Low Distortion Target Speech Extraction Method Using Two-Channel Microphone System

Takaaki Ishibashi^{a,*}, Kosuke Inoue^b, Chiharu Okuma^c, Kei Eguchi^d

^aDepartment of Information, Communication and Electronic Engineering,

^bAdvanced Course of Electronics and Information Systems Engineering,

^cDepartment of Human-Oriented Information Systems Engineering,

National Institute of Technology, Kumamoto College

2659-2 Suya, Koshi, Kumamoto 861-1102, Japan

^dDepartment of Information Electronics, Fukuoka Institute of Technology

3-30-1 Wajirohigashi, Higashi-ku, Fukuoka 811-0295, Japan

*Corresponding Author: ishibashi@kumamoto-nct.ac.jp

Abstract

This paper proposes a low distortion target speech signal extraction method. The authors have been proposed some target speech extraction methods using two-channel microphone system. The methods are based on a variable arbitrary directional pattern for multiple sound signals. And the pattern is directed to the target signal. However, when the directivity becomes a sharp pattern, the musical noise occurs in the estimated signal. Therefore, we propose a new target extraction method with a null pattern to the noise. The distortion of the estimated signal using the proposed method is low, since the influence of the target sound is low. By several simulations, it is found that the proposed method can extract the target speech from observed multi speaker's utterance sounds under a noise.

Keywords: noise reduction, target speech extraction, variable arbitrary directivity pattern, two-channel microphone system

1. Introduction

Many noise reduction methods in a real-time processing such as SS (Spectral Subtraction)⁽¹⁾, SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones)⁽²⁾ and MUSIC (Multiple Signal Classification)⁽³⁾ have been proposed. These methods can estimate the original source signals. However, these algorithms need a lot of data points since the methods are based on the stochastic theory and used the frequency domain information using Fourier transform.

The authors have already proposed some systems to reduce noise from mixture signals observed with two-channel microphone system⁽⁴⁾. These systems are implemented with a microcontroller and can suppress noise in real-time. The systems extract only one target sound, all other sounds was deleted as noise. In actual application, when multiple speakers utter in a noisy environment, the multi-speaker's speech become target sounds. It means that we need multiple speaker's speech extraction methods under a noisy environment. In addition, a simple two-channel microphone system with variable arbitrary directional pattern has proposed by the authors⁽⁵⁾. However, the musical noise occurred in the estimated signal with the method, when the directivity becomes sharp.

This paper proposes a low distortion target sound speech signal extraction method. The proposed method makes a null pattern for the noise. The distortion of the estimated signal with the proposed method is low since the influence of the target sound is low. In addition, the method can make several null patterns for a multi-source condition. A typical microphone forms unidirectional directivity pattern. Therefore, when multiple speakers speak at the same time, it is necessary to place as many microphones as the number of speakers. Our method works well even when the number of sound sources is larger than that of microphones. The target speech of one person or plural persons are emphasized by our method. And other speech and noise are reduced. Since the proposed method is a very simple algorithm and the amount of calculation is small, it can be expected to be applied to industries.

2. Principle of variable directional pattern

When N source signals $s_n(t)$ ($n = 1, 2, \dots, N$) are observed by M microphones, the observed mixture signals $x_m(t)$ ($m = 1, 2, \dots, M$) are expressed as

$$x_m(t) = \sum_{n=1}^N a_{mn} s_n(t) \quad (1)$$

where a_{mn} denotes an unknown mixing parameter, N and M denote the number of the sources and the microphones, respectively.

Using two mixture signals observed by two microphones in the case of two speaker's utterances at the same time, a joint distribution is plotted as Fig. 1(a). The horizontal and vertical axes denote the amplitude of $x_1(t)$ and $x_2(t)$, respectively. In the following, we consider the sound signals as stochastic varieties and omit the time series t .

From the Fig. 1(a), it is found that the joint distribution has two linear components. For clarity the linear components, directions ϕ of distribution are calculated as follows.

$$\phi = \tan^{-1} \frac{x_2}{x_1} \quad (2)$$

Using all ϕ , its histogram $\text{hist}(\phi)$ as shown in Fig. 1 (b) means that the components of the joint distribution are concentrated two directions. The reason is that since the human speech has a silent interval, it becomes only one source in many data points.

In the case of three active sources, a joint distribution and a histogram observed two microphones are shown in Fig. 2(a) and (b), respectively. In these figures, the dense crossing lines are still discernible and three peaks are recognizable. These values of the inclinations of lines depend on the transfer functions from the sources to the microphones.

Fig. 1: Joint distribution and histogram ($N = 2$).

Fig. 2: Joint distribution and histogram ($N = 3$).

From the above discussions, in the case which the target source signal and the noise signals exist at the same time, the joint distributions made from the observed signals have straight lines depend on transfer functions of each source. Then, it means that the target source signal can be estimated by extracting the straight line.

To extract the target speech using the joint distribution means that to reduce the straight line components generated by the noise, and to extract the straight line components created by the target signals. Based on the fact, the authors have proposed a noise reduction method based on the ratio of observed signals by two microphones⁽⁴⁾. We define the error ε as

$$\varepsilon = \Phi - \phi, \quad (3)$$

where Φ denotes the direction of the target source. From the above discussion, Φ is estimated as the mode value (the most frequent value) of ϕ as follows.

$$\Phi = \arg \max_{\phi} \text{hist}(\phi) \quad (4)$$

In the case which the target source signal is multiple speakers speech, we estimate another peak Φ_k again excluded the mode value Φ and neighbor value.

