Active Audio-Visual Integration for Voice Activity Detection based on a Causal Bayesian Network

Takami Yoshida* and Kazuhiro Nakadai†
* Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8552, Japan
† Honda Research Institute Japan Co., Ltd., 351-0114, Saitama, Japan
Email: yoshida@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract—This paper addresses an active audio-visual integration framework which integrates auditory and visual information with a robot's active motion for noise-robust Voice Activity Detection (VAD). VAD is crucial for noise robust Automatic Speech Recognition (ASR) because speech captured by a robot's microphones is usually contaminated with other noise sources. To realize such noise-robust VAD, we propose Active Audio-Visual (AAV) integration framework which integrates auditory, visual and motion information using a Causal Bayesian Network (CBN). CBN is a subclass of Bayesian networks, which is able to estimate the effect on VAD performance caused by active motions. Since CBN is a general framework for information integration, we can naturally introduce various types of information such as the location of a speaker and a noise source which affect VAD performance to CBN, and CBN selects the optimal active motion for better perception of the robot using “intervention” mechanism in CBN. We implemented a prototype system based on the proposed framework on a humanoid robot called Hearbo. The proposed AAV-VAD is compared with three types of AV-VAD; simple AAV-VAD, multi-regression-based AAV-VAD, and stationary (not active) AV-VAD. A preliminary experiment using the prototype system showed that the VAD performance of the proposed AAV-VAD was 14.4, 26.0, and 30.3 points higher than that of the simple active, multi-regression-based active, and stationary AV-VAD, respectively.

I. INTRODUCTION

Human-robot interaction by speech is crucial when a humanoid robot works together with humans, because humans use speech to communicate with each other. Automatic Speech Recognition (ASR) is important in realizing these interactions. In general, ASR consists of two main processes: Voice Activity Detection (VAD) and speech DECodeing (DEC). VAD extracts speech from an input signal and DEC recognizes the extracted speech. To perform ASR correctly, both VAD and DEC should work correctly. However, VAD and DEC performances deteriorate in a noisy environment.

Two approaches may improve the noise-robustness of ASR for a humanoid robot:
1) **Audio-Visual integration**, and
2) **Active audition**.

Audio-Visual (AV) integration improves the noise-robustness of VAD and DEC. Koiwa et al. proposed AV-DEC based on two psychologically-inspired approaches, the missing feature theory, which utilizes only reliable features, and coarse-to-fine recognition, which changes recognition units dynamically [1]. This AV-DEC showed high speech recognition performance even when either audio or visual cues were not available. We proposed two-layered AV integration, which applies AV integration to both VAD and DEC [2], [3]. While this framework improved VAD and DEC when the quality of audio and visual cues was high, their performance dropped when the quality of audio or visual cues was low.

Active audition improves auditory processing using active motion. In the robot audition community, active audition can be applied to, for example, Sound Source Localization (SSL) [4], [5], [6], [7], [8], [9].

Most of the studies mainly utilize the turning motion [4], [5], [6], [7]. Nakadai et al. proposed active audition for SSL, which integrates audition, vision and motor control [4]. Their system controlled the waist of a humanoid robot to align a pair of microphones orthogonal to a sound source. Reid and Milios described an active SSL system that utilizes pan-and-tilt motion [5], generated according to the previous time delay measurement. Berglund and Sitte also proposed active audition for SSL [6], by orienting two microphones towards a sound source using reinforcement learning. Kim et al. proposed an active audition system, consisting of SSL, VAD, face tracking, and sound source tracking, using turning motion [7]. Although these approaches showed high SSL performances, they controlled only the direction of the microphones.

Some researches utilize locomotion [8], [9]. Sasaki et al proposed 2D sound source mapping based on triangulation [8]. They showed the effectiveness of active motion for SSL. However, they mainly focused on the SSL algorithm and did not deal with path planning. Martinson and Brock proposed text-to-speech interface for a robot based on active audition, which tries to go far away from noise sources [9]. This approach is effective to reduce noise levels. However, their approach was aimed at a text-to-speech application. In

![Fig. 1. The proposed active AV integration and conventional stationary robot audition](image-url)
addition, they only considered the effect of acoustic noises. Therefore, this approach has difficulties in applying to Active AV (AAV) integration for VAD.

