



Omnidirectional Audio-Visual Talker Localizer With Dynamic Feature Fusion Based on Validity and Reliability Criteria

Yuki Denda¹, Takanobu Nishiura², and Yoichi Yamashita²

¹Graduate School of Science and Engineering, Ritsumeikan University, Japan

²College of Information Science and Engineering, Ritsumeikan University, Japan

{gr021052@se, nishiura@is, yama@media}.ritsumei.ac.jp

Abstract

Talker localization is indispensable in video conferencing. Statistical audio-visual (AV) talker localizers that fuse AV features based on prior statistical property are ideals. However, statistical property must be estimated prior to the AV feature fusion procedure. To overcome this problem, this paper proposes a novel robust and omnidirectional AV talker localizer that dynamically fuses AV features based on validity and reliability criteria for eliminating prior statistical property. Direction estimation of speech arriving using equilateral triangular microphone array and human position detection using an omnidirectional video camera extract AV features from captured AV signals. Validity criterion, called audio- or visual-localization counter, validates both features. Reliability criterion, called evaluator of directional-speech arriving, acts as weight for dynamic AV feature fusion. The results of talker localization experiments in an actual office room confirmed that the proposed AV localizer based on dynamic feature fusion is superior to that of the conventional localizer that utilizes either audio or visual features.

Index Terms: Dynamic feature fusion, omnidirectional localization, direction estimation of speech arriving, human position detection, video conferencing

1. Introduction

Analysis and transcription of video conferencing attended by multiple users has been widely investigated in recent years [1]. Such works require the talker (active user) to be located and the talker to be isolate from other silent users. However, locating and isolating the talker is very difficult in real noisy environments. This is because the audio signal is contaminated with room reverberations and/or ambient noise, and the visual signal is corrupted by the oscillation of fluorescent lights or drastic scene changes. Thus, only using the audio or visual signal to locate the talker causes errors.

To cope with these problems, audio-visual (AV) talker localizers have been investigated [2, 3, 4]. We have proposed an AV localizer using a microphone array and a pan/tilt/zoom (PTZ) video camera [2]; however, this localizer cannot locate the talker when they exit of view of the PTZ video camera. On the other hand, multiple video cameras distributed throughout a room can acquire visual signals over a wide field of view [3]. However, this system becomes complex because it requires each individual video camera to be calibrated. A statistical approach (ex. Bayesian network) has been used to fuse features [4]. This approach statistically combines AV features based on results in a conditional probability table (CPT). The CPT must be estimated prior to the fusion procedure; therefore, it requires prior supervised training using a large number of training data that

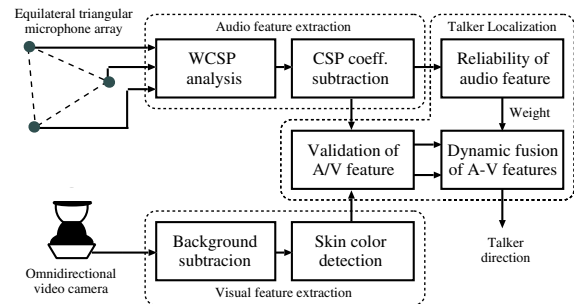


Figure 1: Overview of the proposed omnidirectional audio-visual talker localizer.

have been collected in actual processes. Unfortunately, this approach tends to fail to locate the talker if some of the features are unreliable.

To deal with these problems, we propose a robust omnidirectional AV localizer. The proposed localizer offers two innovations. One is omnidirectional audio features and omnidirectional visual features. The audio features are extracted by estimating the direction of speech arriving based on weighted cross-power spectrum phase (WCSP) analysis [5] and CSP coefficient subtraction [5] using an equilateral triangular microphone array. The visual features are extracted by detecting a human's position based on normalized distance-based background subtraction [6] and skin color detection [7] using an omnidirectional video camera. The other is robust dynamic feature fusion for omitting prior supervised statistical training. In this paper, the talker is assumed to remain virtually stationary for a short-time period, and is also assumed to be talking into AV equipment. Thus, the following two criteria accomplish robust and dynamic feature fusion using observed features. The validity criterion, called the audio- or visual-localization counter, validates both features. This is because the speech arriving or the human's position is continuously observed using audio or visual features observed over a short period of time. The reliability criterion, called the evaluator of speech arriving, acts as a weight for dynamic fusion. This is because the audio features in the speech activity are sufficiently reliable to contribute to localization, and the visual features must assist in localization in non-speech activity.

