9th IEEE-RAS International Conference on Humanoid Robots
December 7-10, 2009 Paris, France

Automatic Speech Recognition Improved by Two-Layered Audio-Visual Integration For Robot Audition

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

Abstract— **The robustness and high performance of ASR is required for robot audition, because people usually speak to each other to communicate. This paper presents two-layered audio-visual integration to make automatic speech recognition (ASR) more robust against speaker's distance and interfering talkers or environmental noises. It consists of Audio-Visual Voice Activity Detection (AV-VAD) and Audio-Visual Speech Recognition (AVSR). The AV-VAD layer integrates several AV features based on a Bayesian network to robustly detect voice activity, or speaker's utterance duration. This is because the performance of VAD strongly affects that of ASR. The AVSR layer integrates the reliability estimation of acoustic features and that of visual features by using a missing-feature theory method. The reliability of audio features is more weighted in a clean acoustic environment, while that of visual features is more weighted in a noisy environment. This AVSR layer integration can cope with dynamically-changing environments in acoustics or vision. The proposed AV integrated ASR is implemented on HARK, our open-sourced robot audition software, with an 8 ch microphone array. Empirical results show that our system improves 9.9 and 16.7 points of ASR results with/without microphone array processing, respectively, and also improves robustness against several auditory/visual noise conditions.**

I. Introduction

In a daily environment where service/home robots are expected to communicate with humans, the robots have difficulty in automatic speech recognition (ASR) due to various kinds of noises such as other speech sources, environmental noises, room reverberations, and robots' own noises. In addition, properties of the noises are not always known in a daily environment. Therefore, 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. To realize such a robot, there are 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 and locations of noise and speech sources dynamically change, the performance of sound source separation sometimes deteriorates, and thus ASR does not always show such high performance. For 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 able to be available. Thus, their methods have difficulties in applying them to robot applications.

To solve the difficulties, we reported AVSR for robots by introducing two psychologically-inspired methods [6]. One is missing feature theory (MFT) which improves noise-robustness by using only reliable acoustic 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 has three issues as follows:

1) The system assumed that voice activity is given.
2) A single audio channel input was still used, while we have reported microphone array techniques to improve ASR performance drastically.
3) Only a closed test was performed for evaluation, that is, a test dataset for evaluation was included in a training dataset for an acoustic model in ASR.

For the first issue, we propose *Audio-Visual Voice Activity Detection (AV-VAD)*. Actually, the performance of VAD strongly affects that of ASR. We consider that VAD also improves with AV integration such as integration of audio-based activity detection and lip movement detection. We, then, integrate AV-VAD with our AVSR system, that is, a ***two-layered AV integration framework*** is used to improve speech recognition. For the second issue, we introduce HARK[1] [7]. HARK is open-sourced software for robot audition we released last year, and it provides a user-customizable total robot audition system including multi-channel sound acquisition, sound localization, separation and ASR. Thus, we integrate our AVSR with microphone-array-based sound source separation in HARK. For the last issue, we performed a word-open test to evaluate our system fairer.

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

[1]HARK stands for Honda Research Institute Japan Audition for Robots with Kyoto University, which has a meaning of "listen" in old English. It is available at http://winnie.kuis.kyoto-u.ac.jp/HARK/.

The rest of this paper is organized as follows: Section II discusses issues in audio and visual voice activity detection (AV-VAD), and Section III shows an approach for AV-VAD. Section IV describes our automatic speech recognition system for robots using two-layered AV integration, that is, AV-VAD and AVSR. Section V shows evaluation in terms of VAD and ASR performance. The last section concludes this paper.

II. Issues in Audio and Visual Voice Activity Detection For Robots

This section discusses issues in voice activity detection (Audio VAD) and lip activity detection (Visual VAD) for robots and their integration (AV-VAD), because VAD is an essential function for AVSR.

