Significance of Speaker Information in Wideband Speech

Gayadhar Pradhan and S R Mahadeva Prasanna
Dept. of ECE, IIT Guwahati, Guwahati 781039, India
Email:{gayadhar, prasanna}@iitg.ernet.in

***Abstract*—In this work, speech signal having information up to 4 kHz is termed as narrowband (NB) speech and the other having information up to 8 kHz is termed as wideband (WB) speech. The objective is to demonstrate the significance of speaker information present in the WB speech. A speaker verification (SV) system is developed using the mel-frequency cepstral coefficients (MFCCs) computed from the WB speech and modeled using Gaussian mixture models (GMM). For comparison, a SV system is also developed from the corresponding NB speech. The experimental results show that the SV performance improves for WB speech and the improvement is significant under degraded conditions. Further, the performance improvement is better for female speakers.**

***index terms*- Wideband speech, narrowband speech, speaker information, speaker verification.**

I. Introduction

In the present work, speech collected over telephone having information up to 4 kHz and sampled at 8 kHz is termed as narrowband (NB) speech and speech having information up to 8 kHz and sampled at 16 kHz is termed as wideband (WB) speech. Most of the available speech databases, especially, for speaker recognition are collected over landline telephone or mobile phone networks resulting in NB speech [1]. This may be motivated from the availability of low cost communication networks and also possible remote person authentication as a potential application for speaker recognition. The 3 kHz (0.3 - 3.3 kHz) telephone bandwidth was initially standardized, since communication channel was a very precious component. With the progress in technology, the communication channel cost has come down drastically. Also from the human perception point of view the quality and intelligibility of WB speech is better compared to NB speech. Motivated by both these observations, recently many efforts are being made to reconstruct wideband (WB) speech having information up to 8 kHz from the NB speech [2] [3]. The third generation partnership project (3GPP) has led to the standardization of a wideband adaptive multirate (WB-AMR) codec for encoding wideband speech (50 Hz to 7 kHz) at rates from 6.6 up to 23.85 kbps [4]. Even though intuitively we feel that extended bandwidth may improve the quality and intelligibility of signal, the fundamental question is how significant it is for the speaker recognition task? The experimental work to answer this is the motivation for this work. If we have simultaneously recorded NB and WB speech signals, then performing a speaker recognition task on both the signals will help in understanding the same.

In a natural conversation, formants and harmonics may not be limited to telephonic passband (0.3-3.3 kHz) [5]. For high pitch speakers, the formants and harmonics may also be extended beyond telephonic passband. Further, the higher order harmonic structure may be different for different speakers. Thus, limiting the bandwidth of the speech signal not only looses the naturalness of the speech signal, but also some speaker information. The effect may be more severe for female speakers. The preliminary signal analysis shows that WB speech contains more information compared to NB speech. If speaker information present in the WB speech is significant, the speaker recognition system should provide robustness in practical conditions like noise, changing environments and sensors. Also, for female speakers, the performance improvement may be significant.

Most of the databases available for speaker recognition are NB speech and the databases available in WB speech are not recorded in practical conditions. Looking all these factors we have developed a 100 persons speaker recognition database in a multi-environment, multi-sensor, multilingual and multi style-condition. The database consists of simultaneous recording of speech data over multi-sensors to create WB and NB speech for the same set of speakers. A speaker verification (SV) system is then developed by processing the WB speech in using standard approach like mel-frequency cepstral coefficients (MFCC) as feature [6] and Gaussian mixture models (GMM) as the modeling technique [7]. SV systems are also developed under degraded conditions. The corresponding SV systems using NB speech are also developed. A comparative study of the respective WB and NB speaker verification systems will reveal the significance of WB speech.

The rest of the paper is organized as follows: Speech signal analysis of WB and NB speech for different sound units is described in Section II. Speaker verification system using WB speech is described in Section III. The experimental studies are described in Section IV. The experimental results are discussed in Section V. Finally, the paper is concluded in Section VI.

