This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Single Frequency Filtering Approach for Discriminating Speech and Nonspeech

G. Aneeja and B. Yegnanarayana, *Fellow, IEEE*

***Abstract*—In this paper, a signal processing approach is proposed for speech/nonspeech discrimination. The approach is based on single frequency filtering (SFF), where the amplitude envelope of the signal is obtained at each frequency with high temporal and spectral resolution. This high resolution property helps to exploit the resulting high signal-to-noise ratio (SNR) regions in time and frequency. The variance of the spectral information across frequency is higher for speech and lower for many types of noises. The mean and variance of the noise-compensated weighted envelopes are computed across frequency at each time instant. Decision logic is applied to the feature derived from the mean and variance values on varieties of degradations, including NTIMIT, CTIMIT and distance speech, besides degradation due to standard noise types. In all cases, the proposed method gives significantly better performance than the standard Adaptive Multi-rate VAD2 (AMR2) method. AMR2 method is chosen for comparison, as the method adapts itself for different degradations, and is seen to give good performance over different SNR situations. The proposed method does not use training data to derive the characteristics of speech or noise, nor makes any assumption on the nonspeech beginning. The SFF method appears promising in other applications of speech processing, such as pitch extraction and speech enhancement.**

***Index Terms*—Speech/nonspeech discrimination, single frequency filtering (SFF), voice activity detection (VAD), weighted component envelope, spectral variance, temporal variance.**

I. Introduction

The objective of voice activity detection (VAD) is to determine regions of speech in the acoustic signal, even when the signal is corrupted by additive or other types of degradations. VAD is an essential first step for development of speech systems such as speech and speaker recognition. Human listeners are able to distinguish speech and nonspeech regions by interpreting the signal in terms of speech characteristics, as well as the context. If a machine has to discriminate these two regions, it has to depend only on the characteristics of speech and degradation. It is difficult to make a machine use the accumulated knowledge of a human listener for this purpose.

Robustness of a VAD algorithm depends on the type of degradation, the features extracted from the signal and the models used to discriminate speech and nonspeech regions. The acoustic features are usually based on the signal energy in different frequency bands, which includes standard melfrequency cepstral coefficients (MFCC's) [1]. Features based on speech characteristics such as voicing and dynamic spectral characteristics have also been explored [2], [3]. In [2], the phase of the Fourier Transform is averaged over a window to compensate for phase wrapping, and then processed over mel-frequency bands. The phase information gives performance similar to MFCCs even in the cases of degradation. But combination of MFCCs and phase information seems to have improved the performance. Some attempts have been made to explore features in the excitation component of speech signal [4]. Features of the discrete wavelet transform and Teager energy operator have also been proposed for VAD with good results [5], [6]. Characteristics of speech and noise can be captured well if the samples are collected over long ($>$1 sec) durations, as some of the studies below indicate. For example, the long-term divergence measure (LTDM) measures the spectral divergence between speech and noise over longer duration [7]. The LTDM measure is calculated as the ratio of the long-term spectral energies of speech and noise over different frequency bands. More recently long-term spectral variability has been suggested for VAD [8]. The long-term feature is the variance across frequency of the entropy computed over 300 msec of speech at each frequency. It was shown to be robust at low signal-to-noise ratio (SNR) conditions for a variety of noise degradations. The long-term signal variability (LTSV) was extended to multi-band long-term signal variability to accommodate multiple spectral resolutions [9]. The long-term spectral variability feature together with contextual, discriminative and spectral cues was shown to give further improvement in performance of VAD [10]. New features like Multi-Resolution cochleagram (MRCG) along with boosted Deep Neural Networks (bDNNs) have been proposed recently for VAD, which are shown to outperform the state-of-the-art VADs even at low SNRs, for babble and factory noises [11], [12]. The MRCG feature is derived using features at multiple spectrotemporal resolutions [11] and the bDNN uses aggregate of predictions of multiple weak classifiers [12].

In [13], a low variance for spectral estimate is assumed for noise, and large amount of data is used for training. But low variance criterion for noise may not be applicable for machine gun noise and some other non-stationary noises, including distant speech. The method proposed in [13] assumes a non-speech beginning to estimate the noise statistics. Other models are also considered for speech and nonspeech discrimination, which include artificial neural networks (ANNs) [14], Gaussian mixture models (GMMs) [15], and deep belief networks (DBNs) [16].

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However permission to use this material fro any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

G. Aneeja is with International Institute of Information Technology, Hyderabad, 500 032, India, Ph:+91-40-6653-1418, E-mail: aneeja.g@research.iiit.ac.in.

B. Yegnanarayana is with International Institute of Information Technology, Hyderabad, 500 032, India, Ph:+91-40-6653-1271, E-mail: yegna@iiit.ac.in.

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Several attempts have been made to improve the performance of VAD, by exploiting the statistics of speech and noise characteristics [17]. One such method is the statistical model-based VAD, and its refinements proposed in [18]. Statistical methods work well if labelled training data for speech and nonspeech in different noise conditions are available for training the models. These are called supervised learning systems [19]. In some cases, the noise model derived from training data is used for initialization process. These methods are called semi-supervised learning [17]. Methods based on universal models of speech, without assuming any specific type of noise, are also proposed. In [20], non-negative matrix factorization (NNMF) approach is used to develop universal speech model. In practice, it is preferable to develop a VAD algorithm that can operate without any training data, i.e., unsupervised learning.

Most of the VAD algorithms are tested on data with simulated degradation, either by adding noise or by passing the clean signal through a degrading channel. This is necessary to evaluate new methods in comparison with known/existing methods. Very few attempts have been made to assess the performance of a VAD algorithm with data collected in practical environments. The degradations in such environments may not fit into any standard model. Moreover, it is difficult to obtain ground truth in practice to evaluate the VAD methods. In general, the characteristics of the environment in which speech signal is produced vary, and hence are not predictable to model. The only option available is to develop VAD algorithm by exploiting the characteristics of speech that may be present even in the degraded signal. For this, the features of excitation source and dynamic vocal tract system need to be explored for robustness against degradation. Also, it is necessary to develop methods to extract those features from degraded signals.

In this paper, a signal processing approach is proposed to highlight the features of speech even in degraded signal. The method extracts the temporal variation of signal energy at each frequency. The characteristics of speech signal (due to correlations among speech samples) at each frequency are distinctly different from the characteristics of noise (due to uncorrelatedness in noise samples in many cases) at each frequency. The SNR of the speech signal is high at some frequencies, compared to noise. The high SNR property of speech at several single frequencies is exploited. Since the method is based on extracting energy at a single frequency, it is called single frequency filtering (SFF) method. Note that single frequency information can also be derived by computing the discrete Fourier transform (DFT) over a block of data at every sampling instant. Other methods of deriving similar information include gammatone filters [21]. The temporal variation of signal energy at each frequency is processed further to compensate for the effects of noise in that band by determining a weighting factor for each band. The mean and variance of the weighted signal energy across frequency at each sampling instant are used to derive a parameter contour as a function of time, to discriminate between speech and nonspeech regions. An adaptive threshold is derived from the parameter contour for each utterance, followed by a decision logic based on the features of speech and noise in the given utterance. The method is tested using simulated degradations on speech signals, and also using speech signals collected in practical environments. Since the method exploits the properties of the speech signal, it is not necessary to have training data of speech and nonspeech signals to build models. The present approach does not rely on the appended silence/noise regions to estimate the noise characteristics.

Many studies in literature compare VAD algorithms with the Adaptive Multi-Rate (AMR) method [22]. The comparison is done mainly at the score level. To have a fair comparison with the AMR method, the VAD algorithms should consider the following other factors of the AMR method into account:

- Adaptability: AMR method is adaptable to various types of noise, SNRs and environments.
- No prior information: It does not require training data or any other prior information about the type of noise.
- Automatic threshold: The threshold estimation does not require nonspeech beginning, and also does not use data for training of statistical models.

In section 2, speech data collected in different types of degradation is described. Section 3 discusses the basis for the proposed single frequency filtering (SFF) method for processing the signals. Section 4 gives the development of the proposed VAD algorithm. Section 5 gives results of evaluation of the SFF-based method of VAD in comparison with the AMR2 method for different types of degradations. This section also includes a discussion on relative performance of SFF, DFT and gammatone filtering methods of deriving information in different frequency bands. Section 6 gives a summary, and indicates how the proposed SFF method can be exploited for other speech processing applications.

II. Different types of degradation

In this section different speech and noise databases and their characteristics are discussed to indicate the variety of degradations considered for evaluation of the proposed VAD algorithm. Note that, although some of the data was collected at 16 kHz sampling rate and other data at 8 kHz sampling rate, the frequencies in the range 300 - 4000 Hz are considered in both the cases as explained in section IV-A.

