Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 41–60
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Noise robust voice activity detection based on periodic to aperiodic component ratio ☆

Kentaro Ishizuka [a,*], Tomohiro Nakatani [a], Masakiyo Fujimoto [a], Noboru Miyazaki [b]

[a] *NTT Communication Science Laboratories, NTT Corporation, Hikaridai 2-4, Seikacho, Sourakugun, Kyoto 619-0237, Japan*
[b] *NTT Cyber Space Laboratories, NTT Corporation, Hikarino-oka 1-1, Yokosuka City, Kanagawa 239-0847, Japan*

Received 28 May 2008; received in revised form 8 July 2009; accepted 20 August 2009

Abstract

This paper proposes a noise robust voice activity detection (VAD) technique called PARADE (PAR based Activity DEtection) that employs the periodic component to aperiodic component ratio (PAR). Conventional noise robust features for VAD are still sensitive to non-stationary noise, which yields variations in the signal-to-noise ratio, and sometimes requires *a priori* noise power estimations, although the characteristics of environmental noise change dynamically in the real world. To overcome this problem, we adopt the PAR, which is insensitive to both stationary and non-stationary noise, as an acoustic feature for VAD. By considering both periodic and aperiodic components simultaneously in the PAR, we can mitigate the effect of the non-stationarity of noise. PARADE first estimates the fundamental frequencies of the dominant periodic components of the observed signals, decomposes the power of the observed signals into the powers of its periodic and aperiodic components by taking account of the power of the aperiodic components at the frequencies where the periodic components exist, and calculates the PAR based on the decomposed powers. Then it detects the presence of target speech signals by estimating the voice activity likelihood defined in relation to the PAR. Comparisons of the VAD performance for noisy speech data confirmed that PARADE outperforms the conventional VAD algorithms even in the presence of non-stationary noise. In addition, PARADE is applied to a front-end processing technique for automatic speech recognition (ASR) that employs a robust feature extraction method called SPADE (Subband based Periodicity and Aperiodicity DEcomposition) as an application of PARADE. Comparisons of the ASR performance for noisy speech show that the SPADE front-end combined with PARADE achieves significantly higher word accuracies than those achieved by MFCC (Mel-frequency Cepstral Coefficient) based feature extraction, which is widely used for conventional ASR systems, the SPADE front-end without PARADE, and other standard noise robust front-end processing techniques (ETSI ES 202 050 and ETSI ES 202 212). This result confirmed that PARADE can improve the performance of front-end processing for ASR.

Keywords: Voice activity detection; Robustness; Periodicity; Aperiodicity; Noise robust front-end processing for automatic speech recognition

1. Introduction

Voice activity detection (VAD), which is a method for detecting periods of speech in observed signals, plays a crucial role in speech signal processing techniques. VAD in the "real world" such as in a car, or on the street, is particularly important in relation to speech enhancement techniques for estimating noise statistics from the non-speech periods (Le Bouquin-Jeannès and Faucon, 1995), speech coding techniques that allow a variable bit-rate for

☆ Portions of this work were presented in "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, Pittsburgh, USA, September 2006, and "Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio," Proceedings of the 10th European Conference on Speech Communication and Technology, Antwerp, Belgium, August 2007.

* Corresponding author. Tel.: +81 774 93 5216; fax: +81 774 93 1945.
E-mail addresses: ishizuka@cslab.kecl.ntt.co.jp (K. Ishizuka), nak@cslab.kecl.ntt.co.jp (T. Nakatani), masakiyo@cslab.kecl.ntt.co.jp (M. Fujimoto), miyazaki.noboru@lab.ntt.co.jp (N. Miyazaki).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.08.003

discontinuous transmission (Srinivasan and Gersho, 1993; ITU-T Recommendation G.729 Annex B, 1996; ETSI TS 101 707, 2000), and automatic speech recognition (ASR) techniques designed to prevent false alarms that occur when noise signals are recognized as speech signals (Junqua et al., 1994). Since these techniques depend strongly on VAD accuracy or sometimes assume ideal VAD, insufficient VAD accuracy seriously affects their practical performance. This fact has made it necessary to develop more robust VAD for real world applications (Lamel et al., 1981; Savoji, 1989; Karray and Martin, 2003).

In general, VAD consists of two parts: ‘acoustic feature extraction’, and a ‘speech/non-speech decision mechanism’. The former part extracts the acoustic features that can appropriately represent the probability of the existence of target speech signals in observed signals. Employing these acoustic features as a basis, the latter part finally decides whether the target speech signals are present in the observed signals using, for example, a well-adjusted threshold that reflects the characteristics of speech (Savoji, 1989; Marzinzik and Kollmeier, 2002), likelihood ratio tests (Sohn et al., 1999; Cho and Kondoz, 2001; Ramírez et al., 2005; Chang et al., 2006; Davis et al., 2006; Górriz et al., 2006; Ramírez et al., 2007; Fujimoto and Ishizuka, 2008), and hidden Markov models (HMMs) (Wilpon and Rabiner, 1987; Basu, 2003). Although the performance of both parts has a strong influence on the VAD performance, this paper focuses particularly on the noise robust acoustic feature for VAD.

The short-term signal energy and zero-crossing rate (Rabiner and Sambur, 1975) have long been used as simple acoustic features for VAD. Although these features are indeed effective under high signal-to-noise ratio (SNR) conditions, they are easily degraded by environmental noise. In addition, environmental noise sometimes possesses an energy and zero-crossing rate similar to those of speech signals. To cope with this problem, various kinds of robust acoustic features have been proposed for VAD as regards noise. Some speech characteristic based features are related to the fact that the distribution of the speech signal spectra is biased to the low frequency region. Such features are composed of the power in the band-limited region or the sum of the outputs from band-pass filter banks (Mak et al., 1992; ITU-T Recommendation G.729 Annex B, 1996; ETSI ES 202 050, 2001; Marzinzik and Kollmeier, 2002). Spectral shapes such as Mel-frequency Cepstral Coefficients (MFCCs) (Kristjansson et al., 2005), long-term spectral envelopes (LTSE) (Ramírez et al., 2004), and delta line spectral frequencies (LSF) (ITU-T Recommendation G.729 Annex B, 1996) have also been used to distinguish speech signals from other sounds. In addition to the speech spectral characteristics model, some methods consider noise spectral characteristics (Lee et al., 2004; Kristjansson et al., 2005; de la Torre et al., 2006; Fujimoto and Ishizuka, 2008), or utilize enhanced speech spectra derived from a minimum mean-square error estimation (MMSE) (Ephraim and Malah, 1984) or Wiener filtering based estimations (Agarwal and Cheng, 1999) of noise statistics (Sohn et al., 1999; ETSI ES 202 050, 2001). To estimate more precise parameters for these features, Davis et al. (2006) employed the Welch–Barlett method to achieve a lower variance spectral estimation than that employed in (Sohn et al., 1999), and Evangelopoulos and Maragos (2006) introduced multiband Teager energy to the LTSE framework (Ramírez et al., 2004). In terms of the utilization of speech signal statistics, some VAD methods have been proposed that are based on higher-order statistics (Nemer et al., 2001; Li et al., 2005; Cournapeau and Kawahara, 2007) and volatility (conditional variance) obtained from GARCH filtering (Tahmasbi and Razaei, 2007; Kato Solvang et al., 2008). In addition, some features have been proposed that are based on the periodicity of speech signals because it is not easily contaminated by background sounds. These features include autocorrelation function (ACF) based features (Rabiner and Sambur, 1975; Atal and Rabiner, 1976; Kingsbury et al., 2002; Basu, 2003; Kristjansson et al., 2005), spectrum based features that utilize harmonicity (Shen et al., 1998; Yantorno et al., 2001; Wu and Wang, 2005), and fundamental frequency (F0) based features (Hamada et al., 1990; Tucker, 1992; Ahmadi and Spanias, 1999).

Although most of the above methods, particularly the periodic feature based methods, are indeed robust as regards environmental noise, they are still sensitive to non-stationary noise, whose power and characteristics change with time. Because environmental noise is usually non-stationary and in practice its power changes greatly very quickly, the conventional acoustic features are affected by this non-stationarity, and consequently deciding the optimum threshold is difficult with VAD. Even with periodic feature based methods, it is difficult to estimate the periodic features of speech signals precisely in the presence of non-stationary noise because the features change corresponding to changes in noise power. This fact degrades the usability of VAD methods in the real world. Therefore, there is a need for VAD features and algorithms that are insensitive to non-stationary noise.

Let us now turn to the sound representation of observed sound signals. Sound signals in the real world usually have both periodic and aperiodic components, and can be decomposed into these two components. For example, speech signals consist not only of periodic components such as the steady parts of vowels and voiced consonants, but also of aperiodic components such as fluctuations that are intrinsic to vowels, voiced consonants, stops, fricatives, and affricates. Such a twofold representation of sounds has long been studied in terms of speech/music analysis/synthesis (e.g. Griffin and Lim, 1988; Serra and Smith, 1990; Boersma, 1993; Laroche et al., 1993; Yegnanarayana et al., 1998; Jackson and Shadle, 2001; Deshmukh et al., 2005), because the aperiodic component is responsible for the perceived synthesized speech/music quality (Krom, 1993; Richard and d’Alessandro, 1996). In addition, in terms of ASR, the word accuracy in noisy environments can be improved by using the decomposed periodic and aperiodic features of observed signals (Jackson et al., 2003; Ishizuka et al., 2006). The above indicates the efficiency of such a rich representation of sound signals. However, numerous VAD methods focusing only on the periodic characteristics of speech signals have long been studied as described above, whereas there has been little work on either the periodic or aperiodic component explicitly for VAD.

In this paper, we propose a feature for VAD that is insensitive to the non-stationarity of noise based on the representation of sound with periodic and aperiodic components. This method first decomposes the observed signals into their periodic and aperiodic components, and then utilizes the ratio of the powers of the periodic component to the aperiodic component (the periodic to aperiodic component ratio; PAR) as an acoustic feature for VAD. When we consider the implication of the PAR, it should be noted that the PAR of an observed signal is independent of its power. In terms of these characteristics, even if the environmental noise power changes dynamically, its PAR is not expected to change insofar as its characteristics do not change. In addition, because the power of voiced speech signals is concentrated in their harmonic components, voiced speech signals exhibit higher PARs than those of environmental noise including non-stationary noise, e.g. clicks and bursts, whose frequency spectra are distributed widely over all frequencies unlike those of speech signals. Therefore, the PAR is expected to be an acoustic feature for VAD that is insensitive to non-stationary noise. Henceforth, we refer to the proposed method as PARADE (PAR based Activity DEtection).

The PAR is a similar measure to the harmonics-to-noise ratio (HNR) (Boersma, 1993), which has been widely used in the field of speech/music analysis for describing the quality or characteristics of speech/music signals (Hillenbrand, 1987; Krom, 1993) and for deciding voiced/unvoiced parts of speech signals (Mousset et al., 1996; Ahmadi and Spanias, 1999; Fisher et al., 2006; Nakatani et al., 2008). However, in this paper, we use the term ‘PAR’ rather than ‘HNR’ to make it clear that we calculate the PAR for all kinds of sounds to utilize the measure for VAD, whereas HNR is basically employed as a measure of the quality (such as hoarseness) of speech/music signals.

In the literature on speech signal analysis, Boersma (1993) has already argued that the HNR derived from ACF can be used for detecting speech in the presence of additive noise. The HNR is indeed useful for VAD in the presence of additive white Gaussian noise because the effect of the noise can be canceled out in the ACF, however, the calculation includes estimation errors in the presence of colored noise such as environmental noise in the real world. This is because the HNR calculation is simply based on the ACF, and does not consider the power of the aperiodic component at the frequencies where the periodic component exists, and this results in errors when estimating the powers of the periodic and aperiodic components. Therefore, a more precise estimation of these components is needed if we are to utilize the periodic to aperiodic component power ratio as an acoustic feature for VAD.

PARADE can estimate the powers of these components more precisely by taking account of the aperiodic component power at the frequencies where the periodic component exists, and employing more relaxed restrictions as regards the frequency distributions of aperiodic components than those used for the HNR calculation. In particular, this power estimation is based on the following two assumptions. (1) The power of the periodic component of the observed signal can be calculated as a sum of the powers of sinusoids that correspond to the F0 multiples of the observed signal. (2) The average power of the aperiodic components at the frequencies of the periodic components is equal to that over the whole frequency range. In addition, PARADE provides a speech/non-speech decision mechanism by considering the estimation errors of the periodic/aperiodic component powers.

This paper is organized as follows. Section 2 provides detailed explanations of the PARADE algorithm. Section 3 describes VAD evaluation experiments designed to show the advantage of PARADE by comparing VAD performance with that achieved with conventional acoustic features. Section 4 describes an application of PARADE to noise robust front-end processing for ASR, and ASR evaluation experiments show the effectiveness of the application. Section 5 concludes this study and mentions future work.

2. Method

This section provides detailed explanations of the PARADE algorithm. Fig. 1 shows a block diagram of PARADE. Let us first define the sound conditions that

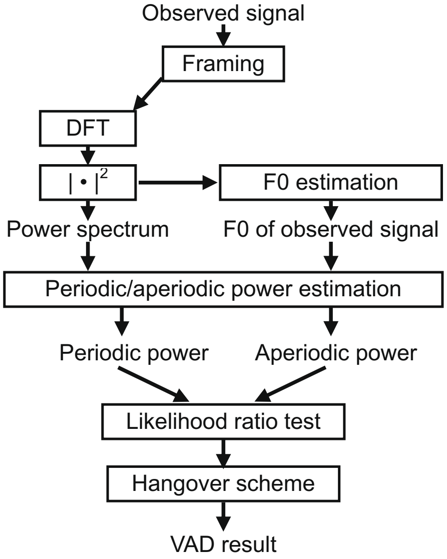

Fig. 1. Block diagram of the proposed voice activity detection method (PARADE).

