Effects of increasing modalities in understanding three simultaneous speeches with two microphones

Hiroshi G. Okuno[*,†], *Kazuhiro Nakadai*[*], *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@symbio.jst.go.jp, kitano@csl.sony.co.jp

Abstract

This paper reports effects of increasing modalities in *understanding three simultaneous speeches* with two microphones. This problem is difficult because the beamforming technique adopted for a microphone array needs at least four microphones, and because independent component analysis adopted for blind source separation needs at least three microphones. We investigate four cases; monaural (one microphone), binaural (two microphones), binaural with independent component analysis (ICA), and binaural with vision (two microphones and two cameras). The performance of word recognition of three simultaneous speeches is improved by adding more modalities, that is monaural, binaural, and binaural with vision.

1. Introduction

"Listening to several things simultaneously", or computational auditory scene analysis (CASA) may be one of the important capabilities for next generation automatic speech recognition systems (ASR) [1, 2, 3]. 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. According to the theory of beamforming, by using n microphones, $n-1$ dead angles can be formulated [4]. If sound sources are mutually independent, n sound sources can be separated by Independent Component Analysis (ICA) with using n microphones [5, 6]. 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.

We investigate the effects of increasing modalities in sound source separation, that is, monaural (one microphone), binaural (two microphones), binaural with ICA, and binaural with vision (two microphones and two cameras). The key aspect in this investigation is how to extract the sound source direction and how to exploit it. The extraction of the direction is based on the difference of some features between two channels, such as interaural phase difference (IPD), and interaural intensity difference (IID).

For monaural input, we use HBSS, which uses harmonic structures as a clue of sound source separation [3]. It first extracts harmonic fragments by using a harmonic structure as clue, and then groups them according to the proximity and continuity of fundamental frequency. This works well when fundamental frequencies do not cross. Suppose that one fundamental frequency is increasing while the other decreasing and they are crossing. With a single microphone, it is difficult to discriminate the case that they are really crossing from one that they are approaching and then departing.

To solve this ambiguity, two microphone, that is, a dummy head microphone is used. For binaural input, we use BiHBSS, which uses directional information as an additional clue [7]. It first extracts a pair of harmonic structure for left and right channels and then calculates the IPD and IID to obtain the sound source direction. Finally it groups them according to the proximity and continuity of the sound source direction and fundamental frequency. This calculation of the sound source direction is based on the fundamental frequency, and thus is called *feature-based matching* borrowed from stereo vision. 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].

Therefore, 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. The idea of obtaining the sound source direction is probabilistic reasoning in terms of the set of IPD and IID obtained from each subband. This calculation is based on the *area-based matching* borrowed from stereo vision.

For binaural input, we also developed the system that integrates BiHBSS and ICA (Independent Component Analysis). BiHBSS generates one harmonic structure and the residue. Since the residue is considered to consist of the remaining harmonic structures and non-harmonic parts, it is separated by ICA. This is based on our observation that ICA works better than BiHBSS for a mixture of two speeches.

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

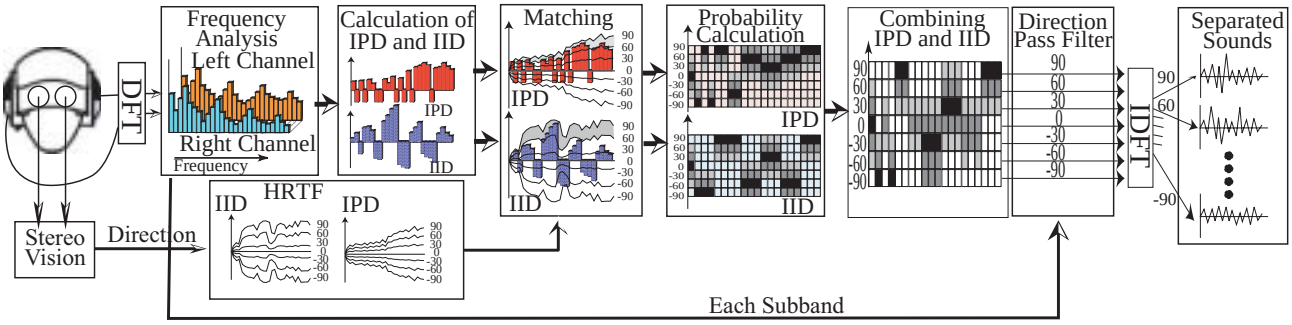

Figure 1: Block diagram of direction-pass filter which extracts sounds originating from the specific direction

analyzed by 1024-point Discrete Fourier Transformation (DFT). Thus, the resolution of DFT is about 11 Hz.

