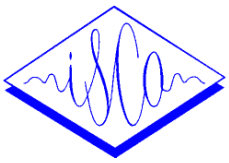


AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4–7, 2003

Improvement of Three Simultaneous Speech Recognition By Using AV Integration and Scattering Theory for Humanoid

Kazuhiro Nakadai† *, Daisuke Matsuura*‡*, Hiroshi G. Okuno**, *and Hiroshi Tsujino*†

†Honda Research Institute Japan Co., Ltd.
‡ Graduate School of Science and Engineering, Tokyo Institute of Technology,
* Graduate School of Infomatics, Kyoto University
nakadai@nakadai.com, matsuurd@mep.titech.ac.jp, okuno@nue.org, tsujino@jp.honda-ri.com

Abstract

This paper presents improvement of recognition of three simultaneous speeches for a humanoid robot with a pair of microphones. In such situations, sound separation and automatic speech recognition (ASR) of the separated speech are difficult, because the number of simultaneous talkers exceeds that of its microphones, the signal-to-noise ratio is quite low (around -3 dB) and noise is not stable due to interfering voices. To improve recognition of three simultaneous speeches, two key ideas are introduced — acoustical modeling of robot head by scattering theory and two-layered audio-visual integration in both name and location, that is, speech and face recognition, and speech and face localization. Sound sources are separated in real-time by an active direction-pass filter (ADPF), which extracts sounds from a specified direction by using interaural phase/intensity difference estimated by scattering theory. Since features of sounds separated by ADPF vary according to the sound direction, multiple Direction- and Speaker-dependent (DS-dependent) acoustic models are used. The system integrates ASR results by using the sound direction and speaker information by face recognition as well as confidence measure of ASR results to select the best one. The resulting system shows around 10% improvement on average against recognition of three simultaneous speeches, where three talkers were located 1 meter from the humanoid and apart from each other by 0 to 90 degrees at 10-degree intervals.

1. Introduction

A robot operating in a normal environment should detect various sound events and pay attention to them to attain robust recognition and to interact with people. So, sound is a very important source of information for both robot and human. To realize such robot audition system in a real world environment, the robot should have the capability to localize speeches, separate one from their mixture and recognize it, because a general sound that a robot hears consists of not a single sound source but multiple ones originating from human, instruments and electronic devices.

Research on *Computational Auditory Scene Analysis* (CASA) focuses on the computer modeling and implementation for the understanding of acoustic events [1]. Approaches for sound source separation by using auditory cues [10], microphone arrays with beam-forming techniques [2], and independent component analysis or blind source separation [3] have been reported. The first approach, however, has been studied only under simulated and off-line environments. Other two approaches based on signal processing and information theory have the theoretical limitation: the number of sound sources should be less or equal to the number of microphones.

In robotics, sound source separation has not studied so much due to difficulties in real world sound source separation. Most robots have a microphone attached near the mouth of each speaker to eliminate sounds other than the signal. Otherwise, they adopt the "stop-hear-act" principle; that is, a robot stops to hear [4]. Some robots with sound source separation [5] require a lot of measurement in advance, and have difficulty in separation during motion, while human can listen to a specific sound during motion. Therefore, robot audition is not enough to be deployed in the real world yet.

One of the most significant key idea to realize such robot audition is "active audition" [6] that improves auditory scene analysis by integration with audition and active motion. We have reported a real-time human tracking system based on the active audition, which localizes and tracks sound sources while in motion [6]. In addition, we attained sound source separation and recognition of simultaneous speeches [7] by using our robot [1]. The system is, however, not always robust against real-world processing as follows:

1. An acoustic model of a robot head that the system uses for both sound localization and separation is getting inaccurate as sound source is away from the front direction of the head or its frequency is higher.
2. The speech recognition process assumes only three directions of speech, that is, 0° and $\pm 60^\circ$.

In this paper, we introduce new acoustical modeling of the robot head by scattering theory to solve the first problem. To relax the second problem, we also introduce two-layered audio-visual (AV) integration in location and name.

The scattering theory[21] can model acoustics of the robot head such as *Interaural Phase Difference* (IPD) and *Interaural Intensity Difference* (IID) mathematically. This enables continuous estimation of IPD and IID against directional changes. The continuous estimation is essential to track sound sources in the real world. Although we have reported continuous estimation of IPD [6], that of IID requires much computational power in case of using conventional methods such as finite/boundary

This work was mainly done with Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp. when the first author had been working there. We thank Dr. Hiroaki Kitano, director of Kitano Symbiotic Systems Project for supporting this work.

