NOISE ROBUST TALKER LOCALIZATION BASED ON WEIGHTED CSP ANALYSIS WITH AN AVERAGE SPEECH SPECTRUM FOR MICROPHONE ARRAY STEERING

Yuki Denda, Takanobu Nishiura, and Yoichi Yamashita

gr021052@se.ritsumei.ac.jp, nishiura@is.ritsumei.ac.jp, yama@media.ritsumei.ac.jp
Graduate School of Science and Engineering, Ritsumeikan University, Japan

ABSTRACT

This paper proposes a noise robust talker localization method based on weighted CSP (Cross-power Spectrum Phase) analysis with an average speech spectrum as the pre-process of microphone array steering. The proposed method consists of two processes. First, CSP coefficients are weighted by analysis weight coefficients based on an average speech spectrum, which is trained with speech database, in advance. Next, the interference noises are reduced on spatial domain by CSP coefficient subtraction. As a result of evaluation experiments in a real room, we confirmed that the proposed method could provide better talker localization performance than the conventional methods.

1. INTRODUCTION

Talker localization is very important process of microphone array steering [1] for high quality sound capture of distant-talking speech. It is because that the desired speech can be selectively acquired by steering the directivity to the target talker direction with the microphone array.

Talker localization method based on CC (Cross Correlation) method [2] has been proposed and it is often used for this purpose. It localizes a target talker based on cross-power spectrum between captured signals. However, it is not enough robust, because cross-power spectrum is directly affected by the amplitude of noise signal. To overcome this problem, CSP (Cross-power Spectrum Phase) analysis [2] has been proposed as an advanced CC method. It can accurately localize the target talker without dependence on spectral characteristics of captured signals, because it only utilizes phase difference between captured signals with a pair of transducers by employing normalized cross-power spectrum with the amplitude of captured signals. Therefore, it is a powerful technique for correct talker localization in higher SNR (Signal to Noise Ratio) environments. However, it cannot sufficiently achieve the effective performance in lower SNR environments, in especially directional-noisy environments, because it cannot acquire only the spatial phase difference of desired speech in real noisy environments, in which many noise sources with various frequency characteristics exits. In addition, while the general short-time speech spectra have weak energy in higher frequency bands and so on, the CSP analysis localizes a target talker by utilizing the spatial phase difference of whole frequency bands of captured signals, generally.

To solve this problem, we study introduction of analysis weight coefficients based on only spectrum of speech signals into the CSP analysis. Analysis weight coefficients weight the spatial phase difference by only the frequency characteristics of speech signal. In contrast, the CC method is realized as the CSP analysis weighted by not only the spectrum of desired speech, but also the one of noise signal. Accordingly, we propose weighted CSP analysis with an average speech spectrum, as a new method of noise robust talker localization. In addition, we propose CSP coefficient subtraction as a noise reduction method on spatial domain. It spatially suppresses influences of the interference noise, by subtracting CSP coefficients acquired in non-speech frame from the ones acquired in noisy speech frame.

2. CONVENTIONAL METHODS

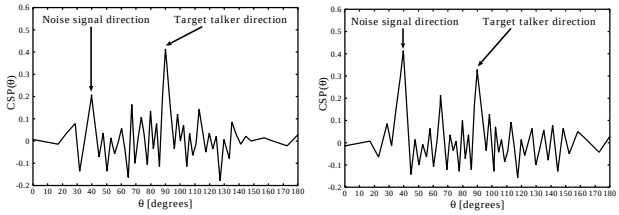
2.1. Cross correlation method

CC (Cross Correlation) method [2] estimates DOAs (Direction Of Arrival) and TDOAs (Time Delay Of Arrival) of target sound sources based on CC coefficients between captured signals, as derived from Equation (1)(2).

$$\mathrm{CC}(k) = \mathrm{IDFT}\left[x_1(\omega) \cdot x_2^*(\omega)\right], \tag{1}$$

$$\theta = cos^{-1}\left(\frac{c \cdot \tau}{d \cdot F_s}\right), \quad \tau = \underset{k}{\mathrm{argmax}}(\mathrm{CC}(k)), \tag{2}$$

where k shows time delay, $x_{[\cdot]}(\omega)$ shows frequency representation of $x_{[\cdot]}(t)$, * shows the complex conjugate, $\mathrm{IDFT}[\cdot]$ is the inverse DFT (Discrete Fourier Transform), $\mathrm{CC}(k)$ is CC coefficients, τ is an estimated TDOA, c is the sound propagation speed, d is the distance between a pair of transducers, F_s is the sampling frequency and θ is an estimated DOA.



