Complex Spectrum Circle Centroid for Microphone-Array-Based Noisy Speech Recognition

Shigeki Sagayama, Takashi Okajima, Yutaka Kamamoto, Takuya Nishimoto
Graduate School of Information Science and Technology
The University of Tokyo
Hongo, Bunkyo-ku, Tokyo, Japan
E-mail: {`sagayama, t-okajima, kamamoto, nishi`}`@hil.t.u-tokyo.ac.jp`

Abstract

We propose a novel principle based on Complex Spectrum Circle Centroid (CSCC) for restoring complex spectrum of the target signal from multiple microphone input signals in a noisy environment. If noise arrives at multiple microphones with different time delays relative to the target signal, the observed noisy signals lie on a circle in the complex spectrum plane from which the target signal is restored by finding the centroid of the circle. Unlike most of existing methods for noise reduction such as ICA, AMNOR and beamforming, this non-linear operation is applicable to any type of noise including non-stationary, moving, signal-correlated, non-planar, and spoken noises, without identifying the noise direction and training parameters.

The proposed method was evaluated with speech recognition experiments in simulated noisy environments and was shown to improve the word accuracy close to the clean speech recognition rate of 89.4% in the case of a single spoken noise, and from 0% with one microphone to 60.6% with 8 microphones in the case of 3 spoken noises. The properties of this new method is further discussed theoretically and experimentally.

1. Introduction

This paper discusses a novel approach to microphone array signal processing based on a geometrical manipulation on the complex spectrum plane and gives some preliminary experimental results.

Microphone array signal processing is actively studied for various purposes such as improving speech recognition performance in noisy environments[1, 2]. The main idea is utilization of differences in path lengths from sources of target and noise signals to multiple microphones.

The simplest technique is Delay-and-Sum (DS) which adjusts delays added to microphone inputs so that the target signal from a particular direction synchronizes across multiple microphones while noises from different directions do not. This technique has an advantage that it requires no training, though it does not give a high performance in noise reduction.

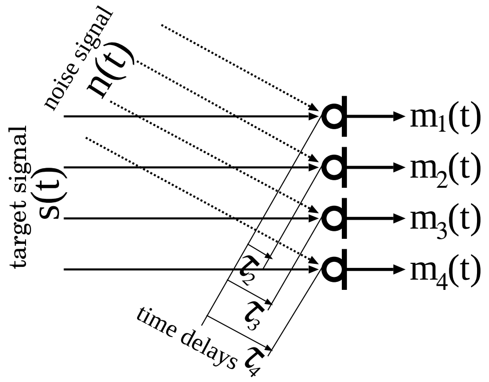

Figure 1: Sound propagation to the microphone array

On the other hand, adaptive types of microphone array signal processing such as Griffiths-Jim[3], AMNOR[4], and other adaptive beamforming methods require training the filter coefficients during a silent interval before its operation, though better performance can be obtained compared with DS. These methods often fail to track rapid changes of environmental characteristics including noise and reverberation, and result in poor improvements in noise reduction even compared with simple DS. Other methods based on blind source separation or independent component analysis assume statistical independence between signal and noise which is not always true.

These methods had mainly aimed at noise cancelation or reduction in the waveform domain. In speech recognition, however, noise reduction in the waveform domain is not necessary; instead, we need noise reduction in feature parameteres such as Mel-Frequency Cepstrum Coefficients (MFCCs) of noise-reduced speech.

In this paper, we focus on microphone array signal processing for noise-reduced spectrum estimation for speech recognition.

