Concurrent processing of voice activity detection and noise reduction using empirical mode decomposition and modulation spectrum analysis

Yasuaki KANAI, Shota MORITA, and Masashi UNOKI

School of Information Science, Japan Advanced Institute of Science and Technology
{y-kanai, s-morita, unoki}@jaist.ac.jp

Abstract

Voice activity detection (VAD) is mainly used to detect speech/non-speech periods in observed noisy signals. The detected periods are used to reduce noise components or enhance speech components in noisy speech. However, current VAD techniques have serious problems in that the accuracy of detection of speech/non-speech periods drastically reduces if they are used for noisy speech and/or for mixtures of non-speech such as those in musical and environmental sounds. Thus, VAD needs to be robust to enable speech periods to be accurately detected in these situations. This paper proposes concurrent processing of VAD and noise reduction (NR) using empirical mode decomposition (EMD) and modulation spectrum analysis (MSA) to simultaneously resolve these problems. The proposed method effectively works on reducing stationary background noise by using EMD without estimating SNR (noise conditions), and then on reducing non-stationary noise including non-speech components by using MSA while this is determining speech/non-speech periods by thresholding the noise-reduced speech. Three experiments on VAD/NR in real environments were conducted to evaluate the proposed method by comparing it with typical methods (Otsu's method, G.729B, and AMR) and our previous methods. The results demonstrated that the proposed method could accurately detect speech/non-speech periods and effectively reduce noise components simultaneously.

Index Terms: voice activity detection, noise reduction, empirical mode decomposition, modulation spectrum analysis

1. Introduction

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 automatic speech recognition (ASR) systems, speech enhancement techniques, and adaptive and effective speech coding [1, 2]. Therefore, practical VAD must be able to accurately and robustly detect speech/non-speech periods from observed signals in real environments in which there are simultaneous non-speech signals and background noise.

There have been many previous studies on robust VAD, and many methods/algorithms have been proposed over the last few decades [1]. Classic VAD methods that utilize thresholding of the signal power [3] and the number of zero crossings [4], for example, can be used to detect complete speech periods in clean environments. However, there is a serious problem in noisy environments in that the accuracy of detection reduces remarkably due to the effect of background noise. Modern methods, based on higher order statistics [5], the power of the low frequency band [6], features based on periodicity or aperiodicity [7], and the non-stationarity properties of speech/noise [8] have been proposed for robust VAD in noisy environments. These are actually robust to background noise, but they have weaknesses to non-speech signals such as musical and environmental sounds because these may have similar characteristics to speech such as periodicity or aperiodicity and non-stationarity properties.

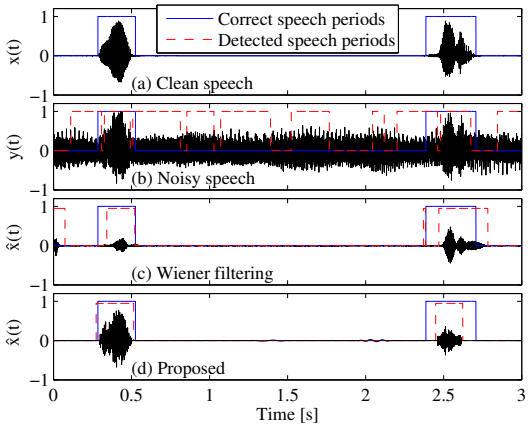

Figure 1: Examples of VAD. (a) Clean, (b) noisy, (c) and (d) noise reduced conditions with Wiener filtering and proposed method. Solid lines indicate correct speech periods. Dashed lines indicate speech periods detected by VAD.

Various applications with noise reduction (NR) methods that need to utilize VAD are generally used. Figure 1 shows, for example, the VAD for ASR in car interior environments in which there is target speech (male speech as shown in Fig. 1(a)) with non-speech signals (background music) in non-stationary background noise (car noise in the signal-to-noise ratio (SNR) of 0 dB), as seen in Fig. 1(b). Although NR techniques seem to be the best way of reducing these effects of stationary/non-stationary noise, the SNR of noise conditions in the short and long terms must be known before VAD for accurate NR. However, speech/non-speech periods could not be accurately detected independently in these environments, as seen in Fig.1(c), and these still remain a challenging problem. Thus, this issue is one of the most demanding tasks that remain to be resolved.

This paper proposes a novel approach to achieving concurrent processing of robust VAD and noise reduction in real environments (only noise conditions, without reverberation effect in this paper) even if they contain background noise and non-speech signals, as shown in Fig. 1(b). The proposed scheme can especially reduce stationary/non-stationary noise components to enhance speech components while it can robustly detect speech/non-speech periods, as shown in Fig. 1(d).

