FREQUENCY DOMAIN MULTI-CHANNEL SPEECH SEPARATION AND ITS APPLICATIONS

Masaki Handa, Takayuki Nagai, and Akira Kurematsu

Course in Electronic Engineering, The University of Electro-Communications, Tokyo, Japan.
Email:{masaki, tnagai, kure}@apple.ee.uec.ac.jp

ABSTRACT

In this paper, a multi-channel speech separation method for real environments is proposed. The proposed method is based on frequency assignment, namely the magnitude of each channel at the same frequency bin is compared with each other and it is assigned to the channel, to which it originally belongs. This method is a direct consequence of frequency domain interpretation of the eigendecomposition method proposed by Cao *et al.*[1]. Furthermore, our proposed method does not require eigendecomposition, which consumes the costs of computation.

We also present two example applications of the proposed method, that is, voice controlled computers in a multi-user environment and a noise removal in cellular phone using two microphones.

1. INTRODUCTION

Multi-channel signal separation has been widely noticed and studied recently[1]-[5]. A lot of applications have been found in area of the engineering such as multi-channel data communications, sonar array processing, biomedical signal processing, and speech processing. Especially, multi-channel speech separation is an interesting and an important challenge, since it is useful for separating competing speakers, denoising, and so forth. Thus, it is successfully applicable to the pre-processing of the speech recognition system, which is being used as a human interface. Among variety of methods, ICA (Independent Component Analysis) is a promising technique, which can solve the problem by assuming the independence of the sources. However, ICA has some drawbacks like high computational cost, ambiguity of the permutation and amplitude, and stability. Ikeda and Murata proposed ICA in time-frequency domain[2]. Although it works well for the speech recorded in a real environment (convolutive mixtures), the costs is not low enough to implement it in a cellular phone, PDA and so on.

In this paper, a multi-channel speech separation method for real environments is proposed. The proposed method is based on frequency assignment, namely the magnitude of each channel at the same frequency bin is compared with each other and it is assigned to the channel, to which it originally belongs. This method is a direct consequence of frequency domain interpretation of the eigendecomposition [1]. Furthermore, our proposed method does not require eigendecomposition, which consumes the costs of computation.

We also present two example applications of the proposed method, that is, voice controlled computers in a multi-user environment and a noise removal in cellular phone using two microphones.

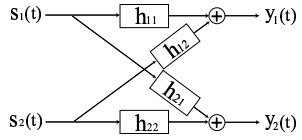

Figure 1: Acoustic model of two channel case.

2. PRELIMINARY

2.1. Multi-Channel Speech Separation

Fig.1 shows the acoustic model of two-channel case, where $s_i(t)$ and $y_j(t)$ represent the i-th source signal and the j-th observation, respectively. h_{ij} denotes the impulse response of the acoustic path from j-th source (speaker) to i-th sensor (microphone). The model can be written in matrix form as

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \otimes \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix} \tag{1}$$

where '$\otimes$' represents linear convolution. The problem of multi-channel speech separation can be formulated as recovering source signals $s_i(t)$ (or $h_{ii} \otimes s_i(t)$) from multi-channel observations $y_j(t)$.

2.2. Signal Separation by Eigendecomposition

Cao *et al.* have proposed two-channel speech separation method based on eigendecomposition[1]. The method processes a speech frame by frame base, therefore it seems to be suitable for real time applications. Here, we refer to the method. For more details, please see [1].

In this method, the correlation matrix of each observation is calculated as

$$\boldsymbol{R}_{Y_i} = \boldsymbol{Y}_i \boldsymbol{Y}_i^*, \tag{2}$$

where

$$\boldsymbol{Y}_i = \begin{pmatrix} y_i(1) & \dots & y_i(N) & 0 & \dots & 0 \\ 0 & y_i(1) & \dots & y_i(N) & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & y_i(1) & \dots & y_i(N) \end{pmatrix}, \tag{3}$$

and '$*$' denotes the conjugate transpose. N represents the length of a frame. Then, new matrix $\boldsymbol{R}_{ratio}$ is built from the correlation matrices.

$$\boldsymbol{R}_{ratio} = \boldsymbol{R}_{Y_1}^{-1} \boldsymbol{R}_{Y_2}. \tag{4}$$

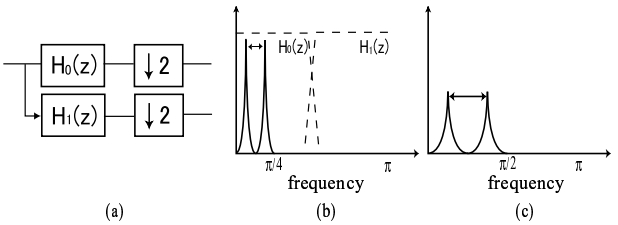

Figure 2: Basic idea of subband eigendecomposition. (a) Analysis filter bank. (b) Filtering by analysis filters. (c) After decimation.