[1] Our project and publications on humanoid *SIG* are described in http://winnie.kuis.kyoto-u.ac.jp/SIG/

element methods, that is, it is difficult to apply to real-time systems. The proposed method for estimation of IPD and IID based on the scattering theory is fast enough to work in real-time.

From the viewpoint of speech recognition, sound source separation is effective to enhance a speech with low signal-to-noise ratio such as simultaneous ones. We use speech separation by an *Active Direction-Pass Filter* (ADPF)[7] that separates sound source originating from a specified direction in real-time. Since the performance of ADPF depends on the accuracy of speech localization due to using directional information, AV integration in location is used to obtain robust speech direction.

Generally, AV integration in speech recognition uses visual speech, that is, lip-reading [11, 12]. In a robot, however, lip-reading is not always available because, when a person is away from the robot, resolution of images from robot's camera is insufficient for detecting the lips. The face is generally detected easier than the lips due to its size. Therefore, face recognition is more convenient for robots than lip-reading. We introduce new AV integration in name level, that is, integration of speech and face recognition. To realize this integration, we introduce multiple *Direction- and Speaker-dependent* (DS-dependent) acoustic models, because features of sounds separated by the ADPF vary according to the sound direction. Then, multiple recognition results obtained by DS-dependent acoustic models and speaker name obtained by face recognition are integrated in name level. ROVER[13] is such an integration method which can integrate multiple *Automatic Speech Recognition* (ASR) systems by a weighted voting method and integration of ASR systems based on confidence measure. When more speech directions are assumed to relax the second problem, the number of DS-dependent acoustic models is getting bigger. On the use of a large number of acoustic models, the integration of simple voting or majority rule often fails because a lot of misclassified results affect the system badly. We introduce new integration of recognition results based on word recognition rate of each DS-dependent acoustic model. Thus, two-layered AV integration is introduced to improve recognition of three simultaneous speeches.

The paper is organized as follows: Section 2 describes acoustic modeling of robot head. Section 3 proposes our system for simultaneous speech recognition by integration of active audition and face recognition. Section 4 evaluates the system. Last section gives conclusion.

2. Acoustic Modeling of Robot Head

The most widely used model for acoustic modeling of a head is a *Head Related Transfer Function* (HRTF). The HRTF is different for each person or embedded system due to the differences in the shape of the head. The HRTF also changes as environment changes. In some cases, the HRTF's change is quite drastic and HRTF based mobile systems may fail in sound localization. The HRTF has some difficulties in real-world processing. HRTF, which is often used for sound source localization in binaural research, is obtained by measurement of a lot of impulse responses. Because HRTF is usually measured in an anechoic room, sound source localization in an ordinary echoic room needs HRTF including room acoustic, that is, the measurement has to be repeated if the system is installed at a different room. However, deployment to the real world means that the acoustic features of the environment are not known in advance. It is infeasible for any practical system to require such extensive measurement of the operating space. Thus, robot audition system without or at least less dependent on HRTF is essential for practical systems. Therefore, sound source localization utilizing a different technology is essential in a real-world environment.

MIT's humanoid Cog has a pair of omni-directional microphones embedded in simplified pinnae [14, 15]. In Cog, auditory localization is trained by visual information. This approach does not use a HRTF, but assumes a single sound source. Humanoids of Waseda University can localize a sound source by using two microphones [16, 17]. These humanoids assume a single sound source but localize a sound source by calculating IID or IPD obtained by HRTF. Sound source localization based on IPD and IID is expected to be effective because it is based on the human sound localization model also known as the Jeffress's cross-correlator [18, 19].

We proposed *auditory epipolar geometry* which can extract directional information of sound sources by using two microphones without using HRTF [6]. The auditory epipolar geometry is defined by applying an idea of epipolar geometry in vision[20] to audition. Since the auditory epipolar geometry extracts directional information geometrically, it can dispense with HRTF. However, the auditory epipolar geometry is able to discriminate only three directions – left, right or center – for IID, while it gives continuous estimation of IPD against a direction of a sound source. Auditory processing at higher frequency has much ambiguity in comparison with lower frequency because IPD and IID are effective in lower and higher frequency, respectively[19]. In addition, it does not consider scattered sounds along the back of the head. When a sound source is located at side direction, the scattered sound along the back of the head strongly affects the sound captured by the microphone on the opposite of the sound source. This means that the auditory epipolar geometry is inaccurate for sound from the periphery.

