Acoust. Sci. & Tech. **22**, 2 (2001)

PAPER

Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones

Mariko Aoki[1], Manabu Okamoto[2], Shigeaki Aoki[3], Hiroyuki Matsui[4], Tetsuma Sakurai[5] and Yutaka Kaneda[6]

[1]*Media Processing Project, NTT Cyber Space Laboratories, 3–9-11 Midorichou. Musasino, 180–8585 Japan*
[2]*Business Communications Headquarters, NTT East Corporation, UrbanNet Otemachi Bldg., 2–2–2 Otenachi, Chiyoda-ku, Tokyo, 100–0004 Japan*
[3]*Media Technology Development Center, NTT Communications Corporation, Tokyp Opera City Tower 21F, 3–20–2 Nishi-Shinjuku, Shinjuku-ku, Tokyo, 163–1421 Japan*
[4]*Solution Business Division, NTT Communications Corporation, Kowa Nishi-shinbashi Bldg. B Tower 14–1, Nishi-shinbashi 2-chome, Minato-ku, Tokyo, 105–0003 Japan*
[5]*Department of Information Science, Faculty of Engineering, Fukui University, 9–1, Bunkyo 3-chome, Fukui, 910–8507 Japan*
[6]*Department of Information and Communication Engineering, Tokyo Denki University, 2–2, Kanda-Nishiki-cho Chiyoda-ku, Tokyo, 101–8457 Japan*

(*Received 5 July 2000, Accepted for publication 27 November 2000*)

Abstract: We have developed a method of segregating desired speech from concurrent sounds received by two microphones. In this method, which we call SAFIA, signals received by two microphones are analyzed by discrete Fourier transformation. For each frequency component, differences in the amplitude and phase between channels are calculated. These differences are used to select frequency components of the signal that come from the desired direction and to reconstruct these components as the desired source signal. To clarify the effect of frequency resolution on the proposed method, we conducted three experiments. First, we analyzed the relationship between frequency resolition and the power spectrum's cumulative distribution. We found that the speech-signal power was concentrated on specific frequency components when the frequency resolution was about 10 Hz. Second, we determined whether a given frequency resolution decreased the overlap between the frequency components of two speech signals. A 10-Hz frequency resolution minimized the overlap. Third, we analyzed the relationship between sound quality and frequency resolution through subjective tests. The best frequency resolution in terms of sound quality corresponded to the frequency resolutions that concentrated the speech signal power on specific frequency components and that minimized the degree of overlap. Finally, we demonstrated that this method improved the signal-to-noise ratio by over 18 dB.

Keywords: Sound source segregation, Phase difference between input signals, Amplitude difference between input signals, Frequency analysis, Discrete Fourier transformation

PACS number: 43.72.Ew

1. INTRODUCTION

The problem of segregating a desired sound from among several concurrent sounds in a sound field has been actively studied, but a satisfactory solution has not yet been obtained. The aim of our research is to develop a signal-processing technique for segregating individual speech signals from received signals consisting of various sounds.

Two main approaches have been used in pursuit of this goal. One uses a single channel input (acquired by a single microphone), and the other uses multi-channel inputs (acquired by multiple microphones).

An example of a single-channel method is to estimate the harmonic structures of speech signals [1, 2]. Much effort has gone into finding a way to detect the pitch pattern of the desired sound. However, tracing the pitch pattern is still difficult, especially when it changes rapidly. Spec-

tral subtraction methods, which enhance the speech signal by subtracting estimated noise from the observed signal, have also been proposed [3–5]. However, these methods are difficult to apply to non-stationary noise.

Methods that use multi-channel inputs use spatial characteristics as additional cues [6–8]. To work, however, these methods require a very precise estimation of the transfer function. Other methods derived from binaural perception have also been proposed [9]. These use two microphones and imitate several auditory processes. The model uses a series of broadband filters as a basilar membrane model, imitates the neuron's excitation pattern, and uses an interaural cross-correlation model to achieve a computational cocktail-party effect. However, which part of this process is the key to the cocktail-party effect is still under investigation.

In this paper, we describe a new method called SAFIA (sound source Segregation based on estimating incident Angle of each Frequency component of Input signals Acquired by multiple microphones). SAFIA segregates objective speech from concurrent sounds by selecting the frequency components judged to be the objective speech. SAFIA is based on sound localization. It uses two channel inputs and is simpler than several previous approaches using multi-channel inputs. In SAFIA, each signal received by two microphones is transformed into the frequency domain by discrete Fourier transformation. For each frequency component, differences between the channels in both amplitude and phase are calculated. These differences are then used to determine which frequency components come from the desired direction and to reconstruct these components as the desired source signal.

