Robust Speaker Recognition Using Spectro-Temporal Autoregressive Models

Sri Harish Mallidi[1], *Sriram Ganapathy*[2], Hynek Hermansky[1]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA.
[2]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA.
{mallidi,hynek}@jhu.edu, ganapath@us.ibm.com

Abstract

Speaker recognition in noisy environments is challenging when there is a mis-match in the data used for enrollment and verification. In this paper, we propose a robust feature extraction scheme based on spectro-temporal modulation filtering using two-dimensional (2-D) autoregressive (AR) models. The first step is the AR modeling of the sub-band temporal envelopes by the application of the linear prediction on the sub-band discrete cosine transform (DCT) components. These sub-band envelopes are stacked together and used for a second AR modeling step. The spectral envelope across the sub-bands is approximated in this AR model and cepstral features are derived which are used for speaker recognition. The use of AR models emphasizes the focus on the high energy regions which are relatively well preserved in the presence of noise. The degree of modulation filtering is controlled using AR model order parameter. Experiments are performed using noisy versions of NIST 2010 speaker recognition evaluation (SRE) data with a state-of-art speaker recognition system. In these experiments, the proposed features provide significant improvements compared to baseline features (relative improvements of 20% in terms of equal error rate (EER) and 35 % in terms of miss rate at 10 % false alarm).

Index Terms: Rate-Scale Filtering, Autoregressive Modeling, Speaker Recognition, Robust Feature Extraction.

1. Introduction

Speech technology works reasonably in matched conditions but rapidly degrades when there is acoustic mis-match between the training and test conditions. Although multi-condition training can improve the performance [1], realistic scenarios can benefit from more robustness without requiring training data from the target acoustic environment. In this paper, we develop a feature extraction scheme which attempts to address robustness in noisy and reverberant environments.

In the past, various feature processing techniques like spectral subtraction [2], Wiener filtering [3] and missing data reconstruction [4] have been developed for noisy speech recognition applications. Feature compensation techniques have also been used in the past for speaker verification systems (feature warping [5], RASTA processing [6] and cepstral mean subtraction (CMS) [7]). With noise or reverberation, the low energy valleys of speech signal have the worst signal to noise ratio (SNR), while the high energy regions are robust and could be well modeled [9]. In general, an autoregressive (AR) modeling approach represents high energy regions with good modeling accuracy [10, 11]. The AR modeling approach of signal spectra is widely used for feature extraction of speech [12]. The AR modeling of Hilbert envelopes [16, 17] have been used with similar goals of preserving peaks in sub-band temporal envelopes and has been successfully applied for speaker verification [27]. 2-D AR modeling was originally proposed for speech recognition by alternating the AR models between spectral and temporal domains [14].

In this paper, we extend our previous approach on two dimensional AR modeling [15] with a modulation filtering framework. Long segments of the input speech signal are decomposed into sub-bands and linear prediction is applied on the sub-band discrete cosine transform (DCT) components to derive Hilbert envelopes [16]. The sub-band envelopes are stacked together to form a time-frequency description and a second AR model is applied across the sub-bands for each short-term frame (25 ms with a shift of 10ms). The output of the second AR model is converted to cepstral coefficients and used for speaker recognition. Modifying either of the AR models, time domain one or the frequency domain one, represents in effect a rate-scale (time-frequency) modulation filtering [18]. The time domain AR model does the rate filtering and the frequency domain AR model does the scale filtering, similar to the approaches discussed in [19].

Experiments are performed on core conditions of NIST 2010 SRE data [20] with various artificially added noise and reverberation. In these experiments, the proposed features provides considerable improvements compared to the conventional features. The rest of the paper is organized as follows. Sec. 2 details the proposed feature extraction scheme using 2-D AR models. This is followed by a discussion of various rate-scale feature streams derived from this framework (Sec. 3.1). Sec. 4 describes the experiments on the NIST 2010 SRE. In Sec. 5, we conclude with a brief discussion of the proposed front-end.

2. Feature Extraction

The block schematic for the proposed feature extraction is shown in Fig. 1. Long segments of the input speech signal (10s of non-overlapping windows) are transformed using a discrete cosine transform [27]. The full-band DCT signal is windowed into a set of 96 over-lapping linear sub-bands in the frequency range of 125-3700 Hz. In each sub-band, linear prediction is applied on the sub-band DCT components to estimate an all-pole representation of Hilbert envelope [16, 17]. This constitutes the temporal AR modeling stage. The FDLP envelopes from various sub-bands are stacked together to obtain a two-dimensional representation as shown in Fig. 1.