Then, *Scattering Theory*[21] is introduced instead of the auditory epipolar geometry. The scattering theory is concerned with the effect that obstacles or inhomogeneities have on incident waves. It provides a method to estimate the scattered field from the knowledge of the incident field and the scattering obstacle, which gives an accurate estimation of IPD and IID without acoustic measurement by calculating the difference between the total fields at the left and right ears.

2.1. Modeling of Humanoid Head by Scattering Theory

Humanoid head is assumed to be a sphere. The spherical polar coordinates (r,θ,φ) are related to the Cartesian coordinates (x,y,z) by

$$\begin{aligned} x &= r\sin\theta\cos\varphi, \\ y &= r\sin\theta\sin\varphi, \\ z &= r\cos\theta, \end{aligned} \tag{1}$$

where $0 \le r < \infty$, $0 \le \theta \le \pi$, $0 \le \varphi < 2\pi$. The radius of the head is a, that is $r = a$.

A sound source at $\boldsymbol{r_0} = (r_0, 0, 0)$ is defined by

$$V^i = \frac{v}{2\pi R f} e^{i\frac{2\pi R f}{v}}, \tag{2}$$

where f and v are the frequency and the velocity of sound, and R is a distance between the source $\boldsymbol{r_0}$ and an observation point $\boldsymbol{r}$ [22].

On the surface $r = a$, the total field of incident and scattered velocity potential is defined by

$$S(\theta, f) = V^i + V^s = -\left(\frac{v}{2\pi a f}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}(\frac{2\pi r_0}{v} f)}{h_n^{(1)'}(\frac{2\pi a}{v} f)}, \quad (3)$$

where P_n and $h_n^{(1)}$ are the first kind Legendre function and the first spherical Hankel function, respectively [22].

The left and right microphones locate at $M_l = (a, \frac{\pi}{2}, 0)$ and $M_r = (a, -\frac{\pi}{2}, 0)$, respectively. When a sound source is located at $r_0 = (r_0, \theta, 0)$, The total velocity potential S_l and S_r in the left and right microphones are defined by

$$S_l(\theta, f) = S(f, \frac{\pi}{2} - \theta), \quad (4)$$

$$S_r(\theta, f) = S(f, -\frac{\pi}{2} - \theta). \quad (5)$$

Thus, the IPD $\Delta\varphi_s$ and IID $\Delta\rho_s$ are calculated by

$$\Delta\varphi_s(\theta, f) = \arg\left(S_l(\theta, f)\right) - \arg\left(S_r(\theta, f)\right), \quad (6)$$

$$\Delta\rho_s(\theta, f) = 20 \log_{10} \frac{|S_l(\theta, f)|}{|S_r(\theta, f)|}. \quad (7)$$

2.2. Auditory Epipolar Geometry

When the influence by a head shape (sphere) is considered, the auditory epipolar geometry is defined by

$$\Delta\varphi_e = \frac{2\pi f}{v} \times r\left(\theta + \sin\theta\right) \quad (8)$$

where f, v, r and θ are the frequency of sound, the velocity of sound, radius of a humanoid's head and the sound direction, respectively. $\Delta\varphi$ is an estimated IPD corresponding to θ. When the distance between a sound source and our humanoid is more than 50 cm, the influence of the distance can be ignored [7]. Because Eq. (8) assumes such sound source distance, it is defined as a function independent from the distance.

It is easy to estimate IPD, but it does not have any mathematical model on IID — only three directions such as the left, the center and the right of the robot can be judged.

3. Simultaneous Speech Recognition by Two-Layered AV Integration

The robot audition system for simultaneous speech recognition by AV integration consists of two stages, that is, "localization and separation stage" and "recognition stage" shown in Fig. 1. To realize robust system, the first and second stages have AV integration in location and name, respectively. The following sections describe these stages by focusing on AV integration.

