The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Two-Layered Audio-Visual Speech Recognition for Robots in Noisy Environments

Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno.

Abstract— **Audio-visual (AV) integration is one of the key ideas to improve perception in noisy real-world environments. This paper describes automatic speech recognition (ASR) to improve human-robot interaction based on AV integration. We developed AV-integrated ASR, which has two AV integration layers, that is, voice activity detection (VAD) and ASR. However, the system has three difficulties: 1) VAD and ASR have been separately studied although these processes are mutually dependent, 2) VAD and ASR assumed that high resolution images are available although this assumption never holds in the real world, and 3) an optimal weight between audio and visual stream was fixed while their reliabilities change according to environmental changes. To solve these problems, we propose a new VAD algorithm taking ASR characteristics into account, and a linear-regression-based optimal weight estimation method. We evaluate the algorithm for auditory- and/or visually-contaminated data. Preliminary results show that the robustness of VAD improved even when the resolution of the images is low, and the AVSR using estimated stream weight shows the effectiveness of AV integration.**

I. Introduction

In a daily environment where service/home robots are expected to communicate with humans, the robots have a difficulty in Automatic Speech Recognition (ASR) due to various kinds of noises such as other speech sources, environmental noise, room reverberations, and robots' own noise. In addition, properties of the noises are not always known. Therefore, an ASR system for a robot should cope with the input speech signals with an extremely low Signal-to-Noise Ratio (SNR) by using less prior information on the environment.

An ASR system generally consists of two main processes. One is Voice Activity Detection (VAD) and the other is ASR. VAD is the process, which detects start and end points of utterances from an input signal. When the duration of the utterance is estimated shorter than the actual one, the beginning and/or the last part of the utterance is missing, thus ASR fails. Also, an ASR system requires some silent signal parts (300-500 ms) before and after each utterance. When the silent parts are too long, it also affects the ASR system badly. Even if VAD detects an utterance correctly, ASR may fail due to a noise. Thus, to make an ASR system robust, both VAD and ASR should be noise-robust.

To realize such a noise-robust ASR system, there are mainly two approaches. One is sound source separation to improve SNR of the input speech. The other is the use of another modality, that is, Audio-Visual (AV) integration.

For sound source separation, we can find several studies, especially, in the field of "Robot Audition" proposed in [1], which aims at building listening capability for a robot by using its own microphones. Some of them reported highly-noise-robust speech recognition such as three simultaneous speeches [2]. However, in a daily environment where acoustic conditions such as power, frequencies, locations of noises and speech sources dynamically change, the performance of sound source separation sometimes deteriorates, and thus ASR does not always show such a high performance.

Regarding AV integration for ASR, many studies have been reported as Audio-Visual Speech Recognition (AVSR) [3], [4], [5]. However, they assumed that the high resolution images of the lips are always available. Thus, their methods have difficulties in being applied to robot applications.

To tackle with these difficulties, we reported AVSR for a robot based on two psychologically-inspired methods [6]. One is Missing Feature Theory (MFT), which improves noise-robustness by using only reliable audio and visual features by masking unreliable ones out. The other is coarse phoneme recognition, which also improves noise-robustness by phoneme groups consisting of perceptually-close phonemes instead of using phonemes as units of recognition. The AVSR system showed high noise-robustness to improve speech recognition even when either audio or visual information is missing and/or contaminated by noises. However, the system assumed that the voice activity was given while VAD affects ASR performance. To cope with this issue, we reported an ASR system based on AV-VAD and AVSR [7]. The AVSR system showed high speech recognition performance when either audio or visual information is contaminated by noises. However, the system has three issues as follows:

1) A priori information was used to integrate audio and visual information,
2) VAD and ASR have been separately studied although these processes are mutually dependent,
3) The evaluation of the system was done without using an actual robot.

For the first issue, we introduce stream weight optimization which can control AV integration to improve AVSR performance depending on acoustic- and visual-noise. For the

T. Yoshida and K. Nakadai are with Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8522, JAPAN. yosihda@cyb.mei.titech.ac.jp

K. Nakadai is also with Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0114, JAPAN, nakadai@jp.honda-ri.com

H. G. Okuno is with Graduate School of Informatics, Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto 606-8501, JAPAN okuno@kuis.kyoto-u.ac.jp

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

second issue, we evaluate the performance of AVSR system, that is, the combination of VAD and ASR. For the third issue, we evaluate the performance of the proposed method using an actual robot.

