# Study of Noise Robust Voice Activity Detection Based on Periodic Component to Aperiodic Component Ratio

*Kentaro Ishizuka and Tomohiro Nakatani*

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikari-dai, Keihanna Science City, Kyoto 619-0237 Japan

{ishizuka, nak}@cslab.kecl.ntt.co.jp

## Abstract

This paper describes a study of noise robust voice activity detection (VAD) utilizing the periodic component to aperiodic component ratio (PAR). Although environmental sound changes dynamically in the real world, conventional noise robust features for VAD are sensitive to the non-stationarity of noise, which yields variations in the signal to noise ratio, and sometimes requires apriori noise power estimations. To overcome this problem, we adopt the PAR as an acoustic feature for VAD that is insensitive to the non-stationarity of noise. Hearing research also suggests that the decomposition of the periodic and aperiodic components plays an important role in the human auditory system. The proposed method first estimates the PAR of the observed signals with a harmonic filter in the frequency region. Then it detects the presence of target speech signals based on the voice activity likelihood defined in relation to the PAR. The performance of the proposed VAD algorithm was examined by using simulated and real noisy speech data. Comparisons confirmed that the proposed VAD algorithm outperforms the conventional VAD algorithms particularly in the presence of non-stationary noise.

## 1. Introduction

Voice activity detection (VAD) plays a crucial role in speech signal processing techniques. In particular, VAD in the “real world”, for example in a car, on the street, or at a railway station, is important for speech processing techniques such as speech enhancement [1], speech coding [2], and automatic speech recognition [3]. Since these techniques depend strongly on VAD accuracy or sometimes assume ideal VAD, insufficient accuracy seriously affects their practical performance. This has made it necessary to develop more robust VAD [4].

In general, VAD consists of two parts: an ‘acoustic feature extraction’ part, and a ‘decision mechanism’ part. The former extracts acoustic features that can appropriately indicate the probability of target speech signals existing in observed signals, which also include environmental sound signals. Based on these acoustic features, the latter part finally decides whether the target speech signals are present in the observed signals using, for example, a well-adjusted threshold [5], the likelihood ratio [6], and hidden Markov models [7]. The performance of the each part seriously influences the VAD performance.

The short-term signal energy and zero-crossing rate [8] have long been used as simple acoustic features for VAD. However, they are easily degraded by environmental noise, and environmental sounds also possess a similar energy and zero-crossing rate to speech signals. To cope with this problem, various kinds of robust acoustic features have been proposed for VAD. As acoustic features representing inherent characteristics of speech signals, such as autocorrelation function based features [9]-[11], spectrum based features that utilize harmonicity [12][13], pitch based features [14], the power in the band-limited region [5][15][16], mel-frequency cepstral coefficients [11], delta line spectral frequencies [15], and features based on higher order statistics [17] have been proposed. On the other hand, some methods employ the model of noise characteristics [18] or enhanced speech spectra derived from Wiener filtering based on estimated noise statistics [6][16].

Most of the above methods assume the stationarity of noise within a certain temporal length, thus they are sensitive to changes in the signal to noise ratios (SNRs) of observed signals and non-stationary noise. However, in practice, environmental sound is not stationary and its power changes dynamically within a short time. This sensitivity makes it difficult to decide the optimum threshold which prevents VAD methods from being used in the real world. Therefore, there is a need for VAD algorithms that are insensitive to the non-stationarity of noise.

Let us now turn to sound representation. Sound signals can be decomposed into their periodic and aperiodic components. For example, speech signals consist not only of periodic signals such as steady parts of vowels and voiced consonants, but also of non-periodic signals such as fluctuations included in vowels, voiced consonants, stops, fricatives, and affricates. With regard to psychoacoustics, findings derived from concurrent vowel identification experiments suggest that the human auditory system can suppress harmonic interferers and perceive the residual target signal [19]. This finding suggests that the human auditory system may process both the harmonic (periodic) component and the residue after canceling the harmonicity (aperiodic) component, which deviates from the dominant periodicity. Such a twofold representation of sounds has been studied with respect to speech/music synthesis [20][21], because the aperiodic component is responsible for the perceived speech/music quality [22]. In addition, in terms of automatic speech recognition, the word accuracy in noisy environments can be improved by using the periodic and aperiodic components of observed signals [23][24]. The above indicates the effectiveness of such a rich representation of sound signals. However, although VAD methods using only periodic characteristics of speech signals have long been studied [9]-[14], there is no VAD method that explicitly utilizes both periodic and aperiodic components.