The eigenvectors and the eigenvalues of $\boldsymbol{R}_{ratio}$ have following two properties;

1. Signal subspace spanned by eigenvectors of $\boldsymbol{R}_{ratio}$ coincide with a subspace spanned by the eigenvectors of $\boldsymbol{R}_{y_1}$ and $\boldsymbol{R}_{y_2}$.
2. Eigenvalues of $\boldsymbol{R}_{ratio}$ equals to the ratio of corresponding power densities for each signal component of the two signals.

From property 1, the eigenfilter built from an appropriate eigenvector can remove the frequency components originated from other signal. The appropriate eigenvector is selected by the property 2.

In [1], IIR notch filter, which removes the undesired signals, is designed from the eigenfilter (selected eigenvector). It should be noted that the design process of IIR filter requires the polynomial rooting.

3. PROPOSED METHOD

In this section, we extend the eigendecomposition method to a subband eigendecomposition in order to improve the performance of the speech separation. The frequency assignment is motivated by the subband eigendecomposition.

3.1. Subband Eigendecomposition

Since the eigendecomposition method uses the IIR notch filter, it fails to remove the frequency component of the other source when undesired source contains really close frequency component to that of the desired source signal. To overcome this difficulty, we consider applying the eigendecomposition to band limited signals, that is, subband signals in a multi-rate filter bank. Fig.2 illustrates this idea for two-channel multi-rate filter bank, which spreads the distance between two frequency components due to the decimation (Fig.2(c)). It makes easier to remove the frequency component of the other source by IIR notch filters. Fig.3(a) shows the example of sinusoidal waves whose frequencies are close each other. From the figure, one can see that the subband eigendecomposition can separate two sinusoidal waves (fig.3(c)) while the conventional eigendecomposition fails to do so (fig.3(b)).

3.2. Speech Separation by Frequency Assignment

The above subband eigendecomposition can improve the performance of speech separation. However, the algorithm still requires eigendecomposition and its extension to three or more channels is not straightforward.

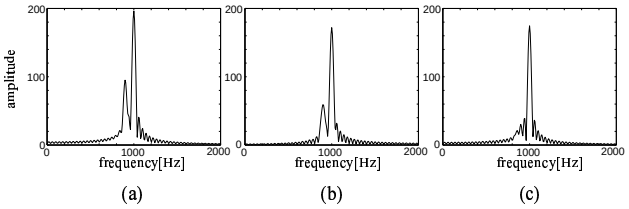

Figure 3: Example of two sinusoidal waves separation. (a) Mixed input signal. (b) Separated by eigendecomposition. (c) Separated by subband eigendecomposition.

Now, assuming that each subband is narrow enough to contain only one frequency component. Under this assumption, IIR notch filter is not required anymore, since the frequency component is simply judged which channel it should belong to. This is the basic idea of our proposed frequency assignment. Obviously, source signals must be assumed to have different frequency components from each other, i.e. they never share the common frequency component. Although this assumption is more restrictive than orthogonality and is not absolutely true, it approximately holds for speech signals if the band width of each subband is narrow enough.

If the delay and the reflections are not too long, Eq.(1) is approximated in time-frequency domain as

$$\begin{pmatrix} y_1(f,t) \\ y_2(f,t) \end{pmatrix} = \begin{pmatrix} h_{11}(f) & h_{12}(f) \\ h_{21}(f) & h_{22}(f) \end{pmatrix} \begin{pmatrix} s_1(f,t) \\ s_2(f,t) \end{pmatrix}, \qquad (5)$$

where f and t denote frequency and time index of the frame, respectively. By using the assumption we made, Eq.(5) can be further rewritten as

$$\begin{cases} y_1(f,t) = h_{11}(f)s_1(f,t) \\ y_2(f,t) = h_{21}(f)s_1(f,t) \end{cases}, \qquad (6)$$

$$or$$

$$\begin{cases} y_1(f,t) = h_{12}(f)s_2(f,t) \\ y_2(f,t) = h_{22}(f)s_2(f,t) \end{cases}. \qquad (7)$$

