



Two-Layered Audio-Visual Integration in Voice Activity Detection and Automatic Speech Recognition for Robots

Takami Yoshida¹ and Kazuhiro Nakadai^{1,2}

¹Graduate School of Information Science and Engineering, Tokyo Institute of Technology

²Honda Research Institute Japan co., ltd.

yoshida@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

Automatic Speech Recognition (ASR) which plays an important role in human-robot interaction should be noise-robust because robots are expected to work in noisy environments. Audio-Visual (AV) integration is one of the key ideas to improve the robustness in such environments. This paper proposes two-layered AV integration for ASR which applies AV integration to Voice Activity Detection (VAD) and ASR decoding process. We implemented a prototype ASR system based on the proposed two-layered AV integration and evaluated the system in dynamically-changing situations where audio and/or visual information is noisy or missing. Preliminary results showed that the proposed method improves the robustness of ASR system even in auditory- or visually-contaminated situations.

Index Terms: audio-visual integration, speech recognition, voice activity detection

1. Introduction

Noise-robust Automatic Speech Recognition (ASR) is essential for home/service robots which are expected to communicate with humans in daily environments. In such an environment, a robot has difficulties in ASR due to various kinds of noises such as environmental noise, other speech sources, and robot's generating ego-noise. In addition, noise properties may change dynamically because the robot and sound sources move. Thus, ASR robust for dynamically-changing noise is necessary for robots and other hands-free applications.

To realize noise-robust ASR, Audio-Visual (AV) integration is crucial. Many studies on AV-integrated ASR (AV-ASR) have been reported. However, most of them considered AV integration only in an ASR decoding process, that is, AV-integrated decoder (AV-DEC)[1, 2]. Indeed, they showed high speech recognition performances in accuracy and robustness for acoustic noises, but they assumed voice activities were given in advance although the performance of Voice Activity Detection (VAD) strongly affects that of ASR system. Only a few AV-ASR studies reported a combination of audio-based VAD (A-VAD) and AV-DEC [3]. We can also find AV-ASR studies on another combination of AV-integrated VAD (AV-VAD) and a normal audio-based ASR decoder (A-DEC) [4]. In any case, these studies have issues as follows:

- 1) The use of AV integration is limited either in VAD or in the decoding process although both should be noise-robust,
- 2) The VAD and ASR performances were evaluated by using ideal image data which is not always available for robots.

In this paper, for 1), we propose two-layered AV integration which applies AV integration to VAD and ASR decod-

ing process, and for 2), we conduct evaluation using visually-contaminated data.

2. Issues in Audio-Visual Integration for ASR

AV integration is promising to improve the robustness of VAD and ASR, and thus, AV integration should be applied to both VAD and ASR. AV-ASR for robots should cope with a crucial problem that audio- and/or visual-features are not always available. An audio feature can be contaminated with noises coming from other speakers and robot's actuators. A visual feature can be missing due to occlusion and the change of facial direction, and/or can be damaged due to low resolution images for a distant speaker.

2.1. Audio-Visual Integration for VAD

In VAD, AV integration methods are mainly classified into two approaches. One is feature-level integration called Early Integration (EI), and the other is decision-level integration called Late Integration(LI). Almajai *et al.* reported EI-based AV-VAD using Mel-Frequency Cepstral Coefficient (MFCC) and a visual feature based on the 2-D discrete cosine transform [5]. The integrated AV feature showed high performance for VAD. However, they assumed that the high resolution images of the lips are always available. Murai *et al.* presented LI-based AV-VAD [4]. They detected a lip activity by using a visual feature based on a temporal sequence of lip shapes and a voice activity based on speech signal power. AV-VAD was performed by extracting the intersection of these detected activities. However when either lip or voice activity is mis-detected, the total system performance easily deteriorates.

2.2. Audio-Visual Integration for ASR

We use EI-based AV-DEC because LI-based AV-DEC like ROVER [1] requires expensive computational cost caused by adjusting alignments between A-DEC and V-DEC recognition results.

