A Study of Weighted CSP Analysis with Average Speech Spectrum for Noise Robust Talker Localization

Yuki Denda, Takanobu Nishiura, Yoichi Yamashita

Graduate School of Science and Engineering,
Ritsumeikan University, Kusatsu, Japan

gr021052@se.ritsumei.ac.jp, nishiura@is.ritsumei.ac.jp, yama@media.ritsumei.ac.jp

Abstract

This paper describes a new method of noise robust talker localization for the front-end processing of microphone array steering. Conventional talker localization methods cannot localize a target talker accurately in higher noisy environments. To deal with this problem, in this paper, we propose an weighted CSP (Cross-power Spectrum Phase) analysis with an average speech spectrum. The proposed method consists of two processes. At first, CSP coefficients are weighted by analysis weight coefficients based on average speech spectrum, which is trained with speech database, in advance. Next, the interference noises are reduced on spatial domain by CSP coefficients 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

The high quality sound capture of distant-talking speech is very important for the hands-free speech acquisition systems, because ambient noise and room reverberations seriously degrade sound capture quality in real acoustical environments. A microphone array steering [1] is an ideal candidate. With the microphone array steering, the desired speech can be selectively acquired by steering the directivity to the target talker direction, sensitively. However, it requires localizing the target talker.

Therefore, talker localization method based on CC (Cross-Correlation) method [2] has been proposed and it is often used for this purpose. The CC method localizes a target talker by utilizing CC coefficients based on cross-power spectrum between captured signals. However, it is not enough robust, if the desired speech and noise signals are simultaneously captured. Because CC coefficients are influenced by the amplitude of noise signal, directly. To overcome this problem, CSP (Cross-power Spectrum Phase) analysis [2] has been proposed as an advanced CC method. It only utilizes phase difference between captured signals with a pair of transducers, by employing normalized cross-power spectrum with amplitude of captured signals instead of cross-power spectrum. Accordingly, it can accurately localize the target talker without dependence on spectral characteristics of captured signals [3].

In noisy environments, the CC coefficients are realized as weighted CSP coefficients with frequency characteristics of desired speech and noise signal. On the other hand, the CSP analysis only utilizes the spatial phase difference. However, although it is a powerful technique for correct talker localization in higher SNR (Signal to Noise Ratio) environments, it cannot sufficiently achieve the effective performance in lower SNR environments, in especially directional-noisy environments. It is because that the CSP analysis cannot only acquire the spatial phase difference of desired speech in real noisy environments to which there are many noise sources with various characteristics.

To cope with this problem, in this paper, we study introduction of analysis weight coefficients into the CSP analysis. Analysis weight coefficients based on the frequency characteristics of speech signals may improve the DOA estimation performance, the CSP analysis localizes a target talker utilizing spatial phase difference of whole frequency bands of captured signals. In addition, the spatial phase difference is only weighted by the characteristics of speech signal. Accordingly, we propose an weighted CSP analysis with average speech spectrum, as a new method of noise robust talker localization. Analysis weight coefficients are realized based on average speech spectrum trained with speech database, in advance. In addition, we propose a spatial noise reduction method to suppress influences of the interference noise, by subtracting CSP coefficients acquired in non-speech frame from the ones acquired in noisy speech frame. We call this approach the CSP coefficients subtraction.

2. Conventional talker localization method

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).

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

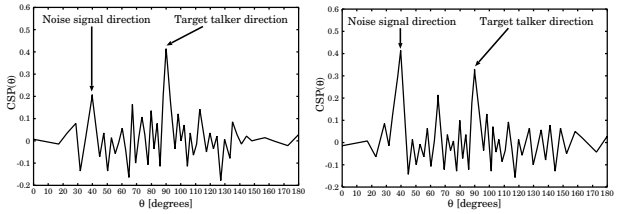
$$\theta = cos^{-1}\left(\frac{c \cdot \tau}{d \cdot F_s}\right), \quad \tau = \underset{k}{\mathrm{argmax}}(\mathrm{CC}(k)), \tag{2}$$

where k are time indexes, $x_{[\cdot]}(\omega)$ 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.

However, 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



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

Figure 1: An example of CSP coefficients.

the amplitude of captured signals, instead of CC coefficients, as derived from Equation (3).

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

The CSP analysis can accurately estimate a DOA without dependence on spectral characteristics of desired signal [3]. It is because that it only utilizes phase difference between captured signals by a pair of transducers on each frequency, as derived from Equation (3). Consequently, 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).

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

On the other hand, DOA estimation performance of the CSP analysis is seriously degraded by noise signal 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 ones of noise signal direction. Accordingly, it is necessary to suppress CSP coefficients affected by the interference noise. In addition, while the general speech spectra have weak energy in higher frequency bands and so on, the CSP analysis localizes a target talker with whole frequency bands of captured signals, generally. Therefore, in this paper, we propose an weighted CSP analysis with average speech spectrum and CSP coefficients 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, we weight the phase difference on each frequency by analysis weight coefficients based on average speech spectrum, respectively. 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

In this paper, we employed analysis weight coefficients based on average speech spectrum.