2. VAD using EMD and MSA

This section explains how we detect speech/non-speech periods from an observed signal, $y(t)$, using empirical mode decomposition (EMD) and modulation spectrum analysis (MSA). In this paper, the observed signal, $y(t)$, is assumed to be $y(t) = x(t) + n(t)$ where $x(t)$ is an original speech signal and $n(t)$ is a mixture of a non-speech signal and background noise.

2.1. Empirical mode decomposition (EMD)

The technique of EMD is used for analyzing non-stationary signals. Analytic signal $y(t)$ is decomposed by EMD into intrinsic mode functions (IMFs), $C_k(t)$, and negligible residue, $r(t)$. The $y(t)$ can be represented as

$$y(t) = \sum_{k=1}^{K} C_k(t) + r(t), \quad (1)$$

where k is the channel number and K is the number of IMFs (decomposition number). K depends on $y(t)$, i.e., IMFs depend on the waveform even if we use a part of the same $y(t)$. Details on the algorithm to calculate EMD are given elsewhere [9, 10].

EMD can also be regarded as common envelope-based decomposition [10]. IMFs are decomposed by the order of stationarity. Moreover, $y(t)$ can be resynthesized by summarizing all the IMFs based on Eq. (1).

2.2. Modulation spectrum analysis (MSA)

A modulation spectrum (MS) of observed signal $y(t)$ can be obtained from the short-term Fourier transform (STFT) of the power envelope of $y(t)$. The frame size is 1000 ms and the frame shift is 10 ms. A Hanning window was used in the STFT. Thus, MS represents the frequency characteristics of temporal fluctuations in the envelope of the signal.

The MSs for five kinds of signals were investigated to establish a diagnostic criterion [11]. These signals were voice [12], stationary noise [13], environmental noise [13], musical sounds [14], and bird calls [15]. The averaged MSs were investigated from 100 kinds of signals for all kinds of signals.

Based on these results, the MSa of speech and stationary noise are characterized in Fig. 2. The peak in the MS of voice has a high value and a broader band, as indicated by the solid line in Fig. 2. This is consistent with the peak in modulation frequency of 2 to 8 Hz that appears as a unique feature of speech [16, 17]. A target signal may be able to be easily distinguished as speech/non-speech periods by using this feature. The peak in the MS of stationary noise has a low value and a narrower band, and the others have flatter slopes as indicated by the dashed line in Fig. 2. These characteristics are used to find stationary IMFs.

2.3. Previous method

We previously proposed a framework to achieve a robust VAD technique that used EMD to reduce stationary noise without estimating SNR and MSA to accurately determine speech/non-speech to fulfill our previously stated purpose [11, 18].

The proposed approach focuses on stationarity in the IMFs as prior knowledge to remove stationary IMFs (stationary noise), and then remaining non-stationary IMFs (speech and non-speech signals) are used for detecting speech periods. An advantage in the use of EMD enabled us to decompose signals into stationary IMFs (lower IMFs) and non-stationary IMFs (higher IMFs), as explained in subsection 2.1. Since speech is a non-stationary signal, $x(t)$, and typical background noise, $n(t)$, is a stationary signal, noisy speech, $y(t)$, could be separated without having to estimate SNR. However, it was difficult to completely determine the boundary in the IMFs to separate lower and higher IMFs since the IMFs depended on signals.

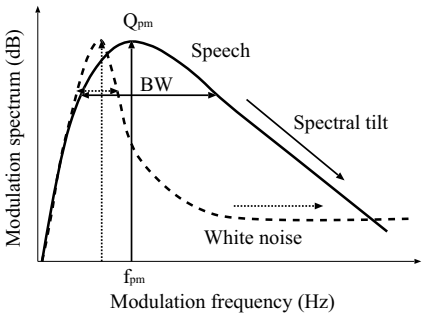

Figure 2: Modulation spectra of speech and stationary noise.

The proposed approach focuses on three characteristics of MS (peak frequency, f_{pm}; Q-value, $Q_{pm} = f_{pm}/\mathrm{BW}$ where BW is the 3-dB bandwidth of the MS; and MS's tilt over 4 Hz) as prior knowledge to find stationary IMFs (to remove stationary noise) and determine speech/non-speech signals. An advantage of the use of MSA enabled us to distinguish various kinds of signals such as speech and background noise, as explained in Subsection 2.2. However, it would be difficult to extract the fine MS of the target signal in noisy environments because the features of the signals are mixed in the time domain.