PARADE is designed to deal with. Observed signals are recorded monaurally, and there is only one target sound, i.e. the target speech, which exists in the presence of background noise whose frequency spectra are widely distributed over all frequencies. There is no assumption regarding the stationarity of the noise power, thus the noise power may change dynamically, and many kinds of environmental sounds may be included in the background noise. In addition, there is no *a priori* knowledge about the kind of background noise or target speech signals.

PARADE copes with the above sound scenes as follows. PARADE first decomposes the periodic/aperiodic components included in all the observed signals to determine the periodic component of the target signals, namely speech signals, whereas the conventional decomposition methods (e.g. Serra and Smith, 1990; Yegnanarayana et al., 1998) aim to decompose the components inherent in a single sound source such as speech/music signals. Therefore, in this paper, the term 'aperiodic component' includes both environmental noise and the aperiodic components of speech signals, and the term 'periodic component' includes a dominant harmonic component in the observed signal. Section 2.1 describes the decomposition method employed in the frequency domain.

PARADE 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. Statistical VAD methods, such as that proposed by Sohn et al. (1999), utilize likelihoods derived from *a priori* SNRs, whereas PARADE calculates likelihoods derived from the error distributions of the powers of periodic to aperiodic components without any knowledge of the SNR. In addition, PARADE does not utilize speech and/or noise characteristics obtained from training data *a priori*, unlike other statistical VAD approaches (Basu, 2003; Lee et al., 2004; de la Torre et al., 2006; Fujimoto and Ishizuka, 2008). Section 2.2 describes the likelihood calculation in detail.

It should be noted that PAR inherently cannot handle aperiodic speech components, namely unvoiced speech, thus a mechanism called a 'hangover scheme' is needed to handle unvoiced (aperiodic) speech components that are close to voiced (periodic) speech components. This hangover scheme has been widely adopted in conventional VAD techniques to prevent the appearance of voiced speech fragments and to avoid short pauses in speech periods. Section 2.3 discusses the requirements of the hangover scheme and describes the procedure in detail.

2.1. Decomposition of powers of periodic and aperiodic components

In this section, we explain how to estimate the power of periodic (harmonic) and aperiodic components from an observed signal. We assume that an observed signal $x(n)$ can be described as the sum of its periodic and aperiodic components, $x_p(n)$ and $x_a(n)$, respectively

$$x(n) = x_p(n) + x_a(n), \tag{1}$$

where n is the sampling index of the observed signal. Henceforth, we denote short time Fourier transform (STFT) representations of the above signals at the kth frequency bin analyzed by a temporal frame, whose index and temporal length are i and L, by $X(i,k)$, $X_p(i,k)$, and $X_a(i,k)$, respectively. And, we denote their short temporal powers in the ith frame by $\rho(i)$, $\rho_p(i)$, and $\rho_a(i)$, respectively.

The short temporal power $\rho(i)$ of a signal $x(n)$ within a frame is defined as:

$$\rho(i) = \sum_{n=0}^{L-1} (g(n)x(n))^2, \tag{2}$$

where $g(n)$ is an analyzing window such as a Hanning window, and it can be considered the 0th-order autocorrelation coefficient. Since the autocorrelation coefficients are equal to the inverse Fourier transform of the power spectrum $|X(i,k)|^2$ of $x(n)$, the short time power of the signal can also be obtained as

$$\rho(i) = \frac{1}{K} \sum_{k=0}^{K-1} |X(i,k)|^2, \tag{3}$$

where K is the maximum number of frequency bins.

Now, we assume the additivity on the powers of the components in the frequency domain as

$$|X(i,k)|^2 = |X_p(i,k)|^2 + |X_a(i,k)|^2. \tag{4}$$

Then the following equation can also be derived from Eqs. (3) and (4)

$$\rho(i) = \rho_p(i) + \rho_a(i). \tag{5}$$

We denote the F0 of the periodic component and the number of harmonics at the ith frame by $f_0(i)$ and $v(i)$, respectively, and an operator for transforming the mth harmonic frequency $mf_0(i)$ to the index of a frequency bin in the corresponding discrete Fourier representation by $[mf_0(i)]$. Then, we assume that the power of the periodic component of the observed signal can be calculated as the sum of harmonic component sinusoids whose F0 is $f_0(i)$ as follows

$$\textbf{Approximation I}: \quad \begin{aligned} \rho_p(i) &= \sum_{m=1}^{v(i)} \rho_{p,m}(i), \\ \rho_{p,m}(i) &= \eta |X_p(i, [mf_0(i)])|^2, \end{aligned} \tag{6}$$

where $\rho_{p,m}(i)$ is the power of a sinusoid corresponding to the mth harmonic component, η is a constant for estimating the power of the sinusoid from its STFT representation (see Appendix A, where it is defined as $\eta = (2\sum_{n=0}^{L-1} g(n)^2)/(\sum_{n=0}^{L-1} g(n))^2)$. By using Eqs. (4) and (6), we can obtain the following equation

$$\begin{aligned}\rho_p(i) &= \sum_{m=1}^{v(i)} (\eta |X_p(i,[mf_0(i)])|^2) \\ &= \eta\left(\sum_{m=1}^{v(i)} |X(i,[mf_0(i)])|^2 - \sum_{m=1}^{v(i)} |X_a(i,[mf_0(i)])|^2\right).\end{aligned} \tag{7}$$

It should be noted that Eq. (7) can be viewed as a kind of comb filtering with a sharp cutoff frequency that can be utilized for extracting the harmonic components of the observed signals.

In addition to Eq. (6), we introduce the following assumption

$$\textbf{Approximation II}: \quad \frac{1}{K}\sum_{k=0}^{K-1} |X_a(i,k)|^2 = \frac{1}{v(i)}\sum_{m=1}^{v(i)} |X_a(i,[mf_0(i)])|^2, \tag{8}$$

which means that the average power of the aperiodic components at the frequencies of the dominant harmonic components is equal to that over the entire frequency range. This represents the power of the aperiodic components which is widely distributed over the entire frequency range in a manner that is independent of the frequencies of the periodic components. It should be noted that this assumption only provides the equality of the average powers, and the assumption does not require any specific distribution of the spectra of the aperiodic components in the frequency region. Here, we can assume the following equation as Eq. (3)

$$\rho_a(i) = \frac{1}{K}\sum_{k=0}^{K-1} |X_a(i,k)|^2. \tag{9}$$

According to Eqs. (8) and (9), we can obtain the following equation

$$\begin{aligned}\sum_{m=1}^{v(i)} |X_a(i,[mf_0(i)])|^2 &= v(i)\left(\frac{1}{v(i)}\sum_{m=1}^{v(i)} |X_a(i,[mf_0(i)])|^2\right) \\ &= v(i)\left(\frac{1}{K}\sum_{k=0}^{K-1} |X_a(i,k)|^2\right) \\ &= v(i)\rho_a(i).\end{aligned} \tag{10}$$

Substituting Eqs. (7) and (10) into Eq. (5), we derive the following equation

$$\begin{aligned}\rho(i) &= \eta\left(\sum_{m=1}^{v(i)} |X(i,[mf_0(i)])|^2 - v(i)\rho_a(i)\right) + \rho_a(i) \\ &= \eta\sum_{m=1}^{v(n)} |X(i,[mf_0(i)])|^2 + (1-\eta v(i))\rho_a(i).\end{aligned} \tag{11}$$

Consequently, we can obtain the following equations for the power decomposition

$$\hat{\rho}_a(i) = \frac{\rho(i) - \eta\sum_{m=1}^{v(i)} |X(i,[mf_0(i)])|^2}{1-\eta v(i)}, \tag{12}$$

$$\hat{\rho}_p(i) = \rho(i) - \hat{\rho}_a(i) = \eta\frac{\sum_{m=1}^{v(i)} |X(i,[mf_0(i)])|^2 - v(i)\rho(i)}{1-\eta v(i)}, \tag{13}$$

where $\hat{\rho}_p(i)$ and $\hat{\rho}_a(i)$ indicate estimations of the true values of $\rho_p(i)$ and $\rho_a(i)$. Eq. (12) can be obtained by transforming Eq. (11) for $\rho_a(i)$, and Eq. (13) can be obtained by using Eqs. (3), (5) and (12). It should be noted that both $\hat{\rho}_p(i)$ and $\hat{\rho}_a(i)$ may take negative values according to the above definition. In practice, such negative values have a floor as described in Appendix B. By using Eqs. (3), (12), and (13), PARADE decomposes observed signals into the powers of their periodic and aperiodic components.

The above decomposition requires $f_0(i)$, and so we estimate it by the autocorrelation method widely used for estimating F0 (Rabiner, 1977; Hess, 1983), that is, we estimate $\hat{f}_0(i)$ by searching for the value that maximizes the ACF of $x(n)$ within the frequency range that includes the F0s of speech (e.g. 50–500 Hz) as follows

$$\hat{f}_0(i) = f_s/\tau_{\max}, \quad \tau_{\max} = \arg\max_{\tau} E\{x(n)x(n+\tau)\}, \tag{14}$$

where τ is a lag for the ACF, and f_s is the sampling rate in Hz. ACF can be calculated by the inverse Fourier transformation of the power spectrum of the observed signal.

In place of the estimation based on Eq. (14), we can estimate F0 as the value that maximizes the numerator of Eq. (13), namely we determine the estimate $\hat{f}_0(i)$ by searching the frequency range that includes the F0 of human speech as follows

$$\hat{f}_0(i) = \arg\max_{f_0(i)}\left(\sum_{m=1}^{v(i)} |X(i,[mf_0(i)])|^2 - v(i)\rho(i)\right). \tag{15}$$

It should be noted that Eq. (15) is related to one adopted by a robust F0 estimator known as REPS (Ripple-Enhanced Power Spectrum) (Nakatani and Irino, 2004). Although Eq. (15) is a simpler implementation than the full implementation of REPS, the advantage of utilizing Eq. (15) is that we can estimate the powers of the periodic and aperiodic components and the F0 simultaneously in the PARADE framework. The VAD performance obtained with the two F0 estimation methods is compared in Section 3.

2.2. Likelihood calculation for VAD decision

If the power 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 because assumptions (6) and (8) or F0s estimated by Eqs. (14) or (15) contain certain errors. Therefore, 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.

The likelihood of a speech segment is defined as a joint probability density function of the segmental signal power, $\rho(i)$, and the decomposed powers, $\hat{\rho}_a(i)$ and $\hat{\rho}_p(i)$, given the state of the segment H_i (=1 or 0) at frame i. If $H_i = 1$, then speech signals are present in the observed signals, and *vice versa*. The likelihood function is mathematically defined as:

$$\begin{aligned}\mathscr{L}(H_i) &= p(\rho(i), \hat{\rho}_a(i), \hat{\rho}_p(i)|H_i) \\ &= p(\hat{\rho}_a(i), \hat{\rho}_p(i)|H_i, \rho(i))p(\rho(i)|H_i). \end{aligned} \tag{16}$$

Henceforth, $p(\rho(i)|H_i)$ in Eq. (16) is disregarded as a constant term.

When $H_i = 0$, i.e. there is no speech signal in the observed signal, assuming $\rho(i) = \rho_a(i)$, we can estimate the error of an aperiodic component $\varepsilon_a(i)$ as

$$\varepsilon_a(i) = \rho(i) - \hat{\rho}_a(i) = \hat{\rho}_p(i). \tag{17}$$

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

$$\begin{aligned}p(\hat{\rho}_a(i), \hat{\rho}_p(i)|H_i, \rho(i)) &\approx p(\varepsilon_a(i)|H_i = 0, \rho(i) = \rho_a(i)) \\ &= \frac{1}{\sqrt{2\pi}\alpha\rho_a(i)} \exp\left(-\frac{\varepsilon_a(i)^2}{2(\alpha\rho_a(i))^2}\right) \\ &\approx \frac{1}{\sqrt{2\pi}\alpha\hat{\rho}_a(i)} \exp\left(-\frac{1}{2\alpha^2}\left(\frac{\hat{\rho}_p(i)}{\hat{\rho}_a(i)}\right)^2\right).\end{aligned} \tag{18}$$

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

$$\varepsilon_p(i) = \rho(i) - \rho_a(i) - \hat{\rho}_p(i). \tag{19}$$

However, because we cannot know the true value of $\rho_a(i)$, it is difficult to determine the distribution of $\varepsilon_p(i)$. Thus, we consider an ideal case of $\rho_a(i) = 0$, then Eq. (19) can be rewritten as

$$\varepsilon_p(i) = \rho(i) - \hat{\rho}_p(i) = \hat{\rho}_a(i). \tag{20}$$

Under this assumption, we again assume that the error distribution follows a Gaussian distribution whose mean and standard deviation are 0 and $\beta\rho_p(i) \approx \beta\hat{\rho}_p(i)$ with a positive constant β. Then, the likelihood of the observed signal for a speech period can be modeled by

$$\begin{aligned}p(\hat{\rho}_a(i), \hat{\rho}_p(i)|H_i, \rho(i)) &\approx p(\varepsilon_p(i)|H_i = 1, \rho(i) = \rho_p(i)) \\ &= \frac{1}{\sqrt{2\pi}\beta\rho_p(i)} \exp\left(-\frac{\varepsilon_p(i)^2}{2(\beta\rho_p(i))^2}\right) \\ &\approx \frac{1}{\sqrt{2\pi}\beta\hat{\rho}_p(i)} \exp\left(-\frac{1}{2\beta^2}\left(\frac{\hat{\rho}_a(i)}{\hat{\rho}_p(i)}\right)^2\right).\end{aligned} \tag{21}$$

Although $\rho_a(i)$ must be larger than zero in practice (particularly in the presence of background noise), we introduce Eq. (21) as an approximation for all cases in this paper. It should be noted that this approximation provides an underestimation of the probability of the presence of speech, because in Eq. (19) we consider an ideal case where $\rho_a(i) = 0$, i.e. there is no background noise. Therefore, if we can introduce a more adequate estimation for $\mathscr{L}(H_i = 1)$, we can expect the VAD performance of the proposed method to improve.