(a) Higher-SNR environment. (b) Lower-SNR environment.

Figure 1: An example of CSP coefficients.

The CC method is sensitive to noise signals, because cross-power spectrum depends on the frequency characteristics of captured signals, as derived from Equation (1).

2.2. Cross-power spectrum phase analysis

CSP (Cross-power Spectrum Phase) analysis [2] has been proposed as an advanced technique of the CC method. It employs CSP coefficients based on normalized cross-poser spectrum by the amplitude of captured signals, instead of CC coefficients, as derived from Equation (3).

$$\mathrm{CSP}(k) = \mathrm{IDFT}\left[\frac{x_1(\omega)\cdot x_2^*(\omega)}{|x_1(\omega)|\cdot|x_2(\omega)|}\right]. \qquad (3)$$

The CSP analysis only utilizes phase difference between captured signals by a pair of transducers on each frequency, as derived from Equation (3). Therefore, it can accurately localize a target talker in higher SNR environments, as shown in Figure 1(a). In Figure 1(a), the CSP coefficient of target talker direction is most largest peak. In Figure 1, CSP coefficients are transformed from time domain ($\mathrm{CSP}(k)$) into directional domain ($\mathrm{CSP}(\theta)$) as derived from Equation (4).

$$\theta = cos^{-1}\left(\frac{c\cdot k}{d\cdot F_s}\right). \qquad (4)$$

On the other hand, the talker localization performance of the CSP analysis is seriously degraded in lower-SNR environments, as shown in Figure 1(b). In Figure 1(b), the CSP coefficient of target talker direction is smaller peak than the one of noise signal direction. To cope with this problem, it is necessary to suppress the CSP coefficients affected by the interference noise. Consequently, in this paper, we propose weighted CSP analysis with an average speech spectrum and CSP coefficient subtraction to spatially suppress noise signal.

3. PROPOSED METHOD

Figure 2 shows an overview of the proposed method. After signal capturing with a paired-transducer in noisy speech frame, the phase difference on each frequency is weighted by analysis weight coefficient based on average speech spectrum. Then, we can acquire the weighted CSP coefficients. Finally, a target talker is localized by subtracting CSP coefficients acquired in non-speech frame from the weighted CSP coefficients.

3.1. Analysis weight coefficients

First, we calculate average speech spectrum with speech database, in advance, as derived from Equation (5).

$$\overline{s}(\omega) = \frac{\sum_{l=1}^{L}\sum_{n=1}^{N_l}|s([l,n],\omega)|}{\sum_{l=1}^{L}N_l}, \qquad (5)$$

where $s([l,n],\omega)$ is the speech spectrum, L is the number of sentences, N_l is the total frame number of l th sentence and $\overline{s}(\omega)$ is the average speech spectrum. In this paper, we employed 503 phoneme-balanced Japanese sentences $\times$ 20 subjects (14 females and 6 males) as training data for the average speech spectrum. The average speech spectrum is calculated one time per speech frame with 32 msec. (Hamming window).

Next, we divide the average speech spectra into subbands with equal bandwidth on mel-frequency. It is because that the general short-time speech spectra have strong energy in lower frequency bands, which including first and second formants. In addition, the long-time average speech spectra have an almost flat inclination in 800 Hz or less, and they have an inclination of $-10\,\mathrm{dB/oct.}$ in 800 Hz or over [3]. Accordingly, subband division with equal bandwidth on mel-frequency provides the ideal subband resolution with more particularly lower frequency bands and more roughly higher frequency bands.