The rest of this paper is organized as follows: Section II discusses issues and Section III explains approaches in audio-visual integration for an ASR system. Section IV describes our ASR system for a robot using two-layered AV integration. Section V shows evaluations in terms of VAD and ASR. The last section concludes this paper.

II. Issues in Audio and Visual Integration for Speech Recognition

AV integration methods for VAD and ASR are mainly classified into two approaches. One is early integration which integrates audio and visual information on the feature level and the other is late integration which integrates audio and visual information on the decision level.

In the following sections, we discuss issues in audio-visual integration methods for VAD and ASR for a robot because VAD and ASR are essential functions for an ASR system.

A. Audio-Visual Integration for VAD

AV integration is promising to improve the robustness of VAD, and thus audio visual integration should be applied to VAD in the real world. Almajai *et al.* presented an AV-VAD based on early integration [8]. They integrated an audio feature (Mel-Frequency Cepstral Coefficient: MFCC) and a visual feature based on 2-D Discrete Cosine Transform (DCT). The integrated AV feature was used to detect voice activities and this approach showed a high performance. However, they assumed that the high resolution image of the lips is always available. Murai *et al.* presented an AV-VAD based on late integration [9]. They detected lip activity by using a visual feature based on a temporal sequence of lip shapes. Then, they detected a voice activity from the detected lip activity by using speech signal power. Therefore, this method detects the intersection of the lip activity and the voice activity. However, in this case, when either the first or the second detection fails, the performance of the total system deteriorates.

B. Audio-Visual Integration for ASR

For AVSR, we use early integration because when we use late integration for AVSR, we have to consider the difference between alignments of two recognition results. The alignments are separately estimated in the audio- and visual-based speech recognizers. Thus, it is time consuming to integrate the recognition results, because a lot of hypotheses are tested to find the best alignment correspondence between the results.

The issue in early integration is how to integrate audio and visual features. When an audio feature is reliable and a visual feature is unreliable, AVSR should place more emphasis on the audio feature and less on the visual feature, and vice versa. AVSR realizes this control by using a weight called stream weight. So, a lot of stream weight optimization methods have been studied in the AVSR community. They mainly used log likelihoods in audio and/or visual speech models. Optimization methods based on discriminative criteria, such as minimum classification error criterion [10], maximum mutual information criterion [11], and maximum entropy criterion [12] have been proposed. These methods have been reported with good performance. However, these methods mainly dealt with acoustic noise and assumed that ideal images are available. Thus, these methods are difficult to apply an ASR system to a robot directly. To apply these methods, we have to cope with dynamic changes of resolution, illumination, or face orientation. Resolution is especially important because the performance of visual speech recognition (lip reading) drops when the resolution is low.

III. Approaches in Audio and Visual Integration for Speech Recognition

A. Audio-Visual Integration for VAD

To solve the issues in AV-VAD, we introduced AV-VAD based on Bayesian network [13], because Bayesian network provides a framework that integrates multiple features with some ambiguities by maximizing the likelihood of the total integrated system. Actually, we used the following features as the inputs of the Bayesian network:

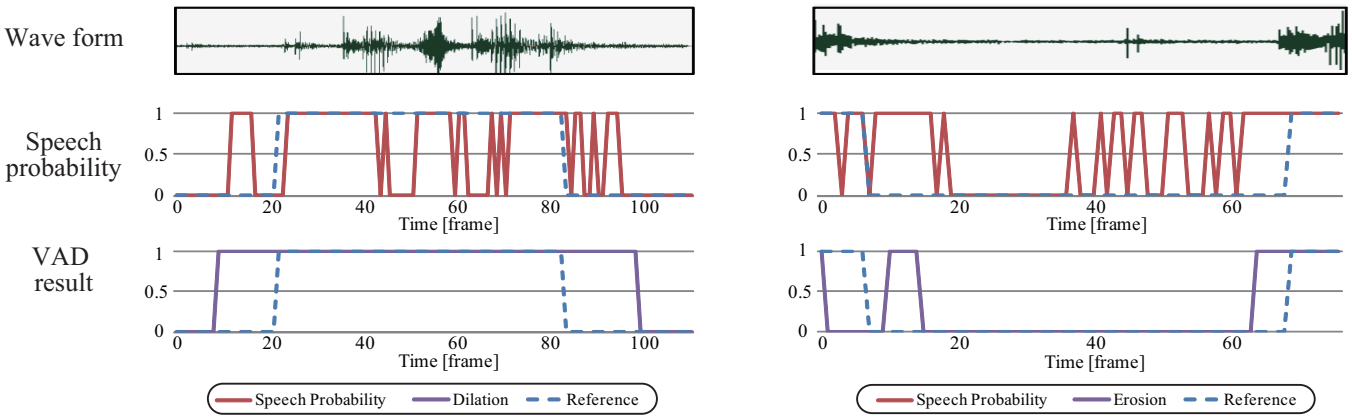
- The score of log-likelihood for silence calculated by speech decoder (x_{dvad}),
- An eight dimensional feature based on the height and the width of the lips (x_{lip}),
- The belief of face detection which is estimated in face detection (x_{face}).

The first feature x_{dvad} is calculated by using an acoustic model of speech recognition and thus, takes the property of voice into account. This feature reported high noise-robustness [14]. The second feature is derived from the temporal sequence of the height and width information by using linear regression [13]. The last feature is calculated in the face detection process. Since these features, more or less, have errors, the Bayesian network is an appropriate framework for AV integration in VAD.

First, we calculate a speech probability by using a Bayesian network. The Bayesian network is based on the Bayes theory defined by

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}, \quad j = 0, 1 \tag{1}$$

where x corresponds to each feature such as x_{dvad}, x_{lip}, or x_{face}. A hypothesis ω_j shows that ω_0 or ω_1 corresponds to a silence or a speech hypothesis, respectively. A conditional probability, $p(x|\omega_j)$, is obtained by using a Gaussian Mixture Model (GMM) which is trained with a training dataset in advance. The probability density functions $p(x)$ and the probability $P(\omega_j)$ are also pre-trained with the training dataset.



a) Effect of dilation b) Effect of erosion

Fig. 1. Hangover processing based on erosion and dilation

A joint probability, $P(\omega_j|x_{dvad}, x_{lip}, x_{face})$, is thus calculated by

$$P(\omega_j|x_{dvad}, x_{lip}, x_{face}) = P(\omega_j|x_{dvad})P(\omega_j|x_{lip})P(\omega_j|x_{face}). \quad (2)$$

By comparing this probability and a threshold, we estimate a voice activity.

Next, we perform hangover processing based on **dilation** and **erosion** for the temporal sequence of estimated voice activity. Dilation and erosion are commonly used in pattern recognition. Figure 1 shows hangover process based on dilation and erosion. In the dilation process, a frame is added to the start- and the end-points of a voice activity as below.

$$\hat{V}[k] = \begin{cases} 1 & \text{if } V[k-1] = 1 \text{ or } V[k+1] = 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $V[k] = \{0(\text{non-speech}), 1(\text{speech})\}$ is the estimated voice activity at k frame and $\hat{V}[k]$ is the result of dilation. This process removes fragmentation as shown in Fig. 1a). In erosion process, a frame is removed from the start- and the end-points of a voice activity as below.

$$\hat{V}[k] = \begin{cases} 0 & \text{if } V[k-1] = 0 \text{ or } V[k+1] = 0 \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

This erosion process removes false detects as shown in Fig. 1b). AV-VAD performs these processes several times and decides a voice activity.

Finally, the detected term and additional margin are extracted as an input data for speech recognition, because a speech model assumes that the first and the last period are silent. These margins before and after the detected period correspond to the silent period in the first and last part of a HMM-based speech model used in ASR. Therefore, when we extract a voiced term strictly, the mismatch between the speech model and the inputted data badly affects to the performance of ASR.

B. Audio-Visual Integration for ASR

To cope with the issues in section II-B, we introduced stream weight optimization. This optimization is based on SNR and a face size which is directly affected by the image resolution.