A. Adding degradation at different SNRs to clean speech signal.

The TIMIT test corpus is used for evaluation [23]. The sampling rate is 16 kHz. A VAD algorithm should ideally accept speech and also reject nonspeech. In a situation where there is more duration of speech than nonspeech, then if the algorithm has a higher speech acceptance, then the algorithm shows better performance even if the performance of nonspeech rejection is poor. A similar situation of better performance would arise for longer duration of nonspeech, with higher nonspeech detection rate and lower speech detection rate of the algorithm. To overcome this problem, each TIMIT utterance is appended with 2 sec of silence at the beginning and end of the utterance as in [8]. Various samples of the thirteen types of noises from NOISEX-92 database [24] are added to the clean TIMIT speech signal at SNRs of -10 dB and

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

5 dB, to create degraded speech signals. The TIMIT data provides boundaries of the phone labels, which are generated automatically and are then hand corrected by experienced acoustic phoneticians. Hence these boundaries are used as ground truth for comparing the results of the proposed VAD algorithm on the noisy speech data. The silence and pause labels are considered as nonspeech.

Most VAD algorithms use post processing techniques like hangover scheme. The hangover scheme is used to reduce the risk of lower energy regions of speech at the ends of speech regions being falsely rejected [13]. This is based on the assumption that speech frames are highly correlated in time [13], [17]. In hangover schemes decisions at the frame level are smoothed by considering sequence of frames to arrive at a final decision. Hangover schemes are applied to the VAD algorithm after the initial VAD decision. In some regions, the features of speech might not be evident even in clean speech, although those regions are labelled as speech in the database. The ground truth given in TIMIT database may not be a perfect reference for comparing results of any VAD algorithm. This may be due to mismatch between the perceptual evidence and speech data in manual labelling. Hence the accuracy will not be 100% even in the case of clean speech.

B. Telephone channel database.

NTIMIT (Network TIMIT) database [25] was collected by transmitting TIMIT data over telephone network. Speech utterances are transmitted from a laboratory to a central office and then back from the central office to the laboratory, thus creating a loopback telephone path from laboratory to a large number of central offices. These central offices were geographically distributed to simulate different telephone network conditions. Half of the TIMIT database was sent over local telephone paths, while the other half was transmitted over long distance paths. All recordings were done in an acoustically isolated room. The NTIMIT test corpus is used for VAD evaluation. The sampling rate is 16 kHz. In the NTIMIT case, 2 sec silence segments are not appended to the data, as this kind of degradation can not be simulated in the appended regions. The ground truth for the NTIMIT is same as for the TIMIT data.

C. Cellphone channel database.

The CTIMIT read speech corpus [26] was designed to provide a large phonetically-labelled database for use in the design and evaluation of speech processing systems operating in diverse, often hostile, cellular telephone environments. CTIMIT was generated by transmitting and redigitizing 3367 of the 6300 original TIMIT utterances over cellular telephone channels from a specially equipped van, in a variety of driving conditions, traffic conditions and cell sites in southern New Hampshire and Massachussetts. The recorded data was played in the van over a loudspeaker and cellular handset combination. Each received call was digitized at 8 kHz, segmented and time-aligned with the original TIMIT utterances. The ground truth of TIMIT labels can be used here also. CTIMIT test corpus is used for VAD evaluation [26]. Note that here also the 2 sec silence segments are not appended to the data, as in the case of NTIMIT database.

D. Distant speech.

The differences between the characteristics of speech signal collected by a distant microphone (DM) and that collected by a close-speaking microphone (CM) are as follows: (a) The effects of radiation at far-field are different from those at the near-field. (b) The SNR is lower in the DM speech signal due to additive background noise. (c) The reverberant component in the DM speech signal is also significant, due to reflections, diffuse sound and reduction in amplitude of the direct sound. (d) The DM speech signal may also be affected due to interference from speech of other speakers present in the room. Hence, the acoustic features derived from the DM speech signal are not same as those derived from the corresponding CM speech signal.

Speech signals from SPEECON database are used for evaluation of the VAD algorithm for distant speech [27]. The signals were collected in three different cases, namely, car interior, office and living rooms (denoted by public). The signals were collected simultaneously using a close-speaking microphone (a microphone placed just below the chin of the speaker), and microphones placed at distances of 1 meter, 2 meters and 3 meters from the speaker. These four cases are denoted by C0, C1, C2 and C3, respectively. Each case has 1020 utterances. Speech signals collected in the office environment are affected by noises generated by computer fans and air-conditioning. Speech signals collected in living rooms are affected by babble noise and music (due to radio or television sets). Reverberation is present mostly in the office and living room environments. The estimated reverberation time in these environments varied from 250 msec to 1.2 sec. The average SNR measured at the close speaking microphone (C0) was around 30 dB, while that measured at distances of 2 meters to 3 meters was in the range 0 - 5 dB. The database consists of speech signals collected from 30 male and 30 female speakers. For each speaker, 17 utterances were recorded, resulting in about one minute of speech data per speaker. People were asked to record free spontaneous items, elicited spontaneous items, read speech and core words. A manual voiced-unvoiced-nonspeech labels are marked for every 1 msec in the SPEECON database for C0 case. Since speech at all the distances are simultaneously collected, the same labels are used for the data at all distances. The manual labels (voiced-unvoiced labels for speech and nonspeech label for the rest) form the ground truth for the data at all distances. The sampling rate is 16 kHz. Since the utterances of each speaker are from different environments, it is not possible to build statistical models with this kind of data. No silence data is appended in this case also.

III. Basis for single frequency filtering approach

Speech signal has dependencies both along time and along frequency. This results in signal to noise power ratio to be a function of time as well as a function of frequency. For an ideal noise of a given total power, the power gets divided equally over frequency, whereas for a signal, the power is distributed

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

nonuniformly across frequency. Thus $\frac{S^2(f)}{N^2(f)}$ is higher in some frequencies and lower in some other frequency regions, where $S(f)$ and $N(f)$ are signal and noise amplitudes as a function of frequency. This gives a much higher value for the average of $\frac{S^2(f)}{N^2(f)}$ over a frequency range, compared to the ratio of total signal power to total noise power over the entire frequency range.

Let

$$\alpha = \int_{f_0}^{f_L} \frac{S^2(f)}{N^2(f)}\, df, \tag{1}$$

$$\beta = \sum_{i=0}^{L-1} \frac{\int_{f_i}^{f_{i+1}} S^2(f)\, df}{\int_{f_i}^{f_{i+1}} N^2(f)\, df}, \tag{2}$$

and

$$\gamma = \frac{\int_{f_0}^{f_L} S^2(f)\, df}{\int_{f_0}^{f_L} N^2(f)\, df}, \tag{3}$$

where $(f_i - f_{i+1})$ is the $(i+1)^{th}$ interval of the L nonoverlapping frequency bands, and $i = 0, 1, ..., L-1$. The following inequality holds good.

$$\alpha \geq \beta \geq \gamma. \tag{4}$$

TABLE I. Values of $\overline{\alpha}, \overline{\beta}, \overline{\gamma}$ for speech signal degraded at -10 dB SNR using DFT approach.

NOISE	$\overline{\alpha}$	$\overline{\beta}$	$\overline{\gamma}$
white	2.289	1.1224	1.103
babble	237.4061	8.518	1.1481
volvo	3698.2985	233.7515	1.4089
leopard	1599.8217	35.0032	1.143
buccaneer1	277.1179	1.1356	1.1166
buccaneer2	5.4785	1.1299	1.0975
pink	2.2999	1.1034	1.1105
hfchannel	132.6712	1.8899	1.1094
m109	79.2124	2.4393	1.1294
f16	34927.2057	1.3335	1.1098
factory1	406.8931	1.2621	1.1199
factory2	66.5143	3.626	1.1703
machine gun	186034.6397	12056.4657	71.9059

TABLE II. Values of $\overline{\alpha}, \overline{\beta}, \overline{\gamma}$ for speech signal degraded at -10 dB SNR using SFF approach.

NOISE	$\overline{\alpha}$	$\overline{\beta}$	$\overline{\gamma}$
white	3.9853	1.1317	1.1034
babble	804.0397	3.992	1.1596
volvo	299.3464	44.2498	1.4496
leopard	117.9238	10.9565	1.156
buccaneer1	72.916	1.1395	1.1184
buccaneer2	3.1214	1.134	1.0978
pink	4.9493	1.1131	1.1112
hfchannel	17.8483	1.6226	1.1113
m109	59.2573	2.0275	1.1414
f16	18.2013	1.2939	1.1126
factory1	4.9478	1.2516	1.1235
factory2	21.4015	2.3495	1.1777
machine gun	10642.4183	753.7107	69.1977

The $S(f)$ and $N(f)$ are computed for degraded speech utterance and for noise using 512-point DFT of Hann windowed segments of size 20 msec for *every sample shift* using $L = 16$. In Table I, the mean values $\overline{\alpha}, \overline{\beta}, \overline{\gamma}$ of α, β and γ respectively, computed over the entire utterance are given. It is clear that $\overline{\alpha} \geq \beta \geq \overline{\gamma}$ for different types of noises. In the case of uniform noise, (eg white), the values of $\overline{\alpha}, \overline{\beta}, \overline{\gamma}$ are lower than the values for the nonstationary noises (eg volvo and machine gun). In the case of some nonstationary noises, the floor value is low at some frequencies which makes the denominator $N(f)$ small. With small values of the denominator, the ratios of α, β, γ are relatively higher as observed in Table I from the values of $\overline{\alpha}$, $\overline{\beta}, \overline{\gamma}$ for volvo and machine gun noises. It is also interesting to note that for nonuniformly distributed noises, such as machine gun, f16 and volvo, the $\overline{\alpha}$ and $\overline{\beta}$ values are much higher than for the more uniformly distributed noises, such as white, pink and buccaneer2, whereas the corresponding $\overline{\gamma}$ values are low in all cases. This is due to regions having high $\frac{S(f)}{N(f)}$ in the time and frequency domains for nonuniformly distributed noises.