Therefore, the ratio matrix Eq.(4) becomes as follows;

$$R_{ratio}(f) = \begin{cases} \frac{|h_{21}(f)|^2}{|h_{11}(f)|^2}, & for\ Eq.(6) \\ \frac{|h_{22}(f)|^2}{|h_{12}(f)|^2}, & for\ Eq.(7) \end{cases}. \qquad (8)$$

It should be noted that eigendecomposition is not necessary, since $R_{ratio}(f)$ is a scalar format in this case. In order to decide which channel the frequency component f belongs to (Eq.(6) or Eq.(7)), we must consider two situations.

First scenario is two speakers (sources) are located close to their own microphones. That means $|h_{11}(f)| > |h_{21}(f)|$ and $|h_{22}(f)| > |h_{12}(f)|$ hold true practically, because of the attenuation of sound. Therefore we can assign the frequency component f according to the value of $R_{ratio}(f)$. When $R_{ratio}(f) > 1$, the frequency component is assigned to $s_2(f,t)$, then $y_1(f,t)$ is set to zero. If $R_{ratio} < 1$, the component is assumed to be originated from $s_1(f,t)$. Therefore $y_2(f,t)$ is set to zero. Fig.4 illustrates the proposed frequency assignment.

Another situation is that we never know the relationship between $R_{ratio}(f)$ and observations. This situation

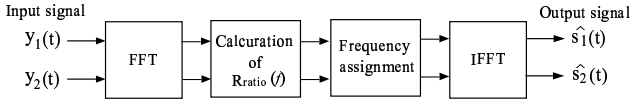

Figure 4: Two channel signal separation by frequency assignment.

happens when both speakers are located close to one microphone or two microphones are closely located. In such case, different assignment rule is required. We consider this problem later (in Sec.4).

The above two channel frequency assignment can be easily extended to three or more channel separation. In this case, multiple $R_{ratio}(f)$'s must be calculated using all combination of the observations. However, this can be obviously replaced by a maximum search as

$$k = \arg\max_i \{y_i(f,t)\}, \tag{9}$$

Then we set $s_i(f,t)$ to zero for $i \neq k$ and leave $s_k(f,t)$ as it stands.

To show the performance of our proposed algorithm, we present the comparison between ICA[2] and frequency assignment. In order to be fair, mixed speech and separated speech by ICA were taken from www[6]. Fig.5(a) and (b) show mixed input signals, which were recorded in a real environment. Fig.5(c) and (d) are separated speech waveforms by ICA. The separated signals by our proposed method are illustrated in Fig.5(e) and (f). Since the original signal is unknown, objective values such as SNR cannot be used. Therefore, we take a following subjective evaluation called MOS (Mean Opinion Score). Ten people, whom are unaware of the experiment, heard the input and output(processed) speech to evaluate the improvement by ranking the result. The ranks are defined as 5 being very good and 1 being bad. The score in each channel of ICA is 1.7 and 1.6, whereas 3.2 and 3.1 is obtained by the proposed method. This results shows that frequency assignment method is more effective.

4. APPLICATIONS

In this section, two experiments using the proposed method are presented. First, an experiment for channel separation in a multi-user environment. Next, a noise removal from cellular phone using two microphones. These experiments are undertaken with an intension to implement the proposed method into an application.

4.1. Voice Controlled Computers in Multi-User Environment

In multi-user environments, like computer rooms, offices, and laboratories, where many people share the same space, people are constantly speaking simultaneously. In these environments, voice of other people may degrade the recognition rate of the speech recognition. By using our proposed method to separate the speech, it is applicable to the pre-processor of a recognition system. We have tested the speech separation system with four microphones in a computer room. Four non-directional microphones are located side by side in a room, size of 7.2[m] by 5.5[m], as