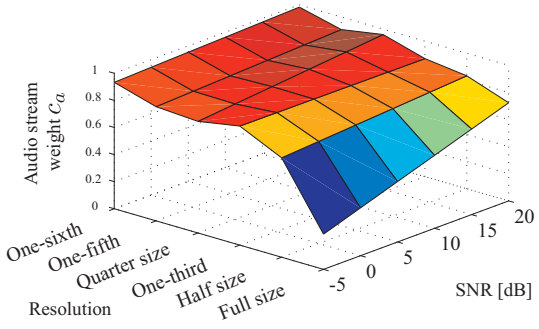


Fig. 2. The linear regressions of optimal stream weight

We first evaluate AVSR by changing the audio stream weight from 0 to 1 at 0.1 increments. From the word correct rate of this test, we decide optimal stream weights for every SNR and image resolution. The estimated audio stream weight is calculated from linear regression of optimal audio stream weights. Figure 2 shows the linear regressions obtained by optimal stream weights.

IV. Automatic Speech Recognition Based on Two-Layered Audio-Visual Integration

Figure 3 shows our automatic speech recognition system for a robot with two-layered AV integration, that is, AV-VAD and AVSR. It consists of four implementation blocks as follows;

- Visual feature extraction block,
- Audio feature extraction block,
- The first layer AV integration for AV-VAD,
- The second layer AV integration for AVSR.

In the following sections, we describe three of these four blocks because AV-VAD is already described in the previous section.

A. Visual Feature Extraction Block

This block consists of four modules, that is, face detection, face size extraction, lip extraction, and visual feature extraction. Their implementation is based on Facial Feature Tracking SDK which is included in MindReader [1]. Using

[1]http://mindreader.devjavu.com/wiki

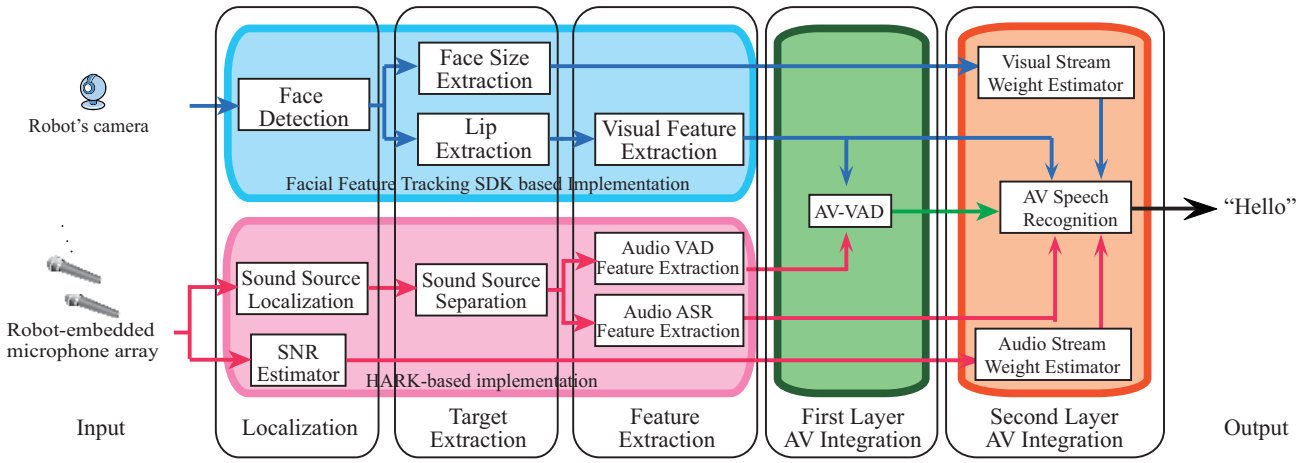


Fig. 3. An automatic speech recognition system with two-layered AV integration for robots

this SDK, we detected a face and facial components like the lips. Because the face and the lips are detected with its left, right, top, and bottom points, we can easily compute the height and the width of the face and lips, and normalize the lips height and width by using the face size estimated in face detection. We use an 8-dimensional visual feature **??** both in VAD and in ASR.