Finally, we calculate the likelihood ratio $\Lambda(i)$ based on Eqs. (18) and (21) at frame i as follows

$$\begin{aligned}\Lambda(i) &= \frac{\mathscr{L}(H_i = 1)}{\mathscr{L}(H_i = 0)} = \frac{\alpha}{\beta} \cdot \frac{1}{\mu(i)} \exp\left(\frac{1}{2\alpha^2}\mu(i)^2 - \frac{1}{2\beta^2} \cdot \frac{1}{\mu(i)^2}\right), \\ \mu(i) &= \frac{\hat{\rho}_p(i)}{\hat{\rho}_a(i)}.\end{aligned} \tag{22}$$

If the likelihood ratio is higher than a threshold decided *a priori*, PARADE determines that there is a speech signal in the frame. Constants $\alpha = \beta = 1$ are used in this paper, thus (22) only includes the PAR, i.e. $\mu(i) = \hat{\rho}_p(i)/\hat{\rho}_a(i)$, and its inverse.

2.3. Hangover scheme

Because the PAR corresponds to the strength of the periodic components included in the observed signals, PARADE inherently cannot deal with unvoiced speech components such as stops, fricatives, and affricates, which do not include periodic components. This drawback is common to all VAD methods that utilize the periodicity. Fig. 2b shows the PAR for clean speech signals that include the voiced and unvoiced speech components shown in Fig. 2a. As shown in this figure, the PAR cannot show high values in the presence of unvoiced speech, e.g. around /h/, /ts/, /t/, /k/, and /s/. However, speech signals consist of syllables, and a syllable usually consists of a syllable nucleus (usually a vowel) and its preceding or succeeding consonants. Based on this characteristics of speech, unvoiced speech components can be handled by employing temporal margins before and after the voiced speech components detected by the periodicity based acoustic features. Such a scheme is included in the 'hangover scheme', which is widely used in conventional VAD methods to prevent the occurrence of short-term speech fragments and avoid short pauses between speech segments. Therefore, we apply the

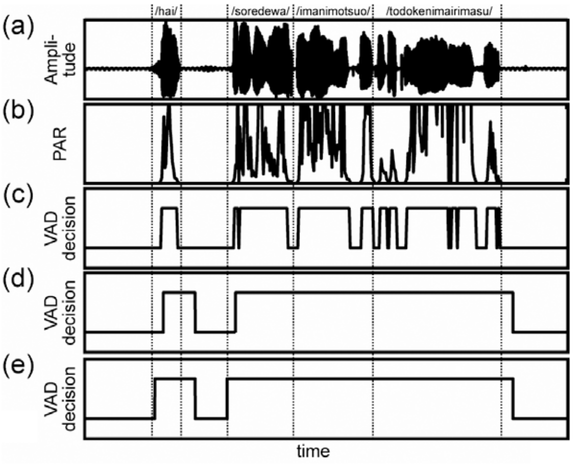

Fig. 2. The hangover scheme used to deal with unvoiced speech, to prevent short-term speech fragments, and to avoid short pauses between speech segments. (a) Speech signal in silence. This signal consists of two utterances ("*Hai. Ima nimotsu o todoke ni mairimasu.*" in Japanese. It means "Yes. I'm bringing your luggage to you now."). (b) PAR for the speech data. (c) VAD results $\mathsf{V}(i)$ obtained with simply thresholding Eq. (22). (d) Revised VAD results $\mathsf{R}(i)$ obtained with applying the hangover scheme to $\mathsf{V}(i)$. (e) Corrected VAD results by taking short-term backward margin from $\mathsf{R}(i)$.

hangover schemes proposed in (ETSI ES 202 050, 2001) to the VAD results obtained in Section 2.2.

The details of the hangover scheme are as follows (ETSI ES 202 050, 2001). Here, $\mathsf{V}(i)$ is a binary variable of 0 (speech does not exist) or 1 (speech exists), indicating the VAD result for frame i obtained by thresholding Eq. (22).

1. Set M as the maximum number of sequences of $\mathsf{V}(j) = 1$ in $i - B < j < i$. In this paper, the buffer size $B = 7$.
2. If $\mathsf{M} >= \mathsf{S}_\mathsf{p}$ and $\mathsf{T} < \mathsf{L}_\mathsf{s}$, then $\mathsf{T} = \mathsf{L}_\mathsf{s}$. In this paper, we used the 'speech possible' threshold $\mathsf{S}_p = 3$, and the short hangover frame length $\mathsf{L}_\mathsf{s} = 5$.
3. If $\mathsf{M} >= \mathsf{S}_\mathsf{L}$ and $i > \mathsf{F}_\mathsf{s}$, then $\mathsf{T} = \mathsf{L}_\mathsf{M}$, or else if $\mathsf{M} >= \mathsf{S}_\mathsf{L}$, then $\mathsf{T} = \mathsf{L}_\mathsf{L}$. In this paper, we used the 'speech likely' threshold $\mathsf{S}_\mathsf{L} = 4$, the medium hangover frame length $\mathsf{L}_\mathsf{M} = 23$, and the failsafe long hangover frame length $\mathsf{L}_\mathsf{L} = 40$.
4. If $\mathsf{M} < \mathsf{S}_\mathsf{p}$ and $\mathsf{T} > 0$, then decrement T. This step is a hangover step when the possibility of speech existing is low.
5. If $\mathsf{T} > 0$, then the revised VAD result is $\mathsf{R}(i) = 1$, or else $\mathsf{R}(i) = 0$,

where T is the hangover timer. This scheme is based on heuristics regarding the nature of speech signals. PARADE uses $\mathsf{R}(i)$ as the final VAD result. Fig. 2c and d shows $\mathsf{V}(i)$ and $\mathsf{R}(i)$. Although this scheme cannot handle unvoiced speech components at the onsets of utterances, Fig. 2d confirmed that PARADE can deal with unvoiced speech components after voiced speech parts with the hangover scheme, and detects entire speech periods. If the application of VAD allows a short delay (e.g. 50 ms), the unvoiced speech component at the utterance onset can be handled by taking backward margin from the detected utterance onset as shown in Fig. 2e. Although this paper applies an existing hangover scheme to PARADE as described above, it should be noted that the aperiodic component power estimation in Eq. (12) might provide useful information about the existence of unvoiced fricatives because the estimation provides larger values when the fricative exists than those in the noise only period. This constitutes future work.

3. Experiment

In this section, we first show qualitatively that PARADE is more robust as regards the non-stationarity of noise than some representative conventional methods through a preliminary experiment using example speech data in the presence of simulated and real non-stationary noise. Then, we conduct an experiment that uses a large number of speech data in the presence of various kinds of environmental noise to determine the validity of PARADE quantitatively. In this section, PARADE usually employs Eq. (14) (autocorrelation method) as the F0 estimation method (Rabiner, 1977; Hess, 1983). The performance of the F0 estimation methods described by Eqs. (14) and (15) is compared in the latter part of Section 3.2.

3.1. Qualitative evaluation of VAD performance

To examine the validity of PARADE qualitatively, we conducted a preliminary experiment using speech data in the presence of simulated and real noise. As the speech data and real noise data, we used example speech data and subway noise from the CENSREC-1 (AURORA-2J) database (Nakamura et al., 2003; Nakamura et al., 2005). The speech and noise were recorded with 16-bit quantization and at a sampling rate of 8 kHz.

We first compare 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. 3a obtained from CENSREC-1.

We examined the decomposition of the periodic/aperiodic components and the likelihood ratio for VAD calculated with Eqs. (3), (12), (13), and (22), respectively. A test signal was created by adding white noise to the speech data shown in Fig. 3a (Fig. 3b). The results obtained from PARADE are shown in Fig. 3c–e. These results indicate that our method decomposes periodic and aperiodic components well in the presence of such an ideal white noise (Fig. 3c and d). In addition, the likelihood ratios (Fig. 3e) correspond to the period in which the speech signals exist. The result indicates the effectiveness of using this likelihood ratio for VAD. The VAD result based on the likelihood ratios is shown in Fig. 3f. For comparison, results obtained with VAD based on log-likelihood ratio

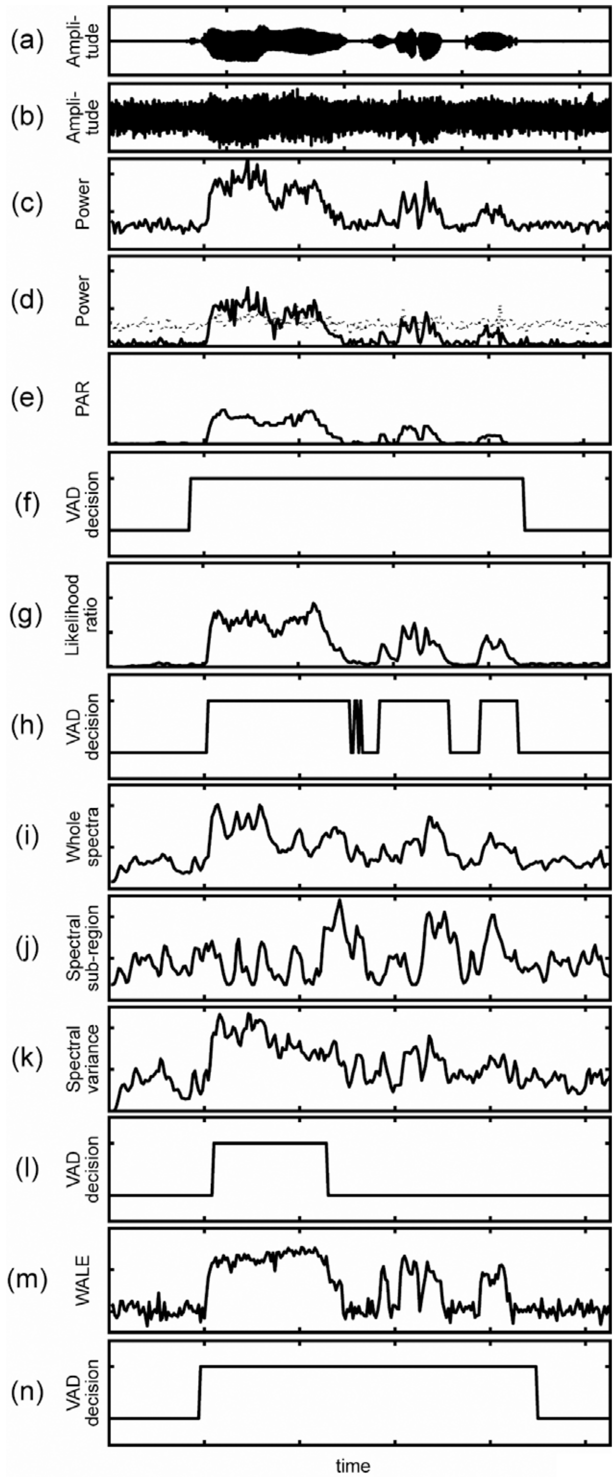

Fig. 3. VAD features of a speech signal in the presence of white noise and a comparison of the VAD results obtained for the speech signal. (a) Speech signal in silence. (b) Speech signal mixed with white noise. (c) Power of (b). (d) Estimated periodic (solid line) and aperiodic (dashed line) components for (b). (e) PAR derived from the periodic and aperiodic components (d). (f) VAD result for (b) obtained with PARADE. (g) Features (log-likelihood ratios) for (b) obtained with statistical VAD (Sohn et al., 1999). (h) VAD result for (b) obtained with Sohn's statistical VAD. (i)–(k) Features (whole spectrum, spectral sub-region, spectral variance) for (b) obtained with (ETSI ES 202 050, 2001). (l) VAD result for (b) obtained with ETSI ES 202 050. (m) WALE features (Kristjansson et al., 2005) for (b). (n) VAD result for (b) obtained with WALE.

tests (LRT) (Sohn et al., 1999), ETSI ES 202 050 VAD for frame dropping (ETSI ES 202 050, 2001), and a windowed autocorrelation lag energy (WALE) (Kristjansson et al., 2005) are also shown in Fig. 3g–n.

The LRT approach (Sohn et al., 1999) utilizes likelihood ratios derived from *a priori* SNRs calculated from the spectral amplitude estimated by the MMSE approach (Ephraim and Malah, 1984). With this method, the likelihood ratio for the *k*th frequency bin is calculated as follows:

$$\Lambda(i,k) = \frac{p(X(i,k)|H_i = 1)}{p(X(i,k)|H_i = 0)} = \frac{1}{1+\xi_k} \exp\left(\frac{\gamma_k \xi_k}{1+\xi_k}\right), \tag{23}$$

where ξ_k and γ_k are *a priori* and *a posteriori* SNRs, respectively. Based on Eq. (23), the geometric mean of the likelihood ratios is calculated as follows:

$$\log \Lambda(i) = \frac{1}{K} \sum_{K=0}^{K-1} \log \Lambda(i,k). \tag{24}$$

LRT-based VAD obtains final decision statistics based on the likelihood ratios derived from Eq. (24) and the HMM-based hangover scheme. Fig. 3g shows the likelihood ratios obtained from applying the HMM hangover scheme to Eq. (24), and Fig. 3h shows the VAD results obtained with this method.