The issue in EI-based AV-DEC is how to control the balance of audio and visual features. When the audio feature is reliable and visual feature is not, AV-ASR should put more weight on the audio feature and less on the visual feature, and vice versa. We use a stream weight to cope with this issue.

A lot of stream weight optimization methods have been studied. They mainly used log likelihoods in audio and/or visual speech models. For instance, optimization methods based on likelihood value normalization [2] and maximum entropy criterion [6] have been proposed. These methods mainly dealt with acoustic noise using only high resolution images. Thus, it is difficult to apply these methods to an ASR system for a robot,

10.21437/Interspeech.2010-716

because resolution and face orientation are dynamically changing. In particular, dealing with resolution changes is a crucial factor for AV-DEC because low resolution images are still considered as informative while most the performance of AV-ASR drops with low resolution images [7].

3. Approaches in Audio-Visual Integration for ASR

3.1. Audio-Visual Integration for VAD

To solve the issue in AV-VAD, we introduce AV-VAD based on a Bayesian network [7], because it provides a framework that integrates multiple features with some ambiguities by maximizing the likelihood of the total integrated system. Asano, *et al.* proposed AV-VAD based on a Bayesian network [8]. Their system integrated sound source localization and visual human tracking and did not use lip movements while they are efficient. We use the following features as the inputs to the Bayesian network:

- The score of log-likelihood for silence calculated by ASR 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 process (x_{face}).

x_{dvad} is calculated by using an acoustic model of speech recognition, and thus, it takes the property of voice into account. This feature shows high noise-robustness as reported in [9]. x_{lip} is derived from the temporal sequence of the height and width information by using linear regression [7]. x_{face} is calculated based on Gabor wavelet and elastic graph matching [10] in a 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

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

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 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 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 thresholding this probability, 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 [11]. In the dilation process, a frame is added to the start- and end-points of voice activity as below.

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

where $V[k] = \{0(\text{non-speech}), 1(\text{speech})\}$ is the estimated voice activity at time frame k and $\hat{V}_d[k]$ is the result of dilation. This process removes false rejects. In erosion process, a

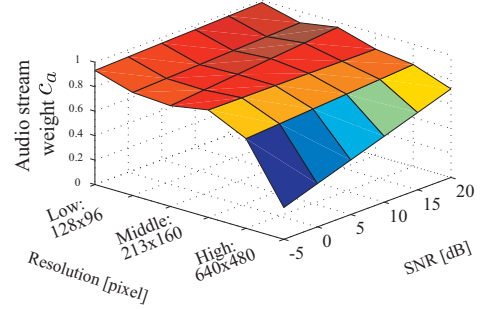


Figure 1: The linear regressions of optimal stream weight frame is removed from the start- and end-points of voice activity as below.

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

where $\hat{V}_e[k]$ is the result of erosion. This erosion process removes false detection of voice activity. AV-VAD performs these processes several times and refines voice activity.

Finally, temporal margins are added just before and after the refined voice activity period to avoid a mismatch problem between a VAD process and a HMM-based speech model in ASR, because ASR decoder assumes that each utterance begins and finishes with silence.

3.2. Audio-Visual Integration for ASR

To cope with the issue in Section 2.2, we introduced stream weight optimization based on Signal-to-Noise Ratio (SNR) and face size corresponding to the lip image resolution. We first evaluated AV-ASR by changing the audio stream weight from 0 to 1 at 0.1 intervals for several audio and visual noise conditions. From a word correct rate of this test, we decided optimal stream weights for every SNR and image resolution. The estimated audio stream weight was calculated from linear regression of optimal audio stream weights. Figure 1 shows the linear regressions obtained by optimal stream weights. We used six functions to estimate optimal stream weights. The log likelihood b_w of AV feature x_{AV} for a word w at time frame t is calculated using audio stream weight c_a ($0 \leq c_a \leq 1$) by

$$b_w(x_{AV}(t)) = c_a b_{Aw}(x_A(t)) + (1 - c_a) b_{Vw}(x_V(t)) \quad (5)$$

$b_{Aw}(x_A(t))$ and $b_{Vw}(x_V(t))$ are likelihoods for an audio and visual features respectively.