SAFIA falls into the same category as methods that enhance the objective speech by weighting the spectrum — for example, conventional harmonic weighting methods or spectral subtraction methods. However, harmonic weighting methods suffer from the drawback of pitch mistracing, which degrades their performance. SAFIA is free from this drawback because pitch tracing is not needed. Moreover, our method uses spatial cues to decide which spectrum-weighting rule to use on a frame-by-frame basis, and does not require a priori knowledge of the power spectrum of noise. Thus, it can be used to reduce even non-stationary noise, unlike the conventional spectral subtraction method.

The procedure of SAFIA is described in Section 2. Then, the effect of the frequency resolution upon SAFIA is analyzed in Section 3. In Section 4, we evaluate the performance of SAFIA in terms of the signal-to-noise ratio.

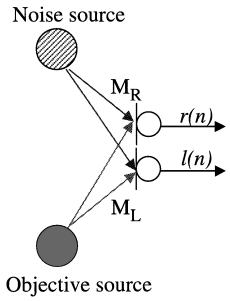

Fig. 1 Arrangement of sound sources and microphones.

2. PROPOSED METHOD (SAFIA)

Consider one objective sound source, one noise source, and two microphones M_{L} and M_{R} set in a field (Fig. 1). To simplify the explanation, we assume that the objective sound source is closer to microphone M_{L} than to microphone M_{R}, and that the noise source is closer to microphone M_{R} than to microphone M_{L}.

A block diagram of SAFIA is shown in Fig. 2. The processing steps of the method are as follows.

In the frequency analysis, each input signal, $r(n)$, $l(n)$, is transformed into frequency components $R(f)$ and $L(f)$ by discrete Fourier transformation. The length of the overlap between analysis frames is half the frame length. Components $R(f)$ and $L(f)$ are expressed by

$$R(f) = |R(f)| \exp(j \arg(R(f))) \tag{1}$$

$$L(f) = |L(f)| \exp(j \arg(L(f))) \tag{2}$$

The inter-channel amplitude difference $\Delta A(f)$ and the inter-channel phase difference $\Delta\phi(f)$ between $R(f)$ and $L(f)$ are then calculated. These are defined as

$$\Delta A(f) = 20 \log_{10} \left(\frac{|L(f)|}{|R(f)|} \right) \tag{3}$$

$$\Delta\phi(f) = \arg(L(f)) - \arg(R(f)) \tag{4}$$

To keep $\Delta\phi(f)$ within $-\pi < \Delta\phi(f) < \pi$, we add or subtract 2π to or from $\Delta\phi(f)$ if necessary.

Both objective speech and noise (including undesired speech) are assumed to have harmonic structures. We then hypothesize that if the frequency resolution is properly determined, these harmonic components hardly overlap. In other words, most of the frequency components of a mixed signal belong to either the objective original speech or the noise. Based on these assumptions, the inter-channel amplitude difference $\Delta A(f)$ for each frequency between $L(f)$ and $R(f)$ is that of either the objective speech or the noise. The inter-channel phase difference $\Delta\phi(f)$ between $L(f)$ and $R(f)$ is that of either the objective speech or the noise. For the arrangement shown

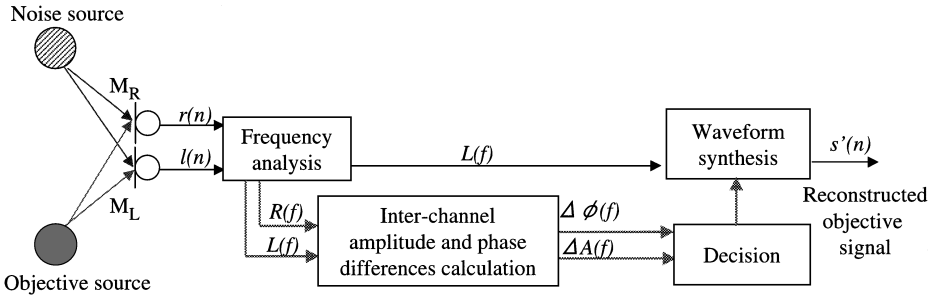

Fig. 2 Block diagram of proposed method, SAFIA.

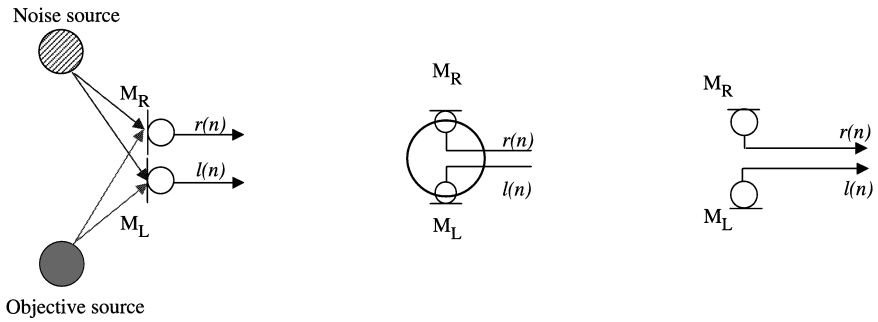

Fig. 3 Examples of sound-acquisition equipment.

in Fig. 1, where the objective source is closer to microphone M_L than M_R, the level of the objective speech contained in $L(f)$ will be greater than that in $R(f)$. The phase of the objective speech in $L(f)$ will be farther advanced than that in $R(f)$. So, in the decision process, a frequency component that has a positive $\Delta A(f)$ or $\Delta\phi(f)$ is judged to contain the objective speech. In the same way, a frequency component with a negative $\Delta A(f)$ or $\Delta\phi(f)$ is judged to contain noise.