In this paper, we propose a VAD algorithm that is insensitive to the non-stationarity of noise and which utilizes an acoustic feature that represents the power ratios of the periodic and aperiodic components in observed signals. This feature is referred to as the periodic component to aperiodic component ratio (PAR). With this method, the likelihood of the existence of target speech signals is calculated from probability distributions defined in relation to the PARs. Section 2 provides detailed explanations of our proposed method. Section 3 describes preliminary evaluation experiments that show the advantage of the proposed method by comparing it with conventional methods. Section 4 concludes this study and outlines future work.

## 2. Method

Let us first define the problem that the present method is designed to solve. Observed signals are recorded monaurally, and there is only one dominant sound, i.e. the target speech, which is in the presence of background noise whose frequency spectra distribute widely over all frequencies. There is no assumption as regards the stationarity of the noise power, thus the power of noise changes dynamically. In addition, there is no apriori knowledge about the kind of background noise.

To cope with this problem, our method first decomposes observed signals into their periodic and aperiodic components. While the conventional decomposition methods [20][21] aim to decompose the components inherently included in speech or music signals, our method aims to decompose the components included in all the observed signals to determine the periodic component of the dominant signals, namely speech signals. Therefore, in this paper, the term 'aperiodic component' includes both environmental noise and aperiodic components of speech signals, and the term 'periodic component' includes a dominant harmonic component in the observed signal. Because the PAR of noise is normally independent of the power of the noise, voice activity detection based on the PAR is expected to be insensitive to any dynamic change in the SNR. Note that the PAR is a similar measure to the harmonic noise ratio (HNR) [21]. However, in this paper, we use the term 'PAR' rather than 'HNR' to make it clear that the observed signals with the proposed method (speech in the presence of noise) differ from those in the speech analysis method (speech/music). Section 2.1 describes our decomposition method performed using sub-band signals in the frequency domain.

Our proposed VAD method then detects the existence of speech based on the statistics of the PARs, which represent the difference between periods that contain only noise and periods that contain both noise and speech. The statistical VAD method proposed by Sohn et al. [6] utilizes likelihoods derived from posteriori SNRs, whereas our method calculates likelihoods derived from probability distributions of the PARs without knowledge of the SNR. In addition, unlike other statistical VAD approaches [7][18] based on noise spectral characteristics, our method does not need these characteristics apriori, and so is expected to be insensitive to variations in noise characteristics. Section 2.2 describes our likelihood calculation in detail.

### 2.1. Decomposition of periodicity and aperiodicity

We first describe how to estimate the power of a sinusoidal component in the frequency region before explaining how to decompose the periodic (harmonic) and aperiodic (inharmonic) components of observed signals. Let us consider the following sinusoidal component:

$$s_p(t) = r\cos(\psi(t)) \quad s.t. \quad \psi(t) = \frac{2\pi f}{f_s}t + \phi$$

where $f_s$, $t$, $r$, $f$, and $\phi$ are the sampling frequency, sampling index, amplitude, frequency, and initial phase of the sinusoid. We assume that $f$ is over 50 Hz in the presence of the frequency components of ordinary speech signals. Here, we employ a bandpass filter $h(t) = g(t)\exp(j2\pi ft/f_s)$ to $s_p(t)$ as

$$\begin{aligned}\hat{s}_p(t) &= \sum_{l=0}^{L-1} s_p(t-l)h(l) \\ &= \frac{r}{2}\exp(j\psi(t))\sum_{l=0}^{L-1} g(l) + \frac{r}{2}\exp(-j\psi(t))\sum_{l=0}^{L-1} g(l)\exp\left(j\frac{4\pi f}{f_s}l\right)\end{aligned}$$

where $L$ is the length of the bandpass filter. When we assume that $g(t)$ is a lowpass filter, whose cutoff frequency is below 50 Hz, corresponding to the analysis window for a short time Fourier transform (STFT), the second term of the above equation can be disregarded. Therefore, we can obtain

$$r = \frac{2|\hat{s}_p(t)|}{\sum_{t=0}^{L-1} g(t)} \tag{1a}$$

Let $g^*(t) = g(L-1-t)$, which is a symmetric temporal window of $g(t)$, and $l' = (L-1) - l$, then $\hat{s}_p(t)$ can be rewritten as

$$\hat{s}_p(t) = \exp\left(j\frac{2\pi f}{f_s}(L-1)\right)S_p(n,m)$$

$$S_p(n,m) = \sum_{l'=0}^{L-1} g^*(l')s_p(l' + (t-(L-1)))\exp\left(-j2\pi\left(\frac{f}{f_s}L\right)l'/L\right) \tag{1b}$$

where $S_p(n,m)$ is an STFT representation of $s_p(t)$ at a frequency bin of $m = L(f/f_s)$ analyzed by a temporal frame whose index is $n$ beginning at $t_n = t-(L-1)$. Furthermore, the short temporal power $\rho_p(n)$ of $s_p(t)$ can be calculated as

$$\begin{aligned}\rho_p(n) &= \sum_{t=0}^{L-1}\left(g(t)s_p(t-t_n)\right)^2 \\ &= \frac{r^2}{2}\sum_{t=0}^{L-1} g(t)^2 + \frac{r^2}{2}\sum_{t=0}^{L-1} g(t)^2\cos(2\psi(t-t_n))\end{aligned}$$

where $n$ is the index of the temporal frame. Because $g(t)^2$ can be considered a lowpass filter similar to $g(t)$, the second term of the above equation can also be disregarded. Therefore, the short temporal power can be calculated as follows using equations (1a) and (1b).