To achieve more sophisticated active audition than conventional approaches, a robot should select an optimal motion from various types of motions. For example, Cooke et al. [10] described three types of active human motions;

- Fine head movements to disambiguate front-back confusion,
- Gross head movements to improve Signal-to-Noise Ratio (SNR) at the best ear, to use head shadow to reduce signals from interferes, and to locate a target in the high resolution part of the azimuthal plane, and
- Body translation to improve target signaling, to reduce the levels of interference and reverberation, and to increase spatial separation between target and interferer.

In addition, if a robot can perform multiple active motions at the same time, the scalability of AAV integration framework is very important.

We sought to provide a framework enabling a robot to determine the optimal active motion from various types of active motions. We called this framework AAV integration, as explained in the following sections.

II. ISSUES IN ACTIVE AUDITION FOR A HUMANOID ROBOT

In considering the application of AAV integration to a humanoid robot, two issues arose:

1) Estimating the effect on VAD performance of a robot’s active motion, and
2) Scalability to represent various types of a robot’s active motions.

A robot should perform the active motion expected to yield the greatest improvement in VAD performance. Therefore, before a robot performs active motions, it must estimate their effect. Indirect estimation of VAD performance is especially necessary, because direct estimation is difficult in a daily environment. For this purpose, we can use SNR, which strongly affects VAD performance (for example, see [11]). However, SNR calculations are based on the powers of speech and non-speech parts, obtained as the output of VAD.

Scalability is essential when a humanoid robot can perform various types of active motions. If we consider only one active motion like conventional approaches [4], [5], [6], [7], we can easily describe the effects of active motion on VAD performance. However, such a system is difficult to apply to a robot that can perform multiple active motions.

In estimating the effects of an active motion on VAD performance, probabilistic models are more feasible than deterministic models because the latter are prone to noises. In a probabilistic model, probabilistic variables for observations and active motions must be separated because the effects of an observation are not always the same as those of an active motion. Augmented Probabilistic Models (APMs) can be used to deal with both observations and active motions.

APMs are extended probabilistic models that can cope with probabilistic variables for an active motion. Therefore, in APMs, additional nodes are introduced to represent active motion. These APMs, however, have difficulties utilizing multiple active motions due to an exponential increase in the number of additional probabilistic variables.

III. AAV INTEGRATION FRAMEWORK BASED ON A CAUSAL BAYESIAN NETWORK

In this section, we describe AAV integration framework, which integrates a robot’s active motion, audio and visual information using a Causal Bayesian Network (CBN) [12].

A. Causal Bayesian Network

CBNs are a subclass of Bayesian networks that satisfy two assumptions: its construction is based on causal knowledge and the system can change one relationship without changing others.

Fig. 2a) shows an example of a CBN with 10 nodes and 11 edges. A node represents a probabilistic variable and an edge represents a causal relationship from a parent to a child. Fig. 2b) shows an example of an APM that corresponds to Fig. 2a). Both models can deal with two kinds of active motions for $v_4, v_5,$ and $v_6$ at the same time. The APM introduces an additional probabilistic variable $F$ for active motion (e.g. $F_i = 1, i = 0, 1$ means “intervene” and $F_i = 0, i = 0, 1$ means “idle”). Thus, the number of probability distributions increases exponentially as the number of active motions becomes large. Actually, the APM in Fig. 2 uses two additional nodes ($F_i, i = \{0, 1\}$) to deal with active motions. For example, since $v_6$ is connected to both $F_0$ and $F_1$, four probability distributions $(P(V_0|F_0, F_1), P(F_0), P(F_1) = (0, 0), (0, 1), (1, 0), (1, 1))$ should be calculated in advance. APMs, thus, have difficulties in coping with AAV integration.

Apart from APMs, CBNs provide a dynamic method, named do-calculus [12], for calculating the effects of active motions. In a CBN, an increase in the number of probability distributions is limited to be in linear order because of do-calculus. An active motion with a probabilistic variable setting $V_i = s_i$ is described as $do(s_i)$. 371
B. A CBN for AAV-VA

In our AAV integration framework, a robot’s active motion is regarded as “intervention,” and AV information is regarded as “observation.”

