Separating Three Simultaneous Speeches with Two Microphones by Integrating Auditory and Visual Processing

Hiroshi G. Okuno[*,†], *Kazuhiro Nakadai*[*], *Tino Lourens*[*], *and Hiroaki Kitano*[*,‡]

[*] Kitano Symbiotic Systems Project, ERATO, Japan Science and Tech. Corp., Tokyo, Japan
[†] Graduate School of Informatics, Kyoto University, Kyoto, Japan
[‡] Sony Computer Science Laboratories, Inc., Tokyo, Japan
okuno@nue.org, {nakadai, tino}@symbio.jst.go.jp, kitano@csl.sony.co.jp

Abstract

This paper addresses the problem of automatic recognition of three simultaneous speeches with two microphones, that is, that of sound source separation where the number of sound sources is greater than that of microphones. The approach used is the *direction-pass filter*, which is implemented by hypothetical reasoning on the interaural phase difference (IPD) and interaural intensity difference (IID). Auditory processing calculates IPD and IID for each subband, and generates hypotheses for precalculated IPD and IID for every direction including one obtained by visual processing. Then the system calculates the belief factor of hypothesis by Dempster-Shafer theory and determines the direction of each subband. Subbands of the specific direction are collected and then converted to a wave form by inverse FFT. With 200 benchmarks of three simultaneous utterances of Japanese words, the average 1-best and 10-best recognition rates of extracted speeches are 60% and 81%, respectively.

1. Introduction

"Listening to several things simultaneously", or computational auditory scene analysis (CASA) may be one of the next goals to automatic speech recognition systems (ASR) which have been widely available recently on personal computers [1, 15, 2]. Since we hear a mixture of sounds under real-world environments, CASA techniques are critical in applying ASR for such applications.

This paper addresses the problem of separation and automatic recognition of three simultaneous speeches with two microphones, partially because it models the recognition of speeches in the presence of speech from interfering talkers, and partially because the number of sound sources is greater than that of microphones. Our approach is to use sound source separation as a hearing aid for ASR; that is, each speech stream is extracted from a mixture of sound and then is recognized by ASR.

According to the theory of beamforming, by using n microphones, $n-1$ dead angles can be formulated [3]. If sound sources are mutually independent, n sound sources can be separated by Independent Component Analysis (ICA) can separate n sound sources by using n microphones [4, 5]. In real-world environments, however, this is often the case that the number of sound sources is greater than that of microphones, and that not all sound sources are mutually independent.

Varga and Moore applied hidden Markov model decomposition to the recognition of two simultaneous speeches with one microphone. One speaker utters the isolated digits, while the other a monosyllabic word [6]. This may be the first experiment of recognition of simultaneous speeches, but the experiment is rather simple and each speech was not separated.

Nakatani *et al.* developed the BiHBSS which separates speech streams from binaural inputs [7]. It firsts extracts harmonic fragments by using a harmonic structure as clue, and then groups them according to the sound source direction and continuity of fundamental frequency. The direction is obtained by calculating the interaural phase difference (IPD) and interaural intensity difference (IID) between the corresponding harmonics of the left and right channels. BiHBSS is applied to recognition of two simultaneous speeches to improve the recognition performance [8].

Although such spatial information improves the accuracy of sound source separation, there remains ambiguities because the direction obtained by BiHBSS carries ambiguity of about $\pm 10^{\circ}$. To overcome this kind of ambiguity in the sound source direction, we exploit the integration of visual and auditory information, since the direction obtained by visual processing is much more accurate [9].

At the same time, there are many research on integration of visual, auditory, and other perceptive information. Most of these studies basically use additional perceptive input in order to provide clue to shift attention of other perceptive input. For example, research of sound-driven gaze are addressing how sound source can be used to control gaze to the object which generates sound [10, 11]. Bimodal speech recognition is exploited by combining visual lipreading with acoustic speech recognition [12]. This system assumes only one speaker and does not separate sound sources from a mixture of sounds.

In this paper, we present the design of a *direction-pass filter* that separates sound signals originating from a specific direction given by visual or auditory processing. The direction-pass filter does not assume either that the number of sound sources is given in advance and fixed during the processing, or that the position of microphones is fixed. This feature is critical for applications under dynamically changing environments.

