A TWO-MICROPHONE BASED VOICE ACTIVITY DETECTION FOR DISTANT-TALKING SPEECH IN WIDE RANGE OF DIRECTION OF ARRIVAL

Yanmeng Guo, Kai Li, Qiang Fu, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences. Beijing 100190, China

ABSTRACT

In this paper, a two-microphone based voice activity detection (VAD) algorithm is proposed to detect the distant-talking speech coming randomly from a wide range of direction of arrival (DOA). The long-term information of inter-channel phase difference (LTIPD) is introduced as a target speech existence measure, which describes the concentration degree of DOA estimations on a sound source with harmonic structure. The proposed algorithm performs robustly on distant-talking speech recorded in several real environments.

Index Terms— Voice activity detection, inter-channel phase difference, direction of arrival, harmonic structure

1. INTRODUCTION

Voice activity detection (VAD) plays an important role in human-machine speech interfaces. The single-channel VADs discriminate speech based on time-frequency (T-F) information, and they are practical for close-talking applications. However, for the distant-talking applications, the single-channel VADs are usually unreliable because of the environment noises, speech attenuation and reverberation. In those conditions, the multi-channel VADs are usually needed, because they can exploit the spatial information.

The multi-channel algorithms extract the spatial information based on a priori knowledge of array geometry and microphone properties. The basic approaches include spatial filtering, correlation and time-delay estimation. Spatial filtering VADs are precise in spatial domain, and they are widely used in the cases with presteered DOA[1][2][3]. However, it will become unreliable or complicated for speech detection for a wide DOA range. The VADs based on correlation[4] or homogeneity[5] are good at detecting directional sound sources, but they cannot discriminate sources from different DOA. The VADs based on time-delay estimation usually get the time-difference-of-arrival (TDOA) through correlation[6] or phase[7], but they are only sensitive to the target speech from a predefined narrow DOA range.

This work is partially supported by the National Natural Science Foundation of China (NO.10925419, 90920302, 10874203, 60875014, 61072124, 11074275)

However, there are still a lot of speech applications, such as the interactive TV, hands-free mobile phone and humanoids, in which the target speech may come randomly from a wide DOA range. The VADs for predefined DOA are not applicable for these applications, while the VADs without target DOA limit are also inappropriate. The sector-based algorithms[2] partition the physical space into several connected sectors to discriminate sound sources from different DOA, but the target DOA range is also narrow.

The inter-channel phase difference (IPD) is a sensitive indicator for DOA, but its reliability is highly related to the local signal-to-noise ratio (SNR). It is also annoyed by the phase-wrapping problem and the non-linear mapping between IPD and TDOA.

In this paper, the long-term information of IPD (LTIPD) is extracted from IPD as a robust target speech existence measure. LTIPD describes the concentration degree of DOA estimations on a sound source with harmonic structure. The sparseness of target speech in spatial and time-frequency domain are utilized in a sector and T-F block-based strategy, and the $\pm 2\pi$ phase wrapping correction as well as the compensation for the nonlinear mapping between IPD and DOA are also proposed to improve the robustness.

The rest of this paper is organized as follows. Section 2 and 3 describes the signal model, the algorithm and the analysis. The performance evaluation is presented in Section 4, and the conclusion is made in section 5.

2. LONG-TERM INFORMATION OF INTER-CHANNEL PHASE DIFFERENCE

2.1. Signal model

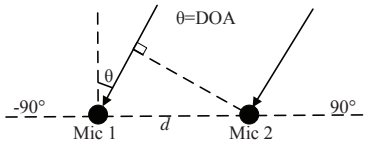


Fig. 1. Signal model

Given two omnidirectional microphones with distance d, as shown in Fig.1, the DOA range is $-90^\circ \leq \theta \leq 90^\circ$, where θ is DOA. For a far field sound source $s(t)$, the signal obtained by the two microphones is $x_1(t) = s(t) + n_1(t)$ and $x_2(t) = s(t-\tau) + n_2(t)$, where $n_1(t)$ and $n_2(t)$ represent the ambient noise, and τ is TDOA. Denote the sound speed as c, then $\tau = \frac{d\sin\theta}{c}$.