A. Audio VAD

VAD detects the start and the end points of an utterance. When the duration of the utterance is estimated shorter than the actual one, that is, the start point is detected with some delay and/or the end point is detected earlier, the beginning and the last part of the utterance is missing, and thus ASR fails. Also, an ASR system requires some silent signal parts (300-500 ms) before and after the utterance signal. When the silent parts are too long, it also affects the ASR system badly. Therefore, VAD is crucial for ASR, and thus, a lot of VAD methods have been reported so far. They are mainly classified into three approaches as follows:

A-1: The use of acoustic features,
A-2: The use of the characteristics of human voices,
A-3: The use of intermediate speech recognition results using ASR.

Common acoustic features for **A-1** are energy and zero-crossing rate (ZCR), but energy has difficulty in coping with an individual difference and a dynamic change in voice volume. ZCR is robust for such a difference/change because it is a kind of frequency-based feature. On the other hand, it is easily affected by noise, especially, when the noise has power in speech frequency ranges. Therefore, a combination of energy and ZCR is commonly used in conventional ASR systems. However, it is still prone to noise because it does not have any prior knowledge on speech signals.

For **A-2**, Kurtosis or Gaussian Mixture Model (GMM) is used. This shows high performance in VAD when it is performed in an expected environment, that is, an acoustic environment for a VAD test is identical to that for GMM training. However, when the acoustic environment changes beyond the coverage of the model, VAD easily deteriorates. In addition, to achieve noise robust VAD based on these methods, a large number of training data is required.

A-3 uses the ASR system for VAD, and thus, this is called decoder-based VAD. An ASR system basically has two stages for recognition. At the first stage, the ASR system computes log-likelihood of silence for an input signal at every frame. By using the computed log-likelihood, VAD is performed by thresholding x_{dvad} defined by

$$x_{dvad} = \log(p(\omega_0|x)) \quad (1)$$

where x is audio input, and ω_0 shows the hypothesis that x is silence.

Actually, this mechanism is already implemented on open-sourced speech recognition software called "Julius" [8]. It is reported that this approach shows quite high performance in real environments. Although this approach sounds like the chicken-or-egg dilemma, this result shows that integration of VAD and ASR is effective.

Thus, each method has unique characteristics, and none of them are suitable for all-purpose use. **A-1** is still commonly-used, **A-3** has the best performance.

B. Visual VAD

Visual VAD means lip activity detection (LAD) in visual speech recognition (VSR) which corresponds to audio VAD in ASR. The issues in visual VAD for integration with audio VAD and AVSR are as follows:

B-1: The limitation of frame rate,
B-2: The robust visual feature.

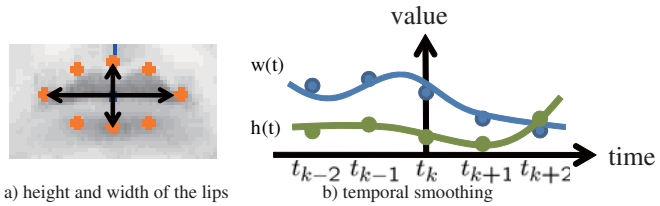
The first issue is derived from the hardware limitation of conventional cameras. The frame rate of a conventional camera is 30 Hz, while that of acoustic feature extraction in ASR is usually 100 Hz. Thus, when we integrate audio and visual features, a high speed camera having a 100 Hz capturing capability or a synchronization technique like interpolation is necessary.

For the second issue, a lot of work has been studied in the AVSR community so far. A PCA-based visual feature [9], and a visual feature based on width and length of the lips[10] were reported. However, these features are not robust enough for VAD and AVSR because visual conditions change dynamically. Especially, the change in a facial size is hard to be coped with, since the facial size is directly related to facial image resolution. Thus, an appropriate visual feature should be explored further.