As a result, we can acquire analysis weight coefficient by respectively smoothing the average speech spectrum on each subband, as derived from Equation (6).

$$W(\omega) = 20\cdot\log_{10}\left(\frac{\sum_{\omega=\omega_{b_L}}^{\omega_{b_H}}\overline{s}(\omega)}{\omega_{b_H}-\omega_{b_L}}\right),\ b=1,\cdots,B, \qquad (6)$$

where $W(\omega)$ is the analysis weight coefficient, ω_{b_L} is the under limitation frequency of b th subband, ω_{b_H} is the upper limitation frequency of b th subband and B is the number of subband divisions.

3.2. Weighted CSP coefficients

Weighted CSP coefficients are acquired by multiplying spatial phase difference by the average speech spectrum. In addition, we conduct the frequency band selection, because the general short-time speech spectra have weak energy in higher frequency bands and so on. The weighted CSP coefficients are derived from Equation (7).

$$\begin{aligned}\text{W-CSP}(k) &= \mathrm{IDFT}\left[W(\omega)\cdot\frac{x_1(\omega)\cdot x_2^*(\omega)}{|x_1(\omega)|\cdot|x_2(\omega)|}\right],\\ \omega &= \omega_L,\cdots,\omega_H, \qquad (7)\end{aligned}$$

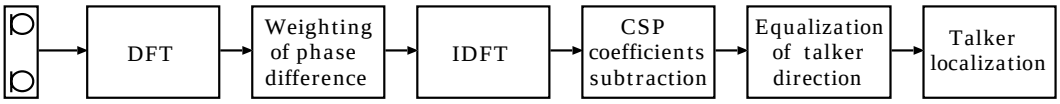


Figure 2: An overview of the proposed method.

where W-CSP(k) shows the weighted CSP coefficients, ω_L is the under limitation of selected frequency and ω_H is the upper limitation of selected frequency.

3.3. CSP coefficient subtraction

SS (Spectral Subtraction) [4] is an effective noise reduction method. It enhances the desired speech by subtracting the spectrum of noise signal from the one of observed signal, as derived from Equation (8).

$$|\hat{S}(\omega)| = |Y(\omega)| - |\overline{N(\omega)}|, \quad (8)$$

where $|\overline{N(\omega)}|$ shows the spectrum of noise signal acquired in non-speech frame, $|Y(\omega)|$ shows the one of observed signal, $|\hat{S}(\omega)|$ shows the one of enhanced speech. In this paper, we propose CSP coefficients subtraction, as an extended technique of the SS.

When the desired speech $s(t)$ and noise signal $n(t)$ are simultaneously captured, a cross-term appears in the numerator of Equation (3) and it is derived from Equation (9).

$$\begin{aligned} x_1(\omega) \cdot x_2^*(\omega) = & \\ & S(\omega)^2 e^{-j\omega(\tau_{s_1}-\tau_{s_2})} + N(\omega)^2 e^{-j\omega(\tau_{n_1}-\tau_{n_2})} \\ & + S(\omega)N(\omega)(e^{-j\omega(\tau_{s_1}-\tau_{n_2})} + e^{-j\omega(\tau_{n_1}-\tau_{s_2})}), \end{aligned} \quad (9)$$

where $\tau_{s_{[\cdot]}}$ and $\tau_{n_{[\cdot]}}$ are TDOAs of $s(t)$ and $n(t)$. In this situation, if $s(t)$ and $n(t)$ are strongly correlated, we cannot simply denote the observed CSP coefficients as derived from Equation (10). However, we simply derive Equation (11) from Equation (10). Thus, we can reduce noise signal on the domain of CSP coefficient as derived from Equation (12).

$$\begin{aligned} \mathrm{CSP}(k) &= w_s \cdot \mathrm{CSP}_s(k) + w_n \cdot \mathrm{CSP}_n(k) \\ &\quad + w_{sn} \cdot \mathrm{CSP}_{sn}(k), & (10) \\ \mathrm{CSP}(k) &= \mathrm{CSP}_s(k) + \mathrm{CSP}_n(k), & (11) \\ \mathrm{CSP}_s(k) &= \mathrm{CSP}(k) - \mathrm{CSP}_n(k), & (12) \end{aligned}$$

where $\mathrm{CSP}_{sn}(k)$ show CSP coefficients of the cross-term, $\mathrm{CSP}_n(k)$ show the ones of noise signal and $\mathrm{CSP}_s(k)$ show the ones of speech signal. However, it is so difficult to acquire only $\mathrm{CSP}_n(k)$ that we employ Equation (13) instead of $\mathrm{CSP}_n(k)$.