ETSI ES 202 050 VAD for frame dropping (ETSI ES 202 050, 2001) simultaneously utilizes the 'whole spectra' shown in Fig. 3i, namely the squared sum of the Mel-warped Wiener filter coefficients, which corresponds to spectral information from speech signals after noise reduction, the 'spectral sub-region' shown in Fig. 3j, namely the average of the second, third and fourth Mel-warped filter coefficients, which corresponds to the speech characteristics that the power of speech signals biased into low frequency region, and the 'spectral variance' shown in Fig. 3k, namely the variance of the values comprising the whole frequency range of the linear-frequency Wiener filter coefficients, which corresponds to the degree to which harmonic signals are present in the observed signals. The VAD result obtained with this method is shown in Fig. 3l.

The WALE approach (Kristjansson et al., 2005) is based on ACF, and outperformed other conventional ACF based VAD methods. WALE is designed to capture the characteristics of vocal tracts on ACF, and is calculated as follows:

$$WALE(i) = \max_{l} \sum_{\tau=l}^{l+W-1} E\{x(n)x(n+\tau)\}, \tag{25}$$

where τ is a lag for the ACF, and W is the length of a lag window for frames. This WALE feature for the observed signal shown in Fig. 3b is illustrated in Fig. 3m.

The former two conventional methods include hangover mechanisms, so we utilized the mechanisms described in the corresponding papers. With the WALE approach, we applied the PARADE hangover mechanisms described in Section 2.3 to WALE. That is, the difference between the

PARADE and WALE approaches is solely their acoustic features. The VAD result obtained from the WALE features is shown in Fig. 3n. In Fig. 3, the comparison indicates that PARADE and WALE perform better than the other two methods even in the presence of white noise. The results suggest that PARADE is at least robust as regards stationary noise such as white noise. This confirms the validity of utilizing the periodic feature of the observed signals in the presence of white noise.

We then compared the VAD performance using another test signal (Fig. 4b) created by adding amplitude-modulated white noise to the speech data shown in Fig. 4a in order to test the robustness of PARADE as regards the non-stationarity of noise. The amplitude modulation rate was 4 Hz, and the modulation depth was 0.4. Fig. 4c–e, g, i–k, and m show the features calculated by PARADE and the three conventional methods described above. Although the conventional features based on the power of observed signals were deteriorated by the amplitude modulation of the noise, the results indicate that the likelihood ratios of PARADE are insensitive to such amplitude modulation. Fig. 4f, h, l, and n compare the VAD results obtained with the four VAD methods. The results suggest that PARADE is also robust as regards non-stationarity of noise as such amplitude modulation. WALE also performed better than the other two conventional methods. These results again indicate the validity of utilizing the periodic feature of the observed signals in the presence of amplitude-modulated white noise.

Finally we compared the VAD performance obtained with the proposed and conventional methods in the presence of noise in the real world, which includes non-stationary sounds. For this comparison, we used noisy speech data taken from CENSREC-1 (Fig. 5b) created by adding subway noise to clean speech data at an SNR of 0 dB. The subway noise includes not only stationary sounds such as train motor sounds but also non-stationary sounds such as train wheel sounds. Fig. 5c–e, g, i–k, and m show the VAD features calculated by PARADE and the three conventional methods mentioned above. 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 by circles in Fig. 5b). It should be noted that the periodic features obtained with WALE and PARADE also indicate estimation errors for the period when non-stationary sounds appear although these sounds do not include periodic components. This is because it is difficult to avoid such estimation errors in the presence of real environmental sounds unlike the case with white noise. This fact indicates that it is difficult to avoid detection errors caused by estimation errors even when such periodic features are utilized for VAD. On the other hand, our proposed likelihood ratios of the periodic and aperiodic components could moderate the influence of the non-stationary sounds and the variation in the noise amplitude, and deteriorated less than conventional features. Fig. 5f, h, l and n

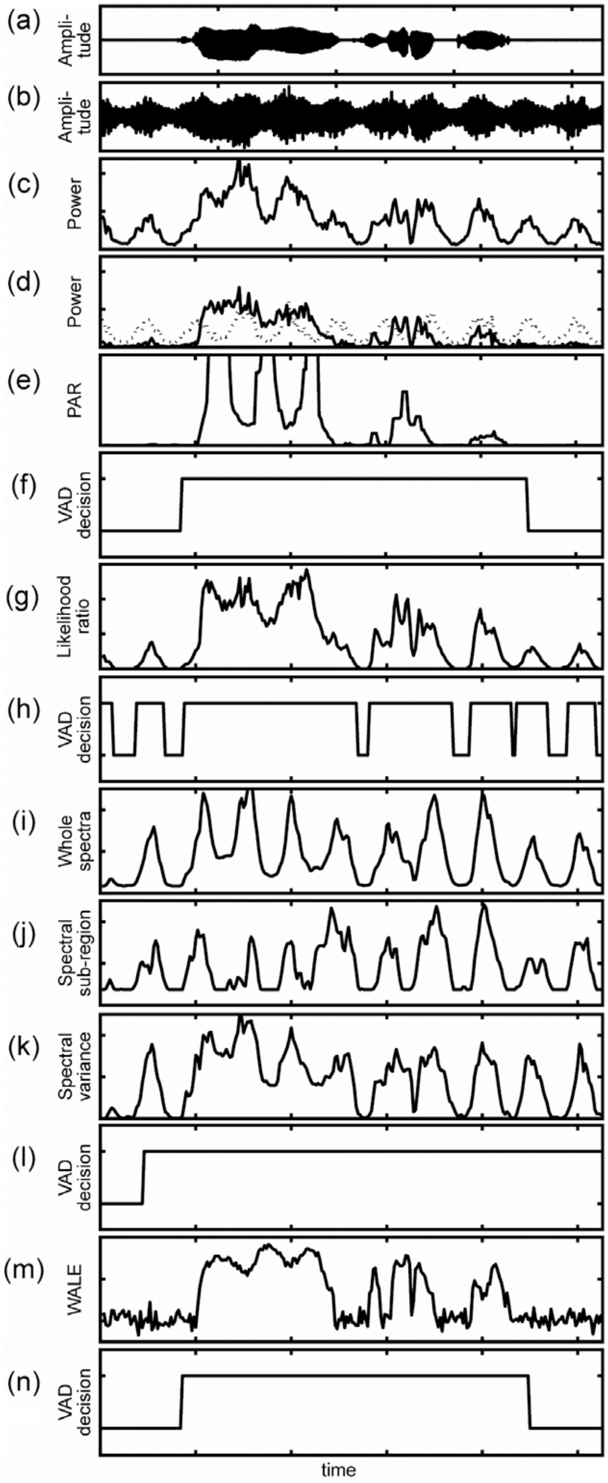

Fig. 4. VAD features of a speech signal in the presence of amplitude-modulated white noise and a comparison of the VAD results obtained for the speech signal. (a) Speech signal in silence. (b) Speech signal mixed with amplitude-modulated white noise. (c) Power of (b). (d) Estimated periodic (solid line) and aperiodic (dashed line) components for (b). (e) PAR derived from the periodic and aperiodic components (d). (f) VAD result for (b) obtained with PARADE. (g) Features (log-likelihood ratios) for (b) obtained with statistical VAD (Sohn et al., 1999). (h) VAD result for (b) obtained with Sohn's statistical VAD. (i)–(k) Features (whole spectrum, spectral sub-region, spectral variance) for (b) obtained with (ETSI ES 202 050, 2001). (l) VAD result for (b) obtained with ETSI ES 202 050. (m) WALE features (Kristjansson et al., 2005) for (b). (n) VAD result for (b) obtained with WALE.

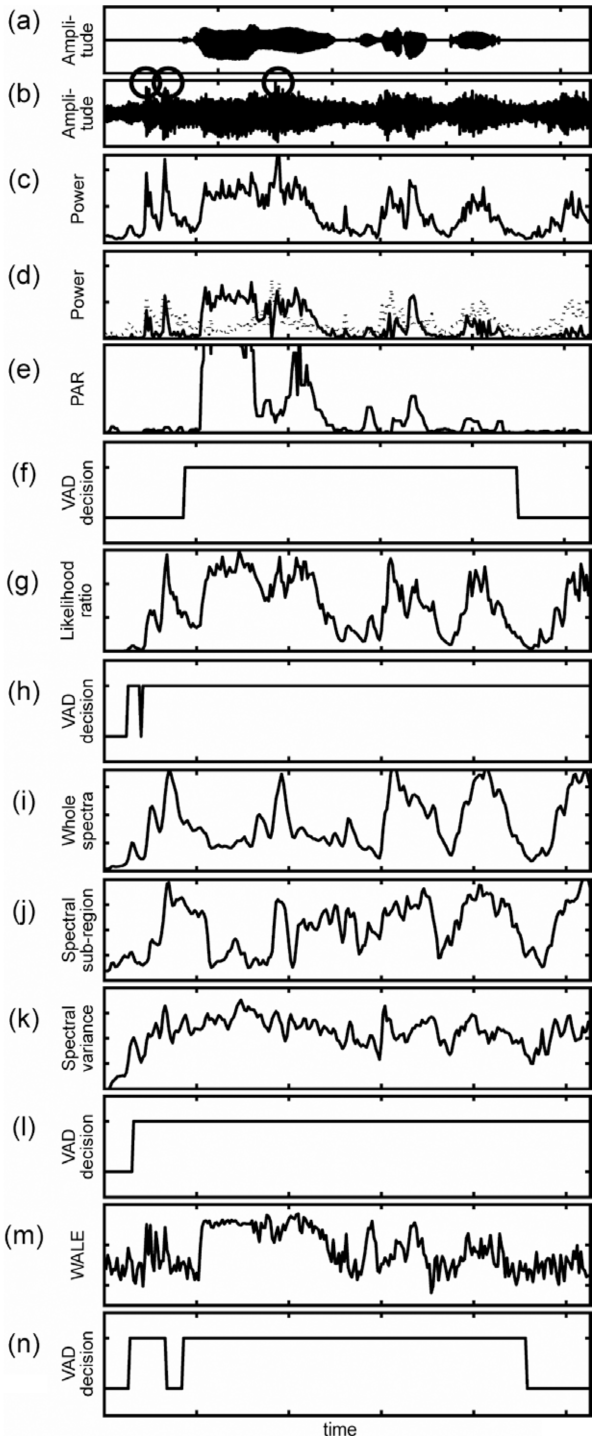

Fig. 5. VAD features of a speech signal in the presence of subway noise and a comparison of VAD results obtained for the speech signal. (a) Speech signal in silence. (b) Speech signal mixed with subway noise at an SNR of 0 dB. Circles indicate non-stationary noise (train wheel sounds). (c) Power of (b). (d) Estimated periodic (solid line) and aperiodic (dashed line) components for (b). (e) PAR derived from the periodic and aperiodic components (d). (f) VAD result for (b) obtained with PARADE. (g) Features (log-likelihood ratios) for (b) obtained with statistical VAD (Sohn et al., 1999). (h) VAD result for (b) obtained with Sohn's statistical VAD. (i)–(k) Features (whole spectrum, spectral sub-region, spectral variance) for (b) obtained with (ETSI ES 202 050, 2001). (l) VAD result for (b) obtained with ETSI ES 202 050. (m) WALE features (Kristjansson et al., 2005) for (b). (n) VAD result for (b) obtained with WALE.

confirm the effectiveness of PARADE compared with the three conventional VAD methods.

In addition to the above results, to show the effectiveness of PARADE as regards temporal changes in the powers of noise in the real world, we compared the VAD performance using two concatenated noisy speech data with different SNRs taken from CENSREC-1. A test signal (Fig. 6b) 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. 6b. Fig. 6c–f compares the VAD results obtained with PARADE and the three conventional VAD methods. Although the conventional methods have mechanisms for updating noise statistics or thresholds, they failed to update correctly, and could not perform well. On the other hand, PARADE performed well regardless of the temporal change in the SNRs despite the lack of threshold updating. The above results suggest that PARADE can detect voice activity robustly even in the presence of noise in the real world, which includes both stationary and non-stationary noise components. The robustness as regards the non-stationarity of noise is a particular advantage of PARADE in terms of practical use.

3.2. Quantitative evaluation of VAD performance

To examine the validity of PARADE quantitatively, we conducted an experiment using a large quantity of speech data in the presence of real noise. Speech data mixed with

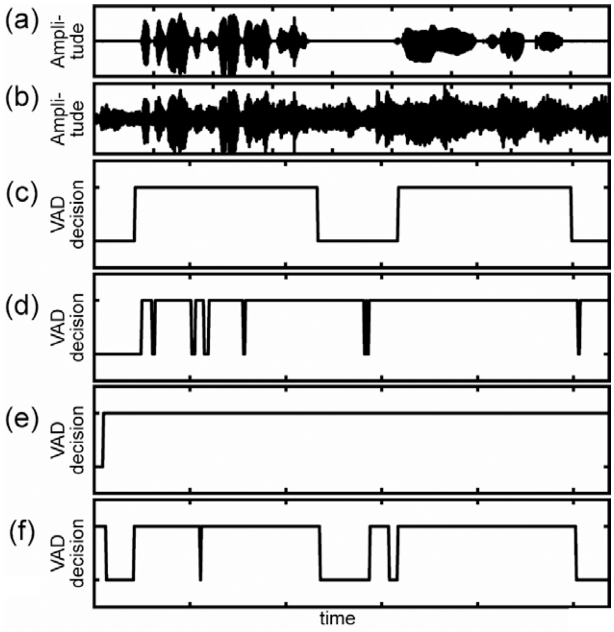

Fig. 6. Comparison of VAD results for a speech signal in the presence of subway noise when the SNR changes temporally from 5 to 0 dB. The SNR of the former half is 5 dB, and that of the latter half is 0 dB. (a) Speech signal in silence. (b) Speech signal mixed with subway noise. (c) VAD result for (b) obtained with PARADE. (d) VAD result for (b) obtained with statistical VAD (Sohn et al., 1999). (e) VAD result for (b) obtained with (ETSI ES 202 050, 2001). (f) VAD result for (b) obtained with WALE features (Kristjansson et al., 2005).

environmental noise recorded in the real world were used in this evaluation. We used travel arrangement dialogue data spoken in Japanese (SDB-L in Nakamura et al., 1996). The data consist of 2292 utterances spoken by 178 speakers. The utterance duration is between 1.4 and 12.1 s. The total length of the speech data was around 117 min. 69.8% of the speech periods were voiced speech, e.g. vowels or voiced consonants, and 30.2% were unvoiced speech periods, e.g. fricatives or stops. We down-sampled the data sampling rate from 48 to 8 kHz. Correct VAD references were generated based on a hand labeled transcription for SDB-L, which includes onset, offset, and pause information. Fig. 8a and b shows an example of clean speech data, and its correct VAD reference, respectively.

