Modulation Features for Noise Robust Speaker Identification

Vikramjit Mitra, Mitchel McLaren, Horacio Franco, Martin Graciarena, Nicolas Scheffer

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.
{vmitra, mitch, hef, martin, scheffer}@speech.sri.com

Abstract

Current state-of-the-art speaker identification (SID) systems perform exceptionally well under clean conditions, but their performance deteriorates when noise and channel degradations are introduced. Literature has mostly focused on robust modeling techniques to combat degradations due to background noise and/or channel effects, and have demonstrated significant improvement in SID performance in noise. In this paper, we present a robust acoustic feature on top of robust modeling techniques to further improve speaker-identification performance. We propose Modulation features of Medium Duration sub-band Speech Amplitudes (MMeDuSA); an acoustic feature motivated by human auditory processing, which is robust to noise corruption and captures speaker stylistic differences. We analyze the performance of MMeDuSA using SRI International's robust SID system using a channel and noise degraded multilingual corpus distributed through the Defense Advance Research Projects Agency (DARPA) Robust Automatic Transcription of Speech (RATS) program. When benchmarked against standard cepstral features (MFCC) and other noise robust acoustic features, MMeDuSA provided lower SID error rates compared to the others.

Index Terms— noise-robust speaker identification, modulation features, noise robust acoustic features

10.21437/Interspeech.2013-695

1. Introduction

Current state-of-the-art speaker identification (SID) systems achieve very high performance (low error rates) in clean and high signal-to-noise-ratio (SNR) conditions; but under noisy conditions (especially at low SNRs) their performance degrades appreciably. Research on noise-robust SID has been rare. This scarcity of research coupled with the requirement to have a real-world SID application capable of performing well in an adverse environment have increased the need for robust SID systems.

Prior work [1, 2] presented studies showing how noise degrades the performance of state-of-the-art SID systems, including Gaussian mixture models (GMM); maximum likelihood linear regression (MLLR) [3] and i-vector probabilistic linear discriminant analysis (PLDA) based SID systems [4]. State-of-the-art SID systems at present primarily focus on channel, session and background mismatch compensation by using suitable techniques at the back-end. Few studies have focused on using robust acoustic features to reduce noise degradation where the standard Mel-Cepstrum features (e.g., MFCC) tend to fail [5, 6]. MFCCs have been so far the feature-of-choice for most SID systems because they are simple to generate and have demonstrated state-of-the-art performance in National Institute of Standards and technology (NIST) Speaker Recognition Evaluation (SRE) tasks. Recent success of robust acoustic features for automatic speech recognition (ASR) systems has stirred up the interest in SID community to explore noise robust acoustic features. Modulation features have been quite successful in ASR [7, 8] which stimulated their exploration in SID tasks. Modulation based features have been explored for SID in [9-12] and results have indicated promise in using such features compared to the standard MFCC features. Modulation features due to their longer term modeling capability captures supra-segmental information [23], which is one of the cues efficiently used by human for speaker recognition tasks. Moreover, studies [7-10] have also demonstrated that modulation features are robust to noise. A comprehensive account on the use of robust features for SID is given in [13]. Finally, studies have also looked into fusion of multiple feature based systems to improve accuracy under noisy conditions [14, 27].

Recently, the SID community has produced a significant surge in performance accuracy from the successful implementation of a factor-analysis-based framework. This framework incorporates an i-vector extractor module [15] along with a Bayesian backend (such as probabilistic linear discriminant analysis (PLDA)), and has become the state-of-the-art for SID systems. I-vector extraction is a transformation where speech utterances of variable durations are projected into a single low-dimensional vector, typically having a few hundred components. The i-vector's low rank enables the use of advanced machine-learning strategies that would otherwise be too costly due to the input space's large dimensionality. PLDA was found to be a powerful technique for producing a good identification score [16, 17]. In i-vector-PLDA model, each i-vector is separated into speaker and channel components, analogous to a Joint Factor Analysis (JFA) framework [18]; where PLDA is a probabilistic model that models the speaker and intersession variability in i-vector space.

In this paper, we present the Modulation features of Medium Duration sub-band Speech Amplitudes (MMeDuSA), which track temporal modulations across frequency bins. Studies [19, 20] have shown that amplitude modulation of speech signals plays an important role in speech perception; hence, several studies [8, 21] have modeled the speech signal as a weighted combination of amplitude-modulated narrow-band signals. For a reliable estimate of amplitude modulation, it is imperative to ensure that the signals are sufficiently band-limited or narrow-band [7], for which we have used a gammatone filter-bank [22].