There were two previous methods that were prototypes to solve drawbacks with both EMD and MSA by effectively combining both advantages with EMD and MSA. The first one [11] was proposed to solve the first issue of how to completely determine the boundary between stationary and non-stationary IMFs. This is referred to as Method A. The second one [18] was proposed to solve the second issue of how to extract the fine MS of the target from non-stationary IMFs (higher IMFs) in noisy environments. This is referred to as Method B.

2.4. Problem

The previous methods correctly detected speech/non-speech periods in which there were background noise and non-speech signals. However, when background noise was non-stationary sound, the speech/non-speech periods might not have been correctly detected because there was no effect of noise reduction (NR) for non-stationary noise in lower IMFs. Therefore, non-stationary background noise could not be removed and the feature in MS was then distorted under the influence of background noise. This is a problem with decisions on speech/non-speech periods under real conditions.

3. Concurrent processing of VAD and NR

This paper proposes a novel approach to concurrent processing of robust VAD and NR in real environments to solve the above problem. Here, the three characteristics (peak frequency, Q-value, and MS's tilts) of MS with the speech signal are used to correctly determine speech periods in each IMF. At the same time, the three characteristics of MS with a non-speech signal are also used to accurately reduce non-stationary noise components in each IMF.

The process for the proposed method is given in Fig. 3. An observed noisy signal, $y(t)$, is decomposed into IMFs by using EMD (Fig. 3(a)), and all IMFs are then analyzed by MSA

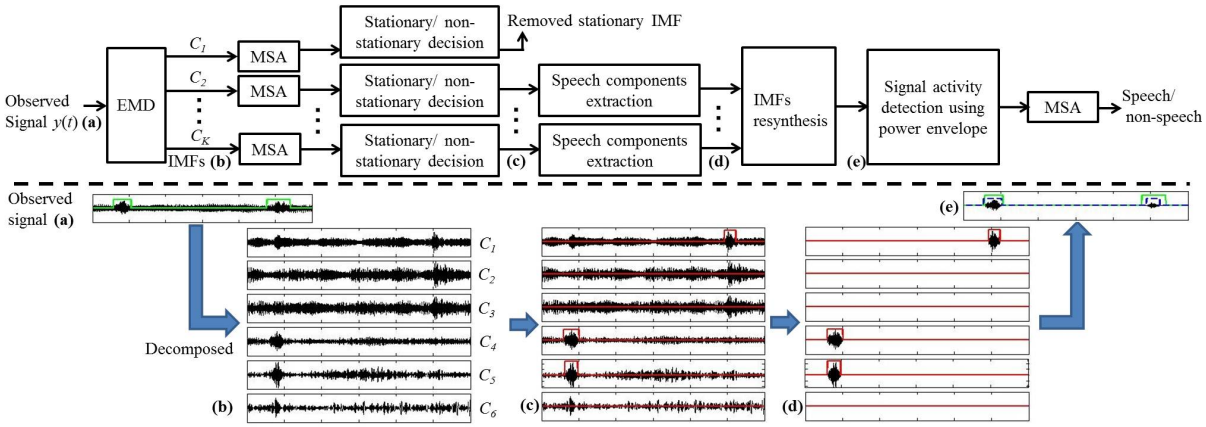

Figure 3: Block diagram of proposed method and processing flow for IMFs. (a) Observed signal, (b) decomposed IMFs (IMFs 1 to 6), (c) non-stationary IMFs, (d) speech periods of non-stationary IMFs, and (e) resynthesized signal and detected speech periods.

to accurately determine the boundary between the stationary and non-stationary IMFs (Fig. 3(b)). Stationary IMFs (lower IMFs) are then removed as stationary NR and non-stationary IMFs (higher IMFs) are used to reduce non-stationary noise and non-speech components as non-stationary NR (Fig. 3(c)). The speech components in each IMF are extracted by using MSA (Fig. 3(d)) and then resynthesized as an enhanced speech signal, (Fig. 3(e)). All periods of signal activities, i.e., candidates of speech periods, are detected from the enhanced speech signal, $\hat{x}(t)$, by thresholding the power envelopes of $\hat{x}(t)$. Finally, speech/non-speech periods are detected by using MSA.