The sub-band envelopes are integrated in short-term frames

This research was funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015 and the Office of the Director of National Intelligence (ODNI). The authors would also like to acknowledge Brno University of Technology, Xinhui Zhou and Daniel Garcia-Romero for software fragments.

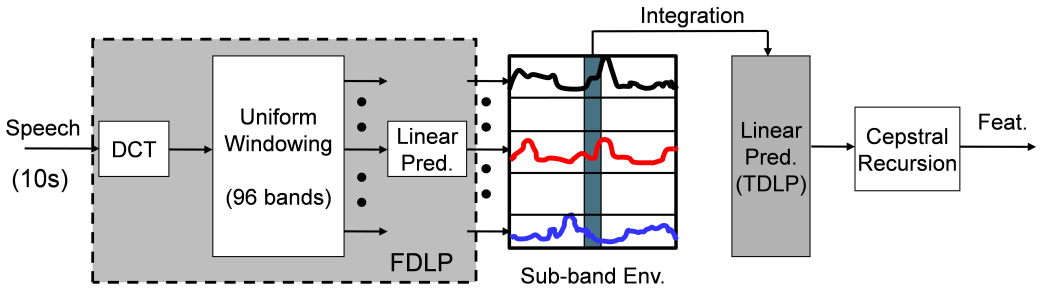

Figure 1: *Block schematic of the proposed feature extraction using spectro-temporal AR Models.*

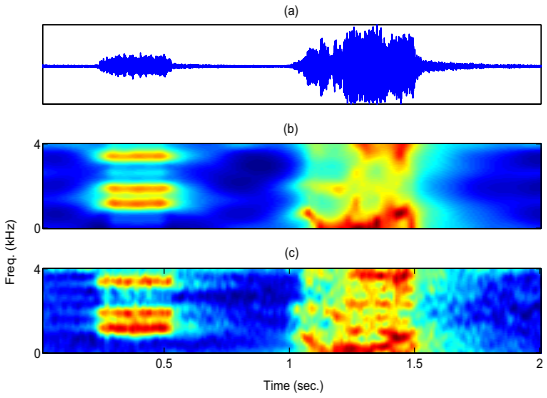

Figure 2: *A portion of the speech signal in (a), and its spectrographic representation with different resolutions obtained using spectro-temporal AR model in (b) low model order in both dimensions and (c) high model order in both dimensions.*

(25ms with a shift of 10ms). The output of the integration process provides an estimate of the power spectrum of signal in the short-term frame level. The frequency resolution of this power spectrum is equal to the initial sub-band decomposition of 96 bands. These power spectral estimates are transformed to temporal autocorrelation estimates using inverse Fourier transform and the resulting autocorrelation sequence is used for time domain linear prediction (TDLP). We derive 13 cepstral coefficients from the all-pole approximation of the 96 point short-term power spectrum. The delta and acceleration coefficients are extracted to obtain 39 dimensional features.

3. Properties of 2-D AR Models

3.1. Rate-Scale filtering using AR Models

A temporal modulation filter is referred to as a rate filter and a spectral modulation filter is referred to as a scale filter [19]. In the proposed feature extraction framework, the AR modeling process represents a filter impulse response, whose frequency response ("time response" in the case of the temporal AR filter) can be controlled by the model order. A lower model order represents more smoothing in a given domain , while the higher model captures finer details . Thus, various streams of spectrographic representations can be generated from the proposed framework using different choices of model order for temporal and spectral AR models as shown in Fig. 2. The low-rate low-scale representations represent broad energy variations in the signal as seen in Fig. 2 (b). The other configuration using higher order for the AR models is shown in Fig. 2 (c) where more details about the various events in the spectrogram are evident. A higher order could also mean that such AR models may carry information about noise or reverberation artifacts that is present in the finer details of the spectrogram in its spectral or temporal directions.

In addition to the configurations shown in Fig. 2, other possibilities include a lower model order for temporal AR model with a higher order for the spectral AR model and vice-versa. Thus, various feature streams which differ in the extent of modulations can be derived from the spectro-temporal AR model framework. In Sec. 4, we provide some experiments showing the effect of model order on the speaker recognition performance.

