Dual Microphone Voice Activity Detection Exploiting Interchannel Time and Level Differences

Jaehoon Park, Yu Gwang Jin, Soojoong Hwang, Student Member, IEEE, and Jong Won Shin, Member, IEEE

Abstract—The two most important spatial cues in human auditory system may be the inter-aural time difference and the inter-aural level difference. There have been many attempts to utilize the time difference of arrival (TDoA) and level difference between two microphone signals for voice activity detection (VAD). In this letter, we propose a dual microphone VAD algorithm based on a support vector machine (SVM) for which the input vector consists of both TDoA-based and level difference-based features. Several candidates for the feature combination have been compared using various TDoA-related and level difference-related features. Experimental results showed that the proposed VAD algorithm outperformed a standardized single microphone VAD, VADs based on the TDoA or level difference, and logical combination of them in various noise environments.

Index Terms—Dual microphone, interchannel level difference, interchannel time difference, support vector machine, voice activity detection.

I. INTRODUCTION

Voice activity detection (VAD) which determines if the current frame of input contains speech or not is indispensable for successful speech enhancement, rate determination of variable-rate codecs, and speech recognition [1]-[12]. Single channel VAD approaches based on statistical models have demonstrated successful performance for rather stationary noises such as white, babble, factory, street and office noises [1]-[4]. However, the performance of VAD using one microphone is limited for highly nonstationary noises such as competing talkers and background music. These days, devices with multiple microphones including smartphones, tablet PCs and smart TVs have become popular, which has made the multichannel signal processing practical. Voice activity detectors using multiple microphones have shown superior performances as they can exploit spatial information on top of the information from each microphone signal [5]-[9]. In this letter, we focus on the VAD using spatial features. VAD based on other characteristics of the signals could be easily combined later to further enhance the performance.

One of the most studied spatial features is the interchannel time difference (ITD) [5], [6], [13]. For the applications in which the range of the direction-of-arrival (DoA) of the desired speech is known in advance, the VAD based on ITD can be easily constructed by just determining if the estimated ITD corresponds to the DoA within the range or not. The generalized cross correlation method with the phase transform weighting function (GCC-PHAT) [13] may be the most popular time difference of arrival (TDoA) estimation technique. Recently, the long-term information of interchannel phase difference (LTIPD) [5] which evaluates the consistency of the estimated DoA in a certain frequency band over several frames was proposed as a test statistic for an ITD based VAD. Interchannel level difference (ILD) has also been exploited in multichannel VAD and speech enhancement [7], [14]-[16], for the near field case in which speech is much stronger in one microphone than others. The most popular use case may be using smartphones with dual microphones in the handset mode, in which the mouth is located much closer to the primary microphone than the secondary one. In [15], the normalized difference of power spectral density (NDPSD) was proposed in order to estimate the noise power spectral density (PSD) and a proper spectral gain for noise reduction. Recently, a VAD based on the two-step power level difference ratio (PLDR) [7] which utilizes short-term and long-term ILD measures has demonstrated reliable performance under various noise environments including nonstationary noises. There have been a few attempts to utilize both TDoA-based and level difference-based information for VAD [8], [9], but the target DoA was strictly assumed to be broad-side [17], which restricts the application of the algorithms. There have also been efforts to utilize both of the information for the blind source separation [18], [19]. In [20], the cross correlation-based features were used as another multi channel features.

In this letter, we propose a dual microphone VAD algorithm using a support vector machine (SVM) [21] which exploits both the TDoA-based and level difference-based features. Various combinations of TDoA-related and level difference-related features have been tested. From the experiments using the data recorded with a commercial smartphone, we could verify the proposed algorithm outperformed not only the VADs based on TDoA or level difference but also the logical combination of them in the presence of stationary and nonstationary noises.

II. SPATIAL FEATURES FOR DUAL MICROPHONE VAD

For the end-fire configuration of the dual microphones [17], both ITD and ILD information are helpful to determine voice activity. One of the popular applications is the handset mode of the smartphones, for which the primary microphone that is usually at the bottom of the phone is much closer to the mouth than the secondary microphone which is typically located on the back or top edge of the device.
Let $Y_1(n,k)$ and $Y_2(n,k)$ denote the short-time Fourier transform (STFT) coefficients of the received signals at the primary and secondary microphones for the $k$-th frequency bin at frame $n$, respectively. The phase difference $\Delta \psi(n,k)$ is defined as the difference between phases of $Y_1(n,k)$ and $Y_2(n,k)$ as follows:

$$
\Delta \psi(n,k) \triangleq \angle Y_1(n,k) - \angle Y_2(n,k) = \angle \{Y_1(n,k)Y^*_2(n,k)\}
$$