4. ASR System Based on Two-Layered AV Integration

Figure 2 shows our ASR system for robots based on two-layered AV integration. It consists of four blocks as follows:

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

In the following sections, we describe two of these four blocks because AV-VAD and AV-DEC are already described in the previous sections. In terms of implementation, we used a Bayesian networks library called OpenPNL¹ for AV-VAD, and Multiband Julius[12] was used for AV-DEC.

¹<http://sourceforge.net/projects/openpnl>

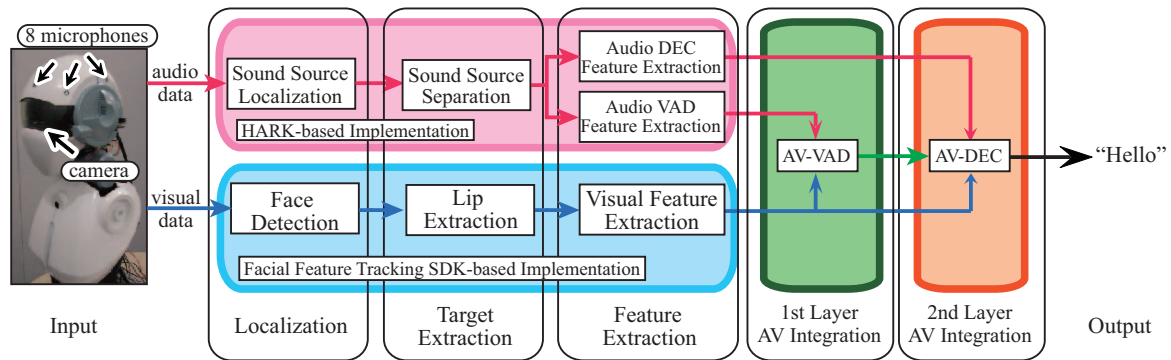


Figure 2: An automatic speech recognition system with two-layered AV integration for robots

4.1. 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 MindReader². Using this SDK, we detected face and facial components like the lips. Because the face and lip are detected with its left, right, top, and bottom points, we can easily compute the height and the width of the faces and lips, and normalize the height and width of the lips by using the face size estimated in face detection. Then, we apply third order polynomial function fitting for temporal sequence of height and width information. We obtain 4 coefficients from each fitting and use these 8 coefficients as visual feature both in VAD and in ASR decoder.

4.2. Audio Feature Extraction Block

This block consists of five modules, that is, sound source localization, sound source separation, audio VAD feature extraction, and audio ASR feature extraction. Their implementation is based on HARK [12]. The audio VAD feature extraction was already explained in Section 2, and thus, the other three modules are described.

We used an 8 channel circular microphone array which is embedded around the top of our robot’s head. For sound source localization, we used Multiple Signal Classification. For sound source separation, we used Geometric Sound Separation (GSS). 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 those of sound sources obtained from sound source localization.

For an acoustic feature for ASR, MFCC is commonly used. However, sound source separation produces spectral distortion in separated sounds, and such distortion spreads over all coefficients in the case of MFCC. Since Mel Scale Logarithmic Spectrum (MSLS) is an acoustic feature in the frequency domain, 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 Δ MSLS, and Δ log power.

5. Evaluation

5.1. Experiments and Evaluation

We evaluated the system through two experiments.

Ex.1: The effectiveness of AV-integration for VAD.

²<http://mindreader.devjavu.com/wiki>

Ex.2: The effectiveness of two-layered AV-integration for ASR.

For **Ex. 1** and **Ex. 2**, we recorded a Japanese word AV dataset. This AV dataset contains speech data from 10 males and 266 words (216 ATR phonemically-balanced words and 50 other 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 and an AV-DEC model, we used 216 clean AV data from 10 males.