In the waveform synthesis process, the objective speech is enhanced by weighting the spectrum of $L(f)$, which contains more of the objective speech than the spectrum of $R(f)$. To enhance the objective speech, the frequency component of $L(f)$ judged not to contain the objective speech is multiplied by 0 or some small number $\alpha(f)$. Also, the component of $L(f)$ judged to contain the objective speech is multiplied by 1. The objective speech is then reconstructed by transforming $L(f)$ from the frequency domain into the time domain by inverse Fourier transformation.

Which criterion ($\Delta A(f)$ or $\Delta\phi(f)$) to use in the judgement for each frequency component depends on the type of sound-acquisition equipment. This is because the degrees of the amplitude and phase differences depend on the sound-acquisition equipment. There are several kinds, such as two omnidirectional microphones, two microphones implemented within a diffraction object (for example, a dummy head or a hard sphere), and two directional microphones (Fig. 3).

In the case of two omnidirectional microphones, inter-channel amplitude differences $\Delta A(f)$ are detected at all frequencies. If the distance between the two microphones is short, however, $\Delta A(f)$ becomes small, and detection errors occur. Therefore, in a low-frequency range, where the phase difference can be determined uniquely and with few errors (except at very low frequencies), the inter-channel phase difference $\Delta\phi(f)$ is used in the decision process. In a high-frequency range, half the spatial wavelength is smaller than the microphone distance, so $\Delta\phi(f)$ cannot be determined uniquely; in this case, the inter-channel amplitude difference $\Delta A(f)$ is used for the judgement.

In the case of two microphones implemented within a diffraction object, the two criteria, $\Delta A(f)$ and $\Delta\phi(f)$, are used for each frequency component as in the previous case. This is because the phase difference is determined

uniquely at low frequencies, and a large amplitude difference can be obtained at high frequencies because of the benefit of diffraction [10, 11].

In the case of two directional microphones, sufficient amplitude difference can be obtained at all frequencies even if the distance between the two microphones is short. However, a precise phase difference is difficult to obtain when the distance between the microphones is short. Thus, in this case, the amplitude difference is used for all frequencies.

A generalized explanation of SAFIA is as follows. It is understood from Fig. 4 that $\Delta A(f)$ and $\Delta\phi(f)$ depend on the azimuth θ between the sound source and the microphones and on the distance r (Fig. 4). We denote these dependencies as functions $\Delta A(f,\theta,r)$ and $\Delta\phi(f,\theta,r)$, respectively. We assume that the ideal amplitude difference of the objective speech $\Delta A(f,\theta,r)$ is known. Then, the frequency component is judged to be that of the objective speech when the absolute value of the difference between $\Delta A(f,\theta,r)$ and the observed $\Delta A(f)$ is smaller than some small number ε_1 as shown in Eq. (5). The frequency component is also judged to be that of the objective speech when Eq. (6) holds for a small value ε_2. Thus, the objective speech can be segregated regardless of the position of its source.

$$|\Delta A(f) - \Delta A(f,\theta,r)| \le \varepsilon_1 \quad (5)$$

$$|\Delta\phi(f) - \Delta\phi(f,\theta,r)| \le \varepsilon_2 \quad (6)$$

To simplify the above explanation, only one noise source was assumed. However, SAFIA also works when there are multiple noise sources with the judgement based on Eqs. (5) and (6).

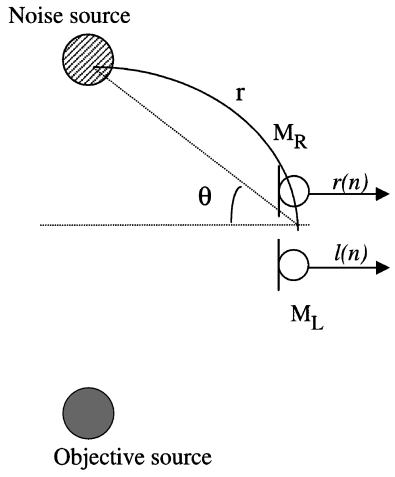

Fig. 4 Dependence of inter-channel amplitude and phase differences upon azimuth θ and distance r.