Here, the three characteristics of MS ($f_{pm} \neq 3$ Hz, $Q_{pm} > 2$, and the slope of MS after 4 Hz > -0.5 dB/Hz) were used to determine the boundary between stationary and non-stationary IMFs, as shown in Fig. 3(c). The three characteristics of MS ($f_{pm} = 3$, $Q_{pm} < 2$, and the slope of MS after 4 Hz < -0.5 dB/Hz) were used to extract speech components in non-stationary IMFs, as shown in Fig. 3(d). These were also used to detect speech periods from the enhanced speech signals, as shown in Fig. 3(e). The MSs were obtained by using STFT (the Hanning window, 1000-ms frame, and 10-ms shift), as mentioned in Subsection 2.2. A threshold value in the signal activity detection in Fig. 3(e) was set to be relatively 3-dB downward and this was determined from the receiver operating characteristic (ROC) curve of VAD.

4. Evaluations

4.1. Evaluations for VAD

Three simulations were carried out to evaluate the proposed method. The stimulus conditions used in these simulations were: (1) clean speech, (2) speech with non-speech in stationary background noise, and (3) speech with non-speech in non-stationary background noise. Four datasets were used in these simulations: the ATR database A-set [12] for clean speech signals (100 samples: five male and five female speakers and 10 speech samples), the Noisex-92 [13] for stationary (one sample) and non-stationary noise (five samples), the real world computing (RWC) music database [14] for musical sounds (five samples), and the Avian Vocalizations Center's database [15] for bird calls (five samples).

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 SNR were 20, 10, and 0 dB in these simulations to control additional noise levels. These were used as observed signal $y(t)$. Typical (Otsu thresholding, G.729B, and AMR methods [19]) and conventional (thresholding method on the power envelope) methods were compared with the proposed method. The previous methods (Methods A and B) were also compared with the proposed method to verify improvements to the proposed approach. Thresholding on the power envelope was done with the same method that was used in the proposed approach. Therefore, we compared the results obtained from these methods to establish the effect of EMD and MSA.

The false acceptance rate (FAR) and false rejection rate (FRR) in VAD were used to evaluate the accuracy of VAD. FAR is the rate at which non-speech periods are detected as speech periods. FRR is the rate at which speech periods are detected as non-speech periods. The trade-off between FAR and FRR is a key concern in designing robust VAD. The correct rates (100−FAR and 100−FRR) were used in three simulations to represent the results of VAD in a trade-off relationship.

First, VAD was assessed under clean conditions. There were a total of 100 stimuli. The average values and standard deviations for 100−FAR [%] and 100−FRR [%] are in Figs. 4(a) and 4(b). These results indicate that the proposed and conventional methods collectively have high correct rates (100−FAR and 100−FRR). We proved that the proposed approach could accurately determine speech/non-speech periods as well as the conventional methods could under clean conditions.

Next, VAD's tolerance to non-speech signals and stationary background noise was assessed. Experimental stimuli were created from two speech signals (male speaker: mau, /a/ and /i/) and non-speech signals (/a/+non-speech signal+/i/). Sixteen non-speech signals were used for background noise conditions: one white noise, five of environment noise, five musical sounds, and five bird calls. The background noise signals were white noise, pink noise, and babble noise, and the conditions for the SNR were 20, 10, and 0 dB. Figures 4(c) and 4(d) present the results for the 100−FAR and 100−FRR of VAD.

The FRRs for all the methods had lower values. The FARs for the proposed method remained at lower values while those for the conventional methods were drastically increased. This was due to the effect of EMD in noise reduction and that of MSA in speech/non-speech decisions. In addition, the FARs for the conventional methods increased as SNR decreased. Even if SNR was too low for the proposed method, 100−FAR hardly