2. Proposed omnidirectional audio-visual talker localizer

Figure 1 is an overview of the proposed omnidirectional AV talker localizer. Omnidirectional audio signals and visual sig-

nals are conveniently captured using equilateral triangular microphone array and an omnidirectional video camera. Following WCSP analysis and CSP coefficient subtraction extract audio features, and background subtraction and skin color detection extract visual features. Finally, the talker is located using dynamic AV feature fusion based on the validity and reliability criteria. This dynamic fusion procedure, hence, omits the need for calculating statistical property beforehand because it is performed using observed features.

2.1. Audio feature extraction

2.1.1. WCSP analysis

The target audio in video conferencing is the speech. In the current research, we assumed that there was no correlation between the spectral characteristics of speech and interference. We have already proposed WCSP analysis using an average speech spectrum as a specialized method to estimate the direction of speech arriving [5]. This is derived:

$$\begin{aligned} WCSP(k) &= \text{IDFT} \left[W(\omega) \frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|} \right] \\ &= \text{IDFT} \left[W(\omega)e^{-j\omega(\varphi_2-\varphi_1)} \right], \end{aligned} \quad (1)$$

$$WCSP(\theta) = \mathcal{F}(WCSP(k)), \quad (2)$$

$$\mathcal{F}: \theta = \cos^{-1} \left(\frac{ck}{dF_s} \right), \quad (3)$$

where $WCSP(k)$ is the time domain (k) WCSP coefficient, $W(\omega)$ is the average speech spectrum-based weight coefficient, $\text{IDFT}[\cdot]$ is the inverse discrete Fourier transform, $X_{[\cdot]}(\omega)$ is the frequency representation of $x_{[\cdot]}(t)$, $x_{[\cdot]}(t)$ is the audio signal captured by the paired-microphones, $\varphi_{[\cdot]}$ is the arrival times of $x_{[\cdot]}(t)$, $*$ denotes a complex conjugate and $WCSP(\theta)$ is the directional domain (θ) WCSP coefficient, respectively. Equation (3) projects the time domain WCSP coefficient into the directional domain WCSP coefficient, where c is the sound propagation speed, d is the distance between paired microphones and F_s is the sampling frequency, respectively. As derived from Eq. (1), the phase information at each frequency is weighted using the average speech spectrum-based weight coefficient. This is because the phase information related to the arrival of the speech is correct only at those frequencies where the speech is greater than the interference. As a result, the weight coefficient can be used as the reliability criterion for the arrival of speech at each frequency. However, our initial WCSP analysis can only deal with frontal 180 degrees of the paired-microphones. To overcome this problem, we used an equilateral triangular microphone array to synthesize omnidirectional WCSP coefficients. As shown in Fig. 2, the equilateral triangular microphone array consists of three paired-microphones: (M_l, M_r) , (M_r, M_c) and (M_c, M_l) ; thus, they provide three WCSP coefficients: $WCSP_{(l,r)}(\theta)$, $WCSP_{(r,c)}(\theta)$ and $WCSP_{(c,l)}(\theta)$. Omnidirectional WCSP coefficients $WCSP_{(omni)}(\theta)$ are synthesized using the following equation based on the arrangement of each microphone:

$$\begin{aligned} WCSP_{(omni)}(\theta) &= WCSP_{(l,r)}(\theta) \\ &+ WCSP_{(r,c)}(\theta + 120) + WCSP_{(c,l)}(\theta - 120). \end{aligned} \quad (4)$$

The omnidirectional WCSP coefficients $WCSP_{(omni)}(\theta)$ are simply described as $WCSP(\theta)$.

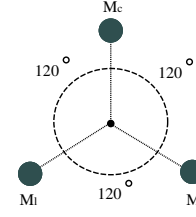


Figure 2: Arrangement of equilateral triangular microphone array.

2.1.2. CSP coefficient subtraction

The CSP coefficient subtraction performs noise-robust direction estimation of the speech arriving by eliminating the spatial distribution of spatially stationary-noise [5]. This is accomplished by subtracting the noise-oriented WCSP coefficients (the spatial distribution of interference) from the observed WCSP coefficients:

$$WCSP_s(\theta) = WCSP(\theta) - \alpha WCSP_{\bar{\pi}}(\theta), \quad (5)$$

$$WCSP_{\bar{\pi}}(\theta) = \frac{\sum_{n=1}^N \max(WCSP(n, \theta), 0)}{N}, \quad (6)$$

$$\alpha = \frac{\max(WCSP(\theta))}{\max(WCSP_{\bar{\pi}}(\theta))}, \quad (7)$$

where $WCSP_s(\theta)$ is the speech-oriented WCSP coefficient, $WCSP_{\bar{\pi}}(\theta)$ is the noise-oriented WCSP coefficient estimated prior to the subtraction procedure and α is the automatically controlled subtraction coefficient. After WCSP coefficient subtraction, the speech-oriented WCSP coefficient can be used as the audio feature $F_A(\theta)$.