Convert the signal to T-F domain by short-time Fourier transform (STFT), then $X_{1,k}(\omega) = S_k(\omega) + N_{1,k}(\omega)$ and $X_{2,k}(\omega) = S_k(\omega)e^{-j\omega\tau} + N_{2,k}(\omega)$, where k is the time index, and ω represents angular frequency. The calculated IPD is

$$\widetilde{\psi}_k(\omega) = \angle X_{1,k}(\omega) - \angle X_{2,k}(\omega) \quad (1)$$

$\angle X_{1,k}(\omega)$ and $\angle X_{2,k}(\omega)$ are the phase of $X_{1,k}(\omega)$ and $X_{2,k}(\omega)$, and they are constrained to $[-\pi, \pi]$ after the $\mathrm{mod}(2\pi)$ operation, so $\widetilde{\psi}_k(\omega)$ is in the range of $(-2\pi, 2\pi)$.

Denote the IPD error as $\nu_k(\omega)$, then

$$\widetilde{\psi}_k(\omega) = \omega\tau + 2n\pi + \nu_k(\omega) = \psi_k(\omega) + 2n\pi + \nu_k(\omega) \quad (2)$$

where $\psi_k(\omega) = \omega\tau$, and it represents the unwrapped IPD. n is an integer number related to phase-wrapping.

$\nu_k(\omega)$ is defined as a random variable related to the local SNR. If the target speech dominates the T-F point, $\nu_k(\omega)$ is Gaussian-distributed with the mean equals 0, and the lower the SNR, the higher the variance. If there is no directional sound source, $\nu_k(\omega)$ is uniformly-distributed in the range of $(-2\pi, 2\pi)$. If there exists other directional sound sources in the T-F point, $\nu_k(\omega)$ follows a Gaussian distribution, and its mean is related to the DOA and energy of the sound sources. Because of the sparsity of speech in T-F domain, a T-F point dominated by more than one directional sound sources are rare if the T-F resolution is high enough.

2.2. LTIPD extraction

Speech is sparse in T-F domain, and in the T-F blocks that are dominated by the target speech, the IPDs usually indicate homogeneously to the DOA of the target speech. LTIPD is defined to measure such concentration of DOA estimations. The basic idea is to divide the signal in T-F and spatial domain to search for the T-F blocks with the obvious concentration of DOA estimates.

As shown in Fig.2, the target DOA range is divided into U overlapped sectors with equal width, and the signal on two channels are divided into overlapped frames and performs Discrete Fourier Transform to get the IPDs on each frequency bin for each frame. A whole T-F block in size $L \times Q$ is divided into Q sub-blocks in size $L \times 1$, denoted as b, $b = 1..Q$. A concentration measure $C_{b,i}$ is calculated for each sub-block for sector $i = 1..U$. If $C_{b,i} > \kappa_i$, the energy in the sub-block b are summarized as E_i for sector i. Finally, LTIPD is got as the maximum of all E_i.

$$LTIPD = max(E_i) \quad i = 1..U \quad (3)$$

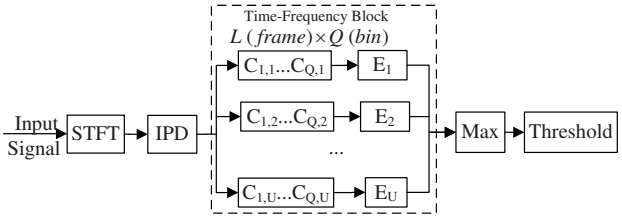


Fig. 2. Target speech detection based on LTIPD

The long-term window of L frames shifts one frame after the LTIPD calculation, and the final LTIPD is a frame-by-frame feature for target speech detection.