We used two kinds of test datasets. One is 50 AV data which is not included in the training dataset. The other is AV data captured by actual robot shown in Fig. 2. For the former data, the audio data was converted to 8 channel data so that each utterance comes from 0 degrees by convoluting the transfer function of the 8 channel robot-embedded microphone array. After that, we generated audio data whose SNR is from 15 to -5 dB at every 5 dB by adding a music signal from 60 degree as a noise source. To evaluate the effectiveness of microphone array processing mentioned in Section 4, we applied microphone array processing when the SNR is the worst (-5 dB). Also, we generated low resolution visual data whose resolution is one-third (middle) and one-fifth (low) compared with the original one (high) by using a down-sampling technique. The latter data is a 24-second AV data including 15 words shown in Fig. 3. Audio data is contaminated by a music and noise coming from robot’s power source, and visual data includes occlusion of the face and dynamic changes of face size and orientation.

For the ground truth of voice activity, we labeled input data by listening to sounds and looking at waveforms.

In ASR decoding process, we conducted isolated word recognition using a monophone HMM which has three states and 32 Gaussian mixtures in each state.

5.2. Results

Table 1 shows Word Detection Rates (WDRs) for **Ex. 1**. We assume that VAD succeeds when both start- and end-point errors are less than 200 ms, and WDR is defined as the ratio of the number of successfully-detected words to that of all words.

When SNR is from 10 to -5 dB and image resolution is middle or low, AV-VAD shows higher performances than the others, and the improvement in these cases was, on average, 11.5 points. So, We can say that the proposed method improves the robustness for both resolution and SNR changes.

However, in some cases, AV integration worsens the performance of VAD. This is because the Bayesian network was trained by using a clean AV dataset. Multi-condition training would be effective to cope with this problem.

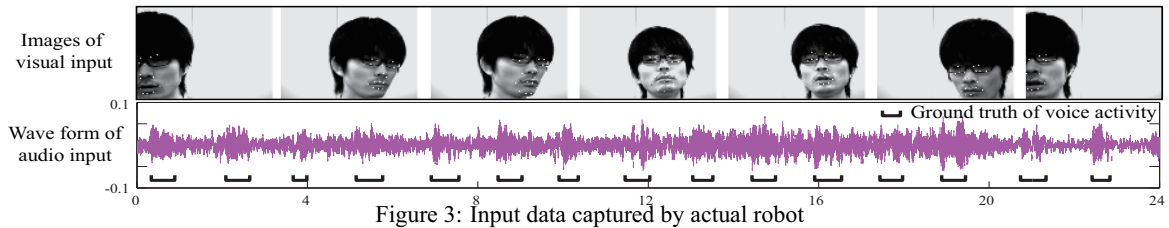


Figure 3: Input data captured by actual robot

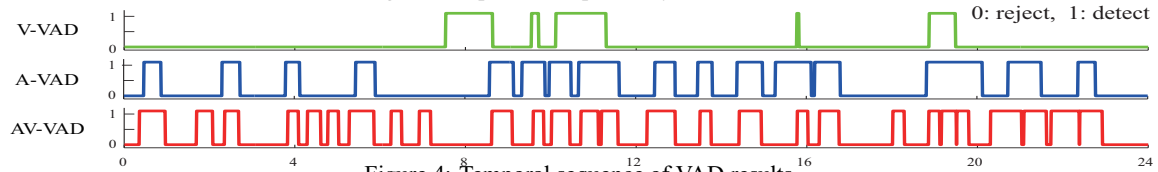


Figure 4: Temporal sequence of VAD results

Table 1: Word detection rates (%) in VAD.