$$\rho_p(n) = \eta|S_p(n,m)|^2 \quad \text{where} \quad \eta = \frac{2\sum_{t=0}^{L-1} g(t)^2}{\left(\sum_{t=0}^{L-1} g(t)\right)^2} \tag{2}$$

On the other hand, the short time power of a general signal $s(t)$ defined as (3) can be considered as the *0*th-order autocorrelation coefficient.

$$\rho(n) = \sum_{l=0}^{L-1}\left(g(l)s(l+t_n)\right)^2 \tag{3}$$

Since the autocorrelation coefficients are equal to the inverse Fourier transform of the power spectrum, the short time power can be also obtained as

$$\rho(n) = \frac{1}{M}\sum_{m=0}^{M-1}|S(n,m)|^2 \tag{4}$$

where $M$ is the maximum number of frequency bins.

Next, based on our estimation of the sinusoidal power described above, we explain how to estimate the power of a periodic (harmonic) component from an observed signal. We assume that an observed signal $s(t)$ can be described as the sum of its periodic and aperiodic components, $s_p(t)$ and $s_a(t)$, respectively.

$$s(t) = s_p(t) + s_a(t)$$

Hereafter, we denote STFT representations of the above signals by $S(n,m)$, $S_p(n,m)$, and $S_a(n,m)$, respectively, and their short time power in a time frame by $\rho(n)$, $\rho_p(n)$, and $\rho_a(n)$, respectively. Now, we assume the additivity on the powers of the components as

$$|S(n,m)|^2 = |S_p(n,m)|^2 + |S_a(n,m)|^2 \tag{5a}$$

Then the following equation can also be derived from the above assumption and equation (4).

$$\rho(n) = \rho_p(n) + \rho_a(n) \tag{5b}$$

Let us denote the fundamental frequency (F0) of the periodic component and the number of harmonics at frame $n$ by $f_0(n)$ and $\nu(n)$, respectively, and an operator to transform the $k$-th harmonic frequency $kf_0(n)$ to an index of a frequency bin in the corresponding frequency domain by $[kf_0(n)]$. Then, from equations (2) and (5a), we can obtain the following equation.

$$\begin{aligned}\rho_p(n) &= \sum_{k=1}^{\nu(n)} \left(\eta |S_p(n,[kf_0(n)])|^2\right) \\ &= \eta\left(\sum_{k=1}^{\nu(n)} |S(n,[kf_0(n)])|^2 - \sum_{k=1}^{\nu(n)} |S_a(n,[kf_0(n)])|^2\right)\end{aligned} \tag{6}$$

In addition, we introduce the following assumption.

**Assumption**: $\frac{1}{M}\sum_{m=0}^{M-1} |S_a(n,m)|^2 = \frac{1}{\nu(n)}\sum_{k=1}^{\nu(n)} |S_a(n,[kf_0(n)])|^2$,

which means that the average power of the aperiodic components at the frequencies of the dominant harmonic components is equal to that over the whole frequency range. According to this assumption, we can obtain

$$\begin{aligned}\sum_{k=1}^{\nu(n)} |S_a(n,[kf_0(n)])|^2 &= \nu(n)\left(\frac{1}{\nu(n)}\sum_{k=1}^{\nu(n)} |S_a(n,[kf_0(n)])|^2\right) \\ &= \nu(n)\left(\frac{1}{M}\sum_{m=0}^{M-1} |S_a(n,m)|^2\right) \\ &= \nu(n)\rho_a(n)\end{aligned} \tag{7}$$

Substituting equations (4), (6), and (7) to equation (5b), we derive the following equation.

$$\rho(n) = \eta\left(\sum_{k=1}^{\nu(n)} |S(n,[kf_0(n)])|^2 - \nu(n)\rho_a(n)\right) + \rho_a(n)$$

Consequently, we can obtain the following equations.

