over delta-cepstral features.

The NIST test data were evaluated in a clean condition as well as with the addition of:

- Gaussian white noise at 30, 20, 10, and 0 dB SNR
- Babble noise at 30, 20, 10, and 0 dB SNR
- Reverberation with 200, 400, 600, 800 ms RT30

Performance in each condition was evaluated based on the detection error tradeoff (DET) curve and equal error rate (EER). The DET curve plots the probability of a "miss" (the test data comes from the target speaker and is falsely rejected) vs. the probability of a false alarm (the test data comes from a non-target speaker and is falsely accepted). A more useful performance measure, the EER, is derived from the DET curves generated in the scoring stage. The EER represents the accuracy of the system at the threshold at which the probability of a miss and the probability of a false alarm are equal. A lower equal error rate indicates a better system [2].

III. Experimental Results

Figure 6 displays the overall equal error rate for MFCC+DCCs and MFCC+DSCCs for all different test conditions.

In general, MFCC+DSCCs were more robust to white noise and reverberation than MFCC+DCCs in condition 2, when training and test data were recorded on the same microphone type. MFCC+DSCCs showed no improvement over MFCC+DCCs in babble noise in condition 2 or 3. In condition 3, when there was a channel mismatch between training and test data, MFCC+DSCCs were actually less robust overall than MFCC+DCCs.

IV. Conclusions

The MFCC+DSCCs used in our feature extractor did not prove to be as robust to noise, reverberation, and channel effects as we had hoped. The feature extraction method devised by Kumar, et al. [3], was optimized for speech recognition; we likely could have improved this front-end system by modifying it to emphasize speaker-specific properties rather than speech-specific properties such as phonetic classes. Additionally, performing a Gaussianization instead of a logarithmic operation may have reduced the effectiveness of cepstral mean subtraction in the DSCC features. Cepstral mean subtraction significantly reduces channel effects; our speaker recognition system was not robust to variations in channel type. Finally, it was illustrated via short-time power plots in [3] that the delta-spectral features with Gaussianization were more robust than the delta-cepstral features in white noise. This same robustness was not illustrated for additive babble noise. It is possible that the delta-spectral features were devised to be more robust to slowly-changing noise, not necessarily to quickly-changing noise.

V. Future Work

Delta-spectral cepstral coefficients may still be utilized in a more robust front-end system for speaker recognition. The DSCC feature extractor could be modified in order to better suit a speaker recognition system. It would be practical to analyze DSCCs using a logarithmic nonlinearity rather than a Gaussianization nonlinearity, with a potential for reduction of channel effects. Also, it would be beneficial to analyze DSCCs in conjunction with other types of features, such as linear predictive cepstral coefficients (LPCCs), as it is possible that DSCCs appended to other types of features will prove more robust than DSCCs appended to MFCCs.

Acknowledgment

Jonathan Deutsche would like to thank the NSF for its support through OCI award #1063035.

References

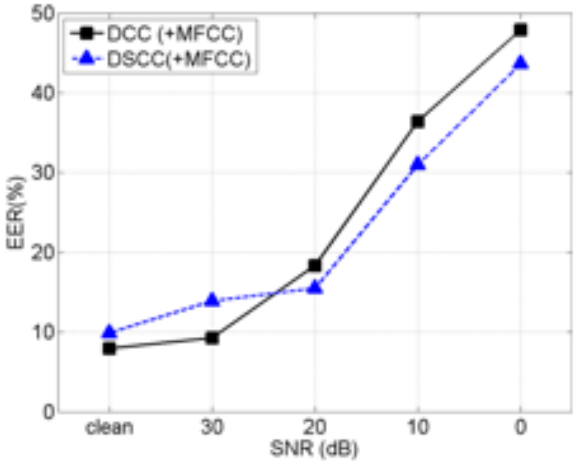
[1] D. Garcia-Romero. "Speaker Recognition Using Gaussian Mixture Models (GMMs)." Powerpoint Presentation. 2006.

[2] T. Kinnunen and H. Li. "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors." *Speech Communication* 52.1 (2010): 12-40. *ScienceDirect*.

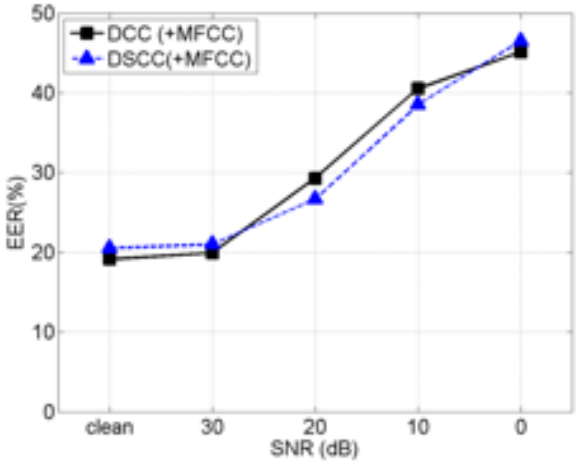
[3] K. Kumar, C. Kim, and R. M. Stern. "Delta-Spectral Cesptral Coefficients for Robust Speech Recognition." *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2011): 4784-787. *Carnegie Mellon University*.