II. WB and NB Speech Analysis

This section reports some studies that have been conducted on the WB and NB speech signals. Different sound units are studied in the frequency domain for male and female speakers. The speech in TIMIT database is recorded at 16 kHz and the speech in NTIMIT is TIMIT speech files passed through a telephone channel and recorded at 16 kHz sampling frequency

[8] [9]. The corresponding speech files in these two corpus is the same sentence uttered by the same speaker. Therefore, these two corpus are best suited to study the difference between WB and NB speech signals. In this section, if not mentioned WB speech corresponds to TIMIT speech file and NB speech corresponds to the corresponding NTIMIT speech file down sampled by a factor of two.

A Hamming windowed 30 msec segment of unvoiced fricative $/sh/$ for a female speaker is taken from WB speech, shown in Fig. 1(a). The short term magnitude spectrum of the signal is plotted in logarithmic scale by computing the discrete Fourier transform (DFT), shown in Fig. 1(b). The corresponding portion of speech file is taken from NB speech. The DFT of the Hamming windowed NB speech in the logarithmic scale is shown in Fig. 1(c). The comparison of the two short term spectra shows that for NB speech the spectrum falls around 3.3 kHz and dies out at 4 kHz. The corresponding spectrum of WB speech extends up to 8 kHz. The Hamming windowed 30 msec segment of vowel $/a/$ for the same speaker is shown in Fig. 1(d). The short term spectra for the WB and NB speech are shown in Fig.1(e) and 1(f), respectively. The spectrum of NB speech falls around 3.3 kHz and noisy up to 4 kHz before it dies out. The corresponding spectrum of WB signal is significant up to 5 kHz. The spectrum from 3.3 to 5 kHz may contain formants and harmonics. These two illustrations show that the information in spectrum is not limited to 4 kHz and extends well beyond 4 kHz.

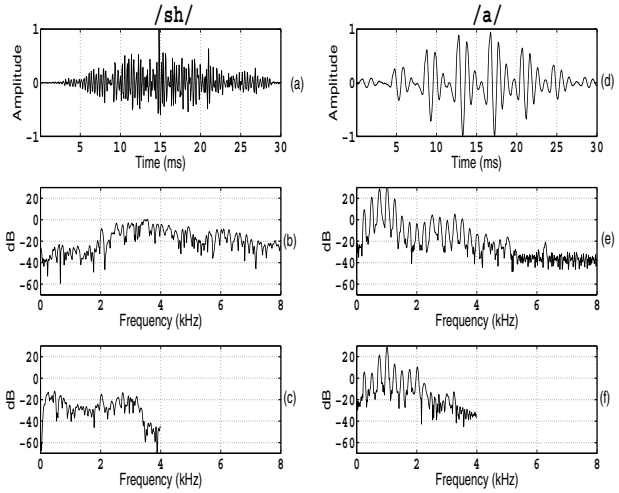

Fig. 1. Log magnitude spectra of WB and NB speech signals of a female speaker for the fricative $/sh/$ and vowel $/a/$.

A Hamming windowed 30 msec segment of fricative $/sh/$ of a male speaker taken from WB speech is shown in Fig. 2(a). The short term magnitude spectra for WB and NB speech are shown in Fig. 2(b) and 2(c), respectively. The spectrum for NB speech falls around 3.3 kHz and dies at 4 kHz. The corresponding spectrum of WB speech rises at 3.5 kHz from -20 dB and reaches around 10 dB at 4 kHz and maintains the same level up to 8 kHz. The Hamming windowed 30 msec segment of vowel $/a/$ for the same speaker is shown in Fig. 2(d). The short-term magnitude spectrum for WB and NB speech are shown in Fig. 2(e) and 2(f), respectively. The spectrum of NB speech falls around 3.3 kHz and dies out at 4 kHz. The corresponding spectrum of WB speech is significant up to 6 kHz. These illustrations indicate the WB speech characteristics are similar in case of male speakers also.

The studies show that the spectral information is present in the higher frequencies for consonants as well as vowels. The spectral magnitude for high frequency consonants is 20 dB higher in high frequency regions compared to lower frequencies. The spectral magnitude for vowel sound falls to 20 dB around 5.5 kHz. During the MFCC computation, the width of filter bank at higher frequencies is generally large compared to the lower frequency. Although the spectral magnitude falls to -20 dB compared to lower frequency for vowels, but due to large width of the filter bank, these spectral regions may contribute to the computation of cepstral coefficients.