$$\rho(n) = \frac{1}{M}\sum_{m-0}^{M-1} |S(n,m)|^2 \tag{8}$$

$$\hat{\rho}_a(n) = \frac{\rho(n) - \eta\sum_{k=1}^{\nu(n)} |S(n,[kf_0(n)])|^2}{1-\eta\nu(n)} \tag{9}$$

$$\begin{aligned}\hat{\rho}_p(n) &= \rho(n) - \hat{\rho}_a(n) \\ &= \eta\frac{\sum_{k=1}^{\nu(n)} |S(n,[kf_0(n)])|^2 - \nu(n)\rho(n)}{1-\eta\nu(n)}\end{aligned} \tag{10}$$

where $\hat{\rho}_p(n)$ and $\hat{\rho}_a(n)$ indicate estimated values for true values of $\rho_p(n)$ and $\rho_a(n)$. Note that both $\hat{\rho}_p(n)$ and $\hat{\rho}_a(n)$ may take negative values according to the above definition. By using equations (8)-(10), our method decomposes observed signals into their periodic and aperiodic components.

The above decomposition method needs an F0 value, and so we estimate it as the value that maximizes the numerator of equation (10), that is, we determine the estimate $\hat{f}_0(n)$ by searching the frequency range that includes the F0 of human speech (e.g. from 50 to 500 Hz) as follows.

$$\hat{f}_0(n) = \underset{f_0(n)}{\arg\max}\left(\sum_{k=1}^{\nu(n)} |S(n,[kf_0(n)])|^2 - \nu(n)\rho(n)\right)$$

It is important to note that the above equation coincides with one adopted by a robust F0 estimator known as REPS [25], therefore, this equation is expected to provide us with fairly reliable F0 estimates even under adverse noise conditions. In addition, the advantage of utilizing this equation is that we can estimate the powers of the periodic and aperiodic components and the F0 simultaneously.

### 2.2. Likelihood calculation

If the decomposition described in section 2.1 can ideally estimate the powers of periodic components, we can detect speech signals based solely on these estimates. However, the decomposition cannot completely avoid power estimation errors. By taking the estimation errors into account, our proposed VAD method statistically detects the existence of speech signals based on the likelihood derived from the error distributions estimated for the periodic and aperiodic components.

We presume the state of the existence of speech signals (1 or 0) at frame $n$ to be random variables, and denote them by $H_n$. If $H_n$ = 1, then speech signals exist in the observed signals, and vice versa. When $H_n$ = 0, i.e. there is no speech signal in the observed signal, assuming $\rho(n) = \rho_a(n)$, we can estimate the error of an aperiodic component $\varepsilon_a(n)$ as

$$\varepsilon_a(n) = \rho(n) - \hat{\rho}_a(n) = \hat{\rho}_p(n)$$

We assume that the error distribution follows a Gaussian distribution whose mean and standard deviation are 0 and $\alpha\hat{\rho}_a(n)$ with a positive constant $\alpha$. Then, the likelihood of the observed signal for non-speech periods can be modeled by

$$\begin{aligned}p(\rho(n) \mid H_n = 0) &= c_1(n)\exp\left(-\frac{\varepsilon_a(n)^2}{2(\alpha\hat{\rho}_a(n))^2}\right) \\ &= c_1(n)\exp\left(-\frac{1}{2\alpha^2}\left(\frac{\hat{\rho}_p(n)}{\hat{\rho}_a(n)}\right)^2\right)\end{aligned} \tag{11}$$

In contrast, when $H_n$ = 1, i.e. a speech signal exists in the observed signal, we can estimate the error of the periodic component $\varepsilon_p(n)$ as

$$\varepsilon_p(n) = \rho(n) - \rho_a(n) - \hat{\rho}_p(n)$$

However, because we cannot know the true value of $\rho_a(n)$, it is difficult to determine the distribution of $\varepsilon_p(n)$. Thus, we consider an ideal case of $\rho_a(n) = 0$. Here, $\varepsilon_p(n)$ can be rewritten as

$$\varepsilon_p(n) = \rho(n) - \hat{\rho}_p(n) = \hat{\rho}_a(n)$$

Under this assumption, we again assume that the error distribution follows a Gaussian distribution whose mean and