As noise data, we recorded real environmental sounds at an airport arrival gate, a railway station platform and ticket gate, a subway station platform and ticket gate, a restaurant, and on a street in Tokyo. The recording equipment consisted of omni-directional microphones (Sony ECM-77B) and portable IC recorders (Marantz PMD670), and the data were sampled at 48 kHz. The example spectrograms of the recorded environmental sounds are shown in Fig. 7. Various sound events such as babble, train sounds, car sounds, ambient music, and the sound of people walking make all the environmental sounds non-stationary as shown in Fig. 7. The recorded data were also down-sampled to 8 kHz, and added to the above speech data at SNRs of 0, 5, and 10 dB. Because environmental sounds are not stationary, we adjusted the SNRs according to the ratio of the power peaks of the speech and noise data within the period of an utterance. Fig. 8c shows examples of noisy speech data obtained by adding street noise to the clean speech data shown in Fig. 8a at an SNR of 0 dB.

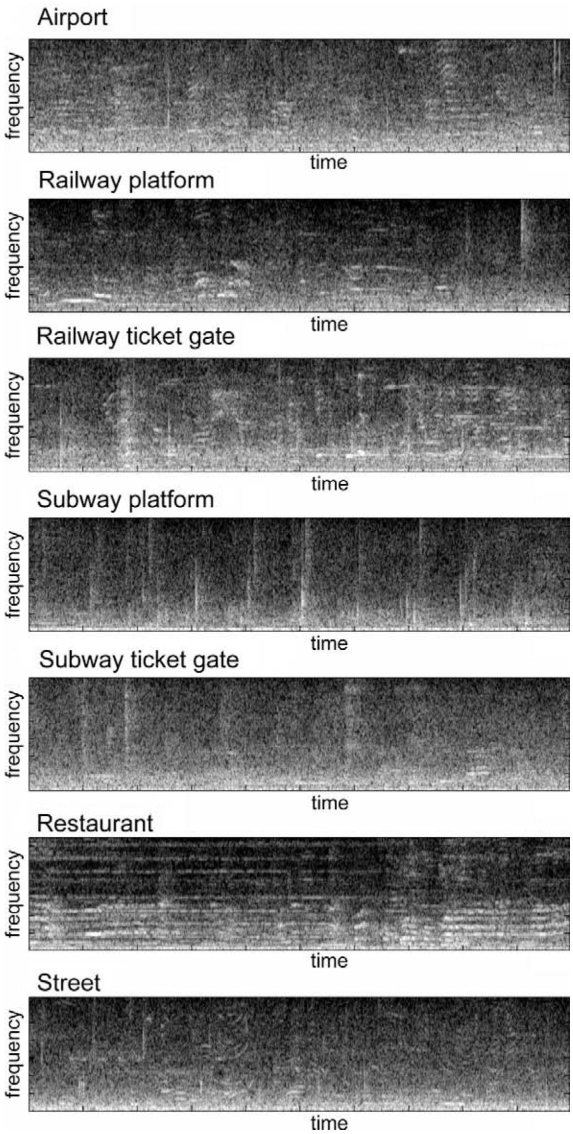

Fig. 7. Example spectrograms for 5 s of data from the recorded environmental sounds. The frequency range is from 0 to 4 kHz.

To assess the performance obtained with the proposed PARADE, we compared the curves of receiver operating characteristics (ROC) obtained with the proposed PARADE and conventional methods. In this paper, the ROC curves were generated from false acceptance rate (FAR) and false rejection rate (FRR) calculations with various thresholds as follows

$$FAR = \frac{\text{Number of frames wrongly detected as speech}}{\text{Number of non-speech frames}}, \tag{26}$$

$$FRR = \frac{\text{Number of frames wrongly detected as non-speech}}{\text{Number of speech frames}}. \tag{27}$$

The superior method will achieve both a lower FAR and a lower FRR. As a comparison, we used four standard methods: ITU-T G.729 Annex B (ITU-T Recommendation G.729 Annex B, 1996), ETSI GSM AMR VAD1 and VAD2 (ETSI EN 301 708, 1999), and ETSI WI008

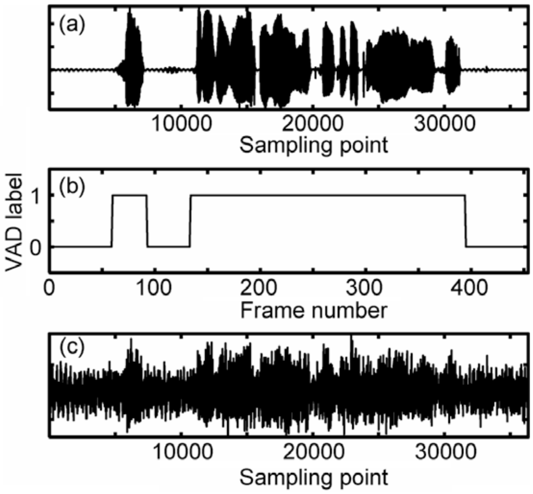

Fig. 8. Samples of test speech data for evaluating VAD performance. (a) Speech waveform under silent conditions. (b) Correct VAD data for (a). (c) Speech waveform (a) mixed with street noise at an SNR of 0 dB.

Advanced Frontend (AFE) VAD for frame dropping (ETSI ES 202 050, 2001), and four conventional methods based on LRT (Sohn et al., 1999), higher-order statistics (HOS) (Nemer et al., 2001), long-term spectral divergence (LTSD) (Ramírez et al., 2004), and WALE (Kristjansson et al., 2005). The HOS approach utilizes biased distributions of the residual speech signals after analyzing the observed signals by linear predictive coding (LPC) analysis (Itakura and Saito, 1968), and employs skewness, kurtosis, skewness-to-kurtosis ratios, SNRs, probability of noise-only frames, and LPC prediction errors of observed signals as acoustic features, and a speech/noise state machine as a detection mechanism. The LTSD approach is inspired by the masking mechanisms of human auditory systems, and employs the following acoustic features:

$$LTSE_N(i,k) = \max\{|X(i+j,k)|\}_{j=-N}^{j=+N}, \tag{28}$$

$$LTSD_N(i) = 10\log_{10}\left(\frac{1}{K}\sum_{k=0}^{K-1}\frac{|LTSE_N(i,k)|^2}{|N(k)|^2}\right), \tag{29}$$

where N indicates the order of the LTSE, and $N(k)$ is the average noise spectrum for the mth frequency bin. $N(k)$ is updated at non-speech frames. In addition to the above conventional methods, we used another PARADE algorithm based on HNR (Boersma, 1993) as an acoustic feature instead of the likelihood ratio obtained from Eq. (22). In this case, HNR is calculated as follows

$$HNR(i) = 10\log_{10}\left(\frac{r_x(\tau_{\max})}{1 - r_x(\tau_{\max})}\right), \tag{30}$$

where $r_x(\tau_{\max})$ indicates the ACF coefficients at lag $\tau_{\max}$ obtained as in Eq. (14). For the HNR calculation, ACF is normalized by $r_x(0)$ and the effect of the window function in ACF is compensated as proposed by Boersma (1993). As described in Section 1, HNR is widely used for deciding the voiced/unvoiced parts of speech signals (Mousset et al., 1996; Ahmadi and Spanias, 1999; Fisher et al., 2006; Nakatani et al., 2008). This comparison aims to measure the improvement obtained with the proposed feature extraction technique. In all cases, PARADE analyzed the observed signals using a 25 ms long analysis window with a 15 ms overlap.

Fig. 9 shows the ROC curves achieved under street noise conditions at an SNR of 10 dB. PARADE achieved better ROC curves than those obtained with the other conventional methods. The results obtained with the standard methods were consistent with the results for noisy speech contained in AURORA-2 reported in (Ramírez et al., 2004). WALE outperformed G.729 Annex B probably owing to the use of the periodicity, whereas HNR and PARADE outperformed WALE by utilizing information from both the periodic and aperiodic components of the observed signals as described in Section 3.1. In addition, PARADE outperformed HNR because of the precise periodic/aperiodic component power decomposition described by Eqs. (3), (12), and (13). Although the HNR calculation in Eq. (30) can avoid the effect of white noise because it is uncorrelated in time, the value of the coefficient $r_x(\tau_{\max})$ is affected by colored noise such as environmental sound. Because PARADE considers the aperiodic component power at frequencies where the periodic component exists, and employs relaxed restrictions regarding the frequency distributions of aperiodic components as in Eq. (8), PARADE outperformed HNR. PARADE also outperformed the AFE, LRT, HOS, and LTSD approaches in the small FAR region. In the large FAR region, PARADE achieved a performance level comparable to those obtained from AMR VAD1 and VAD2.

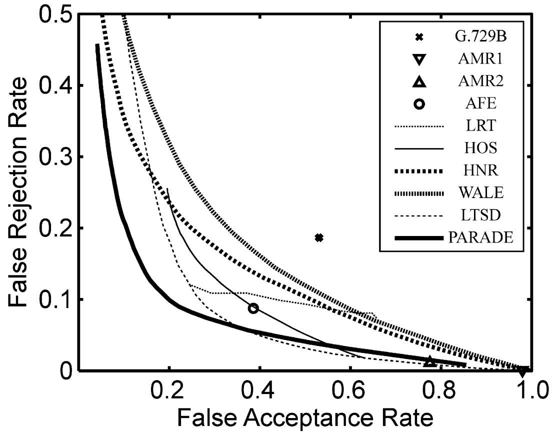

Fig. 9. ROC curves for speech mixed with street sound at an SNR of 10 dB obtained with G.729B (ITU-T Recommendation G.729 Annex B, 1996), ETSI GSM AMR1 and AMR2 (ETSI EN 301 708, 1999), AFE (ETSI ES 202 050, 2001), LRT (Sohn et al., 1999), HOS (Nemer et al., 2001), LTSD (Ramírez et al., 2004), WALE (Kristjansson et al., 2005), HNR (Boersma, 1993), and PARADE (proposed method).

To compare the performance for all kinds of conditions, we introduced the equal error rate (EER), which is the error rate at which FAR and FRR are the same. Table 1 shows the result obtained with the LTSD and WALE approaches, the HNR based VAD, and PARADE. The results confirmed that PARADE always achieved better EERs than those obtained with the other methods.

It should be noted that the performance of PARADE is affected by the F0 estimation method. We also introduced EER to compare the performance of PARADE with Eqs. (14) and (15) for the F0 estimation methods. The results are also shown in Table 1, and indicate that PARADE with Eq. (14) could generally achieve superior performance to that achieved by PARADE with Eq. (15). This may be owing to its simpler implementation of REPS (Nakatani and Irino, 2004) in Eq. (15), whereas it is of interest that PARADE with Eq. (14) achieved good performance regardless of the simple ACF based F0 estimation. ACF based F0 estimation as in Eq. (14) also has the advantage of low computational cost compared with Eq. (15) or other robust F0 estimation methods. However, it should also be noted that this result does not restrict the best combination of PARADE and the F0 estimation method. The PARADE framework can use any kind of F0 estimation method, thus more precise F0 estimation methods than

Table 1
Experimental results evaluating the robustness of VAD methods with noisy speech data. These are the average equal error rates at SNRs of 0–10 dB for seven kinds of environmental sounds and all test sets (overall).

Environmental noise	SNR (dB)	Method (%)				
		PARADE with Eq. (14)	PARADE with Eq. (15)	HNR	LTSD	WALE
Airport	0	**25.8**	27.7	36.0	42.2	35.9
	5	**18.1**	18.8	26.1	32.8	24.5
	10	**14.4**	15.1	20.5	23.0	17.6
Railway platform	0	**25.7**	28.0	36.1	39.2	38.0
	5	**18.6**	19.3	26.1	30.1	27.2
	10	**15.2**	15.4	21.4	22.0	20.2
Railway ticket gate	0	**21.8**	24.0	37.1	41.5	39.0
	5	**15.5**	16.3	25.0	31.6	26.7
	10	**13.6**	13.8	18.8	21.6	19.1
Subway platform	0	**23.1**	30.1	52.0	32.7	55.7
	5	**16.2**	20.4	42.6	22.5	49.2
	10	**13.7**	15.4	35.0	17.2	42.4
Subway ticket gate	0	**24.4**	26.8	35.4	42.4	38.1
	5	**16.9**	18.3	23.6	32.7	26.8
	10	**13.8**	14.1	18.1	22.5	20.0
Restaurant	0	**26.6**	36.4	49.3	35.0	51.5
	5	**18.0**	23.2	38.2	27.2	41.9
	10	**14.2**	16.3	30.0	20.0	32.4
Street	0	**26.4**	33.9	43.4	36.6	47.0
	5	**17.5**	22.1	29.9	26.5	35.4
	10	**14.3**	15.8	21.9	19.3	25.6
Overall	0	**24.8**	29.6	41.3	38.5	43.6
	5	**17.3**	19.8	30.2	29.1	33.1
	10	**14.2**	15.1	23.7	20.8	25.3

Eq. (14) are expected to provide better VAD performance. This will be examined in future work.

