Applied Acoustics 113 (2016) 221–229

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Optimally weighted maximum *a posteriori* probabilities based on minimum classification error for dual-microphone voice activity detection

Seng Hyun Huang, Joon-Hyuk Chang *

School of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea

ARTICLE INFO

Article history:
Received 23 October 2015
Received in revised form 4 April 2016
Accepted 27 June 2016
Available online 7 July 2016

Keywords:
Voice activity detection
Dual-microphone
Discriminative weight training
Minimum classification error

ABSTRACT

The dual-microphone voice activity detection (VAD) technique is proposed by applying discriminative weight training to achieve optimal weighting of spatial features available within the dual-microphone VAD. Since the motivation behind our method is to use the relevant spatial information available from the two microphones, we employ the phase difference, coherence, and power level difference ratio (PLDR) as a feature vector, and then use this feature vector to derive the maximum *a posteriori* (MAP) probabilities. Then, we combine each MAP probability based on a discriminative weight training, i.e., the minimum classification error (MCE) method to offer an optimal VAD decision in a spectral domain, which successfully represents the dynamic evolution of speech over time even in the non-stationary noise environments. The proposed dual-microphone VAD algorithm outperforms conventional dual-microphone VAD methods based on only single feature among the PLDR, phase difference, and spectral coherence.

1. Introduction

Voice activity detection (VAD), whereby speech and non-speech segments in a speech signal are classified, plays an essential role in speech transmission [1], speech recognition, and speech enhancement. For instance, VAD technology is integrated into speech coding systems to suspend their operation in the absence of speech and thus to reduce error rates. Early VAD approaches focused on single-microphone-based algorithms whereas more recently, multiple-microphone-based VAD techniques have been accepted as promising solutions to adverse noise conditions. Traditionally, most VAD algorithms considering a single microphone use only single metric, such methods include the use of the zero-crossing rate (ZCR) [2], linear predictive coding (LPC) parameters [3], MFCCs [4], cepstral coefficients [5], entropy [6], pattern recognition [7], periodicity measures [8] and statistical model approaches [9]. Among them, the statistical-model-based VAD employing the decision-directed (DD) method reported high detection accuracy. The superiority of the statistical-model-based VAD has been verified in most studies in which the likelihood ratio (LR) test is derived given a set of hypotheses [9]. The statistical-model-based VAD was further improved by using the minimum classification error (MCE) algorithm [10], whereby optimally weighted LRs are integrated into the VAD decision. However, these single-channel VAD algorithms do not exhibit acceptable performance especially in non-stationary noise conditions by increasing the number of microphones in the VAD system, further detection accuracy improvement is expected. But, the inclusion of more than two microphones for the mobile device faces serious design difficulties in terms of size, complexity and power consumption. Thus, herein we mainly focus on dual-microphone VAD systems as a trade off.

Notice that one of the dominant dual-microphone VAD techniques adopts a coherence technique, namely the magnitude squared coherence (MSC) method [11,12], which is based on the assumption that the speech signals in the two channels are correlated, but the noisy signals in a diffuse field are relatively uncorrelated. However, in practical situations, the noisy signals are not mutually uncorrelated when the distance between microphones is too short or when the microphones are close to the noise source. Arabi and Shi [13] developed a VAD method in which the time difference of arrival (TDOA) between the input signals of the two microphones were considered. However, this technique cannot estimate the TDOA of the target speech signal when the direction of the noise is identical to that of the target signal. Also, Kim and Cho have studied another technique [14] relying on the phase

* Corresponding author.
E-mail address: jchang@hanyang.ac.kr (J.-H. Chang).

http://dx.doi.org/10.1016/j.apacoust.2016.06.025
0003-682X/

information, this has been used to apply the phase difference to a microphone array to calculate the LRs for VAD. However, in practice, the performance of this technique is sensitive to the estimation error of direction of arrival (DOA) in such a way that more than four microphones are needed.

In addition, a power level difference (PLD) [15] criterion has been developed relies on the fact that speech signals transmitted from the source have different power levels between microphones, whereas the power levels of noise signals are almost equivalent. This technique is effective for highly non-stationary noise, even when the noise source is located in a similar direction as the target speech. However, this method is sensitive to the noise type or noise level because the PLDs of different noise sources are not identical due to differences in acoustical factors, reverberation or azimuth angles from the noise source at each microphone. To mitigate this problem, Choi and Chang [16] presented a novel algorithm, which incorporates the PLD of noise during speech pauses and then proposed the two-step PLD ratio (called PLDR), which is the ratio of the PLDs of speech and noise estimated during noise periods. Especially, the long-term PLDR (LT-PLDR) and short-term PLDR (ST-PLDR) were devised to characterize the long-term evolution and short-term variation of speech, respectively. The two PLDRs are combined into a final decision rule for the reliable VAD performances under various acoustical environments.