3.1. Localization and Separation of Speech and Face

An upper torso humanoid is used for a testbed of this work. The robot has a pair of CCD cameras (Sony EVI-G20) for face localization and recognition, and a pair of microphones for speech localization, separation and recognition. It is driven by 4 DC motors (4 DOFs) with functions of position and velocity control by using potentio-meters. Speeches and images captured by robot's microphones and a camera are sent to speech and face localization modules, respectively.

3.1.1. Speech and Face Localization

In speech localization, IPD and IID of the input speech are calculated , and hypotheses on IPD and IID are created by using

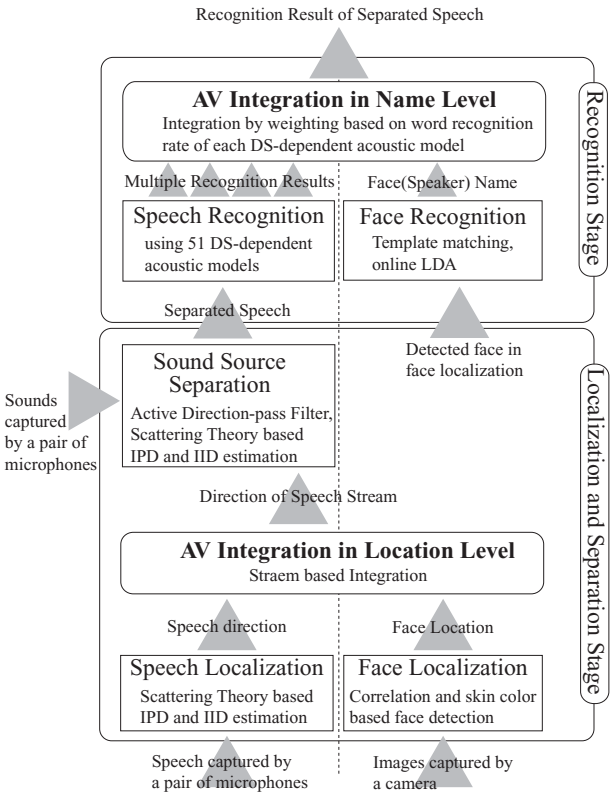


Figure 1: Speech Recognition by two-layered AV Integration

Eqs. (6) and (7) in the range of $\pm 90^\circ$ at 5° intervals. Belief factors of IPD and IID by using a probability density function from Euclidean distances between the input and the hypothesis. They are integrated using Dempster-Shafer theory [9]. Finally, θ for the maximum integrated belief factor is treated as the direction of the input speech [6].

In face localization, face direction is extracted in real-time by using skin-color extraction, correlation based matching, and multiple scale images generation [6].

3.1.2. AV Integration in Location Level

AV integration in location is based on the reported stream based integration method [6]. A speech and a face stream are formed from the extracted speech and face directions by taking their time series into account. A pair of a speech and a face stream which are close for more than a constant time are associated into an associated stream. Because the associated stream includes more accurate directional information by face localization, speech localization improves by the AV integration when face location is available.

3.1.3. Speech Separation

The speech direction obtained by an associated or a speech stream is sent to speech separation by the ADPF which separates sounds originating from a specific direction [7]. Practically, the ADPF separates speech by sub-band selection based on IPD and IID which are estimated by the scattering theory (Eqs. (6) and (7)) as follows:

1. Let θ_s be a direction obtained by AV integration in location, that is, a direction of an associated or a speech stream to be extracted.
2. The pass range $\delta(\theta_s)$ of the ADPF is selected according to a pass range function based on auditory fovea [7]. The pass range function δ has a minimum value in the robot's

front direction because it has maximum sensitivity, and a larger value at the periphery because of lower sensitivity. Let $\theta_l = \theta_s - \delta(\theta_s)$ and $\theta_h = \theta_s + \delta(\theta_s)$.

3. The sub-bands are collected if the IPD, $\Delta\varphi(f)$, and IID, $\Delta\rho(f)$, of the input signal satisfy the specified condition.

$$\begin{aligned} f \le f_{th} &: \Delta\varphi_s(\theta_l, f) \le \Delta\varphi(f) \le \Delta\varphi_s(\theta_h, f), \text{and} \\ f > f_{th} &: \Delta\rho_s(\theta_l, f) \le \Delta\rho(f) \le \Delta\rho_s(\theta_h, f). \end{aligned} \quad (9)$$

where f_{th} is defined as 1500 Hz according to wave length corresponding to a diameter of the robot's head.