Finally, to assess the computational load imposed by PARADE, we measured the real-time factor (RTF), which we define as the total processing time required on a Linux PC with a 3.2 GHz Intel Pentium 4 processor and 2 GB memory, divided by the total computing time needed for the input signals being processed. PARADE was implemented in C language for this RTF calculation. The RTF was 0.0028. This result confirmed that PARADE can perform with a low computational load.

4. Noise robust front-end processing for ASR with PARADE

As described in Section 1, robust VAD can be applied to speech enhancement and ASR techniques (Junqua et al., 1994). This section describes an application of PARADE to noise robust front-end processing for ASR, and shows the effectiveness of the application through evaluation experiments using noisy speech data.

Various front-end processing approaches have been proposed to cope with the poor performance of ASR systems in the presence of environmental noise (e.g. Benítez et al., 2001; Li et al., 2001; Noé et al., 2001; Macho et al., 2002). Generally, such front-end processing for ASR consists of noise reduction methods, channel distortion compensation methods, and feature extraction methods. For example, ETSI ES 202 050 (2001), which is an effective front-end approach, consists of 2-stage adaptive Wiener filtering (Agarwal and Cheng, 1999) for noise reduction, blind equalization that compensates for the channel distortion (Mauuary, 1998), and power based MFCC feature extraction. In addition, ETSI ES 202 050 has VAD mechanisms to estimate speech segments for updating the noise statistics in Wiener filtering (VADNest), and dropping non-speech frames to reduce the number of false alarms that occur in ASR. After the standardization of ETSI ES 202 050, ETSI ES 202 212 (2003) was also standardized to allow improved recognition of tonal languages.

Recently, Ishizuka and Nakatani (2006) proposed a front-end processing technique for ASR that consists of a speech feature extraction method SPADE (Subband based Periodicity and Aperiodicity DEcomposition) (Ishizuka et al., 2006), adaptive Wiener filtering (Adami et al., 2002) for noise reduction, and cepstral normalization methods (Chen et al., 2002) for channel compensation. By normalizing SPADE features for the temporal axis, we can mitigate the effect of the differences between the channel characteristics of training and test data. The blind equalization method in ETSI ES 202 050 (Mauuary, 1998) also aims to mitigate the effect. This front-end is called “SPADE front-end” in this paper. Although SPADE front-end has not been incorporated in effective VAD mechanisms, it could achieve comparable performance to that achieved by ETSI ES 202 050 for ASR in environmental noise (Ishizuka and Nakatani, 2006). Since the performance of ETSI ES 202 050 can also be improved by using an effective VAD method (Ramírez et al., 2004), the ASR performance achieved by SPADE front-end is expected to be improved further if it employs effective VAD mechanisms such as PARADE.

In this section, we describe how we applied PARADE to the SPADE front-end. Unlike the SPADE front-end proposed by Ishizuka andNakatani, 2006, in this work we used RASTA (RelAtive SpecTrA) (Hermansky and Morgan, 1994) rather than cepstral normalization methods (Chen et al., 2002) to compensate for channel distortion. Because the cepstral normalization methods need cepstral sequences with a certain temporal length, they cannot work in a frame-by-frame manner. RASTA filtering primarily aims to extract amplitude modulations inherent in speech signals, while it normalizes the absolute power of the acoustic features. Thus it can mitigate the effect of the difference between the channel characteristics of training and test data. Because RASTA filtering works in real time, namely it only uses information about past frames, it can be used to make the SPADE front-end work in real time instead of the cepstral normalization method. PARADE was used to estimate the noise statistics for Wiener filtering (Adami et al., 2002), and to drop frames that do not include speech segments thus reducing the false alarms in ASR. The block diagrams of ETSI ES 202 050, the SPADE front-end, and the proposed front-end are shown in Fig. 10.

4.1. Applying PARADE to Wiener filtering

To realize adaptive noise compensation, the SPADE front-end employed an adaptive Wiener filtering algorithm (Adami et al., 2002). With this algorithm, a Wiener filter $|H(i,k)|^2$ is estimated by using the following equation for each frame:

$$|H(i,k)|^2 = \max(1 - \gamma(i) \cdot |\widehat{N}(i,k)|^2 / |X(i,k)|^2, \beta), \tag{31}$$

where $|X(i,k)|^2$ and $|\widehat{N}(i,k)|^2$ are the power spectrum estimations of noisy speech and additive noise signals, respectively, at the ith frame, β is the flooring parameter used to avoid negative or small transfer function components, and the noise overestimation factor $\gamma(i)$ varies linearly in proportion to the local *a posteriori* SNR between large and small values. The *a posteriori* SNR is calculated from $|X(i,k)|^2$ and $|\widehat{N}(i,k)|^2$. In Ishizuka and Nakatani (2006), $|\widehat{N}(i,k)|^2$ is estimated from the first several frames, and then updated when $|X(i,k)|^2$ is smaller than double $|\widehat{N}(i,k)|^2$ (this rule can be considered a simple energy-based VAD mechanism) as follows:

$$\log|\widehat{N}(i,k)|^2 = (1 - \alpha)\log|\widehat{N}(i-1,k)|^2 + \alpha\log(1 + |X(i-1,k)|^2), \tag{32}$$

where α is an update parameter, e.g. 0.01. PARADE is applied to this noise update procedure, that is, $|\widehat{N}(i,k)|^2$ is updated as Eq. (32) only when Eq. (22) is smaller than a fixed threshold, e.g. 1.0.

The clean speech power spectrum, $|\widehat{S}(i,k)|^2$, is estimated by multiplying $|H(i,k)|^2$ and $|X(i,k)|^2$

$$|\widehat{S}(i,k)|^2 = \max(|X(i,k)|^2 \cdot |H(i,k)|^2, \beta_{filt}|\widehat{N}(i,k)|^2), \tag{33}$$

where β_{filt} is the noise flooring parameter, e.g. 0.001.

4.2. Applying PARADE to frame dropping

PARADE is also employed for dropping non-speech frames before decoding to prevent false alarms in ASR. In this case, to reflect the noise characteristics in PARs, threshold θ for Eq. (22) is determined from the first several frames, i.e. 20 frames only including noise as follows:

$$\theta = \max(\mu_{A(i)} + c \cdot \sigma_{A(i)}, \beta_{thres}), \tag{34}$$

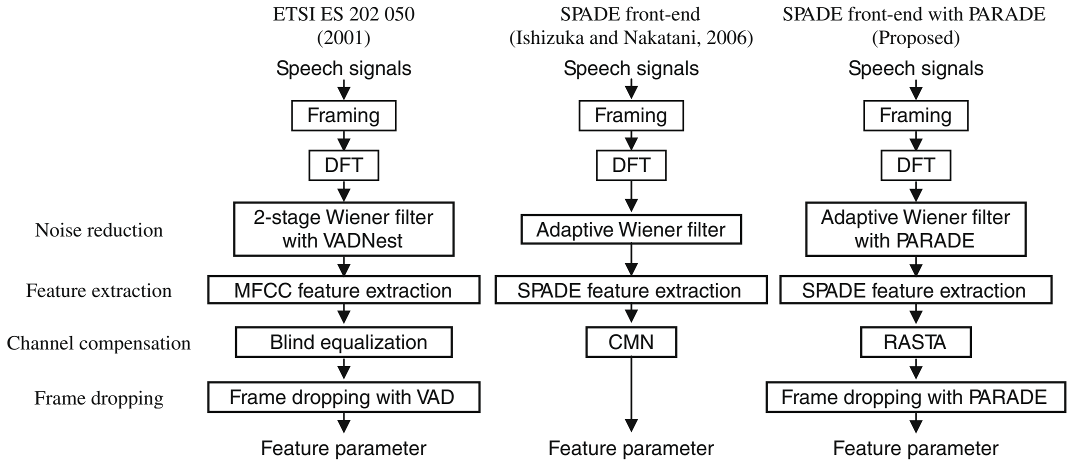

Fig. 10. Block diagrams of (ETSI ES 202 050, 2001), SPADE front-end (Ishizuka and Nakatani, 2006), and the proposed SPADE front-end with PARADE. Each front-end includes noise reduction, feature extraction, channel compensation, and frame dropping stages.

where and $\sigma_{\Lambda(n)}$ are the mean and standard deviation of $\Lambda(n)$, β_{thres} is the flooring parameter, e.g. 0.5, c is a constant, e.g. 4.0.

4.3. Evaluation of ASR performance in noisy environment

We conducted an evaluation experiment to assess the noisy ASR performance of SPADE front-end with PARADE by using CENSREC-1-C (Kitaoka et al., 2007) as test speech data. CENSREC-1-C is a speech database that consists of two sets of speech data: a simulated data set and a real recorded data set. The simulated data set includes connected continuous digit utterances taken from CENSREC-1 (Nakamura et al.,2003, 2005), which is the Japanese version of AURORA-2 (Hirsh and Pearce, 2000; Pearce and Hirsh, 2000). The sampling rate was 8 kHz. The speech data were continuous digit utterances spoken by 104 speakers from the test set in CENSREC-1, and these utterances were connected at intervals of one second for each speaker. The simulated speech data set consists of speech mixed with eight kinds of environmental sounds identical to those used in AURORA-2, i.e. subway, babble, car, exhibition, restaurant, street, airport, and station noises. The speech data included 9 or 10 utterances, and these were mixed with environmental noise at SNRs of 0, 5, 10, 15, and 20 dB. The channel characteristics were G.712. A total of 4992 speech data were used for this evaluation. Fig. 11 shows examples of simulated noisy speech data in CENSREC-1-C. The pauses inserted between the continuous digit utterances mean that the test data is more practical than the test data in CENSREC-1 or other AURORA databases, which include only one utterance for each test. Although the test data in CENSREC-1 or AURORA database was generated manually, one utterance in one test data assumes that the speech segment can be detected from the observed signals, which must include many utterances and various environmental noises, with an ideal VAD mechanism or push-to-talk mechanism. Thus, these databases are not appropriate for evaluating the effect of the VAD mechanism when we use ASR under practical conditions in which some VAD mechanisms are really needed. In other words, we cannot assume the existence of ideal VAD and push-to-talk mechanisms. Therefore, in this section we used CENSREC-1-C as an appropriate database with which to evaluate the front-end processing for ASR, which is incorporated in the VAD mechanism.

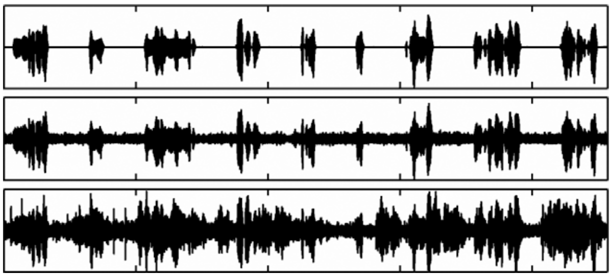

Fig. 11. Example simulated noisy speech data from CENSREC-1-C. (Top) Speech waveform (connected continuous digit utterances) under clean conditions. (Middle, bottom) Speech waveforms mixed with airport noise at SNRs of 10 and 0 dB, respectively.

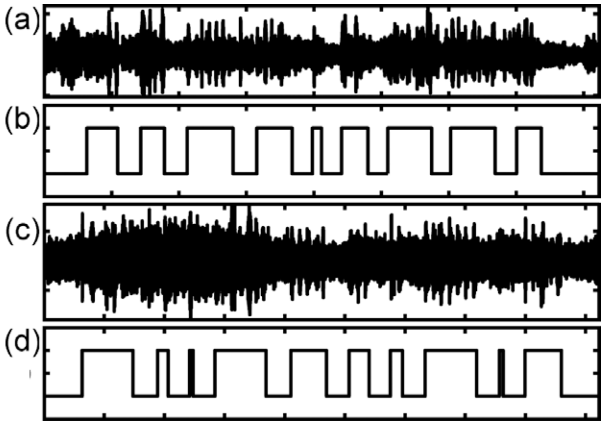

Fig. 12. Example real recorded noisy speech data from CENSREC-1-C. (a) Speech waveform recorded at a student cafeteria (noise sound level of around 69.7 dBA). (b) Hand-labeled target speech periods for (a). (c) Speech waveforms recorded beside freeway (noise sound level of around 69.2 dBA). (d) Hand-labeled target speech periods for (c).

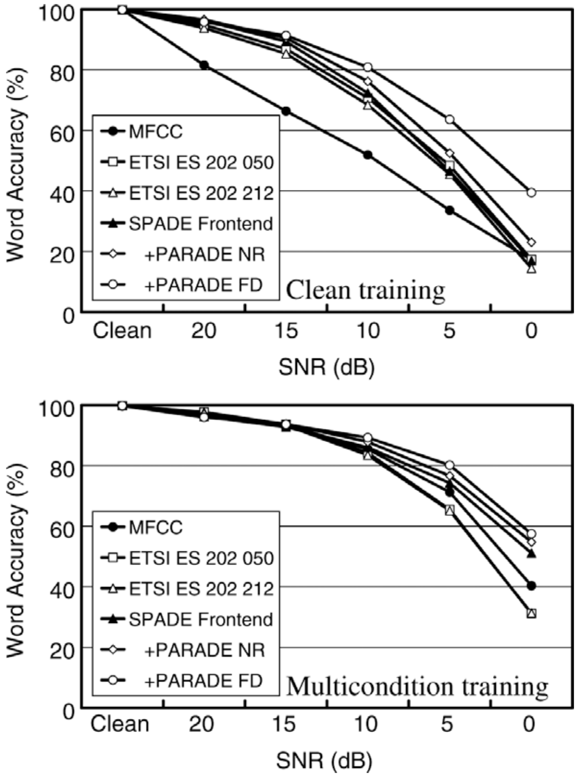

Fig. 13. Experimental results evaluating the robustness of front-end processing with CENSREC-1-C simulated noisy speech data. The results were obtained from MFCC-based front-end (ETSI ES 202 050, 2001; ETSI ES 202 212, 2003), SPADE front-end (Ishizuka and Nakatani, 2006), SPADE front-end with PARADE based noise reduction (NR), and SPADE front-end with PARADE based NR and frame dropping (FD). Average word accuracies for eight noise conditions. Experimental results obtained under clean (top) and multi-condition (bottom) training conditions are shown.