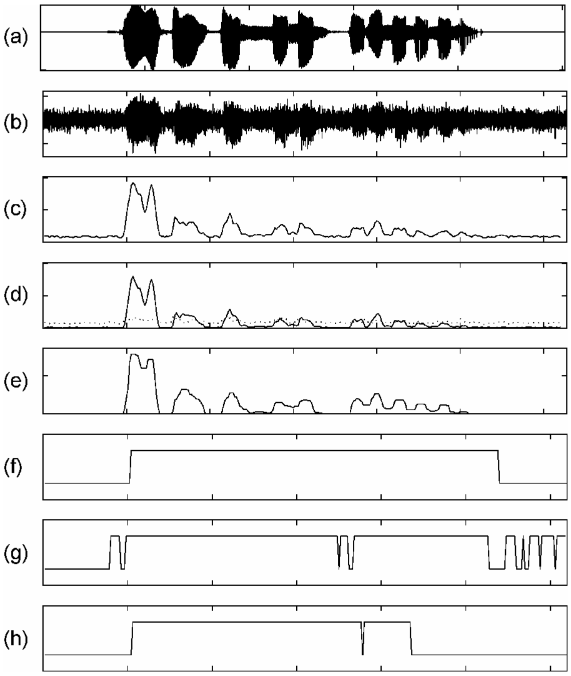

***Figure 1**: Behavior with the proposed method and a comparison of VAD results for a speech signal in the presence of white noise. (a) Speech waveform in silence. (b) Noisy speech waveform created by adding white noise to a speech (a) at an SNR of 0 dB. (c) Power of (b). (d) Decomposed periodic (solid line) and aperiodic (dashed line) components for (b). (e) Log likelihood ratios for VAD derived from the periodic and aperiodic components (d). (f) VAD result for (b) obtained with the proposed VAD method. (g) VAD result for (b) obtained with a statistical VAD method [6]. (h) VAD result for (b) obtained with ETSI AFE VAD [16].*

standard deviation are 0 and $\beta\hat{\rho}_p(n)$ with a positive constant $\beta$. Then, the likelihood of the observed signal for speech period can be modeled by

$$p(\rho(n)\,|\,H_n=1)=c_2(n)\exp\left(-\frac{\varepsilon_p(n)^2}{2(\beta\hat{\rho}_p(n))^2}\right)$$
$$=c_2(n)\exp\left(-\frac{1}{2\beta^2}\left(\frac{\hat{\rho}_a(n)}{\hat{\rho}_p(n)}\right)^2\right) \quad (12)$$

Although $\rho_a(n)$ must be larger than zero in practice (especially in the presence of noise), we introduce equation (12) as an approximation for all cases. It should be noted that this approximation provides an underestimation of the probability of the existence of speech. Therefore, if we can introduce a more adequate estimation for $p(\rho(n)\,|\,H_n=1)$, the VAD performance of the proposed method is expected to improve.

Finally, we calculate the likelihood ratio $\Lambda(n)$ of equations (11) and (12) at frame *n* as follows.

$$\Lambda(n)=\frac{p(\rho(n)\,|\,H_n=1)}{p(\rho(n)\,|\,H_n=0)} \quad (13)$$

If the likelihood is higher than a threshold decided apriori, our VAD algorithm decides that a speech signal exists in the frame. Because the threshold does not depend greatly on SNRs, we use a fixed value for the threshold. In addition, since this method does not use noise statistics, it is unnecessary to set a certain fixed period in the observed signal that contains only noise components for estimating the noise statistics.

## 3. Experiment

To examine the validity of the proposed method, we conducted preliminary experiments using simulated and real noisy speech data. We used the AURORA-2J database [26] for the speech data. The speech and noise were recorded with 16-bit quantization and at a sampling rate of 8 kHz. In this section, our proposed method always used $\alpha=\beta=1$ to calculate the likelihood. We also estimated the F0s by the method described in section 2.1 with trajectory smoothing based on dynamic programming. In addition, our proposed method applied the hangover methods proposed in [16] to the VAD results obtained from the likelihood ratios described in section 2.2.

### 3.1. Comparison of VAD performance in the presence of simulated noise

This section compares the behavior of the proposed and conventional VAD algorithms in the presence of white and amplitude modulated white noise. For this comparison, we used the speech data spoken by a male speaker shown in Fig. 1(a).

We examined the decomposition of the periodic/aperiodic components and the likelihood ratio for VAD calculated with equations (8)-(10), and (13), respectively. A test signal was created by adding white noise to the speech data shown in Fig. 1(a) at an SNR of 0 dB (Fig. 1(b)). The result is shown in Fig. 1(c)-(e). Even at such a low SNR, the results indicate that our method well decomposes periodic and aperiodic components. In addition, the likelihood ratios (Fig. 1(e)) correctly indicate the period in which speech signals exist. This result suggests the effectiveness of using this likelihood ratio for VAD.

The VAD result based on the likelihood ratios is shown in Fig. 1(f). For comparison, the VAD results obtained with Sohn's statistical VAD [6] and ETSI WI008 Advanced Front End (AFE) VAD for frame dropping [16] are also shown in Fig. 1(g) and (h). The former utilizes likelihoods derived from posteriori SNRs calculated from the spectral amplitude estimated by the minimum mean square error approach, and the latter simultaneously utilizes the whole spectra to design Wiener filters, the spectral sub-region, and spectral variance. Both conventional methods include the hangover mechanisms. The comparison indicates that our proposed method performs better than conventional methods in the presence of white noise.