The effect of active motion \( \text{do}(s), s = [s_1, \ldots, s_n] \) on the target variables \( y = [y_1, \ldots, y_n] \) can be calculated by the following truncated factorization.

\[
P(y|x, \text{do}(s)) = \prod_i P(y_i|p_a(y_i)) \prod_i P(x_i|p_a(x_i))
\]

if \( s \) consistent with \( \text{do}(s) \),
\[
= 0, \text{ otherwise}
\]

where \( x = [x_1, \ldots, x_n] \) are intermediate variables. Note that we can apply the do-calculus described in Eq. (1) to any CBN, as described in details in [12].

The effect on \( Y = V_{11} \) caused by an active motion \( \text{do}(s), S = \{V_s\} \) in Fig. 2a can be calculated as

\[
P(y|x, \text{do}(s)) = P(v_{12}|v_9)P(v_{11}|v_7, v_9)P(v_{10}|v_7)P(v_9|v_6, v_8)P(v_8)
\]

\[
\times P(v_7|v_6, \ldots, v_5)P(v_5|v_4)P(v_3, v_2)P(v_1, v_0).
\]

The optimal active motions \( s^* \) can be selected as

\[
s^* = \arg \max_s P(y_1, \ldots, y_n|x_1, \ldots, x_n, \text{do}(s)).
\]

IV. SYSTEM IMPLEMENTATION

Fig. 3 shows the system architecture of the proposed AAV integration framework.

A. Hardware

We used the humanoid robot “Hearbo” shown in Fig. 3 as a testbed. Hearbo’s lower body is an omnidirectional-cart, with its upper body placed on the cart.

1) Sensors: We used a camera in the position of its left eye and an 8-channel circular microphone array embedded around the top of the head. Audio signals are captured at 24 bits and 16 kHz sampling. Visual images are captured in 8 bit grayscale, 640×480 pixels, and 30 Hz. The head’s joint and cart’s wheels are equipped with encoders. The encoder values of the head and cart are observed at 100 Hz and 30 Hz, respectively.

2) Actuator: We controlled the head’s pitch \( \psi_{\text{pitch}} \) and yaw \( \psi_{\text{yaw}} \), and the cart’s position \( \xi_x, \xi_y \) and direction \( \xi_{\theta} \). The cart has four wheels, each equipped with steering and driving motors.

B. Software

The prototype system consists of four main blocks: Audio feature extraction, Visual feature extraction, Robot controller, and AV-VA

1) Audio Feature Extraction: In this block, the system performs microphone array processing such as Generalized-EigenValue-Decomposition-based Multiple Signal Classification (GEVD-MUSIC) for SSL and Geometric High-order Diсorrelation-based Source Separation (GHDSS) for Sound Source Separation (SSS). Subsequently, a Mel-Scale Log Spectrum (MSLS) is extracted from the separated sounds. These methods are available in HARK [13].

SSL results are sent to the robot controller to be integrated with active motion. GEVD-MUSIC produces a so-called spatial spectrum \( I_{ss}(\phi) \) in every time frame. When \( I_{ss}(\phi) \) has a high value, a sound source is likely to be present in direction \( \phi \). In our implementation, \( I_{ss} \) was obtained every 5 degrees in the azimuth plane \( \{\phi \in \{-180, -175, \ldots, 175\}\} \). Then we extracted the local peaks of the spatial spectrum as

\[
I_{\text{pass}}(\phi) = \begin{cases} 
I_{ss}(\phi) & \text{if } I_{ss}(\phi) \geq I_{ss}(\phi - 5) \\
\cup I_{ss}(\phi) \geq I_{ss}(\phi + 5) & \text{otherwise}.
\end{cases}
\]

We converted the coordinates of each spectrum to a 2D plane using the robot’s position.

\[
\phi(x, y, \xi_x, \xi_y, t) = \arctan \left( \frac{y - \xi_y(t)}{x - \xi_x(t)} \right),
\]

where \( (x, y) \), \( (\xi_x, \xi_y) \) and \( t \) are the location of a sound in a room, the location of the robot, and time, respectively. A 2D sound map can be obtained by determining the temporal average over \( T \) frames.

\[
\pi(x, y, \xi_x, \xi_y, t) = \sum_{t=t-T}^t I_{\text{pass}}(\phi(x, y, \xi_x, \xi_y, t)).
\]

A sound source location is extracted by \( (x, y) = \arg \max \pi(x, y, \xi_x, \xi_y, t) \). Fig. 4 shows an example of a 2D
sound map. A fixed sound source is present at \((3.0, 3.5)\). The robot moved for 5 [s], as shown by the green arrow in Fig. 4. The color shows the average spatial spectrum for 5 [s], and the white pixels have a high probability for the presence of a sound source. Although the localization results are ambiguous, the pixels around the actual sound source are white in color.

