Voice Liveliness Identification Assisted By Noise Categorization

Jayashree Magadum[1] Sandhya Bevoor[2]
[1]PG Student [2]Assistant Professor
[1,2]Dept of DECS
[1,2]MMEC, Belgaum

Abstract— Voice Activity Detection (VAD) is a very important front end processing in all Speech and Audio processing applications. The performance of most if not all speech/audio processing methods is crucially dependent on the performance of Voice Activity Detection. Voice activity detection (VAD), is to detect the presence of speech in an audio signal degraded by noise, is widely applied in numerous modern speech communication systems. Since speech signals are non-stationary and contain many transient components, it is appropriate to use, perceptual wavelet packet transform (PWPT) as a tool for feature extraction especially in noisy environments. Voice activity detection (VAD) is a process, which can detect speech and non-speech segments from a audio signal. This method combines a noise robust speech processing feature extraction process together with SVM models trained in different background noises for speech/non-speech classification. A multiclass SVM is also used to classify background noises in order to select SVM model for VAD.

Key words: Voice Activity Detection, Perceptual Wavelet Packet Transform, Noise Classification, Support Vector Machine

I. INTRODUCTION

Voice activity detection (VAD) is a process, which can detect speech and non-speech segments from a speech signal. A typical conversational speech is characterized by a speech to-non-speech ratio of forty to sixty (1). Hence, the use of VAD could improve the channel capacity as well as the power consumption of voice communication systems. VAD can also help in many speech-related applications such as speech coding (2), automatic speech recognition (3), and speech enhancement systems. The basic procedure of most VADs in use today consists of a feature extraction step followed by a decision part. The feature extraction step extracts acoustic parameters from the input speech signal for discrimination of speech and non speech segments. The conventional acoustic parameters are the short-time energy levels, zero-crossing rates, pitch period, and spectral difference. Then, the decision part makes use of these acoustic parameters with some decision rules to determine the VAD result. The decision rules could be simple threshold values or complex statistical models. It is possible to use a trained classifier such as support vector machines (SVM) for the decision rule part. This paper shows an effective method employing SVM for VAD in noisy environments. In noisy environments, the performance of VADs is severely affected. Commonly, there are two main methodologies to deal with noise in VADs. In the first approach, a speech enhancement method is usually used for noise reducing (4), and in the second one, noise robust features are extracted from noisy speech for VADs (5). There are many different acoustical noises in the environment (such as babble, street, car, etc.), which result in performance degradation of VADs. Usually, the effect of different. noises is not considered in VADs. By modifying the processing according to the type of background noise, the performance of VAD can be enhanced. This requires noise classification, which has been used in many applications, such as robust speech recognition , and speech enhancement. VAD has received considerable attention from the research community. In high quality recording conditions, energy-based methods perform well. In noisy conditions however, energy-based measures often produce a considerable number of false alarms. For this reason, a large variety of other features have been investigated for use in noisy environments . Such techniques often require tuning parameters to a particular noise environment for them to be effective, and have difficulties dealing with non-stationary or instantaneous types of noises that are frequent in our task.

II. LITERATURE REVIEW

Voice activity detection (VAD) refers to the ability of distinguishing speech from noise and is an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hand-free telephony, and echo cancellation. Although the existed VAD algorithms performed reliably, their feature parameters are almost depended on the energy level and sensitive to noisy environments.

As a key module for many speech processing applications , VAD has received considerable research interests in the last couple of decades. It started as early as in the 1970s (6), when VAD was often referred to as speech endpoint detection or word boundary detection problem. Back then, VAD algorithms often dealt with only little or no noise corruption in speech coding applications, and with separate recording utterances in speech recognition systems . Up to recently , advances in various speech applications require the detection of human speech in a continuous real-time fashion, and is often corrupted by a wide variety classes of noise. Algorithms for VAD had grown accordingly over the years.