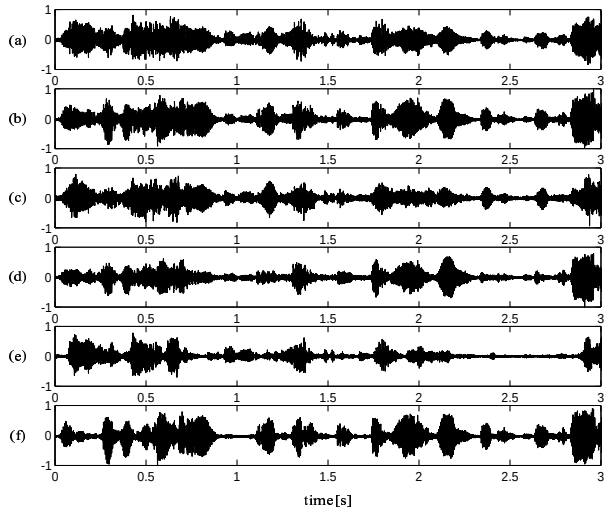

Figure 5: Comparison between ICA and proposed method. (a) Mixed input signal $y_1(t)$. (b) Mixed input signal $y_2(t)$. (c) Separated speech $\hat{s_1}(t)$ by ICA. (d) Separated speech $\hat{s_2}(t)$ by ICA. (e) Separated speech $\hat{s_1}(t)$ by proposed method. (f) Separated speech $\hat{s_2}(t)$ by proposed method.

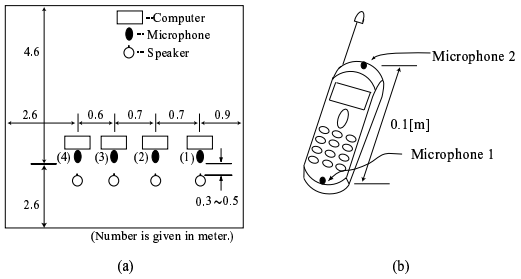

Figure 6: Location of microphones (a) In a room. microphones are named (1), (2), (3) and (4) from right to left. (b) Placed in cellular phone.

shown in Fig.6(a). The distances between the microphones are 0.6[m], 0.7[m], 0.7[m] starting from left. Male speakers are seated in front of each microphone, and the distance between a microphone and a person is ranging from 0.3 to 0.5[m]. Condition of this experiment is mentioned below. Signals are sampled at 8000[Hz], with a frame length of 25[ms] and an overlap length of 10[ms]. Each frame is pre-processed with a hamming window.

Because each microphone is located directly in front of the speakers, voice of the speaker seated in front is captured as a main source. In this case, the frequency assignment is used. Captured signals from the microphones are shown in Fig.7(a)-(d).The waveform of the separation result is shown in Fig.8(a)-(d). Waveform of the output signals present that our proposed method is effective.

4.2. Noise Removal in Cellular Phone

Next, the proposed algorithm is applied to the noise removal in a cellular phone. The experimental conditions and results are described below. The microphones and conditions

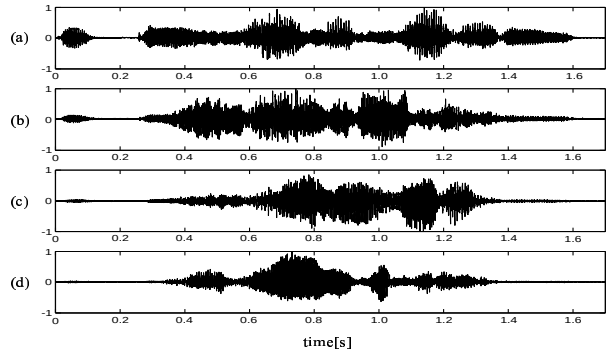

Figure 7: The waveform of input signals. (a) $y_1(t)$ of microphone(1). (b) $y_2(t)$ of microphone(2). (c) $y_3(t)$ of microphone(3). (d) $y_4(t)$ of microphone(4).

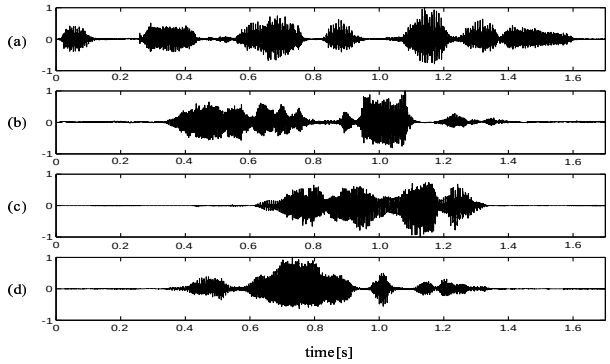

Figure 8: The waveform of output signals. (a) $\hat{s_1}(t)$. (b) $\hat{s_2}(t)$. (c) $\hat{s_3}(t)$. (d) $\hat{s_4}(t)$.