visual condition	modal-ity	SNR [dB]					
		15	10	5	0	-5	-5*
-	A	58.8	35.6	16.8	16.8	12.4	62.4
high resolution	AV	64.4	55.2	50.8	46.0	44.4	74.0
	V	52.0	52.0	52.0	52.0	52.0	52.0
middle resolution	AV	55.6	49.6	42.0	42.8	42.0	70.0
	V	36.0	36.0	36.0	36.0	36.0	36.0
low resolution	AV	49.6	45.6	38.0	41.2	42.4	68.0
	V	24.0	24.0	24.0	24.0	24.0	24.0

* VAD with microphone array processing

Table 2: Word correct rates (%) in ASR.

visual condition	modal-ity	SNR [dB]					
		15	10	5	0	-5	-5*
-	A	75.6	68.5	60.9	42.4	29.9	72.4
high resolution	AV	84.4	76.0	68.0	52.8	36.8	74.8
	V	25.6	25.6	25.6	25.6	25.6	25.6
middle resolution	AV	78.4	72.8	63.2	46.0	30.0	75.6
	V	24.4	24.4	24.4	24.4	24.4	24.4
low resolution	AV	73.6	68.8	62.4	45.6	29.6	71.2
	V	24.0	24.0	24.0	24.0	24.0	24.0

* ASR with microphone array processing

Figure 4 shows a VAD performance using an actual robot. We can see that the proposed method worked robustly even when V-VAD had a poor performance.

Table 2 shows the results of Ex. 2 in Word Correct Rates (WCRs). When SNR is 10, 5, or 0 dB, the AV-ASR performances are better than Audio- or Visual-ASR. Even when SNR is 15 dB, AV integration improves WCR if high or middle resolution images are available.

By comparing Tables 1 and 2, WCR is higher than WDR, that is, the ASR decoder can recognize incorrectly-detected words properly in some cases. This is the effect of margin addition described in Section 3.2. When a voice activity is estimated shorter than actual one, margin addition recovers missing parts, and ASR decoder succeeds.

6. Conclusion and Future Work

In this paper, we proposed two-layered AV integration to improve the robustness of ASR for a robot. This framework includes AV-VAD based on a Bayesian network and AV-DEC based on stream weight optimization. We implemented a prototype system based on the proposed method and evaluated the performance of the whole ASR system in both auditory- and visually-contaminated situations. Experimental results show that two-layered AV integration improves the robustness of ASR system.

Future work is to cope with dynamic changes of facial orientation, sound source location, and the number of sound sources. Another is stream weight optimization without SNR

estimation because SNR estimation is difficult.

7. Acknowledgment

We thank Dr. R. Kaliouby and 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. 19100003), and a Grant-in-Aid for Scientific Research on Innovative Areas (No. 22118502).

8. References

- [1] J. Fiscus, "A post-processing systems to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of ASRU*, 1997, pp. 347–354.
- [2] S. Tamura, *et al.*, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. of ICASSP*, 2005, pp. 469–472.
- [3] G. Potamianos, *et al.*, "Far-field multimodal speech processing and conversational interaction in smart spaces," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 119–123.
- [4] 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.
- [5] I. Almajai, *et al.*, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. of EUSIPCO*, 2008.
- [6] G. Gravier, *et al.*, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. of ICASSP*, 2002, pp. 853–856.
- [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*, 2009, pp. 604–609.
- [8] F. Asano, *et al.*, "Fusion of audio and video information for detecting speech events," in *Proc. of the Int. Conf. of Information Fusion*, 2003, pp. 386–393.
- [9] S. Kuroiwa, *et al.*, "Robust speech detection method for telephone speech recognition system," *Speech Communication*, vol. 27, pp. 135–148, 1999.
- [10] K. Okada, *et al.*, "The Bochum/USC face recognition system and how it fared in the FERET phase III test," *NATO ASI series. Series F: computer and system sciences*, 1998.
- [11] K. W. Bowyer, *et al.*, "Image understanding for iris biometrics: A survey," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 281 – 307, 2008.
- [12] K. Nakadai, *et al.*, "Design and implementation of robot audition system 'HARK' –open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.