C. Audio-Visual VAD

AV integration is promising to improve the robustness of VAD, and thus, audio and visual VAD should be integrated to improve AVSR performance in the real world. In this case, we have two main issues. One is AV synchronization as described above. The other is the difference between audio and visual VAD. The ground truth of visual VAD is not always the same as that of audio VAD, because extra lip motions are observed before and after an utterance to open/close the lips. AV-VAD which integrates audio and visual VAD should take their differences into account. To avoid this problem, Murai *et al.* proposed two-stage AV-VAD [11]. First, they extract lip activity based on a visual feature of inter-frame energy. Then, they extract voice activity by using speech signal power from the extracted lip activity. However, in this case, when either the first or the second stage fails, the performance of the total system deteriorates.

In robotics, AV-VAD and AVSR have not been studied well although VAD is essential to cope with noisy speech. Asano *et al.* used AV integration for speech recognition, but their AV integration was limited to sound source localization [12].



a) height and width of the lips b) temporal smoothing

Fig. 1. Visual feature extraction

Nakadai *et al.* also reported that AV integration in the level of speaker localization and identification indirectly improved ASR in our robot audition system [13]. However, in their cases, VAD was just based on signal power for a speaker direction which is estimated in AV sound source localization, that is, they indirectly used AV integration for VAD.

III. Approaches for AV-VAD

This section describes an approach for AV-VAD in our two-layered AV integration.

A. Audio VAD

For audio VAD, three approaches are described in the previous section, and the **A-3** approach has the best performance. Thus, we used decoder-based VAD as one of **A-3** approaches.

B. Visual VAD

We use a visual feature based on width and length of the lips, because this feature is applicable to extract viseme feature in the second layer of AV integration, i.e., AVSR.

To extract the visual feature, we, first, use Facial Feature Tracking SDK which is included in MindReader[2]. Using this SDK, we detect face and facial components like the lips. Because the lips are detected with its left, right, top, and bottom points, we easily compute the height and the width of the lips, and normalize them by using a face size estimated in face detection shown in Fig. 1a).

After that, we apply temporal smoothing for the consecutive five-frame height and width information by using a 3rd-order polynomial fitting function as shown in Fig. 1b). The motion of the lips is relatively slow, and the visual feature does not contain high frequency components. Such high frequency components are regarded as noise. This is why temporal smoothing is performed to remove the noise effect. Let the feature values at time frame t_i be x_{t_i}. When $S_i(t)$ is the 3rd-order polynomial function for a section $[t_i, t_{i+1}]$, the cubic spline interpolation using this function is defined by

$$\begin{aligned} S_i(t) &= a_i + b_i(t-t_i) + c_i(t-t_i)^2 + d_i(t-t_i)^3, \quad (2)\\ S(t_i) &= p_i,\\ S'_{i+1}(t_{i+1}) &= S'_i(t_{i+1}),\\ S''_{i+1}(t_{i+1}) &= S''_i(t_{i+1}),\\ S''(t_1) &= S''(t_n) = 0. \end{aligned}$$

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

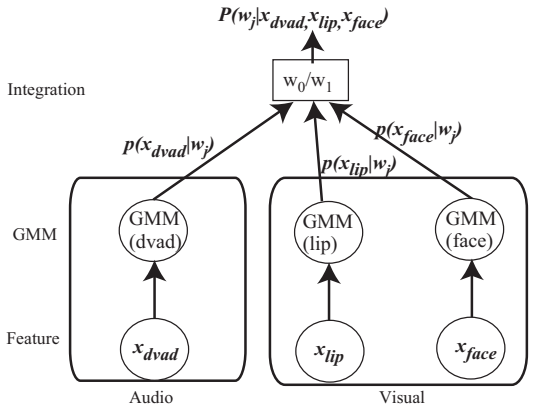


Fig. 2. AV-VAD based on a Bayesian network

Thus, we can get four coefficients such as $a_i - d_i$ for height and another four for width. In total, eight coefficients are obtained as a visual feature vector.

For the frame rate problem mentioned in Section II-B, we propose to perform up-sampling for the extracted eight coefficients so that they can easily synchronize with audio features. As a method of up-sampling, we used another cubic spline interpolation based on a 3rd-order polynomial function.

