Three Simultaneous Speech Recognition by Integration of Active Audition and Face Recognition for Humanoid

Kazuhiro Nakadai, Daisuke Matsuura, Hiroshi G. Okuno, and Hiroshi Tsujino

Kitano Symbiotic Systems Project, ERATO, Japan Science and Technology Corp.
Graduate School of Science and Engineering, Tokyo Institute of Technology,
Graduate School of Informatics, Kyoto University
Honda Research Institute Japan Co., Ltd.
nakadai@nakadai.com, okuno@nue.org, matsuurdm@mep.titech.ac.jp, tsujino@jp.honda-ri.com

Abstract

This paper addresses listening to three simultaneous talkers by a humanoid with two 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. Humanoid audition system consists of sound separation, face recognition and ASR. Sound sources are separated by an active direction-pass filter (ADPF), which extracts sounds from a specified direction in real-time. Since features of sounds separated by ADPF vary according to the sound direction, ASR uses multiple direction- and speaker-dependent acoustic models. 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 improves word recognition rates against three simultaneous utterances.

1. Introduction

Fifty years ago, Cherry discovered cocktail party effect which means that at a crowded party one can attend one conversation and then switch to another one [1]. The essence of cocktail party effect resides in selective attention to one sound stream included in an input mixture of sounds. By elaborating upon the cocktail party effect, one may expect that people or robots can listen to at most two petitions and give appropriate decisions at the same time. Psychological study showed that people can listen to at most two things at the same time [2].

Research on computational auditory scene analysis (CASA) focuses on the computer modeling and implementation for the understanding of acoustic events [3]. One common way for CASA is sound source separation by using auditory cues such as common onset, AM, FM harmonic structure, formants, and sound source localization. Other approaches to sound source separation are based on signal processing and information theory; microphone arrays with beam-forming techniques [4], and independent component analysis or blind source separation [5]. The problem of latter approaches is the theoretical limitation that the number of sound sources should be equal to or less than that of microphones.

In robotics, source separation has not studied so much. Most robots have a microphone attached near the mouth of each speaker to avoid motor noise in motion and other sounds. Otherwise, they adopt the “stop-hear-act” principle: that is, a robot stops to hear [6]. Some robots have a sound source separation function [7]. However, because their system requires a lot of measurement in advance, and has difficulty in separation during motion, while human can hear during motion, it is not enough for robot audition to be deployed in the real world yet.

Sound source separation is effective for front-end processing to improve speech recognition under noisy environments [8, 9]. For example, recognition of three simultaneous speech with two microphones by integrating visual and auditory information has been reported [9]. Although their system separates speech streams by using interaural phase difference (IPD), interaural intensity difference (IID) and sound direction by image, it has studied with synthesized voices under simulated and offline environment, and neither microphones nor sound sources are movable. Such assumptions are too strong to apply real-world systems such as mobile robots.

From the viewpoint of improving speech recognition, various integration based approaches have been studied. An approach of improving speech recognition is audio-visual integration. Most of audio-visual integration in speech recognition use visual speech, that is, lip-reading [10, 11]. In a robot, however, lip-reading is not always available because the robot is too far from a person to detect his lips. A face is generally detected easier than the lips due to its size. Therefore, face recognition is more convenient for robots than lipreading. In this paper, improvement of speech recognition by integration of speech and face recognition is reported. The other effective approach of such integration is the use of multiple results obtained from same or different ASRs. For example, ROVER [12] that integrates different ASRs by a weighted voting method and integration of ASRs based on confidence measure have been reported.

Our approach uses integrates of recognition results obtained by 51 direction- and speaker-dependent acoustic models to improve recognition of speech separated by the active direction-pass filter (ADPF) [13] that separates sound source originating from a specified direction. On the use of a large number of acoustic models, the integration of simple voting or majority rule often fails because a lot of wrong and same result affect the system badly. Then, we integrate the results based on word recognition rate of each acoustic model.