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

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

where $s([l,n],\omega)$ is the Fourier spectrum, L is the number of sentence, 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) and frame interval with 16 msec.

Next, we divide the average speech spectra into subbands with equal bandwidth on mel-frequency. Because, the general 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 [4]. 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. Finally, we can acquire analysis weight coefficient by smoothing the average speech spectrum on each subband, respectively. Equation (6) derives from analysis weight coefficients.

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

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 division. Figure 3 shows an example of analysis weight coefficients.

3.2. Weighted CSP coefficients

Weighted CSP coefficients are acquired by multiplying spatial phase difference with the analysis weight coefficients. In addition, we conduct the frequency band selection, because the general speech spectra have weak energy in higher frequency bands and so on. Equation (7) derives from the weighted CSP coefficients.

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

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

3.3. CSP coefficients subtraction

In this paper, we propose the CSP coefficients subtraction, as an extended technique of SS (Spectral Subtraction) [5].

The SS is a conventional noise reduction method. It enhances the desired speech by subtracting the spectrum of noise signal from the spectrum of observed signal, as derived from Equation (8).

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

where $|Y(\omega)|$ shows the spectrum of observed signal, $|\overline{N(\omega)}|$ shows the spectrum of noise signal acquired in non-speech frame, $|\hat{S}(\omega)|$ shows the spectrum of enhanced speech.

In this paper, we assume that the desired speech $s(t)$ and noise signal $n(t)$ are simultaneously captured. In this situation, the cross-term appears in the numerator of Equation (3) and it

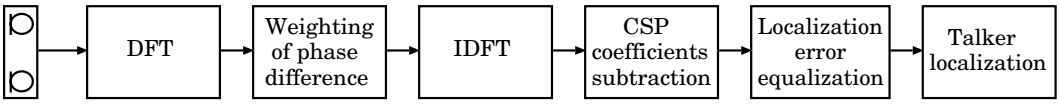


Figure 2: An overview of the proposed method.

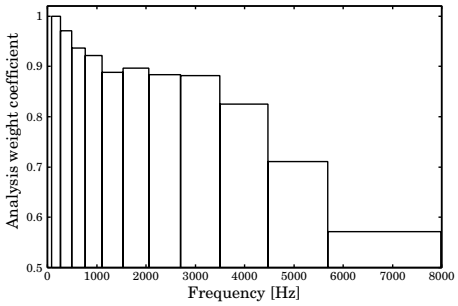


Figure 3: An example of analysis weight coefficients.

is derived from Equation (9).

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

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

$$\mathrm{CSP}(k) = w_s\cdot\mathrm{CSP}_s(k) + w_n\cdot\mathrm{CSP}_n(k) + w_{sn}\cdot\mathrm{CSP}_{sn}(k), \quad (10)$$

$$\mathrm{CSP}(k) = \mathrm{CSP}_s(k) + \mathrm{CSP}_n(k), \quad (11)$$

$$\mathrm{CSP}_s(k) = \mathrm{CSP}(k) - \mathrm{CSP}_n(k). \quad (12)$$

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

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

where $\mathrm{CSP}_{n'}(n,k)$ are CSP coefficients acquired in n th non-speech frame, $\mathrm{CSP}_{\hat{n}}(k)$ are estimated CSP coefficients affected by noise signal. As a result, we can denote the CSP coefficients subtraction as derived from Equation (15).

$$\alpha = \frac{max(\text{W-CSP}(k))}{max(\mathrm{CSP}_{\hat{n}}(k))}, \quad (14)$$

$$\mathrm{CSP}_{\hat{s}}(k) = \text{W-CSP}(k) - \alpha\cdot\mathrm{CSP}_{\hat{n}}(k), \quad (15)$$

where α is the normalization coefficient.

3.3.1. Localization error equalization

In this paper, we assume that the target talker may not move so rapidly. Therefore, we average 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 talker localization error, because it smoothes rapid shift of peak of the ones on time sequences. Finally, the target talker is localized with the proposed method as derived form Equation (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.

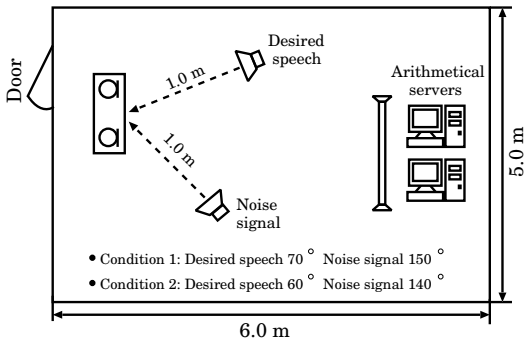


Figure 4: Experimental environment.