3. EFFECT OF FREQUENCY RESOLUTION ON SAFIA

Certain frequency components of harmonic signals are more powerful when the frequency resolution is high. As the concentration of power increases, the performance of our method rises. However, to increase the frequency resolution, the analysis window must be made longer. Therefore, for speech signals, whose pitch and harmonics change over time, a substantial increase in the frequency resolution does not necessarily lead to a concentration of power in specific components. Thus, optimum selection of the frequency resolution is very important for methods that segregate speech based on spectrum weighting. However, there have been few studies on how the frequency resolution can be optimized. We studied this through three experiments.

First, we analyzed whether the selected frequency resolution affects the concentration of power in specific frequency components. Second, we analyzed whether the selected frequency resolution reduces the degree of overlap between the frequency components from two speech signals. Third, we investigated the relationship between the frequency resolution and quality of sound segregated by SAFIA.

3.1. Relationship between Frequency Resolution and Concentration in the Power Spectrum

3.1.1. Measure of power concentration

As a measure to evaluate the power concentration, we used the cumulative distribution of the power spectrum. The procedure for obtaining it (Fig. 5) was as follows:

① Speech signal $s(n)$ was transformed into frequency components by discrete Fourier transformation. The power spectrum $|S(f)|^2$ was then calculated.
② Power spectrum $|S(f)|^2$ was sorted in order of size.
③ The sorted power spectrum from ② was accumulated.
④ The cumulative distribution of the power spectrum was calculated by normalizing the accumulated power spectrum from ③. The calculation from ① to ④ was repeated while shifting the time window, and the average curve of the cumulative distribution of the power spectrum was calculated.
⑤ From the average curve calculated above, the percentage of frequency components P that corresponded to an accumulated power of 80% was calculated.

If the frequency component power is concentrated within specific frequency components, the cumulative distribution of the power spectrum increases rapidly and the value of P becomes smaller.

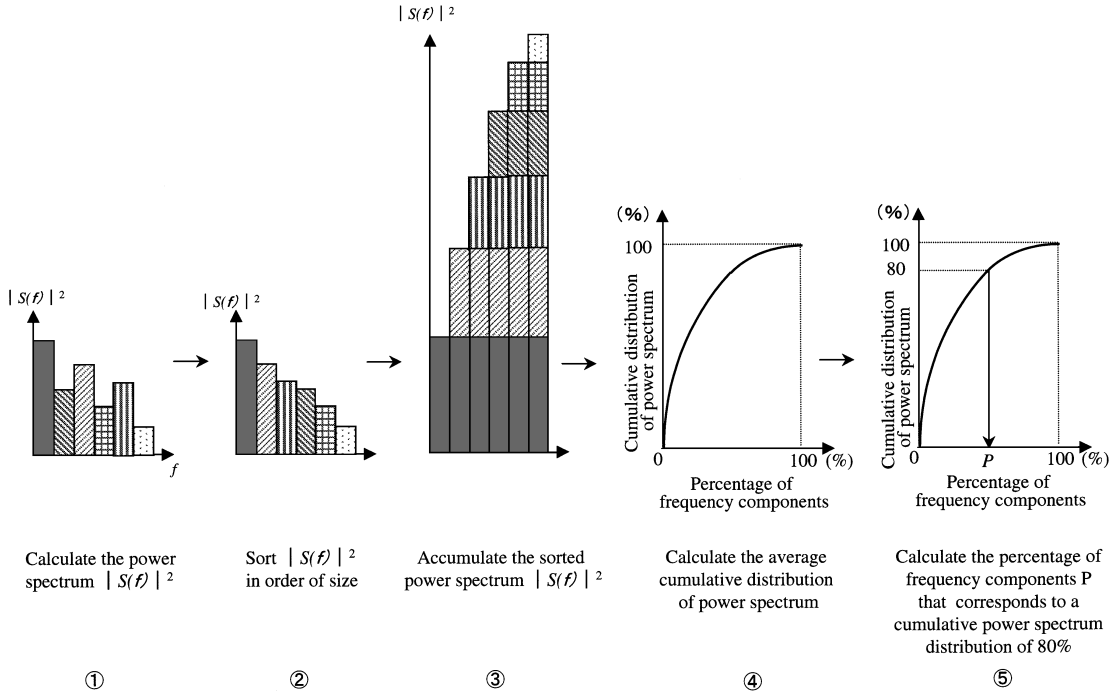

Fig. 5 Procedure for obtaining the percentage of frequency components *P* that corresponds to an 80% cumulative distribution of the power spectrum.

In this test, *P* was calculated for six different frequency resolutions (2.5, 5, 10, 20, 40, and 80 Hz), and the values were compared. We used 60-s-long speech signals from male and female subjects as the test signal. The sampling frequency was 11.025 kHz. A Hanning window was used for discrete Fourier transformation. The six different frequency resolutions mentioned above were achieved by using six different frame lengths, which were 4,096, 2,048, 1,024, 512, 256, and 128 points, respectively. The frequency ranges of the speech signals were from 20 Hz to 5 kHz. Periods of silence were removed from the signals.

