

Speech Recognition of Double Talk using SAFIA-based Audio Segregation

Toshiyuki Sekiya, Tetsuji Ogawa and Tetsunori Kobayashi

Dept.EECE, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
{sekiya,ogawa,koba}@tk.elec.waseda.ac.jp

Abstract

Double-talk recognition under a distant microphone condition, a serious problem in speech applications in a real environment, is realized through use of modified SAFIA and acoustic model adaptation or training.

The original SAFIA is a high-performance audio segregation method based on band selection using two directivity microphones. We have modified SAFIA by adopting array signal processing and have realized optimal directivity for SAFIA. We also used generalized harmonic analysis (GHA) instead of FFT for the spectral analysis in SAFIA to remove the effect of windowing which causes sound-quality degradation in SAFIA.

These modifications of SAFIA enable good segregation in a human auditory sense, but the quality is still insufficient for recognition. Because SAFIA causes some particular distortion, we used MLLR-based acoustic model adaptation and immunity training to be robust to the distortion of SAFIA. These efforts enabled 76.2% word accuracy under the condition that the SN ratio is 0 dB, this represents a 45% reduction in the error obtained in the case where only array signal processing was used, and a 30% error reduction compared with when only SAFIA-based audio segregation was used.

1. Introduction

Hands-free speech recognition, in which the microphone is mounted on the terminal side rather than the user's body, has a wide range of applications, including situations in which many users share the system and may speak simultaneously. A personal robot serving a family in a household is a good example of this. To realize hands-free speech recognition in a real environment requires speech enhancement or separation of speech recorded with a microphone at a distance. Speech separation is particularly difficult in a situation where the signal-to-noise ratio is zero or worse. Several efforts have been made to solve this problem [1][2] [3][4][5][6][7].

Aoki et al. have proposed a method of sound-source segregation, called SAFIA [1], which is an effective way to suppress an interference sound to obtain a desired sound.

We used a microphone array to apply SAFIA to a speech signal coming from an arbitrary direction. With the array signal processing, the input speech is controlled to provide directivity for each sound source. We can realize precise sound-source segregation by using SAFIA, but there are some problems with SAFIA.

One is spectral distortion. Speech separated by SAFIA, even if a human can hear it clearly, has spectral distortion. Thus, the recognition performance is still not very high. To improve the performance, we tried to adapt the acoustic model by MLLR and training of the acoustic model with separated speech containing the characteristic of SAFIA. In this way, we sought to

absorb the spectral distortion and improve the recognition performance.

Another problem occurs when a window function, such as a Hanning window, is used. When we use fast Fourier transformation (FFT) for spectral analysis, the original spectrum cannot be observed because of the influence of the window function. This may degrade the sound-source segregation performance. Therefore, we used GHA for spectral analysis. We could thus extract the true frequency components without using the window function and analyze the smaller frequency components. In this way, we sought to improve the sound-source segregation.

In this paper, we report the results that we obtained when we attempted double-talk speech recognition using a microphone array and SAFIA. We also examined which signal-processing array, the DCMP adaptive array [8] or the delayed-sum array [9], is most suitable for SAFIA preprocessing.

Furthermore, we examined the improvement in SAFIA performance when GHA was used, and the improvement in recognition performance enabled by adapting the acoustic model through MLLR and training of the acoustic model with separated speech.

2. SAFIA

Aoki et al. have proposed a method of sound-source segregation based on estimating the incident angle of each frequency component of input signals acquired by multiple microphones. This method is called SAFIA. (A block diagram of SAFIA is shown in Fig. 1.) The processing steps of SAFIA are as follows. In the frequency analysis, each input signal, $x_1(n), x_2(n)$, is transformed into frequency components $X_1(f)$ and $X_2(f)$ by FFT.