The signal and noise power as a function of frequency can be computed using either by block processing as in the DFT, or by filtering through SFF, as described in the next section. Table II shows that the inequality (4) holds good for SFF approach also. Both the DFT and SFF based approaches are expected to give similar results. The SFF approach is used here, as it may avoid some effects due to block processing. Also, the computation of SFF is faster compared to the computation of DFT at each sampling instant.

IV. Proposed VAD algorithm

A. Envelope of speech signal at each frequency.

The discrete-time speech signal $s(n)$ is differenced, and the differenced signal is denoted by $x(n) = s(n) - s(n-1)$. The sampling frequency is f_s. The signal $x(n)$ is multiplied by a complex sinusoid of a given normalised frequency $\bar{\omega}_k$. The resulting operation in the time domain is given by

$$x_k(n) = x(n) e^{j\bar{\omega}_k n}, \tag{5}$$

where

$$\bar{\omega}_k = \frac{2\pi \bar{f}_k}{f_s}. \tag{6}$$

Since we multiplied $x(n)$ by $e^{j\bar{\omega}_k n}$, the resulting spectrum of $x_k(n)$ is a shifted spectrum of $x(n)$. That is,

$$X_k(\omega) = X(\omega - \bar{\omega}_k), \tag{7}$$

where $X_k(\omega)$ and $X(\omega)$ are spectra of $x_k(n)$ and $x(n)$, respectively.

The signal $x_k(n)$ is passed through a single-pole filter, whose transfer function is given by

$$H(z) = \frac{1}{1 + rz^{-1}}. \tag{8}$$

The single-pole filter has a pole on the real axis at a distance of r from the origin. The the location of the root is at $z = -r$ in the z-plane, which corresponds to half the sampling frequency i.e., $f_s/2$. The output $y_k(n)$ of the filter is given by

$$y_k(n) = -r y_k(n-1) + x_k(n). \tag{9}$$

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

The envelope of the signal $y_k(n)$ is given by

$$e_k(n) = \sqrt{{y_{kr}}^2(n) + {y_{ki}}^2(n)}, \tag{10}$$

where $y_{kr}(n)$ and $y_{ki}(n)$ are the real and imaginary components of $y_k(n)$. Since the filtering of $x_k(n)$ is done at $f_s/2$, the above envelope $e_k(n)$ corresponds to the envelope of the signal $x_k(n)$ filtered at a desired frequency of

$$f_k = \frac{f_s}{2} - \bar{f}_k. \tag{11}$$

The above method of estimating the envelope of the component at a frequency f_k is termed as single frequency filtering (SFF) approach. The choice of the filter with a pole at $z = -r$ for estimating the envelopes of the filtered signals is likely to be more accurate, as the envelopes are computed at the highest frequency ($f_s/2$) possible. Also, choosing a filter at a fixed frequency for any desired frequency f_k avoids scaling effects due to different gains of the filters at different frequencies. If the pole is chosen on the unit circle, i.e., $z = r = -1$, it may result in the filtered output becoming unstable. The stability of the filter is ensured by pushing the pole slightly inside the unit circle. Hence r is chosen as 0.99.

In this study, the envelope is computed at every 20 Hz in the range 300 Hz to 4000 Hz as a function of time. The frequency range 300 - 4000 Hz is chosen, as it covers the useful spectral band of speech. Thus we have envelopes for 185 frequencies as a function of time. In principle, the envelope can be computed at any desired frequency.

B. Weighted component envelopes of speech signal.

Since speech signal has large dynamic range in the frequency domain, the signal may have high power at some frequencies at each instant. At those frequencies the SNR will be higher, as the noise power is likely to be less due to more uniform distribution of the power. Even for noises with nonuniform distribution of power, the lower correlations of noise samples result in a lower dynamic range in the spread of noise power across frequencies, compared to speech. Note that the spectral dynamic range gives an indication of the correlation of the samples in the time domain.

The noise power creates a floor for the envelope at each frequency, and the floor level depends on the power distribution of noise across frequency. The floor is more uniform across time if the noise is nearly stationary. Even if the noise is nonstationary, it is relatively stationary over larger intervals of time than in speech. In such cases, the floor level can be computed over long time intervals at each frequency, if needed.

To compensate for the effect of noise, a weight value at each frequency is computed using the floor value. For each utterance, the mean (μ_k) of the lower 20% of the values of the envelope at each frequency f_k is used to compute the normalised weight value w_k at that frequency. The choice of 20% of the values is based on the assumption that there is at least 20% of silence in the speech utterance. The normalised weight value at each frequency is given by

$$w_k = \frac{\frac{1}{\mu_k}}{\sum_{l=1}^{N} \frac{1}{\mu_l}}, \tag{12}$$

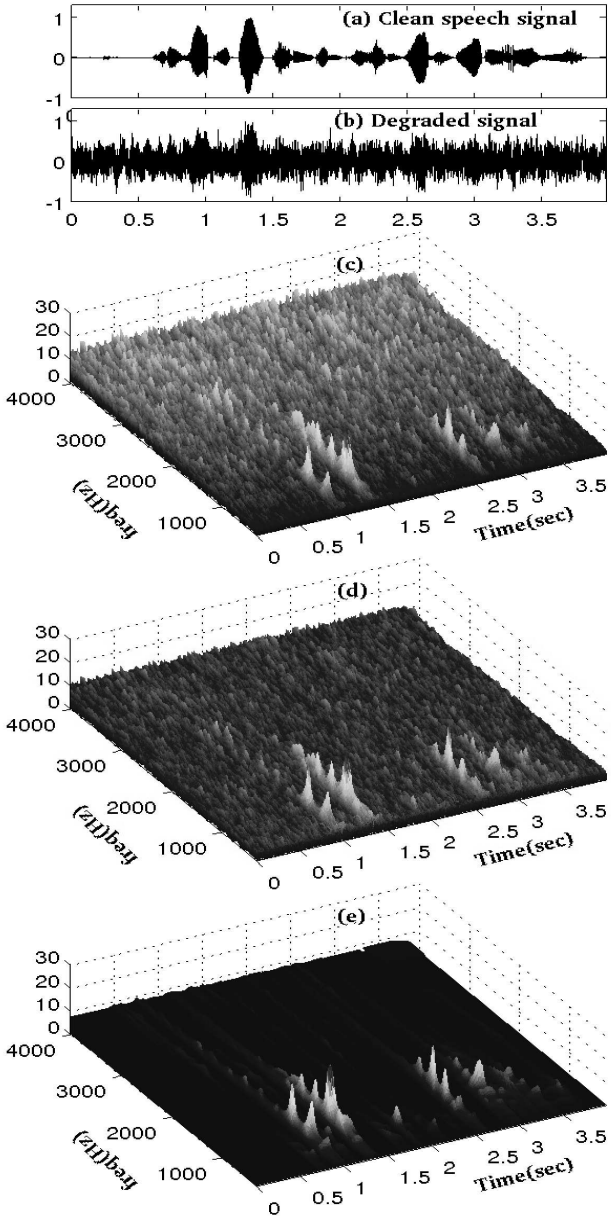

Fig. 1. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at -10 dB SNR. (c) Envelopes as a function of time. (d) Corresponding weighted envelopes. (e) Envelopes as a function of time for clean speech shown in (a).

where N is the number of channels. The envelope $e_k(n)$ at each frequency f_k is multiplied with the weight value w_k to compensate for the noise level at that frequency. The resulting envelope is termed as weighted component envelope. Note that by this weighting, the envelope at each frequency is divided by the estimate of the noise floor (μ_k). Fig. 1 shows the envelopes and the corresponding weighted envelopes at different frequencies for a speech signal degraded by pink noise at -10 dB SNR, along with the envelopes for clean speech. It is observed that features of speech are reflected better in the weighted envelopes (Fig. 1(d)), as the weighting reduces the effects of noise. The envelopes are scaled to the same value for comparison.