3.1.2. Analysis Results

The analysis results are shown in Fig. 6. As the frequency resolution increased from 80 to 10 Hz, the value of *P* decreased; *i.e.*, the frequency component power became more concentrated on specific components. However, *P* increased below 10 Hz. The difference between *P* at 10 and 20 Hz was slight. The female speech was more sensitive to the frequency resolution, but both female and male speech showed similar behavior.

If the signal is harmonic and stationary, then as the frequency resolution improves the frequency component power becomes more concentrated on specific components. Thus, *P* becomes smaller. However, most speech signals are not stationary. The average stationary period of a speech signal is about 40 ms [12], and a time window length of 40 ms corresponds to a frequency resolution of 25 Hz. Therefore, as the frequency resolution rises from 80 to 20 Hz, *P* will become smaller. However, our results showed that *P* was lowest at 10 Hz. This indicates that although the signal included a small number of non-stationary components, the improved frequency resolution concentrated the power on specific frequency components.

When the resolution was 5 or 2.5 Hz, although the frequency resolution was high enough, the time window lengths were too long (they were about 200 and 400 ms, respectively). As a result, each window included many non-stationary components. Each source's frequency components fluctuated more widely in a longer time window. That is, the power of the frequency components was more dispersed among the components.

For the resolutions of 40 and 80 Hz, the window length was suitable for stationary speech, but the poor frequency resolution caused the power to be dispersed among the components.

Thus, the most suitable frequency resolution concentrates the speech-signal power spectrum on specific frequency components. A suitable frequency resolution for segregating objective speech signals by SAFIA is likely to be between 10 and 20 Hz.

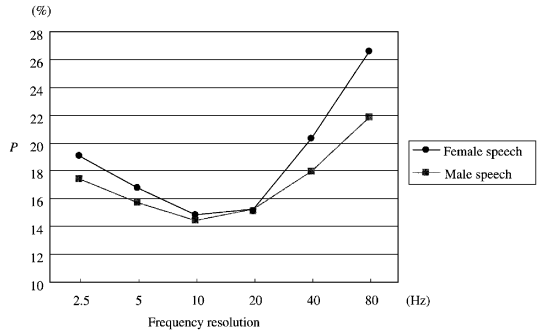

Fig. 6 Dependence of P on frequency resolution, where P is the percentage of frequency components that corresponds to an 80% cumulative distribution of the power spectrum.

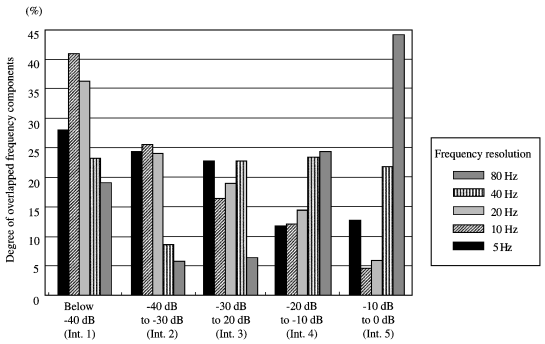

Fig. 7 Relationship between frequency resolution and degree of overlap (female speech with female speech).

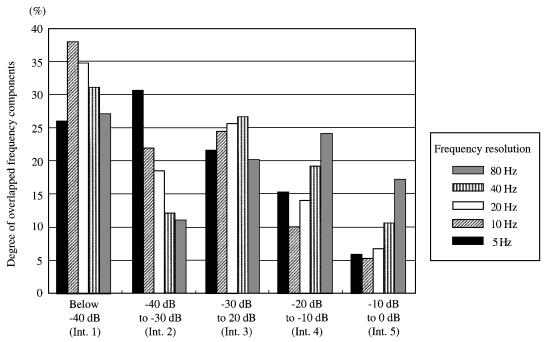

Fig. 8 Relationship between frequency resolution and degree of overlap (male speech with female speech).

3.2. Frequency Component Overlap between Two Speech Signals

3.2.1. Measure of the degree of overlap

We investigated the degree of frequency component overlap between two speech signals as follows.

(1) Two speech signals, denoted by $s_a(n)$ and $s_b(n)$, with equal amplitudes were transformed into frequency components by discrete Fourier transformation. These frequency components were denoted by $S_a(f)$ and $S_b(f)$.

(2) For each frequency component, the ratio of the amplitude of $S_a(f)$ to that of $S_b(f)$ was calculated; this was denoted by $\Delta Lev(f)$.

$$\Delta Lev(f) = -\left|20\log_{10}\left(\frac{|S_a(f)|}{|S_b(f)|}\right)\right| \tag{7}$$

From this definition, $\Delta Lev(f)$ reaches 0 dB when the amplitudes of $S_a(f)$ and $S_b(f)$ are equal. This means that $S_a(f)$ and $S_b(f)$ are highly overlapped for that frequency component.