where $\angle X$ and $X^*$ denote the phase and complex conjugate of a complex value $X$, respectively. The GCC-PHAT algorithm proposed in [13] is arguably the most widely used approach for the TDOA estimation. By the GCC-PHAT method, the TDOA $\tau_{PHAT}$ is obtained as

$$
\tau_{PHAT}(n) = \arg \max_{\tau} \gamma_{12}(\tau) \quad (2)
$$

$$
\gamma_{12}(\tau) \triangleq \frac{1}{N} \sum_{k=0}^{N-1} \frac{Y_1(n,k)Y^*_2(n,k)}{|Y_1(n,k)Y^*_2(n,k)|} e^{j 2 \pi \kappa \tau / N}
$$

where $|\cdot|$ denotes the absolute value, and $N$ is the length of the STFT. The estimate of the voice activity for the frame $n$ based on ITD information, $V_{ITD}(n)$, can be determined by comparing $\tau_{PHAT}(n)$ with the lower and upper limits of the target TDOA range, $\tau_L$ and $\tau_H$, i.e.,

$$
V_{ITD}(n) = \begin{cases} 1, & \text{if } \tau_L \leq \tau_{PHAT}(n) \leq \tau_H \, , \\ 0, & \text{otherwise} \end{cases} \quad (3)
$$

In [5], the LTIPD was proposed as another measure reflecting ITD information, which describes the degree of the concentration of the signal energy within a small DoA range. First, the target DoA range is divided into $U$ overlapped sectors with equal width. For each frequency bin, the concentration measure $C_i(n,k)$ is calculated by counting the number of frames among the last $L$ frames for which DoA estimates correspond to the sector $i$, for $i = 1, \ldots, U$. The energy concentrated in the sector $i$, $E_i(n)$, is defined as

$$
E_i(n) = \sum_{k:C_i(n,k) > \kappa_i} |Y_1(n,k)|^2 \quad (4)
$$

where $\kappa_i$ is the threshold of the concentration of the DoAs. Then, another ITD-based VAD can be constructed:

$$
V_{ITD}(n) = \begin{cases} 1, & \text{if } E_M(n) \geq \xi_E \\ 0, & \text{otherwise} \end{cases} \quad (5)
$$

where $M = \arg \max_i E_i(n)$, and $\xi_E$ denotes the threshold of energy concentration.

Another important feature for the dual microphone VAD is ILD. The level difference between two microphone signals reflects how close one microphone is located to the sound source compared with the other microphone, which can be beneficial in near-field scenarios. In [15], the NDPSD $\Delta \Phi(n,k)$ is computed from $\Phi_Y(n,k) = |Y_1(n,k)|^2$ and $\Phi_{Y_2}(n,k) = |Y_2(n,k)|^2$ which are the instantaneous estimates of the PSDs of $Y_1(n,k)$ and $Y_2(n,k)$, respectively, as follows:

$$
\Delta \Phi(n,k) \triangleq \frac{|\Phi_Y(n,k) - \Phi_{Y_2}(n,k)|}{|\Phi_Y(n,k) + \Phi_{Y_2}(n,k)|} \, . \quad (6)
$$

The normalization with the overall input power is introduced for the robustness with respect to the input power variation. Using the NDPSDs for all frequency bands, the estimate of the voice activity based on ILD information, $V_{ILD}(n)$, can be decided as

$$
V_{ILD}(n) = \begin{cases} 1, & \text{if } \frac{2}{N} \sum_{k=1}^{N/2} \Delta \Phi(n,k) \geq \xi_{NDPSD} \\ 0, & \text{otherwise} \end{cases} \quad (7)
$$

where $\xi_{NDPSD}$ denotes a threshold.

[7] proposed the long-term and short-term PLDRs as test statistics for VAD, which are basically the ratios between the input power differences and the noise power differences between microphone signals for each frequency using two different smoothing factors. Let $\Delta P_Y(n,k) \triangleq |\Phi_Y(n,k) - \Phi_{Y_2}(n,k)|$ and $\Delta P_Y(n,k)$ denote the level difference of input PSDs in two microphones for the frame $n$ and frequency $k$ and the smoothed version of it, respectively. The PLDR $Q(n)$ is defined as follows:

$$
Q(n) = \frac{2}{N} \sum_{k=1}^{N/2} \log \left( \frac{\Delta P_Y(n,k)}{\Delta P_N(n,k)} \right) \quad (8)
$$

where $\Delta P_N(n,k)$ represents the smoothed version of the difference of the noise PSDs in two microphone signals estimated in a way similar to the minima-controlled recursive averaging (MCRA) [22]. Two different smoothing factors are used to capture both long-term and short-term information. Final voice activity decision is made by comparing the $a$ posteriori probabilities $p_L(n)$ and $p_S(n)$, which are modeled as logistic functions [21] of the long-term and short-term PLDRs, $Q_L(n)$ and $Q_S(n)$, respectively, with thresholds $\xi_L$ and $\xi_S$ as follows:

$$
V_{ILD}(n) = \begin{cases} 1, & \text{if } p_L(n) \geq \xi_L \text{ and } p_S(n) \geq \xi_S \\ 0, & \text{otherwise} \end{cases} \quad (9)
$$

$V_{ITD}$ or $V_{ILD}$ results can be enhanced by introducing a hangover scheme which requires several consecutive instantaneous decisions of 0 to make the final decision 0.