3.2. Robustness to Noise

When a speech signal is corrupted with noise or reverberation, the valleys in the sub-band envelopes are dominated by noise. Even with moderate amounts of distortion, the low-energy regions are substantially modified and cause acoustic mis-match with the clean training data. Since the AR modeling tends to fit the high energy regions with good accuracy [11], the spectro-temporal AR modeling approach described in Sec. 2 could be more robust to noise and reverberation artifacts. This is illustrated in Fig. 3 where we plot a portion of clean speech signal, speech with additive noise (babble noise at 10 dB SNR) and speech with artificial reverberation (reverberation time of 300 ms). The spectrographic representation obtained from mel frequency representation is shown in the second panel and the corresponding representation obtained from spectro-temporal AR models is shown in the bottom panel. In comparison with the mel spectrogram, the representation obtained from AR modeling emphasizes the high energy regions. Thus, such a representations can be more similar for the clean and the noisy versions of the same signal. This is desirable and contributes to improved robustness when these features are used for speaker recognition in noisy environments.

4. Experiments and Results

The proposed features are used for speaker recognition using the core conditions of the NIST 2010 speaker recognition evaluation (SRE) [20]. The baseline features consist of 39 dimensional MFCC features [8] containing 13 cepstral coefficients, their delta and acceleration components. These features are computed on 25ms frames of speech signal with a shift of 10ms. We use 37 Mel-filters in the frequency range of 125-3700 Hz for the baseline features.

We use a GMM-UBM based speaker verification system [22]. The input speech features are feature warped [5] which forms a normalization of the mean,variance and higher order moments. Gender dependent GMMs with 1024 mixture

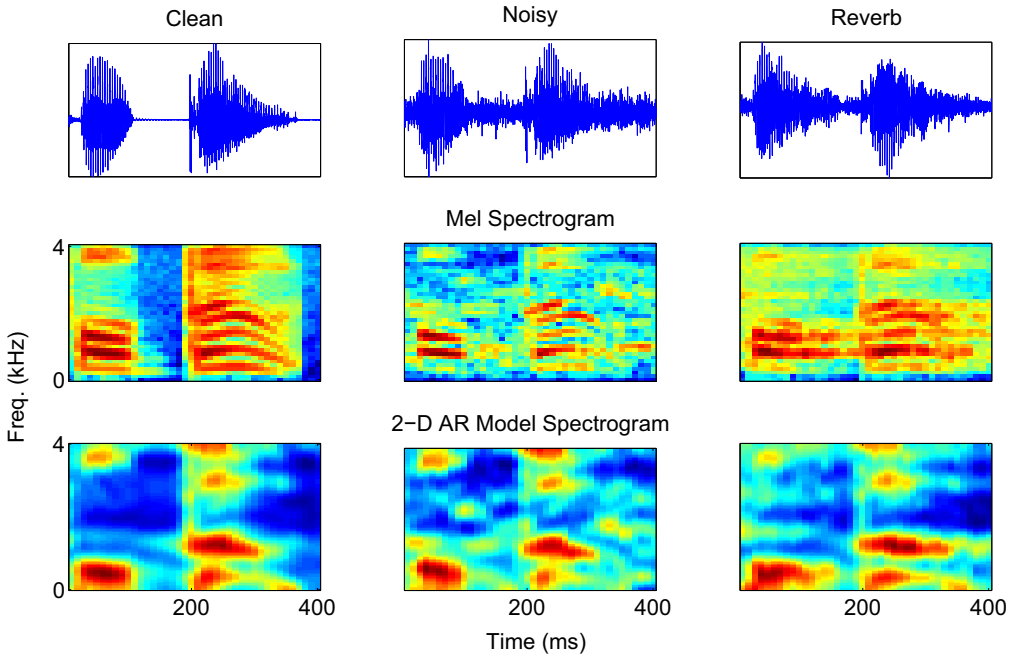

Figure 3: *Comparison of spectrographic representations obtained from clean speech, noisy speech (babble noise at* 10 *dB) and reverberant speech (reverberation time of* 300 *ms for the mel-spectrogram and the proposed 2-D AR model spectrogram.*

components are trained on the development data. The development data set consists of a combination of audio from the NIST 2004 speaker recognition database, the Switchboard II Phase 3 corpora, the NIST 2006 speaker recognition database, and the NIST08 interview development set. There are 4324 male recordings and 5461 female recordings in development set.