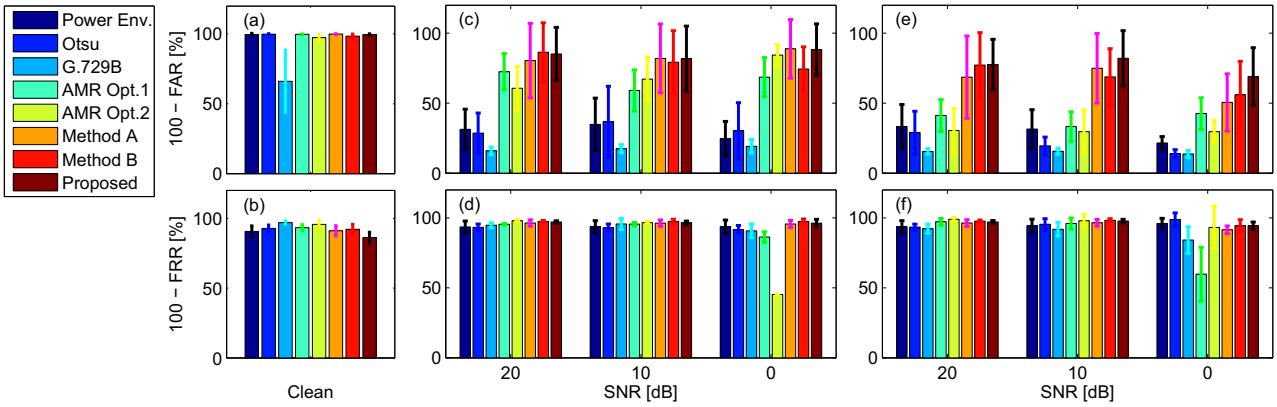

Figure 4: Results of (a) and (b) under clean conditions, results of (c) and (d) under stationary conditions, and results of (e) and (f) under non-stationary conditions. Top panels (a), (c), and (e) indicate $100 - \mathrm{FAR}$ and bottom panels (b), (d), and (f) indicate $100 - \mathrm{FRR}$.

increased. These results confirmed that the VAD we propose was able to operate accurately, even if non-speech signals and background noise existed simultaneously.

Finally, VAD's tolerance under conditions in which non-stationary background noise and non-speech signals existed simultaneously with target speech was assessed for the first base for real evaluations. The form of the stimuli we created changed background noise into three kinds of non-stationary noise (environment noise (factory 1), musical sound (rock), and a bird call (Cettia+diphone)), which were added to the stimuli in the second simulation. Figures 4(e) and 4(f) present the results for the $100-\mathrm{FAR}/100-\mathrm{FRR}$ of VAD. The FRRs all have lower values. The conventional methods have the highest FARs. Moreover, their FARs increased as SNR decreased. Even if SNR was too low for the proposed method, $100-\mathrm{FAR}$ maintained outstanding values. These results confirmed that VAD was able to operate accurately in real environments, even if non-speech signals and non-stationary background noise existed simultaneously.

4.2. Evaluations for noise reduction

We evaluated what effect reduced noise had on resynthesized signals, which is discussed in this section. The stimuli used in this evaluation were the same as those in the last evaluation of VAD. The measures we used were signal to error ratio (SER), log-spectral distortion (LSD), and the perceptual evaluation of speech quality (PESQ). Higher SERs and PESQs were favorable, and lower LSDs were also favorable. We investigated the SER, LSD, and PESQ of resynthesized speech and clean speech, and compared them with the SER, LSD, and PESQ of noisy speech and clean speech. After noise had been reduced, we evaluated how the three measures changed. The measures were investigated over the entire periods of the observed signal to evaluate improvements in the noise level of the whole signal, and not only the speech periods.

The SER, LSD, and PESQ of signals from noisy environments and a resynthesized signal with reduced noise were calculated with the proposed method, and the average of the difference was summarized for all SERs of background noise. The results are summarized in Table 1. Here, typical NR methods (spectral subtraction [20], MMSE-STSA [21], and Wiener filtering [22, 23]) were compared with the proposed method to verify improvements to the proposed approach. By comparing it with others, we found that the values of SER and PESQ increased favorably and the value of LSD decreased favorably after noise had been reduced with the proposed approach.

Table 1: Results for noise reduction.

Method		SNR conditions [dB] 20	10	0
No processing	SER [dB]	2.19	1.37	0.35
	LSD [dB]	19.29	20.12	20.66
	PESQ	2.27	1.97	1.71
Spectral subtraction [20]	SER [dB]	5.64	-3.82	-17.92
	LSD [dB]	17.04	18.29	20.38
	PESQ	2.55	2.03	1.84
MMSE log-STSA [21]	SER [dB]	5.63	-5.23	-10.90
	LSD [dB]	17.41	18..80	20.35
	PESQ	2.50	2.21	1.91
Wiener Filtering (Scalart, [22])	SER [dB]	5.95	-11.43	-21.41
	LSD [dB]	15.93	17.62	19.22
	PESQ	2.67	2.20	2.03
Wiener Filtering (2-step, [23])	SER [dB]	-2.44	-2.76	-4.14
	LSD [dB]	15.66	15.98	16.51
	PESQ	2.47	2.46	2.39
Proposed	**SER [dB]**	**4.80**	**4.19**	**2.59**
	LSD [dB]	**9.78**	**10.46**	**11.73**
	PESQ	**2.85**	**2.68**	**2.25**

5. Conclusion

We proposed concurrent processing of robust VAD and NR using EMD and MSA. We evaluated the efficiency of VAD with the proposed method in comparison with typical approaches. The results we obtained from three experiments indicated that the proposed method could work better than the conventional methods. The results from the first and second experiments indicated the FAR and FRR obtained with the proposed method were radically reduced in comparison with those with the typical methods. Moreover, the results from the last experiment revealed that the proposed method was vastly superior to robust VAD under non-stationary noise conditions than the typical methods. Therefore, the results revealed that the proposed approach could accurately detect speech/non-speech periods and effectively reduce noise components simultaneously, even if background noise and non-speech signals were included.