The real recorded data set in CENSREC-1-C was recorded in two real noisy environments (a student cafeteria and beside a freeway) with two different noise sound levels (around 60 dBA: high SNR and around 70 dBA: low SNR). The data was first recorded at a sampling rate of 48 kHz, and then down-sampled to 8 kHz. The microphone was around 50 cm from the speaker during the recording. There were ten speakers (five males and five females). Four speech data were recorded per subject. One item of speech data included 8–10 utterances of continuous 1–12 digit numbers with two-second intervals for each utterance in each noisy environment and each SNR condition. The channel characteristics are different from those of the simulated data set. Fig. 12 shows examples of the real recorded speech data set and hand-labeled speech periods for each example. The total number of speech data was 1532.

This ASR performance evaluation task was continuous digit ASR, and the number of words in the vocabulary was 11. 16-state 24-Gaussian mixture HMMs were used as pattern-classifiers for each one-digit speech. The HMMs were trained with a clean training data set and a multi-condition training data set included in CENSREC-1. The clean training data set consists of speech data spoken in a noiseless environment, and includes 8440 continuous digit utterances spoken by 110 speakers. The multi-training data set also includes 8440 utterances, and a part of the training set is mixed with four kinds of noise, i.e. subway, babble, car, and exhibition noises, at SNRs of 5, 10, 15, 20 dB. The channel characteristics were identical for both the simulated data set in CENSREC-1-C and the training sets in CENSREC-1 (G.712). The HMMs were trained and ASR was performed with HMM Toolkit (HTK). For the simulated data set, the ASR performance was measured in terms of the average word accuracies for SNRs of 0–20 dB. This evaluation criterion is identical to that employed for AURORA-2 and CENSREC-1. For the real recorded data set, word accuracies were calculated for each noise and sound level combination.

To evaluate the robustness, we compared the ASR performance obtained using the proposed SPADE front-end

Table 2
Experimental results evaluating the robustness of front-end processing with CENSREC-1-C. (Above) Results with simulated noisy speech data in CENSREC-1-C. The results were obtained from MFCC-based front-end (ETSI ES 202 050, 2001; ETSI ES 202 212, 2003), SPADE front-end (SPADE FE) (Ishizuka and Nakatani, 2006), SPADE FE with PARADE based noise reduction (NR), and SPADE FE with PARADE based NR and frame dropping (FD). These are the average word accuracies for SNRs of 0–20 dB for each noise condition and all test sets (average). (Bottom) Results obtained with real recorded noisy speech data in CENSREC-1-C. These are average word accuracies under high and low SNR conditions for each noise condition and all test sets (average).

	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Average (%)
ASR word accuracy (%): simulated data, clean training condition									
MFCC	51.30	51.63	48.79	49.74	45.20	56.05	54.17	45.45	50.29
ETSI ES 202 050	55.78	60.70	82.00	62.22	44.21	73.91	65.66	64.79	63.66
ETSI ES 202 212	55.27	56.28	81.29	60.25	39.51	73.97	62.57	62.95	61.51
SPADE Front-end	77.25	48.86	79.41	63.73	41.53	75.75	62.39	65.28	64.28
+PARADE NR	73.53	57.24	78.18	64.54	49.19	74.50	64.53	64.41	65.76
+PARADE FD	**81.09**	**73.88**	**82.33**	**73.69**	**64.23**	**78.67**	**69.73**	**70.16**	**74.22**
ASR word accuracy (%): simulated data, multi-condition training condition									
MFCC	90.68	**87.95**	87.18	90.74	59.01	71.07	69.70	63.92	77.53
ETSI ES 202 050	76.51	75.30	90.96	88.83	53.29	76.14	65.77	68.81	74.45
ETSI ES 202 212	78.75	74.79	90.36	87.34	50.97	77.15	65.52	68.44	74.17
SPADE Front-end	91.56	77.89	91.83	90.55	63.74	85.39	69.98	77.83	81.10
+PARADE NR	91.80	82.64	**92.33**	**91.11**	63.88	**88.00**	70.14	76.11	82.01
+PARADE FD	**92.84**	83.71	91.40	90.43	**70.28**	86.82	**71.90**	**79.09**	**83.31**

	Student's cafeteria		Roadside of freeway		Average (%)
	SNR high	SNR low	SNR high	SNR low	
ASR word accuracy (%): real recorded data, clean training condition					
MFCC	45.17	1.28	34.43	25.23	26.53
ETSI ES 202 050	49.18	−61.66	59.93	38.16	21.40
ETSI ES 202 212	41.26	−78.87	55.19	33.06	12.66
SPADE Front-end	64.85	−18.94	44.44	30.24	30.15
+PARADE NR	**71.13**	0.46	59.84	37.80	42.31
+PARADE FD	69.95	**6.28**	**84.06**	**61.20**	**55.37**
ASR word accuracy (%): real recorded data, multi-condition training condition					
MFCC	72.50	15.48	46.36	23.32	39.42
ETSI ES 202 050	56.10	−96.54	−1.18	−31.15	−18.19
ETSI ES 202 212	53.19	−93.35	−6.47	−31.69	−19.58
SPADE Front-end	80.33	36.61	63.48	46.45	56.72
+PARADE NR	**83.42**	**45.72**	72.40	56.92	64.61
+PARADE FD	74.77	23.50	**88.43**	**81.42**	**67.03**

with PARADE for noise reduction and frame dropping, the SPADE front-end with PARADE for noise reduction (without frame dropping), the SPADE front-end without PARADE (Ishizuka and Nakatani, 2006), ETSI ES 202 050 (2001), ETSI ES 202 212 (2003), and a conventional MFCC feature extraction technique (no noise reduction, channel compensation, or frame dropping). As shown in Fig. 10, the functions of the components in ETSI ES 202 050 are comparable to those in the proposed front-end.

Fig. 13 and Table 2 show the ASR results. These results confirmed that SPADE front-end without PARADE achieves comparable performance to that achieved by ETSI ES 202 050 and ETSI ES 202 212, particularly under a clean training condition, whereas SPADE front-end with PARADE has significantly better ASR performance than the other front-ends. The result confirmed that using PARADE for noise estimation could substantially improve on the ASR performance obtained with the SPADE front-end, and that using PARADE for frame dropping could improve the ASR performance, particularly at low SNRs. The proposed front-end could improve average word accuracies for the simulated data by 10% under a clean training condition. However, it should be noted that PARADE for frame dropping also slightly degrades the ASR performance at high SNRs because the frame dropping increases the deletion errors in ASR. With clean training, ETSI ES 202 050 and ETSI 202 212 substantially improved the word accuracies compared with the MFCC front-end, whereas they degraded its ASR performance with multi-condition training in spite of the fact that ETSI ES 202 050 and ETSI 202 212 could achieve word accuracies of over 90% for CENSREC-1 with multi-condition training (Nakamura et al., 2005). This poor performance resulted from frame dropping failures that led to deletion of a large number of false alarms in the pause periods between the digit utterances. This fact suggests that the MFCC feature extraction in ETSI ES 202 050 and ETSI ES 202 212 may not be suitable for speech data that includes long pauses between utterances. These results also suggest that ASR evaluations using practical long speech data are needed to evaluate the front-ends for practical use. By contrast, the proposed front-end and the SPADE front-end without PARADE achieved superior ASR performance with both clean and multi-condition training. This may be due to the robustness of the PARADE and SPADE feature extraction. Even with multi-condition training, PARADE based VAD improved the ASR performance of the SPADE front-end.

For a real recorded data set, by utilizing PARADE the proposed front-end improved the ASR performance in terms of average word accuracies. As with the simulated data set, ETSI front-ends were able to improve the ASR performance for a clean training condition except under the low SNR condition in the student cafeteria, whereas the performance was degraded with multi-condition training. This also results from the large number of false alarms that occurred during non-speech periods. The results indicate that the proposed method is also robust as regards real recorded speech data in a noisy environment, which may include the phenomena that occur in real recorded data such as the Lombard reflex (Summers et al., 1988; Junqua, 1993).

5. Conclusion

This paper proposed a feature for VAD based on the power ratios of the periodic and aperiodic components of observed signals. Assuming a relaxed condition where the average power of the aperiodic components at the frequencies of the dominant harmonic components is equal to that over the whole frequency range, the proposed feature can reveal voice activity more precisely than HNR, which requires a stricter noise condition. Experiments with large speech data in the presence of various environmental sounds confirmed that the proposed method (PARADE) could achieve superior VAD performance to that obtained with conventional methods. In addition, an ASR experiment utilizing CENSREC-1-C confirmed that the combination of PARADE with the SPADE front-end could achieve high ASR performance in the presence of real environmental noise. This fact also indicates the importance of robust VAD for noise robust ASR front-end processing.

Although PARADE is indeed robust as regards noise as shown above, because PARADE utilizes periodicity as a measure of speech presence, it fails to detect unvoiced speech parts. In addition, PARADE is not robust as regards noise that includes periodic sounds such as background music owing to its inherent mechanism. These problems can be solved by utilizing multiple acoustic features with the PAR (Fujimoto et al., 2008). An investigation of effective mutually complementary combinations of multiple acoustic features is our future work.

Acknowledgements

We thank the members of the Signal Processing Research Group at NTT Communication Science Laboratories for their useful suggestions regarding this research. We also thank the members of the Spoken Dialogue System Group at NTT Cyber Space Laboratories for their valuable comments on practical applications. This research employed the noisy speech recognition evaluation environment CENSREC-1 and CENSREC-1-C produced and distributed by the IPSJ SIG-SLP Noisy Speech Recognition Working Group. The ASR experiments using CENSREC-1 and CENSREC-1-C described in this paper employed HMM Toolkit (HTK) developed by the Speech Vision and Robotics Research Group of Cambridge University Engineering Department.

Appendix A

This appendix describes a method for esti- mating the power of a sinusoidal component in the frequency region, which is utilized in the decomposition in PARADE. Let us consider the following sinusoidal component $x_s(n)$:

$$x_s(n) = r\cos(\psi(n)) \quad \text{s.t. } \psi(n) = \frac{2\pi f}{f_s}n + \phi, \tag{A.1}$$

where f_s, n, r, f, and ϕ are the sampling frequency, sampling index, amplitude, frequency, and initial phase of the sinusoid. We assume that f is included in the frequency range where ordinary speech signals exist. Here, we apply a band-pass filter $h(n) = g(n)\exp(j2\pi fn/f_s)$ to $x_s(n)$ as

$$\begin{aligned}\hat{x}_s(n) &= \sum_{l=0}^{L-1} x_s(n-l)h(l) \\ &= \frac{r}{2}\exp(j\psi(n))\sum_{l=0}^{L-1} g(l) + \frac{r}{2}\exp(-j\psi(n)) \\ &\quad \times \sum_{l=0}^{L-1} g(l)\exp\left(j\frac{4\pi f}{f_s}l\right),\end{aligned} \tag{A.2}$$

where L is the temporal length of the sampling points of the band-pass filter. When we assume that $g(n)$ is a low-pass filter corresponding to the analysis window for an STFT, the second term of Eq. (A.2) can be disregarded. Therefore, we can obtain the following equation

$$r = \frac{2|\hat{x}_s(n)|}{\sum_{n=0}^{L-1} g(n)}. \tag{A.3}$$

Let $g^*(n) = g(L-1-n)$, which is a symmetric temporal window of $g(n)$, and $l' = (L-1) - l$, then $\hat{x}_s(n)$ can be rewritten as

$$\begin{aligned}\hat{x}_s(n) &= \exp\left(j\frac{2\pi f}{f_s}(L-1)\right)X_s(i,k), \\ X_s(i,k) &= \sum_{l'=0}^{L-1} g^*(l')x_s(l' + (n-(L-1))) \\ &\quad \times \exp\left(-j2\pi\left(\frac{f}{f_s}L\right)l'/L\right),\end{aligned} \tag{A.4}$$

where $X_s(i,k)$ is an STFT representation of $x_s(n)$ at a frequency bin of $k = L(f/f_s)$ analyzed by a temporal frame whose index is n beginning at $n_L = n - (L-1)$. Furthermore, the short temporal power $\rho_s(i)$ of $x_s(n)$ can be calculated as follows

$$\begin{aligned}\rho_s(i) &= \sum_{n=0}^{L-1}(g(n)x_s(n-n_L))^2 \\ &= \frac{r^2}{2}\sum_{n=0}^{L-1} g(n)^2 + \frac{r^2}{2}\sum_{n=0}^{L-1} g(n)^2\cos(2\psi(n-n_L)).\end{aligned} \tag{A.5}$$

Because $g(n)^2$ can be considered a low-pass filter similar to $g(n)$, the second term of Eq. (A.5) can also be disregarded. Therefore, the short temporal power of the sinusoid $x_s(n)$ can be calculated as follows using Eqs. (A.3) and (A.4)

$$\rho_s(i) = \eta|X_s(i,k)|^2 \quad \text{where} \quad \eta = \frac{2\sum_{n=0}^{L-1} g(n)^2}{\left(\sum_{n=0}^{L-1} g(n)\right)^2}. \tag{A.6}$$

Appendix B

To cope with the negative value estimations of $\hat{\rho}_p(i)$ and $\hat{\rho}_a(i)$ obtained with Eqs. (12) and (13), we introduce a small positive value ε to ensure that they possess positive values. If we first estimate $\hat{\rho}_a(i)$ before estimating $\hat{\rho}_p(i)$,

$$\begin{cases}\hat{\rho}_a(i) = \rho(i) - \varepsilon \\ \hat{\rho}_p(i) = \varepsilon\end{cases} \quad \text{if } \hat{\rho}_a(i) \geqslant \rho(i). \tag{B.1}$$

On the other hand, if we first estimate $\hat{\rho}_p(i)$ before estimating $\hat{\rho}_a(i)$,

$$\begin{cases}\hat{\rho}_p(i) = \rho(i) - \varepsilon \\ \hat{\rho}_a(i) = \varepsilon\end{cases} \quad \text{if } \hat{\rho}_p(i) \geqslant \rho(i). \tag{B.2}$$

We used $\varepsilon = 1$ for the quantization of the waveform with a 16-bit integer in this paper.