MMeDuSA is a combination of amplitude modulation (AM) based cepstral features and summary AM based cepstral feature; where the summary AM signal is obtained by summing the estimated AM signals across the frequency channels, where modulation information between 5 to 200 Hz is retained. The AM energies are root compressed before being transformed using Discrete Cosine Transform (DCT) as conventional log compression is known to be susceptible to noise corruption [23]. The final MMeDuSA feature is obtained by taking the first few DCT coefficients along with their

velocity (Δ) and acceleration (Δ^2) coefficients. The summary modulation information in MMeDuSA helps to capture information such as vowel stress and prominence, which adds speaker stylistic cue into the features.

We compared the MMeDuSA's performance with traditional MFCC features and previously proposed noise robust features on retransmitted channel and noise corrupted DARPA RATS data [24]. The DARPA RATS program aims to develop robust speech processing techniques for highly degraded transmission channels and contains four broad tasks: speech activity detection (SAD), language identification (LID), key word spotting (KWS), and SID. The data was collected by Linguistic Data Consortium (LDC) by retransmitting conversational telephone speech through eight different communication channels [24]. The RATS rebroadcasted data is unique in the sense that the noise and channel degradations were not artificially introduced by performing simple mathematical operations on the speech signal, but by transmitting clean source signals through eight different radio channels (more detailed description of the retransmission process is given in [24]), where variation of channel to channel introduced a wide variety of distortion modes. The distortion modes include band limitation, strong channel noise, nonlinear speech distortions, frequency shifts, intermittent no-transmission bursts, variable SNR, diverse noise characteristics etc. The data retransmission process included a wide array of target signal transmitters/transceivers, interference signal transmitters, listening station receivers and signal collection and digitization apparatus [24], also the data contained speech from multiple languages specified in section 3. Note that the NIST-SREs and IARPA BEST evaluation of speaker technology dealt with SID under noisy and reverberated conditions, where degradations were mostly artificially simulated, (except the latest NIST-SRE which contained some speech examples recorded in noisy environment), and had high SNRs.

2. MMeDuSA Feature

The proposed MMeDuSA feature was obtained by using the signal processing steps outlined in Figure 1. First, the speech signal is pre-emphasized (using a pre-emphasis filter with coefficient 0.97) and then analyzed using a hamming window of 51.2 ms with a 10 ms frame rate. The windowed speech signal s$[n]$ is passed through a gammatone filter-bank having 34 critical bands, with center frequencies spaced equally in the equivalent rectangular bandwidth (ERB) scale between 250 Hz and 3750 Hz. Note that for all experiments presented in this paper, we assume that the input speech signal has useful information up to 4000 Hz. The filters' bandwidths are characterized by the ERB scale, where the ERB for channel c (where c = 1 … 34) given by-

$$ERB_c = \frac{f_c}{Q_{ear}} + BW_{min} \quad (1)$$

where f_c represents the center frequency for filter c and Q_{ear} and BW_{min} are constants set to 9.26449 and 24.7 according to Glasberg & Moore specifications [22]. The time signal from the c^{th} gammatone filter with impulse response $h_c(n)$ is given as

$$s_c(n) = s(n) * h_c(n). \quad (2)$$

For each of these 34 subband signals, their AM signals are computed using the Teager Energy Operator (TEO) [25]. TEO is a nonlinear energy operator, Ψ, which tracks the instantaneous energy of a band-limited signal. While formulating the operator Ψ, Teager assumed [25] that a signal's energy is not only a function of its amplitude but also of its frequency. Let us consider a discrete sinusoid $x[n]$, where A = a constant amplitude, Ω = digital frequency, f = frequency of oscillation in Hertz, f_s = sampling frequency in Hertz, and θ = initial phase angle -

$$x[n] = Acos[\Omega n + \theta];\ \Omega = 2\pi\,(f/f_s). \quad (3)$$

If $\Omega \leq \pi/4$ and is sufficiently small, Ψ takes the form

$$\Psi\{x[n]\} = \{x^2[n] - x[n-1]x[n+1]\} \approx A^2\Omega^2 \quad (4)$$

where the maximum-energy-estimation error in Ψ will be 23% if $\Omega \leq \pi/4$ or $f/f_s \leq 1/8$. Maragos *et al.*, [26] used Ψ to formulate the discrete energy separation algorithm (DESA), and showed that the algorithm can instantaneously separate the AM/FM components of a narrow-band signal. However, AM/FM signals computed from the DESA may contain discontinuities or instantaneous spikes (that substantially increase their dynamic range), for which median filters or low-pass filters have been used. In order to remove such artifacts from the DESA algorithm, we assume that the sub-band signals are sufficiently band-limited that their instantaneous frequency signal (Ω) is approximately equal to the center frequency of the corresponding gammatone filter

$$\Omega \approx f_c. \quad (5)$$

Given (5), the estimation of the instantaneous AM signal from (4) becomes straight forward

$$A_c \approx \sqrt{\frac{|s_c{}^2[n] - s_c[n-1]s_c[n+1]|}{f_c{}^2}}. \quad (7)$$