C. Audio-Visual VAD

AV-VAD integrates audio and visual features using a Bayesian network shown in Fig. 2, because the 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 Julius (x_{dvad}),
- Eight coefficients regarding the height and the width of the lips (x_{lip}),
- The belief of face detection which is estimated using Facial Feature Tracking SDK (x_{face}).

Since these features have errors more or less, the Bayesian network is an appropriate framework for AV integration in VAD.

The Bayesian network is based on the Bayes theory defined by

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}, \qquad j = 0, 1 \qquad (3)$$

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 using a 4-mixture GMM which is trained with a training dataset in advance. The probability density function $p(x)$ and probability $P(\omega_j)$ are also pre-trained with the training dataset.

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

$$\begin{aligned} &P(\omega_j|x_{dvad}, x_{lip}, x_{face}) = \\ &\quad P(\omega_j|x_{dvad})P(\omega_j|x_{lip})P(\omega_j|x_{face}). \end{aligned} \qquad (4)$$

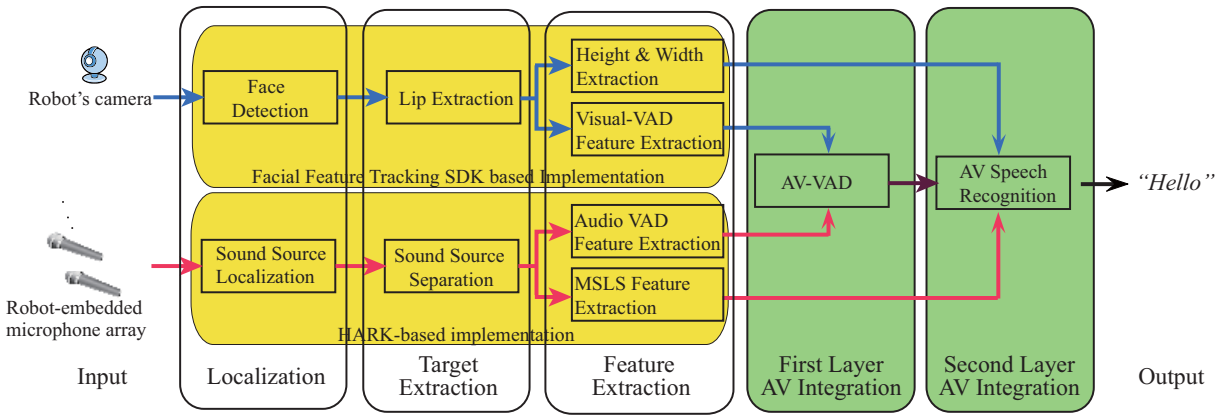


Fig. 3. An Automatic Speech Recognition System with Two-Layered AV Integration for Robots

By thresholding $P(\omega_j|x_{dvad}, x_{lip}, x_{face})$, AV-VAD decides voice activity.

IV. Automatic Speech Recognition System with Two-Layered AV Integration

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

- Facial Feature Tracking SDK based implementation for visual feature extraction,
- HARK-based implementation for microphone array processing to improve SNR and acoustic feature extraction,
- The first layer AV integration for AV-VAD,
- The second layer AV integration for AVSR.

Four modules in *Facial Feature Tracking SDK based implementation block* were already described in Section III-B, and *the first layer AV integration for AV-VAD* was also explained in Section III-C. Thus, the remaining two blocks are mainly described in this section.

A. HARK-based implementation block

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

For sound source localization, we used MUltiple SIgnal Classification (MUSIC) [14]. 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) [15]. 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 ASR systems, Mel Frequency Cepstrum Coefficient (MFCC) is commonly used. However, sound source separation produces spectral distortion in the-separated sound, and such distortion spreads over all coefficients in the case of MFCC. Since Mel Scale Logarithmic Spectrum (MSLS) [16] is an acoustic feature in a frequency domain, and thus, the distortion concentrates 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.

