Robust Voice Activity Detection Using the Spectral Peaks of Vowel Sounds

In-Chul Yoo and Dongsuk Yook

ABSTRACT—This letter proposes the use of vowel sound detection for voice activity detection. Vowels have distinctive spectral peaks. These are likely to remain higher than their surroundings even after severe corruption. Therefore, by developing a method of detecting the spectral peaks of vowel sounds in corrupted signals, voice activity can be detected as well even in low signal-to-noise ratio (SNR) conditions. Experimental results indicate that the proposed algorithm performs reliably under various noise and low SNR conditions. This method is suitable for mobile environments where the characteristics of noise may not be known in advance.

Keywords—Voice activity detection (VAD), spectral peak, mobile environment.

I. Introduction

With rapid advances in mobile devices and increasing demand for voice interfaces for such devices, the ability to distinguish human speech from other sounds is becoming crucial. The current state-of-the-art voice activity detection (VAD) methods are statistical pattern classification approaches [1], [2], such as Gaussian mixture models. In statistical VAD methods, both speech and noise models are trained via corresponding training data. Log likelihood ratio tests are applied to input data for speech and noise discrimination. While these VAD methods exhibit superior performance, noise characteristic estimation is required to train the noise models. This limits their use in unknown noisy environments. Noise can be non-stationary and can vary with time, especially in outdoor environments in which mobile devices, such as personal digital assistants (PDAs) and cellular phones, are frequently used. In such cases, obtaining noise estimates may be difficult or impossible because we cannot know in advance which types of noise may occur.

Many works have attempted to discover characteristic features of human voices that are present only in speech. Since such characteristic features have not yet been discovered, short-time energy, zero crossing rate (ZCR), low-variance spectrum (LVS), spectral entropy (SE), periodicity, and so on have been used instead [2]-[7]. While it is true that speech has such characteristics, the problem is that some non-speech sounds can also have such characteristics. This leads to a high false acceptance rate for specific kinds of noise. For example, loud white noise can also have high energy and ZCR, and noise with non-uniform spectral distributions, such as train noise, can have low spectral entropy. We propose a new VAD algorithm that uses the spectral peaks of vowel sounds. Since the spectral peaks of vowel sounds have distinctive characteristics for speech and noise discrimination, the proposed method is robust to unknown noise even in low signal-to-noise ratio (SNR) conditions. The main difference between the proposed approach and previous approaches utilizing vowel sound characteristics is that we focus on individual vowel sound characteristics, rather than characteristics of whole vowel sounds, as in [7].

II. VAD Using the Spectral Peaks of Vowel Sounds

We reduce the problem of detecting voice activity to the problem of detecting the presence of vowels. It is assumed that

Manuscript received Mar. 5, 2009; revised May 11, 2009; accepted June 15, 2009.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2006-311-D00822). It was also supported by the MKE (Ministry of Knowledge Economy), Rep. of Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute for Information Technology Advancement) (IITA-2008-C1090-0803-0006).

In-Chul Yoo (phone: +82 2 3290 3641, email: icyoo@voice.korea.ac.kr) and Dongsuk Yook (phone: +82 2 3290 3202, email: yook@voice.korea.ac.kr) are with the Speech Information Processing Laboratory, Department of Computer and Communication Engineering, Korea University, Seoul, Rep. of Korea.

doi:10.4218/etrij.09.0209.0104

vowel sounds are nearly unique to speech. It is highly unlikely that vowel-like sounds are present in non-speech sounds. Also, because consonant sounds are always accompanied by vowel sounds and their duration is short, the presence of vowel sounds can be directly related to the presence of speech. Vowel sounds are known to have distinctive spectral peaks. These are highly concentrated peaks of energy at specific spectral bands. The spectral peaks can remain higher than their surroundings even after severe types of noise corruption. To completely diminish a given peak, either severe attenuation must be applied to the peak, or severe noise must be added to the surroundings until the energy of the peak and its surroundings equalize.

Figure 1 shows an example in which the vowel sound of '*a*' is corrupted by car noise. From the figure, it is clear that many of the original spectral peaks of '*a*' remain higher than their surroundings even in the case of a noisy condition of -15 dB SNR. Therefore, if we can develop a method to detect the characteristic peaks of vowel sounds in corrupted signals, the presence of vowel sounds can be detected even in low SNR conditions. Once vowel sounds are detected, a hang-over-like scheme can be employed to include adjacent consonant sounds because consonant sounds are accompanied by vowel sounds. This implies that the ability to detect the spectral peaks of vowel sounds can be directly used for speech detection.