Once the UBM is trained, the mixture component means are MAP adapted and concatenated to form supervectors. We use the i-vector based factor analysis technique [23] on these supervectors in a gender dependent manner. For the factor analysis training, we use the development data from Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2, NIST04-05 and extended NIST08 far-field data. There are 17130 male recordings and 21320 female recordings in this sub-space training set. Gender specific i-vectors of 450 dimensions are extracted and these are used to train a PLDA system [24]. The output scores are obtained using a 250 dimensional PLDA subspace for each gender.

For evaluating the robustness of these features in noisy conditions, the test data for Cond-2 is corrupted using (a) babble noise, (b) exhibition hall noise, and (c) restaurant noise from the NOISEX-92 database, each resulting in speech at 5, 10, 15 and 20 dB SNR. These noises are added using the FaNT tool [25]. For simulating reverberant recording conditions, we also convolve the test data for Cond.-2 with three artificial room responses [26] with reverberation time of 100, 300 and 600 ms. Cond-2 has interview microphone recordings with the highest number of trials among NIST 2010 core conditions (2.8M) and it contains 2402 enrollment recordings and 7201 test recordings. In our experiments, the enrollment data consists of "clean" speech data present in NIST 2010 and the test data may be clean speech data or noisy data. The voice-activity decisions provided by NIST are used in these experiments. The GMM-UBM, i-vector and the PLDA sub-spaces trained from the development data are used without any modification.

Table 1: EER (%) clean and noisy version (babble at 5 dB SNR for Cond.-2 of NIST 2010 SRE for baseline MFCC features and 2-D AR features for various choices of model order for temporal AR model in terms of poles per sec (pps) and spectral AR model in terms of poles per frame (ppf).

Feat.	Clean	Noisy
MFCC	3.0	12.5
2-D AR (10pps, 6ppf)	4.8	15.4
2-D AR (90pps, 6ppf)	3.7	14.4
2-D AR (10pps, 24ppf)	4.0	12.8
2-D AR (90pps, 24ppf)	2.7	10.5
2-D AR (30pps, 12ppf)	2.7	9.8
2-D AR (60pps, 12ppf)	2.8	9.7
2-D AR (15pps, 12ppf)	3.0	11.4
2-D AR (30pps, 18ppf)	2.6	10.2

The performance metric used is the EER (%) and the false-alarm rate at a miss-rate of 10 % (Miss10). The initial set of experiments discuss the selection of model order using the clean data for Cond.-2 as well as validation data from babble noise at 5 dB SNR. This choice of validation data was not optimized in any manner and the performance on other types of noise and SNR levels relates to the generalization of the parameter selection process. The results for various choices of model order (described in terms of number of peaks per second for temporal model or number peaks per frame across all bands for the spectral AR model) is shown in Table. 1.

Based on the results provided in Table. 1, we select a model order of 30 poles per sec (pps) for the temporal AR model and an order of 12 poles per frame (ppf) for the spectral AR model. The comparison of the performance for various noisy and reverberant conditions (average of three types of noise) for the baseline features as well as the 2-D AR features is shown in Fig. 4.

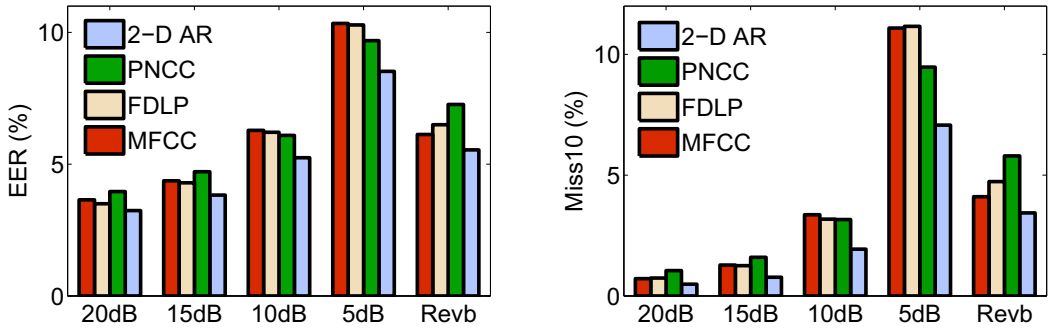

Figure 4: *Average performance over three types of noise (Babble, Restaurant, Exhall) and three reverberant room responses (with reverberation time of* 100*,* 300 *and* 600 *ms) in terms of EER (%) and False Alarm (%) at 10% Miss Rate (Miss10) for core condition* 2*.*

Table 2: EER (%) and False Alarm (%) at 10% Miss Rate (Miss10) in parantheses for core evaluation conditions in NIST 2010 SRE.