2) Visual feature extraction: In this part, the system detects a face from an input image, extracts the height and width of lips, and calculates 8-dimensional lip-shape-related visual features (for details, see [11]).

First, the system detects a face from an input image and calculates the most likely position and size of the detected face. The system subsequently estimates the distance from the robot to the user using the detected face size \(d = c_1 r + c_0\), where \(c_0 = -0.0106\), \(c_1 = 4.04\). (7)

The parameters \(c_0, c_1\) can be calculated from face-detection results of training data. The average face sizes for speakers located 1.5 [m] and 2.5 [m] away from the robot are 239 [pixels] and 145 [pixels], respectively. The system extracts the height and width of the lips from the detected face and normalizes height and width relative to detected face size. Polynomial function fitting is performed to the temporal sequence of the normalized height and width of the lips. Finally, the coefficients of the fitting function are used as visual features.

This part is implemented based on Facial-Feature-Tracking SDK, which is included in MindReader\(^1\).

3) AV-VD: In this part, we simply used our AV-VD system proposed in [11]. This model estimated an utterance using four utterance states: non-speech (\(s_0\)), start-motion (\(s_1\)), speech (\(s_2\)), and end-motion (\(s_3\)). This block estimates an utterance state (\(\tilde{s}\)) from a posterior probability for the four utterance states \((s_0, s_1, s_2, s_3)\) using the lip-shape-related visual feature \((x_v)\) and the audio feature \((x_u)\):

\[
\tilde{s} = \arg \max P(s_i | x_u, x_v),
\]

\[
P(s_i | x_u, x_v) = \frac{p(x_u | s_i)p(x_v | s_i) P(s_i)}{p(x_u, x_v)}. \tag{8}
\]

The system performs hangover processing to fix fragmentation based on erosion and dilation \([14]\). Finally, it extracts the part of the input signal corresponding to the spoken state, (for details, see [11]).

<table>
<thead>
<tr>
<th>TABLE I NOTATION AND MEANING OF VARIABLES IN FIG. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(v_0, v_1)</td>
</tr>
<tr>
<td>(v_2, v_3)</td>
</tr>
<tr>
<td>(v_4, v_5, v_6)</td>
</tr>
<tr>
<td>(v_7)</td>
</tr>
<tr>
<td>(v_8)</td>
</tr>
<tr>
<td>(v_9)</td>
</tr>
<tr>
<td>(v_{av})</td>
</tr>
<tr>
<td>(v_{v})</td>
</tr>
<tr>
<td>(v_{avv})</td>
</tr>
</tbody>
</table>

4) Robot controller: In this part, a sound source at the position of the detected face is regarded as a target and the other sound sources are assumed to be noise.

This motion planning is based on the proposed CBN. Among the variables described in Table I, \(v_4, v_5, \text{ and } v_6\) are control variables, and \(v_{av}\) is the target variable. Therefore, the robot selects the optimal active motion based on the effect on AV-VAD caused by active motion \(s^2\):

\[
s^2 = \arg \max P(v_{av} | v_0, \ldots, v_3, v_7, \ldots, v_9, do(v_4, v_5, v_6))
\]

\[
= \arg \max P(v_{av} | v_0, \ldots, v_3, v_7, \ldots, v_9, do(\tilde{\xi}_x, \tilde{\xi}_y, \tilde{\xi}_\theta)), \tag{10}
\]

where \(\tilde{\xi}_x, \tilde{\xi}_y, \tilde{\xi}_\theta\) represents possible active motions. Since the cart is omnidirectional, the robot can perform the following active motion.

\[
\tilde{\xi}_x = \xi_x + \Delta_x,
\]

\[
\tilde{\xi}_y = \xi_y + \Delta_y,
\]

\[
\tilde{\xi}_\theta = \xi_\theta + \omega,
\]

where \(\Delta_x, \Delta_y\) and \(\omega\) are the velocity \((\Delta^2_x + \Delta^2_y = \Delta^2)\) and angular velocity of the cart, respectively. For implementation, we used the Robot Operating System (ROS)\(^2\).