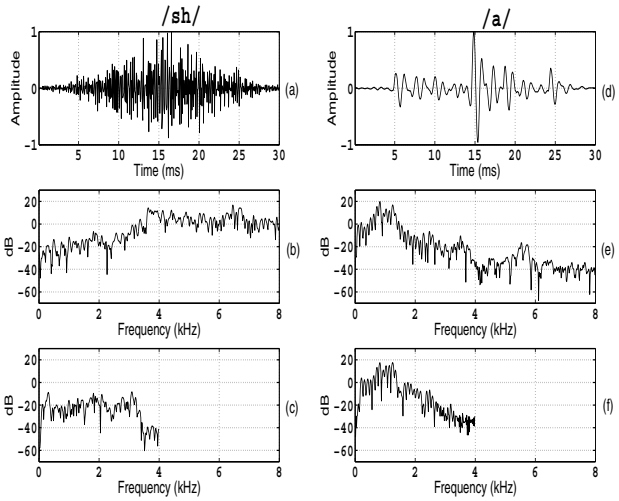

Fig. 2. Log magnitude spectra of WB and NB speech of a male speaker for fricative */sh/* and vowel */a/*.

III. Speaker Verification System using WB Speech

A. Database

We have used a subset of **IITG-DIT** Multi- Sensor- Environment- Language- Style (M4) Speaker Recognition database developed in house for these studies. IITG-DIT M4 database is collected in a setup having five different sensors, two different environments, different Indian languages and two different styles. The five different sensors include headphone microphone mounted close to the speaker, inbuilt tablet PC microphone, two mobile phones and one digital voice recorder. Except for the headphone microphone, all the other four sensors are placed at a distance of about 2-3 feet from the speaker. Speech was recorded simultaneously over these sensors. Speech recorded in headphone microphone and inbuilt tablet PC microphone are at 16 kHz and stored with 16 bits/sample resolution. Speech recorded in digital voice recorder is at 44.4 kHz and stored with 16 bits/sample, which is later resampled to 16 kHz and stored at 16 bits/sample. The

speech recorded in two mobiles are at 8 kHz and sampled at 16 bits/sample. The recording was done in two different environments, namely, office/laboratory and hostel rooms. The recording was done in two languages, namely, English and favorite language of the speaker which happens to be one of the Indian languages like Hindi, Telugu, Kannada, Oriya and so on.

B. Feature Extraction

The silence regions are removed based on energy threshold ($0.06 \times averageenergy$) of the speech file. In the training and testing process, the speech signal is processed in frames of 20 ms at 10 ms frame rate. For each 20 ms Hamming windowed frame, MFCC are calculated using 22 logarithmically spaced filters [6]. The first 13 coefficients excluding zeroth coefficient value are used as a feature vector. Delta (Δ) and delta-delta ($\Delta\Delta$) of MFCC are also computed using two preceding and two succeeding feature vectors from the current feature vector. Thus the feature vector will be of 39 dimension with 13 MFCC, 13 ΔMFCC and 13 $\Delta\Delta$MFCC.

C. Parameter normalization

The blind deconvolution like cepstral mean subtraction (CMS) reduces the performance when there is not much variability in the recording sensor and environment, and it improves the performance when there is variation [10]. In the present experimental setup for sensor mismatch experiments, there is variation in sensor and environment. In the sensor match experiments, although there is no variation in recording sensor, still there is lot of environmental variation present from training to testing session. Further, the models are built by adapting a sensor mix universal background model (UBM). Looking all these factors, in the present experimental setup the feature vectors are normalized to fit a zero mean and unit variance distribution.

D. Speaker modeling and testing

The main motivation of this work is to study the discriminating information present in the WB speech for speaker modeling and testing. Except band extension, there is no difference in the steps of speaker verification system development. Hence, the extensively used GMM-UBM based speaker modeling is employed [7]. The UBM is a large GMM which represents the speaker independent distribution of features. The UBM is generally built using large population speech. The UBM is the core part of GMM-UBM speaker verification system. The UBM should balance with respect to male and female speakers, and the speech should come from every possible sensor which will be encountered at the time of speaker verification. The UBM is represented by a weighted sum of C component densities as $U = \{\eta_c, \mu_c, \Sigma_c\},\ c = 1, ...C$, where μ_c, Σ_c and η_c are the mean vector and covariance matrix of each mixture, and weight associated with mixture c, respectively. The speaker dependent models are built by adapting the components of UBM with the speakers training speech using maximum a posteriori (MAP) algorithm [7]. During the testing stage the log likelihood scores are calculated between the claimed model and UBM.

