MULTI-LAYER PERCEPTRON BASED SPEECH ACTIVITY DETECTION FOR SPEAKER VERIFICATION

Sriram Ganapathy[1], *Padmanabhan Rajan*[2], *and Hynek Hermansky*[1]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, USA.
[2] Dept. of Computer Science and Engg., Indian Institute of Technology Madras, India.
{ganapathy,hynek}@jhu.edu, padman@cse.iitm.ac.in

ABSTRACT

In this paper, we present a speech activity detection (SAD) technique for speaker verification in noisy environments. The proposed SAD is based on phoneme posteriors derived from a multi-layer perceptron (MLP). The MLP is trained using modulation spectral features, where long temporal segments of the speech signal are analyzed in critical bands. In each sub-band, temporal envelopes are derived using the autoregressive modelling technique called frequency domain linear prediction (FDLP). The robustness of the sub-band envelopes is achieved by a minimum mean square envelope estimation technique. We also experiment with MFCC features processed with cepstral mean subtraction. The speech features are input to the trained MLP to estimate phoneme posterior probabilities. For SAD, all the speech phoneme probabilities are merged to one speech class to derive speech/non-speech decisions. The proposed SAD is applied for a speaker verification task using noisy versions of NIST 2008 speaker recognition evaluation (SRE) data, where the proposed SAD provides significant improvements (relative equal error rate (EER) improvement of about 9 % in additive noise and about 19 % in reverberant conditions). Furthermore, the improvements are consistent for the two different front-ends (FDLP and MFCC) considered here.

Index Terms— Frequency Domain Linear Prediction (FDLP), Speech Activity Detection, Speaker Verification.

1. INTRODUCTION

In most speech processing systems, the first step in dealing with a speech signal is the reliable detection of speech activity. Speech activity detection (SAD) has been studied for many decades now and various algorithms have been developed for speech recognition, speech coding, speaker verification and other applications. In the case of speaker verification, the main challenge is to obtain speech segments with low amounts of false alarms for signals embedded in noise and reverberation. In low signal-to-noise ratio (SNR) and non-stationary environments, conventional approaches often fail and speaker recognition performances can degrade significantly.

Several approaches have been proposed in the past for SAD. Earlier methods like [1] use parameters from waveform like time domain linear prediction coefficients, zero crossing rate etc. Statistical approaches to SAD have also been explored in the past [2], where a likelihood ratio test is applied on the noise to signal ratio. The most common SAD is the adaptive, energy-based speech detector [3]. Other approaches like estimation of noise energy from the speech signal and using a threshold on the frame level mel spectrum SNR are used in speech recognition [4]. In a more recent approach, multi-scale spectro-temporal modulations which emulate human auditory processing have been also investigated for SAD [5].

In this paper, we address the problem of robust SAD using multi-layer perceptrons (MLPs). MLPs are widely used in automatic phoneme recognition tasks where they are used to estimate phoneme posterior probabilities [6]. For SAD, the speech phoneme posteriors are merged to single speech class. This gives a two class posterior probability vector with speech/non-speech probabilities. These probabilities are hard thresholded to speech/non-speech decisions and a Viterbi decoder is used to smooth the decisions. A minimum duration of 7 consecutive frames is imposed on the speech/non-speech classes.

We propose to train MLPs using robust feature extraction schemes based on frequency domain linear prediction (FDLP) [7]. In this process, the speech signal is analyzed in critical bands and a minimum-mean square error (MMSE) estimation of the sub-band Hilbert envelopes is performed [9]. The robust sub-band envelopes are used for deriving using autoregressive models (FDLP) [8]. The FDLP envelopes are compressed using static and dynamic compression and are converted to modulation frequency components [7]. In order to test the performance of standard front-ends on the MLP SAD system, we also experiment with MFCC features processed with cepstral mean subtraction.

The proposed SAD technique is evaluated on noisy and reverberated versions of a subset of the NIST 2008 speaker recognition evaluation (SRE) dataset. In these experiments, the MLP-based SAD results in robust detection of speech segments compared to other SAD techniques. Using various SAD techniques, we also perform speaker verification experiments using a Gaussian mixture model-universal background model (GMM-UBM) with the i-vector Gaussian probabilistic linear discriminant analysis (PLDA) system [10]. In these experiments, the MLP based SAD system (with either FDLP or MFCC features) outperforms the other approaches in terms of both SAD error as well as the resulting speaker recognition equal error rate (EER).