Table 1: Experimental conditions

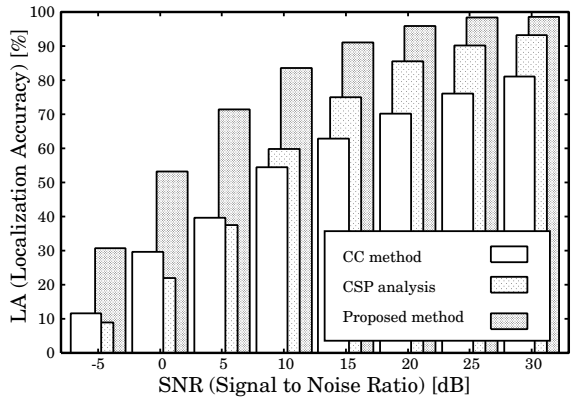
Room reverberation $T_{[60]}$	0.47 sec.
Ambient noise	50.1 dBA
Sampling frequency	16 kHz
Paired microphone	148.75 mm spacing
Test date	
Speech (open)	216 words × 2 subjects (1 female and 1 male)
Noise signal	White Gaussian noise, HSLN [6]
Talker localization	
Frame length	64 msec.
Window function	Hanning window
Frequency band division	12 divisions
Frequency band selection	300 ∼ 5,000 Hz
Frame averaging number	10

4. Evaluation experiments

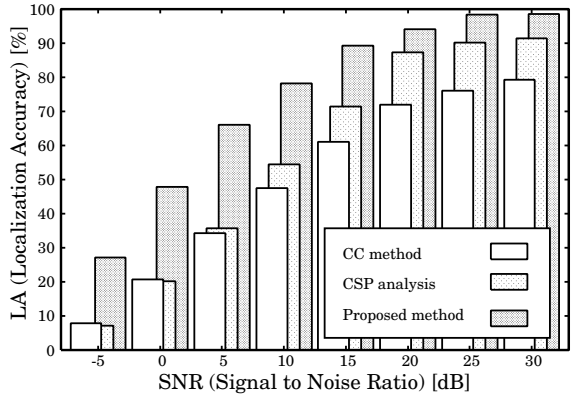
4.1. Experimental conditions

We carried out evaluation experiments in a real room. Figure 4 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 × 2 subjects (1 fe-



(a) White Gaussian noise environment.



(b) HSLN environment.

Figure 5: The experimental results of talker localization.

male and 1 male) as speech test data. We employed white Gaussian noise and HSLN (Human Speech Like Noise) [6] as noise signal. HSLN is a kind of bubble noise generated by superimposing independent speech signals. By changing the number of superposition, we can simulate various noise conditions. In this paper, number of superposition is 256 times.

Talker localization is conducted in following two conditions. **Condition 1:** The desired speech comes form 70 degrees and noise signal comes from 150 degrees. **Condition 2:** The desired speech comes form 60 degrees and noise signal comes from 140 degrees. The distance between the paired-transducer and each sound source is 1.0 m. The distance between two transducers is 148.75 mm. Talker localization is conducted one time per speech frame with 64 msec. Talker localization method employed frequency band of 300 $\sim$ 5,000 Hz.

In these situations, we evaluated talker localization performance, subject to SNR of -5 dB, $\sim$, 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 word, 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 5 shows the experimental results of talker localization. Figure 5(a) shows the experimental results in white Gaussian noise environment and Figure 5(b) shows the experimental results in HSLN environment. Each experimental results are the average of **Condition 1.** and **Condition 2.** In Figure 5, "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 5(a), we can confirm that the proposed method accurately localized a target talker than the conventional methods in white Gaussian noise environments.

On the other hand, comparing Figure 5(a) with Figure 5(b), the talker localization performance in HSLN environment was degraded than the talker localization performance in white Gaussian noise environment. It is because that the HSLN has similar spectral inclination to speech signal, while the white Gaussian noise has flat spectral inclination. However, the proposed method also provides better talker localization performance than the conventional methods in HSLN environments.

5. Conclusions

This paper proposes a new method of noise robust talker localization based on an weighted CSP analysis with average speech spectrum. The proposed method consists of two processes. At first, CSP coefficients are weighted by analysis weight coefficients based on average speech spectrum, which is trained with speech database, in advance. Next, the interference noises are reduced on spatial domain by CSP coefficients 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.

6. Acknowledgments

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

7. 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 and G.C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. ASSP, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[3] I. Yamada and N. Hayashi, "Improvement of the performance of cross correlation Method for identifying aircraft noise with pre-whitening of signals," J. Acoust. Soc. Jpn. (E), vol. 13, no. 4, pp. 241–252, 1992.

[4] H. K. Dunn and S.D. White, "Statistical measurements on conversational speech," J. Acoust. Soc. Am., vol. 11, no. 3, pp. 449-476, 1956.

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

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