$$\begin{aligned} \mathrm{CSP}_{\hat{n}}(k) &= \frac{\sum_{n=1}^{N} \mathrm{CSP}_{n''}(n,k)}{N}, & (13) \\ \mathrm{CSP}_{n''}(n,k) &= \begin{cases} \mathrm{CSP}_{n'}(n,k) & \mathrm{CSP}_{n'}(n,k) > 0 \\ 0 & \mathrm{CSP}_{n'}(n,k) \leq 0 \end{cases}, \end{aligned}$$

where $\mathrm{CSP}_{n'}(n,k)$ show CSP coefficients acquired in n th non-speech frame, $\mathrm{CSP}_{\hat{n}}(k)$ show the estimated ones of noise signal. As a result, the CSP coefficient subtraction is derived from Equation (14)(15).

$$\begin{aligned} \alpha &= \frac{max(\text{W-CSP}(k))}{max(\mathrm{CSP}_{\hat{n}}(k))}, & (14) \\ \mathrm{CSP}_{\hat{s}}(k) &= \text{W-CSP}(k) - \alpha \cdot \mathrm{CSP}_{\hat{n}}(k), & (15) \end{aligned}$$

where α is the normalization coefficient.

3.4. Equalization of estimated talker direction

In this paper, we assume that the target talker may not move so rapidly. Therefore, we average the CSP coefficients acquired in current frame and the ones acquired in previous frames as derived form Equation (16). Averaging of CSP coefficients on time sequences may equalize the estimated talker direction, because it smoothes rapid shift of peak of the ones. As a result, the target talker is localized with the proposed method as derived form Equation (16)(17).

$$\overline{\mathrm{CSP}_{\hat{s}}}(k) = \frac{\sum_{l=1}^{L} \mathrm{CSP}_{\hat{s}}(n-l,k)}{L}, \quad (16)$$

$$\theta_s = cos^{-1}\left(\frac{c \cdot \tau_s}{d \cdot F_s}\right), \quad \tau_s = \underset{k}{\mathrm{argmax}}(\overline{\mathrm{CSP}_{\hat{s}}}(k)), \quad (17)$$

where θ_s is the estimated DOA of the desired speech, that is the estimated talker direction.

4. EVALUATION EXPERIMENTS

4.1. Experimental conditions

We carried out evaluation experiments in a real room. Figure 3 shows the experimental environment. Room reverberation ($T_{[60]}$) was 0.47 sec. and ambient noise level was 50.1 dBA. Thus, this room is a higher noisy environment. Table 1 shows experimental conditions. We employed 216 phoneme-balanced isolated Japanese words $\times$ 2 subjects (1 female and 1 male) as speech test data. We employed HSLN (Human Speech Like Noise) [5] as noise signal. HSLN is a kind of bubble noise generated by superimposing independent speech signals. By changing the number of superpositions, we can simulate various noise conditions. In this paper, number of superpositions is 256 times. Talker localization is conducted in following two conditions. **Condition 1:** The desired speech comes form 70

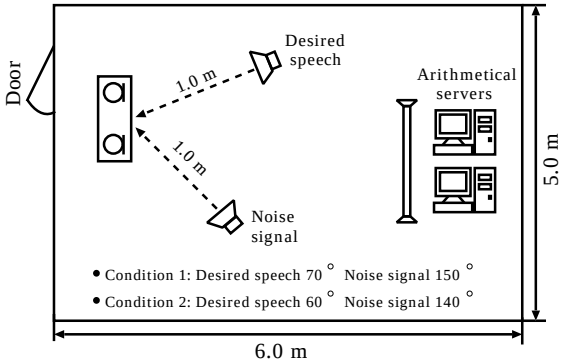


Figure 3: Experimental environment.