2. Direction-Pass Filter

The direction-pass filter has two microphones and two cameras embedded in the head of a robot. The block diagram of its processing is shown in Fig. 1. The flow of information in the system is sketched as follows:

1. Input signal is sampled by 12KHz as 16 bit data, and analyzed by 1024-point Discrete Fourier Transformation (DFT). Thus, the resolution of DFT is about 11 Hz.

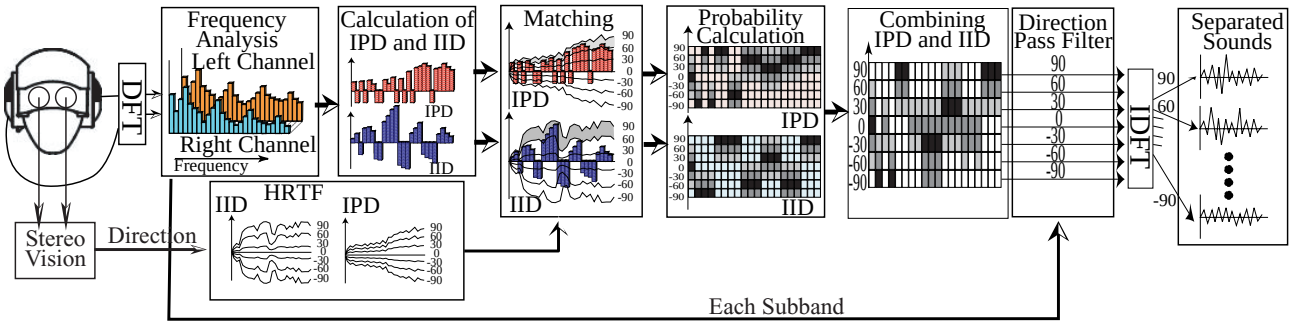

Figure 1: Block diagram of direction-pass filter which extracts sounds originating from the specific direction

2. Left and right channels of each point (subband of 11Hz) are used to calculate the IPD, $\Delta\phi$, and IID, Δp. Please note that the suffix indicating subband is not specified.
3. The hypotheses are generated by matching $\Delta\phi$ and Δp with the reference data of a specific direction or every direction.
4. Satisfying subbands are collected to reconstruct a wave form by Inverse DFT (IDFT).

2.1. Stereo Visual Processing

The visual processing calculates the direction by the common matching in stereo vision based on the corner detection algorithm [13]. It extracts a set of corners and edges, and then constructs a pair of graphs. A graph matching algorithm is used to find corresponding left and right images to obtain the depth, that is, the distance and direction.

From this direction, the corresponding IPD and IID are extracted from the database, which are calculated in advance from the data of the head-related transfer function (HRTF). In this paper, the HRFT is measured at every 10° in the horizontal plane.

2.2. Hypothetical Reasoning on the Direction

The integration system first generates hypotheses IPD and IID, $Phs_h(\theta)$ and $Int_h(\theta)$ of the direction, θ for each subband. The suffix of subband is not specified due to readability. The distance of IPD hypothesis, $Phs_h(\theta)$, and the actual value $\Delta\phi$, is calculated as follows:

$$d_p(\theta) = \left(Phs_h(\theta) - \Delta\phi\right)^2 \tag{1}$$

Similarly, the distance of IID hypothesis, $Int_h(\theta)$ and Δp, is calculated as follows:

$$d_i(\theta) = \left(Int_h(\theta) - \Delta p\right)^2 \tag{2}$$

Then, two belief factors are calculated from the distances using probability density function as shown in Eq. (3), instead of taking the minimum value of $d_P(\theta)$ and $d_I(\theta)$.

$$P_k(\theta) \quad = \quad \int_{-\infty}^{\frac{d_k(\theta)-m}{\sqrt{s/n}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \tag{3}$$

where k indicates p (for IPD) or i (for IID). m and s is the average and variance of $d_k(\theta)$, respectively. n is the number of candidates of direction. In this paper, only each 10° is measured and thus $n = 36$.

