EVALUATION OF FEATURES FOR VOICE ACTIVITY DETECTION USING DEEP NEURAL NETWORK

[1,2]SUCI DWIJAYANTI, [1]MASATO MIYOSHI

[1]Graduate School of Natural Science and Technology, Kanazawa University, Japan

[2]Department of Electrical Engineering, Universitas Sriwijaya, Indonesia

E-mail: [1]dwijas@stu.kanazawa-u.ac.jp

ABSTRACT

Voice activity detection (VAD) is implemented in the preprocessing stage of various speech applications to identify speech and non-speech periods. Recently, deep neural networks (DNNs) have been utilized for VAD given their superior performance over other methods. When used to identify speech and non-speech periods, DNNs depend on the input of different features to discriminate speech from noise. Hence, different features have been used as input for DNN-based VAD. However, the contribution and effectiveness of such features have not been thoroughly evaluated. In this paper, we address these aspects by comparing five features, namely, log power spectra, filter bank, mel-frequency cepstral coefficients, relative spectral perceptual linear predictive analysis, and amplitude modulation spectrogram, which are widely used on speech processing, to evaluate their performance in a DNN-based VAD. Experiments on the TIMIT speech corpus show that the amplitude modulation spectrogram is the feature with the best performance given its high accuracy even when processing speech data with low signal-to-noise ratio. The next feature showing high performance is log power spectra, which can be considered as a raw feature because it does not require as many calculations or processing as the other features. This suggests that raw features may be suitable inputs for DNN-based VAD. Moreover, limiting the number and processing of features for DNNs may foster system performance, real-time application, and portability of VAD by reducing the computational cost, required memory and storage.

Keywords: *DNN, Speech Period, Speech Features, Voice Activity Detection, Amplitude Modulation Spectrogram, Log Power Spectra*

1. INTRODUCTION

Voice activity detection (VAD) is an important stage in speech applications as it discriminates the presence of speech periods from background noise in an audio signal. VAD has been used in different applications for a variety of purposes. For instance, in speech coding, VAD allows to deactivate transmission in the absence of speech, and hence reduce the amount of transmitted data while maintaining quality [1]. In speech enhancement, VAD is employed to estimate noise during non-speech periods and subsequently remove noise during speech [2]. In speech recognition, VAD is utilized to identify speech in audio signals, as it should be fed to a recognition engine, thus avoiding the processing of non-speech periods that do not convey information and may even undermine the recognition process [3].

Essentially, VAD is a binary classification problem, with one indicating speech and zero non-speech. The successful classification depends on extracting speech features from audio signals to separate speech from noise [4]. Hence, several features have been proposed for VAD. Some of the previous studies considered features in the time domain, such as energy and zero-crossing rates. In particular, Rabiner *et al.* [5] proposed to trace the endpoints of an utterance based on these features. This type of technique is suitable for clean signals, but its performance degrades under a low signal-to-noise ratio (SNR). To increase robustness against noise, some methods have been proposed to modify energy-related features. Prasad *et al.* [6] used a detector based on adaptive linear energy to update a threshold that adjusts the detector operation according to different acoustic environments, and Sakhnov *et al.* [7] used the root-mean-square energy of speech signals. Moreover, VAD methods that rely on energy features have been adopted in certain standards, such as the ITU-T

Recommendation G.729, given its low computational complexity. Other widely used features have been developed in the frequency domain, including some based on the entropy estimation of time–frequency quantities [8]. Ramirez *et al.* [9] proposed a method using long-term spectral divergence, which provides a metric between speech and noise; Ma and Nishihara [10] used the long-term spectral flatness measure. In addition, mel-frequency cepstral coefficients (MFCCs) also serve as features for VAD, and they have been used with classifiers such as support vector machine [11] and Gaussian mixture model [12]. Other approaches include the use of long-term temporal information [13], an acoustic feature that represents the power ratio between periodic and aperiodic components in a signal [14], and spectral autocorrelation under co-channel condition [15]. Pek *et al.* [16] investigated effective modulation frequency ranges and used the modulation spectrum to detect speech and non-speech periods for VAD. Statistical approaches have also been used for VAD. For instance, Sohn *et al.* [17] modeled speech and noise probability density functions, and VAD was based on a likelihood ratio test. Davis *et al.* [18] employed a low-variance spectrum and determined an optimal detection threshold based on noise estimation.

Feature classification of speech signals is the final stage of VAD. The simplest way to classify a signal is by using a detection threshold, which defines the features as speech or non-speech. However, using a threshold is not effective for classification when the feature space is not linearly separable. To overcome this drawback, machine learning approach may be useful, such as support vector machines [11, 19, 20] and neural networks [21, 22]. Recently, deep neural networks (DNNs) have been successfully applied in speech processing, including VAD, given their capabilities. In fact, Mohamed *et al.* [23] describe DNNs as a flexible model that does not require information on the specific data distribution. In addition, a DNN can include several nonlinear hidden layers, thus increasing its flexibility and discrimination capabilities. Furthermore, a generative pre-training allows a strong, domain-dependent regularization of the network weights.

A variety of DNN applications have been proposed in speech processing to detect speech and non-speech periods. For instance, Zhang *et al.* [24] utilized a DNN to explore the advantages of features such as pitch, discrete Fourier transform, MFCCs, linear prediction coefficient, relative spectral perceptual linear predictive (RASTA-PLP) analysis, and amplitude modulation spectrogram (AMS) with its delta features as DNN input for VAD. Likewise, Ryant *et al.* [25] used MFCCs as inputs of a DNN for speech activity on YouTube, and Espi *et al.* [26] used spectro–temporal features as inputs of a DNN to detect non-speech acoustic signals (e.g., the sound of a moving chair). Although using a single feature or a combination of features has been considered in such DNNs, their selection and combination is not a trivial problem, and these aspects should be thoroughly considered to obtain a high classification performance, which is highly dependent on the features used as input. Therefore, in this paper we evaluate the effectiveness and contribution of different speech features to improve the performance of a VAD DNN. Moreover, we aim to reduce computational complexity by discarding features with low contribution and using those that require less processing. To achieve this goal, we investigate five speech features, namely, log power spectra, mel filter bank, MFCCs, RASTA-PLP, and AMS, which have shown a high performance in different speech processing applications, such as speech recognition [27]. After the evaluation, we aim to obtain the best features regarding classification accuracy and provide guidelines on the most effective implementation of VAD using DNNs. Likewise, by considering feature characteristics, we intend to evaluate the performance of that with less computational cost (i.e., log power spectra) and compare it with its counterparts that require more processing and calculations. This way, we can verify whether a raw feature can outperform "hand crafted" feature, which would lead to an increased efficiency for VAD.

This paper is organized as follows. In Section 2, we summarize the DNN-based VAD used for our study. Section 3 presents descriptions of the speech features that we used as DNN inputs and the evaluation method to compare the features. The results and discussion are detailed in Section 4. Finally, we draw our conclusions in Section 5.