4. A wave consisting of collected sub-bands is constructed.

The improvement of about 9 dB in noise reduction has been reported in separation of three simultaneous speeches with the same loudness [7]. This stage works in real-time with a small latency of 200 ms by distributed processing with 3 PCs, networked through Gigabit Ether network.

3.2. Recognition of Separated Speech and Face

The recognition stage consists of three modules. The first module is speech recognition by using multiple DS-dependent acoustic models. ASR systems of which the number is the same as that of the acoustic models are processed in parallel. The second one is face recognition, 3-best name list of a detected face and their belief factors are estimated. The last one is AV integration in name level, that is, integration of speech and face recognition. All results by ASRs are integrated with results of face recognition in the AV integration module.

3.2.1. Speech Recognition by Multiple Acoustic Models

The Japanese automatic speech recognition software "Julian" is used for ASR engine[23]. Hidden Markov Model (HMM) based DS-dependent acoustic models are used. To make DS-dependent acoustic models, 150 words including numbers, colors and fruits by two men (Mr. A and Mr. C) and a woman (Ms. B) are used. Every word is played by loudspeakers of B&W Nautilus 805, and recorded by a pair of humanoid's microphones. The loudspeakers and the humanoid are installed in a 3 m×3 m room with about 0.2 sec of reverberation time, the distance between each loudspeaker and the humanoid is 1 m. Three kinds of speeches are recorded as follows:

1. **single:** A loudspeaker is used for recording. The direction of the loudspeaker varies from -90° to 90° by 10° steps.
2. **double:** Two loudspeakers are used for recording simultaneously. The direction θ_2 of one loudspeaker is among 10°, 20°, $\cdots$, 80° and 90°. The direction of the other loudspeaker is 0° or $-\theta_2$.
3. **triple:** Three loudspeakers are used for recording simultaneously. The direction of the first loudspeaker is fixed to 0°. The direction θ_3 of the second loudspeaker is among 10°, 20°, $\cdots$, 80° and 90°. The direction of the last loudspeaker is $-\theta_3$.

To create training datasets for acoustic models, each speech is separated from recorded data (single, double and triple) by the ADPF under the condition that the directions of loudspeakers are given. The separated speeches are clustered by speaker and direction. As a result, 51 data sets (17 directions $times$ 3 speakers) are obtained as training datasets. By using these training datasets, 51 acoustic models are trained. In this paper, each acoustic model is a triphone model trained 10 times by using Hidden Markov Model Toolkit (HTK). [2]

[2] http://htk.eng.cam.ac.uk/

Julian generates a score which represents logarithmic likelihood of the result. Each score is transformed to a belief factor P_s by using probability density function. Since 51 results are created per input, 51 recognition results with belief factors are sent to the AV integration module.

3.2.2. Face Recognition

Face recognition is basically the same as a method reported in [6]. The face recognition module (see Fig. 1) projects each extracted face into the discrimination space, and calculates its distance to each registered face. Since this distance depends on the degree (the number of registered faces) of discrimination space, it is converted to a parameter-independent belief factor P_v by using a probability density function. The discrimination matrix is created in advance or on demand using a set of variation of the face with an ID (name). This analysis is done by Online Linear Discriminant Analysis. Finally, the face recognition module sends 3-best face ID (Name) with its belief factor P_v to the AV integration module.