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 [10]. 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 Sigals 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 satisfying 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. BiHBSS&ICA

BiHBSS and blind source separation are integrated to exploit each merits and overcome each weak points. The idea is very simple. Since blind source separation can separate better than BiHBSS for a mixture of two sounds, BiHBSS generates a mixture of sounds. The whole system is depicted in Fig 2. We use the "on-line algorithm" for blind source separation developed by Murata and Ikeda [6], and have confirmed that it separates each speech from a mixture of two speeches with successful results.

3.1. On-line algorithm for Blind Source Separation

Blind source separation by ICA is sketched roughly. Let source signals consisting of n components (sound sources) be denoted by the vector (1), and observed signals by n sensors (microphones) be denoted by the vector (2) specified as below:

$$\boldsymbol{s}(t) = (s_1(t), \cdots, s_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{5}$$

$$\boldsymbol{x}(t) = (x_1(t), \cdots, x_n(t))^T, \quad t = 0, 1, 2, \ldots \tag{6}$$

Each component of $\boldsymbol{s}(t)$ is assumed to be independent of each other, that is, the joint density function of the signals is factor-

ized by their marginal density function

$$p(s_1(t), \cdots, s_n(t)) = p(s_1(t)) \times \cdots \times p(s_n(t)).$$

In addition, observations are assumed to be linear mixtures of source signals:

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t)$$

Note that A is an unknown linear operator.

Let $a_{aj}(\tau)$ be a unit impulse response from source j to sensor i with time delay τ. The observation at sensor i can be represented as

$$\boldsymbol{x}(t) = \left(\sum_k a_{ik} * s_k(t) \right),$$

$$\text{where, } a_{ik} * s_k(t) = \sum_{r=0}^{\tau_{max}} a_{ik}(\tau) * s_k(t - \tau)$$

Thus, A can be represented in matrix form as $[a_{ij}(t)]$.

The goal of ICA is to find a linear operator $B(t)$, such that the components of reconstructed signals

$$\boldsymbol{y}(t) = B * \boldsymbol{x}(t)$$

are mutually independent, *without* knowing the operator $A(t)$ and the probability distribution of source signal $\boldsymbol{s}(t)$.

Ideally we expect $B(t)$ to be the inverse operator of $A(t)$, but there remains indefiniteness of scaling factors and permutation due to lack of information on the amplitude and the order of the source signals.

"*On-line ICA*" algorithm [6] separates source signals from a mixture of signals in the following steps:

1. First, mixed signals are converted to the spectrogram by the windowed DFT with the Hamming window of 128 points.
2. Then, on-line ICA (Independent Component Analysis) is applied to the frequency components of the non-symmetric 65 points.
3. Next, the correspondence of separated components in each frequency is determined based on temporal structure of signals.

   Since the output of ICA carries ambiguities in permutation of the frequent components and in the amplitudes, the permutation of components is determined on the basis of correlation between their envelops.
4. Finally, separated spectrogram of the source signals is constructed.

3.2. Integration of BiHBSS and ICA

The flow of processing is roughly sketched as below:

1. When BiHBSS gets input signals, its Harmonic Fragment Extractor extracts harmonic fragments, which Harmonic Grouping Agent groups to harmonic groups.
2. The Coordinator agent always watches the processing of Harmonic Grouping Agent and bookkeeping information on harmonic groups. When Harmonic Grouping Agent finishes all processing, Coordinator generates a mixture of two speeches, and gives it to ICA. This mixture usually consists of the latestly separated two speeches. Independent Component Extractor extracts independent components and Permutation/Grouping Agent calculates a correct combination of independent components and reconstructs speeches. In this stage, information on independent components is fed back to Coordinator, which bookkeeps the information.
3. Finally, speeches separated by BiHBSS and ICA are given to Conflict Resolver, which checks whether speech separated by ICA has a corresponding speech separated by BiHBSS. If found, BiHBSS's output is adopted. Otherwise, Conflict Resolver calls Harmonic Grouping Agent to do regrouping according to ICA.