The power of the estimated AM signals were computed (refer to Figure 1) and non-linear compression (for the experiments reported here we have used $1/15^{th}$ root compression as it is found to be more noise robust compared to logarithmic compression) was performed on it. The power of the AM signal $a_{k,j}[n]$ for k^{th} channel and j^{th} frame is given as

$$P_{k,j}^{AM} = a_{k,j}^T a_{k,j}. \quad (8)$$

For a given analysis window 34 power coefficients were obtained for each of the 34 channels, which were then transformed using DCT and their first 20 coefficients were retained. Note that in our experiments we have used these 20 coefficients by themselves along with their velocity (Δ) and acceleration (Δ^2) coefficients, which are named as the medium duration modulation cepstra (MDMC) features, it acquired its name 'medium duration' because of its larger analysis window size- 52ms compared to traditionally used 20ms~25ms windows.

In parallel, each of the 34 estimated AM signals (as shown in Figure 1) were band-pass filtered using DCT, retaining information only within 5 Hz to 200 Hz. These are the medium duration modulations (represented as: $a\_mod_{k,j}[n]$), which were summed across the frequency scale to obtain medium duration modulation summary

$$\overline{a\_mod_j} = \sum_{k=1}^{N} a\_mod_{k,j}[n]. \quad (9)$$

The power of the medium duration modulation summary was obtained, followed by $1/15^{th}$ root compression. The resultant was transformed using DCT and the first 3 coefficients were retained. These 3 coefficients were combined with the 20 DCT coefficients obtained from the other branch of the MMeDuSA processing (refer to Figure 1), to yield a 23 dimension feature vector. Velocity (Δ) and acceleration (Δ^2) coefficients were computed for each of the

23 feature dimensions, yielding a final 69-dimensional feature set. This is the final MMeDuSA feature set used in our SID experiments presented below.

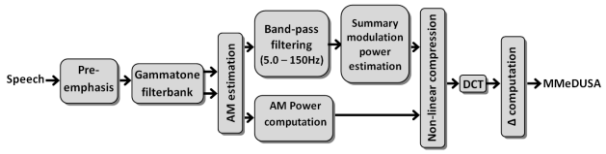

Figure 1. Flow diagram of MMeDuSA feature extraction from speech.

3. Data

The training and test data for the experiments presented here were taken from DARPA RATS Rebroadcast Example (RATS-RE) for the RATS SID task, distributed by LDC [24]. The data was collected by retransmitting telephone speech through eight different communication channels. These channels have a range of distortions associated with them. The RATS program specified multiple duration configurations for speaker enrollment and testing, including a total of eight conditions with input file durations of 3, 10, 30 and 120 sec [27]. In experiments presented here, we considered the SID speech data available at the time of the DARPA-RATS phase-1 evaluation and focused on matched enrollment and testing durations of 30 sec, 10 sec and 3 sec data. The data also contained 120 sec durations, where most of the features performed well in our earlier experiments [27] hence we decided to focus only on the more challenging durations. Note that enrollment duration of 30 sec denotes that speaker models were trained using six sessions, each containing at least 30 sec of speech activity. For phase-1 of the RATS SID task, LDC released three datasets (LDC2012E49, LDC2012E63 and LDC2012E69) containing five languages: Levantine Arabic, Farsi, Dari, Pashto and Urdu. This data was divided into training and development sets used in the experiments presented below. The DARPA RATS dataset is unique in the sense that noise and channel degradations were not artificially introduced by performing mathematical operations on the clean speech signal, but the signals were in fact rebroadcasted through a channel and noise degraded ambience and then rerecorded. Consequently, the data contained several unusual artifacts such as nonlinearity, frequency shifts, modulated noise, intermittent bursts etc., and traditional noise robust approaches developed in the context of additive noise may not work so well.