Cond.	MFCC-baseline	FDLP	2-D AR Feat.
1. Int.mic - Int.mic-same-mic.	2.0 (0.1)	2.1 (0.1)	1.9 (0.1)
2. Int.mic - Int.mic-diff.-mic.	3.0 (0.4)	2.9 (0.5)	2.7 (0.4)
3. Int.mic - Phn.call-tel	3.9 (1.1)	3.6 (0.8)	3.8 (0.9)
4. Int.mic - Phn.call-mic	3.3 (0.5)	2.8 (0.3)	2.9 (0.3)
5. Phn.call - Phn.call-diff.-tel	2.9 (0.4)	2.9 (0.6)	3.8 (0.9)
6. Phn-call - Phn.call-high-vocal-effort-tel	4.3 (1.5)	5.1 (2.2)	5.1 (2.4)
7. Phn-call - Phn.call-high-vocal-effort-mic	7.6 (4.9)	5.8 (2.5)	4.7 (2.2)
8. Phn-call - Phn.call-low-vocal-effort-tel	2.1 (0.3)	2.6 (0.5)	2.8 (0.6)
9. Phn-call - Phn.call-low-vocal-effort-mic	2.1 (0.1)	2.1 (0.1)	1.8 (0.1)

Table 3: False Alarm (%) at 10% Miss Rate (Miss10) for evaluation conditions in IARPA BEST 2011 task.

Cond.	MFCC	2-D AR Feat.
1. Int.mic - Int.mic-noisy.	15.5	11.3
2. Int.mic - Phn-call-mic	3.7	2.8
3. Int.mic - Phn.call-tel	3.3	2.8
4. Phn-call-mic - Phn.call-mic	7.4	6.7
5. Phn.call-mic - Phn.call-tel	7.5	6.3
6. Phn.call-tel - Phn.call-tel	1.3	1.8

We also compare these results the FDLP features which involves one dimensional temporal AR model [27] and the power normalized cepstral coefficients (PNCC) [28]. The PNCC features provide improvements over the baseline MFCC features on low SNR additive noise conditions. However, on all the noise types and reverberant conditions, the proposed approach improves over the other feature extraction methods considered here. On the average, the proposed features provide about 35 % relative Miss10 improvement over the baseline MFCC system. These improvements are mainly due to the robust representation of the high energy regions by 2-D AR modeling and the rate-scale modulation filtering.

In the next set of experiments, we compare the proposed 2-D AR model features for all the 9 core conditions in NIST 2010 SRE. These results are reported in Table 2. From these results, it can be seen that the proposed 2-D features provides good improvements in mis-matched far-field microphone conditions like Cond. 1,2 7 and 9). In these conditions the modeling of high-energy regions in time-frequency domain is beneficial. However, the baseline MFCC system performs well in telephone channel matched conditions (Cond. 5, 6 and 8). The degradation in Cond. 5, 6 and 8 may be attributed to the reduced resolution caused by the 2-D AR modeling. In the final set of experiments, we measure the speaker verification performance using the IARPA BEST 2011 data [29]. The database contains 83198 recordings (25822 enrollment utterances and 57376 test utterances) with a wide-variety of intrinsic and extrinsic variabilities like language, age, noise and reverberation. There are 38M trials which are split into various conditions as shown in Table 3. Condition 1 contains majority of the trials (20M trials) recorded using interview microphone data with varying amounts of additive noise and artificial reverberation.

The performance (Miss10) for the baseline MFCC system is compared with proposed features in Table 3. In these experiments, the proposed features provide noticeable improvements for all conditions except the matched telephone scenario (Cond. 6). On the average, the proposed features provide improvements of about 18% in the Miss10 metric relative to the baseline.

5. Summary

In this paper, we have proposed a two-dimensional autoregressive model for robust speaker recognition. An initial temporal AR model is derived from long segments of the speech signal. This model provides Hilbert envelopes of sub-band speech which are integrated in short-term frames to obtain power spectral estimates. The estimates are used for a spectral AR modeling process and the output prediction coefficients are used for speaker recognition. Various experiments are performed with noisy test data on NIST 2010 SRE where the proposed features provide significant improvements. These results are also validated using a large speaker recognition dataset from BEST. The results are promising and encourage us to pursue the problem of joint 2-D AR modeling instead of a separable time and frequency linear prediction schemes adopted in this paper.