B. The second layer AV integration block

This block performs AVSR. We simply introduced our reported AVSR for robots [6] as mentioned in Section I, because this AVSR system showed high noise-robustness to improve speech recognition even when either audio or visual information is missing and/or contaminated by noises. This kind of high performance is derived from missing feature theory (MFT) which drastically improves noise-robustness by using only reliable acoustic and visual features by masking unreliable ones out. In this paper, this masking function is used to control audio and visual stream weights which are decided to be optimal manually in advance. For ASR implementation, MFT-based Julius [17] was used.

V. Evaluation

We performed two experiments for evaluation as follows:

Ex.1: VAD performance for acoustic noises,
Ex.2: ASR performance for acoustic noises and face size changes.

In each experiment, we used a Japanese word AV dataset. This dataset contains 10 male speech data and 266 words for each male. Audio data is sampled at 16 kHz and 16 bits, and visual data is 8 bit monochrome and 640x480 pixels in size recorded at 100 Hz using BASLER A602fc. For training an AV-VAD model, we used 216 clean AV data by 5 males in this AV dataset. For AVSR acoustic model training, we used 216 clean AV data by 10 males in this AV dataset.

The audio data is converted to 8 ch data so that each utterance comes from 0 degrees by convoluting a transfer function of the 8 ch robot-embedded microphone array. After

