Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information

Yuuki Tachioka*, Tomohiro Narita† and Tomohiro Iwasaki‡

Information Technology R & D Center, Mitsubishi Electric Corporation, 5–1–1 Ofuna, Kamakura, 247–8501 Japan
(Received 8 August 2011, Accepted for publication 6 September 2011)

Keywords: Sound source localization, DOA estimation, CSP analysis, Noise robustness, Prior distributions

PACS number: 43.60.Fg, 43.60.Jn [doi:10.1250/ast.33.68]

1. Introduction

Sound source localization is important as a pre-process for target detection, speech enhancement, and speech recognition. To estimate the direction of arrival (DOA), the cross-power spectrum phase (CSP) analysis that uses two microphones is widely known as an effective estimation method [1]. However, the accuracy of the CSP analysis decreases at low SNR and owing to directional noise. Denda proposed a weighted CSP analysis, that weighs the spectrum in speech bands, and CSP coefficient subtraction that extracts estimated noise components from CSP coefficients [2]. These methods are applied only for speech with stationary noise. Nishihara proposed synchronous addition of a CSP coefficient, which uses three or more microphones and synchronously adds paired CSP coefficients [3]. This increases device size and computational load. In this paper, we propose a CSP analysis using prior distributions of source direction and voice activity detection (VAD) information in order to eliminate noise from CSP coefficients. This method can be adopted for any sound sources if activity of the source can be detected somehow, and requires minimal computation because the algorithm is simple and realized by two microphones.

2. Eliminating noises from CSP coefficients

2.1. Outline of CSP analysis

In the CSP analysis, DOA is estimated from the arrival time delay $\tau$ [s] using a cross-power spectrum between the two microphones. The distance between the microphones is $d$ [m/s], and INT is a function that returns the integer parts of $d$. $c$ denotes a sampling frequency [Hz], and $\sin^{-1}$ denotes a function that returns the angle whose sine is $x$. $\text{INT}(x)$ is a function that returns the integer part of $x$. $\text{MAX}$ is a function that returns a maximum value of $x$. $\arg\max(x)$ is a function that returns the argument of $x$.

$\arg\max(x)$ (CSP($i$, $k$)) [1]. Finally, the direction of the sound source $\theta$ is obtained as $\theta = \sin^{-1}(tc/(Lmf))$.

Figure 1 shows a schematic diagram of the proposed method. In this paper, we define “CSP(I)” and “CSP(II)” as baseline methods. “CSP(I)” is an original CSP analysis. “CSP(II)” is a conventional CSP analysis with a peak-hold process [4] and noise component suppression, which sets the cross power spectrum to zero when the estimated SNR is under 0 dB. The proposed method, which we name “CSP F,” has a CSP coefficient filtering process that is added to “CSP(I)” and “CSP(II)” respectively in Sects. 2.2 and 2.3.

2.2. Using prior distributions

We assume that sources do not move substantially and the duration of sources is longer than that of noise. For example, when you operate hands-free devices by voice, the location of speaker would not move substantially. Smoothed CSP coefficient $\overline{\text{CSP}}(i, k)$ is obtained as in Eq. (2) by averaging CSP($i$, $k$) during $2d + 1$ frames (here, $d = 5$).

$$\overline{\text{CSP}}(i, k) = \frac{1}{2d + 1} \sum_{j=-d}^{d} \text{CSP}(j, k).$$ (2)

We assume that $\overline{\text{CSP}}(i, k)$ is a likelihood of sound source directions, which correspond to a delay time $k$, then we obtain an accumulated likelihood $L(i, k)$ as follows: $L(i, k) = \sum_{j=0}^{d} \overline{\text{CSP}}(j, k)$. Prior distribution $P(i, k)$ ($0 \leq P(i, k) \leq 1$) is normalized by the maximum value of $L(i, k)$, as in Eq. (3):

$$P(i, k) = \frac{\text{MAX}(L(i, 0), L(i, 1), \ldots, L(i, k_{\text{max}}))}{L(i, k_{\text{max}})},$$ (3)

where $\text{MAX}$ is a function that returns a maximum value of arguments. Finally, the filtered CSP coefficient $\overline{\text{CSP}}(i, k)$ is obtained by combining a weighted CSP coefficient whose weight is $P(i, k)$ with an original coefficient at the combination ratio $r$, as in Eq. (4):

$$\overline{\text{CSP}}(i, k) = (r + (1 - r)P(i, k))\overline{\text{CSP}}(i, k).$$ (4)