(3) The value of $\Delta Lev(f)$ varies depending on frequency and time. To obtain the distribution of $\Delta Lev(f)$, we calculated its histogram. The intervals of the histogram were defined as follows.

(Int. 1) $\Delta Lev(f) \leqq -40$ dB
(Int. 2) $-40\,\text{dB} < \Delta Lev(f) \leqq -30$ dB
(Int. 3) $-30\,\text{dB} < \Delta Lev(f) \leqq -20$ dB
(Int. 4) $-20\,\text{dB} < \Delta Lev(f) \leqq -10$ dB
(Int. 5) $-10\,\text{dB} < \Delta Lev(f) \leqq 0$ dB

When the value of $\Delta Lev(f)$ fell into Int. 4 or Int. 5, the estimated degree of overlap for the frequency components between two signals was high. On the other hand, when the value of $\Delta Lev(f)$ fell into Int. 1 or Int. 2, the estimated degree was low. We calculated the histogram of $\Delta Lev(f)$ for five different frequency resolutions: 5, 10, 20, 40, and 80 Hz.

Two pairs of simultaneous speech signals were used as test signals. One pair consisted of male speech and female speech; the other pair consisted of female speech, with the same person speaking different phrases. The signal length was about 60 s.

3.2.2. Analysis results

Figures 7 and 8 show the calculated histograms of the overlap for the two types of speech pairs. (The vertical axis shows the percentages of distribution $\Delta Lev(f)$ for each interval.) Both histograms had some common features as follows.

(1) In Int. 4 and Int. 5, where the degree of overlap was high, the distribution of $\Delta Lev(f)$ increased as the frequency resolution fell from 10 to 80 Hz. On the other hand, in Int. 1 and Int. 2, where the degree of overlap was low, the distribution of $\Delta Lev(f)$ decreased as the frequency resolution became lower. Thus, the degree of frequency component overlap between the two speech signals increased as the frequency resolution became lower.

(2) In Int. 4 and Int. 5, the percentage of $\Delta Lev(f)$ decreased as the frequency resolution rose from 80 to 10 Hz. However, it increased when the frequency resolution was even higher (5 Hz). In Int. 1, on the other hand, it

increased as the frequency resolution rose from 80 to 10 Hz, but decreased when the frequency resolution became 5 Hz.

These results indicate that the optimal frequency resolution that minimizes the degree of frequency component overlap between two speech signals is about 10 Hz. This is consistent with the optimal resolution obtained in Section 3.1 (Fig. 6).

3.3. Relationship between Frequency Resolution and Sound Quality

3.3.1. Evaluation method

To establish the relationship between the sound quality with SAFIA and the frequency resolution, we evaluated the sound quality through a subjective test where the frequency resolution was varied from 5 to 80 Hz. The input signals were mixed to simulate the situation shown in Fig. 9. In this situation, a free-field was assumed. Two omnidirectional microphones, separated by 23 cm, were used. Each sound source was arranged at 45° to the line through the microphones. The distance between the microphones and the sound sources was 60 cm. The mean energies of the two source signals were assumed to be equal. This simulated situation represents two speakers sitting side by side.

Two mixed speech signals (Table 1) were used in the test. One was a mixed signal containing male speech and female speech. The other was a mixed signal containing only female speech (the same person speaking different phrases). The frequency range of the signals was from 20 Hz to 5 kHz. The parameter $\alpha(f)$ was set to zero. The parameters ε_1 and $\Delta A(f,\theta,r)$ were set to 3.37 dB. The parameters ε_2 and $\Delta\phi(f,\theta,r)$ were set to 0.676 ms.

In the test, five subjects listened to six kinds of speech: the original speech and the speech segregated at a frequency resolution of 5, 10, 20, 40, or 80 Hz. The signal length was about 4 s. The subjects used headphones to listen to the six kinds of speech in random order. Then they ranked the speech quality on a five-point scale from 1 (bad) to 5 (excellent). The five subjects were Japanese men in their twenties or thirties.

3.3.2. Results

The results are shown in Fig. 10. In the S1 case, the highest quality was achieved at a frequency resolution of 10 Hz, and there were significant differences ($\alpha \leqq 0.05$) between the frequency resolutions. In the S2, S3, and S4 cases, although the highest-quality frequency resolution was 20 Hz, there were no significant differences between the test results of 10 and 20 Hz. There were significant differences between 20 and 5 Hz, between 20 and 40 Hz, and between 20 and 80 Hz.

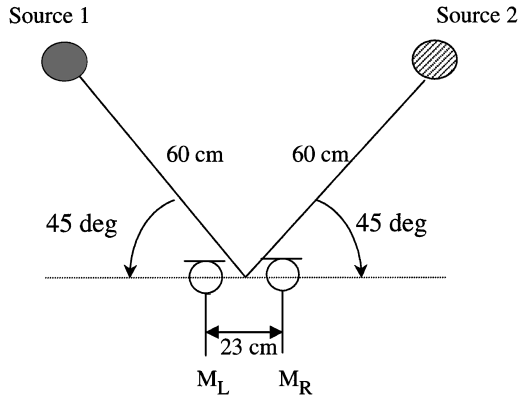

Fig. 9 Arrangement for the subjective test.