3. ALGORITHM DESCRIPTION

Four processes are introduced in LTIPD calculation to extract the long-term information and improve robustness. 1. More valid IPD samples are acquired by correction of the $\pm 2\pi$ wrapped IPDs. 2. The time resolution is increased by using short frame shift. 3. The sensitivity for a point sound source is improved by the sector-based DOA partition, in which nonlinear mapping between IPD and DOA are compensated by a weighting parameter. 4. The sensitivity for the concentration of DOA estimations are improved by dividing the T-F block based on the voice harmonic structure. In general, the LTIPD reflects the existence of a small-sized sound source in the target DOA range with harmonic structure in T-F domain.

3.1. IPD correction for $\pm 2\pi$ phase wrapping

In the noise-free case without spatial aliasing, where the frequency $f < \frac{c}{2d}$, $\widetilde{\psi}(\omega)$ may equal to $\psi(\omega)$ or $\psi(\omega) \pm 2\pi$, which is the $\pm 2\pi$ wrapped value. Such wrapped value exists for time τ in every signal period $T = \frac{2\pi}{\omega}$, so it appears more in higher frequency. If a wrapped IPD values is got on a T-F point, the DOA of the point cannot be estimated correctly.

However, in condition of $f < \frac{c}{2d}$, the range of $\psi(\omega)$ is $(-\pi, \pi)$, so the wrapped $\widetilde{\psi}(\omega)$ can be discriminated if it is outside this range, and $\psi(\omega)$ can be derived from $\widetilde{\psi}(\omega)$ by a $\pm 2\pi$ processing. Then the correct DOA estimation is got by $\widetilde{\psi}(\omega) \rightarrow \psi(\omega) \rightarrow \theta$.

3.2. Frame shift selection

Speech usually has many short pauses and silent segments, so it is sparse in time domain. If we analysis the signal with the traditional frame shift, which is 10ms or even longer, there are less than 10 T-F points for each frequency bin in a 100ms syllable. The concentration degree of DOA estimations varies little between the speech and non-speech segments.

In this paper, we set the frame shift to be 2ms or even less to get more T-F samples, then we observe the DOA estimations in a time duration of about one syllable. The concentration degree of the DOA estimations will increase obviously when the speech appears.

3.3. Spatial sector-based analysis

Speech is sparse in spatial domain. The target speech could be looked as a point sound source in the DOA range. As a wide-band signal, the target speech usually improves the SNR in the whole frequency band. This can be observed as higher concentration of DOA estimations on full frequency band. This property is utilized in this paper as a spatial sector-based analysis. The whole target DOA range is divided into several overlapped sectors to discriminate sound sources from different DOA, and each sector is a narrow DOA range. When the target speech appears, the DOA estimations on a wide frequency band would be concentrated to a sector obviously.

The concentration degree of T-F block b to sector i is evaluated as $C_{b,i} = \frac{H_{b,i}}{D_b}$, where $H_{b,i}$ represents the number of T-F points that indicate to sector i, and D_b represents the total T-F point number in block b. $C_{b,i}$ reflects the existence probability of a directional sound source from sector i. If it is higher than a threshold κ_i, the T-F block b is detected as dominated by the sound from sector i.

However, if we use equal κ_i for each sector, a nonlinear bias to the 0° DOA will be caused because of the mapping between IPD and DOA based on $\tau = \frac{d\sin\theta}{c}$. Therefore, we get κ_i by $\kappa_i = R\lambda_i$, where R is a constant, and λ_i is a weighting parameter to compensate the bias. Suppose the DOA range of sector i is $\alpha_i < \theta < \beta_i$, then λ_i is calculated as $\lambda_i = \frac{\sin\beta_i - sin\alpha_i}{2}$, where the constant 2 is got by $\sin 90^\circ - \sin(-90^\circ)$.

3.4. Harmonic structure-based T-F block partition

Speech is sparse in frequency domain, especially for the voiced speech. Most of the voice energy are concentrated to the harmonic structure. This property is utilized in this paper as a harmonic structure-based detection by dividing the T-F block into sub-blocks with long time and narrow frequency band, as shown in Fig.2. If there exists a harmonic in a sub-block b, the concentration parameter $C_{b,i}$ would be higher than a threshold κ_i for sector i.