A small amount of white noise (at 100 dB SNR) is added to all the signals (after appending with zeros in the case of

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

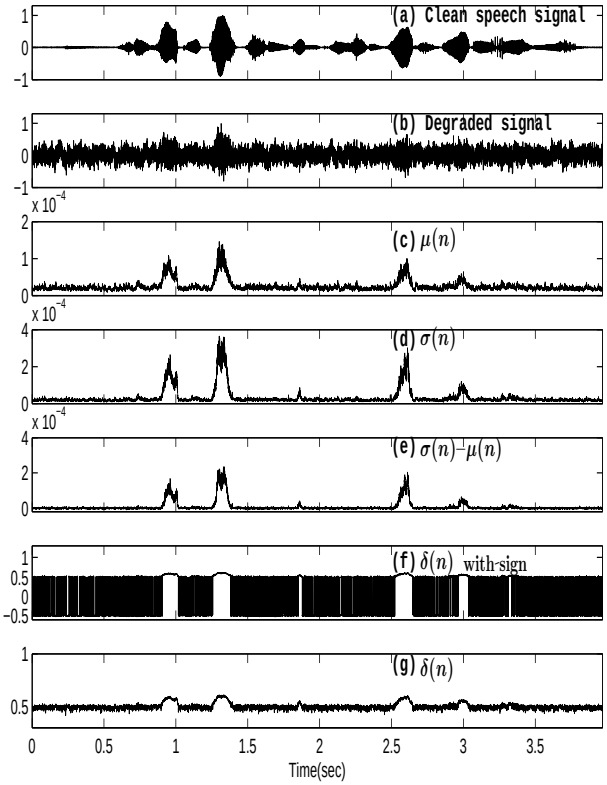

Fig. 2. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at **-10 dB** SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

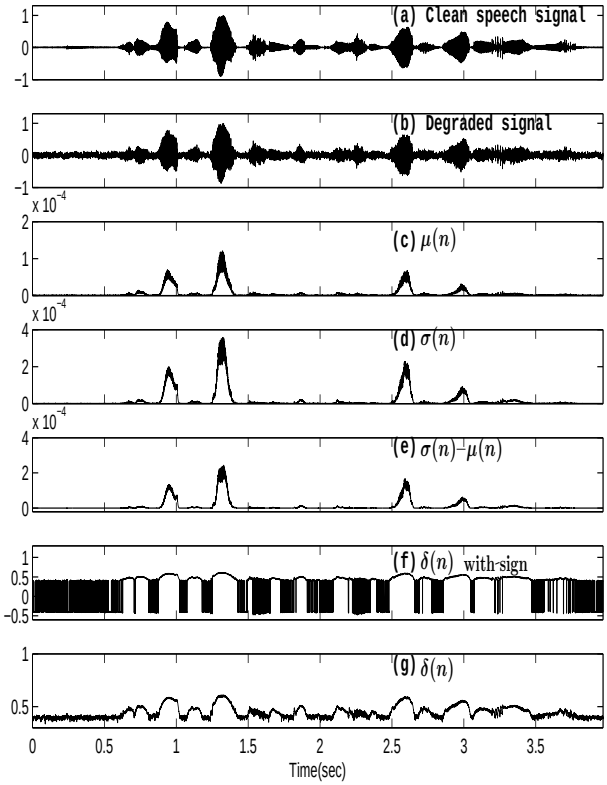

Fig. 3. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at **5 dB** SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

TIMIT utterances) to ensure that the floor value is not zero. For the computation of w_k, the values in the appended silence regions are not considered.

At each time instant, the mean ($\mu(n)$) of the square of the weighted component envelopes computed across frequency corresponds approximately to the energy of the signal at the instant (Fig. 2(c)). The $\mu(n)$ is expected to be higher for speech than for noise in the regions where speech signal is present, as the noise components are deweighted. At each time instant, the standard deviation ($\sigma(n)$) of the square of the weighted component envelopes computed across frequency will also be relatively higher for speech than for noise in the regions of speech due to formant structure (Fig. 2(d)). Hence $(\sigma(n) + \mu(n))$ is generally higher in the speech regions, and lower in the nonspeech regions. Since the spread of noise (after compensation) is expected to be lower, it is observed that the values of $(\sigma(n) - \mu(n))$ are usually lower in the non-speech regions compared to the values in the speech regions (Fig. 2(e)). Multiplying $(\sigma(n) + \mu(n))$ with $(\sigma(n) - \mu(n))$ gives $(\sigma^2(n) - \mu^2(n))$, which highlights the contrast between speech and nonspeech regions. Figs. 2 and 3 illustrate the features of $\mu(n)$, $\sigma(n)$ and $(\sigma(n) - \mu(n))$ for an utterance corrupted by pink noise at SNR = -10dB and SNR = 5dB, respectively.

Due to large dynamic range of the values of $(\sigma^2(n) - \mu^2(n))$, it is difficult to observe the speech regions with small values of $(\sigma^2(n) - \mu^2(n))$. To highlight the contrast between speech and nonspeech regions, the dynamic range is reduced by computing

$$\delta(n) = \sqrt[M]{|(\sigma^2(n) - \mu^2(n))|}, \tag{13}$$

where M is chosen as 64.

The value of M is not critical. Any value of M in the range of 32 to 256 seems to provide good contrast between speech and nonspeech regions in the plot of $\delta(n)$. In computing $\delta(n)$, only the magnitude of $(\sigma^2(n) - \mu^2(n))$ is considered. If the sign of $(\sigma^2(n) - \mu^2(n))$ is assigned to $\delta(n)$, the values will be fluctuating around zero in the nonspeech regions for most types of noise (see Fig. 2(f) for pink noise), but the short time (20 - 40 msec) temporal average value will be small and fluctuating, making the noise floor uneven. This makes it difficult to set a threshold for deciding nonspeech regions. The values of $\delta(n)$ will have a high temporal mean value in the nonspeech regions, with small temporal variance (Fig. 2(g)). This helps to set a suitable threshold to isolate nonspeech regions from speech regions. The range of $\delta(n)$ with sign value (Fig. 2(f)) is different from $\delta(n)$ values (Fig. 2(g)). The small temporal spread of $\delta(n)$ values in the nonspeech regions and its mean value helps to fix a suitable threshold. The $\delta(n)$ values in the nonspeech regions is dictated by the noise level. The $\delta(n)$ values in nonspeech regions are high for pink noise degradation at -10 dB SNR (Fig. 2(g)) than at 5 dB SNR (Fig. 3(g)). Note that, by considering the $\delta(n)$ values without sign, we are losing some advantage in the discrimination of nonspeech regions, which has both positive and negative values, compared to speech regions which have

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

mostly positive values. The $\delta(n)$ values with $M = 64$ are used for further processing for decision making. Note the changes in the vertical scales in Figs. 2(f) and 2(g), and also in Figs. 3(f) and 3(g), to understand the significance of using the absolute value, i.e., $\delta(n)$ without sign.

C. Decision logic.

The decision logic is based on $\delta(n)$ for each utterance, by first deriving the threshold over the assumed (20% of the low energy) regions of noise, and then applying the threshold on temporally smoothed $\delta(n)$ values. The window size l_w used for smoothing $\delta(n)$ is adapted based on an estimate of the dynamic range (ρ) of the energy of the noisy signal in each utterance, assuming that there is at least 20% silence region in the utterance. The binary decision of speech and nonspeech at each time instant, denoted as 1 and 0, respectively, is further smoothed (similar to hangover scheme) using an adaptive window, to arrive at the final decision. The following 5 steps describe the implementation details of the decision logic:

1) Computation of threshold (θ):
Compute the mean (μ_θ) and variance (σ_θ) of the lower 20% of the values of $\delta(n)$ over an utterance. A threshold of $\theta = \mu_\theta + 3\sigma_\theta$ is used in all cases. The θ value depends on each utterance. Thus the threshold value, corresponding to the floor value of $\delta(n)$, is adapted to each utterance, depending on the characteristics of speech and noise in that utterance.
2) Determination of smoothing window l_w:
The energy E_m of the signal $x(n)$ is computed over a frame of 300 msec for a frame shift of 10 msec, where m is the frame index. The dynamic range (ρ) of the signal is computed as

$$\rho = 10 \log_{10} \frac{\max_m (E_m)}{\min_m (E_m)}. \quad (14)$$

The window length parameter l_w for smoothing is obtained from the dynamic range (ρ) of the signal. Table III gives the ρ values for degraded speech at SNRs of -10 dB and 5 dB for different noises. The ρ values are high at 5 dB SNR compared to the values at -10 dB SNR for the same noise. The ρ values vary for different noises for the same SNR, because the degradation characteristics of noises vary. For distance speech, the histogram of ρ values for utterances in the C3 case is shown in Fig. 4.
The SNR for distant speech depends on the environmental conditions and on the distance of the speaker from microphone. It is observed that the ρ values for the distant speech are spread out, compared to the ρ values for different noises. This is mainly due to the effects of reverberation. The distribution of ρ values depends on the distance as well. The ρ value for each utterance is used to determine some parameter values for further processing of $\delta(n)$ and for arriving at the decision logic. In cases where the $\delta(n)$ represent the discriminating characteristics of speech and nonspeech well, the corresponding ρ values are high, as observed for volvo, leopard and machine gun noises. In such cases, small value of the smoothing window parameter l_w is used. The following values of l_w are chosen based on experimentation with speech degraded by different types of noises at different SNR levels:

TABLE III. Values of ρ for speech signal degraded at SNRs of -10 dB and 5 dB for different types of noises. The value for clean speech is 65.28.