2. DNN-BASED VAD

A DNN can be used for accurate classification of speech and non-speech periods in VAD. We considered the DNN proposed in [28] as the basis for our study. The activation vector for the first L layers of the DNN is expressed by

$$\mathbf{v}^{\ell} = f(\mathbf{z}^{\ell}) = f(\mathbf{W}^{\ell}\mathbf{v}^{\ell-1} + \mathbf{b}^{\ell}),\ 0 < \ell < L, \quad (1)$$

where $\mathbf{z}^{\ell} = \mathbf{W}^{\ell}\mathbf{v}^{\ell-1} + \mathbf{b}^{\ell} \in \mathbb{R}^{N_{\ell}\times 1}, \mathbf{v}^{\ell} \in \mathbb{R}^{N_{\ell}\times 1}$, $\mathbf{W}^{\ell} \in \mathbb{R}^{N_{\ell}\times N_{\ell-1}}, \mathbf{b}^{\ell} \in \mathbb{R}^{N_{\ell}\times 1}$, and $N_{\ell} \in \mathbb{R}$ are the excitation vector, activation vector, weight matrix, bias vector, and number of neurons at layer ℓ, respectively, and $f(\cdot): \mathbb{R}^{N_{\ell}\times 1} \to \mathbb{R}^{N_{\ell}\times 1}$ is the elementwise activation function. The input layer is denoted by $\ell = 0$, and thus $\mathbf{v}^{0} = \mathbf{o} \in \mathbb{R}^{N_0\times 1}$ corresponds to the speech features of the DNN, and $N_0 = D$ is the number of features. The logistic sigmoid function,

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (2)$$

is used as activation function of the DNN.

VAD is a binary classification problem (i.e., it contains two classes) to identify speech and non-speech. Observation vector $\mathbf{o}$ belongs to the class whose output unit $y_k, k = 1, 2,$ has a value of 1. The output unit is defined by the following decision function:

$$y_k = \begin{cases} 1, \text{if } s_k > s_i, \forall i = 1, 2, i \neq k \\ 0, \text{ otherwise} \end{cases}, \quad (3)$$

where s_k is the softmax output that represents the probability, $P_{\text{dnn}}(k|\mathbf{o})$, that observation vector $\mathbf{o}$ belongs to class k:

$$s_k = P_{\text{dnn}}(k|\mathbf{o}) = \frac{e^{z_k^L}}{\sum_{i=1}^{I} e^{z_i^L}}, \quad (4)$$

where z_k^L is the k-th element of excitation vector $\mathbf{z}^L$. Therefore, the prediction function of the VAD DNN is given by

$$y_{out} \triangleq s_2 - s_1 \underset{H_d \in H_0}{\overset{H_d \in H_1}{\lessgtr}} \eta, \quad (5)$$

where H_1 and H_0 denote the speech and noise hypotheses, respectively, and η is an adjustable decision threshold.

The DNN training process consists of two stages. First, pre-training is performed using greedy layer-wise unsupervised learning. Second, fine tuning is performed to the whole network [29]. In this study, we considered a DNN composed of five layers of restricted Boltzmann machines (RBMs) that constitute the visible and hidden layers. Specifically, Bernoulli (visible)–Bernoulli (hidden) RBMs were used. After completing the learning process of each RBM, the activity values of its hidden units were used as inputs for the learning process of the subsequent RBM [30]. For pre-training, we used the contrastive divergence algorithm to approximate the gradient of the negative log-likelihood of the data with respect to the RBM parameters [31]. Finally, we used backpropagation techniques through the whole DNN to fine-tune the weights and thus obtain optimal results [32]. In this study, object oriented MATLAB toolbox, namely DeebNet toolbox [32], is used to train the DNN. Detailed parameters used inside the network are described in Section 3.6.

3. VAD FEATURES AND EVALUATION METHOD

3.1 Log Power Spectra

A speech signal can be analyzed using the short-time Fourier transform, which is defined as

$$X(m, k) = \sum_{n=-\infty}^{n=\infty} h(m - n)x(n)W_K^{kn}, \quad (6)$$

where $x(n)$ is a discrete speech signal, $h(n)$ is an analysis window, which is time-reversed and shifted by m frames, k is a frequency variable, K is the number of frequency bins, and $W_K = e^{-j\left(\frac{2\pi}{K}\right)}$. We considered an analysis window of 20 ms with a 10 ms window shift.

The transform in Equation (6) represents a spectrogram which is a graphical display of speech power spectrum over time. Hence, log power spectra provide information on the frequencies of a speech signal, which is updated over appropriate time frames. Consequently, this feature can be used for real-time VAD.

3.2 Mel Filter Bank

Equation (6) shows that the short-time Fourier transform reflects the amount of energy at different frequencies. However, human hearing is not equally sensitive to all frequency bands. In fact, it becomes less sensitive at frequencies above 1000 Hz. Moreover, human sensitivity and the perceived loudness of audio signals can be considered as approximately logarithmic functions [33]. This behavior can be suitably represented by the mel scale, which is approximately linear below 1000 Hz and logarithmic onwards. The mel scale can be obtained as follows:

$$Mel(f) = 2595\ \log_{10}\left(1 + \frac{f}{700}\right), \qquad (7)$$

where f and $Mel(f)$ denote the signal frequency and the perceived level for that frequency, respectively.

A filter bank can be implemented in the frequency domain, where their center frequencies are usually evenly spaced. However, to mimic human perception, we implemented the warped frequency distribution according to the nonlinear function of Equation (6). In addition, triangular filters are commonly used for speech processing. Hence, we computed the mel spectrum of $X(k)$ by multiplying the spectrum magnitude by the corresponding triangular mel weighting filters:

$$s(m) = \sum_{k=0}^{N-1}\left[|X(k)|^2 H_m(k)\right], 0 \le m \le M-1, \quad (8)$$

where M is total number of triangular mel weighting filters and $H_m(k)$ is the weight of the k-th power spectrum bin contributing to the m-th output band [34]. $H_m(k)$ is expressed as

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{2(k - f(m-1))}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1) - k)}{f(m+1) - f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (9)$$

with m ranging from 0 to $M-1$. The output of the mel filter bank is presented in a logarithmic scale to compress the dynamic range and reduce the sensitivity of frequency estimates to slight input variations.

3.3 MFFCs

MFCCs have been the dominant feature used in speech processing given their ability to compactly represent the speech amplitude spectrum. MFCCs correspond to the determination of cepstrum. The cepstrum is the spectrum of the log of the spectrum. Given the vocal tract softness, the components of mel-spectral vectors are highly correlated among frames. Hence, to decorrelate components and reduce the number of parameters, we applied a transform to the mel-spectral vectors using the discrete cosine transform. Thus, we calculated the MFCCs as the inverse DFT using the following equation:

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad (10)$$

for $n = 0,1,2, \dots C-1$, where $c(n)$ are the cepstral coefficients and C is the number of MFCCs. Since log power spectrum is real and symmetric, the inverse DFT is equivalent to a discrete cosine transform. Figure 1 shows the block diagram to obtain MFCCs.