In this work, a novel dual-microphone VAD technique is proposed using optimally spatial weighted features including the PLDR, phase difference, and coherence to exploit spatial information able to successfully represent the dynamic evolution of speech. In addition to the PLDR proposed in [16], we consider more spatial features such as the phase difference and coherence and apply the MCE scheme in an attempt to incorporate the different contributions of the spatial features under dynamic acoustic environments. Above all, the maximum *a posteriori* (MAP) probabilities are first obtained from each feature by means of a model-trust minimizing algorithm to classify periods of speech presence and absence. Then, optimal weights are derived by means of the generalized probabilistic descent (GPD) technique based on the contribution of each MAP probability for VAD and are applied to each MAP probability to be optimally adjusted in the final VAD decision rule. The performance of the proposed algorithm is evaluated by means of extensive objective tests, under various acoustic conditions consisting of different noises, with different azimuths and distances between the source and the microphones. The results of a number of experiments demonstrated that the proposed VAD technique combining several models yielded a better performance than models solely utilizing the PLDR, phase difference, or spectral coherence, this superior performance was observed under various acoustical circumstances and input SNRs.

The rest of the paper is organized as follows. Section 2 firstly gives a brief review of traditional VAD algorithms based in the use of two microphones, and Section 3 presents the technique used to derive the optimal weighting based on an MCE scheme. Section 4 describes the experimental setup and results in detail. Conclusions are presented in Section 5.

2. Review of two-microphone VAD techniques

In this section, we first briefly review the notion of two-microphone VAD algorithms as depicted in Fig. 1. Input signals at the two microphones are denoted as $y_i(t) = x_i(t) + n_i(t)$, that is to say, the noisy signal $y_i(t)$ is considered to be the sum of a clean speech signal $x_i(t)$ and a noise signal $n_i(t)$, where i denotes the microphone index and t denotes the sample index. By taking the discrete Fourier transform (DFT) of $y_i(t)$, the equation in the time-frequency domain is obtained as

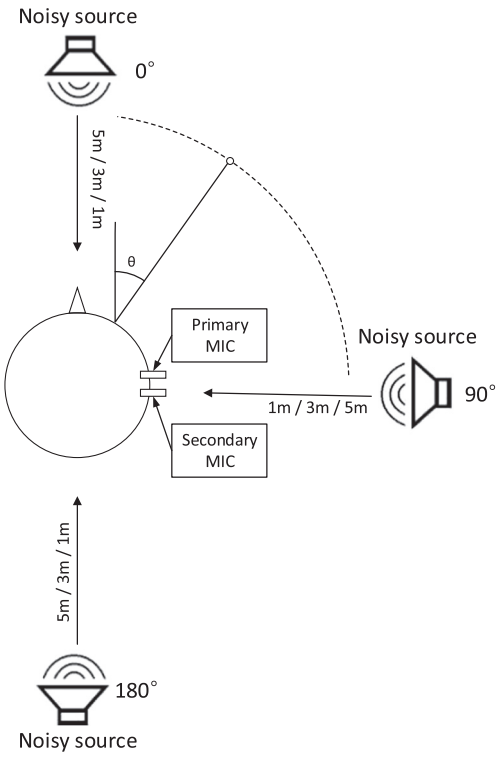

Fig. 1. The location of the sound source about dummy head.

$$Y_i(k,n) = X_i(k,n) + N_i(k,n), \quad i = 1,2 \tag{1}$$

where $k(=0,1,\ldots,K-1)$ is a frequency-bin index and n is a frame index. For two hypotheses, $H_0(k,n)$ and $H_1(k,n)$, which respectively indicate speech absence and presence, we assume that

$$\begin{aligned} H_0(k,n) &: Y_i(k,n) = N_i(k,n) \\ H_1(k,n) &: Y_i(k,n) = X_i(k,n) + N_i(k,n). \end{aligned} \tag{2}$$

2.1. Review of PLDR-based VAD

Speech and noise are assumed to be independent in [16], thus the power spectral density (PSD) of each microphone can be represented as

$$\begin{aligned} P_{Y_1}(k,n) &= P_{X_1}(k,n) + P_{N_1}(k,n) \\ P_{Y_2}(k,n) &= P_{X_2}(k,n) + P_{N_2}(k,n). \end{aligned} \tag{3}$$

The PLD that subtracted from the PSD of the other microphone with an absolute operator is given by

$$\Delta P_Y(k,n) = |P_{Y_1}(k,n) - P_{Y_2}(k,n)|. \tag{4}$$

Based on the PLD obtained for the current frame and the PLD estimated at the noise periods, the long-term PLD (LT-PLD) can be expressed in a first-order recursive way as

$$\widehat{\Delta P}_Y^{\mathrm{LT}}(k,n) = \alpha_{\mathrm{LT}} \widehat{\Delta P}_Y^{\mathrm{LT}}(k,n-1) + (1-\alpha_{\mathrm{LT}}) \Delta P_Y(k,n) \tag{5}$$

where $\alpha_{\mathrm{LT}}(=0.9)$ is a smoothing parameter. Then, the LT-PLDR $Q_{\mathrm{LT}}(k,n)$ can be expressed as

$$Q_{\mathrm{LT}}(k,n) = \frac{\widehat{\Delta P}_Y^{\mathrm{LT}}(k,n)}{\widehat{\Delta P}_N^{\mathrm{LT}}(k,n)} \tag{6}$$

where the PLD of the noise $\widehat{\Delta P}_N^{\mathrm{LT}}(k,n)$ is obtained by using the minima controlled recursive averaging (MCRA) scheme in the period of speech absence as given by