IV. SV Experimental Studies using WB and NB Speech

Speaker verification system validates the identity claim of a person [11]. A good SV should accept all the true claims and reject all the false claims. In practical applications, some of the true trials may be rejected and some false trials may be accepted. The SV performance is measured in terms of false rejection rate (FRR) and false acceptance rate (FAR). When the FRR equals FAR, the error is termed as equal error rate (EER).

In order to compare the performance obtained using WB speech, we have developed another SV system using NB speech which is termed as *baseline system*. The NB speech used for baseline system is the original speech recorded at 16 kHz down sampled by a factor of two. The only difference between baseline system and proposed system lies in the bandwidth of speech signal. Therefore, if the proposed SV system gives better performance compared to the baseline system, the performance improvement is only due to WB speech. The robustness of WB speech for SV system is required to be tested in practical conditions like environment and sensor mismatch conditions.

For the present work, we consider 100 speakers set of IITG-DIT M4 database which include 75 male speakers and 25 female speakers. The initial 2 minutes of speech data recorded in the first session is used for building the models. For each speaker, 10 speech segments between 30-45 sec duration from the second session are taken as test utterances. Therefore for 100 speakers set, there are in total 1000 test trials. In the testing process, each test segment is tested against 11 models, out of which one is genuine model and rest are impostor models. Out of the five sensors, the speech recorded in the two mobile phones are only sampled at 8 kHz. Therefore these two sensors are not considered for present experiments. Speech recorded over digital voice recorder (D01) is worst affected by the environmental noise like air conditioner, fan sound and room reverberation due to its high sensitivity. The speech recorded in the headphone microphone (H01) is more clean compared to the other two sensors. Accordingly, the speech recorded in D01 is considered as noisy speech and speech recorded in H01 is considered as clean speech. The speech recorded in inbuilt tablet PC (T01) is not so clean as sensor H01 and not so noisy as sensor D01.

In this experimental setup ten hours of UBM speech were selected from 17 male and 17 female speakers those who are not belonging to the present 100 speakers set. This 10 hours of speech contains five hours of male speech and five hours of female speech. For each speaker, the UBM speech is distributed equally among the three sensors H01, T01 and D01. Using the sensor mixed data, two gender dependent 512 mixture size GMM are built, one for the male and other for the female speech. Finally, a 1024 mixture size gender independent UBM is built by pooling the two models and normalizing

the weights [7]. Two such gender independent sensor mixed UBMs are built one for the NB speech and another for the WB speech. During the time of model adaptation and testing, the respective UBM is used.

Keeping the language as English and conversational style, experiments are conducted on IITG-DIT M4 database as follows:

1) *Sensor matched condition*: Training and testing speech are collected over the same sensor.
2) *Sensor mismatched condition*: Training and testing speech are collected over different sensors.

Finally, the SV performance is evaluated for male and female cases separately to study the effectiveness of WB speech for each gender.

V. Experimental Results and Discussion

TABLE I
PERFORMANCE OF SPEAKER VERIFICATION SYSTEM USING NB SPEECH (BASELINE) AND WB SPEECH IN TERMS OF EER FOR MATCHED AND MISMATCH CONDITIONS.

Train Sensor	Test Sensor					
	H01		T01		D01	
	NB	WB	NB	WB	NB	WB
H01	2.96	2.55	10	7.24	15.3	14.3
T01	9.59	6.32	6.63	4.18	19.59	16.83
D01	15.51	13.16	20.81	17.55	20.4	15.61

TABLE II
PERFORMANCE OF SPEAKER VERIFICATION SYSTEM FOR MALE SPEAKERS USING NB SPEECH (BASELINE) AND WB SPEECH IN TERMS OF EER FOR MATCHED AND MISMATCH CONDITIONS.