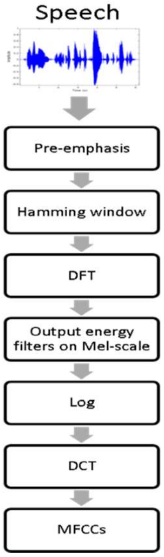


Figure 1: Block Diagram of MFCCs

Speech is not constant frame-to-frame, dynamics features that show how the cepstral coefficients change over time are added. These features are delta and double delta features. Delta features represent the change between frames in the corresponding cepstral and double delta features represent the change between frames in the corresponding delta features. Delta coefficients are calculated as follows

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|}, \quad (11)$$

where $c_m(n)$ is the m-th feature for the m-th time frame, k_i is the i-th weight and T, which is generally taken as 2, is the number of successive frames used for computation. The double delta coefficients are computing by taking the first order derivative of the delta coefficients.

3.4 RASTA-PLP

Perceptual linear prediction, which relies on the psychophysics of hearing, is a technique that warps speech spectra to minimize the differences among speakers while preserving important speech information [35].

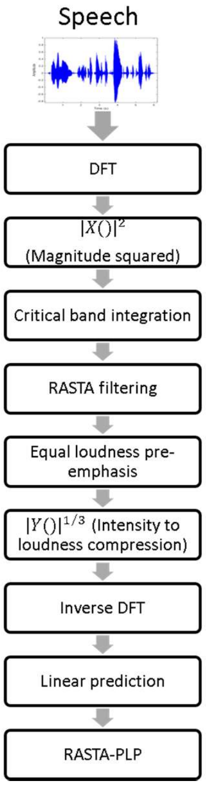

Speech
DFT
$|X()|^2$
(Magnitude squared)
Critical band integration
RASTA filtering
Equal loudness pre-emphasis
$|Y()|^{1/3}$ (Intensity to loudness compression)
Inverse DFT
Linear prediction
RASTA-PLP

Figure 2: Block Diagram of RASTA-PLPs

Meanwhile, the relative spectra approach [36] is based on a bandpass time applied to a log spectral representation of the speech, such as the log filter bank energy. Thus, RASTA-PLP analysis combines these two methods.

This analysis is based on a sequence of calculations for each analysis frame. First, the critical-band spectrum is computed. Then, the temporal derivative of the log of this spectrum is estimated by using a linear regression considering five consecutive spectral values. A nonlinear processing can also be performed in this domain. Next, the derivative is reintegrated by using a first-order infinite impulse response (IIR) filter. The pole position of this filter can be adjusted to define the effective window size. Following perceptual linear prediction, equal loudness is added and multiplied by 0.33 to represent the hearing power law. Subsequently, a relative auditory spectrum is obtained by taking the inverse logarithm of the relative log power spectrum. Finally, an all-pole model is computed. As mentioned in [37], the derivative–reintegration process is equivalent to a bandpass filtering of each frequency channel through the following IIR filter:

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})}. \qquad (12)$$

The detailed process to obtained RASTA-PLP is shown in Figure 2.

3.5 AMS

AMS is inspired on neurophysiology and psychoacoustics, and an AMS-based algorithm for speech was proposed in [38]. In this algorithm, a noisy speech signal is bandpass filtered into 25 channels according to a mel-frequency spacing. Next, the envelopes at each band are computed using full-wave rectification and then decimated by a factor of 3. where the authors use 128 samples acquired every 32 ms with a 50% overlap. Each segment is Hann windowed and transformed into a 256-point fast Fourier transform including zero padding. The modulation spectrum is calculated for each channel by using the transform, with a frequency resolution of 15.6 Hz. Within each band, the spectrum magnitudes are multiplied by 15 triangular-shaped windows evenly spaced in the 15.6–400 Hz range and summed up to produce 15 amplitude values. These values represent the AMS feature vector. The AMS is a two-dimensional representation of the spectral and temporal properties of an acoustic signal [39]. Processing stage to obtain AMS is shown in Figure 3.

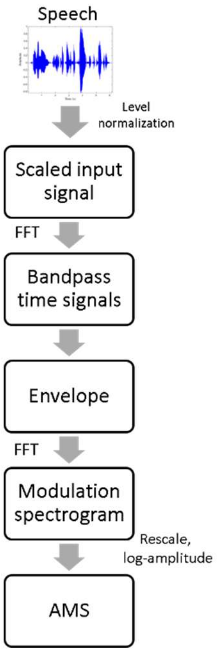


Figure 3: Block Diagram of AMS

3.6 Evaluation Method

To evaluate and compare the above mentioned features on the VAD DNN, we used the data of 150 speakers from the TIMIT speech corpus [40], which contains phonemically and lexically transcribed speech of American English speakers with different dialects. From the speaker records, 100 were used for training and the remaining 50 for evaluation. To increase distortion, we injected five types of noises in the speech signals, namely, white, babble, factory, car, and pink noises from the NOISEX-92 database [41]. Each noise type and SNR values of 10, 5, 0, and –5 dB were randomly selected for injection in the speech signals.

The above mentioned features were extracted from clean and noisy speech signals. The input signals were sampled at 8 kHz, and we selected a frame size of 20 ms and a Hamming analysis window with a 10 ms frame shift. In addition, we considered a 40-channel mel-scale filter bank, 13 MFCCs, 12th-order RASTA-PLP, and 15 modulation spectra for the AMS. All the features were normalized to zero mean and unit variance at each dimension.

Then, we trained the DNN using five RBMs, which were stacked to obtain the hidden layers. The consecutive hidden layers contained 512, 512, 256, 256, and 128 neurons. We also considered a learning rate of 0.001, fixed the maximum epochs of both pre-training and fine tuning to 100, and used equal parameters and settings for the evaluation of every feature as DNN input.

To quantify the performance of each feature as DNN input for VAD, we generated receiving operating characteristic (ROC) curves, which correspond to true positive rate *TPR* against false positive rate *FPR* given by

$$TPR = \frac{TP}{TP + FN}, \tag{13}$$

$$FPR = \frac{FP}{FP + TN}, \tag{14}$$

where *TP* represents the number of true positives, i.e., the number of correctly identified speech frames, *TN* represents the true negatives, i.e., the number of correctly identified non-speech frames, *FP* represents the false positives, i.e., the number of non-speech frames incorrectly identified as speech, and *FN* represents the false negatives, i.e., the number of speech frames incorrectly identified as non-speech. In addition, to obtain a performance measure, we calculated the area under the curve (AUC) of the ROC curves, where higher AUC values indicate more accurate classification.