3.2.3. AV Integration in Name Level

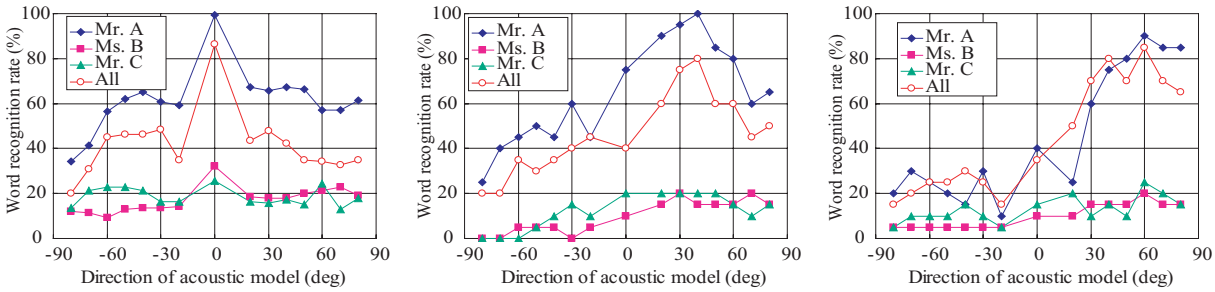
The AV integration module receives 51 speech recognition results with belief factors and face name with a belief factor, and integrates them to output the most reliable result. Although we tried integration of speech and face recognition by utilizing majority rule and voting such as ROVER [13], such integration was effective only when the number of sound directions is at most three, that is, the number of DS-dependent acoustic models is around ten. The number of misclassified results increase, as the number of sound direction is large. Because a set of such misclassified results affect the integration badly, the effectiveness of the integration based on majority rule or voting is weak in the situation where 17 sound directions are assumed.

To define a suitable algorithm for the integration of a large number of acoustic models, we measured word recognition rate against DS-dependent acoustic models when a speaker and a direction of input speech are fixed. Figures 2a), b) and c) are distributions of the results against Mr. A's speech from 0°, 30° and 60°, respectively. In these Figs, the x axis is direction of acoustic model, and the y axis is word recognition rates. The same speech data as a training dataset is used for recognition. The lines labeled "Mr. A", "Ms. B", "Mr. C" are the results by using acoustic models of "Mr. A", "Ms. B", "Mr. C". The line labeled "All" means the results by direction-dependent and speaker-independent acoustic model. When both of the person and the direction are correct, the word recognition rate is more than 90%, and better than that using speaker independent acoustic models. When direction of speech is correct, speaker independent acoustic models are useful for an unknown speaker and in case that the face name is unavailable. By taking the results in Fig. 2 into account, the AV integrator uses a cost function by Eq. (10) to integrate the results.

$$\begin{aligned} V(p_e) &= P_v(p_e) \cdot \left(\sum_d r(p_e, d) \cdot v(p_e, d) \cdot P_s(p_e, d) \right. \\ &\quad + \sum_p r(p, d_e) \cdot v(p, d_e) \cdot P_s(p, d_e) \\ &\quad \left. - r(p_e, d_e) \cdot P_s(p_e, d_e) \right). \end{aligned} \quad (10)$$

$$v(p, d) = \begin{cases} 1 & if \quad Res(p, d) = Res(p_e, d_e), \\ 0 & if \quad Res(p, d) \neq Res(p_e, d_e). \end{cases}$$

where $r(p, d)$ and $Res(p, d)$ are recognition rate shown in Fig. 2 and recognition result against input speech when an acous-



a) Mr. A, 0° b) Mr. A, 30° c) Mr. A, 60°

Figure 2: word recognition rates based on DS-dependent acoustic models (closed test)

tic model of person p and sound direction d is used. The d_e is the sound source direction estimated by the real-time tracking system, and the p_e is a person to be evaluated. $P_v(p_e)$ is a probability in the face recognition module, and it is set to 0.5 which represents "unknown face" when face recognition is unavailable. Finally, the AV integration module selects person p_e and result $Res(p_e, d_e)$ with the largest $V(p_e)$.

If the largest $V(p_e)$ is too small (less than 1) or close to the second largest one, the humanoid turns to the sound source and asks the person corresponding to the sound source again to make sure what he/she said.

Thus, the system can recognize simultaneous speeches and identify the speaker by using multiple acoustic models and face recognition.

4. Evaluation

The efficiency by the AV integration in the system is evaluated through recognition of "three" simultaneous speeches with/without using face recognition. In sound source separation, to prove the efficiency of the scattering theory based estimation of IPD and IID, two other estimation methods based on the auditory epipolar geometry and and HRTF measured in an anechoic room are evaluated as well. In case of the auditory epipolar geometry based estimation, $\Delta\varphi_e(\theta)$ defined by Eq. (8) is used in Eq. (9) instead of using $\Delta\varphi_s(\theta)$ defined by Eqs. (6) and (7) for every sub-band. In case of HRTF, IPD $\Delta\varphi_h(\theta)$ and IID $\Delta\rho_h(\theta)$ obtained from HRTF are used instead of $\Delta\varphi_s(\theta)$.