Train Sensor	Test Sensor					
	H01		T01		D01	
	NB	WB	NB	WB	NB	WB
H01	2.15	2.04	9.49	7.34	15.06	14.1
T01	8.22	6.32	6.83	4.55	20.63	17.59
D01	14.55	13.03	21.39	17.72	21.59	16.58

TABLE III
PERFORMANCE OF SPEAKER VERIFICATION SYSTEM FOR FEMALE SPEAKERS USING NB SPEECH (BASELINE) AND WB SPEECH IN TERMS OF EER FOR MATCHED AND MISMATCH CONDITIONS.

Train Sensor	Test Sensor					
	H01		T01		D01	
	NB	WB	NB	WB	NB	WB
H01	2.63	1.05	10.52	4	11	5
T01	10.3	4.1	4.21	2.1	13.15	10.52
D01	19	14.21	17.89	14.26	13.68	11.1

A. Sensor matched conditions

In this set of experiments the trained and test speech data are collected through the same sensor. In the sensor matched speech although there is no sensor variation from trained to test speech, but in our recording setup the recording environment is captured differently in three sensors due to their position and sensitivity. The aim of this experiment is to study the usefulness of speaker discriminating information present beyond 4 kHz for SV system under different noise levels. The DET curves in Fig.3(a) show the performance of SV system using WB signal and baseline system for clean and sensor matched condition (trained and test speech collected through headphone microphone(H01)) [12]. The DET curves show that for the most favoring condition of a SV system, the WB speech gives best performance compared to NB speech.

Moving slightly towards the noisy data, the second sensor matched experiment is conducted on the inbuilt tablet PC microphone (T01) recorded speech. In our recording setup, the tablet pc is placed about 2 feet from the speaker and the inbuilt tablet PC microphone is omnidirectional in nature. The speech recorded in sensor T01 is noisy compared to sensor H01, but the complete environment is not reflected in the recorded speech. The DET curves in Fig.3(b) show that in such a condition the WB speech gives significantly better performance compared to NB speech. This result shows that for moderate noisy and environment changing condition, the speaker information at higher frequency provides robustness. The recording environment is more pronounced in the digital voice recorder (D01) recorded speech due to its high sensitivity and placement. The final sensor matched experiment is conducted on this noisy data to investigate the proposed SV performance in the degraded environments. The DET curves in Fig.3(c) shows that in a more pronounced noisy condition also, the WB speech gives significantly better performance compared to NB speech.

The above three experiments show that the improvement in the speaker verification performance using WB speech is significantly better compared to the corresponding NB speech. This is specifically true for the degraded conditions.

B. Sensor Mismatched conditions

In these experiments the trained and test speech is recorded from different sensors. Among all other factors, the SV performance is greatly affected by the sensor mismatch between training and testing sessions. In our recording setup, recorded speech is affected by sensor and recording environment. The sensor matched experiments are conducted to study the effect of environment on WB speech signal. The sensor mismatched experiments are done to study the performance of WB speaker verification system for gross mismatch conditions in terms of both sensor and also environment. The performance of SV system using NB speech and WB speech for various sensor mismatch conditions are summarized in Table. I. By comparing the EER given in Table. I, it can be seen that starting from a close sensor mismatch (sensor H01 and sensor T01) to gross mismatch (sensor T01 and sensor D01), the performance of SV system using WB speech is always better than SV system using NB speech.

All these experiments show that the WB SV system gives better performance compared to NB SV system and the performance improvement is also maintained for noisy and sensor mismatched speech.