4. EVALUATION AND ANALYSIS

4.1. An example of LTIPD

Fig.3 shows an example of LTIPD used for in-car VAD, where the DOA range is $\pm 20^\circ$, and there is the wind and whistle noise from $\theta = 80^\circ$. The waveform and the spectrum are given in the left column. The right column gives the $\sin\theta$ derived from $\widetilde{\psi}_k(\omega)$ and the final LTIPD, in which the frame

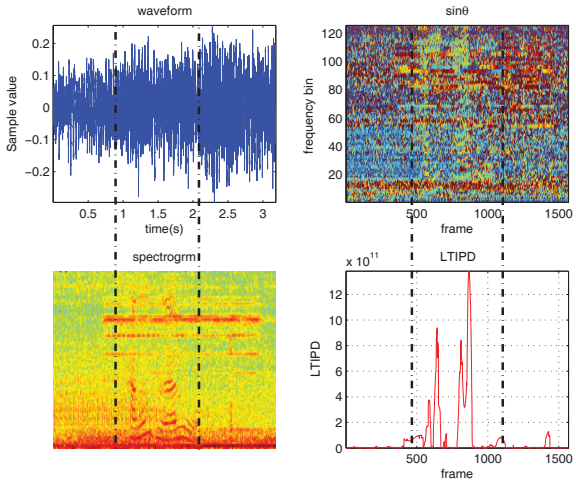


Fig. 3. An example of LTIPD VAD for in-car use

Table 1. Database Description

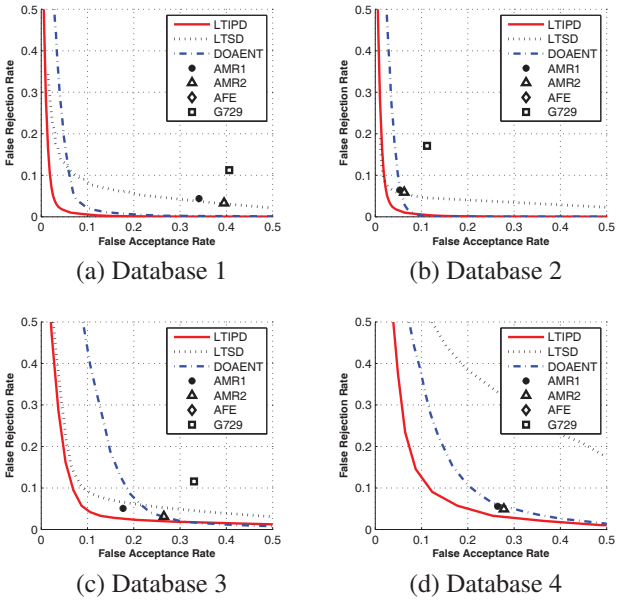
No	Speech	Speaker	$\theta(^\circ)$	D(m)	Len(h)	Environment
1	words	2F2M	± 30	0.2-0.3	2	Supermarket
2	words	2F2M	± 20	0.4-0.5	1	Meeting room
3	sentences	5F5M	± 50	1.5-3.0	3	Living room
4	words	4F4M	± 20	0.5-0.6	2	In car,60km/h

shift is 2ms. The target speech exists in 0.8-2.2s, which is labeled by the dash dot line, and a whistle from outside exits during 0.7-2.8s. As can be seen, LTIPD are only sensitive to the harmonic structure of the target speech. A hang-over scheme will be helpful to improve the VAD performance by preserving the unvoiced speech. For example, the short segments between the voiced speech segments could be detected as unvoiced speech.

4.2. Evaluation in ROC curves

In order to evaluate the performance of LTIPD VAD, 4 databases were recorded in real environments by 2 omni-directional microphones with $d = 5$cm, and all databases are hand-labeled. As shown in Table 1, the target speech are Chinese words and sentences spoken by male(M) and female(F) speakers from a priori known DOA range (θ) and distance range(D). The simulated applications include hands-held devices (Database 1 and 2), in-vehicle speech command (Database 4), and the humanoid robots or interactive TV (Database 3). To cover the real driving scenario, the in-car database was recorded with the window opened during half of the time. The sample rate is 8kHz, and the time length(Len) are 1 to 3 hours.