[4] "The NIST Year 2008 Speaker Recognition Evaluation Plan."
http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf. (2008).

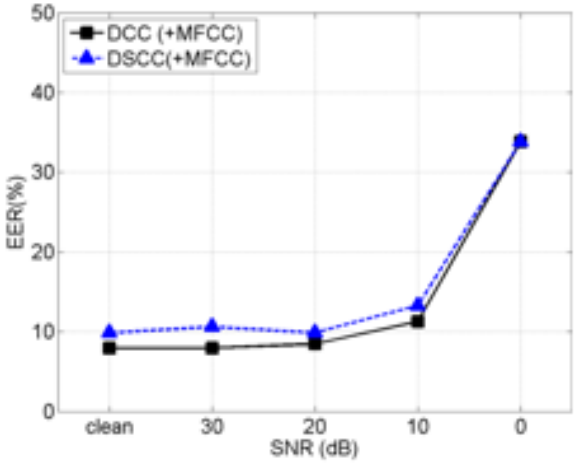
[5] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Ed. Alan V. Oppenheim. Upper Saddle River, NJ: Prentice Hall PTR, 2002.



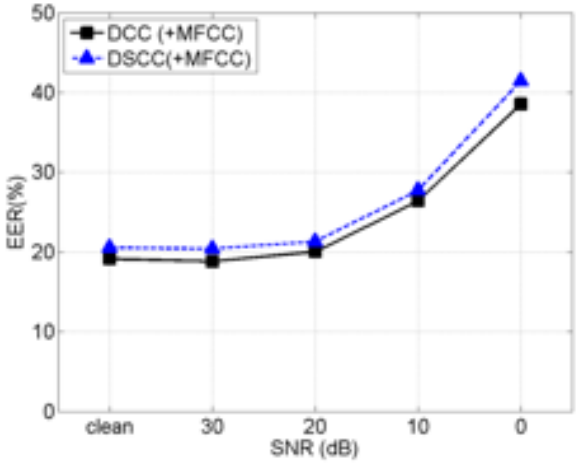
(a) EER for cond. 2 in Gaussian white noise



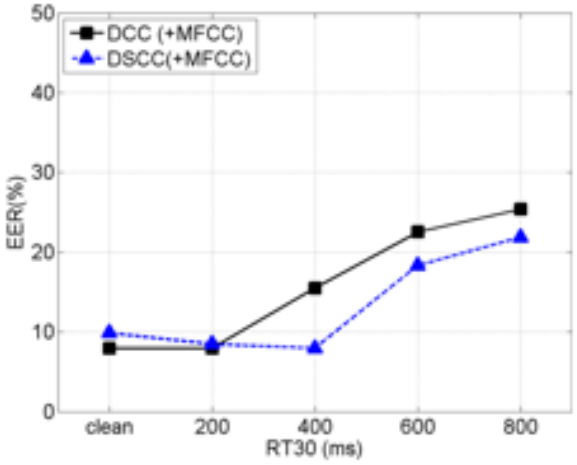
(b) EER for cond. 3 in Gaussian white noise



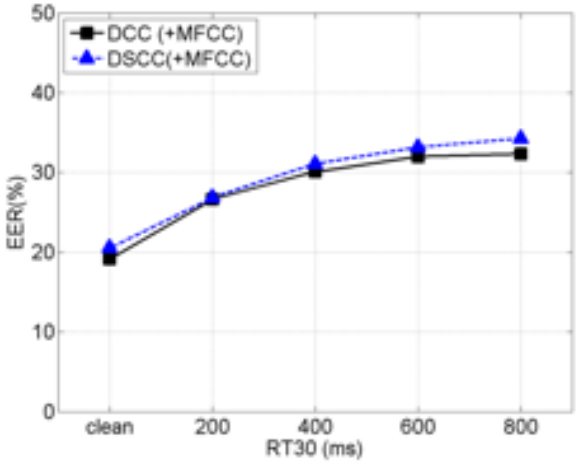
(c) EER for cond. 2 in babble noise



(d) EER for cond. 3 in babble noise



(e) EER for cond. 2 in reverberation



(f) EER for cond. 3 in reverberation

Fig. 6. Comparision of EERs. “Condition 2” indicates that all trials involved interview speech with the same microphone type used in training and test. “Condition 3” indicates that all trials involved interview speech with different microphone types used in training and test.