4. The SID System

For the SID experiments, we used a standard i-vector - PLDA architecture as our speaker recognition system [27, 28]. For the i-vector framework we used universal background models (UBMs) with 512 diagonal covariance Gaussian components trained in a gender independent fashion. The i-vector dimensions of 400 were reduced to 200 dimensions by LDA followed by length normalization and PLDA. For PLDA training, segments in the training set had a single 30 sec cut from each recording to better represent the i-vector distribution of test data.

The RATS SID task was defined as a speaker verification task where each speaker model was trained using six different sessions. A trial was designed using one speaker model and one test session. The transmission channels of the six different enrollment sessions were picked randomly to have speaker models trained on multiple transmission types. Some of the trials were thus performed on channels seen in enrollment, while others were not. While enroll and test durations where restricted to 3, 10 and 30 sec durations, full segments were used for i-vector and UBM training.

The primary metric was defined as the percentage of misses at a 4% false alarm rate. Note that multiple duration configurations for the enrollment and tests were of interest in the RATS phase-1 evaluation, however for the sake-of-simplicity we present only the results using the matched durations and the trend is typically consistent for the other durations as well.

5. Experiments

Our train set consisted of 55982 retransmitted recordings from 8 channels. This data was sourced from 1788 female and 4124 male speakers distributed across languages in the following manner: Levantine Arabic (636), Dari (1096), Urdu (1779), Pashto (1823) and Farsi (579). The UBM was trained from a subset of 9429 of these recordings with an even distribution across languages and channels. Evaluation data contained 26907 retransmitted segments from 305 speakers distributed across languages similar to the train set and contained altogether 106 female and 199 male speakers. Six original recordings were selected per speaker for the purpose of enrollment while their remaining segments set aside for testing. Each speaker had up to 10 models trained by randomly selecting a channel for each of the enrollment segments. Testing models against a pool of 9415 test segments and restricting to same-language trials resulted in 80607 target trials and 5.5 million impostor trials.

We present the results in three different metrics: percentage misses at 4% false alarm (FA), percentage FA at 10% misses and equal error rate (EER). Individual feature based systems were trained using the following features- (a) MFCC features, previously proposed noise robust features (b) PNCC [28], (c) MHEC [9], (d) MDMC (which is MMeDuSA feature excluding the three summary modulation coefficients) and finally the proposed (d) MMeDuSA feature. Figures 2-4 present the percentage misses obtained from the different systems at 4% FA, figures 5-7 present the percentage FA at 10% miss and finally Tables 1 presents the EER obtained at 30s, 10s and 3s trials. All features contained 20 cepstral information padded with their Δ and Δ^2 yielding a 60D feature set with the exception of MMeDuSA which contained 23 base features (as explained in section 2) padded with their Δ and Δ^2 yielding a 69D feature set. The performance of these five features is reported under three different conditions; (1) seen: where the test channel was observed in at least one of the six enrollment segments, (2) unseen: where the test channel was not seen during enrollment, and (3) both: which is the combination of trial subsets (1) and (2). Note that we have performed a random selection of the channels (in accordance with the RATS task), where a channel can potentially be seen more than once during enrollment. Figures 2, 3, 5, 6 and Table 1 show that for both 30sec-30sec and 10sec-10sec durations the MMeDuSA feature consistently outperformed the alternate features in all the testing criteria, however at 3sec-3sec duration MDMC performed the best, where MMeDuSA was the second best for FA (%) and EER and a close 3rd at Misses (%).Overall, MMeDuSA demonstrated a relative EER improvement of 14.4%, 10.5% and 16.6% at 30s-30s, 10s-10s

and 3s-3s conditions compared to the baseline MFCCs and demonstrated a relative EER improvement of 4.9% and 2.7% at 30sec-30sec and 10sec-10sec conditions compared to MDMC which was the second best feature. At 3s-3s duration the EER from MMeDUSA is marginally worse than that of MDMC and is the second best EER out of the five feature sets. For unseen trials MMeDuSA provided 13.9%, 10.9% and 14.6% relative reduction in EER w.r.t MFCCs features at 30s, 10s and 3s duration, and 5.6% and 2.8% relative EER

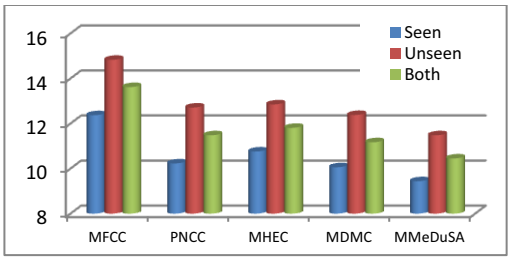

Figure 2. Misses (%) at 4% FA for different features at 30s trial.