We then compared the VAD performance using another test signal (Fig. 2(b)) created by adding amplitude modulated white noise to the speech data shown in Fig. 2(a) in order to test the robustness of the present method as regards the non-stationarity of noise. The amplitude modulation rate was 4 Hz, and its modulation depth was 0.4. Figure 2(c), (e)-(g), and (i) show the features calculated by the two conventional methods and our proposed method. Although the conventional features were deteriorated by the amplitude modulation of the noise, the results indicate that the likelihood ratios of the proposed method are insensitive to such amplitude modulation. Figure 2 (d), (h), and (j) compare the VAD results obtained with the three VAD methods. The results suggest that the proposed method is also robust as regards non-stationarity of noise.

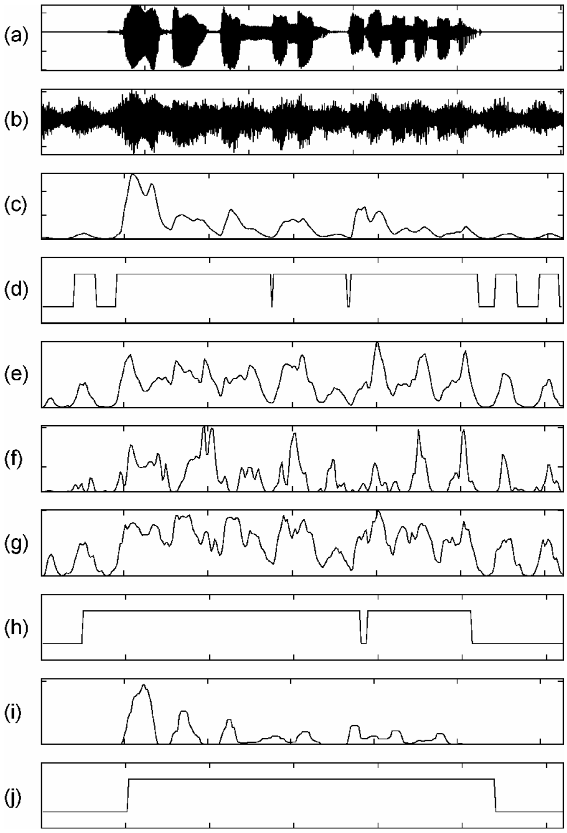

*Figure 2: VAD features for a speech signal in the presence of amplitude modulated white noise and a comparison of VAD results for the speech signal. (a) Speech signal in silence. (b) Speech signal (a) mixed with amplitude modulated white noise at an SNR of 0 dB. (c) Feature (log likelihood ratios) for (b) obtained with a statistical VAD [6]. (d) VAD result for (b) obtained with a statistical VAD method [6]. (e)-(g) Features (whole spectrum, spectral sub-region, spectral variance) for (b) obtained with ETSI AFE VAD [16]. (h) VAD result for (b) obtained with ETSI AFE VAD [16]. (i) Feature (log likelihood ratios) for (b) obtained with the proposed method. (j) VAD result for (b) obtained with the proposed VAD method.*

### 3.2. Comparison of VAD performance in the presence of noise in the real world

Finally we compared the VAD performance obtained from the proposed and conventional methods in the presence of noise in the real world. For this comparison, we used noisy speech data in AURORA-2J (Fig. 3(b)) created by adding subway noise to clean speech data at an SNR of 0 dB. The subway noise includes non-stationary sounds such as wheel sounds of trains.

Figure 3(c), (e)-(g), and (i) show the VAD features calculated by the two conventional methods and our proposed method. The results indicate that there was deterioration in the conventional features that was largely due to the amplitude variation of the noise and non-stationary sounds appearing in the first part of the test data (indicated as circles in Fig. 3(b)). On the other hand, our proposed likelihood ratios of the periodic and aperiodic components could moderate the influence of the non-stationary sounds and amplitude variation of the noise, and deteriorated less than conventional features. Figure 3(d), (h), and (j) confirm the effectiveness of the proposed VAD method compared with the two conventional VAD methods.

