Robust Voice Activity Detection And Denoising Directed By Noise Classification

Ajith Kumar S
PG Scholar, Dept. of ECE, GMIT, Davangere

Harisha G C
Assistant Professor, Dept. of ECE, GMIT, Davangere

Abstract: In this paper voice activity detection (VAD) is devised as a two-class classification issue using support vector machines (SVM). The suggested scheme combines a noise vigorous speech processing feature extraction process together with SVM prototypes trained in diverse background noises for speech/non-speech grouping. A multi-class SVM is as well used to categorize background noises in order to select SVM prototype for VAD. The recommended VAD is verified with speech databank artificially distorted by dissimilar additive noise varieties. Investigational effects indicate that the suggested VAD can extract speech activity under various SNR circumstances and it is moreover insensitive to variable amounts of noise. De-noising is also performed by using Short time Spectral Amplitude Minimum Mean Square Error Method.

Keywords: Voice activity detection · Perceptual wavelet packet transform · Noise classification · Support vector machine

I. INTRODUCTION

Voice activity detection (VAD) is a method, which can notice speech and non-speech sectors from a speech signal. A characteristic casual speech is characterized by a speech to non-speech fraction of forty to sixty. Henceforth, the use of VAD might increase the channel capacity in addition to the power intake of voice communication systems. VAD can moreover help in stacks of speech linked applications such as speech coding [14], spontaneous speech recognition [10], & speech augmentation systems [13].The elementary technique of most VADs in practice nowadays comprises of a feature extraction phase traced by a decision part. The feature extraction phase abstracts acoustic constraints from the input speech signal for discrimination of speech and non-speech fragments. The conservative acoustic constraints are the short-time energy levels, zero-crossing rates, pitch period, & spectral difference. At that moment, the decision part makes use of these acoustic constraints with some decision guidelines to decide the VAD result. The decision guidelines might be simple threshold values or compound statistical models. It is possible to make use of a trained classifier for example support vector machines (SVM) for the decision rule part.

This paper displays an effective way hiring SVM for VAD in noisy atmospheres. Irrespective of the decision guidelines, using suitable features is very vital in the performance of VAD. In the meantime speech signals are non-stationary& enclose plenty of transient portions, it is not suitable to make use of a static time- frequency resolution system for feature extraction in VAD, in noisy backgrounds. Wavelet transform is founded on time-frequency signal analysis. The wavelet investigation implements a windowing technique with variable-sized sections. It permits the use of long time intervals, when they desire more exact low-frequency (LF) information, & shorter regions, wherever they desire high-frequency (HF) information. At this time, perceptual wavelet packet transform (PWPT) is used as a device for feature extraction. PWPT is exploited to modify the decomposition tree arrangement of the standard wavelet packet transform (WPT) so as to approximate the critical bands of the psychoacoustic prototype as close as possible.

The chief motive for inserting the psychoacoustic prototype in the PWPT is that humans can detect the desired speech in a loud environment without the former knowledge of the noise. Thus, the human’s auditorium arrangement can discriminate between diverse acoustical noises. In noisy atmospheres, the performance of VAD is rigorously affected. Every so often, there's main rule to deal with noise in VADs. In the first method, a speech enrichment procedure is generally used for noise reducing [7], and in the second, noise vigorous features are extracted from noisy speech for VADs [5].There are ample of diverse acoustical noises in the environment (such as babble, street, automobile, etc.), which lead to performance degradation of VADs. Typically, the outcome of dissimilar noises is not believed about in VADs. By varying the processing according to the style of background noise, the performance of VAD can be boosted. This involves noise classification, which has been used in
A. PERCEPTUAL WAVELET PACKET TRANSFORM