Table 1: Experimental conditions

Room reverberation $T_{[60]}$	0.47 sec.
Ambient noise	50.1 dBA
Sampling frequency	16 kHz
Paired microphone	148.75 mm spacing
Test date	
Speech (open)	216 words × 2 subjects (1 female and 1 male)
Noise signal	Huma speech like noise [5]
Talker localization	
Frame length	64 msec. (Hannig window)
Frequency band division	12 divisions
Frequency band selection	300 ∼ 5,000 Hz
Frame averaging number	10

degrees and noise signal comes from 150 degrees. **Condition 2:** The desired speech comes form 60 degrees and noise signal comes from 140 degrees. Talker localization is conducted one time per speech frame with 64 msec. The distance between two transducers is 148.75 mm.
In these situations, we evaluated talker localization performance, subject to SNR of -5 dB, ∼, 30 dB, and clean, respectively. The talker localization performance is evaluated by LA (Localization Accuracy) as derived from Equation (18).

$$\mathrm{LA} = \frac{\sum_{l=1}^{L}\sum_{n=1}^{N_l} \mathrm{I_{cor}}(l,n)}{\sum_{l=1}^{L} N_l}, \quad (18)$$

$$\mathrm{I_{cor}}(l,n) = \begin{cases} 1 & |\mathrm{D_{cor}}(l,n) - \mathrm{D_{est}}(l,n)| \leq \mathrm{Err}. \\ 0 & |\mathrm{D_{cor}}(l,n) - \mathrm{D_{est}}(l,n)| > \mathrm{Err}. \end{cases},$$

where L is the number of words, N_l is the total frame number of l th word, $\mathrm{D_{cor}}(l,n)$ is the correct talker direction, $\mathrm{D_{est}}(l,n)$ is the estimated talker direction and $\mathrm{Err}.$ is the admissible error (in this paper, $\mathrm{Err}.$ = 10 degrees).

4.2. Experimental results

Figure 4 shows the experimental results of talker localization, that are the average of **Condition 1.** and **Condition 2.** In Figure 4, "CC method" represents the experimental results with the CC method, "CSP analysis" represents the experimental results with the conventional CSP analysis, "Proposed method" represents the experimental results with the proposed method. As shown in Figure 4, we can confirm that the proposed method accurately localized the target talker than the conventional methods.

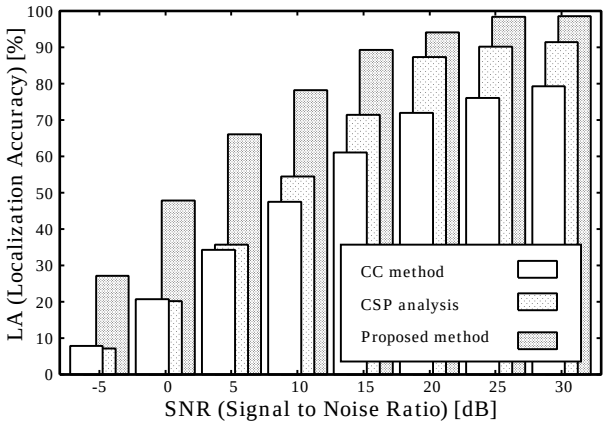


Figure 4: The experimental results of talker localization.

5. CONCLUSIONS

This paper proposes a noise robust talker localization method based on weighted CSP (Cross-power Spectrum Phase) analysis with an average speech spectrum. The proposed method consists of two processes. First, CSP coefficients are weighted by analysis weight coefficients based on an average speech spectrum, which is trained with speech database, in advance. Next, the interference noises are reduced on spatial domain by CSP coefficient subtraction. As a result of evaluation experiments in a real room, we confirmed that the proposed method could provide better talker localization performance than the conventional methods. In future work, we will attempt to conduct speaker adaptation of analysis weight coefficients.

Acknowledgement: This work was partly supported by The Leading Project "e-Society" funded by The Ministry of Education, Culture, Sports, Science and Technology of Japan.

6. REFERENCES

[1] J.L. Flanagan, et al., "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[2] C.H. Knapp, et al., "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[3] H.K. Dunn, et al. "Statistical measurements on conversational speech," *J. Acoust. Soc. Am.*, vol. 11, no. 3, pp. 449-476, 1956.

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[5] D. Kobayashi, et al., "Extracting speech features from human speech like noise," *Proc. ICSLP96*, vol. 1, pp. 418–421, 1996.