Table 1 Segregated signals evaluated subjectively.

Mixed test signal	Segregated signal
Maie speech + female speech	Male speech (S1)
	Female speech (S2)
Female speech + female speech	Female speech (S3)
(same person speaking different phrases)	Female speech (S4)

These results imply that the optimal frequency resolution was between 10 and 20 Hz. This optimal resolution was almost the same as the frequency resolution that gave the highest concentration of speech-signal power on specific frequency components (Section 3.1). Moreover, this optimal resolution was almost the same as the resolution that minimized the degree of overlap (Section 3.2).

4. IMPROVEMENT OF THE SIGNAL-TO-NOISE RATIO

We also evaluated the performance of SAFIA in terms of the signal-to-noise ratio (SNR). This indicated how much of the objective speech was segregated from the mixed speech. The SNR of segregated speech ($s'_a(t)$) is defined by Eq. (8), where $s_a(t)$ represents the original speech. The SNR of mixed speech (speech signal before segregation) ($s_a(t)+s_b(t)$) is defined by Eq. (9), where $s_b(t)$ represents the undesired signal. From these definitions, the improvement in the SNR between before and after segregation can be calculated by Eq. (10).

$$S/N_{\text{aft}} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} s_a(n)^2}{\sum_{n=0}^{N-1}(s'_a(n)-s_a(n))^2}\right) \quad (8)$$

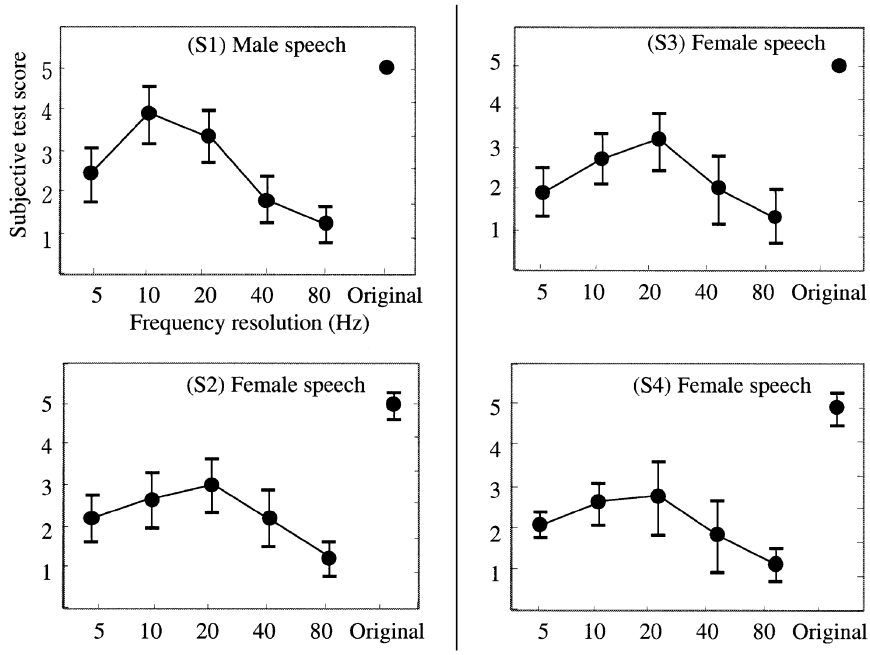

Fig. 10 Relationship between sound quality and frequency resolution.

$$S/N_{\text{bef}} = 10\log_{10}\left(\frac{\sum_{n=0}^{N-1} s_{\text{a}}(n)^2}{\sum_{n=0}^{N-1}((s_{\text{a}}(n)+s_{\text{b}}(n))-s_{\text{a}}(n))^2}\right) \quad (9)$$

$$S/N_{\text{imp}} = S/N_{\text{aft}} - S/N_{\text{bef}} \quad (10)$$

The frequency resolutions were 10 and 20 Hz (which achieved high scores in the subjective test). The speech signals were the same as those used in Section 3.3. As shown in Table 2, SAFIA improved the SNR by more than 18 dB.

5. CONCLUSION

We have developed a method called SAFIA, which segregates desired speech from signals created from concurrent sounds received by multiple microphones. In SAFIA, each signal received by two microphones is transformed into the frequency domain by discrete Fourier transformation. Differences in both the amplitude and phase between channels are calculated for each frequency component. These differences are used to select the frequency components that come from the same direction, and to reconstruct those components as the desired source signal.

Table 2 Improvement in signal-to-noise ratio.