NOISE	-10 dB SNR	5 dB SNR
white	14.90	22.61
babble	19.64	36.36
volvo	41.62	56.79
leopard	27.77	43.22
buccaneer1	16.13	28.36
buccaneer2	15.67	22.75
pink	16.73	27.60
hfchannel	16.68	28.46
m109	22.44	35.87
f16	17.84	28.88
factory1	20.48	36.28
factory2	24.13	36.30
machine gun	40.52	64.84

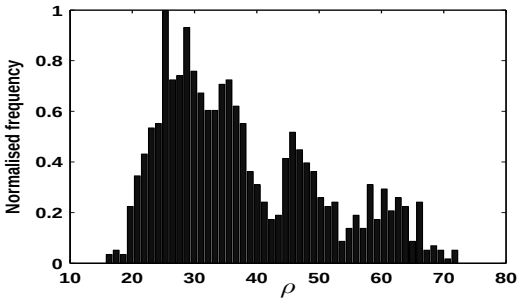

Fig. 4. Histogram of ρ values for distant speech (C3).

$$l_w = 400 \text{ msec, for } \rho < 30. \quad (15)$$

$$l_w = 300 \text{ msec, for } 30 <= \rho <= 40. \quad (16)$$

$$l_w = 200 \text{ msec, for } \rho > 40. \quad (17)$$

3) Decision logic at each sampling instant:
The values of $\delta(n)$ are averaged over a window of size l_w to obtain the averaged value $\bar{\delta}(n)$ at each sample index n. The decision $d(n)$ is made as follows:

$$d(n) = 1, \text{ for } \bar{\delta}(n) > \theta. \quad (18)$$

$$d(n) = 0, \text{ for } \bar{\delta}(n) \leq \theta. \quad (19)$$

4) Smoothing decision at sample level:
The decision $d(n)$ at each sample is processed over windows of size 300 msec, 400 msec and 600 msec, respectively, for the 3 ranges of ρ indicated in (15), (16) and (17). Let η be the threshold on the proportion (in percentage value) of $d(n)$ values which are 1 in the window. If the percentage of $d(n)$ values which are 1 in the window is above the η value, then the final decision $d_f(n)$ is made 1 at the sampling instant n, otherwise it is 0. The value assigned to η is 60%.
5) Decision at frame level:
The decision of the AMR methods is given for every 10 msec frame [28]. In order to compare the proposed

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

method with the AMR method, the decision $d_f(n)$ is converted to a 10 msec frame based decision. For each 10 msec nonoverlapping frame, if majority of $d_f(n)$ values are 1, then the frame is marked as speech, otherwise it is marked as nonspeech. The ground truth of speech signals is also derived for each 10 msec frame.

V. Evaluation of proposed approach.

The proposed method is compared with the state-of-the-art AMR1 and AMR2 methods [28]. AMR1 and AMR2 methods extract subband energies using filter banks. Several acoustical features like pitch, tone, etc., are used to arrive at the decision. Post-processing techniques like hangover are also used [22]. In this paper, we use the version 3GPP TS 26.104 of the AMR methods [28].

We use 5 parameters to evaluate our approach against AMR methods [29] for comparison.

- CORRECT: Correct decisions made by the VAD.
- FEC (front end clipping): Clipping due to speech being misclassified as noise in passing from noise to speech activity.
- MSC (mid speech clipping): Clipping due to speech being misclassified as noise during a speech region.
- OVER (carry over): Noise interpreted as speech in passing from speech activity to noise.
- NDS (noise detected as speech): Noise interpreted as speech within silence/noise region.

All the above parameters are divided by the total number of frames (both speech and nonspeech frames), and then multiplied by 100 to get the percentage value (%). Combining FEC and MSC gives true rejection (TR). Combining OVER and NDS gives false acceptance (FA). The TR indicates the percentage of speech regions not detected as speech, whereas the FA indicates the percentage of nonspeech regions accepted as speech. For good performance, CORRECT should be high, and both TR and FA should be low.

The AMR2 method performs better than AMR1 method in all cases, which is evident from the averaged scores across all noise types for the two different SNRs given in Table IV. Hence we only consider AMR2 scores for comparison.

TABLE IV. Averaged scores across all noise types for two SNR levels for TIMIT database.

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
-10	Proposed	**79.11**	0.06	15.21	0.03	5.49
	AMR2	72.60	0.07	19.31	0.04	7.86
	AMR1	51.70	0.02	2.50	0.10	45.56
5	Proposed	**95.36**	0.02	2.27	0.05	2.20
	AMR2	88.85	0.04	2.31	0.09	8.58
	AMR1	76.05	0.04	1.52	0.09	22.17

Tables V, VI, VII and VIII show the performance of the proposed method in comparison with the AMR2 method for different type of degradations and at different SNR conditions. The best performance in each case is indicated by boldface for CORRECT score.

In the following, the performance of the proposed method is discussed for different types of degradation.

A. Performance on TIMIT database for different types of noises.

Performance of the proposed method under different noise conditions of NOISEX database is given in Table V for two different SNR values, i.e., at -10 dB and 5 dB.

TABLE V. Results for TIMIT database for different types of noises at two SNR levels in comparison with AMR2 method.

NOISE (SNR in dB)	Method	CORRECT	FEC	MSC	OVER	NDS
white (-10)	Proposed	**77.60**	0.08	21.99	0.01	0.23
	AMR2	63.23	0.11	34.32	0.01	2.24
white (5)	Proposed	**97.01**	0.02	1.81	0.05	1.04
	AMR2	87.47	0.08	8.55	0.06	3.74
babble (-10)	Proposed	**67.72**	0.04	12.06	0.05	20.04
	AMR2	61.67	0.05	13.10	0.07	25.01
babble (5)	Proposed	**93.27**	0.03	2.56	0.05	4.01
	AMR2	72.43	0.03	0.52	0.11	26.80
volvo (-10)	Proposed	**98.04**	0.02	0.53	0.08	1.26
	AMR2	95.93	0.02	0.24	0.11	3.59
volvo (5)	Proposed	**96.39**	0.04	2.38	0.06	1.03
	AMR2	94.37	0.00	0.54	0.11	4.89
leopard (-10)	Proposed	**97.09**	0.02	1.18	0.06	1.58
	AMR2	95.92	0.05	0.88	0.10	2.95
leopard (5)	Proposed	**97.82**	0.02	0.78	0.07	1.22
	AMR2	95.61	0.01	0.16	0.11	4.00
buccaneer1 (-10)	Proposed	**69.76**	0.09	25.90	0.01	4.15
	AMR2	65.97	0.11	33.10	0.01	0.71
buccaneer1 (5)	Proposed	**95.59**	0.02	2.23	0.05	2.03
	AMR2	93.92	0.07	3.81	0.08	2.01
buccaneer2 (-10)	Proposed	**76.54**	0.08	21.41	0.01	1.87
	AMR2	64.46	0.11	34.00	0.00	1.33
buccaneer2 (5)	Proposed	**96.89**	0.02	1.81	0.05	1.16
	AMR2	90.78	0.07	6.28	0.06	2.69
pink (-10)	Proposed	**74.23**	0.09	25.43	0.01	0.16
	AMR2	66.79	0.11	32.30	0.00	0.69
pink (5)	Proposed	**97.07**	0.02	1.75	0.05	1.04
	AMR2	94.52	0.07	3.22	0.08	1.99
hfchannel (-10)	Proposed	**75.03**	0.08	23.48	0.02	1.30
	AMR2	71.22	0.09	27.22	0.03	1.33
hfchannel (5)	Proposed	**96.94**	0.02	1.57	0.06	1.35
	AMR2	94.69	0.05	2.57	0.09	2.49
m109 (-10)	Proposed	**89.68**	0.04	6.49	0.04	3.69
	AMR2	82.80	0.08	15.03	0.04	1.94
m109 (5)	Proposed	**97.32**	0.01	0.72	0.07	1.81
	AMR2	95.63	0.04	0.40	0.11	3.70
f16 (-10)	Proposed	**75.94**	0.08	22.37	0.02	1.51
	AMR2	69.88	0.10	29.18	0.01	0.72
f16 (5)	Proposed	**97.10**	0.02	1.43	0.06	1.34
	AMR2	95.64	0.06	2.06	0.09	2.04
factory1 (-10)	Proposed	**67.56**	0.05	13.53	0.05	18.74
	AMR2	58.80	0.06	17.42	0.06	23.57
factory1 (5)	Proposed	**91.72**	0.02	1.85	0.06	6.28
	AMR2	74.12	0.04	1.42	0.10	24.21
factory2 (-10)	Proposed	**82.20**	0.05	9.55	0.04	8.09
	AMR2	82.16	0.07	14.02	0.06	3.59
factory2 (5)	Proposed	**95.09**	0.02	0.91	0.07	3.85
	AMR2	94.45	0.04	0.36	0.11	4.94
machine gun (-10)	Proposed	**77.13**	0.07	13.85	0.03	8.81
	AMR2	64.97	0.01	0.26	0.11	34.56
machine gun (5)	Proposed	**87.55**	0.08	9.78	0.03	2.45
	AMR2	71.43	0.00	0.24	0.11	28.12