The rest of the paper is organized as follows. In Section 2, we describe the proposed modulation feature extraction. The MLP-based SAD system is described is Section 3. Experiments performed with the SAD system are reported in Section 4. Finally, we conclude in Section 5.

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), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). Any opinions or findings expressed in this material are only those of the author(s). The authors would like to thank Xinhui Zhou, Daniel Garcia-Romero and Hans Guenter Hirsch for baseline SAD system source code.

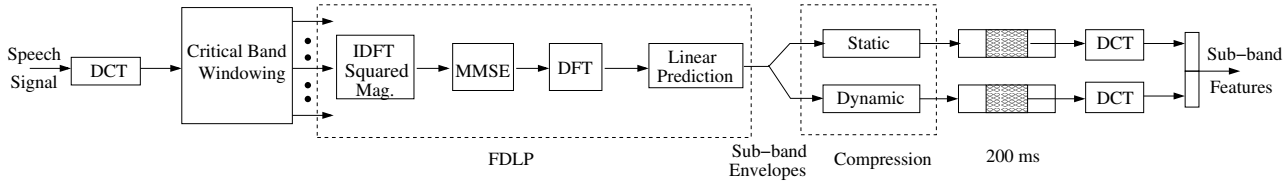

Figure 1: Block schematic for the modulation spectrum based feature extraction technique.

2. FEATURE EXTRACTION

2.1. Extraction of non-parametric Hilbert envelope

The block schematic for the modulation spectrum based feature extraction technique is shown in Fig. 1. Long segments of the speech signal are decomposed into bark-spaced sub-bands by windowing the discrete cosine transform (DCT). For deriving the sub-band Hilbert envelope [8], the squared magnitude of the discrete Fourier transform (DFT) is used. The MMSE technique is applied to estimate the clean envelope from the noisy speech envelope [9].

2.2. MMSE Hilbert envelope estimation

When speech signal is corrupted by uncorrelated additive noise, the signal that reaches the microphone can be written as

$$x[m] = s[m] + n[m], \tag{1}$$

where $x[m]$ is the discrete representation of the input signal, $s[m]$ represents the clean speech signal which is corrupted by noise $n[m]$. By virtue of the orthogonality property of the DCT matrix, the speech and noise signals continue to be additive and uncorrelated in the DCT domain. Further, the application of DFT on the zero padded DCT signal [8] gives

$$A_X(m,i) = A_S(m,i) + A_N(m,i), \tag{2}$$

where $A_X(m,i)$, $A_S(m,i)$ and $A_N(m,i)$ are the discrete time analytic signal representations of the noisy speech, clean speech and noise respectively for the sub-band i. The MMSE estimator [9] can be used for the estimation of the magnitude of the analytic signal (similar to the spectral amplitude estimator). Thus, the plug-in estimate for the squared magnitude can be written as,

$$\hat{E_S}(m,i) = G(m,i)^2 \times E_X(m,i), \tag{3}$$

where E_S,E_X denote the squared magnitude (Hilbert envelope) of A_X, A_S respectively and $G(m,i)$ denotes noise suppression rule. We use the decision directed approach [9] to obtain $G(m,i)$ as

$$G(m,i) = \frac{\zeta(m,i)}{1+\zeta(m,i)} \tag{4}$$

$$\zeta(m,i) = \alpha \frac{\hat{E_S}(m-\delta,i))}{\hat{E_N}} + (1-\alpha)\big(\gamma(m,i)-1\big) \tag{5}$$

$$\gamma(m,i) = \frac{E_X(m,i)}{\hat{E_N}} \tag{6}$$

where $\hat{E_N}$ denotes the noise floor obtained as mean sub-band envelope in noisy segments (identified by using short-term energy estimates [11]), δ is the hangover constant, $\zeta(m,i)$ and $\gamma(m,i)$ denote the apriori and aposteriori SNR in the sub-band envelope. In our case, we set α as 0.9 and δ as 25 ms.