2. Complex Spectrum Circle Centroid

2.1. Complex Spectrum Representation of Microphone Inputs

We assume that acoustic characteristics (gains, directivities, etc.) of microphones are equal. If the target signal $s(t)$ propagates and arrives at K microphones simultane-

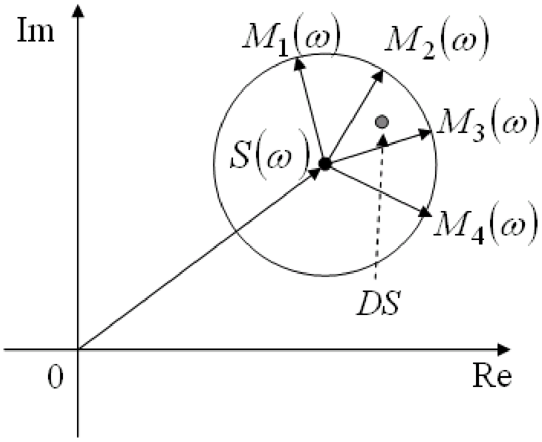

Figure 2: Microphone signals $M_i(\omega)$ located on a circle in the complex spectrum plane with the target signal $S(\omega)$ being the circle centroid

ously at time t while the noise signal $n(t)$ arrives with different time delays $\tau_1, \cdots, \tau_K$ as shown in Figure 1, the observed signal $m_i(t)$ at the i-th microphone is represented by:

$$m_i(t) = s(t) + n(t - \tau_i), \quad i = 1, 2, \cdots, M \tag{1}$$

where τ_i denotes the time delay at the i-th microphone in respect to the noise signal. Even if microphones are not layed out as in Figure 1 and their characteristics are not equal, microphone signals can be calibrated in terms of gain and time delay so as to synchronize to each other in respect to the target signal.

The short-time Fourier transform of the i-th microphone input signal is given by

$$M_i(\omega) = S(\omega) + N(\omega)e^{-j\omega\tau_i} \tag{2}$$

according to the basic properties of Fourier transform where ω denotes angular frequency and $M_i(\omega), S(\omega)$, and $N(\omega)$ denote Fourier transforms of $m_i(t), s(t)$, and $n(t)$, respectively, if the frame length is relatively longer than time differences of noise observations. From microphone signals, we easily obtain framewise complex spectrum of the short period (typically multiplied by a short-time window).

Geometrically, Eq. (2) means that $M_i(\omega)$ is lie on a circle of radius $\|N(\omega)\|$ with a centroid at $S(\omega)$ on the complex spectrum plane. The complex spectrum of target signal $S(\omega)$ is restored by finding the centroid of the circle on which K complex points $M_i(\omega)$ lie. We call this method of estimating the target signal spectrum “Complex Spectrum Circle Centroid (CSCC) method.”

In contrast, the Delay-and-Sum (DS) method uses the center of gravity (arithmetic mean) of microphone inputs: $\bar{M} = \frac{1}{K}\sum_{i=1}^{K} M_i(\omega)$.

2.2. Theoretical Properties of CSCC

Theoretically, the complex spectrum circle centroid (CSCC) has the following interest properties:

(1) Non-linear operation

It should be noted that finding the circle centroid from K complex points is not a linear operation on input signals. In this respect, this method is an entirely different approach from other approaches based on linear filtering.

(2) Frequency independence

The CSCC principle holds at any of frequency points independently without assuming any frequency characteristics of target and noise signals and microphones. The noise source direction need not to be the same across all frequencies. There is a future possibility of further improvement by assuming the same noise source direction over all frequencies.

(3) Correlated signal and noise

Even if the target signal and noise are statistically correlated, the above discussion still holds: i.e., this method does not need any assumption concerning independence between the target and noise signals. This feature significantly distinguishes the CSCC method from Independent Component Analysis (ICA).

(4) Non-planar wave propagation

Since the present principle is solely based on time differences between target and noise, it is applicable not only to planar, but also to spherical and any other wave propagations if differences in gain is negligible.

(5) Multiple noise sources

In principle, the circle centroid can handle only with a single noise source per frequency point. This means that different frequency components are allowed to come from different noise sources as stated in (2). Therefore, even if multiple noise sources exist and if one source is predominant over others per frequency point, the principle also works and the circle centroid is expected to be noise-reduced spectrum. This situation may really happen when multiple speech signals overlap where powers, formants, and pitch frequencies may differ from others.