The speech false acceptance rate (FAR) and false rejection rate (FRR) are presented as the receiver operating



(a) Database 1 (b) Database 2

(c) Database 3 (d) Database 4

Fig. 4. ROC curves for Database 1-4

characteristics (ROC) curves, as shown in Fig.4. The compared VAD systems include single-channel VAD standards (ITU-T G.729B, ETSI Adaptive Multi-Rate (AMR1 and AMR2) and ETSI Advanced Front-End (AFE)), an improved single-channel VAD based on the long-term information in time-frequency domain (LTSD[8]) and the dual-channel algorithm based on the homogeneity of DOA estimations (DOAENT[5]).

In LTIPD VAD, the signal is analyzed in frame shift of 2ms and frame length of 256 points, which is also the FFT size. The sub-block length is $L = 30$ frames, and the analysis frequency band is 300 to 3000Hz, so $Q = 87$ bins. R is 6, and the sector width is 10°. The sector number $U = 10$ for database 1,2 and 4, and $U = 30$ for database 3.

All of the single-channel standard VADs have low FRR and high FAR except for Database 2, which has very stationary background noise. The work points of G.729 and AFE are even outside the scope in some cases. The LTSD works better than the standard VADs because of its good noise-tracking ability, but its error rate increases with the target distance and the intensity of environment noise. The DOAENT VAD outperforms all the single-channel algorithms on database 1, 2 and 4 by rejecting the non-directional noise. However, for the applications with directional noise, such as database3, the homogeneity of DOA estimation is not robust enough.

The proposed LTIPD VAD outperforms the other algorithms on all databases. For the applications with high SNR and narrow DOA range, such as database 1 and 2, LTIPD takes advantage of the high concentration degree of the DOA estimations. In the cases of low SNR and distant talking, LTIPD keeps robust by capturing the harmonic structures of speech in the target DOA range, as shown in ROC curves of Database3 and 4. Further experiments were also carried out with different U, and the performance as well as computational complexity kept stable.

5. CONCLUSION

A voice activity detection algorithm is proposed for distant-talking speech appears randomly in wide range of DOA. The VAD is based on the long-term information of inter-channel phase difference (LTIPD), which is a robust feature extracted from IPD by utilizing the long-term information of target speech, especially the sparsity in time, frequency and spatial domain. LTIPD is most sensitive to the point sound source with harmonic structure. The algorithm shows robust performance in several real environments.

6. REFERENCES

[1] Tao Yu and J.H.L. Hansen, "An efficient microphone array based voice activity detector for driver's speech in noise and music rich in-vehicle environments," in *ICASSP*, 2010, pp. 2834–2837.

[2] G.Lathoud, J. Bourgeois, and J. Freudenberger, "Sector-based detection for hands-free speech enhancement in cars," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[3] Israel Cohen and Baruch Berdugo, "Two-channel signal detection and speech enhancement based on the transient beam-to-reference ratio," in *ICASSP*, 2003, vol. 5, pp. 233–236.

[4] R.Le Bouquin-Jeannes and G.Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, no. 3, pp. 245–254, 1995.

[5] J.E.Rubio, K.Ishizuka, H.Sawada, S.Araki, T.Nakatani, and M.Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," in *ICASSP*, 2007, vol. IV, pp. 385–388.

[6] Hyun-Don Kim, K.Komatani, T.Ogata, and H.G.Okuno, "Two-channel-based voice activity detection for humanoid robots in noisy home environments," in *IEEE ICRA*, 2008, pp. 3495–3501.

[7] Gibak Kim and Nam Ik Cho, "Voice activity detection using the phase vector in microphone array," in *InterSpeech*, 2007, pp. 2957–2960.

[8] J.Ramirez, J.C.Segura, C.Benitez, A.de la Torre, and A.Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.