Most of the speech activity detectors are based on either time domain or frequency domain approach. Various time domain features like short-time average energy (STAE), shorttime average magnitude (STAM), zero-crossing rate (ZCR) and so on are used in time domain. On the other hand, in frequency domain, various spectral information are used for designing a SAD. There are numerous examples where these time and frequency domain information and their statistical properties are used to develop robust speech activity detector. Speech activity detectors based on periodicity measure of speech signal is used in (7). Cepstral information based SAD is proposed in (8). SAD based on long term speech information is proposed for automatic speech recognition. Transformed domain characteristics of speech signal are used to design SAD in . Entropy based SAD is also proposed.

Using appropriate features is very important in the performance of VAD. Since speech signals are non-stationary and contain many transient components, it is not appropriate to use a fixed time– frequency resolution method for feature extraction in VAD, especially in noisy environments. Wavelet transform is based on time–frequency signal analysis. The wavelet analysis adopts a windowing technique with variable-sized regions. It allows the use of long time intervals, when we want more precise low-frequency (LF) information, and shorter regions, where we want high-frequency (HF) information. Here, perceptual wavelet packet transform (PWPT) is used as a tool for feature extraction. That is how applying SVM for VAD is different from other methods

III. PERCEPTUAL WAVELET PACKET TRANSFORM

The mathematical work of the WPT was first proposed by Coifman . WPT is a wavelet transform where the discrete-time (sampled) signal is passed through more filters than the DWT, and therefore, there are more HF sub-bands to appropriately represent the signal. The PWPT method is developed to adjust the decomposition tree structure of the conventional WPT in order to approximate the critical bands of the psychoacoustic model. In the psychoacoustic model, frequency components of sounds can be integrated into critical bands that refer to bandwidths at which subjective response becomes significantly different . Critical bands are important in understanding many auditory phenomena, such as perception of loudness, pitch, and timbre. One class of critical band scales is called Bark scale. Based on the measurements by Zwicker et al. , the Bark scale z can be approximately expressed in terms of the linear frequency by:

$$z(f) = 13\arctan\left(7.6\times10^{-4}f\right) + 3.5\arctan\left(1.33\times10^{-4}f\right)^2 \text{ [Bark]} \quad (1)$$

where f is the linear frequency in Hertz. The corresponding critical bandwidth (CBW) of the centre frequencies can be expressed by

$$\text{CBW}(f_c) = 25 + 75\left(1 + 1.4\times10^{-6}f_c^2\right)^{0.69} \text{ [Hz]} \quad (2)$$

where *fc* is the center frequency. Theoretically, the range of human's auditorium frequency spreads from 20 to 20,000Hz and covers approximately 25 Barks.

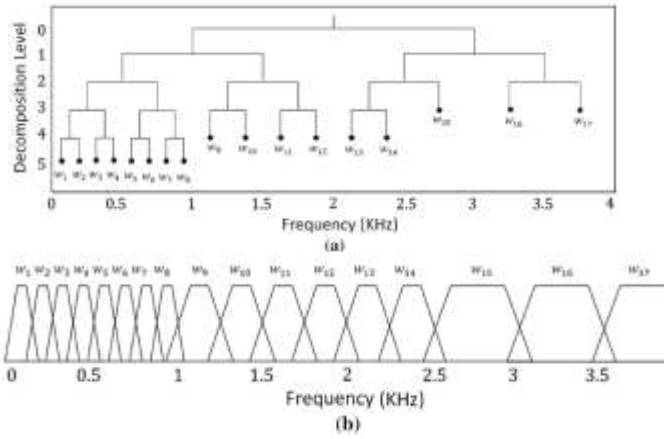
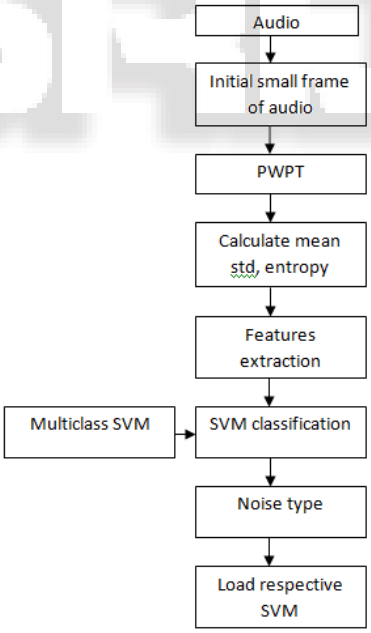