III. VAD EXPLOITING ITD AND ILD

Most of VAD approaches using dual microphones including the aforementioned methods have shown better performances than those using a single microphone as they can exploit spatial diversity. However, ITD-based VADs have an obvious weakness to the directional noise coming from the similar direction to the target speech since they solely rely on the DoA. ILD-based VADs could discriminate the same direction noise as long as the target speaker is in the near-field and the noise is from far-field, but in general it is hard to consolidate a threshold which can deal with various DoAs of noises, noise types, and signal-to-noise ratios (SNRs). Even though a few approaches have been studied to incorporate the ITD-related and ILD-related information at once [8], [9], the strict assumption that the target acoustic source is located at the broad-side of the microphones limits the applications.
The simplest way to consider both the ITD and ILD information simultaneously is logical combination of two VADs from different methods. The logical combination using ‘AND’ or ‘OR’ operation can make the VAD results more precise and robust with extremely low computational complexity. Combining the ILD-based VADs such as (7) and (9) and the ITD-based ones like (3) and (5), the VADs can be given by

\[
V_{\text{AND}}(n) = \begin{cases} 
1, & \text{if } V_{\text{ILD}}(n) = 1 \text{ and } V_{\text{ITD}}(n) = 1 \\
0, & \text{otherwise}
\end{cases}
\]

\[
V_{\text{OR}}(n) = \begin{cases} 
1, & \text{if } V_{\text{ILD}}(n) = 1 \text{ or } V_{\text{ITD}}(n) = 1 \\
0, & \text{otherwise}
\end{cases}
\]

It is noted that the thresholds and numbers of hangover frames for the individual VADs, ‘AND’-combination and ‘OR’-combination can be different.

The SVM is one of the most widely used classifiers based on machine learning approach and has shown impressive performance in a variety of tasks [21]. It was also successfully applied to summarize statistics for VAD [3]. The SVM classifier builds optimal hyperplanes that maximize the margin between different classes. In this work, we adopt a linear SVM classifier in order to detect frames which contain the target speech components, by using both ITD-based and ILD-based features as the input vectors. Given training data consisting of feature vectors \(x(n)\) which contain the ITD- and ILD-related components introduced in Section II and corresponding class labels \(z(n) \in \{+1, -1\}\), the equation for the hyperplane is given as \(\langle w, x \rangle + b = 0\) where \(w\) is the weight vector, \(b\) is the bias, and \(\langle \cdot, \cdot \rangle\) denotes the inner product. After finding the optimized weight \(w_o\) and the bias \(b_o\) through SVM training, the VAD based on the linear SVM for an input vector \(\tilde{x}(n)\) is given according to the sign of

\[
y(\tilde{x}(n)) = \langle w_o, \tilde{x}(n) \rangle + b_o.
\]

In [23], the \textit{a posteriori} probability that the input belongs to one class is modeled as a logistic sigmoid function of the output of the SVM given by

\[
p(V_{\text{SV}} = 1 | \tilde{x}(n)) = \frac{1}{1 + \exp(Ay(\tilde{x}(n)) + B)}
\]

and the final VAD which allows the control of the trade-off between false rejection rate (FRR) and false acceptance rate (FAR) can be constructed as

\[
V_{\text{SV}}(n) = \begin{cases} 
1, & \text{if } p(V_{\text{SV}} = 1 | \tilde{x}(n)) \geq \xi_{\text{SV}}, \\
0, & \text{otherwise}
\end{cases}
\]

where \(A\) and \(B\) are the parameters fitted in the process of the training of the SVM. Contrary to the logical combinations which utilize the individual VAD results, the SVM-based VAD uses the ITD- and ILD-based features directly for the final decision. Two different kinds of features can provide complementary information and may improve the performance over various noisy conditions. It is also noted that an additional hangover scheme can be used after the decision made by SVM. The best combination of the ITD- and ILD-related features for the proposed SVM-based VAD is decided through the experiments which will be explained in the next section.