Next, a combined belief factor of IID and IPD is defined by using Dempster-Shafer theory as is shown in Eq. (4).

$$\begin{aligned} P_{p+i}(\theta) = \quad & P_p(\theta)P_i(\theta) + (1 - P_p(\theta))P_i(\theta) \\ & + P_p(\theta)(1 - P_i(\theta)) \end{aligned} \tag{4}$$

Finally, θ with the maximum P_{p+i} is selected as the sound source direction. This is the way how to determine the direction of each subband.

2.3. Reconstruction of Singals by Subband Selection

When the direction θ is given, the system determines that the subband originates from θ if $P_{p+i}(\theta)$ is greater than 0.7. The value of this constant is empirically determined. The system collects satsifying subbands and converts them to a wave form by applying Inverse DFT.

Usually, the direction is given by visual processing. In some cases where such information is not available due to occlusion, the direction is determined solely by auditory processing. That's why this complicated way of determining the sound source direction and extracting sounds originating the specific direction is adopted.

3. Experiments

3.1. Benchmark Sounds

The task is to separate simultaneous three sound sources using direction-pass filter defined by the previous section. The benchmark sound set consists of 200 mixture of three concurrent utterances of Japanese words, which is used for the evaluation of sound source separation and recognition. Although a small set of benchmarks were actually recorded in an anechoic room, most mixture of sounds were created analytically by using HRTF. Of course, we confirmed that the synthesized and actually recorded data don't cause a significant difference in speech recognition performance [16].

1. All speakers are located at about 1.5 meters from the pair of microphones installed on a dummy head.
2. The first speaker is a woman located at 30° to the left from the center (-30°).
3. The second speaker is a man located in the center.
4. The third speaker is a woman located at 30° to the right from the center.
5. The order of utterance is from left to right with about 150ms delay. This delay is inserted so that the mixture of sounds was to be recognized without separation.

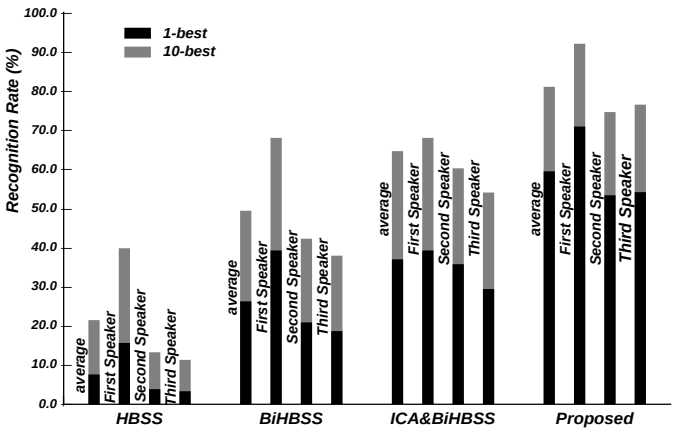

Figure 2: [Experiment 1] Comparison of 1-best/10-best recognition rates by four systems

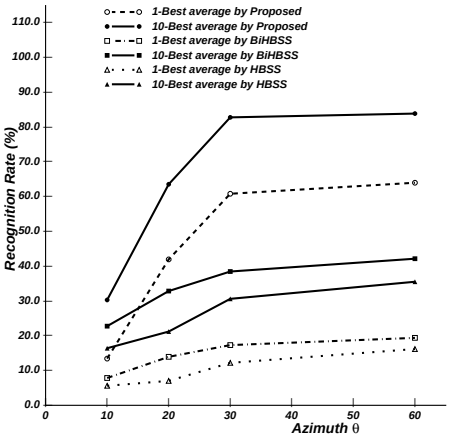

Figure 3: [Experiment 2] Influence of speakers nearness on the average 1-best/10-best recognition rates

Each separated speech stream is recognized by a Hidden Markov Model based automatic speech recognition system [14]. The parameters of HMM were trained by a set of 5,240 words uttered by five speakers. More precisely, each training data is analytically converted to five directions, $\pm 60^\circ$, $\pm 30^\circ$, and 0°, by using HRTF. The training data is disjoint from the utterances included in the above benchmarks.

3.2. Systems to be Compared

The performance of sound source separation is compared among the following four systems.

Proposed Direction-pass filter with the direction given by Vision.

HBSS (Harmonic Based Stream Segregation System) separates sound streams by spectral subtraction and by using harmonic structures as clue [15].