If source does not move and SNR is high, simple averaging as $\overline{\text{CSP}}(i, k) = \sum_{j=-d}^{d} \text{CSP}(j, k)/(i + d + 1)$ is effective. However, if large CSP value of noise inputs, estimation accuracy decreases significantly because the peak of noise hardly diminishes by averaging. The proposed method only suppresses the noise component and does not increase the peak value of CSP coefficient because $P(i, k)$ is 1 or less. Hence, the
2.3. Using voice activity detection information

If the target is speech, peaks of CSP coefficients in non-speech areas are attributed to noise. A modified likelihood $L(i, k)$ is obtained as in Eq. (5) according to VAD information.

$$L(i, k) = \sum_{j=0}^{i+d} ((1 + \alpha) \delta(j) - \alpha) CSP(j, k),$$

where $\delta(j)$ is a function that returns unity in speech areas and zero in non-speech areas at frame $j$, and $\alpha (\alpha > 0)$ is a penalty.

This leads to a sign inversion of CSP coefficients in non-speech areas. Peaks of speech are enhanced by suppressing noise peaks, which are dominant in non-speech areas. Figure 3 illustrates this procedure. The first three frames are noise according to VAD. The speech peak appears at the center in the 4th-frame. However, the noise peak on the left side indicated as a closed asterisk is higher. In this case, likelihood $L(i, k)$ is obtained as in Eq. (5) by inverting any sign of CSP coefficients in the noise area and $P(i, k)$ is estimated from these likelihoods. The speech peak is enhanced and a source is located as a closed circle using the same procedure as in Eq. (4).

3. Experiments

3.1. Experimental setup

Impulse responses were measured at every 30° at the center and near the wall of the room as shown in Fig. 4. The minimum and maximum distances between sources and the center of a microphone array (8 ch circle array) $L_{ar}$ were 1 and 2 m, respectively. Reverberation time $T_{90}$ was 0.68 s and the reverberation decay curve was not bent. Evaluation data was produced by convolving speech (control words for air conditioner) with these impulse responses. Speaker direction was located using two diagonal microphones. Recorded noise (Table 1) was added to the evaluation data at SNR of 6 and 24 dB. Sampling frequency $f_s$ was 16 kHz, and the window length and frame shift of STFT were 60 ms and 30 ms, respectively. According to preparatory studies, $r$ was 0.3 [Eq. (4)] and $\alpha$ was 1.0 [Eq. (5)]. Note that performance depended little on these parameters.

3.2. Results and discussions

3.2.1. Comparison with the two-microphone method

Estimation accuracy in the speech area is calculated by each frame when the microphone array is located at the center of the room. The error tolerance is ±15°. The estimation average accuracy in all directions is shown in Figs. 5 and 6, which represent the easiest case (SNR of 24 dB, $L_{ar}$ of 1 m) and the most difficult case (SNR of 6 dB, $L_{ar}$ of 2 m), respectively. A combination of “CSP(I)” and the proposed method (“CSP F”) improves the accuracy in Fig. 5 but not in
Fig. 6, because of less accurate prior distributions when SNR is low. Accuracy increases with a combination of “CSP(II)” and “CSP F,” because the noise components of the cross power spectrum are reduced in advance. In addition, the use of VAD [5] improves the performance of “CSP F” because the estimation accuracy of prior distributions increases as noises are learned (“CSP F with VAD”). The difference in accuracy between automatic VAD and manually tagged VAD (“CSP F with Ideal VAD”) is not large.

Figure 7 shows CSP coefficients on a time-angle plane with directional noise at “180°” and speech at 60° (SNR of 6 dB, $L_{sr}$ of 2 m). Using the proposed method, maximal peaks created by foot noises near 180° are suppressed, maintaining the speaker peaks near 60°, as seen on the right graph, unlike that observed for the peaks of the conventional method, as seen on the left graph. Figure 8 shows the section at A–B (1.02 s) of Fig. 7. Originally, the peak appears close to 160°, but with the proposed method, it is correctly estimated close to 60°.
3.2.2. Comparison with four- and eight-microphone methods

Using three or more microphones reduces noise by synchronously adding paired CSP coefficients [3]. Figure 9 shows the results using four microphones (all six pairs) and eight microphones (four diagonal pairs). The accuracy of the proposed method is superior to the method that uses four microphones and is equivalent to the method that uses eight microphones.

3.2.3. Effect of receiving point

Figure 10 shows the accuracy when the microphone array is near the wall (SNR of 6 dB, $L_{sr}$ of 2 m). Sources and receivers are located near the wall.

4. Conclusions

We propose a method that reduces the effect of noise for the estimation of DOA by CSP analysis. The proposed method uses prior distributions estimated from accumulated CSP coefficients. We demonstrated that this method was effective for both diffusive and directional noise and that using VAD information improved estimation accuracy.

References