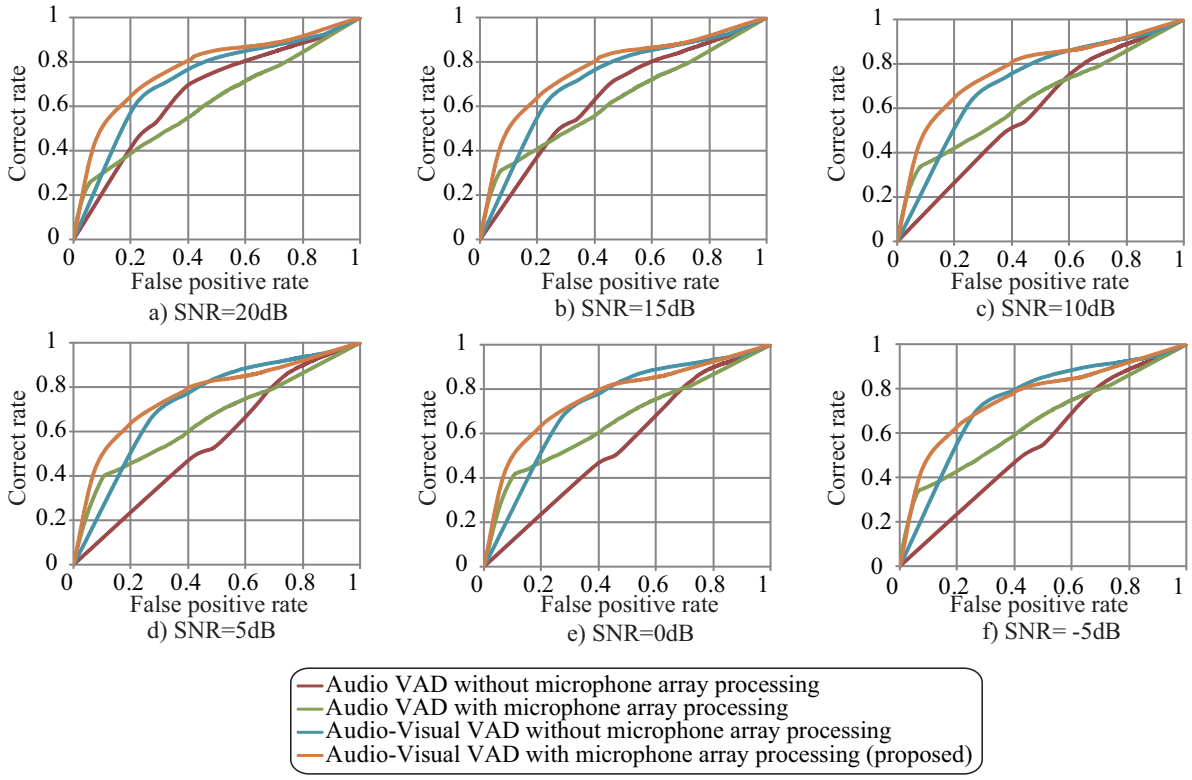


Fig. 4. Results of Voice Activity Detection

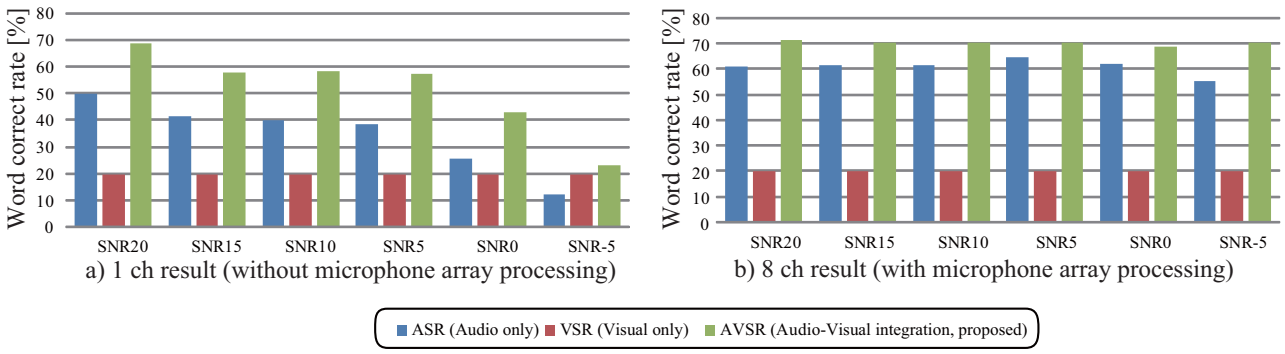


Fig. 5. The effect of AV integration in ASR

that, we added a music signal from 60° as a noise source. The SNR changed from 20 dB to -5 dB at 5 dB increments. Also, we generated visual data whose resolutions are 1/2, 1/3, 1/4, 1/5, and 1/6 compared with the original one by using a down-sampling technique. For the test dataset, another 50 AV data which are not included in the training dataset are selected from the synthesized 8 ch AV data.

In **Ex.1**, four kinds of VAD conditions were examined, that is, audio VAD/audio-visual VAD with/without microphone array processing. For ground truth, the result of visual VAD is used when the resolution of face images is high enough.

In **Ex.2**, performance of ASR, VSR and AVSR was compared through isolated word recognition.

Fig.4 shows VAD results in various conditions using ROC curves. Audio VAD got worse when SNR was low. Our microphone array processing improved VAD performance because it improves SNR. Audio-Visual VAD drastically improved VAD performance. This shows the effectiveness of AV integration in the VAD layer. In addition, the combination of Audio-Visual VAD and microphone array processing, that is, our proposed method improves VAD performance more. This indicates that information integration is a key idea to improve robustness and performance when we cope with real-world data.

Fig.5 shows speech recognition results. The performance of AVSR was better than that of ASR or VSR. Although word-open tests were performed, the word correct rates reached around 70% with our proposed method. The effect of AV integration was 16.7 points when we used a single channel audio input. When we used microphone array processing, it improved ASR performance, but the effect of AV