6. References

[1] Ming, J., Hazen, T.J., Glass, J.R. and Reynolds, D.A., "Robust Speaker Recognition in Noisy Conditions", *IEEE Tran. on Audio Speech Lang. Proc.*, Vol 15 (5), 2007, pp. 1711 - 1723.

[2] Boll, S.F.,"Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27 (2), Apr. 1979, pp. 113-120.

[3] "ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2002.

[4] Cooke, M., Morris, A., Green, P., "Missing data techniques for robust speech recognition", *Proc. ICASSP*, 1997, pp. 863-866.

[5] Pelecanos, J. and Sridharan, S., "Feature warping for robust speaker verification", *Proc. Speaker Odyssey 2001 Speaker Recognition Workshop*, Greece, pp. 213-218, 2001.

[6] Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, Vol. 2, pp. 578-589, 1994.

[7] Rosenberg, A.E., Lee, C. and Soong, F.K., "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification," in *Proc. ICSLP*, pp. 1835-1838, 1994.

[8] Davis, S. and Mermelstein, R., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28 (4), Aug. 1980, pp. 357-366.

[9] Guruprasad, S., "Significance of processing regions of high signal-to-noise ratio in speech signals", PhD Thesis, 2011.

[10] Atal, B.S., Hanauer, L.S, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. America*, Vol 50 (28), 1971, pp. 637-655.

[11] Makhoul, J., "Linear Prediction: A Tutorial Review",in *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[12] Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738-1752, 1990.

[13] Ganapathy, S., Pelecanos, J. and Omar, M.K., "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, 2011, pp. 4836-4839.

[14] Athineos, M. and Hermansky, H. and Ellis, D., "PLP2 Autoregressive modeling of auditory-like 2-D spectro-temporal patterns", *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Processing SAPA04*, pp. 3742, 2004.

[15] Ganapathy, S., Thomas, S. and Hermansky, H., "Feature Extraction Using 2-D Autoregressive Models For Speaker Recognition", *ISCA Speaker Odyssey*, 2012.

[16] Athineos, M. and Ellis, D., "Autoregressive modelling of temporal envelopes," *IEEE Tran. Signal Proc.*, Vol. 55, pp. 5237-5245, 2007.

[17] Kumerasan, R. and Rao, A., "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journal of Acoustical Society of America*, Vol. 105, no 3, pp. 1912-1924, Mar. 1999.

[18] Chi,T., Ru,P. and Shamma,S.A., "Multiresolution spectrotemporal analysis of complex sounds", *The Journal of the Acoustical Society of America*, Vol. 118, 2005, pp. 887-906.

[19] Nemala, S., Patil, K. and Elhilali, M. "A multistream feature framework based on bandpass modulation filtering for robust speech recognition", *IEEE Trans. on Audio, Speech and Lang. Proc.*, Vol. 21 (2), 2013, pp. 416-426.

[20] "National Institute of Standards and Technology (NIST)," speech group website, *http://www.nist.gov/speech*, 2010.

[21] Ganapathy, S., Pelecanos, J. and Omar, M.K., "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, 2011, pp. 4836-4839.

[22] Reynolds, D., "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Comm.* Vol. 17, Aug. 1995, pp. 91-108.

[23] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P and Ouellet, P., "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19(4), pp. 788-798, 2011.

[24] Romero, D. and Espy-Wilson, C.Y., "Analysis of i-vector Length Normalization in Speaker Recognition Systems", *Proc. Interspeech*, 2011.

[25] Hirsch, H.G., "FaNT: Filtering and Noise Adding Tool", *http://dnt.kr.hsnr.de/download.html*.

[26] "ICSI Room Responses," http://www.icsi.berkeley.edu/speech/papers/asru01-meansub-corr.html.

[27] Ganapathy, S., Pelecanos, J. and Omar, M.K., "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, 2011, pp. 4836-4839.

[28] Kim, C., and Stern R., "Power-normalized cepstral coefficients (PNCC) for robust speech recognition", *Proc. ICASSP*, 2012, pp. 4101-4104.

[29] "IARPA BEST Speaker Recognition Challenge 2011", *http://www.nist.gov/itl/iad/mig/best.cfm* , 2011.