2.3. Finding the Complex Spectrum Circle Centroid

It is obvious that the target signal spectrum $S(\omega)$ is restored by finding the centroid of the circle on which three or more microphone inputs $M_i(\omega)$ lie. In the case of $K = 3$, the circle centroid is uniquely determined from three distinct points on the circle. In the case of $K > 3$ microphone inputs, the circle centroid can be determined as a point of nearly equal distance from observed microphone inputs. We estimate the centroid as a point $\tilde{S}(\omega)$ by minimizing the variance of K squared distances from $M_i(\omega)$, i.e.,

$$\tilde{S}(\omega) = \underset{Z(\omega)}{\operatorname{argmin}} \operatorname{Var}\Big[\|Z(\omega) - M_i(\omega)\|^2\Big] \tag{3}$$

Table 1: Word recognition rates (%) compared with clean speech and Delay-and-Sum (DS) in simulated experiments

#noise sources (incident angles [deg])	clean (none)	1 noise (30)		2 noises (30, 270)		3 noises (30, 60, 270)		5 noises (30,60,120,180,270)	
#microphones	–	DS	CSCC	DS	CSCC	DS	CSCC	DS	CSCC
1	89.4	42.2		-5.7		-27.4		-27.5	
3	89.4	51.1	**87.1**	13.2	**57.0**	-12.2	**1.0**	-14.8	**-14.0**
4	89.4	54.1	**87.9**	10.5	**21.6**	-2.6	**20.1**	-8.1	**0.3**
8	89.4	55.1	**88.0**	22.9	**73.6**	7.7	**60.6**	8.2	**49.0**
16	89.4	55.7	**88.0**	25.3	**75.1**	11.2	**60.6**	18.6	**55.5**

for arbitrary $K = 4, 5, \cdots$. We can include cases of $K = 1, 2$, or 3 where the minimum variance is 0.

To solve this equation, let X and jY be real and imaginary parts of $Z(\omega)$, i.e., $Z = X + jY$, and let x_i and jy_i be those of $M_i(\omega) = x_i + jy_i$. Then, we have

$$\begin{aligned}
&\mathrm{Var}\left[\|Z - M_i\|^2\right] \\
&= \frac{1}{K}\sum_{i=1}^{K}\|Z - M_i\|^4 - \Big(\frac{1}{K}\sum_{i=1}^{K}\|Z - M_i\|^2\Big)^2 \\
&= \frac{1}{K}\sum_{i=1}^{K}\left((X - x_i)^2 + (Y - y_i)^2\right)^2 \\
&\qquad -\Big(\frac{1}{K}\sum_{i=1}^{K}\left((X - x_i)^2 + (Y - y_i)^2\right)\Big)^2 \quad (4)
\end{aligned}$$

from which its partial differentials in respect to X and Y is derived as:

$$\begin{aligned}
&\begin{pmatrix}\frac{\partial}{\partial X}\\ \frac{\partial}{\partial Y}\end{pmatrix}\mathrm{Var}\left[\|Z - M_i\|^2\right] \\
&= 8\begin{pmatrix}\mathrm{Var}[x_i] & \mathrm{Cov}[x_i,y_i]\\ \mathrm{Cov}[x_i,y_i] & \mathrm{Var}[y_i]\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix} - 4\begin{pmatrix}\mathrm{Cov}[x_i,x_i^2]+\mathrm{Cov}[x_i,y_i^2]\\ \mathrm{Cov}[y_i,y_i^2]+\mathrm{Cov}[y_i,x_i^2]\end{pmatrix} \quad (5)
\end{aligned}$$

denoting the covariance of a and b by $\mathrm{Cov}[a, b]$. Letting the lefthand side of the above equation be 0, we obtain a linear equation to obtain the centroid that minimizes the variance of squared distances in Eq. (3):