2.2. Visual feature extraction

2.2.1. Normalized distance-based background subtraction

Normalized distance-based background subtraction robustly detects the potential human areas against changes in lighting conditions [6]. The image is divided into $N \times N$ pixel blocks, then, N^2 dimensional vector composed of the intensity at each pixel is acquired. As a result, the normalized distance is derived:

$$ND(\mathbf{I}(u, v)) = \left| \frac{\mathbf{I}(u, v)}{|\mathbf{I}(u, v)|} - \frac{\mathbf{I}_B(u, v)}{|\mathbf{I}_B(u, v)|} \right|, \quad (8)$$

where u indicates the horizontal position of the input image, v indicates the vertical position of the image input, $\mathbf{I}(u, v)$ is the intensity vector of the image, $\mathbf{I}_B(u, v)$ is the intensity vector of the background model and $|\cdot|$ indicates the norm of a vector, respectively. The blocks in which the normalized distance is above a certain threshold are assumed as potential human areas, and the following skin color detection is conducted against these blocks.

2.2.2. Skin color detection

Skin color is an important feature for detecting a human's position visually. Results show that the skin color Gaussian model in the TS plane of a tint/saturation/lightness (TSL) color space decreases most modeling errors [7]. The red/green/blue (RGB)

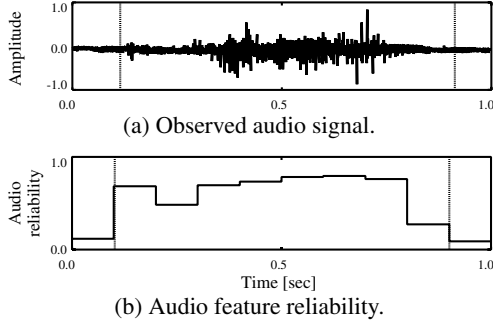


Figure 3: Observed audio signal and audio feature reliability.

colors are projected onto the TS plane:

$$r = R/(R + G + B) - 1/3, \quad (9)$$

$$b = B/(R + G + B) - 1/3, \quad (10)$$

$$S = \sqrt{(9.0/5.0)(r^2 + g^2)}, \quad (11)$$

$$T = \begin{cases} \tan^{-1}(r/g)/2\pi + 1/4, & g > 0 \\ \tan^{-1}(r/g)/2\pi + 3/4, & g < 0 \\ 0 & g = 0 \end{cases}. \quad (12)$$

Then, visual feature $F_V(\theta)$ is extracted by summing the log-likelihood of pixel input along a vertical line of input image at talker candidate areas:

$$\mathbf{I}(\theta, \phi) = [S(\theta, \phi), T(\theta, \phi)], \quad (13)$$

$$\mathbf{I}_S = [\bar{S}, \bar{T}], \quad (14)$$

$$\boldsymbol{\lambda} = [\mathbf{I}(\theta, \phi) - \mathbf{I}_S], \quad (15)$$

$$P[\mathbf{I}(\theta, \phi)|S] = -\ln(e^{-\frac{1}{2}[\boldsymbol{\lambda}^T \mathbf{R}^{-1} \boldsymbol{\lambda}]} / 2\pi |\mathbf{R}_s|^{1/2}), \quad (16)$$

$$F_V(\theta) = \sum_{\phi=0}^{\phi_I} (P[\mathbf{I}(\theta, \phi)|S]), \quad (17)$$

where $\mathbf{I}(\theta, \phi)$ is the input vector, \mathbf{I}_S is the mean vector of the skin color Gaussian model, \mathbf{R}_s is the covariance matrix of the skin color Gaussian model, T indicates transposition, $^{-1}$ indicates the inversion of the matrix and ϕ_I is the vertical resolution of the image input, respectively.