Finally, to visually assess the separability of features, we obtained their distribution. When feature values that belong to the same class have a strong relation, they can be classified more easily through a DNN. Hence, separability can be used as a measure to analyze feature performance. To obtain the feature distributions, we used the t-distributed stochastic neighbor embedding (t-SNE) [42], which represents data in a plane and clusters similar data points. This algorithm is useful to visualize high dimensional data and assess their degree of separability. t-SNE is a variation of SNE which converts pairwise Euclidean distances in N-dimensional to joint probability distribution. Given a set of N -dimensional data points $X = \{x_1, x_2, \cdots, x_n\}$, the joint probability distributions P are computed as

$$p_{ij} = \frac{e^{-d_{ij}^2/\sigma}}{\sum_k \sum_{l \neq k} e^{-d_{kl}^2/\sigma}}, \tag{15}$$

where $d_{ij} = \|x_i - x_j\|^2$ is the *N*-dimensional norm, σ is the variance of the Gaussian distribution centered on data point x_i and $p_{ii} = 0$. Input data is high dimensional data. The similarity of data point $\boldsymbol{x_j}$ to data point $\boldsymbol{x_i}$ is the conditional probability, that $\boldsymbol{x_i}$ would pick $\boldsymbol{x_j}$ as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at $\boldsymbol{x_i}$. The bandwidth of Gaussian kernels is set in such a way that the perplexity of the conditional distribution equals as predefined perplexity using the bisection method. As a result, the bandwidth is adapted to the density of the data, smaller values of σ are used in denser parts of the data space. t-SNE aims to learn a d -dimensional map that reflects the similarities p_{ij} as well as possible. Next, it measures similarities q_{ij} between two points in the map y_i and y_j in low dimensional map using similar approach. The low dimensional mapping obtained by t-SNE, $Y = \{y_1, y_2, \cdots, y_n\}$ uses a student-t distribution with a single degree of freedom to model the similarity between two data points

$$q_{ij} = \frac{(1 + d_{ij})^{-1}}{\sum_k \sum_{l \neq k} (1 + d_{kl})^{-1}}, \tag{16}$$

where $d_{ij} = \|y_i - y_j\|^2$ is the low dimensional norm, and $q_{ii} = 0$. The obtained mapping minimizes the Kullback-Leibler divergence with respect to the high dimensional distribution, using a gradient descent method.

4. RESULTS AND DISCUSSION

The VAD on the noisy signals for the different features is illustrated in Figure 4, where the red dashed lines represent the speech (high level) and non-speech (low level) periods, and the blue solid lines represent the VAD DNN output for the corresponding feature. In addition, the top graph in Figure 4 shows the VAD from a typical clean signal and the noisy signal polluted by car noise at SNR of –5 dB. For the noisy signal, it can be seen that the DNN-based VAD output using the log power spectra provides the best results, which are close to the ground truth, followed by AMS and filter bank. Hence, this feature might be suitable for distinguishing speech and non-speech periods, and contain more discriminative information for VAD than the other evaluated features.