B. Audio Feature Extraction Block

This block consists of five modules, that is, sound source localization, SNR estimation, sound source separation, audio VAD feature extraction, and audio ASR feature extraction. Their implementation is based on HARK mentioned in Section I. The audio VAD feature extraction module was already explained in Section II, and thus, the other three modules are described. We use an 8 ch circular microphone array which is embedded around the top of our robots head.

For sound source localization, we use MUltiple SIgnal Classification (MUSIC) [15]. This module estimates sound source directions from a multi-channel audio signal input captured with the microphone array.

For sound source separation, we used Geometric Sound Separation (GSS) [16]. GSS is a kind of hybrid algorithm of Blind Source Separation (BSS) and beamforming. GSS has high separation performance originating from BSS, and also relaxes BSS's limitations such as permutation and scaling problems by introducing "geometric constraints" obtained from the locations of microphones and sound sources obtained from sound source localization.

For an acoustic feature for an ASR system, Mel Frequency Cepstrum Coefficient (MFCC) is commonly used. However, sound source separation produces spectral distortion in a separated sound, and such distortion spreads over all coefficients in the case of MFCC. Since Mel Scale Logarithmic Spectrum (MSLS)[17] is an acoustic feature in the frequency domain, and thus, the distortion is concentrated only on specific frequency bands. Therefore MSLS is suitable for ASR with microphone array processing. We used a 27-dimensional MSLS feature vector consisting of 13-dim MSLS, 13-dim ΔMSLS, and Δlog power.

C. The Second Layer AV Integration Block

This block consists of two modules, that is, stream weight estimation and AVSR. We introduced a stream weight optimization module which is mentioned in Section II. For AVSR implementation, MFT-based Julius [18] was used.

V. Evaluation

A. Experiments and Evaluations

We evaluated the system through three experiments.

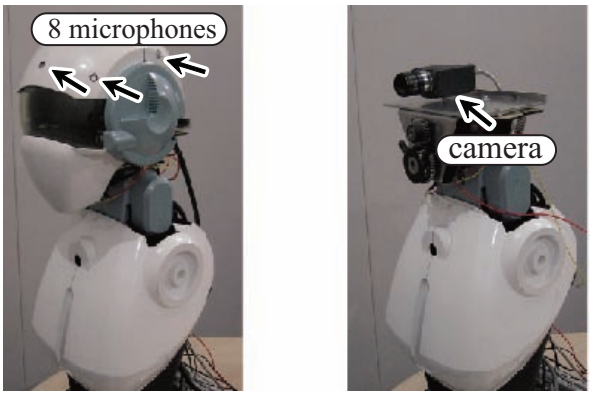
Ex.1: The effectiveness of AV-integration for VAD.
Ex.2: The effectiveness of two-layered AV-integration for ASR.
Ex.3: The robustness against an auditory- and/or visually-contaminated data using an actual robot.

In **Ex. 1** and **Ex. 2**, we used a Japanese word AV dataset, and in **Ex. 3**, we used a scenario captured by using an actual robot.

AV dataset contains speech data from 10 males and 266 words for each male. Audio data was sampled at 16 kHz and 16 bits, and visual data was 8 bit monochrome and 640×480 pixels in size, recorded at 33 Hz.

For training, we used acoustically- and visually-clean AV data. To train an AV-VAD model, we used 216 clean AV data from 5 males. To train an AVSR acoustic model, we used 216 clean AV data from 10 males.

For evaluation, we used two kinds of datasets. One is 50 AV data which is not included in the training dataset. The other is AV data captured by the actual robot. For the former data, the audio data was converted to 8 ch data so that each utterance comes from 0 degrees by convoluting the transfer function of the 8 ch robot-embedded microphone array. After that, we generated two kinds of audio data whose SNR is 20 and 5 dB by adding a musical signal from 60 degrees as a noise source. Also, we generated low resolution visual data whose resolution is 1/3 compared with the original one by using a down-sampling technique. The latter data is a 20-second scenario shown in Fig. 7. Audio data was contaminated by a musical noise from $t = 0$ sec. to $t = 16$ sec as shown in Fig. 8. Visual data includes occlusion of the face and dynamic changes of face size and orientation



a) 8 ch microphone array embeded in SIG

b) Camera location

Fig. 4. Audio and vision sensors for humanoid, SIG

as shown in Fig. 7, 8. As a testbed, we used an upper-torso humanoid called *SIG* shown in Fig 4. SIG has 8 ch microphone array around its head and a camera at its right eye.