2.3. Validity criterion

Audio features or visual features continuously detect speech arriving or a human's position from their feature distribution observed in a short period of time. This is because the talker is assumed to be a spatially stationary-sound source observed in a short period of time. Consequently, both features are validated using an audio- or a visual-localization counter:

$$V_{(A \text{ or } V)}(i, \theta) = \frac{\sum_{t=1}^T lc_{(A \text{ or } V)}(i - t, \theta)}{T}, \quad (18)$$

$$lc_{(A \text{ or } V)}(i, \theta) = \begin{cases} 1, & F_{(A \text{ or } V)}(i, \theta) \geq TH \\ 0, & F_{(A \text{ or } V)}(i, \theta) < TH \end{cases}, \quad (19)$$

where $V_{(A \text{ or } V)}(i, \theta)$ is a validity criterion of each feature, $lc_{(A \text{ or } V)}(i, \theta)$ is an audio- or a visual-localization counter and T is the number of counting frames, respectively. As derived from Eq. (18), the validity criterion is the normalized histogram of audio- or visual-based localization.

2.4. Reliability criterion for dynamic feature fusion

The audio features in the speech activity are sufficiently reliable to contribute to talker localization, and the visual features

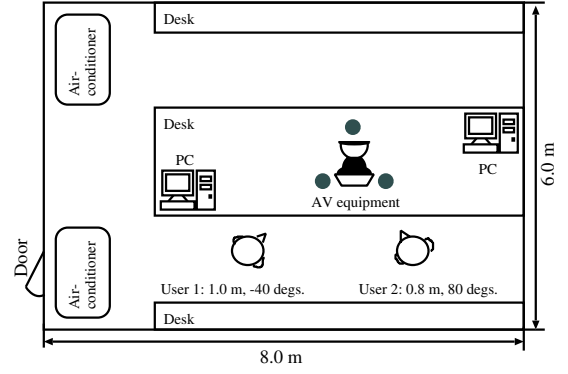


Figure 4: Experimental environment.

Table 1: Experimental conditions.

| Audio conditions | |
|-----------------------|--------------------------|
| Microphone array | 150.0 mm apart |
| Sampling frequency | 16 kHz |
| Room reverberation | 0.41 sec. ($T_{[60]}$) |
| Ambient noise | 44.1 dBA |
| Visual conditions | |
| Omnidirectional image | 640 × 480 pixels |
| Panoramic image | 750 × 230 pixels |
| Frame rate | 15 fps |
| Interference | skin colored envelope |
| Localization | |
| User 1 | -40 degrees, 1.5 m |
| User 2 | +80 degrees, 1.0 m |
| Frame length | 1/15 sec. |

must assist in talker localization in non-speech activity. This is because the talker is assumed to be a user who is speaking into AV equipment. As a result, AV feature fusion with a fixed weight is prone to fail in localization. To deal with this problem, we fused audio features and visual features by dynamically updating the weight

$$R_A(i) = \frac{\max(F_A(i, \theta))}{\sum_{\theta=0}^{\theta_I} F_A(i, \theta)}, \quad (20)$$

$$F_{AV}(i, \theta) = R_A(i)(V_A(i, \theta)F_A(i, \theta)) + (1 - R_A(i))(V_V(i, \theta)F_V(i, \theta)), \quad (21)$$

$$Talker = \begin{cases} \text{Presence,} & F_{AV}(i, \theta) \geq TH \\ \text{Absence,} & F_{AV}(i, \theta) < TH \end{cases}, \quad (22)$$

where $R_A(i)$ is the reliability criterion that is the ratio of the maximum value of the audio feature to the total value of the audio features. As derived from Eq. (21), the reliability criterion acts as the weight for the dynamic feature fusion. Figure 3 shows an observed audio signal (a) and audio feature reliability (b). As shown in Fig. 3, the audio feature reliabilities have high values in speech activity and have small values in non-speech activity or at the end of speech activity that contain barely audible speech. In other words, the audio feature reliability is a criterion for assessing whether a talker is present. As a result, validity criterion and reliability criterion based on only observed features and not statistical property estimated beforehand were successfully used to dynamically locate talkers.

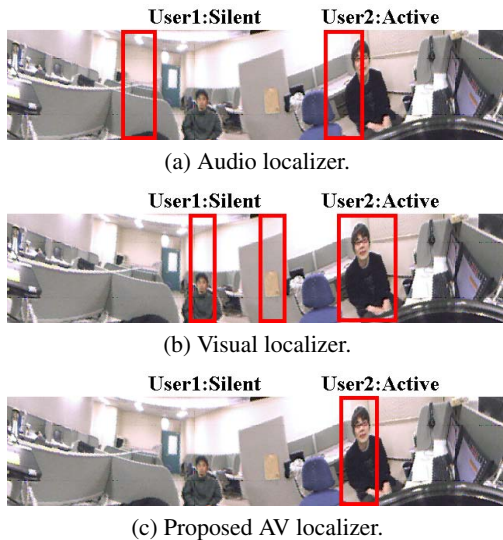


Figure 5: Experimental results of talker localization.