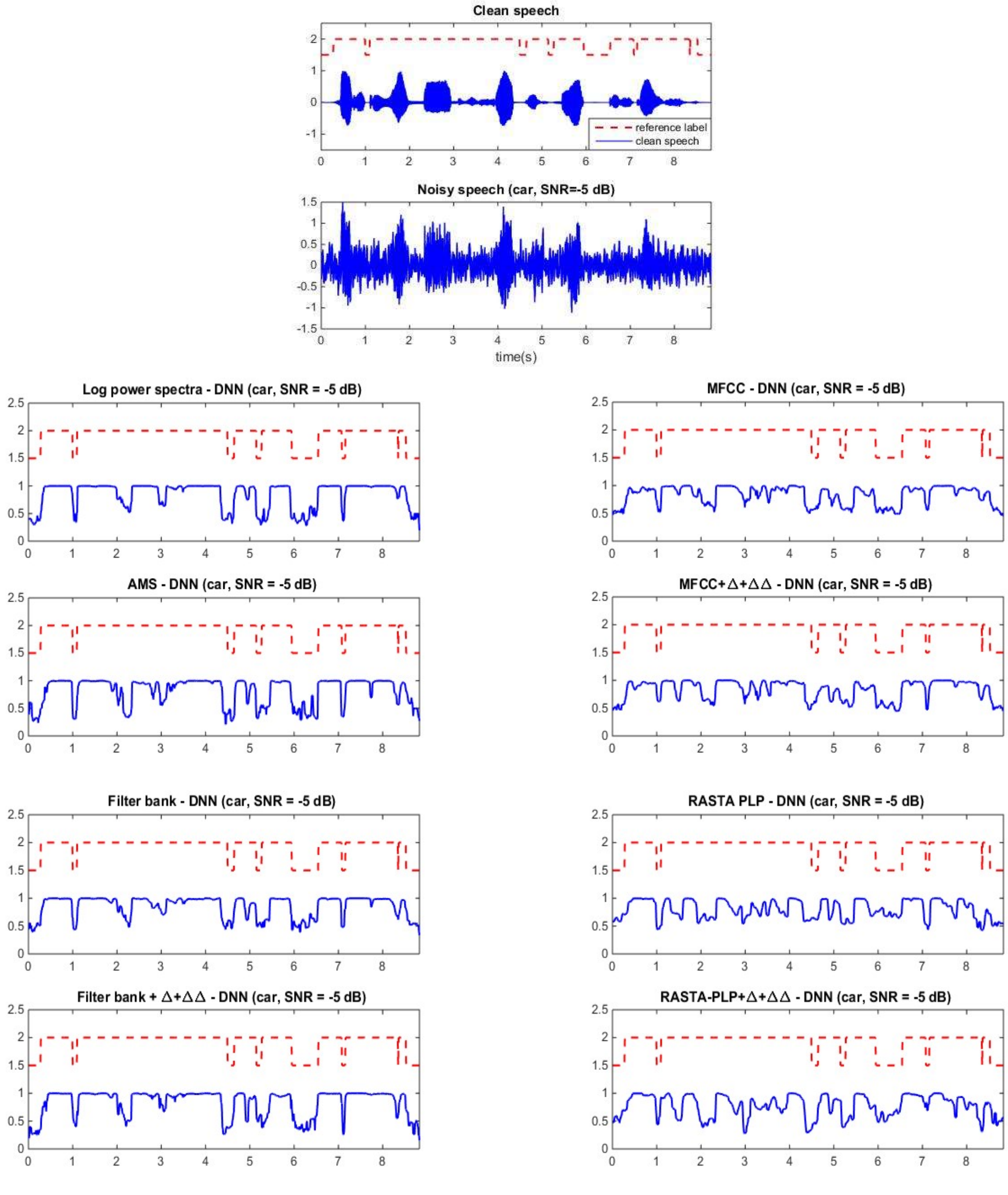
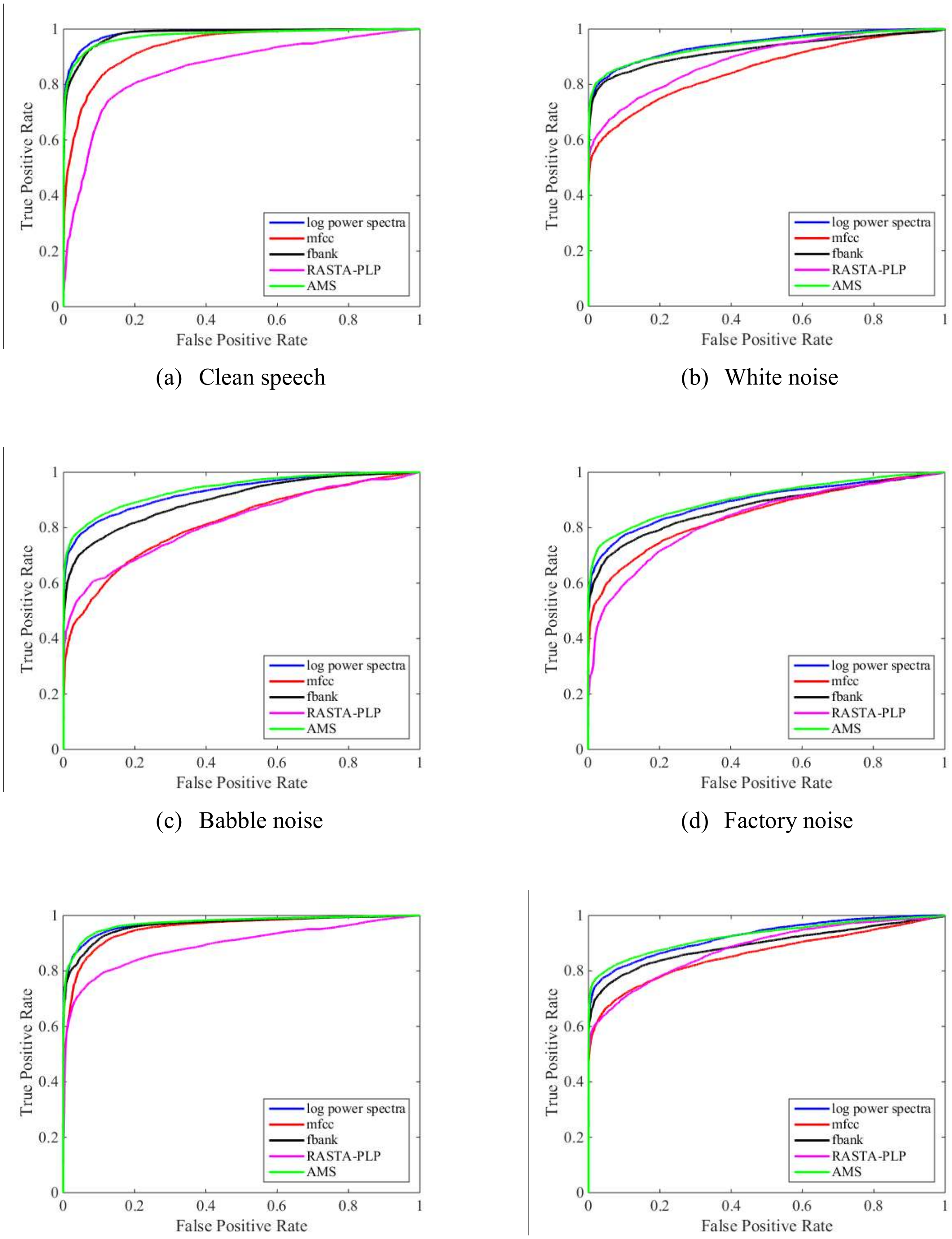


Figure 4: Typical Speech Signal And VAD DNN Output For Different Features



(a) Clean speech

(b) White noise

(c) Babble noise

(d) Factory noise

(e) Car noise

(f) Pink noise

Figure 5: ROC Curves Of The VAD DNN Considering Each Feature When Using The Clean Signal And A 10-dB SNR With Different Types Of Noises

Table 1: AUC For Feature ROC Curves. The Numbers In Bold Indicate The Best Performance

Noise	SNR (dB)	AUC (%)							
		Filter bank	MFCC	Log Power Spectra	RASTA-PLP	AMS	MFCCs+△+△△	RASTA-PLP +△+△△	Filter bank +△+△△
No noise		98.16	93.99	**98.57**	85.84	97.71	95.02	86.65	98.27
White	10	92.41	85.40	**94.54**	89.00	94.24	86.44	89.24	93.03
	5	89.61	79.28	91.50	86.31	**91.84**	80.56	85.88	90.15
	0	85.18	72.68	87.14	82.48	**87.38**	74.68	81.15	85.97
	–5	79.52	67.06	82.33	76.95	**82.34**	70.32	74.67	80.87
Babble	10	90.10	81.67	92.96	81.93	**93.98**	84.66	85.33	91.84
	5	84.53	74.49	88.65	77.36	**90.56**	78.55	80.50	86.59
	0	76.20	66.48	81.92	71.18	**84.21**	71.29	74.07	79.26
	–5	66.24	59.14	72.97	63.88	**75.64**	64.34	66.80	71.52
Factory	10	87.36	84.63	89.40	83.14	**90.61**	86.25	85.12	90.44
	5	81.18	77.76	84.10	78.31	**86.95**	79.86	80.04	84.96
	0	73.75	71.23	77.55	71.74	**80.75**	73.65	73.41	77.88
	–5	64.76	64.98	70.16	64.21	**73.40**	68.04	66.31	70.79
Car	10	96.61	95.52	97.16	89.38	**97.39**	95.83	92.45	97.86
	5	96.66	95.88	96.93	89.29	**97.02**	96.01	92.18	97.44
	0	96.25	95.64	96.42	88.81	96.36	95.76	91.53	**96.78**
	–5	94.81	94.73	95.26	87.59	94.53	94.75	90.04	**95.49**
Pink	10	89.36	85.97	92.49	88.01	**92.65**	87.49	89.73	91.88
	5	84.89	79.37	88.21	84.86	**89.18**	81.14	86.20	87.74
	0	78.49	71.32	**82.27**	79.54	80.32	73.96	80.60	82.08
	-5	69.76	63.55	74.46	71.78	**76.43**	67.18	72.76	75.65