V. EVALUATION

In this section, we evaluate the proposed AAV integration framework through VAD performance.

We compared three methods:

- **Active (Prop.):** Our proposed method.
- **Active (MReg.):** A comparative approach, using multiregression-based estimation of VAD performance. By selecting variables empirically, the position of the sound source \((v_2, v_3)\) and the directional interval between the speaker and the noise source \((v_7)\) are selected as explaining variables.

\[
y = c_0 + c_1 v_2 + c_2 v_3 + c_3 v_7, \tag{14}
\]

\[
c_0 = 0.190, c_2 = 0.160, c_3 = 0.0735, c_7 = 0.171.
\]

This regression model showed a higher determination coefficient \((R^2 = 0.93)\) than a regression model with \(v_7\) and \(v_9\) \((R^2 = 0.78)\) used in our proposed method.

- **Active (Linear):** Simple AAV integration. The robot linearly approaches the speaker and stops when the size of the detected face is almost the same as the average face size in the training data (about 300 [pixels]).

- **Baseline:** The robot stays at its initial position. This approach corresponds to conventional static robot audition.

As a metric of VAD performance, we used F-measure:

\[
F = 100 \times \frac{2N_c}{Nd+N_c} \% \tag{15}
\]

where \(N_c, N_d, \text{ and } N_a\) are the number of correctly-detected, all detected, and all spoken utterances, respectively. We defined an utterance as “correctly-detected” if both its start and end points were estimated within 300 [ms].

\(^1\)http://trac.media.mit.edu/mindreader/

\(^2\)http://www.ros.org/wiki
A. Dataset

We recorded two kinds of AV datasets, an AV word dataset and an AV command sentence dataset. The AV word dataset consisted of 60 words, each uttered by 3 male speakers and captured in a controlled room (clean speech and no occlusion in vision). The robot was placed in front of the speaker, at distances of 1.5 [m] and 2.5 [m]. This AV dataset was used to train the CBN described in the previous sections.

The AV command sentence dataset consisted of 100 command sentences uttered by 1 male speaker in a noisy room. The 100 sentences were randomly-selected Japanese translations of the 6-word command sentences (e.g. “Set blue at A 0 now”) in GRID corpus\(^3\). We used a music signal (RWC Music Database Jazz No. 41\(^4\)), played on a loud speaker, as a source of noise.

B. Experimental condition

The effectiveness of our proposed method was evaluated under two room conditions.

- Exp 1: The speaker was close to the noise source.
- Exp 2: The speaker was far away from the noise source.

We assumed that 1) there was one speaker and one source of noise, 2) the speaker and the robot were facing each other, 3) the sound source locations of the speaker and the noise source were fixed, 4) the initial position of the robot was (0.5, 0.5), and 5) the area within 1 [m] from the target speaker and the noise source has 0 values of the expected VAD performance to avoid collision.

Note that the target speech signal was contaminated by noises coming from the noise source, the robot itself, and its electric power supply.

C. Results

1) Experiment 1: Fig. 5a), b), and c) shows the estimated results of Active(Prop.), Active(MReg.), and actual VAD performance, respectively. In Active (Prop.) and Active (Linear), the robot moved along the curved and straight lines in Fig. 5a), respectively, stopping at T4s. Since both Active (Prop.) and Active (MReg.) showed that the optimal VAD point is (3.0, 3.5), we evaluated VAD performance of Active (Prop.) and Active (Linear), as shown in Fig. 5c). The AV-VAD performance was almost 60% for Baseline, corresponding to a stationary robot audition approach. Active (Linear) improved the performance while the robot was approaching the target speaker. Since, it disregards the noise source, however, the improvement was limited to 10-20 points. In contrast, Active (Prop.) takes noise source into account, and automatically planned the best path for improving VAD performance, with the latter being, over 80%.