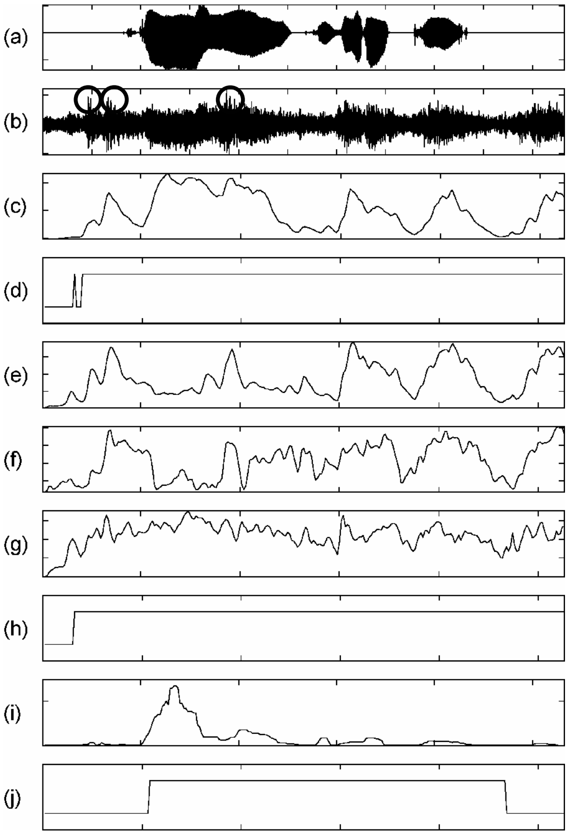

*Figure 3: VAD features for a speech signal in the presence of subway noise and a comparison of VAD results for the noisy speech signal. (a) Speech signal in silence. (b) Speech signal (a) mixed with subway noise at an SNR of 0 dB. Circles indicate non-stationary noise. (c) Feature (log likelihood ratios) for (b) obtained with a statistical VAD [6]. (d) VAD result for (b) obtained with a statistical VAD method [6]. (e)-(g) Features (whole spectrum, spectral sub-region, spectral variance) for (b) obtained with ETSI AFE VAD [16]. (h) VAD result for (b) obtained with ETSI AFE VAD [16]. (i) Feature for (b) obtained with the proposed method. (j) VAD result for (b) obtained with the proposed VAD method.*

In addition, to show the effectiveness of our proposed feature as regards temporal changes in SNRs, we compared the VAD performance using two concatenated noisy speech data whose SNRs differ from each other in AURORA-2J. A test signal (Fig. 4(b)) was created by concatenating noisy speech data in the presence of subway noise at an SNR of 5 dB and the noisy speech data shown in Fig. 3(b). Figure 4(c)-(e) compares the VAD results for the three VAD methods. Since it is difficult to update noise statistics or thresholds correctly when SNRs change within a short time, the conventional methods could not perform well. On the other hand, our proposed method performed well regardless of the temporal change in the SNRs. The above results suggest that the proposed method

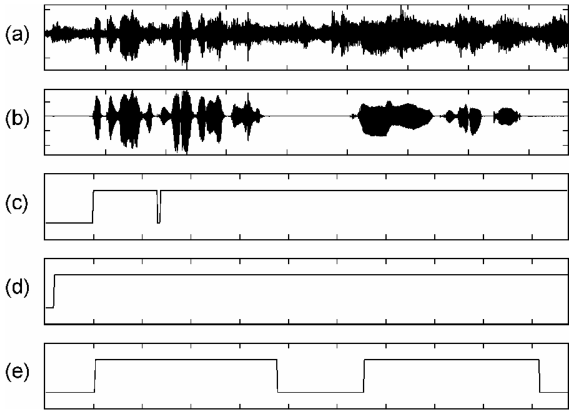

***Figure 4**: Comparison of VAD results for a speech signal in the presence of subway noise when the SNR changes temporally from 5 to 0 dB. The SNR of the first half is 5 dB, and that of the latter half is 0 dB. (a) Speech waveform mixed with subway noise. (b) Speech signal included in (a). (c) VAD result for (a) obtained with a statistical VAD method [6]. (d) VAD result for (a) obtained with ETSI AFE VAD [16]. (e) VAD result for (a) obtained with the proposed VAD method.*

can detect voice activity robustly even in the presence of noise in the real world. The robustness as regards the non-stationarity of noise is a particular advantage of the proposed method in terms of practical use.

## 4. Conclusion

We proposed a noise robust VAD method based on the ratios of the periodic and aperiodic components of observed signals. By utilizing this feature representation that is insensitive to variation in the SNR, the proposed method performs well in the presence of non-stationary noise. We conducted preliminary experiments to evaluate the advantages of the proposed method in terms of VAD performance. These experiments confirmed that the proposed method performed better than conventional VAD methods particularly in the presence of non-stationary noise. In the future we will undertake a more adequate probability estimation of speech existence for VAD features and evaluation experiments using large-scale data.

**Acknowledgement:** The authors thank Noboru Miyazaki (NTT Cyber Space Labs.) for suggesting the basic idea of this work. The authors also thank members of the Signal Processing Research Group at NTT Communication Science Labs. for their regular useful comments about VAD. This research employed the noisy speech recognition evaluation environment AURORA-2J produced and distributed by the IPSJ SIG-SLP Noisy Speech Recognition Working Group.