References

Adami, A., Burget, L., Duponi, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S., 2002. QUALCOMM-ICSI-OGI features for ASR. In: Proc. Interspeech, pp. 21–24.

Agarwal, A., Cheng, Y.-M., 1999. Two-stage Mel-warped Wiener filter for robust speech recognition. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 67–70.

Ahmadi, S. and Spanias, A.S., 1999. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. IEEE Trans. on Speech and Audio Process. 7, 333-338.

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

Basu, S., 2003. A linked-HMM model for robust voicing and speech detection. In: Proc. ICASSP, Vol. 1, pp. 816–819.

Benítez, C., Burget, L., Chen, B., Dupont, S, Garudadri, H., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S., 2001. Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks. In: Proc. Interspeech, pp. 429–432.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. Institute of Phonetic Sciences, Vol. 17, pp. 97–110.

Chang, J.-H., Kim, N.S., Mitra, S.K., 2006. Voice activity detection based on multiple statistical models. IEEE Trans. Signal Process. 54, 1965–1976.

Chen, C.P., Filali, K., Bilmes, J.A., 2002. Frontend post-processing and backend model enhancement on the AURORA 2.0/3.0 databases. In: Proc. Interspeech, pp. 241–244.

Cho, Y.D., Kondoz, A., 2001. Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Process. Lett. 8, 276–278.

Cournapeau, D., Kawahara, T., 2007. Evaluation of real-time voice activity detection based on high order statistics. In: Proc. of Interspeech, pp. 2945–2948.

Davis, A., Nordholm, S., Togneri, R., 2006. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. IEEE Trans. Audio Speech Lang. Process. 14, 412–424.

de la Torre, Á., Ramírez, J., Benítez, C., Segura, J.C., García, L., Rubio, A.J., 2006. Noise robust model-based voice activity detection. In: Proc. Interspeech, pp. 1954–1957.

Deshmukh, O., Espy-Wilson, C.Y., Salomon, A., Singh, J., 2005. Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech. IEEE Trans. Speech Audio Process. 13, 776–786.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoustic Speech Signal Process. ASSP-32, 1109–1121.

ETSI EN 301 708, 1999. Digital cellular telecommunications systems (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description (GSM 06.94 version 7.1.1 Release 1998). V7.1.1.

ETSI ES 202 050, 2001. Speech processing, transmission and quality aspects (STQ). Distributed speech recognition. Advanced front-end feature extraction algorithm. Compression algorithms. V1.1.1.

ETSI ES 202 212, 2003. Distributed speech recognition. Extended advanced front-end feature extraction algorithm. Compression algorithms. Back-end speech reconstruction. V1.1.1.

ETSI TS 101 707, 2000. Digital cellular telecommunications system (Phase 2+). Discontinuous Transmission (DTX) for Adaptive Multi-Rate (AMR) speech traffic channels (3GPP TS 06.93 version 7.5.0 Release 1998). V7.5.0.

Evangelopoulos, G., Maragos, P., 2006. Multiband modulation energy tracking for noisy speech detection. IEEE Trans. Audio Speech Lang. Process. 14, 2024–2038.

Fisher, E., Tabrikian, J., Dubnov, S., 2006. Generalized likelihood ratio test for voiced–unvoiced decision in noisy speech using the harmonic model. IEEE Trans. Audio Speech Lang. Process. 14, 502–510.

Fujimoto, M., Ishizuka, K., 2008. Noise robust voice activity detection based on switching Kalman filter. IEICE Trans. Inf. Syst. E91-D, 467–477.

Fujimoto, M., Ishizuka, K., Nakatani, T., 2008. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. In: Proc. ICASSP, pp. 4441–4444.

Górriz, J.M., Ramírez, J., Puntonet, C.G., Segura, J.C., 2006. Generalized LRT-based voice activity detector. IEEE Signal Process. Lett. 13, 636–639.

Griffin, D., Lim, J.S., 1988. Multiband excitation vocoder. IEEE Trans. Acoustic Speech Signal Process. ASSP-36, 1223–1235.

Hamada, M., Takizawa, Y., Norimatsu, T., 1990. A noise robust speech recognition system. In: Proc. ICSLP, pp. 893–896.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2, 578–589.

Hess, W., 1983. Pitch Determination of Speech Signals. Springer-Verlag, New York.

Hillenbrand, J., 1987. A methodological study of perturbation and additive noise in synthetically generated voice signals. J. Speech Hearing Res. 30, 448–461.

Hirsh, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA Tutorial and Research Workshop on Automatic Speech Recognition, pp. 181–188.

Ishizuka, K., Nakatani, T., 2006. A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. Speech Commun. 48, 1447–1457.

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

Itakura, F., Saito, S., 1968. Analysis synthesis telephony based on the maximum likelihood method. In Report. Int. Cong. Acoust., Vol. C-5-5.

ITU-T Recommendation G.729 Annex B, 1996. A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70.

Jackson, P.J.B., Shadle, C.H., 2001. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. IEEE Trans. Speech Audio Process. 9, 713–726.

Jackson, P.J.B., Moreno, D.M., Russell, M.J., Hernando, J., 2003. Covariation and weighting of harmonically decomposed streams for ASR. In: Proc. Interspeech, pp. 2321–2324.

Junqua, J.C., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93, 510–524.

Junqua, J.C., Mak, B., Reaves, B., 1994. A robust algorithm for word boundary detection in the presence of noise. IEEE Trans. Speech Audio Process. 2, 406–412.

Karray, L., Martin, A., 2003. Towards improving speech detection robustness for speech recognition in adverse conditions. Speech Commun. 40, 261–276.

Kato Solvang, H., Ishizuka, K., Fujimoto, M., 2008. Voice activity detection based on adjustable linear prediction and GARCH models. Speech Commun. 50, 476–486.

Kingsbury, B., Saon, G., Mangu, L., Padmanabhan, M., Sarikaya, R., 2002. Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system. In: Proc. ICASSP, Vol. 1, pp. 53-56.

Kitaoka, N., Yamamoto, K., Kusamizu, T., Nakagawa, S., Yamada, T., Tsuge, S., Miyajima, C., Nishiura, T., Nakayama, M., Denda, Y., Fujimoto, M., Takiguchi, T., Tamura, S., Kuroiwa, S., Takeda, K., Nakamura, S., 2007. Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 607–612.

Kristjansson, T., Deligne, S., Olsen, P., 2005. Voicing features for robust speech detection. In: Proc. Interspeech, pp. 369–372.

Krom, G., 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Hearing Res. 36, 254–266.

Lamel, L.F., Rabiner, L.R., Rosenberg, A.E., Wilpon, J.G., 1981. An improved endpoint detector for isolated word recognition. IEEE Trans. Acoustic Speech Signal Process. ASSP-29, 777–785.

Laroche, J., Stylianou, Y., Moulines, E., 1993. HNS: speech modification based on a harmonic + noise model. In: Proc. ICASSP, Vol. 1, pp. 550–553.

Le Bouquin-Jeannès, R., Faucon, G., 1995. Study of voice activity detector and its influence on a noise reduction system. Speech Commun. 16, 245–254.

Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H., Shikano K., 2004. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In: Proc. Interspeech, Vol. 1, pp. 173–176.

Li, Q., Soong, F.K., Siohan, O., 2001. An auditory system-based feature for robust speech recognition. In: Proc. Interspeech, pp. 619–622.

Li, K., Swamy, M.N.S., Ahmad, M.O., 2005. An improved voice activity detection using higher order statistics. IEEE Trans. Speech Audio Process. 13, 965–974.

Macho, D., Mauuary, L., Noé, B., Cheng, Y.M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of a noise-robust DSR front-end on AURORA databases. In: Proc. Interspeech, pp. 17–20.

Mak, B., Junqua, J.-C., Reaves, B., 1992. A robust speech/non-speech detection algorithm using time and frequency-based features. In: Proc. ICASSP, Vol. 1, pp. 269–272.

Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. 10, 109–118.

Mauuary, L., 1998. Blind equalization in the cepstral domain for robust telephone based speech recognition. In: Proc. EUSIPCO, Vol. 1, pp. 359–362.

Mousset, E., Ainsworth, W.A., Fonollosa, J.A.R., 1996. A comparison of several recent methods of fundamental frequency and voicing decision estimation. In: Proc. ICSLP, Vol. 2, pp. 1273–1276.

Nakamura, A., Matsunaga, S., Shimizu, T., Tonomura, M., Sagisaka, Y., 1996. Japanese speech databases for robust speech recognition. In: Proc. ICSLP, Vol. 4, pp. 2199–2202.

Nakamura, S., Yamamoto, K., Takeda, K., Kuroiwa, S., Kitaoka, N., Yamada, T., Mizumachi, M., Nishiura, T., Fujimoto, M., Saso, A., Endo, T., 2003. Data collection and evaluation of AURORA-2

Japanese corpus. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 619–623.

Nakamura, S., Takeda, K., Yamamoto, K., Yamada, T., Kuroiwa, S., Kitaoka, N., Nishiura, T., Sasou, A., Mizumachi, M., Miyajima, C., Fujimoto, M., Endo, T., 2005. AURORA-2J: an evaluation framework for Japanese noisy speech recognition. IEICE Trans. Inf. Syst. E88-D, 535–544.

Nakatani, T., Irino, T., 2004. Robust and accurate fundamental frequency estimation based on dominant harmonic components. J. Acoust. Soc. Am. 116, 3690–3700.

Nakatani, T., Amano, S., Irino, T., Ishizuka, K., Kondo, T., 2008. A method for fundamental frequency estimation and voicing decision: application to infant utterances recorded in real acoustical environments. Speech Commun. 50, 203–214.

Nemer, E., Goubran, R., Mahmoud, S., 2001. Robust voice activity detection using higher-order statistics in the LPC residual domain. IEEE Trans. Speech Audio Process. 9, 217–231.

Noé, B., Sienel, J., Jouvet, D., Mauuary, L., Boves, L., de Veth, J., de Wet, F., 2001. Noise reduction for noise robust feature extraction for distributed speech recognition. In: Proc. Interspeech, pp. 433–436.

Pearce, D., Hirsh, H.G., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. In: Proc. ICSLP, Vol. 4, pp. 29–32.

Rabiner, L.R., 1977. On the use of autocorrelation analysis for pitch detection. IEEE Trans. Acoust. Speech Signal Process. 25, 24–33.

Rabiner, L.R., Sambur, M.R., 1975. An algorithm for determining the endpoints of isolated utterances. Bell Syst. Tech. J. 54, 297–315.

Ramírez, J., Segura, J.C., Benítez, C., de la Torre, Á., Rubio, A., 2004. Efficient voice activity detection algorithms using long-term speech information. Speech Commun. 42, 271–287.

Ramírez, J., Segura, J.C., Benítez, C., García, L., Rubio, A., 2005. Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process. Lett. 12, 689–692.

Ramírez, J., Segura, J.C., Górriz, J.M., 2007. Revised contextual LRT for voice activity detection. In: Proc. ICASSP, Vol. 4, pp. 801–804.

Richard, G., d'Alessandro, C., 1996. Modification of the aperiodic component of speech signals for synthesis. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Text-to-Speech Synthesis. Springer-Verlag, New York, pp. 41–56.

Savoji, M.H., 1989. A robust algorithm for accurate endpointing of speech. Speech Commun. 8, 45–60.

Serra, X., Smith, J., 1990. Spectral modeling and synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition. Comput. Music J. 14, 12–24.

Shen, J.-L., Hung, J.-W., Lee, L.-S., 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. In: Proc. ICSLP.

Sohn, J., Kim, N.-S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6, 1–3.

Srinivasan, K., Gersho, A., 1993. Voice activity detection for cellular networks. In: Proc. IEEE Workshop on Speech Coding for Telecommunication, pp. 85–86.

Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M., 1988. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am. 84, 917–928.

Tahmasbi, R., Razaei, S., 2007. A soft voice activity detection using GARCH filter and variance Gamma distribution. IEEE Trans. Audio Speech Lang. Process. 15, 1129–1134.

Tucker, R., 1992. Voice activity detection using a periodicity measure. IEE Proc.-I 139, 377–380.

Wilpon, J.G., Rabiner, L.R., 1987. Application of hidden Markov models to automatic speech endpoint detection. Comput. Speech Lang. 2, 321–341.

Wu, B.-F., Wang, K.-C., 2005. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. IEEE Trans. Speech Audio Process. 13, 762–775.

Yantorno, R.E., Krishnamachari, K.L., Lovekin, J.M., 2001. The spectral autocorrelation peak valley ratio (SAPVR) – a usable speech measure employed as a co-channel detection system. In: Proc. IEEE Internat. Workshop on Intelligence Signal Process.

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