This improvement mainly comes from audio-cue. The SSS performance of GHDSS was strongly affected by the difference between two sound sources because GHDSS utilizes geometric information. When the robot performed AV-VAD based on Baseline, Active (Linear), and Active (Prop.) the differences between the target speaker and the noise source were 13.7 [deg], 26.6 [deg], and 90.0 [deg], respectively, with a larger degree indicating a better quality of audio-cue and better VAD performance. Thus, we confirmed the effectiveness of our proposed method. We also found that the proposed AV-VAD outperformed A-VAD and V-VAD.

2) Experiment 2: Fig. 6a), b), and c) shows the estimation results of Active (Proc.), Active (MReg.), and actual VAD performance, respectively. In this condition, both Active (Proc.) and Active (MReg.) tried to go to a place on the line \(y = 2.0\) [m]. After that, Active (Prop.) tried to get closer to the speaker (Fig. 6a), whereas Active (MReg.) stopped as shown in Fig. 6b).

Note that when we used \(v_7\) and \(v_9\) for multi regression analysis, the determination coefficient (\(R^2 = 0.78\)) was smaller than that of Active (MReg.).

The VAD performances are shown in Fig. 6c). Since the robot was close to the noise source, the VAD performance was worse than in Exp. 1. VAD performance gradually increased as the robot moved in Active (Prop.), and it was about 50 % at T4, 22.0 and 28.0 points higher than Active (MReg.) and Baseline, respectively.

D. Discussion

In this work, we assumed that the speaker was facing to the robot, and we conducted the experiment in a room with fairly constant illumination and background color. However, in more realistic environments, audio and visual noises affect the performance of the AAV integration framework badly.

1) Effect of an acoustic noise: The effect of an acoustic noise is limited to be small because of the following reason. If the noise is stationary, the robot moves to reduce the effect of the noise. If the noise is impulsive, the AV-VAD error due to the noise is fixed in the hangover processing.

2) Effect of a visual noise: Visual noises affect the performance of the AAV integration framework badly due to face detection failure.

When the robot detects no face, the system regards the detected face size is zero, and the robot tries to improve the quality of the audio cue. When the robot detects a face with errors, the effect is limited to be small because of the following reason. The typical error is miss-detection of a jaw as a bottom lip. When such miss-detection occurs, the robot underestimates the distance between the robot and a speaker and stops before the robot arrives at the optimal position. When the robot detects a background image as a face, the system can not recover from such miss-detection. If the face detection reliability is low, the system ignores such miss-detection and regards it as no detected face. If the face detection reliability is high, the problem is crucial. For example, when the robot captures a face on a television display, the robot tries to move as if the TV is an actual human speaker. To solve such a problem, additional sensors

\(^3\)http://spandh.dcs.shef.ac.uk/gridcorpus/
\(^4\)http://staff.aist.go.jp/m.goto/RWC-MDB/
such as an infrared camera and a laser range finder should be used.

VI. SUMMARY

We have proposed a CBN-based AAV integration framework to estimate the effects of active motion on VAD performance. This framework integrates audio-cue, visual-cue, and active motion by using a CBN. A CBN is used to estimate VAD performance from intermediate outputs of AV-VAD. A robot selects the best place to move according to the estimation result. We implemented a prototype system in a humanoid robot called Hearbo and showed its effectiveness through voice activity detection experiments. Our preliminary results showed that simple AAV-VAD and multiregression-based AAV-VAD improved the AV-VAD performance compared with that at the initial position, and the proposed AV-VAD improved the VAD performance further.

Much future work remains, including a detailed evaluation of the proposed AV-VAD under more realistic conditions and the use of a more interactive approach than the proposed approach. For example, a robot that can ask a speaker to speak louder or move away from the sources of noise. Incorporation of such interactive functions into this framework is challenging. Coping with hidden variables is also challenging, since we considered only observable variables.

VII. ACKNOWLEDGMENTS

The authors thank Professor Jun-ichi Imura and Professor Tomohisa Hayakawa of TITech, Professor Hiroshi G. Okuno of Kyoto University, and Dr. Rana el Kaliouby and Professor Rosalind W. Picard of MIT. This work was partially supported by Grants-in-Aid (No.24118702, No.22700165) and a JSPS fellowship (No.239496).

REFERENCES