The inter-channel amplitude difference $\Delta A(f)$ and the inter-channel phase difference $\Delta\phi(f)$ are then calculated.

$$\Delta A(f) = 20 \log_{10} \left(\frac{|X_1(f)|}{|X_2(f)|} \right) \quad (1)$$

$$\Delta\phi(f) = \arg(X_1(f)) - \arg(X_2(f)) \quad (2)$$

In the decision process, we decide which frequency components come from the desired direction. In Fig. 1, where the desired source is closer to microphone Mic.1 than Mic.2, the level of the desired speech contained in $X_1(f)$ will be greater than that in $X_2(f)$. Also, the phase of the desired speech in $X_1(f)$ will be farther advanced than that in $X_2(f)$. Therefore, a frequency component that has a positive $\Delta A(f)$ or $\Delta\phi(f)$ is judged to contain the desired speech. In the same way, a frequency component with a negative $\Delta A(f)$ or $\Delta\phi(f)$ is judged to contain undesired speech.

In the waveform synthesis process, to enhance the desired speech, the frequency component judged to contain the desired

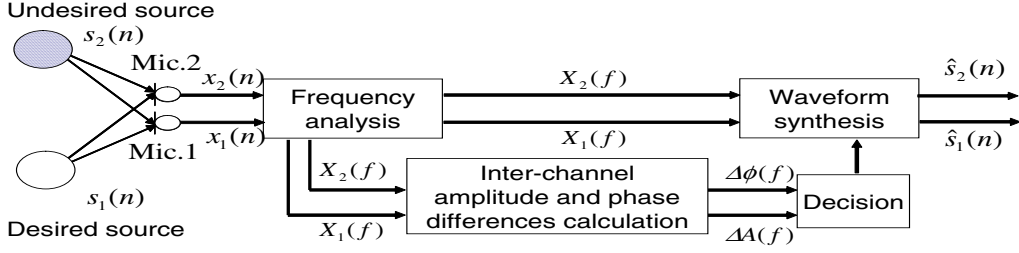


Figure 1: Block diagram of SAFIA.

speech is multiplied by 1 and that judged not to contain it is multiplied by 0. The desired speech is reconstructed by transforming from the frequency domain into the time domain by IFFT. In the same way, the undesired speech is also reconstructed.

3. Proposed method

3.1. Array signal processing and SAFIA

Two directional microphones are usually used in SAFIA. When two sound sources are located in front of them, sound-source segregation goes very well. However, when a sound source is located beyond the range of directivity of a directional microphone, the sound-source segregation deteriorates. We use a microphone array to utilize SAFIA. Based on sound-source positions, the input speech data is selectively emphasized or suppressed by array signal processing such as with a delayed-sum array or a DCMP adaptive array. Then two speech signals having directivity for each sound source are separated by SAFIA.

In this process, a microphone array enables ideal directivity control for applying SAFIA to a speech signal coming from an arbitrary direction. In this way, sound-source segregation is improved.

3.2. GHA-SAFIA

In SAFIA, the sound source is reconstructed by multiplying each frequency component by 1 or 0. This assumes that if the frequency resolution is properly determined the two frequency components will overlap very little in each frequency band. However, under the double-talk condition, the SNR is 0 dB, and this assumption does not hold. Further, when we use FFT for spectral analysis, the spectrum is transformed because of the influence of the window function. The original spectrum cannot be observed. This may cause the deterioration of sound-source segregation.

We try to separate the speech by using GHA for spectral analysis. Because GHA doesn't need a window function in the spectral analysis, we can extract original frequency components. Furthermore, we can analyze smaller frequency components and can segregate a sound source more precisely. The process we use is as follows.

1. Analyze the speech input into a microphone array by GHA
2. From the sound-source positions, each sound source is emphasized by a delayed-sum array.
3. Calculate the power spectrum in the emphasized speech in every frequency extracted by GHA and separate by SAFIA

We call this method GHA-SAFIA.

3.3. Noise adaptation

The speech recognition system performs well for the clean speech recorded by a microphone mounted on a user's body. The recognition performance deteriorates, though, for speech containing spectral distortion.