In experiments, room conditions are the same as those described in Sec. 3.2.1. The three loud speakers are attached photographs of speakers for face recognition instead of real humans. For recording of three simultaneous speeches, a three-word combination is selected from a list of three-word combinations in training datasets. Then, three loudspeakers play the three words according to the combination. The mixture of sounds is captured by humanoid's microphones and sent to the system.

The direction of first speaker is fixed to 0°. The second speaker direction θ varies from 0° to 90° by 10°. The direction of the last loudspeaker is $-\theta$. Figures 3 and 4 show word recognition rate without and with face recognition, respectively. In Figure 3, P_v in Eq. (10) is fixed to 0.5 because face recognition is unavailable. The X and Y axes of figures mean direction difference between loudspeakers θ and average word recognition rate of left, center and right speech in percentage. The lines labeled "AEG", "HRTF" and "ST" are recognition rates obtained by using the auditory epipolar geometry, HRTF and the scattering theory, respectively.

Figures 3 and 4 show the efficiency of the AV integration. The changes of word recognition rate against direction difference between loudspeakers are smaller in Fig. 4. Especially, when direction difference between loudspeakers is from 20° to 70°, the performance of speech recognition by using the scattering theory and HRTF is about 80% in Fig. 4. In the auditory epipolar geometry based method, we can find some improvement of speech recognition by AV integration. This proves that face recognition is also efficient for improvement of speech recognition by AV integration although lip-reading is often used for such visual information.

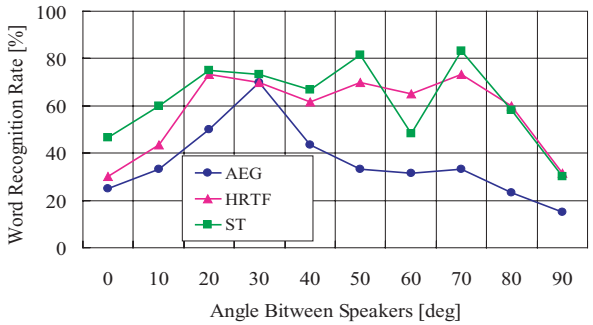


Figure 3: Word recognition rate (baseline)

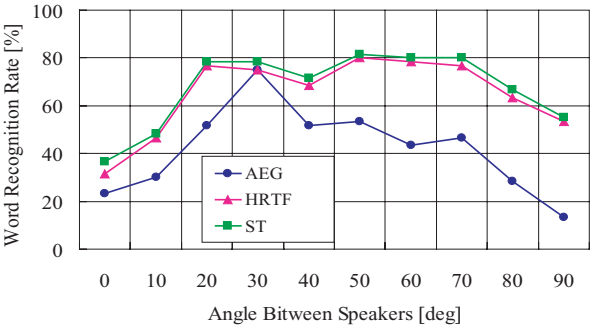


Figure 4: Word recognition rate improved by AV Integration

Figures 3 and 4 shows that the scattering theory based estimation method has the best performance. This means that it estimates IPD and IID well and improves the performance of the ADPF. Compared with the auditory epipolar geometry based method, the word recognition rate is getting much better as the left and the right loudspeakers become close to 90°. This shows that the IPD and IID estimation by the scattering theory is still accurate even when the sound source direction is away from the front direction. The difference between HRTF and the scattering theory based methods is small. The HRTF based separation underperforms because the HRTF is measured in an anechoic room while it is used in an echoic room. Therefore, the scattering theory based method estimates IPD and IID better in the

room with small reverberation time of about 0.2 sec. In addition, scattering theory based method can estimate IPD and IID without measurement in advance. That is the point that the scattering theory based method is superior to HRTF based one.

5. Future Work

In this paper, we still assume azimuth direction, known speakers and isolated word recognition. Robust system against elevation and distance can be obtained by using multiple elevation-dependent and distance-dependent acoustic models with compensation method to obtain an acoustic model for any direction. Introduction of speaker independent acoustic models and speaker adaptation function would be efficient to realize more general system which works properly for unknown persons. The word spotting techniques are effective for speech recognition of sentences. The ASR engine should be considered to cope with characteristics of the ADPF, front-end processing of speech recognition. Because the ADPF separates sound sources based on sub-band selection in frequency domain, some sub-bands can be dropped by separation errors. A sudden and loud noise can affect across wide frequency ranges. To cope with these cases, missing feature theory [24] and missing data theory [25, 26] would be effective.