The mathematical work of the wavelet packet transform (WPT) was first projected by Coifman. WPT is a wavelet transform where the discrete-time (sampled) signal is passed all the way through added filters than the DWT, and consequently, there are additional HF sub-bands to suitably characterize the signal. The perceptual wavelet packet transform (PWPT) scheme is developed to fine-tune the decomposition tree arrangement of the conventional WPT so as to approximate the vital bands of the psychoacoustic representation. In the psychoacoustic representation, frequency components of sounds can be incorporated into critical bands that refer to bandwidths upon which individual reaction becomes significantly different. Critical bands are important in understanding many acoustic phenomena, for example perception of loudness, pitch, and resonance. One category of critical band scales is called Bark scale. Based on the measurements by Zwicker et al., the Bark scale z can be roughly spoken in terms of the linear frequency by:[1]

$$Z(f) = 13\arctan(7.6\times10^{-4}f) + 3.5\arctan(1.33\times10^{-4}f)^2$$

Where $f$ is the linear frequency in Hertz. The equivalent critical bandwidth (CBW) of the core frequencies can be expressed by:

$$CBW(fc) = 25 + 75(1 + 1.4 \times 10^{-6} fc^2)^0.69 \, [Hz]$$

Where $fc$ is the center frequency. Theoretically, the range of human’s auditorium frequency spreads from 20 to 20kHz and covers roughly 25 Barks. In view of the fact that the Bark scale is a function of linear frequency, the first piece of constructing the PWPT is to locate the sampling rate of speech signals so as to find out the valid Bark numbers. In this project, the underlying sampling rate was selected to be 8 kHz, resulting in a bandwidth of 4 kHz, inside this bandwidth, there are roughly 17 critical bands.

The tree composition of the PWPT can be constructed as shown in Figure 1(a). The equivalent frequency bandwidths of the PWPT tree are shown in Figure 1(b). It contains 16 disintegrated cells with 5 decomposition levels to approximate these 17 critical bands.

![Tree Structure of PWPT](image)

**Figure 1:** (a) The tree structure of the PWPT and (b) the frequency bandwidths for the PWPT tree, where $w_j$ defines the wavelet coefficient of the $j$th sub-band of PWPT, where $j = 1−17$. (The sampling frequency is chosen to be 8 kHz, resulting in a bandwidth of 4 kHz, inside this bandwidth, there are roughly 17 critical bands.)

II. PROPOSED METHOD

The proposed robust VAD uses a classification-based technique, in which classification models are trained using noisy speech signals in specific environments. Given a speech signal, a set of features for noise classification is extracted from a short period of silence at the beginning of signal. Features extracted from the silence portion are then used to identify the type of environment. Once knowing the environment type, the recognizer selects a corresponding model for classifying the rest of signal as speech or non-speech. First, the environment or noise classification module is constructed using PWPT and SVM. The computational overhead of the noise classification module should be kept as low as possible, so that the overall system can achieve an acceptable processing time. Then, a particular SVM is trained on noisy speech signals with various levels of SNR. In the following subsections, we have provided more detailed explications of the VAD process.

B. NOISE CLASSIFICATION

The goal of noise classification is to identify the type of speech environment. Here, a simple model based on multiclass SVM classifier is used to identify the type of noise.[1]

II. FEATURE EXTRACTION

The choice of signal features is habitually made on a priori awareness of the nature of the signals to be categorized. A diversity of signal features have been used for this purpose, comprising low-level factors for instance the zero-crossing rate, signal bandwidth, spectral centroid, signal energy, and mel frequency cepstral constants. PWPT is nominated as a tool for feature mining. For healthier refinement amongst dissimilar noises in the PWPT field, 3 features counting mean, standard deviation, and entropy are mined from each sub-band as[1]:
Where \( w_j(k) \) describes the \( k \)th constant of the \( j \)th sub-band of PWPT, where \( j = 1–17 \), \( N_j \) is the number of coefficients in \( j \)th sub-band, and \( k = 1, 2, \ldots, N_j \). \( h_j \) is normalized histogram of absolute values of wavelet coefficients at \( w_j \) sub-band, and \( L \) is the number of equivalent histogram altitudes[1].