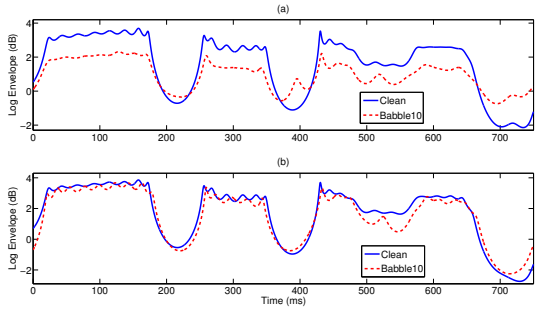

Figure 2: Gain normalized sub-band FDLP envelopes for clean and noisy speech signal (babble 10 dB) (a) without and (b) with MMSE noise suppression.

2.3. Modulation feature extraction

The noise suppressed Hilbert envelope is transformed using DFT into spectral autocorrelations of the sub-band signal, which are used for linear prediction. The order of the linear prediction corresponds to 40 poles per second per sub-band. The steps involved in converting the sub-band DCT signal into AR envelope parameters are referred to as FDLP [8]. In our experiments, we use the gain normalized FDLP envelopes as these are found to be more robust to channel noise [7]. An illustration of the use of the MMSE noise suppression rule on sub-band FDLP envelopes is shown in Fig. 2, where we plot the envelopes from clean speech and noisy speech (babble noise at 10 dB SNR) of a sub-band (500-700Hz) with and without the MMSE noise suppression rule. When MMSE noise suppression is applied, the match between sub-band envelopes extracted from clean and noisy speech is improved.

The sub-band FDLP envelopes are then compressed using a static compression which is a logarithmic function and a dynamic compression scheme [7]. The compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. The temporal envelopes from the two compression streams are then converted into modulation spectral components using DCT, corresponding to the static and the dynamic modulation spectrum. The modulation components form a 420 dimensional feature vector [7].

3. MLP BASED SAD

3.1. MLP training

The MLP estimates the posterior probability of phonemes given the acoustic evidence $P(q_t = i|x_t)$, where q_t denotes the phoneme index at frame t, x_t denotes the feature vector at frame t. We train the MLP using a conversational telephone speech (CTS) database which consists of 130 hours of conversational speech recorded over a telephone channel at 8 kHz [13]. The training data consists of 100 hours of speech and cross-validation data set consists of 30 hours of speech. It is labeled using 45 phonemes (44 speech classes and 1 silence class). The phoneme labels are obtained by force align-

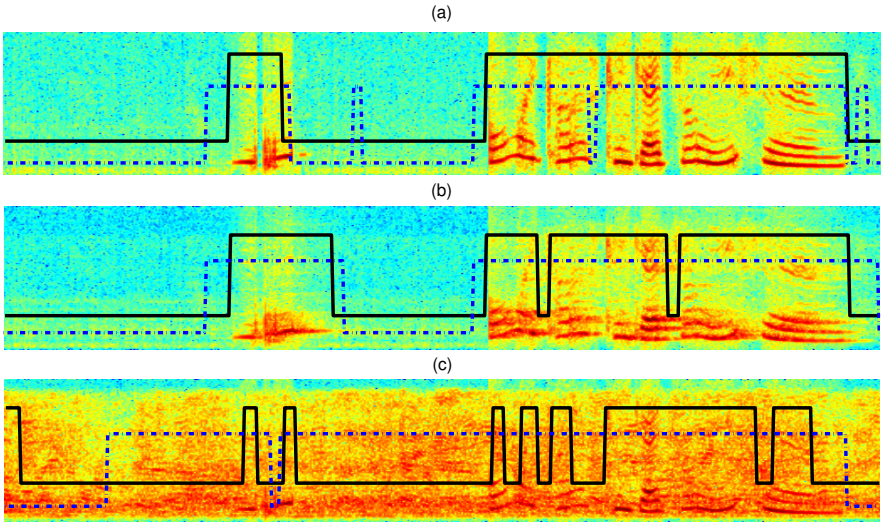

Figure 3: Spectrogram and two SAD outputs for (a) clean speech, (b) its reverberated version (convolved with artificial room response of reverberation time 300 ms), and (c) its noisy version (added with babble noise at 10 dB SNR). The dotted line is adaptive energy-based SAD [3] and the solid line is the proposed MLP-based SAD.

ing the word transcriptions to the previously trained HMM-GMM models [13]. Here, the MLP consists of 5000 hidden neurons, and 45 output neurons (with soft max nonlinearity) representing the phoneme classes. The training data for MLP consists mainly of clean speech segments and thus does not make any assumption of the noisy or reverberated speech.