6. Conclusion

Improvement of recognition of three simultaneous speeches by two microphones is described. We proved that speech separation improved by the proposed scattering theory based method which estimates IPD and IID better than HRTF and the auditory epipolar geometry based ones under weak reverberation. The recognition results show that two-layered AV integration – AV integration in location to obtain robust speech direction and AV integration in name by combination of face recognition and multiple speech recognition results obtained by DS-dependent acoustic models – are efficient and essential to improve speech recognition in a real environment.

7. References

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

[2] H. Saruwatari *et al*, "Speech enhancement using nonlinear microphone array based on complementary beamforming," *IEICE Trans. fundamentals*, E82-A (8), 1999.

[3] N. Murata and S. Ikeda, "An on-line algorithm for blind source separation on speech signals," *Int'l Sym. on Nonlinear Theory and its Applications*, 1998, pp. 923–927.

[4] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *IJCAI-99*, pp. 1146–1151.

[5] F. Asano *et al*, "Real-time sound source localization and separation system and its application to automatic speech recognition.," *Eurospeech 2001*. pp. 1013–1016, ESCA.

[6] K. Nakadai *et al*, "Real-time auditory and visual multiple-object tracking for robots," *IJCAI-01*. pp. 1424–1432.

[7] K. Nakadai *et al*, "Auditory fovea based speech separation and its application to dialog system," *IROS-2002*. pp. 1314–1319, IEEE.

[8] K. C. Yen and Y. Zhao, "Robust automatic speech recognition using a multi-channel signal separation front-end," *ICSLP-96*. vol. 3, pp. 1337–1340, ISCA.

[9] S. A. Shafer *et al*, "An Architecture for Sensor Fusion in a Mobile Robot," *IEEE Conf. on Robotics and Automation*. pp. 2002–2011, 1986.

[10] Y. Nakagawa *et al*, "Using vision to improve sound source separation," *AAAI-99*. pp. 768–775.

[11] J. Luettin and S. Dupont, "Continuous audio-visual speech recognition," *ECCV-98*. vol. II of *Lecture Notes in Computer Science*, pp. 657–673, Springer Verlag.

[12] P.L. Silsbee and A.C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[13] J.G. Fiscus, "A post-processing systems to yield reduced word error rates: Recognizer output voting error reduction (rover)," *ASRU-97*. pp. 347–354, IEEE.

[14] R. Brooks *et al*, "The cog project: Building a humanoid robot," *Computation for metaphors, analogy, and agents*, C.L. Nehaniv, Ed. 1999, pp. 52–87, Spriver-Verlag.

[15] R. E. Irie, "Multimodal sensory integration for localization in a humanoid robot," *CASA'97*. pp. 54–58, IJCAI.

[16] Y. Matsusaka *et al*, "Multi-person conversation via multi-modal interface — a robot who communicates with multi-user," *EUROSPEECH-99*. pp. 1723–1726, ESCA.

[17] A. Takanishi *et al*, "Development of an anthropomorphic auditory robot that localizes a sound direction (*in japanese*)," *Bulletin of the Centre for Informatics*, vol. 20, pp. 24–32, 1995.

[18] L.A. Jeffress, "A place theory of sound localization," *Journal of Comparative Physiology, Psycology*, vol. 41, pp. 35–39, 1948.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, ACADEMIC PRESS, 1989.

[20] O. D. Faugeras, *Three Dimensional Computer Vision: A Geometric Viewpoint*, The MIT Press, 1993.

[21] P. Lax and R. Phillips, *Scattering Theory*, Academic Press, 1989.

[22] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, *Electromagnetic and Acoustic Scattering by Simple Shapes*, Hemisphere Publishing Co., 1987.

[23] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," *EUROSPEECH-01*. pp. 1691–1694.

[24] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," *ICASSP-1997*. pp. 1255–1258.

[25] P. Renevey, R. Vetter, and J. Kraus, "Robust speech recognition using missing feature theory and vector quantization," *EUROSPEECH-01*. vol. 2, pp. 1107–1110, ESCA.

[26] J. Barker, M.Cooke, and P.Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," *EUROSPEECH-01*. vol. 1, pp. 213–216, ESCA.