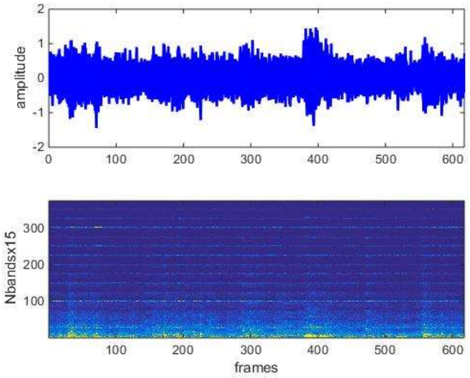
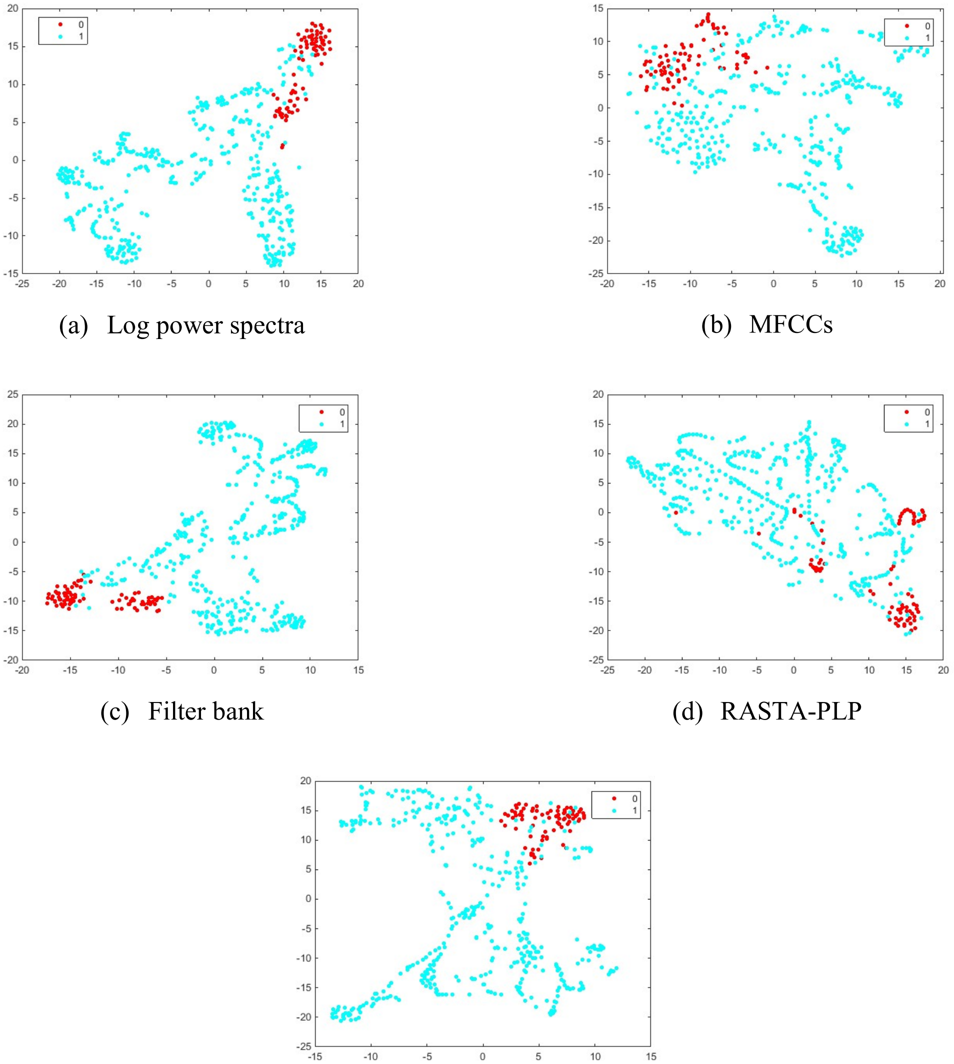


Figure 6: AMS Of Noisy Signal (Factory Noise, SNR of –5 dB).

To quantify the effectiveness of each feature, we generated the ROC curves from the clean and different types of noisy signals, as shown in Figure 5. For the clean signal (Figure 5a), the VAD DNN results in a high and similar performance for the log power spectra, AMS, and filter bank, as these features produce higher true positive rate *TPR* and lower false positive rate *FPR* than the MFCCs and RASTA-PLP. For noisy signals (Figures 5b–5f), the AMS and log power spectra maintain a high performance. In contrast, the MFCCs and RASTA-PLP show a high sensitivity to noise, especially non-stationary noise such as the babble noise (Figure 5c). Overall, the AMS and log power spectra show the best performance among all the features in noisy conditions.

Then, to measure the classification accuracy of each feature, we calculated the AUC for each ROC curve. Table 1 lists the AUC values for the different features used in the VAD DNN. In the clean speech signals, the best performance was obtained using the log power spectra, followed in descending order by the mel filter bank, AMS, MFCCs, and RASTA-PLP. In the noisy speech signals, the AMS showed the best performance for most cases. Moreover, the performance of log power spectra approaches that of AMS, especially with the signal polluted by stationary noise, such as white noise. Interestingly, log power spectra can be considered as a raw feature because, unlike the others, it can be directly obtained from the observed signal without intensive processing or calculations. Among the other features, MFCCs provided the lowest performance, but it improved when combined to dynamic features, i.e., delta and delta–delta cepstral features. In addition, RASTA-PLP outperforms MFCCs when considering noisy signals. These results suggest that a raw feature may be more useful as input for DNN-based VAD compared to counterparts that require more processing and calculations.

Overall, Table 1 shows the superior performance of AMS as feature for VAD, especially when considering noisy signals. This can be due to the fact that AMS contains information of both center and modulation frequencies within each analysis frame, as illustrated in Figure 3, where the feature distribution might be helpful to discriminate speech from noise, and hence provide a high performance as input for DNN-based VAD. Furthermore, AMS encodes modulation spectra that are computed for each channel, and the harmonicity of speech is clearly exhibited as peaks on such encoding. Moreover, AMS is based on perception and physiology, and hence it provides an accurate speech representation with a wide resolution. In addition, regarding the DNN, the AMS input information is not influenced by its dimension, because the hidden layers can unveil higher-order data. Even in the worst-case scenario, when considering the –5 dB signal with babble noise, Table 1 shows that the highest performance of 75.64% is achieved when using AMS.



(a) Log power spectra

(b) MFCCs

(c) Filter bank

(d) RASTA-PLP

(e) AMS

Figure 7: t-SNE Planar Map Of Feature Vectors

Another high-performance feature is the log power spectra, as it shows better results than the mel filter bank and MFCCs. Furthermore, the log power spectra feature has a similar performance to AMS for stationary noise and even outperforms AMS when considering white noise at 10 dB and pink noise at 0 dB. This high performance may be due to the detailed time–frequency information provided by log power spectra. On the contrary, the mel filter bank and MFCCs contain information of specific sub-bands caused by the dimension reduction and discrete cosine transform, respectively, which may cause a notable information loss. The performance of log power spectra reflects that raw data can be superior to highly processed features for the VAD DNN. In fact, the log power spectra feature is directly derived from the speech signal and does not require complicated calculations. Hence, this feature may lead to lower computational time and complexity than the other features.