We assume that the positions of major spectral peaks are the most important factor in recognizing the vowel sounds rather than the relative sizes of peaks or the shapes in spectral valleys, which are vulnerable to noise. Under this assumption, we propose the peak-valley difference (PVD), which calculates the similarity between the peak signature vector S of a registered vowel sound and the spectrum X of an input sound as

$$\mathrm{PVD}(X,S)=\frac{\sum_{k=0}^{N-1}\left(X[k]\times S[k]\right)}{\sum_{k=0}^{N-1}S[k]}-\frac{\sum_{k=0}^{N-1}\left(X[k]\times\left(1-S[k]\right)\right)}{\sum_{k=0}^{N-1}\left(1-S[k]\right)}. \quad (1)$$

The peak signature vector S contains peak position information for a vowel sound. It is a binary vector obtained during a training step by dividing the spectral bands into peak and valley bands for the given vowel sound and assigning values of 1 to the peak band and 0 to the valley band. The PVD calculates the average energy difference between peak and valley bands for an input sound. In [8], a similar measure was proposed to detect the presence of mechanical sounds, such as doorbell sounds, to assist the hearing impaired. The PVD can be considered as a modified version for the task of voice activity detection. Since the value of the PVD is high for vowel sounds and low for other sounds, it can be directly used for voice activity detection as follows:

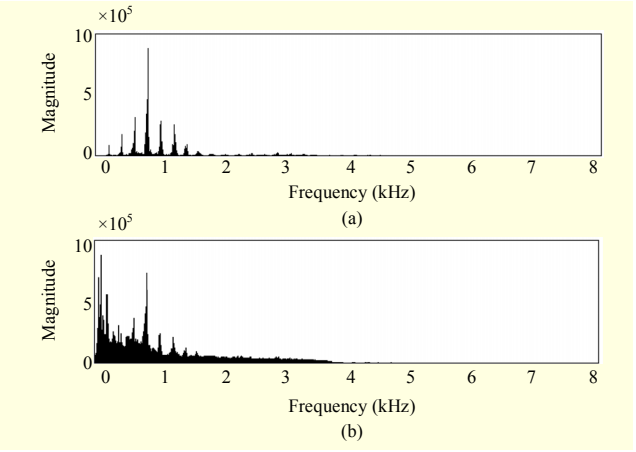

Fig. 1. Spectra of a vowel sound '*a*': (a) clean speech and (b) corrupted speech due to car noise with an SNR of -15dB.

$$\begin{cases} X \text{ belongs to voice, if } \max\limits_{S}\left(\mathrm{PVD}(X,S)\right)\geq \text{threshold} \\ X \text{ belongs to noise, otherwise.} \end{cases} \quad (2)$$

The advantage of PVD over other distance measures can be summarized as follows.

i) The resemblance between the input and reference signals in irrelevant spectral bands, which is caused by similar background noise, has much less effect on the overall similarity.

ii) The use of the average energy in spectral peaks and valleys makes the algorithm robust to changes in relative sizes of characteristic peaks caused by channel distortion, as long as the sum of the energy at peak areas remains relatively constant.

iii) PVD does not require the extraction of spectral peaks from input signals, which is a very difficult task in noisy environments.

III. Experiments

The 20,911 vowel segments from the entire test set of the TIMIT corpus were used to extract the binary vowel peak signature vectors. Each sound was divided into 128 ms blocks at 10 ms intervals. Fourier analysis with a Hamming window was applied to each block. Average spectra of each sound were grouped using the *k*-means clustering algorithm. This created 120 clusters. A simple energy-based peak finding scheme was applied to each cluster to extract the peak signature vectors.

To estimate the rejection threshold used in (2), the initial ten blocks of each utterance were regarded as non-speech segments. Using these segments, the rejection threshold values were calculated as

$$\text{threshold}=\frac{1}{10}\sum_{t=1}^{10}\max_{S}\left(\mathrm{PVD}\left(X_t,S\right)\right)+\alpha, \quad (3)$$

where X_t is the t-th spectrum of an input sound. We used the test set of the TIMIT corpus to find an optimal value for α.