Table 2: Talker localization performance.

| | FRR [%] | FAR [%] |
|-----------------------|---------|---------|
| Audio localizer | 15.1 | 17.2 |
| Visual localizer | 9.8 | 69.6 |
| Proposed AV localizer | 4.2 | 7.6 |

3. Evaluation experiments

3.1. Experimental conditions

Figure 4 is a description of the experimental environment we used and Tab. 1 lists the experimental conditions. All audio signals were captured using an equilateral triangular microphone array spaced 150.0 mm apart. All visual signals were captured using an omnidirectional video camera that provided panoramic image was 750×230 pixels. A skin colored envelope was used as visual interference; however, it was not contained in background model. Two users sat near the AV equipment. One user (User 1) was located 1.0 m from the AV equipment and at an angle of -40 degrees to their direction, the other user (User 2) was located 0.8 m from the equipments and at an angle of $+80$ degrees. The talker localization performance was objectively evaluated using a false rejection rate (FRR) and a false acceptance rate (FAR) in speech activity. The localization frame length was $1/15$ seconds because the visual frame rate was 15 fps.

3.2. Experimental results

Figure 5 shows the experimental results when User 1 was silent and User 2 was speaking. Figure 5(a) shows the localization result when only using the audio features, and Fig. 5(b) shows the localization results when only using the visual features. As shown in Fig. 5(a), when the User 2 was located, audio interference cause incorrect talker localization. In contrast, as shown in Fig. 5(b), User 1 who was silent and a skin colored envelope were also incorrectly located. Finally, as shown in Fig. 5(c), the AV localizer more accurately located the speaking talker when compared to the results achieved based on either audio or visual features. Table 2 lists the talker localization performance based on a FRR and a FAR. The results listed in Tab. 2 indicated that the AV localizer effectively improved the FRR and FAR when compared to the results achieved with a conventional localizer based on either audio or visual features.

The experimental results thus have shown that the proposed AV talker localizer based on dynamic AV feature fusion performs robust talker localization. It is hence a promising noise-robust localizer in real noisy environments.

4. Conclusion

This paper proposed a novel robust and omnidirectional audio-visual (AV) talker localizer based on dynamic AV feature fusion using validity criterion and reliability criterion and no prior statistical property. Omnidirectional audio features were extracted by estimating the direction of speech arriving based on weighted cross-power spectrum phase (CSP) analysis and CSP coefficient subtraction using an equilateral triangular microphone array. Omnidirectional visual features were extracted by detecting a human's position based on normalized distance-based background subtraction and skin color detection using an omnidirectional video camera. The talker is then located by dynamically fusing AV features based on validity and reliability criteria. The validity criterion, called the audio- or visual-localization counter, validates both features, whereas the reliability criterion, called evaluator of directional-speech arriving, acts as a weight to dynamically fuse AV features. To evaluate the effectiveness of the proposed omnidirectional AV localizer, we conducted talker localization experiments in an actual office. The experiments revealed that the talker localization performance of the proposed AV localizer using dynamic AV feature fusion was superior to that of conventional localizers that used either only audio or visual features. Future research will involve evaluating the proposed AV localizer in other conditions, for example, double-talker condition.

5. Acknowledgements

This work was supported in part by The Leading Project for an "e-Society" and Grants-in-Aid for Scientific Research Nos. 17700216 and 17200014 funded by The Ministry of Education, Culture, Sports, Science and Technology of Japan.

6. References

- [1] T. Hain, et al., "Transcription of conference room meetings," Proc. Eurospeech05, pp. 1611–1614, 2005.
- [2] Y. Denda, et al., "A design of audio-visual talker tracking system based on CSP analysis and frame difference in real noisy environments," Proc. MMSP04, pp. 63–66, 2004.
- [3] K. Wilson, et al., "Audio-video array source separation for perceptual user interfaces," Proc. WPUI01, pp. 1–7, 2001.
- [4] A. Garg, et al., "Boosted learning in dynamic Bayesian networks for multimodal speaker detection," Proc. IEEE, Vol. 91, No. 9, 2003.
- [5] Y. Denda, et al., "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," IEICE Trans. on Inform. & Syst., Vol. E89-D, No. 3, pp. 1050–1057, Mar, 2006.
- [6] S. Nagaya, et al., "Moving object detection by time-correlation-based background judgment method," IEICE Trans. on Inform. & Syst. Vol. J9-DII, No. 4, pp. 568–576, 1996.
- [7] J.C. Terrillon, et al., "Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images," Proc. ICVI, pp. 180–187, 1999.