of this experiment were same as those in the previous experiment. Location of microphones is illustrated in Fig.6(b). The Distance between the microphones was set to approximately 0.1[m]. In this experiment, two male speakers are spoke in a normal manner. One male is making a phone call, while the other, standing about 1[m] to his left, is speaking to the person with the phone. The experiment was carried out in our laboratory (indoor). In this application, we cannot assume that the two sources are located close to their own microphones. Therefore the second scenario, must be considered as pointed out in the previous section.

Hence, we now can argue that the source $s_2(t)$ position is much farther than the distance between microphones, it is possible to assume the transfer functions $h_{12}(f)$, $h_{22}(f)$ are similar. Thus, the amplitudes of background signals into two microphones are also similar. While, the source $s_1(t)$ is located by the microphone, making transfer function $h_{11}(f)$, $h_{21}(f)$ distinct. Here, we define frequency subtraction $\Delta y(f,t)$.

$$\Delta y(f,t) = |y_1(f,t)| - |y_2(f,t)| \tag{10}$$

By using the properties of $h_{11}(f)$, $h_{12}(f)$, $h_{21}(f)$, and $h_{22}(f)$, classification of frequency is possible with $\Delta y(f,t)$. We use k-means algorithm to classify the $\Delta y(f,t)$ into two classes, namely, large and small classes. If $\Delta y(f,t)$ is classified as a large class, we can assume that the frequency component

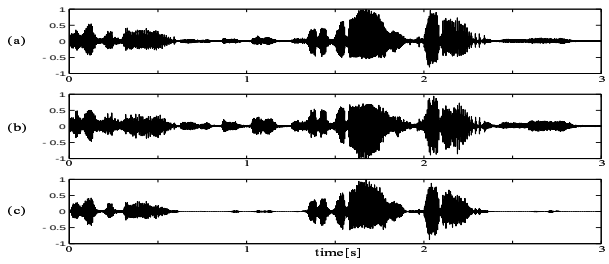

Figure 9: The waveform of signals. (a) Input signal from microphone1 $y_1(t)$. (b) Input signal from microphone2 $y_2(t)$. (c) Separated signal $\hat{s_1}(t)$.

f is originated in the desired source. Therefore, we leave $y_1(f,t)$ as is. When $\Delta y(f,t)$ belongs to small class, the frequency component f is considered to be an interference. Thus, we set $y_1(f,t)$ to zero. The waveform of the result is shown in Fig.9. From Fig.9(c), it can be seen that the interference is removed almost completely. The MOS of our proposed method is 4.3, whereas 2.8 and 2.1 are obtained by the conventional eigendecomposition and Fast ICA[4], respectively. The result shows that our proposed method is effective.

5. CONCLUSION

We have described a method of multichannel signal separation in frequency domain. The proposed method is based on the frequency assignment and is motivated by the eigendecomposition. We also present two applications of the proposed method. The experimental results show the validity of our proposed method. Speech recognition evaluation using data, processed by our proposed algorithm, is left for future research.

6. REFERENCES

[1] Y.Cao, S.Sridharan and M.Moody, "Multichannel Speech Separation by Eigendecomposition and Its Application to Co-Talker Interference Removal", IEEE Trans. on Speech & Audio Processing, vol.5, no.3, pp.209-219 (May 1997)

[2] S.Ikeda and N.Murata, "A Method of ICA in Time-Frequency Domain", Proc. of International Workshop on ICA and BSS, pp.365-371 (Jan. 1999)

[3] S.Shamsunder and G.B.Giannakis, "Multichannel Blind Signal Separation and Reconstruction", IEEE Trans. on Speech & Audio Processing, vol.5, no.6, pp.515-528 (Nov. 1997)

[4] A.Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", IEEE Trans. on Neural Networks, vol.10, no.3, pp.626-634 (May 1999)

[5] E.Weinstein, M.Feder and A.V.Oppenheim, "Multi-Channel Signal Separation by Decorrelation", IEEE Trans. on Speech & Audio Processing, vol.1, no.4, pp.405-413 (Oct. 1993)

[6] http://www.mns.brain.riken.go.jp/%7eshiro/blindsep.html (V-1 Speech - Speech)