At the end of feature mining phase, a stack of 51-dimensional feature vector is achieved. At this moment, PCA is used so as to extract the utmost important features. PCA has been extensively used for feature mining in pattern identification. The key idea of PCA is to project the new feature vector against principal component axes. These axes are perpendicular and agree to the directions of extreme variance in the original feature space. So, projecting input vectors against this principal subspace permits decreasing the redundancy in the original feature space along with the dimension of input vectors[1].

### III. NOISE GROUPING OUTCOMES

5 sorts of noise from NOISEX-92 counting factory, white, room, speech and car stood pre-processed by decreasing the sampling rate to 8 kHz. A full of 34,590 segments, correspondingly spread among the 5 classes, have been used for the grouping.

![Figure 2: Block diagram of the proposed VAD method](image)

B. VAD DIRECTED BY NOISE GROUPING

Once classifying the noise type by means of noise grouping procedure, a strong prototype based on noise type is created for a diversity of signal-to-noise ratios (SNRs).

### IV. FEATURE EXTRACTION

The feature extraction phase is used to raise discrimination among noise (non-speech) and speech for the grouping job. The procedure for feature extraction is specified as follows. The input signal \( x(n) \) sampled at 8kHz is divided into 32-ms overlapped segments with a 15-ms window shift. At that moment, 4 kinds of features are extracted from each frame for the grouping job[1]: (1) sum of autocorrelation (SAC) sequence, (2) entropy, (3) sum of local maxima (SLM) of power spectral density (PSD), and (4) mean of PWPT sub-bands.

![Figure 3: flow chart of the proposed VAD](image)
A. SUM OF AUTOCORRELATION SEQUENCE

The periodic property is an in-built characteristic of speech signals and is normally used to describe speech. The periodic properties of speech signals are exploited to precisely extract speech activity. Indeed, voiced or vowel speech sounds have a robust periodic property than unvoiced sounds and noise signals. Therefore, the famous autocorrelation function (ACF) is defined in the time domain to assess the periodic intensity of each frame. The subjective estimate of the ACF is shown as[1]:

$$R[k] = \frac{1}{N} \sum_{n=0}^{N-K-1} x(n)x(n+k), \forall k = 0,1,...,N-1$$  \hspace{2cm} (6)

The SAC sequence is used as the first feature:

$$SAC = \sum_{k=0}^{N-1} R(k)$$  \hspace{2cm} (7)

B. ENTROPY

Entropy is an arithmetical measure of randomness and amounts information content in a signal. As of periodic property of speech signal and arbitrary nature of noise, the entropy metric can efficiently distinguish them. So, we can use this quantity to distinguish noise and speech in each frame. The entropy index is defined as[1]:

$$H = -\sum_{k=0}^{K} h(k) \times \log_2(h(k))$$  \hspace{2cm} (8)

Where h is the normalized histogram of the unconditional value of the speech signal x(n) in a frame with length N, n = 0, 1, . . . , N – 1, and K is the number of analogous histogram stages[1].

C. SUM OF LOCAL MAXIMA OF POWER SPECTRAL DENSITY

The PSD of a static arbitrary procedure is statistically connected to the correlation sequence by the discrete-time Fourier transform. Generally, the more interrelated or anticipated a signal, the more determined its power spectrum, and contrariwise, the more arbitrary or unpredictable a signal, the more spread its power gamut. Consequently, the power gamut of a signal can be used to assume the presence of monotonous structures or interrelated patterns in the signal process[1]. Welch’s technique is used for PSD valuation, which is a nonparametric process. Welch’s technique is achieved by averaging modified periodograms from overlaid and windowed segments. Once attaining PSD for each frame, SLM of it is employed as the third feature for the grouping task. Fig.4.2.1.3.1noticeablydemonstrates that local maxima of PSD can successfully distinguish noise and speech signals[1].