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

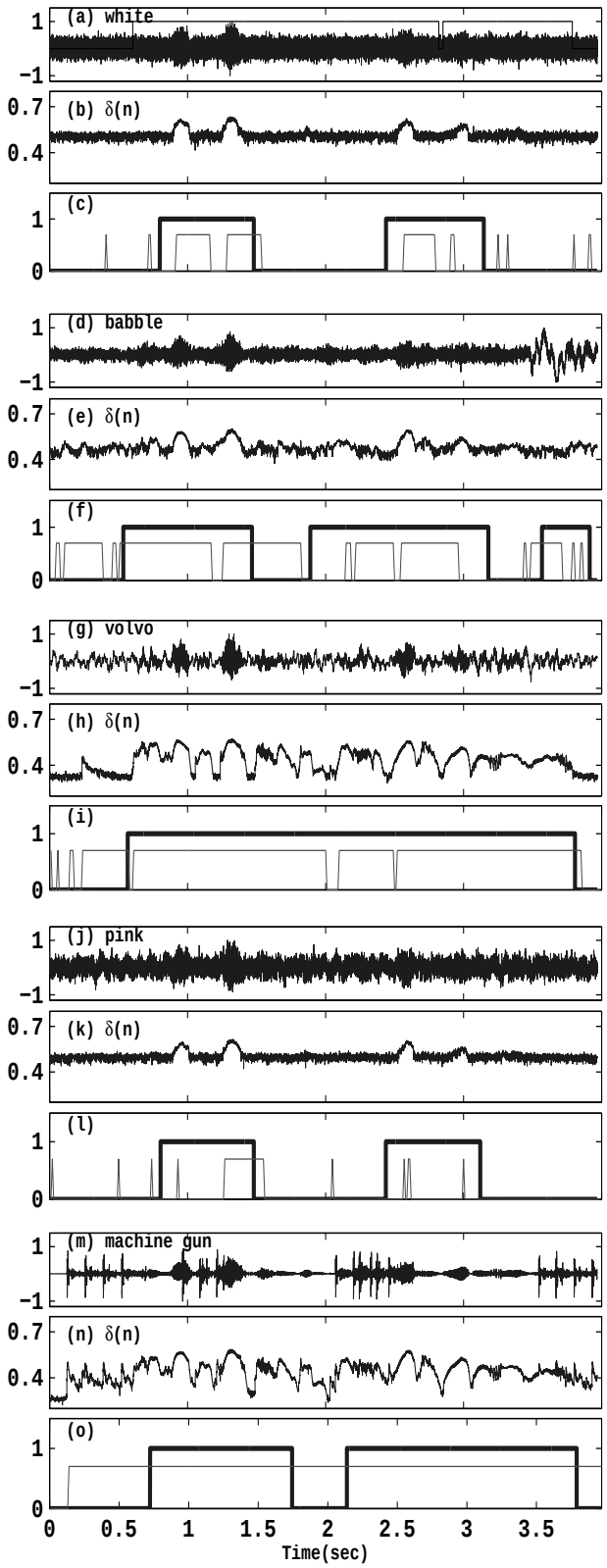

Fig. 5. Illustration of results of VAD for different types of NOISEX data at **-10 dB** SNR. Each noise type has three subfigures: Degraded signal at **-10 dB** SNR, $\delta(n)$, and decision for the proposed method (thick line) and for AMR2 method (thin line). White noise (a, b, c), Babble noise (d, e, f), Volvo noise (g, h, i), Pink noise (j, k, l), Machine gun noise (m, n, o). The ground truth is indicated on top of the degraded speech signal in (a).

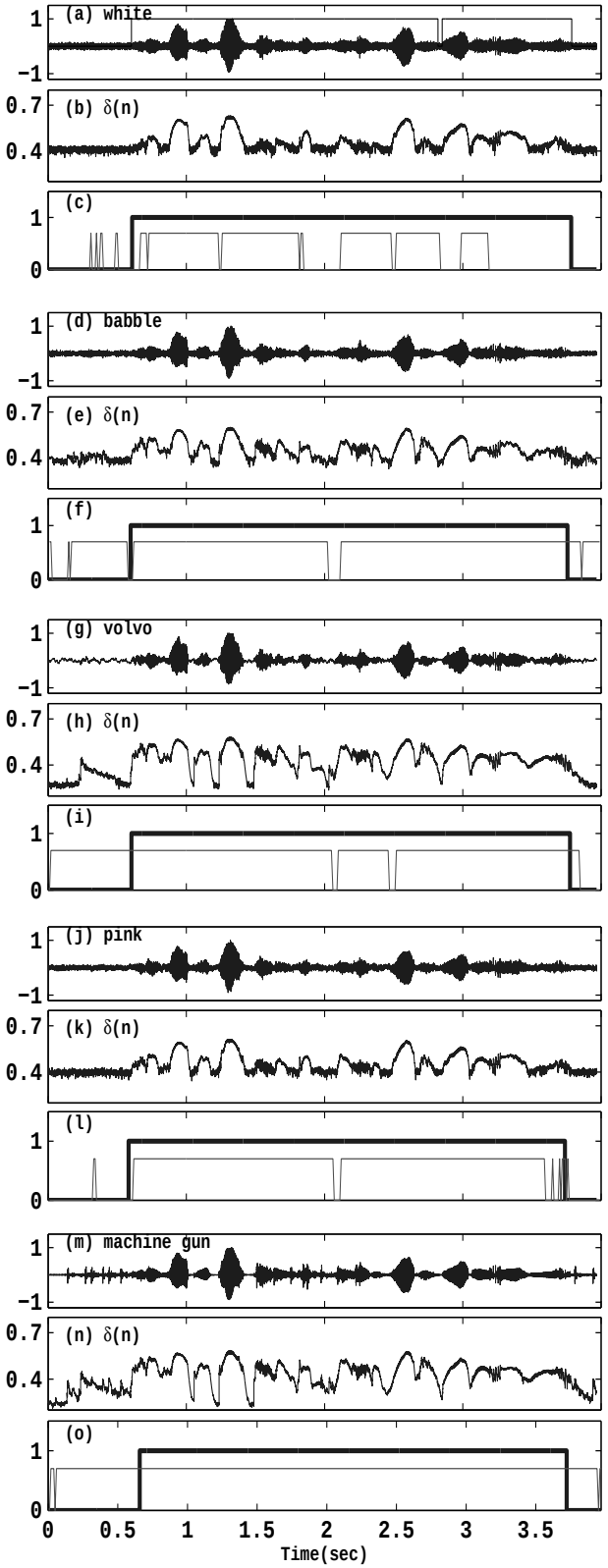

Fig. 6. Illustration of results of VAD for different types of NOISEX data at **5 dB** SNR. Each noise type has three subfigures: Degraded signal at **5 dB** SNR, $\delta(n)$, and decision for the proposed method (thick line) and for AMR2 method (thin line). White noise (a, b, c), Babble noise (d, e, f), Volvo noise (g, h, i), Pink noise (j, k, l), Machine gun noise (m, n, o). The ground truth is indicated on top of the degraded speech signal in (a).

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

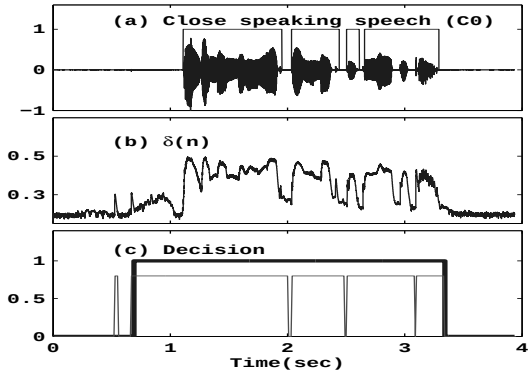

Fig. 7. (a) Close speaking speech (C0) with ground truth indicated on top. (b) $\delta(n)$. (c) Decision of the proposed method at η=90% (thick line) and the AMR2 method (thin line).

