Study on Robust Voice Activity Detection Using Empirical Mode Decomposition and Modulation Spectrum Analysis

Yasuaki Kanai (0910013)

School of School of Information Science,
Japan Advanced Institute of Science and Technology

February 6, 2012

Keywords: VAD, EMD, MSA, correct rate.

Voice activity detection (VAD) is a key technology for automatically detecting speech and non-speech periods from observed signals. VAD is widely used for various signal processes such as those in robust ASR systems, speech enhancements, and adaptive speech coding. Therefore, practical VAD must be able to accurately detect speech periods from observed signals in real environments in which there are non-speech signals and background noise.

Classic VAD methods have been carried out by thresholding the signal power or the number of zero crossings. These techniques can detect complete speech periods in clean environments (where there is only target speech). However, in noisy environments, there is a problem in that detection accuracy reduces remarkably due to the effect of noise. Modern methods, based on higher order statistics, the power of the low frequency band, and features based on periodicity/aperiodicity, are robust to noise, but they have weaknesses to non-speech signals such as musical, animal, and environmental sounds because these may have similar characteristics to speech such as periodicity/aperiodicity. Various applications that need to utilize VAD are generally used in such environments in which there is target speech with non-speech signals in background noise. However, speech periods must be accurately detected independently of these environments. This paper proposes a robust method of VAD in real environments even if they contain background noise and non-speech signals.

We propose a robust method of VAD that uses empirical mode decomposition (EMD) for reducing stationary noise and modulation spectrum analysis (MSA) for accurately determining speech/non-speech to achieve the previously stated purpose. EMD is used to decompose the signals into intrinsic mode functions (IMFs). Since speech is a non-stationary signal and typical background noise is a stationary signal, the components of their mixture (noisy speech) are decomposed by using EMD, into stationary IMFs and non-stationary IMFs. However, it is difficult to completely distinguish between stationary and non-stationary components in IMFs. The modulation spectrum of the signal, on the other hand, represents temporal fluctuations in the signal envelope. Some kinds of signals (e.g., voice, stationary noise, and environmental noise) may be distinguished by using this characteristic. However, it would be difficult to distinguish the modulation spectrum of the target signal in noisy environments because the features of the signals are mixed. The proposed method solves these issues by using two-step processing. Non-stationary components are extracted from noisy speech by using EMD, and then speech/non-speech periods are determined from them by using MSA. The proposed method where it decomposes a noisy signal into IMFs in the first step by using EMD, and then analyzes the modulation spectrum of each IMF to remove stationary IMFs from them. The remaining IMFs are used to resynthesize the non-stationary signal as a target signal. Then, signal periods are obtained in the second step by thresholding the power envelope of resynthesized signal. After this, speech/non-speech periods are detected from the signal periods by using MSA to find specific features of speech in the modulation spectrum.

Five simulations were carried out to evaluate the proposed method. The stimulus conditions used in the five simulations were: (1) clean speech, (2) speech in the background (stationary) noise, (3) speech with other non-speech, (4) speech with non-speech in background (stationary) noise, and (5) speech with non-speech in background (non-stationary) noise. Four datasets were used in these simulations: the ATR database A-set for speech signals, the Noisex–92 for stationary and environmental (nonstationary) noise, the RWC music database for musical sounds, and the Avian Vocalizations Center's database for bird calls. The stimuli used in the simulations were created from these datasets according to stimulus conditions. The same sampling frequency of 20 kHz was used in all stimuli. The conditions for the signal-to-noise ratio (SNR) were 20, 10, and 0 dB in these simulations to control additional noise levels. These were used as observed signal $y(t)$. Traditional (Otsu thresholding and G729 methods) and conventional (thresholding method on the power envelope) methods were compared with the proposed method. The correct rate, FAR(False acceptance rate) and FRR(False rejection rate) were used to evaluate the accuracy of VAD. The correct rate was calculated from the number of matched speech/non-speech periods and the total number of periods. FAR is rate of false acceptance is a speech periods and non-speech periods and FRR is rate of false rejection is a non-speech periods and speech periods. The FAR was calculated from the number of non-speech periods in detected speech periods and the number of detected speech periods. The FRR was calculated from the number of speech periods in detected speech periods and the number of detected non-speech periods.

We proposed a robust method of VAD using EMD and MSA, which provided a correct rate of about 90%, FAR of about 0% and FRR of about 20% under all experimental conditions. The results we obtained from the first experiment indicated that the new method could work well under clean conditions. It also indicated that the proposed approach with VAD had accuracy that was not inferior to the conventional methods. The results from the second experiments yielded a correct rate for the conventional method decreased as SNR decreased, and FAR increased. Even if SNR was too low for the proposed method, the correct rate drastically decreased and FAR drastically increased. These results confirmed that VAD was able to operate accurately, even if there was background noise. This was due to the effect of EMD in reducing background noise. The results from the third experiments yielded a correct rate for the conventional methods decreased and the correct rate for the proposed method drastically decreased, and FRR increased. These results confirmed that VAD was able to operate accurately, even if there were non-speech signals. This was due to the effect of MSA in making speech/non-speech decisions. The results from the fourth experiments yielded a correct rate for the conventional method decreased as SNR decreased, and FAR increased. Even if SNR was too low for the proposed method, the correct rate hardly decreased and FAR hardly increased. These results confirmed that VAD was able to operate accurately, even if non-speech signals and background noise existed simultaneously. These results indicated that the proposed method was robust against practical noise environments as well as the existence of non-speech signals. From the fifth experiments, the conventional method had the lowest correct rate. Moreover, VAD was not able to operate as accurately in a low SNR environment. The proposed method had the highest correct rate, even though it was in a low SNR environment. Moreover, the results from the last experiment revealed that the new method was effective under all conditions even if background noise was non-stationary. The results also demonstrated that the VAD method was made robust and accurate by combining two ideas of removing stationary noise by using EMD and extracting features from the modulation spectrum by using MSA, even if background noise and non-speech signals were included in real environments.