3.2. SAD system

The trained MLP is used for the SAD system by forward passing the features extracted from the noisy speech. The output 45 dimensional posterior vector is converted to 2 dimensional vector by summing all the probability values for the speech classes. This is hard-thresholded to obtain 0-1 speech activity decisions. These outputs are smoothed using a Viterbi decoder which uses a minimum duration of 7 consecutive frames for speech/non-speech class [6].

Fig. 3 illustrates the performance of the proposed SAD system on a 5 second portion of NIST 2008 test utterance. We also plot the output of the SAD system on the noisy version (additive noise with babble at 10 dB) and a reverberated version (artificial reverberation of 300 ms). The solid line shows the output of the MLP-SAD system and the dotted line shows the same for the energy based SAD [3]. This plot shows that the proposed SAD has less amount of false alarms and is quite robust in noise and reverberation. The energy based SAD creates more false alarms for noisy speech.

4. EXPERIMENTS AND RESULTS

4.1. Speaker verification setup

Speaker verification experiments are performed on the telephone subset of the NIST 2008 speaker recognition evaluation (SRE) dataset, containing 1819 utterances, with both male and female speakers. Verification performance is evaluated on three different conditions (Conditions 6, 7 and 8) of the NIST SRE [12], consisting of telephone speech. Enrollment data is unaltered NIST 2008 clean speech data whereas the test data is corrupted using (a) babble, (b) exhibition hall, (c) restaurant and (d) car noises from the NOISEX-92 database, each resulting in speech at 5, 10, 15 and 20 dB SNR. Three reverberant versions of this dataset are also created by convolving the speech with different room responses obtained from [14] with reverberation time of 300 ms.

The speaker verification system in our experiments is a state-of-the-art GMM-UBM system using 400 dimensional i-vectors for speech representation [10]. Gaussian PLDA is applied to reduce the dimension of the i-vectors to 150 dimensions and likelihood scores are computed on these. In order to train the UBM and the total variability matrix used in the i-vector estimation, development data from NIST 2004 SRE, Switchboard II Phase III and NIST 2006 SRE is used. We use the short-term FDLP features for the speaker verification system [15]. During evaluation, the enrollment data is processed using the SAD decisions obtained from NIST (speech recognition outputs) whereas the SAD for the test data is derived using the proposed MLP system as well as the other SAD techniques. All results were obtained without any score normalization.

4.2. Performance evaluation

Performance of five SAD systems are compared here: (a) adaptive energy-based [3], (b) mel spectrum based [4], (c) multi-scale spectro-temporal modulations based on auditory processing [5], (d) the proposed MLP based SAD system with 9 frame context of MFCC features processed with cepstral mean subtraction (MLP1) and (e) the proposed MLP based SAD system with FDLP features (MLP2). The SAD threshold for these systems were kept at the preset value provided (calibrated to provide good performance in noisy conditions).

Since the SAD decision threshold in a speaker verification system is fixed in practice, the SAD performance is measured in terms of the speaker verification equal error rate (EER). Furthermore, we also evaluate the accuracy of the SAD decisions at the frame-level by comparing them with those provided by NIST. The NIST SAD (speech recognition output) is computed on clean speech and is considered to be the ground-truth for SAD evaluation. The SAD error is defined as the average of the false alarm rate and miss rate computed over all test utterances.