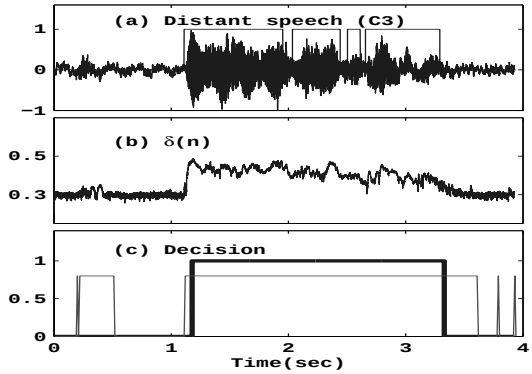

Fig. 8. (a) Distance speech (C3) with ground truth indicated on top. (b) $\delta(n)$. (c) Decision of the proposed method at η=90% (thick line) and the AMR2 method (thin line).

It is observed that performance of the proposed method is higher than that of AMR2 method for all types of noises. For five types of noises, the performance is illustrated for an utterance in the form of plots shown in Figs. 5 and 6 at SNRs of -10 dB and 5 dB, respectively. For each type of noise, the degraded signal, the corresponding $\delta(n)$ values and the derived VAD decision (thick line) are shown. In addition, the AMR2 decision is also shown by thin solid lines for comparison. The ground truth is marked in Figs. 5(a) and 6(a) by a thin line.

As can be seen from Figs. 5(a) and 6(a) for white noise case, many speech regions are missed in the AMR2 method, resulting in high TR. In the case of babble noise at 5 dB SNR, the features in the speech regions stand out over the nonspeech regions (Fig. 6(e)), and hence the FA is lower for the proposed method than for the AMR2 method (Fig. 6(f)).

Since most of the energy is concentrated in the low frequency regions for volvo noise, it is relatively easier to reduce the effect of this type of noise, and hence the proposed method performs better at the two noise levels (Figs. 5(i) and 6(i)).

A significant lower TR is seen in the case of pink noise for the proposed method compared to the AMR2 method. This is due to attenuation of noise regions by weighting (Figs. 5(l)). This can also be seen in the 3D plots given in Fig. 1.

Due to its high temporal variance, most VAD algorithms detect the machine gun chunks as speech. The high temporal resolution of the features in the proposed method gives better performance for the proposed method than for the AMR2 method as indicated in Table V. It is interesting to see in Figs. 5(o) and 6(o) that the nonspeech regions affected by the machine gun noise are identified as nonspeech by the proposed method, whereas the AMR2 method accepts them as speech. The LTSV method proposed in [8] shows poor performance for this noise. The multi-band LTSV method [9] also fails to discriminate transient noise from speech.

B. Performance on NTIMIT and CTIMIT databases.

Performance of the proposed method is similar to the AMR2 method for the NTIMIT data (Table VI), and is higher than for the CTIMIT data (Table VI). This may be due to the cellphone (coding) effects, which degrade speech more than the telephone channel (NTIMIT). The l_w value is 200 msec for most of the utterances in these cases because of high ρ value (see (17)).

TABLE VI. Results for NTIMIT and CTIMIT database in comparison with AMR2 method.

Data	Method	CORRECT	FEC	MSC	OVER	NDS
NTIMIT	Proposed	**94.78**	0.02	1.66	0.18	3.18
	AMR2	93.59	0.12	2.91	0.21	2.91
CTIMIT	Proposed	**91.14**	0.04	4.21	0.15	4.31
	AMR2	87.68	0.15	8.28	0.19	3.46

C. Performance on distant speech.

Distant speech is an amalgam of unknown degradations, and the data for a given environment may be limited. The reverberation present in the distant speech signals has high variance in the time domain, as does the speech. So VAD algorithms often confuse reverberation component for speech. The VAD algorithms which bank on temporal variance ([8]) may not perform well, because the distant speech is highly nonstationary, and even the nonspeech regions may have significant temporal variance.

Fig. 7 illustrates the decision obtained by the proposed method and by the AMR2 method for the case of close speaking speech (C0). The errors in the AMR2 method and the proposed method are mostly due to FA (Fig. 7(c)). Note that the $\delta(n)$ values (Fig. 7(b)) have large fluctuations in the speech region, and also it has low floor values as for any clean speech. It is to be noted, that for distant microphone case the performance of the proposed method gives results similar to the AMR2 method, indicating that the proposed method does not fail. Table VII indicates that by proper choice of the value of the η parameter, there can be slight improvement. But the improvement may not be significant. The interesting aspect is that most of the errors in this case are due to false acceptance (FA). This occurs because the degradation in silence regions is not uniform in the case of distant speech, making it difficult to set proper threshold either in the proposed method or in the AMR2 method. One would notice larger fluctuations in the values of $\delta(n)$ in the nonspeech regions, which would result in higher FA rate. It appears that reverberant effects also may be playing a significant role in producing large fluctuations in the values of $\delta(n)$, as it is difficult to compensate those effects by noise deweighting.

It is also interesting to note that even for the relatively cleaner speech (i.e., C0 case in distant speech), there will be large fluctuations in the $\delta(n)$ values in the silence regions,

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

making it difficult to set the thresholds properly. Hence the performance by both the proposed method and the AMR2 method is poorer for C0 case than for the more degraded case of C1.

Fig. 8 illustrates the decision obtained by the proposed method and by the AMR2 method for the distance speech (case C3) for the same utterance shown in Fig. 7. The error is mostly in FA for the AMR2 method (Fig. 8(c)). Note that the $\delta(n)$ values (Fig. 8(b)) have lower dynamic range in the speech region. Also, it has high floor value, as for most degraded speech.

Performance of distant speech can be improved by increasing the η value, as it reduces FA. Table VII shows the improvement in the performance of the distant speech with increase in the η value for the proposed method in comparison with the AMR2 method. Note that the large values of η can also cause increase in the true rejection (TR), which may result in overall reduction in correct decision.

TABLE VII. Results for distant speech for different values of η in the decision logic in comparison with AMR2 method.

Data	Method	CORRECT	FEC	MSC	OVER	NDS
C0	η=60%	84.54	0.00	0.11	0.26	14.92
	η=70%	86.73	0.02	0.26	0.23	12.61
	η=80%	88.34	0.07	0.67	0.20	10.57
	η=90%	**88.93**	0.10	1.64	0.17	8.99
	AMR2	88.87	0.07	0.39	0.26	10.23
C1	η=60%	89.40	0.01	0.56	0.22	9.65
	η=70%	90.91	0.04	1.20	0.17	7.53
	η=80%	91.03	0.13	2.53	0.14	6.02
	η=90%	89.76	0.14	4.53	0.12	5.27
	AMR2	**91.40**	0.11	0.67	0.25	7.35
C2	η=60%	87.03	0.02	0.87	0.22	11.71
	η=70%	88.50	0.06	1.70	0.18	9.42
	η=80%	**88.63**	0.13	3.14	0.15	7.78
	η=90%	87.91	0.14	5.21	0.13	6.44
	AMR2	87.81	0.11	0.98	0.25	10.64
C3	η=60%	86.89	0.03	1.56	0.21	11.16
	η=70%	**87.89**	0.07	2.77	0.17	8.95
	η=80%	87.58	0.13	4.61	0.14	7.37
	η=90%	86.12	0.14	7.35	0.12	6.08
	AMR2	87.61	0.13	2.99	0.23	8.82

D. Performance on TIMIT database for clean speech.

Performance of the proposed method on clean speech is given in Table VIII. It is interesting to note that smoothing and threshold logic for degraded speech smear the information across time, thus reducing the temporal resolution of the final decision. Hence when the decision logic is applied to clean data, it appears to give poor performance. Due to the high dynamic range in both time and frequency domains, the clean speech signal needs to be treated differently in order to obtain good performance.

In contrast to the C0 case of distant speech, for the clean TIMIT data, the error is more in the true rejection (TR) as in Table VIII. This is because for the clean TIMIT data in the silence region, the $\delta(n)$ values are very low and are fluctuating, making it difficult to set the proper threshold. In this case the TR can be reduced by reducing the threshold value, or equivalently reducing the η value.

The scores given in Tables V and VI are for fixed values of the parameters in the decision logic (section IV-C). The η value has been fixed at 60% for most of the cases. A better performance may be achieved, if the parameters θ, η, l_w are adapted suitably for each type of degradation.

TABLE VIII. Results for TIMIT clean case for different values of η in the decision logic in comparison with AMR2 method.

Method	CORRECT	FEC	MSC	OVER	NDS
η=40%	**95.72**	0.02	2.71	0.06	1.37
η=50%	94.92	0.05	3.96	0.06	0.92
η=60%	93.55	0.07	5.61	0.04	0.62
η=70%	91.66	0.08	7.66	0.04	0.46
η=80%	89.42	0.09	10.01	0.03	0.35
AMR2	93.59	0.12	2.91	0.21	2.91

E. Performance comparison with DFT and gammatone filters.

The proposed method is evaluated using filterbank energy contours using DFT and 128 gammatone filters [21]. After deriving the band energy contours, the subsequent processing, including weighting, the energy contours, computation of $\delta(n)$, thresholding and decision logic, are all same in these cases as in the SFF method described before.