Frequency resolution	Male speech (S1)	Female speech (S2)	Female speech (S3)	Female speech (S4)
20 Hz	20.19 dB	18.01 dB	18.10 dB	18.26 dB
10 Hz	21.66 dB	18.82 dB	20.22 dB	18.35 dB

The choice of which frequency resolution to use in SAFIA is critical. To clarify the effect of the frequency resolution on the effectiveness of our method, we conducted three experiments.

First, we analyzed the relationship between frequency resolution and the cumulative distribution of the power spectrum. We found that the power of speech signals is concentrated on specific frequency components with a frequency resolution of about 10 Hz. In our second experiment, we investigated whether a given frequency resolution decreased the degree of overlap between the frequency components of two speech signals. We found that a frequency resolution of 10 Hz minimized the degree of overlap. In our third experiment, we clarified the relationship between sound quality and frequency resolution through subjective tests. The most efficient frequency resolution for sound quality, in this case about 10 Hz, was close to the resolution that concentrated the power of speech signals on specific frequency components. Moreover, it was close to the frequency resolution that minimized the degree of overlap.

We also demonstrated that SAFIA improved the signal-to-noise ratio by more than 18 dB.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE* **67**, 1586–1604 (1979).

[2] H. Nagabuchi, "Speech enhancement and suppression in mixed speech", *J. IEICE* **J62-A**, 627–634 (in Japanese) (1979).

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust. Speech, Signal Process.* **ASSP-27**, 113–120 (1979).

[4] C. He and G. Zweig, "Adaptive two band spectral subtraction with multi-window spectral estimation", *Proc. ICASSP* 99 Vol. 2, 793–796 (1999).

[5] C. D. Yoo and J. S. Lim. "Speech enhancement based on the generalized dual excitation model with adaptive analysis window", *Proc. ICASSP* 95, 832–835 (1995).

[6] M. Yanagida, O. Kakusho, A. Ueda and Y. Nomura, "Realization of the cocktail party effect by generalized inverse of matrixes", *Proc. Autumn Meet. Acoust. Soc. Jpn.* 2-1-15 (in Japanese) (1980).

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics", *IEEE Trans. Acoust. Speech Signal Process.* **36**, 145–152 (1988).

[8] A. J. Bell and T. J. Sejowski, "An information-maximization approach to blind separation and blind deconvolution", *Neural Comput.* **7**, 1129–1159 (1995).

[9] M. Bodden, "Modeling Human Sound-Source Localization and the Cocktail-party-effect", *Acta Acust.* **1**, 43–55 (1993).

[10] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 3rd Ed. (Academic Press, London, 1989).

[11] J. Blauert, *Spatial Hearing* (MIT Press, Cambridge, Mass., 1996).

[12] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, Electrical Engineering and Electronics* (Marcel Dekker, New York, 1989).

Mariko Aoki received her B. Sci. and M. Sci. degrees in mathematics from Waseda University, Tokyo, Japan in 1993 and 1995, respectively. In 1995, she joined the Cyber Space Laboratory of Nippon Telegraph and Telephone Corporation (NTT), Japan. Her current research interest includes signal processing of sound source segregation. She is a member of ASA and ASJ.

Manabu Okamoto received his M. E. degrees from Kyushu Institute of Design, Japan in 1991. From 1991 to 1999, he has been research engineer at the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT). He is currently an associate manager at NTT EAST Corporation. He is a member of ASJ and IEICE.

Shigeaki Aoki received his Dr. Eng. degrees from Nagoya University, Japan in 1985 with nonlinear interaction between finite amplitude sounds. In 1985 he joined the Electrical Communication Laboratories (ECL) of Nippon Telegraph and Telephone Corporation (NTT). He is currently a manager in the media technology factory of NTT Communications Corporation. Dr. Aoki is a member of the IEEE, ASA, AES, IEICE, JAS, and ASJ.

Hiroyuki Matsui received the B.E. degree from the Kyoto Institute of Technology, Kyoto, JAPAN, in 1978. Since joining NTT in 1978, he has been engaged in research and developments of customer equipments. Currently, he is a senior manager in the Solution Business Division of NTT Communications Corporation. He is a member of the IEICE.

Tetsuma Sakurai received the B.S. and Dr. Eng. degrees in electronic engineering from Nagoya University, Nagoya, Japan, in 1972 and 1985, respectively. In 1972, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan. He is currently a Professor at Faculty of Engineering, Fukui University. He is a member of ASJ, IEICE, JSISE and ISCIE.

Yutaka Kaneda received the B.E., M.E. and Doctor of Engineering degrees from Nagoya University, Nagoya, Japan, in 1975, 1977 and 1990. From 1977 to 2000, he was with Nippon Telegraph and Telephone Corporation (NTT), Musashino, Tokyo, Japan. He is now a professor in acoustics signal processing at the Department of Information and Communication Engineering, Tokyo Denki University, Tokyo, Japan.