The ground truth of voice activity is shown in Fig. 9. For the ground truth, we hand-labeled input data by listening to sounds and looking at wave form. Therefore, this graph shows the voice activity, not the lip activity.

B. Results

1) **Ex. 1:** The results of **Ex. 1** are shown in Fig. 5 as word detect rate. We compared four kinds of VAD methods, that is, A-VAD, V-VAD, AV-VAD(proposed), and AV-VAD(and). AV-VAD (and) is the AV integration method described in section II-A which detects the intersection of Audio-VAD and Visual-VAD results. When SNR is low as shown in Fig. 5c), AV-VAD(proposed) and AV-VAD(and) improve VAD performance. On the other hand, when the resolution is low as shown in Fig. 5b), the performance of AV-VAD(proposed) shows high performance while AV-VAD(and) deteriorates. So, we can say that the proposed method improves robustness for both resolution and SNR changes.

2) **Ex. 2:** The results of **Ex. 2** are shown in Fig. 6 as word correct rate. When both SNR and resolution are high, AVSR improves the word correct rate. In addition, when ether of SNR and resolution is low, AVSR shows the best performance. So, we can say that two-layered AV integration is effective to improve robustness of the ASR system.

3) **Ex. 3:** The performance of Audio-VAD, Visual-VAD, and Audio-Visual-VAD are shown in Fig. 10, 11, 12, respectively. By comparing the results, we can see that only AV-VAD detects word A and B indicated in Fig. 9. Even in the case of AV-VAD, the detection result of word A and B have error. However, the margin addition described in Section III-A makes up the missing parts and almost all utterance of these two wards are detected.

VI. Conclusion and Future Work

In this paper, we proposed a two-layered AV integration framework to improve automatic speech recognition for a robot. The framework includes Audio-Visual Voice Activity Detection (AV-VAD) based on a Bayesian network and AVSR based on stream weight optimization to improve performance and robustness of ASR. We showed that the robustness improved for both auditory- and visually-contaminated input data. In addition, to solve a problem that a priori stream weight was used for so far, we proposed linear-regression-based optimal stream weight estimation. We showed the effectiveness of proposed method in terms of VAD and ASR.