We created a noisy version of the TIMIT corpus to evaluate the performance of the proposed PVD-based VAD method in various noise conditions. The entire training set of the TIMIT corpus was mixed with ten types of background noise. They included white noise, pink noise, and eight types of noise from the Aurora2 corpus, namely, airport, babble, car, exhibition, restaurant, street, subway, and train noises, at seven SNR levels between 0 dB and 30 dB with 5 dB increments. This created 323,400 test utterances (4,620 TIMIT train utterances×10 noise types×7 SNR levels). We added one second of silence to the beginning and end of each utterance before mixing it with the background noise to balance the length of speech and non-speech segments. Table 1 shows the VAD performance for various noise types (averaged over SNR levels) and SNR levels (averaged over noise types). For the purpose of comparison, conventional VAD methods, namely, low-variance spectrum-based VAD [2], spectral-entropy-based VAD [4], International Telecommunication Union G.729 Annex B VAD [5], and European Telecommunications Standards Institute adaptive multi-rate (AMR) VAD options 1 and 2 [6] were tested together with the proposed VAD method.

As seen in Table 1, the PVD-based method outperforms the conventional methods across various noise types and SNR conditions. In particular, the PVD-based method shows stable performance across various noise types in severely corrupted signal conditions. The low false-acceptance error of the PVD-based VAD method supports our assumption that the spectral peaks of vowel sounds are nearly unique to speech so that these peaks are unlikely to appear in non-speech noise.

Table 1. VAD accuracy for various SNR levels and noise types (%).

		G.729	AMR1	AMR2	LVS [2]	SE [4]	PVD
Noise types	Airport	62.5	69	74	90.2	72.4	96.6
	Babble	68.5	62	65.7	91.8	75.6	94
	Car	65.9	73.7	93.1	90.5	81.4	96.7
	Exhibition	66.5	69.3	89.9	87	91.5	96.6
	Pink noise	65.8	75.2	94.5	88.2	87.9	97.3
	Restaurant	63.2	63.7	63.7	92.7	73.3	92.5
	Street	58.1	62.1	83.2	82.3	59.4	90.8
	Subway	68	60.9	66.6	87.9	90.9	94.4
	Train	69.8	80.9	93.2	94.4	41.8	97.5
	White noise	63.7	70.9	94.8	86.6	92	97.5
SNR levels	0 dB	55.1	57.1	79.8	69	65.3	88.8
	5 dB	60.3	57.1	79.5	81.8	72	93
	10 dB	64.4	58.9	79.2	90.9	76.1	95.7
	15 dB	67.3	63.6	79.6	95	78.8	97.2
	20 dB	69.1	72.5	81.5	96.2	80.5	97.6
	25 dB	69.9	83.1	84.5	96	81.5	97.7
	30 dB	70.3	88.9	89.1	95.2	82.1	97.7
	Average	65.2	68.8	81.9	89.1	76.6	95.4

IV. Conclusion

In this letter, the problem of voice activity detection was mapped to vowel sound detection. Vowel sounds have distinctive spectral peaks that are robust to various corruptions. The proposed algorithm uses the peak signature of vowel sounds, which can be retrieved in the training step, to detect the presence of desired peaks in noisy input spectra. This direct calculation effectively avoids the problem of finding peaks from noisy spectra. Experiments under various conditions showed that the proposed method is robust to various kinds of noise. It also works well at low SNRs. Since the proposed algorithm does not require any prior knowledge of the noise characteristics, the algorithm can be used for mobile environments where unknown noise often occurs.

References

[1] J. Sohn, N.S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, 1999, pp. 1-3.

[2] A. Davis, S. Nordholm, and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, 2006, pp. 412-424.

[3] L.F. Lamel et al., "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 4, 1981, pp. 777-785.

[4] J.L. Shen, J.W. Hung, and L.S. Lee, "Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Environments," *Proc. Int. Conf. Spoken Language Process.*, paper 0232, 1998.

[5] ITU-T, *A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70*, ITU-T Rec. G. 729 Annex B, 1996.

[6] ETSI, *Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, GSM 06.94 v7.1.1 (ETSI EN 301 708), 1999.

[7] I.D. Lee, H.P. Stern, and S.A. Mahmoud, "A Voice Activity Detection Algorithm for Communication Systems with Dynamically Varying Background Acoustic Noises," *Proc. Veh. Technol. Conf.*, vol. 2, 1998, pp. 1214-1218.

[8] I.C. Yoo and D. Yook, "Automatic Sound Recognition for the Hearing Impaired," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, 2008, pp. 2029-2036.