$$\begin{pmatrix}X\\ Y\end{pmatrix} = \frac{1}{2}\begin{pmatrix}\mathrm{Var}[x_i] & \mathrm{Cov}[x_i, y_i]\\ \mathrm{Cov}[x_i, y_i] & \mathrm{Var}[y_i]\end{pmatrix}^{-1}\begin{pmatrix}\mathrm{Cov}[x_i, x_i^2]+\mathrm{Cov}[x_i, y_i^2]\\ \mathrm{Cov}[y_i, y_i^2]+\mathrm{Cov}[y_i, x_i^2]\end{pmatrix} \quad (6)$$

whose solutions X and Y give the estimated complex spectrum centroid for each frequency as $\tilde{S}(\omega) = X(\omega) + jY(\omega)$.

Eq. (6) has a solution if the covariance matrix:

$$C_{xy} \equiv \frac{1}{2}\begin{pmatrix}\mathrm{Var}[x_i] & \mathrm{Cov}[x_i, y_i]\\ \mathrm{Cov}[x_i, y_i] & \mathrm{Var}[y_i]\end{pmatrix} \quad (7)$$

between x_i and y_j is regular. Since its determinant is given by

$$|C_{xy}| = \mathrm{Var}[x_i]\,\mathrm{Var}[y_i](1 - r_{xy}^2),\ r_{xy} = \frac{\mathrm{Cov}[x_i, y_i]}{\sqrt{\mathrm{Var}[x_i]\,\mathrm{Var}[y_i]}} \quad (8)$$

where r_{xy} is the correlation coefficient between x_i and y_j, the solution of Eq. (6) is guaranteed to exist unless $r_{xy} = 1$, i.e., all spectrum points $M_i(\omega),\ i = 3, 4, \ldots, K$ lie on a line in the complex plane.

Even though r_{xy} is always guaranteed to be no greater than 1, in numerically bad conditions such as $r_{xy} > 0.99$, we use the center of gravity of K points $M_i(\omega)$, i.e., the delay-and-sum solution, instead of the circle centroid here.

2.4. Noisy Speech Recogniton Using CSCC

As the complex spectrum of the target signal is restored from signals of mutiple microphones for each of frequency points, the mel-filter bank outputs of the target signal (clean speech) are calculated by making weighted sums of restored spectrum $\tilde{S}(\omega)$ of the target signal according to the mel-scaling. They are Fourier transformed to Mel-Frequency Cepstrul Coefficients (MFCCs) which is widely used as the feature vector for speech recognition.

3. Experimental Evaluation of CSCC

3.1. Experimental Conditions

Continuous speech recognition experiments were performed to evaluate the performance of the CSCC method to recognize Japanese sentense speech in noisy environment using an microphone array. We used "IPA-testset" consisting of 100 sentenses each uttered by male and female speakers excerpted from ASJ-JNAS corpus of read newspaper articles as the test set and other 10 sentense utterances from the same database as 1 to 5 additive noises with a signal-to-noise ratio of 10dB per noise, i.e., $\log_{10}(10/\#\mathrm{Noises})$ if more than one noise sources exist.

Input speech were analyzed with a 25-mS frame rate and 10-mS frame shift. 12-th order MFCCs, their ΔMFCCs and Δlog-power were used as acoustic feature vector. Using "Julius3.3p3"[5] as the speech recognition platform, word accuracy was evaluated as the measure of speech recognition performance. Performances for clean speech and delay-and-sum method were compared as references.

Table 2: Word accuracy in preliminary experiments in a reverberant real environment

	1 microphone	DS	CSCC
clean	37.4	**61.1**	59.5
noisy	9.3	34.1	**38.2**

3.2. Simulated microphone array

In simulated microphone array experiments, 3 to 16 micorphones were equally spaced on a circle with a diameter of 30cm. The target and noise sources were assumed to be within the same plane with the microphone array.