The best combination of the ITD- and ILD-related features for the proposed SVM-based VAD is decided through the experiments which will be explained in the next section.

IV. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed approach in realistic conditions, we have recorded spoken utterances and noise signals with a commercial smartphone, Samsung Galaxy S5, SM-900K, which has two microphones about 150 mm away from each other. One male person stood in the center of a reverberant room with the size of 3119 × 3232 × 2080 mm³ and the reverberation time of \(RT_{60} = 120\,ms\), and held a phone with the right hand, exactly in the same way as in a usual telecommunication scenario in the handset mode. Various Korean sentences were spoken by the person and recorded using the device under noise-free condition. Competing talks in Korean by another male participant were recorded as directional interference from four different locations 45°/135°/225°/315° at the distance of 1000 mm as illustrated in Fig. 1 while the person holding the phone remained silent. The diffused noise fields were also simulated using four loudspeakers in Fig. 1 located at the corners of the room. All the loudspeakers face corners instead of the desired speaker to simulate more diffuse noise through complex reflections. The white, babble and car noises from the NOISEX-92 database [24] were replayed at 4 loudspeakers simultaneously. Each signal was sampled at 8 kHz, and the Hamming window of 256 samples with the frame shift of 10 ms was applied to analyze the signals.

To train the SVM and find optimal parameters for competing VAD approaches, training data was constructed with 480 seconds of speech corrupted by directional interferences from 4 different directions and 360 seconds of speech mixed with diffuse noise at -5, 0, 5, 10, 15, and 20 dB SNR. A popular SVM library LIBSVM [25] was used for the training and the testing with the SVM classifier. In order to evaluate the performance of each VAD approach, the overall error rate \(E_{\text{OVR}}\) which is the weighted combination of the FRR and the FAR given by

\[
E_{\text{OVR}} = (\alpha \times \text{FRR}) + \{(1 - \alpha) \times \text{FAR}\}
\]

is computed where \(0 \leq \alpha \leq 1\) is a factor that controls the trade-off between the FRR and FAR. In this work, we set...
In this letter, we have proposed an SVM-based dual microphone VAD technique exploiting both the ITD and ILD information. Compared with logical combinations of ITD- and ILD-based VAD results, the proposed approach could combine the TDoA-related and level difference-related statistics more flexibly. Various combinations of ILD-related and ITD-related features have been compared to find the best pair of information. Experimental results for various types of noises and SNRs demonstrated that the proposed method outperformed the individual ITD- and ILD-based VADs, the ‘AND’ combination of them, and the AMR VAD option 2 over most of the noise conditions. The performance improvement made by the proposed algorithm may be due to the flexible combination of features which reflect different spatial characteristics of the signals.

V. Conclusions

To verify the performance of the proposed VAD algorithm, it is compared with the performances of the LTIPD-based VAD in (5), the two-step PLDR in (9), the logical ‘AND’ combination of the LTIPD-based VAD and NDPSD-based VAD in (7), and the SVM-based VAD using the aforementioned combination of ITD- and ILD-related features. In addition, the performance of 3GPP TS 26.104 adaptive multi-rate (AMR) VAD option 2 [10] which is a standardized single-microphone VAD was also compared. The test material consisted of 4 directional noises of 20 seconds each and three diffused noises of 20 seconds each mixed with speech at -5 dB to 20 dB SNR with 5 dB steps, which made the total length of the test data 840 seconds. Both the speech and noise were not included in the training database, but from the same speakers. Fig. 2 shows the receiver-operating-characteristics curves for the whole test data. It is clear that the proposed SVM-based combination of the ITD- and ILD-related features brought about significant performance improvement over a simple logical combination of ITD- and ILD-related VADs. Fig. 3 illustrates the performance of VADs for each noise type averaged over all SNRs. As can be seen from Fig. 3, the proposed SVM-based VAD outperformed the individual ITD- and ILD-based VADs, the ‘AND’ combination of them, and the AMR VAD option 2 overall most of the noise conditions. The performance improvement made by the proposed algorithm may be due to the flexible combination of features which reflect different spatial characteristics of the signals.

V. Conclusions

In this letter, we have proposed an SVM-based dual microphone VAD technique exploiting both the ITD and ILD information. Compared with logical combinations of ITD- and ILD-based VAD results, the proposed approach could combine the TDoA-related and level difference-related statistics more flexibly. Various combinations of ILD-related and ITD-related features have been compared to find the best pair of information. Experimental results for various types of noises and SNRs demonstrated that the proposed method outperformed the AMR VAD option 2, VADs based on individual spatial features and the logical combination of them. Although we focus on VADs based on spatial features in this letter, the performance can be further enhanced if features based on other properties of the signals are incorporated on top of the spatial features.
REFERENCES