BiHBSS Binaural HBSS disambiguates the crossing of harmonic structures by using the direction of sound sources obtained by IPD and IID [16].

ICA&BiHBSS Independent Component Analysis system called "on-line blind source separation" [5] is combined with BiHBSS. BiHBSS extracts only one sound streams and the remaining signals are given to ICA system, because the remaining signals are expeccted to consist of two sound sources [17].

HBSS takes monaural inputs, while the other three systems take binaural inputs.

3.3. Experiment 1: Recognition of Three Simultaneous Speeches

In Experiment 1, 200 benchmarks are separated by direction-pass Filter with Vision, HBSS, BiHBSS, and ICA&BiHBSS. Then, separated speeches are recognized by automatic speech recognition system. The 1-best and 10-best recognition rates for each speaker are shown in Fig. 2.

The proposed system shows the best performance. The recognition rates for the first speaker are almost the same as those for a single speaker. Those for the third speaker are better than for the second speaker unlike the other three systems. This reason will be investigated by Experiment 3.

The second best system is ICA&BiHBSS. The recognition rates for the first speaker are the same in BiHBSS and ICA&BiHBSS. However, those for the other speakers are much improved, because the remaining signals given to the ICA are distorted due to spectral subtraction in BiHBSS. By comparing the performance of HBSS and BiHBSS, the effect of sound source direction, or monaural vs binaural, is apparent.

3.4. Experiment 2: Robustness against Closer Speakers

In Experiment 2, we investigate the robustness of the three speech stream separation systems, direction-pass filter with vision, BiHBSS, and HBSS, against closer speakers by changing the directions of each speaker. The azimuth between the first and second speakers and that between the second and third speakers are the same, say "θ". We measured the 1-best and 10-best recognition for 10°, 20°, 30°, and 60°.

The result of recognition rates by proposed system, HBSS, and BiHBSS is shown in Fig. 3. Recognition rates saturate around the azimuth of more than 30°. For the azimuth of 10° and 20°, recognition rates for the second (center) speaker are quite poor compared with the other speakers (this data is not shown in Fig. 3).

3.5. Experiment 3: Sensitivity to the Direction

In Experiment 3, the sensitivity to the direction in the direction-pass filter is investigated by specifying different directions. The direction given to the system varies by 10° from 60° to the left to 60° to the right from the center.

The 1-best and 10-best recognition rates of separated sound for every 10° azimuth are shown in Fig. 4. The correct azimuth for this benchmark is 30° to the left (specified by -30° in Fig. 4), 0°, and 30° to the right. For these correct azimuths, recognition rates are reduced significantly. The sensitivity of recognition rates to the accuracy of the sound source depends on how other speakers are close to. That's why the curve of recognition rates for the center speaker is the steepest in Fig. 4.

This experiment proves that if the correct direction of the speaker is available, separated speech is of a high quality at least from the viewpoint of automatic speech recognition. In addition, the recognition rates is quite sensible to the accuracy of the sound source direction if speech is interfered by closer speakers.

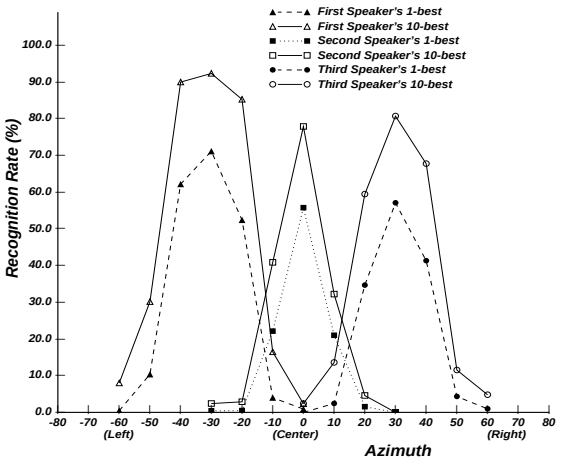

Figure 4: [Experiment 3] 1-best and 10-best recognition rates of direction-pass filter with a given direction

While binaural microphone provides direction information at certain accuracy, it is not enough to separate sound source in realistic situations. There are inherent difficulties in obtaining accurate direction by solely auditory processing. Thus, visual directional information is essential in the proposed system.