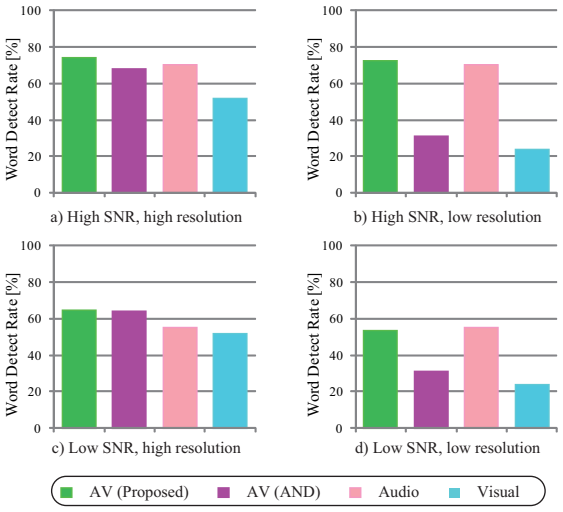


Fig. 5. The result of voice activity detection

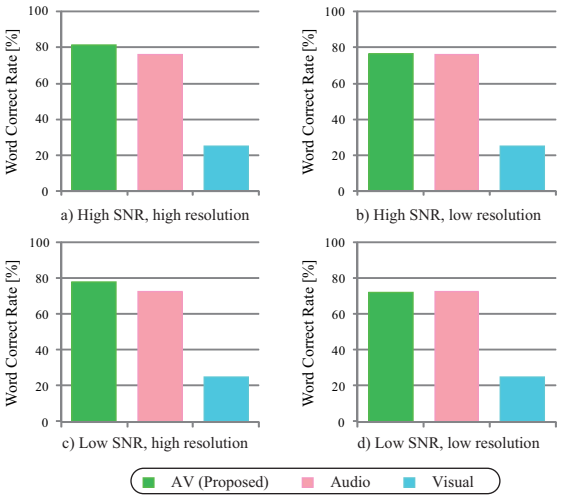


Fig. 6. The result of speech recognition

We have a lot of future work. In this paper, we evaluated the robustness for acoustical noises and face size changes, but other dynamic changes such as reverberation, illumination, and facial orientation exist in a daily environment where robots are expected to work. To cope with such dynamic changes is a challenging topic. Another challenge is to exploit the effect of robot motions actively. Since robots are able to move, they should make use of motions to recognize speeches better.

VII. Acknowledgment

We thank Dr. Rana el Kaliouby, Prof. Rosalind W. Picard for allowing us to use their system. This work is partially supported by a Grant-in-Aid for Young Scientists (B), (No. 22700165), a Grant-in-Aid for Scientific Research (S), (No.