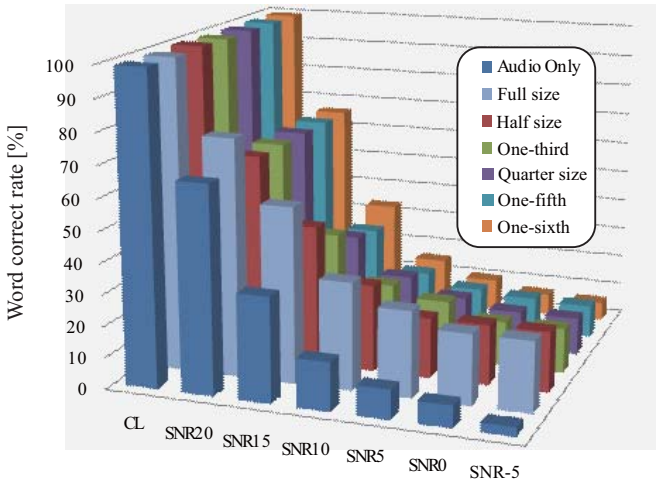


Fig. 6. The robustness for face size changes

integration was still 9.8 points.

Fig.6 shows the robustness for face size changes in ASR performance. Even when face resolution was 1/6 compared with the original resolution, AV integration sometimes improved ASR performance, especially in lower SNR cases. When face resolution and SNR were low, the performance dropped. In this case, a robot should detect that the current situation is not good for recognition, and should take another action such as approaching the target speech source.

VI. Conclusion

We proposed a two-layered AV integration framework which consists of Audio-Visual Voice Activity Detection (AV-VAD) based on a Bayesian network and Audio-Visual Speech Recognition (AVSR) using a missing feature theory to improve performance and robustness of automatic speech recognition (ASR). We implemented an ASR system with the proposed two-layered AV integration framework on HARK, which is our open-sourced robot audition software. Thus, the AV integrated ASR system was integrated with microphone array processing such as sound source localization and separation included in HARK to improve SNR of input speech signals. The total ASR system was evaluated through word-open tests. We showed that 1) our proposed AV integration framework is effective, that is, a combination of AV-VAD and AVSR showed high robustness for input speech noises and facial size changes, 2) microphone array processing improved ASR performance by improving SNR of input speech signals, and 3) a combination of two-layered AV integration and microphone array processing further improved noise-robustness and ASR performance.

We still have a lot of future work. In this paper, we evaluate 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 speech better.

VII. Acknowledgments

We thank Prof. R. W. Picard and Dr. R. E. Kaliouby, MIT for allowing us to use their system. We thank Prof. J. Imura and Dr. T. Hayakawa, Tokyo tech. for their valuable discussions. This research was partially supported by "Binaural Active Audition for Humanoid Robots (BINAAHR)" project, strategic Japanese-French cooperative program.

References

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of 17th National Conference on Artificial Intelligence (AAAI)*, pp. 832–839, 2000.

[2] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J.-M. Valin, K. Komatani, T. Ogata, and H. G. Okuno, "Real-time robot audition system that recognizes simultaneous speech in the real world," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5333–5338, 2006.

[3] G. Potamianos, C. Neti, G. Iyengar, A. Senior, and A. Verma, "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, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream hmms based on likelihood value normalization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), SP-P5.2*, 2005.

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

[6] T. Koiwa, K. Nakadai, and J. Imura, "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)*. pp. 1751–1756, 2007.

[7] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition HARK and its evaluation," in *Proc. of IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. pp. 561–566, 2008.

[8] "http://julius.sourceforge.jp/."

[9] P. Liu and Z. Wang, "Voice activity detection using visual information," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 609–612, 2004.

[10] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, no. 7-8, pp. 667–677, 2007.

[11] K. Murai and S. Nakamura, "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.

[12] F. Asano, Y.Motomura and S. Nakamura, "Fusion of audio and video information for detecting speech events," in *Proc. International Conference on Information Fusion*, pp. 386–393, 2003.

[13] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots," *Speech Communication*, vol. 44, pp. 97–112, 2004.

[14] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition." in *Proc. of International Conference on Speech Processing (Eurospeech)*. pp. 1013–1016, Sep. 2001.

[15] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 2123–2128, 2004.

[16] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," in *Proc. of 148th Acoustical Society of America Meetings*, no. 1aSC7, 2004.

[17] Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, and H. Tsujino, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," in *Proc. of 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. pp. 26–33, 2006.