The rest of the paper is organized as follows: Section 2 presents the system for speech recognition by integration of active audition and face recognition. Section 3 evaluates the sys-
2. Simultaneous Speech Recognition by Audio-Visual Integration

The robot audition system for simultaneous speech recognition consists of a humanoid and three sub-systems – real-time human tracking, sound source separation by active direction-pass filter (ADPF) and speech recognition by AV integration. The architecture of the system is shown in Fig. 1.

As a testbed of this work, an upper torso humanoid is used. It has a pair of CCD cameras (Sony EVI-G20) for face recognition and stereo vision, and a pair of microphones for sound source separation and speech recognition. It is driven by 4 DC motors (4 DOFs) with functions of position and velocity control by using potentiometers.

Sounds and images captured by robot’s microphones and cameras are sent to the real-time human tracking subsystem [14]. The subsystem extracts directions of multiple sound sources from auditory and visual streams formed by fusing information obtained by sound source localization, multiple face localization and object localization by stereo vision. It also tracks one of the extracted sound sources according to focus-of-attention. The subsystem works in real time with a small latency of 200 ms by distributed processing with 5 PCs, networked through gigabit ether net.

The sound source directions are sent to the sound source separation by the ADPF. It separates sounds originating from the direction by using a pair of microphones [13]. The filtering process is implemented by hypothesis matching for each sub-band of interaural intensity difference and interaural phase difference which are calculated from input spectra of left and right channels. The performance of the ADPF shows the difference of resolution in sound localization and separation. The resolution of localization and separation of the center of the humanoid is much higher than that of peripherals, indicating similar property of visual fovea (high resolution in the center of human eye). To exploit this auditoryfovea, the ADPF controls the pass range of the filter according to the sound direction and the direction of a head by motor movement. The improvement of about 9 dB in noise reduction is reported in separation of three simultaneous speech with the same loudness.

The speech recognition subsystem recognizes the extracted speech by multiple acoustic models and speaker ID by face recognition.

2.1. Speech Recognition by AV Integration

The speech recognition subsystem consists of three processes. The first process is speech recognition by using multiple acoustic models. The acoustic models are direction- and speaker-dependent (DS-dependent). ASRs of which the number is the same as that of the acoustic models are processed in parallel. The second one is face detection and recognition, 3-best name list of a detected face and their belief factors are estimated. The last one is integration of speech and face recognition.

2.1.1. Speech Recognition by Multiple Acoustic Models

In the speech recognition subsystem, Hidden Markov Model (HMM) based acoustic models are used. The acoustic models are DS-dependent. The Japanese automatic speech recognition software “Julian” is used for ASR. Multiple ASRs are processed in parallel, and all results are integrated in the integrator with results of the face recognition module.

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, the distance between each loudspeaker and the humanoid is 1 m. Three kinds of speech are recorded as follows:

1. **Single**: A loudspeaker is used for recording. The direction of the loudspeaker varies from -90° to 90° by 10°.
2. **Double**: Two loudspeakers are used for recording simultaneously. The direction $\theta_1$ of one loudspeaker is among 20°, 30°, ..., 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 $\theta_3$ of the second loudspeaker is among 20°, 30°, ..., 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 х 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).

Julian generates a score which represents logarithmic likelihood of the result. Each score is transformed to a belief factor $P_r$ by using probability density function. Since the subsystem creates 51 results per input, 51 recognition results with belief factors are sent to the AV integrator.

2.1.2. Face Recognition

Face recognition is basically the same as a method reported in [14]. Since the visual processing detects several faces simultaneously, extracts, identifies and tracks each face, the size, direction, and brightness of each face changes frequently. The key idea is the combination of skin-color extraction, correlation based matching, and multiple scale images generation. 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 $P_r$ by using 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 $P_r$ to the AV integrator.

2.1.3. Integration of speech and face recognition

The AV integrator receives 51 speech recognition results with belief factors and face name with a belief factor, and integrate them to output the most reliable result.