Fig. 1 A: The Tree Structure Of The PWPT And B The Frequency Bandwidths For The PWPT Tree,

Since the Bark scale is a function of linear frequency, the first step of constructing the PWPT is to set the sampling rate of speech signals in order to determine the valid Bark numbers. In this paper, the underlying sampling rate was chosen to be 8 kHz, yielding a bandwidth of 4 kHz. Within this bandwidth, there are approximately 17 critical bands . The tree structure of the PWPT can be constructed as shown in Fig. 1a. The corresponding frequency bandwidths of the PWPT tree are shown in Fig. 1b. It contains 16 decomposition cells with five decomposition stages to approximate these 17 critical bands.

IV. 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. Figure 2 shows block diagram of the proposed method, which consists of a number of essential stages:



(a)

The choice of signal features is usually based on a priori knowledge of the nature of the signals to be classified. A variety of signal features have been used for this purpose, including low-level parameters such as the zero-crossing rate, signal bandwidth, spectral centroid, signal energy, and melfrequency cepstral coefficients. PWPT is selected as a tool for feature extraction. For better discrimination between different noises in the PWPT domain, three features including mean, standard deviation, and entropy are extracted from each subband as:

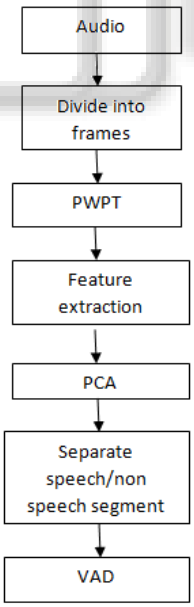
$$M_j = \frac{1}{N_j}\sum_{k=1}^{N_j}|w_j(k)| \quad (3)$$

$$\mathrm{Std}_j = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}\left(|w_j(k)| - |\overline{w_j}|\right)^2} \quad (4)$$

$$\mathrm{En}_j = -\sum_{l=0}^{L} h_j(l) \times \mathrm{Log}_2\left(h_j(l)\right) \quad (5)$$

Where *wj (k)* defines the *k*th coefficient of the *j* th sub-band of PWPT, where *j* = 1–17, *Nj* is the number of coefficients in *j* th sub-band, and *k* = 1, 2, . . . , *Nj* . *h j* is normalized histogram of absolute values of wavelet coefficients at *wj* sub-band, and *L* is the number of corresponding histogram levels. At the end of feature extraction step, a stack of 51- dimensional feature vector is obtained. Now, PCA is used in order to extract the most significant features. PCA has been widely used for feature extraction in pattern recognition. The main concept of PCA is to project the original feature vector onto principal component axes. These axes are orthogonal and correspond to the directions of greatest variance in the original feature space. Therefore, projecting input vectors onto this principal subspace allows reducing the redundancy in the original feature space as well as the dimension of input vectors. Here we are considering different types of noises like car, factory, room and white noises. A trained multiclass SVM classifier is used in our project to find the type of noise.

The SVM model has been trained using LIBSVM software tool .