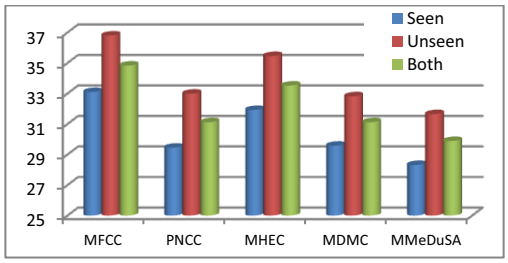

Figure 3. Misses (%) at 4% FA for different features at 10s trial.

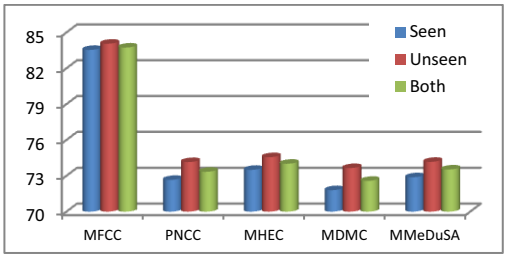

Figure 4. Misses (%) at 4% FA for different features at 3s trial.

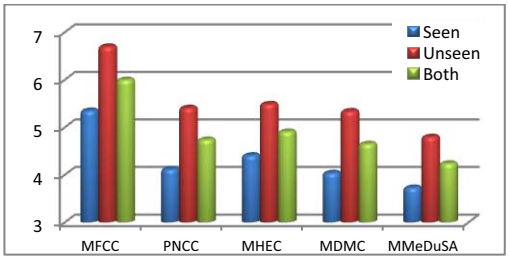

Figure 5. FA (%) at 10% miss for different features at 30s trial.

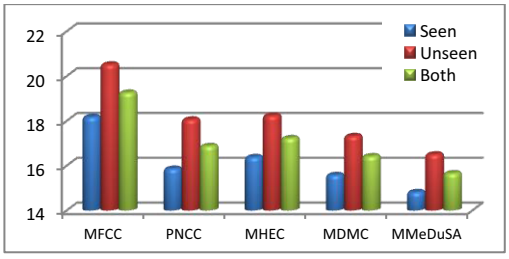

Figure 6. FA (%) at 10% miss for different features at 10s trial.

reduction w.r.t the MDMC feature (which is the second best feature) at 30s and 10s duration, whereas at 3s duration MMeDuSA produced second best EER for unseen trials, providing EER marginally worse than the MDMC features. Overall, both MMeDuSA and MDMC features provided the best result in all trials and all measuring conditions in our experiments.

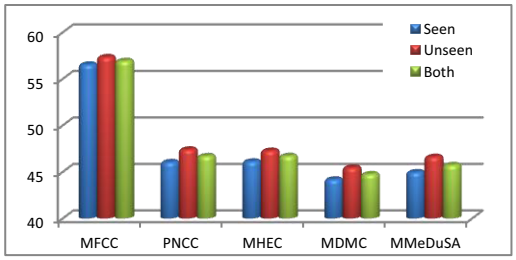

Figure 7. FA (%) at 10% miss for different features at 3s trial.

Table 1. EER from the different feature-based systems

Conditions		MFCC	MHEC	PNCC	MDMC	MMeDuSA
30s-30s	Seen	7.53	6.82	6.60	6.74	**6.41**
	Unseen	8.27	7.60	7.49	7.54	**7.12**
	Both	7.90	7.21	7.03	7.11	**6.76**
10s-10s	Seen	13.66	13.04	12.69	12.63	**12.30**
	Unseen	14.66	13.90	13.65	13.42	**13.05**
	Both	14.14	13.44	13.14	13.00	**12.65**
3s-3s	Seen	31.22	26.30	26.18	**25.20**	25.77
	Unseen	31.48	26.91	26.92	**25.98**	26.56
	Both	31.35	26.59	26.54	**25.58**	26.15

6. Conclusion

We presented MMeDuSA, a modulation-based noise-robust feature for SID, and demonstrated that it offered noise robustness in SID experiments. Our results show that MMeDuSA significantly improved SID performance at low durations compared to the baseline MFCC system and also consistently outperformed two of the previously proposed noise robust features: PNCC and MHEC in most of the conditions. At 3sec-3sec duration MDMC performed the best which was closely matched by the proposed MMeDuSA feature. The experiments presented in this paper dealt with SID tasks for speech degraded with real-world noise and channel artifacts using speakers pooled from multiple languages. Given the difficulty of the task the proposed feature provided consistent improvement with respect to the baseline features and demonstrated that it is competitive even at lower durations. In future we intend to explore feature level combination to see if that can further improve the results beyond what is reported here.

7. Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.
Disclaimer: Research followed all DoD data privacy regulations.
Approved for Public Release, Distribution Unlimited

8. References

[1] P. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. of ICASSP*, Vol. 1, pp. 117-120, Atlanta, 1996.

[2] Y. Pan and A.Waibel, "The effects of room acoustics on MFCC speech parameter," *in of Proc. ICSLP*, Beijing, pp. 129-132, 2000.

[3] M. Graciarena, S. Kajarekar, A. Stolcke, and E. Shriberg, "Noise robust speaker identification for spontaneous Arabic speech," *Proc. of ICASSP* 2007. IEEE, 2007, vol. IV.

[4] Y. Lei, L. Burget, L. Ferrer, M. Graciarena and N. Scheffer, "Towards Noise Robust Speaker Recognition Using Probabilistic Linear Discriminant Analysis", *Proc. of ICASSP*, 2012.

[5] Y. Shao and D.L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," *Proc. of ICASSP*, IEEE, 2008.

[6] Y. Shao, S. Srinivasan, and D.L.Wang, "Incorporating auditory feature uncertainties in robust speaker identification," *Proc. of ICASSP*, IEEE, 2007, vol. IV.

[7] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager energy cepstrum coefficients for robust speech recognition", *in Proc. of Interspeech*, pp. 3013–3016, 2005.

[8] V. Mitra, H. Franco, M. Graciarena and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition", *in Proc. of ICASSP*, pp. 4117-4120, Japan, 2012.

[9] J.-W. Suh, S. O. Sadjadi, G. Liu, T. Hasan, K. W. Godin and J. H. L. Hansen, "Exploring Hilbert envelope based acoustic features in i-vector speaker verification using HT-PLDA", *Proc. of NIST 2011 Speaker Recognition Evaluation Workshop*, Atlanta, GA, USA, 2011.

[10] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition", *Proc. of ICASSP*, vol. I, pp. 665–668, 2006.

[11] T. Kinnunen, K.-A. Lee and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification", *in Proc. of The Speaker and Language Recognition Workshop*, Odyssey 2008.

[12] T. Thiruvaran, E. Ambikairajah and J. Epps, "Extraction of FM components from speech signals using all-pole model", *Electronics Letters* 44, 6, 2008.

[13] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm*., vol. 52, 1, pp. 12–40, Jan. 2010.

[14] N. Thian and S. Bengio, "Noise-robust multi-stream fusion for textindependent speaker authentication," *The Speaker and Recognition Workshop*, 2004.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, May 2010.

[16] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," *in Proc. of ICCV*, 11th IEEE, 2007, pp. 1–8.

[17] P. Kenny, "Bayesian speaker verification with heavy-tailed priors*," in Odyssey 2010-The Speaker and Language Recognition Workshop,* IEEE, 2010.

[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, July 2008.

[19] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception", J. Acoust. Soc. of Am., 95(5), pp. 2670–2680, 1994.

[20] O. Ghitza, "On the upper cutoff frequency of auditory critical-band envelope detectors in the context of speech perception", *J. Acoust. Soc. of Am.*, 110(3), pp. 1628–1640, 2001.

[21] V. Tyagi, "Fepstrum features: Design and application to conversational speech recognition", *IBM Research Report*, 11009, 2011.

[22] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, 47, pp.103–138, 1990.

[23] S. Ravindran, D. V. Anderson and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," *in proc of SAPA*, Pittsburgh, PA, September 2006.

[24] K.Walker and S. Strassel, "The RATS radio traffic collection system," *Proc. of ISCA*, Odyssey, 2012.

[25] H. Teager, "Some observations on oral air flow during phonation", *IEEE Trans. ASSP*, pp. 599–601, 1980.

[26] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis", *IEEE Trans. Signal Processing*, 41, pp. 3024–3051, 1993.

[27] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer and Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion", in review, *ICASSP* 2013.

[28] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," *Proc. of ICASSP*, pp. 4574–4577, 2010.

[29] M. Markaki and Y. Stylianou, "Evaluation of modulation frequency features for speaker verification and identification," *Proc. of EUSPICO*, pp. 549-553, 2009.