4. Conclusion

In this paper, we presented the design of direction-pass filter based on HRTF and reported the recognition performance of three simultaneous speeches. As far as the sound sources are at least at the angle of more than 30°, any number of sound sources up to 12 can be separated by the proposed system. The major contribution of this work is that the effect of visual information in improving sound stream separation was made clear. While many research has been performed on integration of visual and auditory inputs, this may be the first study to clearly demonstrate that information from a sensory input (e.g. vision) affects processing quality of other sensory inputs (e.g. audition).

The remaining work includes real-time processing for direction-pass filter and stereo vision, and real-world applications where HRTF may change drastically. We are currently attacking real-time processing with promising results [18].

4.1. Acknowledgments

We thank Tomohiro Nakatani of NTT Communication Science Laboratories for his help with HBSS and BiHBSS, and Dr. Shiro Ikeda for providing us his on-line blind source separation system. We also thank Dr. Takeshi Kawabata of NTT Cyber Space Laboratories, and Yukiko Nakagawa of Japan Science and Technology Corp. for their valuable discussions.

5. References

[1] M. P. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, **17**(4): 186–190, 1993.

[2] D. Rosenthal and H. G. Okuno, Eds., *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[3] V. K. Madisetti and D. B. Williams, Eds., *The Digital Signal Processing Handbook*, IEEE Press, 1997.

[4] S. Makeig, S. Enghoff, T-P Jung, and T.J. Sejnowski, "A natural basis for efficient brain-actuated control," *IEEE Trans. on Rehabi. Eng.*, vol. 8, 208–211, 2000.

[5] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," in *Proc. of 1998 International Symposium on Nonlinear Theory and its Applications*, 1998, 923–927.

[6] A. P. Varga and R. K. Moore, "Simultaneous recognition of concurrent speech signals using hidden markov model decomposition," in *Proc. of European Conference on Speech Communcation and Technology (EUROSPEECH-91)*. 1991, 1175–1178, ESCA.

[7] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3-4, 209–222, 1999.

[8] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communication*, vol. 27, no. 3-4, 281–298, 1999.

[9] Y. Nakagawa, H. G. Okuno, and H. Kitano, "Using vision to improve sound source separation," in *Proc. of 16th National Conference on Artificial Intelligence (AAAI-99)*. 1999, 768–775, AAAI.

[10] S. Ando, "An autonomous three–dimensional vision sensor with ears," *IEICE Transactions on Information and Systems*, vol. E78–D, no. 12, 1621–1629, 1995.

[11] G. J. Wolff, "Sensory fusion: integrating visual and auditory information for recognizing speech," in *Proc. of IEEE International Conference on Neural Networks*, March 1993, vol. 2, 672–677.

[12] C. Bregler, S. M. Omohundro, and Y. Konig, "A hybrid approach to bimodal speech recognition," in *Proc. of 28th Annual Asilomar Conference on Signals, Systems, and Computers*, 1994, 556–560.

[13] T. Lourens, K. Nakadai, H. G. Okuno, and H. Kitano, "Selective attention by integration of vision and audition," in *Proc. of First IEEE-RAS International Conference on Humanoid Robot (Humanoid-2000)*. 2000, IEEE/RSJ.

[14] K. Kita, T. Kawabata, and K. Shikano, "HMM continuous speech recognition using generalized LR parsing," *Transactions of Information Processing Society of Japan*, vol. 31, no. 3, 472–480, 1990.

[15] T. Nakatani, T. Kawabata, and H. G. Okuno, "A computational model of sound stream segregation with the multi-agent paradigm," in *Proc. of 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP-95)*. 1995, vol. 4, 2671–2674, IEEE.

[16] T. Nakatani, M. Goto, and H. G. Okuno, "Localization by harmonic structure and its application to harmonic sound stream segregation.," in *Proc. of 1996 International Conference on Acoustics, Speech and Signal Processing (ICASSP-96)*. 1996, vol. II, 653–656, IEEE.

[17] H.G. Okuno, S. Ikeda, and T. Nakatani, "Combining independent component analysis and sound stream segregation," in *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*. 1999, 92–98, IJCAI.

[18] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for robots," in *Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI-01). to appear*, 2001, IJCAI.