TABLE IX. Averaged scores across different noise types for two SNR levels for TIMIT database.

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
-10	Proposed	81.25	0.05	12.36	0.03	6.20
	DFT	**81.69**	0.02	4.31	0.06	13.83
	Gammatone	81.63	0.06	11.98	0.03	6.20
	AMR2	74.63	0.06	15.19	0.05	9.94
5	Proposed	**94.77**	0.02	2.37	0.05	2.68
	DFT	93.38	0.01	1.12	0.06	5.33
	Gammatone	93.24	0.03	3.78	0.03	2.82
	AMR2	88.75	0.04	1.52	0.09	9.46

The results are given in Table IX in terms of averaged performance over 11 different noise types (except white and pink noises), using 50 utterances of TIMIT data, for two different noise levels (-10 dB and 5 dB). It is interesting to note that all the three methods of preprocessing namely, SFF, DFT and gammatone filters, give similar results. All of them are significantly better than the results using the AMR2 method.

Note that the three methods of preprocessing may perform differently for different noise types. We have observed that for synthetic noises like white and pink noises, the performance by DFT and gammatone filtering is better than by SFF. This is due to some temporal and spectral averaging of noises in the high frequency region ($>$ 2000 Hz) due to temporal averaging in the case of DFT and due to spectral smoothing in the case of gammatone filters. The performance improvement for all the three methods will be similar even for these two types of noises, if in the SFF method some smoothing is done in the time and frequency domains, especially in the higher frequency region, before computing mean and variance across frequency. Note that the performance improvement of these three preprocessing methods over the AMR2 method is due to the subsequent processing of the energy contours in each band, especially the weighting in (12). The effect of weighting can be seen in the performance of the proposed method with and without weighting as given in Table X. The average scores across all noise types for two different SNR values (-10 dB and 5 dB) are given using unweighted and weighted SFF output for 50 utterances of TIMIT data.

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE X. Averaged scores across all noise types for two SNR levels of unweighted and weighted SFF output for TIMIT database.

SNR (dB)	Method	CORRECT	FEC	MSC	OVER	NDS
−10	Unweighted SFF	70.73	0.07	19.45	0.03	9.60
	Weighted SFF	**78.04**	0.07	15.82	0.03	5.94
5	Unweighted SFF	93.18	0.03	3.80	0.05	2.83
	Weighted SFF	**95.15**	0.02	2.25	0.06	2.41

VI. Summary

A new VAD method is proposed based on single frequency filtering (SFF) approach introduced in this paper. The method exploits the fact that speech has high SNR regions at different frequencies and at different times. The variance of speech across frequency is higher than that for noise, after compensating for spectral characteristics for noise. The spectral characteristics of noise are determined using the floor of the temporal envelope at each frequency, computed by the SFF approach.

The $\delta(n)$ feature proposed for VAD decision is robust against degradation, as evidenced by the high CORRECT percentage scores obtained for all types of noises. The proposed method is tested over standard TIMIT, NTIMIT and CTIMIT databases, as well as for distance speech, thus covering varieties of degradations.

While the results show significant improvement in performance of the proposed method, in comparison with the AMR2 method, better results may be obtained, if the decision logic parameters (θ, η, l_w) are made degradation-specific. It was noticed that adapting the parameters θ, η, l_w based on the degradation characteristics estimated from ρ has improved the overall performance. Adapting the threshold with time in each utterance may also improve the performance. Further improvement can be expected if other characteristics of speech, such as voicing, are also included in the decision logic.

The SFF method yields envelopes at any desired frequency, with high temporal and spectral resolution. This property can be exploited for many other applications in speech processing, such as robust pitch extraction, speech enhancement, and deriving robust features for speech and speaker recognition. Our preliminary studies indicate that the SFF method is indeed showing promise in some of these applications.

VII. Acknowledgements

The authors thank the members of ECESS Consortium Siemens AG, Corporate Technology, Germany, for granting permission to use SPEECON database.

References

[1] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and support vector machine," *in Proc. Int. Conf. on Speech and Computer (SPECOM07)*, vol. 2, pp. 556–561, 2007.

[2] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 19, no. 7, pp. 2026–2038, Sep. 2011.

[3] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *Proc. TENCON'93, IEEE*, 1993, pp. 321–324.

[4] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Process. Lett.*, vol. 17, no. 3, pp. 273–276, Mar. 2010.

[5] T. Pham, M. Stark, and E. Rank, "Performance analysis of wavelet subband based voice activity detection in cocktail party environment," in *Proc. Int. Conf. on Computing and Communication Technologies*, Oct. 2010, pp. 85–88.

[6] Z. Song, T. Zhang, D. Zhang, and T. Song, "Voice activity detection using higher-order statistics in the teager energy domain," in *Proc. Wireless Communications Signal Process.*, Nov. 2009, pp. 1–5.

[7] J. Ramrez, J. C. Segura, C. Bentez, A. D. L. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 3–4, 2004.

[8] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 19, no. 3, pp. 600–613, Mar. 2011.

[9] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Proc. Interspeech*, Aug. 2013, pp. 718–722.

[10] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice," in *Proc. Interspeech*, Aug. 2013, pp. 704–708.

[11] Y. W. Jitong Chen and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE Trans. Speech Audio Process.*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[12] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proc. Interspeech*, Sep. 2014, pp. 1534–1538.

[13] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[14] T. Pham, C. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Int. Conf. on Computing and Communication Technologies, RIVF*, July 2009, pp. 1–8.

[15] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 19, no. 8, pp. 2624 –2633, Nov. 2011.

[16] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Acoust., Speech, Language Process.*, vol. 21, no. 4, pp. 697–710, 2013.

[17] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[18] J. Ramirez, J. Segura, C. Benitez, A. De la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1119–1129, 2005.

[19] P. Harding and B. Milner, "On the use of machine learning methods for speech and voicing classification," in *Proc. Interspeech*, Sep. 2012.

[20] F. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection," in *Proc. Interspeech*, Aug 2013, pp. 732–736.

[21] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 433–442, Jan. 2002.

[22] ETSI, Voice activity detector (VAD) for adaptive multirate (AMR) speech traffic channels, ETSI EN 301 708 v.7.1.1, Dec. 1999.

[23] John S. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, USA, 1993.

[24] A. Varga and J. H. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[25] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP*, Apr. 1990, pp. 109–112.

[26] K. Brown and E. George, "CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition," in *Proc. ICASSP*, May 1995, pp. 105–108.

[27] R. Siemund, H. Huge, S. Kunzmann, and K. Marasek, "SPEECON - speech data for consumer devices," in *Proc. LREC*, 2000, pp. 883–886.

[28] http://www.3gpp.org/ftp/Specs/html-info/26104.htm.

[29] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *Proc. ICASSP*, May 1989, pp. 369–372.

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2015.2404035, IEEE/ACM Transactions on Audio, Speech, and Language Processing

G. Aneeja received the B.Tech. in electronics and communications engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2009. She is currently pursuing the Ph.D. degree at the International Institute of Information Technology, Hyderabad, India. Her research interests include signal processing, speech analysis, voice activity detection and speaker verification.

B. Yegnanarayana (M'78SM'84F'13) received the B.Sc. degree from Andhra University, Waltair, India, in 1961, and the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science (IISc) Bangalore, India, in 1964, 1966, and 1974, respectively. He is currently an Institute Professor at the International Institute of Information Technology (IIIT), Hyderabad. He was professor and Microsoft chair at IIIT, Hyderabad from 2006 to 2012. Prior to joining IIIT, Hyderabad, he was a professor at the Indian Institute of Technology, Madras, India (1980 to 2006), a visiting associate professor at Carnegie-Mellon University, Pittsburgh, USA (1977 to 1980), and a member of the faculty at the IISc, Bangalore (1966 to 1978). His research interests are in signal processing, speech, image processing, and neural networks. He has published over 350 papers in these areas in IEEE journals and other international journals, and in the proceedings of national and international conferences. He is also the author of the book Artificial Neural Networks (Prentice-Hall of India, 1999). He has supervised 30 Ph.D. dissertations and 36 M.S. theses. He is a Fellow of the Indian National Academy of Engineering (INAE), a Fellow of the Indian National Science Academy (INSA), a Fellow of the Indian Academy of Sciences, a Fellow of IEEE (USA) and a Fellow of the International Speech Communications Association (ISCA). He was the recipient of the third IETE Prof. S. V. C. Aiya Memorial Award in 1996. He received the Prof. S. N. Mitra memorial Award for the year 2006 from the INAE. He was awarded the 2013 Distinguished Alumnus award from IISc, Bangalore. He was awarded the "Sayed Husain Zaheer Medal (2014)" of INSA in 2014. He was the Associate Editor for the IEEE Transactions on Audio, Speech and Language Processing during 2003 - 2006. Currently he is an Associate Editor for Springer's international journal Circuits, Systems and Signal Processing.

2329-9290 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.