SAFIA achieves good segregation in a human auditory sense, but the quality is still insufficient for speech recognition. To improve the recognition performance, we use MLLR-based acoustic model adaptation and immunity training with separated speech containing spectral distortion. In this way, we try to absorb the spectral distortion.

4. Experiment

4.1. Conditions

First, we recorded the speech data to enable continuous speech recognition. We used two loudspeakers as sound sources instead of human speakers. The two sound sources (loudspeakers) were separated by an angle of θ degrees ($\theta = 45, 70^\circ$). Source1 (the desired source) was kept stationary while source2 (the undesired source) was moved to vary the experimental conditions. The two loudspeakers were arranged radially with respect to the center microphone array at distances of 100 and 150 cm. The arrangement of the microphone array and two loudspeakers is shown in Fig. 2. Also, the details of the microphone array and the experimental conditions are shown in Table 1.

As speech data, we selected one hundred sentences spoken by twenty male speakers from the ASJ continuous speech corpus [10] and simultaneously played different speech from the two loudspeakers. The SNR was 0 dB.

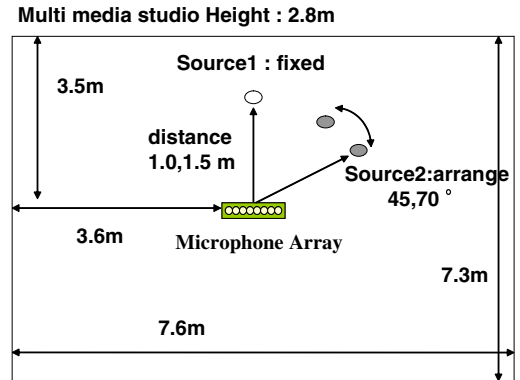


Figure 2: Microphone array and source positions.

Table 1: Microphone array and experimental conditions.

array form	linear and consistent spacing 8 elements spaced 3cm apart
element sampling	non-directional condenser microphone 32 kHz, 16 bit
frame length	1024 samples (32 ms) Hanning window
frame shift	256 samples
voice	two male voices, 100 sentences
voice volume	desired:undesired = 1:1 SNR = 0 dB
voice length	desired:undesired = 1:1
mode vector	65536 point measured with TSP [11] impulse length 1024 samples

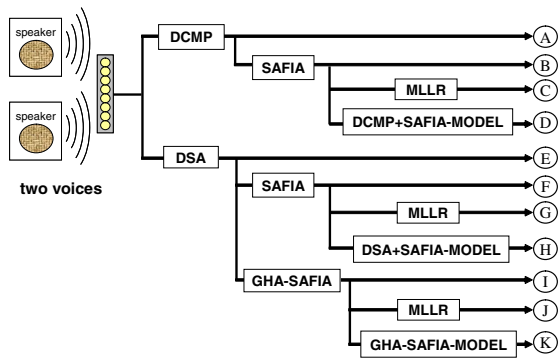


Figure 3: Method used to process the speech data.

4.2. Speech-data processing

The method we used for speech-data processing is shown in Fig. 3. In Fig. 3, DCMP means we processed the speech using a DCMP adaptive array, and DSA means we used a delayed-sum array. MLLR means that we used the acoustic model adapted through MLLR-based adaptation. DCMP+SAFIA-MODEL, DSA+SAFIA-MODEL, and GHA-SAFIA-MODEL mean that we used the acoustic model trained with speech data separated by DCMP+SAFIA, DSA+SAFIA, or GHA-SAFIA, respectively.

As the adaptation data for MLLR, we selected phoneme balance sentences spoken by male speakers from the ASJ continuous speech corpus, excluding the previous twenty speakers. This data was recorded under the same conditions as shown in Fig. 2.

For immunity training of the acoustic model, we used the speech data spoken by about 130 male speakers from the ASJ continuous speech corpus. First, we convoluted the impulse responses in the speech data and input the speech data into a microphone array in a situation where two sound sources existed in a room. Next, we emphasized or suppressed the speech data with the DSA or DCMP and separated it by SAFIA. In this way, we created training data which contained the characteristic of each sound-source segregation method. To learn only the characteristic of each sound-source segregation method, we arranged the speaker positions at random.