On the other hand, the performance of RASTA-PLP is similar to that of MFCCs. In addition, MFCCs show a high performance only for clean signals, whereas this feature is outperformed by RASTA-PLP when using noisy signals, especially those with stationary noise. Table 1 also shows that both MFCCs and RASTA-PLP features have a similar performance when the speech signal is polluted by non-stationary noise, e.g., factory noise at 0 dB. These results suggest that RASTA-PLP may contain more information on VAD for the DNN when considering signals with stationary noise. This more detailed information may be attributed to the three components of hearing psychophysics represented by RASTA-PLP, namely, critical-band spectral selectivity, equal loudness curve, and intensity loudness power law. To further improve performance, we included dynamic features that represent time-varying signal characteristics. In our analysis, the addition of dynamic features (i.e., delta and delta–delta features) improved the performance of the VAD DNN. In fact, Table 1 shows that the dynamic features can boost the performance of VAD for MFCCs, RASTA-PLP, and mel filter bank, with the latter obtaining the highest improvement, with a performance close to that of log power spectra and AMS. Hence, dynamic features may be also important for VAD.

Finally, we evaluated the distribution of each feature to visualize its separability. Hence, we used t-SNE because log power spectra, mel filter bank, MFCCs, RASTA-PLP, and AMS convey high dimensional data, and t-SNE allows to preserve the intrinsic data structure by grouping similar data points in a lower-dimensional space. Figure 4 shows that the feature values are clustered according to their classes, which are represented by different colors. Log power spectra (Figure 4a) and AMS (Figure 4e) show stronger clustering than the other features, as the data points from the two classes (i.e., speech and non-speech) are clearly separated, which implies that these features exhibit a high variability that is required to retrieve suitable classification. Hence, both features may be the best suited as inputs for the VAD DNN by carrying more discriminant information at every dimension. Next, the mel filter bank (Figure 4c) shows stronger separability than both MFCCs (Figure 4b) and RASTA-PLP (Figure 4d), which show more scattered points. Hence, these “hand crafted” features have the lowest separability among the evaluated features, thus indicating a lower variability for classification than log power spectra, which is a raw feature.

From the abovementioned experiments, we reached to some important insights regarding VAD using DNNs. First, AMS and log power spectra were the most effective features for DNN-based VAD, clearly outperforming the other evaluated features. In addition, we confirmed that a raw feature, such as log power spectra, can be more suitable than features that demand several calculations. Therefore, utilizing such raw features may reduce the complexity related to VAD, and thus allow its widespread use in real-time and portable speech applications.

5. CONCLUSION

DNNs can be used as a robust method for VAD. In fact, a DNN may be one of the most accurate classifiers for VAD, but the features used as input can notably affect its performance. Hence, we evaluated the performance of different features in a VAD DNN. We aimed to identify the features that have the most discriminant information to provide a high classification accuracy, and thus improve VAD. Furthermore, we intended to identify and select the best features for a VAD DNN to reduce the related computational cost.

In this paper, we analyze five speech features: log power spectra, mel filter bank, MFCCs, RASTA-PLP, and AMS, to determine their contribution to a VAD DNN. As a result, we found that a feature that does not require much processing, i.e., log power spectra, outperforms features involving several calculations such as mel filter bank, MFCCs, and RASTA-PLP. Hence, log power spectra can be useful in reducing the computational cost and obtaining a high performance to detect speech and non-speech periods. However, for speech signals polluted by non-stationary noise, such as babble noise, the perception-based feature AMS showed the best performance as input for the VAD DNN. Thus, AMS appears to have more discriminative information for the VAD DNN than its counterparts.

Overall, the results from this study suggest the importance of feature selection, and that raw features may be more suitable than its sophisticated counterparts as input for a VAD DNN. Nevertheless, future studies will require evaluating different features and DNN structures such as convolutional neural networks. Likewise, the DNN hidden layer processes should be thoroughly investigated to improve efficiency, and dynamic features should be evaluated with more detail. We expect that selecting the most appropriate DNN structures and features will ultimately lead to a highly efficient VAD that can be applied in real time and embedded in portable devices with low computational cost, thus spreading the use and benefits of speech-based applications in our daily life.

REFERENCES:

[1] J. Ramirez, Juan Manuel Górriz, and José Carlos Segura, “Voice Activity Detection Fundamentals and Speech Recognition System Robustness”, *Robust Speech Recognition and Understanding*. InTech, 2007.

[2] E. Verteletskaya, and Kirill Sakhnov, “Voice Activity Detection for Speech Enhancement Applications”, *Acta Polytechnica*, Vol. 50, No. 4, 2010, pp. 100–105.

[3] K.T. Sreekumar, Kuruvachan K. George, K. Arunraj, and C. Santhosh Kumar, “Spectral Matching Based Voice Activity Detector for Improved Speaker Recognition”, *2014 IEEE International Conference on Power Signals Control and Computations (EPSCICON)*, (India), January 6–11, 2014, pp. 1–4.

[4] S. Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt, “Features for Voice Activity Detection: A Comparative Analysis”, *EURASIP Journal on Advances in Signal Processing*, Vol. 2015, No. 1, p. 91.

[5] L.R. Rabiner, and Marvin R. Sambur, “An Algorithm for Determining the Endpoints of Isolated Utterances”, *Bell Labs Technical Journal*, Vol. 54, No. 2, 1975, pp. 297–315.

[6] R.V. Prasad, Abhijeet Sangwan, H.S. Jamadagni, M.C. Chiranth, Rahul Sah, and Vishal Gaurav, “Comparison of Voice Activity Detection Algorithms for VoIP”, *Proceedings of the Seventh IEEE International Symposium on Computers and Communications (ISCC),* (Italy), July 1–4, 2002, pp. 530–535.

[7] K. Sakhnov, Ekaterina Verteletskaya, and Boris Simak, “Approach for Energy-Based Voice Detector with Adaptive Scaling Factor”, *IAENG International Journal of Computer Science*, Vol. 36, No. 4, 2009.

[8] P. Renevey, and Andrzej Drygajlo, “Entropy Based Voice Activity Detection in Very Noisy Conditions”, *Seventh European Conference on Speech Communication and Technology*, (Scandinavia), 2001.

[9] J. Ramırez, José C. Segura, Carmen Benıtez, Angel De La Torre, and Antonio Rubio, “Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information”, *Speech Communication,* Vol. 42, No. 3, 2004, pp. 271–287.

[10] Y. Ma, and Akinori Nishihara, “Efficient Voice Activity Detection Algorithm Using Long-Term Spectral Flatness Measure”, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2013, No. 1, 2013, p. 87.