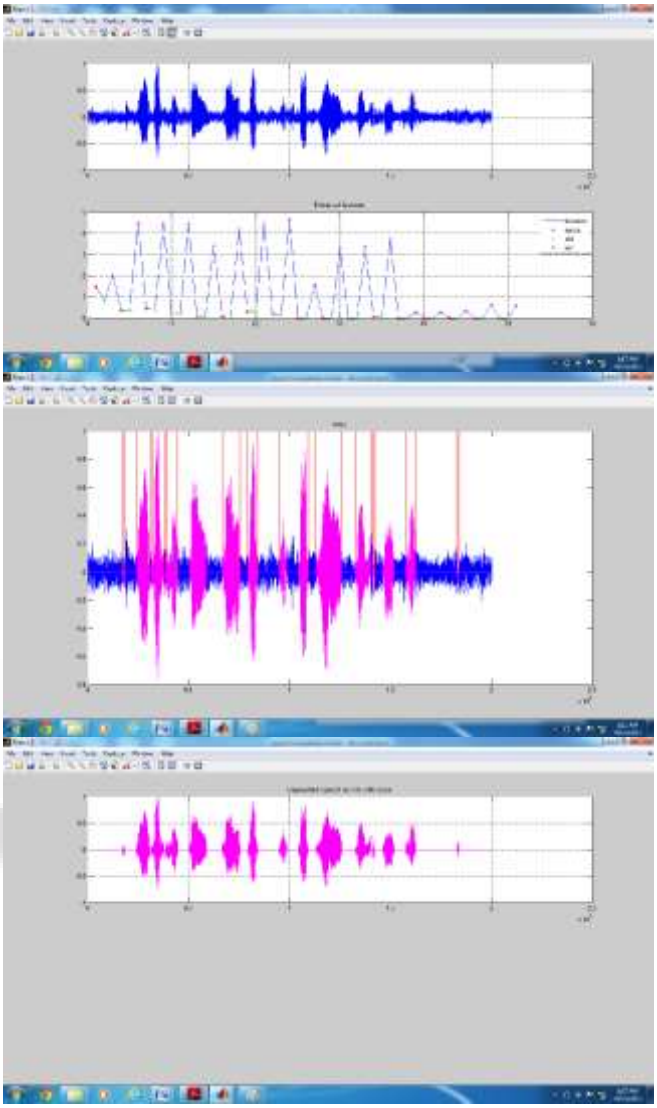


(b)

After identifying the noise type using noise classification algorithm, a robust model based on noise type is constructed.

The feature extraction step is used to increase discrimination between noise (non-speech) and speech for the classification task. The algorithm for feature extraction is stated as follows. The input signal *x(n)* sampled and is decomposed into 32-ms overlapped frames with a 10-ms window shift. Then, four types of features are extracted from each frame for the classification task: (1) sum of autocorrelation (SAC) sequence, (2) entropy, (3) sum of local maxima (SLM) of power spectral density (PSD), and (4) mean of PWPT subbands.

V. Simulation Results



VI. Conclusion

We are trying to provide a simple model for VAD based on a noise classification as the first step of the algorithm. We have also proposed a new robust feature vector based on the PWPT for both noise and speech/non-speech classification. VAD show that the performance of the proposed algorithm is good than other VADs, especially in low SNRs. We are also considering improving the VAD performance by using other classification algorithms. Taking into account more noise types in the proposed VAD can improve the performance in real-world applications.

Reference

[1] Beritelli, F., Casale, S., Ruggeri, G.: Performance evaluation and comparison of ITU-T/ETSI voice activity detectors. In: Proceedings ICASSP, pp. 1425–1428 (2001)

[2] Srinivasant, K., Gersho, A.: Voice activity detection for cellular networks. In: Proceedings IEEE Speech Coding, Workshop, pp. 85–86 (1993)
[3] Karray, L., Martin, A.: Towards improving speech detection robustness for speech recognition in adverse environment. Speech Commun. **40**, 261–276 (2003)
[4] Ramírez, J., Segura, J.C., Benítez, M.C., Torre, Á.D., Rubio, A.J.: An effective subband OSF-based VAD with noise reduction for robust speech recognition. IEEE Trans. Speech Audio Process. **13**(6), 1119–1129 (2005)
[5] Wu, B.F., Wang, K.C.: Voice activity detection based on autocorrelation function using wavelet transform and Teager energy operator. Comput. Linguist. Chin. Lang. Process. **11**(1), 87–100 (2006)
[6] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," Bell System Tech. Jour., vol. 54, no. 2, pp. 297‑315, 1975.
[7] R. Tucker, "Voice activity detection using a periodicity measure," *Communications, Speech and Vision, IEE Proceedings I*, vol. 139, no. 4, pp. 377–380, Aug. 1992.
[8] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering (TENCON'93)*, no. 0, Oct 1993, pp. 321–324.