C6: Train/test in multiple languages						C7: English train/test					C8: Native U.S. English train/test				
Cond.	Egy.	Mel.	Aud.	MLP1	MLP2	Egy.	Mel.	Aud.	MLP1	MLP2	Egy.	Mel.	Aud.	MLP1	MLP2
Clean	10.7	12.4	12.5	8.6	8.9	4.0	5.1	7.1	2.9	2.9	2.9	4.3	6.8	1.4	1.4
5 dB	23.2	23.5	26.3	23.1	21.0	19.0	17.7	22.1	16.6	14.7	20.3	17.9	23.1	18.6	16.0
10 dB	17.9	16.5	16.5	15.2	13.5	12.3	10.2	11.0	9.1	7.5	13.6	9.1	11.9	9.9	8.5
15 dB	14.5	12.4	11.9	11.6	11.6	8.1	5.9	5.8	5.4	5.3	9.6	5.6	5.4	5.2	4.5
20 dB	12.4	10.7	10.6	9.9	10.5	6.0	4.4	3.9	3.4	3.5	6.3	3.8	2.9	2.7	3.2
Revb.	18.8	21.4	21.3	16.4	16.0	12.7	14.8	17.0	9.7	10.1	14.9	17.5	20.1	11.2	11.6

Table 1: Speaker verification performance in terms of EER, utilizing various SAD systems. For noisy and reverberant conditions, performance is averaged over four noise types and three different room responses respectively.

Cond.	Egy.	Mel.	Aud.	MLP1	MLP2
Clean	23.5	28.5	34.0	10.5	11.5
5 dB	31.0	28.5	29.0	25.0	23.0
10 dB	27.5	21.5	21.5	17.5	17.0
15 dB	25.0	22.0	19.0	12.5	13.5
20 dB	23.5	15.0	19.0	11.3	12.0
Revb.	22.0	32.0	36.5	11.0	15.0

Table 2: SAD error computed as average of false alarms rate and miss rate for different systems under various conditions.

4.3. Results and discussion

The speaker verification results obtained using the four SAD systems considered in this paper are given in Table 1. The performance in noisy conditions for a particular SNR is averaged across the four different types of noise. Similarly, performance in reverberation is averaged across the three room responses. From Table 1, it can be seen that the proposed MLP-based SAD (with FDLP features or MFCC features) outperforms the other SAD methods in speaker verification performance (relative EER improvements of 9 % in additive noise, 19% in reverberant conditions and about 31 % in clean conditions). However, it is also interesting to note that the improvements for the MLP-SAD system is relatively less for C6 where the test data can come from multiple languages as the MLP is trained only on English CTS.

The SAD errors for different techniques (obtained at the operating threshold) are reported in Table 2. It can be seen that percentage SAD error is considerably less for the proposed MLP-SAD system. This is because of the reduced false alarm rate for a fixed miss rate as compared with other systems. As false alarms in SAD decisions are reduced, the resulting speaker verification system is able to perform speaker validation more on the speech regions as opposed to non-speech regions.

5. CONCLUSIONS

In this paper, we have investigated an MLP-based speech activity detector and applied it for a speaker verification task. The posterior probabilities obtained from the MLP are merged into two classes to form SAD outputs. The proposed SAD system results in improved speaker verification performance, when compared to other SAD systems. These systems are also evaluated based on SAD error where the MLP system shows good robustness.

6. REFERENCES

[1] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," *Proc. ICASSP*, pp. 323–326, 1977.

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters* , Vol. 6 (1), pp. 1-3, 1999.

[3] D. Reynolds et al. "The 2004 MIT Lincoln laboratory speaker recognition system", *Proc. ICASSP*, pp. 177-180, 2005.

[4] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *Proc. ICASSP*, pp. 153-156, 1995.

[5] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech and Language Process.*, Vol. 14(3), pp. 920-930, 2006.

[6] H. Boulard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[7] S. Ganapathy, S. Thomas and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum", *Jnl. Acoust. Soc. of America*, Vol. 128 (6), pp. 3769-3780, 2010.

[8] M. Athineos and D.P.W. Ellis, "Autoregressive modelling of temporal envelopes", *IEEE Trans. Signal Proc.*, Vol. 55 (11), pp. 5237-5245, 2007.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-32, pp. 1109-1121, 1984.

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

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

[12] The NIST 2008 Evaluation Plan, available online (`http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf`)

[13] T. Hain *et al.*, "The Development of AMI System for Transcription of Speech in Meetings", *Proc. MLMI*, pp. 344-356, 2005.

[14] R. Dhillon, S. Bhagat, R. Carvey, and E. Shriberg, The ICSI Meeting Recorder Project, `http://www.icsi.berkeley.edu/Speech/mr`, 2002.

[15] S. Ganapathy, J. Pelecanos and M.K. Omar, "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, Prague, 2011.