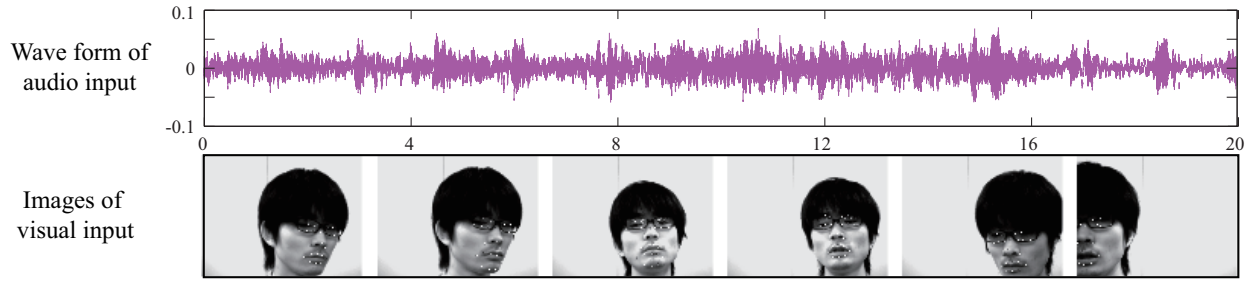


Fig. 7. Input of audio and vision sensors

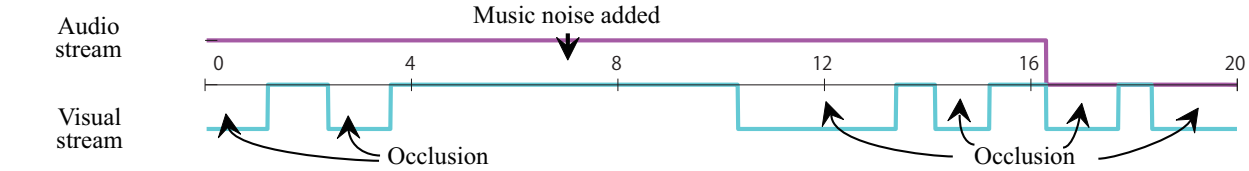


Fig. 8. Events in audio and visual stream

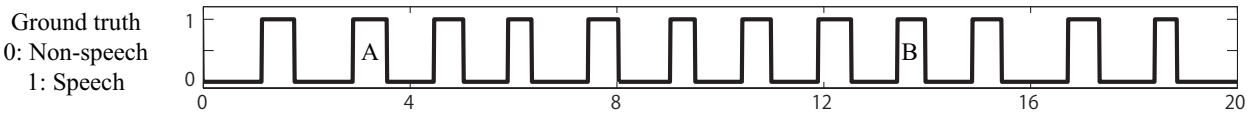


Fig. 9. Ground truth of voice activity

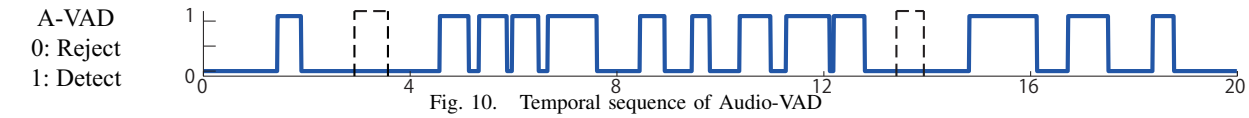


Fig. 10. Temporal sequence of Audio-VAD

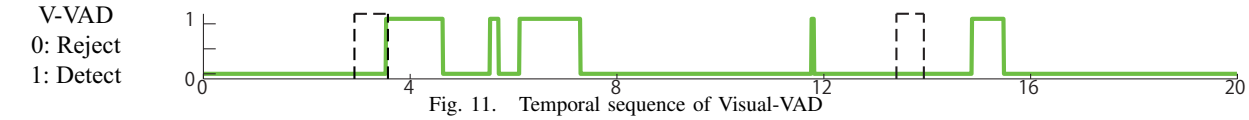


Fig. 11. Temporal sequence of Visual-VAD

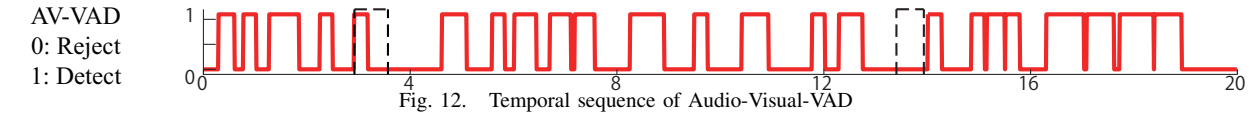


Fig. 12. Temporal sequence of Audio-Visual-VAD

19100003), and a Grant-in-Aid for Scientific Research on Innovative Areas (No. 22118502).

REFERENCES

[1] K. Nakadai, *et al.*, "Active audition for humanoid," in *Proc. of National Conference on Artificial Intelligence (AAAI)*, 2000, pp. 832–839.

[2] S. Yamamoto, *et al.*, "Real-time robot audition system that recognizes simultaneous speech in the real world," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2006, pp. 5333–5338.

[3] G. Potamianos, *et al.*, "A cascade visual front end for speaker independent automatic speechreading," *Speech Technology, Special Issue on Multimedia*, vol. 4, pp. 193–208, 2001.

[4] S. Tamura, *et al.*, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 469–472.

[5] J. Fiscus, "A post-processing systems to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1997, pp. 347–354.

[6] T. Koiwa, *et al.*, "Coarse speech recognition by audio-visual integration based on missing feature theory," in *Proc. of IEEE/RAS Int. Conf. on Intelligent Robots and Systems (IROS)*, 2007, pp. 1751–1756.

[7] T. Yoshida, *et al.*, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *Proc. of IEEE Int. Conf. on Humanoid Robots (Humanoids)*, 2009, pp. 604–609.

[8] I. Almajai, *et al.*, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2008.

[9] K. Murai *et al.*, "Face-to-talk: audio-visual speech detection for robust speech recognition in noisy environment," *IEICE Trans. Inf. & Syst.*, vol. E86-D, no. 3, pp. 505–513, 2003.

[10] G. Potamianos, *et al.*, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. 3733–3736.

[11] Y. Chow, "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1990, pp. 701–704.

[12] G. Gravier, *et al.*, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. 853–856.

[13] T. Yoshida, *et al.*, "An improvement in audio-visual voice activity detection for automatic speech recognition," in *Proc. of Int. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE)*, 2010 (To be appeared).

[14] S. Kuroiwa, *et al*, "Robust speech detection method for telephone speech recognition system," *Speech Communication*, vol. 27, pp. 135–148, 1999.

[15] F. Asano, *et al.*, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Proc. of European Conference on Speech Processing (Eurospeech)*, 2001, pp. 1013–1016.

[16] J. M. Valin, *et al*, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2004, pp. 2123–2128.

[17] Y. Nishimura, *et al.*, "Noise-robust speech recognition using multi-band spectral features," in *Proc. of Acoustical Society of America Meetings*, no. 1aSC7, 2004.

[18] Y. Nishimura, *et al*, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," in *Proc. of IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, 2006, pp. 26–33.