We tried integration based on majority rule and voting such as ROVER [12]. Such integration methods was effective only when the number of sound direction is a few, that is, the number of DS-dependent acoustic models is up to ten. The number of wrong and same results increase, as the number of sound direction is large. Because a set of such wrong and same results affect the integration badly, the effectiveness of the integration based on majority rule or voting is useless 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 speaker and 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. These results show that the influence by direction is less than by speaker. 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. By taking the results in Fig. 2 into account, the AV integrator uses a cost function by Eq. 1 to integrate the results.

$$V(p_r) = \left( \sum_{d_s} r(p_r, d_s) \cdot v(p_r, d_s) \cdot P_r(p_r, d_s) \right) + \sum_{p} r(p, d_s) \cdot v(p, d_s) \cdot P_r(p, d_s) - r(p_r, d_s) \cdot P_r(p_r, d_s) \cdot P_r(p_r) \cdot P_r(p_r).$$

where $r(p, d)$ and $Re s(p, d)$ are recognition rate shown in Fig. 2 and recognition result against input speech when an acoustic model of person $p$ and sound direction $d$ is used. The $d_s$ is the sound source direction estimated by the real-time tracking system, and the $p_r$ is a person to be evaluated. $P_r(p_r)$ is a probability in the face recognition module, and it is set to 0.5 when face-recognition is unavailable. Finally, the AV integrator selects person $p_r$, and result $Re s(p_r, d_s)$ with the largest $V(p_r).$

If the largest $V(p_r)$ 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 speech and who spoke each speech by using multiple acoustic models and face recognition.

3. Evaluation

The efficiency by the AV integration in the system is evaluated through “three” simultaneous speech recognition with/without using face recognition.

In experiments, room conditions are the same as those described in Sec. 2.1.1. The three loud speakers are attached photographs of speakers for face recognition instead of real humans. For recording of three simultaneous speech, 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 $\theta_2$ varies from 20° 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 Exp. 2, $P$, in Eq.1 is fixed to 0.5 because face recognition is unavailable. The X and Y axes of figures mean direction difference between loudspeakers $\Delta \theta$ and word recognition rate in percentage. The lines labeled “left”, “center” and “right” are 1-best recognition rates of left, center and right loudspeakers, respectively. The dotted lines are 3-best recognition rates.

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. This indicates that the AV integration compensates such changes. In Fig. 4, the recognition rates of 1-best and 3-best are close. This means that the AV integration encourages a belief factor of a correct answer properly.

Figures 3 and 4 prove the efficiency of the AV integration. The changes of word recognition rate against direction difference between loudspeakers are smaller in Fig. 4. This indicates that the AV integration compensates such changes. In Fig. 4, the recognition rates of 1-best and 3-best are close. This means that the AV integration encourages a belief factor of a correct answer properly.

3. Results

Three simultaneous speech recognition by two microphones is described. The results show that various kinds of integration – the use of multiple DS-dependent acoustic models, AV integration by combination of face recognition and DS-dependent acoustic models, and active audition combining audition with motion – is efficient and essential to improve speech recognition.

In the integration of a large number of results, simple voting or majority rule are of less use. The system proves that our integration based on belief factor is effective.

However, in the ADPF, some sub-bands can be dropped by separation errors. A sudden and loud noise can affect across wide frequency ranges in a moment. In such cases, happens cannot be used for speech recognition. To cope with such situation, introduction of missing feature/data theory[15, 16] and word spotting is our future work.

4. Conclusion

Three simultaneous speech recognition by two microphones is described. The results show that various kinds of integration – the use of multiple DS-dependent acoustic models, AV integration by combination of face recognition and DS-dependent acoustic models, and active audition combining audition with motion – is efficient and essential to improve speech recognition.

In the integration of a large number of results, simple voting or majority rule are of less use. The system proves that our integration based on belief factor is effective.

However, in the ADPF, some sub-bands can be dropped by separation errors. A sudden and loud noise can affect across wide frequency ranges in a moment. In such cases, happens cannot be used for speech recognition. To cope with such situation, introduction of missing feature/data theory[15, 16] and word spotting is our future work.

5. References