D. MEAN OF PWPT SUB-BANDS

By means of PWPT, the input speech signal can be disintegrated into 17 sub-bands, which are equivalent to wavelet coefficient collections. The white noise occurs in all frequency sub-bands; yet, this is not true for other noises. Consequently, for improved refinement amongst noise and speech, mean of noisy speech in every PWPT sub-band is used as fourth feature[1]. Besides the 4features specified above, delta of every feature is used to exploit the association among adjacent frames in speech signal. The delta function for each feature is expressed as follows[1]:

$$F = 2F(n) - (F(n-1) + F(n+1))$$  \hspace{2cm} (9)

Where n is the frame number. At the end of feature mining step, we will have a heap of feature vector (FV) for each frame to organize it as speech or non-speech[1]:

$$FV = [SAC, H, SLM, M, SAC, H, SLM, M]$$  \hspace{2cm} (10)

Where both of the M and M comprises of 17 features equivalent to the 17 sub-bands of PWPT.

As a last step, we similarly applied PCA to these features so as to extract the utmost noteworthy ones to be used[1].

V. CONSTRUCTION OF SVM MODEL BASED ON NOISE TYPE

Having the feature vector, SVM is cast off for the grouping issue. By means of SVM, a vigorous prototype built on noise variety is made-up for a diversity of SNRs. In other lyrics, a specific SVM prototype is trained on noisy speech with several levels of SNR[1]. noiseless speech, whose SNR surpasses 30 dB, is too pooled in the training set of every single noisy acoustic prototype[1].

So as to build a precise SVM prototype for dissimilar noise types, speech utterances are used, in which each speech sample is artificially distorted by adding a particular noise type such as car, white, factory, room, and speech noise provided in, at different SNR levels.[1]It should be mentioned that the speech utterances are visually labeled into speech and non-speech classes. The SVM prototype has been trained by means of LIBSVM software device[1].

Support vector machines (SVM) for VADs are examined in the literature. The idea is very simple, by means of a feature

![Figure 4: Local maxima of power spectral density using Welch’s method of two sample frames of (a) noise and (b) speech signals](image)
VI. EXPERIMENTAL RESULTS

The suggested VAD was assessed in terms of the capability to distinguish speech from non-speech at diverse SNRs. By decreasing the sampling rate to 8 kHz, speech utterances were preprocessed and used for assessing the suggested VAD procedures. Each speech example was artificially distorted by adding 5 types of noise from NOISEX-92, i.e., factory, white, car, room, and speech-noise, at diverse SNR levels (20, 15, 10, 5, 0 dB).

Steps followed to get final VAD results along with denoising results from the beginning are as follows:

✓ CREATION OF NOISY SPEECH DATABASE:
A clear speech signal has to be taken and it has to be mixed with the noisy signal in order to get the noisy speech signal as shown in the Figure 5.

✓ TRAINING THE MULTICLASS SUPPORT VECTOR MACHINE(SVM):
A noisy signal such as car, room etc., from the noise database has to be taken to train the multiclass SVM and 3 types of feature such as mean, standard deviation and entropy are calculated for the type of background noise classification and are plotted as shown in the Figure 6.

✓ TRAINING VAD SVM:
The training of VAD SVM has to be done by taking any noisy speech signal from the noisy speech database as shown in Figure 7.

After training, the VAD output will be produced as shown in the Figure 8.

✓ DE-NOISING THE VAD OUTPUT:
The segmented speech activity with noise (VAD output) has to be taken as an input for the de-noising algorithm i.e., Short time Spectral Amplitude Minimum Mean Square Error Method. The output produced will be the de-noised version of the original noisy speech signal as shown in the Figure 9.
VII. CONCLUSION

In this project, a modest prototype for VAD centered on a noise grouping as the first phase of the procedure has been provided. Also, a new robust feature vector founded on the PWPT for both noise and speech/non-speech grouping has been proposed. One aspect that can be considered to explore in the future is to have speech phase information in the feature extraction process. Improving the VAD performance by using other classification algorithms is also considered. In view of additional noise forms in the suggested VAD can increase the performance in real-life applications. Forthcoming work must be done on these encouraging matters.

REFERENCES