6. Acknowledgements

This work was supported by a Grant-in-Aid for Challenging Exploratory Research (No. 23650086) and an A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

7. References

[1] Ramriez, J., Gorriz, J. M., and Segura, J. C., "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," *Robust Speech Recognition and Understanding*, 1–22, 2007.

[2] Maleh, K. E. and Kabul, P., "A voice activity detection based on the adaptive integration of multiple speech features and a signal detection scheme," *Proc. IEEE 1997 Canadian Conference*, **2**, 470–473, 1997.

[3] Otsu, N., "A threshold selection method from gray-level histogram," *IEEE Trans. Syst. Man.*, SMC(9), 62–66, 1979.

[4] Rabiner, L. R. and Sambur M. R., "Algorithm for determing the endpoints of isolated utterance," *J. Acoust. Soc. Am.*, **56**(S1), S31, 1974.

[5] Sohn, J., Kim, N. S., and Sung, W., "A statistical model-based voice activity detection," *IEEE Signal Proc. Lett.*, **6**(1), 1–3, 1999.

[6] Benyassine, A. E., Shlomot, E., Su, H. Y., Massaloux, D., Lamblin, C., and Petit, J. P., "ITU-T recommendation G.729 annex B: A silencee compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applocation," *IEEE Trans. Commum. Mag.*, **35**, 64–73, 1997.

[7] Ishizuka, K. and Kato, H., "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," *Proc. ICASSP2006*, **1**, 789–792, 2006.

[8] Lu, X., Unoki, M., Isotani, R., Kawai, H., and Nakamura, S., "Adaptive regularization framework for robust voice activity detection," *Proc. Interspeech2011*, 2653–2656, 2011.

[9] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H., "The Empirical Mode Decomposition and the Hilbert Spectrum for nonlinear and nonstationary time series analysis," *Proc. the Royal Society: Mathematical, Physical and Engineering Sciences.*, **A454**, 903–995, 1998.

[10] Sawaguchi, T. and Unoki, M., "Investigation of a method of speech signal analysis using empirical mode decomposition and its application," *J. Signal Processing*, **14**(4), 273–276, 2010.

[11] Kanai, Y. and Unoki, M., "Study on Robust Voice Activity Detection Using Empirical Mode Decomposition and Modulation Spectrum Analysis," *J. Signal Processing,* **16**(4), 315–138, 2012.

[12] Takeda, K., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H., "Speech Database User's Manual", *ATR Technical Report TR-I-0028*, 1988.

[13] Varga, A., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, **12**(3), 247–251, 1993.

[14] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R., "RWC Music Database: Database of Copyright-cleared Musical Pieces and Instrument Sounds for Research Purposes," *IPSJ Journal*, **45**(3), 728–738, 2004.

[15] Avian Vocalizations Center's database: http://avocet.zoology.msu.edu/

[16] Atlas, L., Greenberg, S., and Hermansky, H., "The modulation spectrum and its application to speech science and technology," *Interspeech2007 Tutorial*, 2007.

[17] Kanedera, N., Arai, T., Hermansky, H., and Pavel, M., "On the importance of various modulation frequencies for speech recognition," *Proc. EuroSpeech 97*, 1079–1082, 1997.

[18] Kanai, Y. and Unoki, M., "Robust voice activity detection using empirical mode decomposition and modulation spectrum analysis," *Proc. ISCSLP2012*, 400–404, 2012.

[19] ETSI EN 301 708 v7.1.1, Digital cellular telecommunications system; Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channeles, 1999.

[20] Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP.*, **27**(2), 113–120, 1979.

[21] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, **33**, 443–445, 1985.

[22] Scalart, P. and Filho, J., "Speech enhancement based on a priori signal to noise estimation.," *Proc. ICASSP96*, 629–632, 1996.

[23] Plaous, C., Marro, C., and Scalart, P., "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, **14**(6), 2098–2108, 2006.