## References

[1] Le Bouquin-Jeannès R. and Faucon, G. “Study of voice activity detector and its influence on a noise reduction system,” Speech Communication, **16**, 245-254, 1995.

[2] Srinivasan, K. and Gersho, A. “Voice activity detection for cellular networks,” *Proc. of IEEE Workshop on Speech Coding for Telecommunications*, 85-86, 1993.

[3] Junqua, J.-C., Mak, B., and Reaves, B. “A robust algorithm for word boundary detection in the presence of noise,” IEEE Trans. Speech Audio Process., **2**, 406-412, 1994.

[4] Karray, L. and Martin, A. “Towards improving speech detection robustness for speech recognition in adverse conditions,” Speech Communication, **40**, 261-276, 2003.

[5] Marzinzik, M. and Kollmeier, B. “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Trans. Speech Audio Process., **10**, 109-118, 2002.

[6] Sohn, J., Kim, N.-S., and Sung, W. “A statistical model-based voice activity detection,” IEEE Signal Process. Lett., **6**, 1-3, 1999.

[7] Basu, S. “A linked-HMM model for robust voicing and speech detection,” *Proc. ICASSP*, **1**, 816-819, 2003.

[8] Rabiner, L. R. and Sambur, M. R. “An algorithm for determining the endpoints of isolated utterances,” The Bell Syst. Tech. Journal, **54**, 297-315, 1975.

[9] Atal, B. S. and Rabiner, L. R. “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,” IEEE Trans. Acoust., Speech, and Signal Process., **ASSP-24**, 201-212, 1976.

[10] Kingsbury, B., Saon, G., Mangu, L., Padmanabhan, M., and Sarikaya, R. “Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system,” *Proc. ICASSP*, **1**, 53-56, 2002.

[11] Kristjansson, T., Deligne, S., and Olsen, P. “Voicing features for robust speech detection,” *Proc. Interspeech*, 369-372, 2005.

[12] Shen, J.-L., Hung, J.-W., and Lee, L.-S. “Robust entropy-based endpoint detection for speech recognition in noisy environments,” *Proc. ICSLP*, 1998.

[13] Yantorno, R. E., Krishnamachari, K. L., and Lovekin, J. M. “The spectral autocorrelation peak valley ratio (SAPVR) – A usable speech measure employed as a co-channel detection system,” *Proc. IEEE Int. Workshop Intell. Signal Process.*, 2001.

[14] Tucker, R. “Voice activity detection using a periodicity measure,” IEE Proceedings-I, **139**, 377-380, 1992.

[15] ITU-T Recommendation G.729 Annex B., 1996.

[16] ETSI standard document, ETSI ES 202 050 V1.1.3., 2003.

[17] Li, K., Swamy, N. S., and Ahmad, M. O. “An improved voice activity detection using higher order statistics,” IEEE Trans. Speech Audio Process., **13**, 965-974, 2005.

[18] Lee, A., Nakamura, K., Nisimura, R., Saruwatari, H., and Shikano K. “Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs,” *Proc. Interspeech*, **1**, 173-176, 2004.

[19] de Cheveigné, A. “Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing,” J. Acoust. Soc. Am., **93**, 3271-3290, 1993.

[20] Serra, X. and Smith, J. “Spectral modeling and synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition,” Comp. Music J., **14**, 1990.

[21] Yegnanarayana, B., d'Alessandro, C., and Darsinos, V. “An iterative algorithm for decomposition of speech signals into periodic and aperiodic components,” IEEE Trans. Speech Audio Process., **6**, 1-11, 1998.

[22] Richard, G. and d'Alessandro, C. “Modification of the aperiodic component of speech signals for synthesis,” in *Progress in Speech Synthesis*, van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J. Eds. New-York: Springer-Verlag, 41-56, 1996.

[23] Jackson, P. J. B., Moreno, D. M., Russell, M. J. and Hernando, J. “Covariation and weighting of harmonically decomposed streams for ASR,” *Proc. Interspeech*, 2321-2324, 2003.

[24] Ishizuka, K., Nakatani, T., Minami, Y., and Miyazaki, N. “Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition,” J. Acoust. Soc. Am., **120**, 443-452, 2006.

[25] Nakatani, T. and Irino, T., “Robust and accurate fundamental frequency estimation based on dominant harmonic components,” J. Acoust. Soc. Am., **116**, 3690-3700, 2004.

[26] Nakamura, S., Takeda, K., Yamamoto, K., Yamada, T., Kuroiwa, S., Kitaoka, N., Nishiura, T., Sasou, A., Mizumachi, M., Miyajima, C., Fujimoto, M., and Endo, T. “AURORA-2J: An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. on Inf. & Syst., **E88-D**, 535-544, 2005.