As the other method, we used an acoustic model trained with speech data, recorded with a microphone mounted on the user's body, spoken by 100 male speakers from the ASJ continuous speech corpus. The acoustic features and analysis conditions are shown in Table 2.

Table 2: Parameters of acoustic feature.

pre-emphasis	0.97
frame length	25 ms
frame shift	10 ms
acoustic feature	MFCC+ Δ MFCC+ Δ power

Table 3: Word accuracy for each method.

method	100 cm		150cm	
	speaker interval 45 °	70 °	speaker interval 45 °	70 °
A	37.2	39.6	33.9	40.8
B	60.9	63.8	48.4	61.1
C	69.3	72.0	61.9	68.1
D	73.4	76.2	66.8	72.9
E	18.5	28.3	17.6	24.4
F	61.7	65.4	47.9	61.1
G	69.1	70.8	62.3	68.6
H	70.9	74.5	63.2	69.5
I	63.0	65.6	49.5	63.0
J	68.7	72.3	61.1	67.8
K	71.2	74.1	63.7	71.0

5. Results

The word accuracy when a microphone was mounted on a user's body was over 94%. However, when the microphone was located away from a speaker, accuracy fell to about 80%. Furthermore, when there were two sound sources, the word accuracy was close to 0% (Fig. 4).

The speech-recognition results with double-talk are shown in Table 3. First, we look at the results of processing in the DCMP adaptive array. Processing in only a DCMP (A) did not enable adequate performance, and a large improvement in the recognition performance was achieved by also using SAFIA (B). This reduced the error rate by about 33% compared to that with only DCMP. This shows that SAFIA is an effective method of sound-source segregation.

The method using MLLR adaptation (C) and immunity training of the acoustic model (D) reduced the error rate by about 22% and 33% compared with the DCMP+SAFIA processing method. These results show the effectiveness of training the acoustic model with the separated speech to improve recognition performance.

Next, we compare the results of the methods using DCMP (A-D) with those using the delayed-sum array (E-H). With only array signal processing (A and E), the DCMP adaptive array enabled higher recognition than the delayed-sum array. Comparing the processing methods with SAFIA (B and F), we found that the two methods performed almost equally well. In addition, after MLLR adaptation (C and G) and immunity training (D and H), the two methods again showed almost the same recognition performance.

Finally, we compare the results for the method using FFT (F-H) with that using GHA (I-K) for spectral analysis (Fig. 5). Compared with when no processing was added to the acoustic model (F and I), we were able to slightly improve the sound-source segregation by using GHA for spectral analysis. Compared with the case where we added processing to the acoustic model, such as the adaptation by MLLR (G and J) and immunity training (H and K), there was almost no difference in recognition performance between the GHA-SAFIA and SAFIA methods. We thus found that precise analysis of the frequency components by adding processing to the acoustic model is inef-

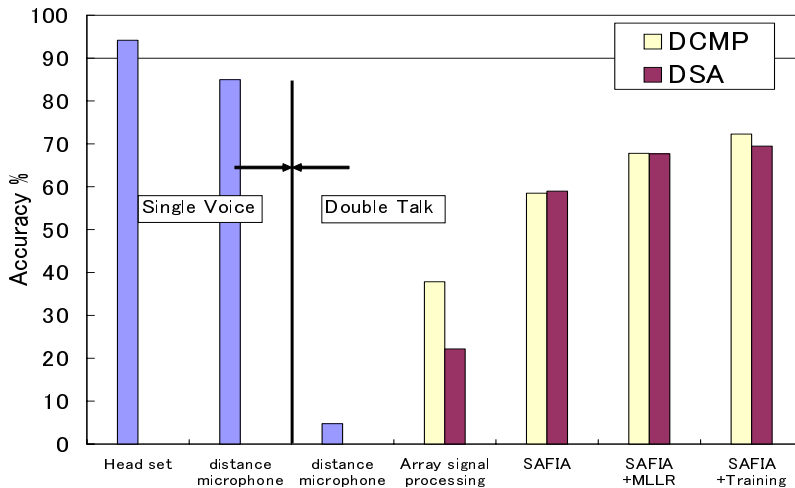


Figure 4: Word accuracy.