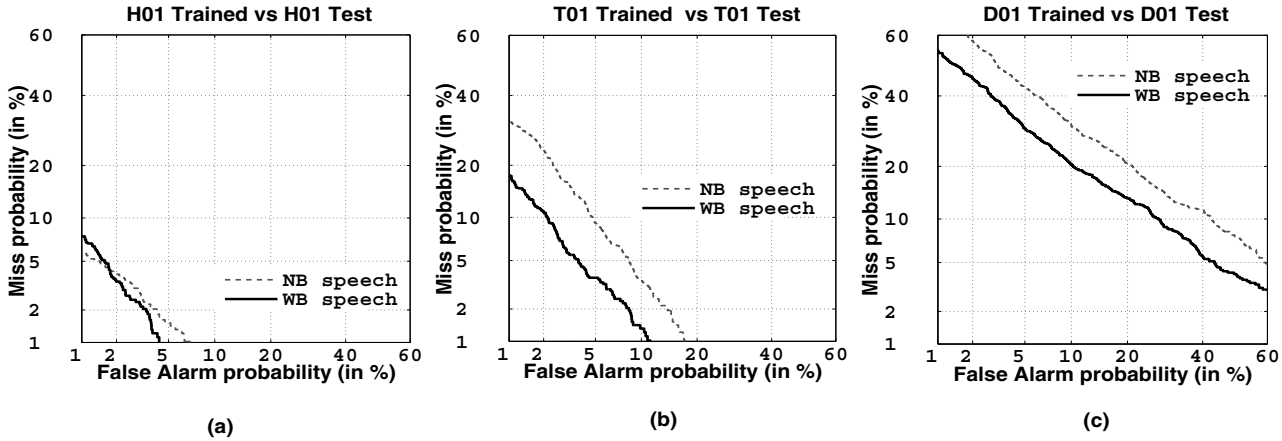

Fig. 3. Det curves for different sensor matched conditions of IITG-DIT M4 database

C. Gender dependent experiments

In this experiment, the SV performance is measured for male and female test files separately. The 100 speakers set used for the present case contains 750 test files for male speakers and 250 test files for female speakers. Although the number of test files in case of female is less, but the comparative performance of NB speech and WB speech can be studied. The performance of SV system for male test files using NB speech and WB speech is summarized in Table. II. These experimental results show that even for low pitch male speakers, the spectrum beyond 4 kHz contains speaker information.

The performance of female test files is summarized in Table. III for NB and WB speech. From the table, it can be observed that the performance improvement in WB speech is significant. The gender dependent experiments show that for male and female speakers, the SV system using WB speech gives better performance compared to NB speech. The relative performance improvement in female speakers is more compared to male speakers. This shows that the pitch and formants of female speakers may be extended beyond 4 kHz.

VI. Summary and Conclusions

In this paper we have shown the significance of wideband speech for speaker verification in different environmental conditions, like clean, noisy and sensor mismatch. The performance of speaker verification system using wideband and narrowband speech is also compared for gender dependent test files. The experimental results show that the performance of speaker verification system for wideband speech is better compared to narrowband speech. The relative improvement in performance of the proposed system compared to the baseline system is approximately same for clean, noisy and sensor mismatched speech. The performance of proposed speaker verification system is better compared to the baseline system for both male and female speakers and the improvement in performance for female speakers is significant.

This work illustrated the significance of speaker information in the wideband speech. Future work should focus on bandwidth expansion of narrowband speech signal and use the extended bandwidth for speaker recognition studies.

Acknowledgements

The authors would like to thank the Department of Information Technology (DIT), New Delhi for sponsoring this work. Special thanks to Prof. B. Yegnanarayana for the technical discussions related to this work.

References

[1] NIST, "NIST-Speaker Recognition Evaluations." in *[Online]. Available: http://www.nist.gov/speech/tests/spk.*

[2] L. Laaksonen, H. Pulakka, V. Myllyl, and P. Alku, "Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal," *IEEE Trans. on Consumer Electronics*, vol. 55, no. 2, pp. 780–787, May 2009.

[3] H. Gustafsson, U. A. Lindgren, and I. Claesson, "Low-complexity feature-mapped speech bandwidth extension," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 577–588, March 2006.

[4] *"AMR wideband speech codec; general description, 3GPP TS 26.171", 3rd Generation Partnership Project (3GPP), 2001, version 5.0.0.*

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, NJ, 1978.

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans on Acoust, Speech and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[7] D. Reynolds, T. Quateri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, Jan 2000.

[8] *TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS Order PB91-505065, NIST, Gaithersburg, MD, 1990, Speech Disc 1-1.1.*

[9] C. Jankowski, A. Kalyanwamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephon bandwidth speech database," in *ICASSP*, April 1990.

[10] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, March 1995.

[11] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sept 1997.

[12] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in asssessment of detection task performance," in *Proc. Eur. Conf. Speech Communication Technology, Rhodes, Greece*, 1997, pp. 1895–1898.