In order to reduce sounds in a direction away from the target source, we use some functions $f_i(\varepsilon)$ that gain decreases as the ε increases as

$$f_1(\varepsilon) = \frac{1}{1 + \alpha \varepsilon^2} \quad (5)$$

$$f_2(\varepsilon) = -0.5 \tanh(\beta \varepsilon^2 - \gamma) + 0.5 \quad (6)$$

$$f_3(\varepsilon) = \begin{cases} 1 & \text{if } \varepsilon^2 \leq \delta \\ 0 & \text{if } \varepsilon^2 > \delta \end{cases} \quad (7)$$

where α, β, γ and δ denote learning parameters for determining the amount of noise reduction. Fig. 3 (a), (b) and (c) show

(a) $f_1(\varepsilon)$

(b) $f_2(\varepsilon)$

(c) $f_3(\varepsilon)$

Fig. 3: Noise reduction functions.

orthogonal coordinates and polar coordinates of $f_1(\varepsilon)$, $f_2(\varepsilon)$ and $f_3(\varepsilon)$, respectively.

As a function with multiple directivity pattern, we adopt the Rose curve g_l as follows,

$$g_l = \begin{cases} \sin l(\theta - \varphi) & \text{if } l : \text{odd} \\ \sqrt{\sin l(\theta - \varphi)} & \text{if } l : \text{even} \end{cases} \quad (8)$$

where θ denotes the angle from 0 to 360 degrees and φ denotes the phase.

Using the noise reduction functions $f_i(\varepsilon)$ and the Rose curve g_l , a variable arbitrary directivity pattern has proposed as follows⁽⁵⁾ and some directivity patterns are shown in Fig. 4.

$$y = g_l f_i(\varepsilon) x_m \quad (9)$$

Fig. 4: Multiple target extractable functions.

From these figures, the noise reducing functions $g_l f_i(\varepsilon)$ mean spatial filters on the ratio of the transfer functions. By the noise reducing function, only the signal having the specific direction according to the transfer functions is extracted. Signals having other directions of the transfer functions are suppressed. Even when the number of sound sources is larger than the number of observed mixture signals, our proposed method can reduce noise signals. In addition, since the algorithm of our noise reduction method is very simple, the method works well in real-time. However, when the directivity becomes a sharp pattern, a distortion occurs in the estimated speech signal. Therefore, we propose a low distortion target sound speech signal extraction method in the next section.

3. Variable null pattern

In the previous method, the directivity was directed to the target signal. It means that the spatial filtering was applied around the target signal. Therefore, in order to reduce the influence on the target signal, a null pattern is formed in the noise direction.

The null patterns as shown in Fig. 5 is proposed using Eq (10) as follows.

Fig. 5: Null pattern functions.

Fig. 6: Simulation results

$$h = 1 - g_l f_i(\varepsilon) \quad (10)$$

The source signal is estimated using the null pattern as

$$y = h x_m. \quad (11)$$

4. Simulation

In order to verify our proposals, several simulations were carried out. The sources were speaker speech data⁽⁶⁾ and car noise⁽⁷⁾. Fig. 6 (a) shows these source signals. The signals were sampled at the rate of 8 [kHz] with 16 [bit] resolution. Using these sources, the mixture signals are generated by Eq (1) are shown in (b). The separated signal $y(t)$ using the proposed method is shown in (c). It is found that noise is removed in this waveform. As a results of using 30 mixing patterns (a combination of 6 utterance signals and 5 noises), it is conformed that the proposed method works well.

5. Conclusion

This paper proposes the two-channel microphone system with the variable null pattern. The proposed method can reduce the noise signal by deleting the distribution except a linear component depending on the target speech. The distortion of the estimated signal was lower than that of the previous method. The method can extract the target speech, even when

the number of source signals is larger than that of observed mixture signals. Furthermore, the method can estimate multi-speakers speech without noise using multi-directional pattern, when multiple speaker's utterance under a noisy environment. In this paper, we simulated under the condition that the direction of the noise is known. Direction of the noise is expected to be estimated using the histogram. The direction estimation of noise will be examined in the future.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 16K21583.

References

- (1) S. F. Boll: "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, No. 2, pp. 113-120, 1979
- (2) M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda: "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," Acoustical Science and Technology, Vol. 22, No. 2, pp. 149-157, 2001
- (3) Y. Sugimoto, S. Miyabe, T. Yamada, S. Makino and B.-H. Juang: "An extension of MUSIC exploiting higher-order moments via nonlinear mapping," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E99-A, No. 6 pp. 1152-1162, 2016
- (4) K. Hayama, T. Ishibashi, C. Okuma and H. Gotanda: "Implementation of directional characteristics by real-time processing of sounds observed by two microphones," ICIC Express Letters, Vol. 10, No. 1, pp. 251-254, 2016
- (5) T. Ishibashi, K. Higuchi, C. Okuma and K. Hayama: "Noise reduction method for multiple sound source signals and its application to two-channel microphone system," Proceedings of the 5th IIAE International Conference on Industrial Application Engineering 2017, pp. 336-339, 2017
- (6) Acoustical Society of Japan: "ASJ continuous speech corpus Japanese newspaper article sentences", JNAS Vols. 1-16, 1997
- (7) NTT Advanced Technology Corporation: "Ambient noise database for telephony 1996", 1996