[11] T. Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li, “Voice Activity Detection Using MFCC Features and Support Vector Machine”, *International Conference on Speech and Computer (SPECOM07),* (Russia), Vol. 2, 2007, pp. 556–561.

[12] T. Kinnunen, and Padmanabhan Rajan, “A Practical, Self-Adaptive Voice Activity Detector for Speaker Verification with Noisy Telephone and Microphone Data”, *ICASSP*, 2013, pp. 7229–7233.

[13] T. Fukuda, Osamu Ichikawa, and Masafumi Nishimura, “Phone-Duration-Dependent Long-Term Dynamic Features for a Stochastic Model-Based Voice Activity Detection”, *Ninth Annual Conference of the International Speech Communication Association*, (Australia), September 22–26, 2008, pp. 1293–1296.

[14] K. Ishizuka, Tomohiro Nakatani, Masakiyo Fujimoto, and Noboru Miyazaki, "Noise Robust Voice Activity Detection Based on Periodic to Aperiodic Component Ratio", *Speech Communication,* Vol. 52, No. 1, 2010, pp. 41–60.

[15] K.R. Krishnamachari, Robert E. Yantorno, Daniel S. Benincasa, and Stanley J. Wenndt, "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments under Co-Channel Conditions", *IEEE International Symposium Intelligent Signal Processing and Communication Systems*, 2000.

[16] K. Pek, Takayuki Arai, and Noboru Kanedera "Voice Activity Detection in Noise Using Modulation Spectrum of Speech: Investigation of Speech Frequency and Modulation Frequency Ranges", *Acoustical Science and Technology,* Vol. 33, No. 1, 2012, pp. 33–44.

[17] J. Sohn, Nam Soo Kim, and Wonyong Sung, "A Statistical Model-Based Voice Activity Detection", *IEEE Signal Processing Letters,* Vol. 6, No. 1, 1999, pp. 1–3.

[18] A. Davis, Sven Nordholm, and Roberto Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 2, 2006, pp. 412–424.

[19] Q-H. Jo, J-H. Chang, J. W. Shin, and N. S. Kim, "Statistical Model-Based Voice Activity Detection Using Support Vector Machine", *IET Signal Processing*, Vol. 3, No. 3, 2009, pp. 205–210.

[20] D. Enqing, Liu Guizhong, Zhou Yatong, and Zhang Xiaodi, "Applying Support Vector Machines to Voice Activity Detection", *6th IEEE International Conference on Signal Processing*, (China), Vol. 2, August 26–30, 2002, pp. 1124–1127.

[21] T. Hughes, and Keir Mierle, "Recurrent Neural Networks for Voice Activity Detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Canada), May 26–31, 2013, pp. 7378–7382.

[22] T.V. Pham, Chien T. Tang, and Michael Stadtschnitzer, "Using Artificial Neural Network for Robust Voice Activity Detection under Adverse Conditions", *IEEE International Conference on Computing and Communication Technologies,* (Vietnam), July 13–17, 2009, pp. 1–8.

[23] A. Mohamed, Geoffrey Hinton, and Gerald Penn, "Understanding How Deep Belief Networks Perform Acoustic Modelling", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (Japan), March 25–30, 2012, pp. 4273–4276.

[24] X-L. Zhang, and Ji Wu, "Deep Belief Networks Based Voice Activity Detection", *IEEE Transactions on Audio, Speech, and Language Processing,* Vol. 21, No. 4, 2013, pp. 697–710.

[25] N. Ryant, Mark Liberman, and Jiahong Yuan, "Speech Activity Detection on YouTube Using Deep Neural Networks", *INTERSPEECH*, 2013, pp. 728–731.

[26] M. Espi, Masakiyo Fujimoto, Daisuke Saito, Nobutaka Ono, and Shigeki Sagayama, "A Tandem Connectionist Model Using Combination of Multi-scale Spectro-Temporal Features for Acoustic Event Detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (Japan), March 25–30, 2012, pp. 4293–4296.

[27] M. Cutajar, Edward Gatt, Ivan Grech, Owen Casha, and Joseph Micallef, "Comparative Study of Automatic Speech Recognition Techniques", *IET Signal Processing*, Vol. 7, No. 1, 2013, pp. 25–46.

[28] D. Yu, and Lee Deng, "Automatic Speech Recognition: A Deep Learning Approach", *Springer*, 2014.

[29] H. Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin, "Exploring Strategies for Training Deep Neural Networks", *Journal of Machine Learning Research*, Vol. 10, Issue Jan, 2009, pp. 1–40.

[30] G.E. Hinton, "Learning Multiple Layers of Representation", *Trends in Cognitive Sciences,* Vol. 11, No. 10, 2007, pp. 428–434.

[31] M.A. Carreira-Perpinan, and Geoffrey E. Hinton, "On Contrastive Divergence Learning", *Aistats*, Vol. 10, 2005, pp. 33–40.

[32] M.A. Keyvanrad, and Mohammad Mehdi Homayounpour, "A Brief Survey on Deep Belief Networks and Introducing a New Object Oriented Toolbox (DeeBNet) ", *arXiv preprint,* 2014, arXiv:1408.3264.

[33] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling", *ISMIR*, 2000.

[34] K.S. Rao, and Anil Kumar Vuppala, "Speech Processing in Mobile Environments", *Springer Science & Business Media*, 2014.

[35] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech”, *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, 1990, pp. 1738–1752.

[36] H. Hermansky, and Nelson Morgan, “RASTA Processing of Speech”, *IEEE Transactions on Speech and Audio Processing,* Vol. 2, No. 4, 1994, pp. 578–589.

[37] H. Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn, “RASTA-PLP Speech Analysis Technique”, *IEEE International Conference on Acoustics, Speech, and Signal Processing,* (USA), Vol. 1, March 23–26, 1992, pp. 121–124.

[38] G. Kim, Yang Lu, Yi Hu, and Philipos C. Loizou, “An Algorithm that Improves Speech Intelligibility in Noise for Normal-Hearing Listeners”, *The Journal of the Acoustical Society of America*, Vol. 126, No. 3, 2009, pp. 1486–1494.

[39] J. Tchorz, and Birger Kollmeier, “SNR Estimation Based on Amplitude Modulation Analysis with Applications to Noise Suppression”, *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 3, 2003, pp. 184–192.

[40] J.S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, “TIMIT Acoustic–Phonetic Continuous Speech Corpus”, *Linguistic Data Consortium*, Vol. 10, No. 5, 1993.

[41] The Rice University, “Noisex-92 Database”, http://spib.linse.ufsc.br/noise.html (accessed 22/02/2017).

[42] L. Maaten, and Geoffrey Hinton “Visualizing Data Using t-SNE”, *Journal of Machine Learning Research*, Vol. 9, Issue Nov, 2008, pp. 2579–2605.