Word accuracies are shown in Table 1. "1 microphone" means the performance without using microphone array. It has been shown that the CSCC method outperforms the Delay-and-Sum method in all conditions and yields high performances near to clean speech recognition.

3.3. Preliminary experiment in a reverberant room

For evaluation in a realistic environment, we used 16 microphones placed at 4 × 4 mesh points with spans of 10cm in a reverberant room with computer noise present. The target sound arrived from a 1m distance in the direction vertical to the microphone plane. Noises sounded at the 10dB signal-to-noise ratio disregarding reverbaration and computer noises in the room.

The recognition results are shown in Table 2. Compared with the case of one microphone, CSCC method remarkably improved the performance, though it did not work significantly better than the Delay-and-Sum method, probably due to the reverberant condition of the room.

4. Discussion and Future Works

The proposed method has high potentials of further modification for even higher performance.

(1) Interpolation between circle and gravity centroids

Depending on microphone spacing, there may be frequencies at which the input spectrum points $\{ M_i(\omega) \}$ overlap or gather and do not form a circle. In addition, the input spectrum points may deviate from the ideal points due to inaccurate layout and unequal characteristics microphones, fast change in the noise signal compared to time differences between microphones, and other sources of errors. These deviations may distort the circle and cause inaccurate centroids.

Such ill-conditioned situation is diagnosed through multiple clues such as correlation coefficient in Eq. (8) or spanning angles between input points from the estimated centroid:

$$\cos\theta = \frac{(Z - M_i, Z - M_j)}{|Z - M_i| \cdot |Z - M_j|}$$

and is relieved by interpolating the circle centroid and center of gravity (arithmetic mean) $\frac{1}{K}\sum M_i(\omega)$.

(2) Microphone layout

Suppose that microphones are equally spaced in a line. If noise arrived with an incident angle yielding the time difference τ between adjacent microphones, the input complex spectrum points $M_i(\omega)$ gather on a circle in the complex plane at a frequency $\omega = 2\nu\pi/\tau$, $\nu = 1, 2, 3, \cdots$ in Eq. (2). The microphone layout can be improved to avoid such ill conditions. In this paper, microphones were equally spaced on a circle in the simulated speech recognition.

(3) Calibration and normalization

In this paper, microphones are supposed to have identical characteristics. If the gain characteristics along frequency is not equal among microphones, they can be equalized by normalizing the gain of each microphone at each frequency point. Directivity of the microphone is supposed to be identical but need not be omni-directional. Otherwise, calibration is not simple in Eq. (2).

5. Conclusion

We proposed the Complex Spectrum Circle Centroid (CSCC) method for restoring complex spectrum of the target signal from multiple microphone input signals in noisy emvironments. Unlike most existing methods, this method can handle with correlated signal and noise, non-planar wave propagations, and any type of noise without locating the noise sources or training filter coefficients. The noise reduction process is non-linear to the input and independent between different frequencies. It has been experimentally shown to be effective for multiple noises.

The proposed method was evaluated in simulated noisy speech recognition experiments and shown to be significantly effective not only in the case of single noise but also in multiple noises.

This new method is still an on-going work to be further explored both theoretically and experimentally.

6. References

[1] J. Bitzer, K.U. Simmer, K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," Speech Communication, vol. 34, pp. 3–12, 2001.

[2] M.L. Seltzer, B. Raj, "Speech-Recognizer-Based Filter Optimization for Microphone Array Processing," IEEE Signal Processing Letters, vol. 10, no. 3, 2003.

[3] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propag., vol. AP-30, no. 1, pp. 27–34, 1982.

[4] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," IEEE Trans. Acoust. Speech Signal Process., vol. ASSP-34, no. 6, pp. 1391–1400, 1986.

[5] A. Lee et al., "Julius - an Open Source Real-Time Large Vocabulary Recognition Engine," Proc. Eurospeech2001, pp. 1691–1694, 2001.