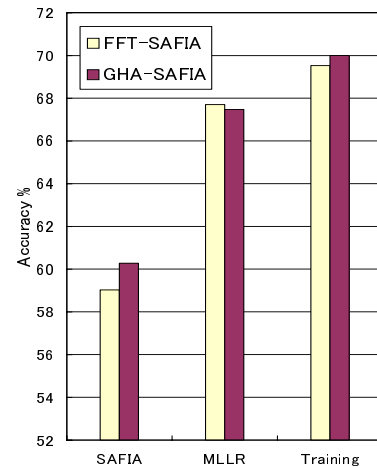


Figure 5: Word accuracy of SAFIA and GHA-SAFIA.

fective.

6. Conclusion

We have evaluated the results of double-talk speech recognition using array signal processing and SAFIA-based audio segregation. We also tried to improve the recognition performance by using MLLR-based acoustic model adaptation and immunity training to be robust to the spectral distortion of SAFIA. We used GHA instead of FFT for spectral analysis to improve the sound-source segregation performance.

Through our experiments, we found that SAFIA is effective for sound-source segregation. Comparing the processing in only a DCMP or a delayed-sum array, we found that DCMP enabled higher recognition. However, when processing with SAFIA, there was no difference between using DCMP or the delayed-sum array for SAFIA preprocessing.

Recognition performance was significantly improved by immunity training of the acoustic model with the speech separated by SAFIA. This method reduced the error rate by about 30% compared to the case SAFIA with no processing added to the acoustic model, and resulted in the best score (76.2%).

We were able to slightly improve the SAFIA performance by using GHA for the spectral analysis. However, after adaptation through MLLR and training of the acoustic model, GHA provided no benefit compared to the use of FFT. The performance of SAFIA and GHA-SAFIA was almost the same.

7. References

- [1] M.Aoki et al., "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones", *J. Acoustic Soc.* vol.22, No.2, 2001.3.
- [2] Hiroshi G. Okuno et al., "A new speech enhancement: speech stream segregation", In *Proceedings of 1996 International Conference on Spoken Language Processing*, pp.2356-2359, ASA.
- [3] S. F. Boll, "Suppression of acoustic noise in speech using

spectral subtraction", *IEEE Trans. ASSP*, ASSP-33, Vol. 27, pp.113-120, 1979.

- [4] E. Weinstein, M. Feder and A. V. Oppenheim, "Multichannel signal separation by decorrelation", *IEEE Trans. on Speech & Audio Processing*, vol.1, no.4, pp.405-413, Oct.1993.
- [5] S. Shamsunder and G. B. Giannakis, "Multichannel blind signal separation and reconstruction", *IEEE Trans. on Speech & Audio Processing*, vol.5, no.6, pp.626-634, Nov.1997.
- [6] T. Gao, 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.
- [7] Anthony J. Bell and Terrence J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution" *Neural Computation*, 7:1129-1159, 1995.
- [8] K. Takao, M. Fujita and T. Nishi, "An adaptive antenna array under directional constraint", *IEEE Trans. Antennas & Propag.* vol.AP-24, No.5, pp.662-669, Sept.1976.
- [9] J. L. Flanagan, J. D. Johnston, R. Zahn, G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", *J. Acoustic. Soc. Am.* 78 (5),pp.1508-1518, 1985.
- [10] K. Itou et al., "The design of the newspaper-based japanese large vocabulary continuous speech recognition corpus", *Proc. ICSLP98*, pp.3261-3264, Nov. 1998.
- [11] Y. Suzuki, F. Asano, H. Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses", *J. Acoustic. Soc. Am.* vol.97 (2), pp.1119-1123, 1995.