   Since a mixture of two speeches Coordinator gives to ICA may contains errors, the system gives precedence to BiHBSS over ICA.

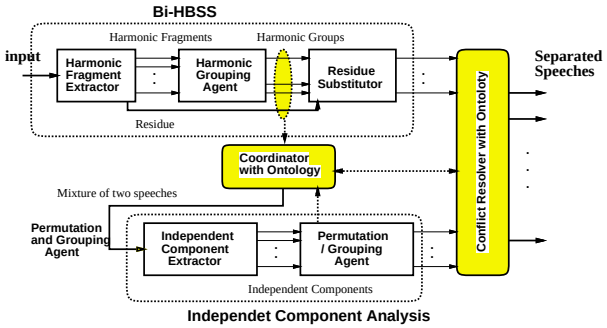

Figure 2: Integrated Systems with BiHBSS and ICA

4. Experiments

4.1. Benchmark Sounds

The task is to separate simultaneous three sound sources by HBSS, BiHBSS, BiHBSS&ICA, and direction-pass filter. 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.

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.

Each separated speech stream is recognized by a Hidden Markov Model based automatic speech recognition system [11]. 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, ±60°, ±30°, and 0°, by using HRTF. The training data is disjoint from the utterances included in the above benchmarks.

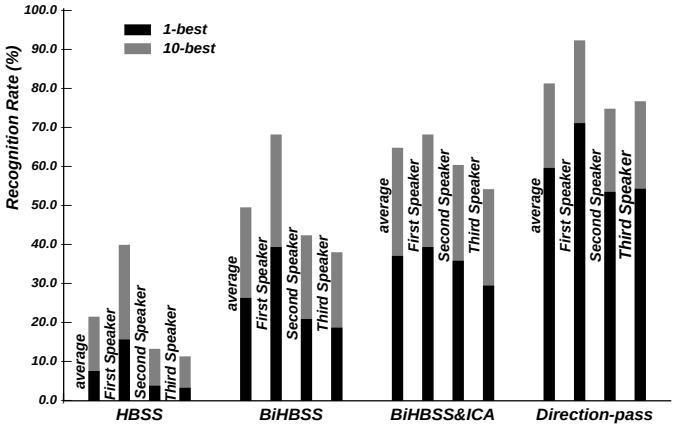

Figure 3: Comparison of four systems in 1-best/10-best recognition rates

4.2. Recognition of Three Simultaneous Speeches

200 benchmarks are separated by direction-pass Filter with Vision, HBSS, BiHBSS, and BiHBSS&ICA. 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. 3.

The direction-pass filter 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.

The second best system is BiHBSS&ICA. The recognition rates for the first speaker are the same in BiHBSS and iHBSS&ICA. 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.

5. Conclusion

In this paper, we report that increasing modalities, from monaural, binaural, to binaural with vision, improve the performance of understanding three simultaneous speeches by using two microphones. The major contribution of this work is that the effect of visual information in improving sound stream separation is 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 searching for other modalities to improve sound source separation. ICA with directional information may be one of promising candidates. Higher lever information should be also exploited; speaker identification, speaker adaptation, and domain-dependent language model for automatic speech recognition. Real-time processing is essential to apply real-world problems. We are currently attacking real-time processing and hierarchical integration of audition and vision, and have already obtained some promising results [12].

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.

6. References

[1] M. P. Cooke, G. J. Brown, M. Crawford, and P. Green, "Computational auditory scene analysis: Listening to several things at once," *Endeavour*, vol. 17, no. 4, pp. 186–190, 1993.

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

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

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

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

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

[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, pp. 209–222, 1999.

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

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

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

[11] 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, pp. 472–480, 1990.

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