COLLABORATIVE STEERING OF MICROPHONE ARRAY AND VIDEO CAMERA TOWARD MULTI-LINGUAL TELE-CONFERENCE THROUGH SPEECH-TO-SPEECH TRANSLATION

Takanobu Nishiura, Rainer Gruhn, and Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan

ABSTRACT

It is very important for multi-lingual tele-conferencing through speech-to-speech translation to capture distant-talking speech with high quality. In addition, the speaker image is also needed to realize a natural communication in a such conference. A microphone array is an ideal candidate for capturing distant-talking speech. Uttered speech can be enhanced and speaker images can be captured by steering a microphone array and a video camera in the speaker direction. However, to realize automatic steering, it is necessary to localize the talker.

To overcome this problem, we propose collaborative steering of the microphone array and the video camera in real-time for a multi-lingual tele-conference through speech-to-speech translation. We conducted experiments in a real room environment. The speaker localization rate (i.e., speaker image capturing rate) was 97.7%, speech recognition rate was 90.0%, and TOEIC score was 530∼540 points, subject to locating the speaker at a 2.0 meter distance from the microphone array.

1. INTRODUCTION

To achieve multi-lingual tele-conferencing through speech-to-speech translation, the high-quality sound capture of distant-talking speech is very important. However, background noise and room reverberations seriously degrade the sound capture quality in real acoustical environments. A microphone array is an ideal candidate as an effective method for capturing distant-talking speech. With the microphone array, the desired speech signal can be selectively acquired by precisely steering the directivity in the desired speech direction. The following directivity patterns can be executed by the microphone array:

- Delay-and-sum beamformers [1, 2]
  Steer the directivity to a sound source direction.
- Multiple beamformers [3]
  Steer the directivity not only to a direct sound direction but also to reflection sound directions.
- Adaptive beamformers [4]
  Steer the Null directivity to noise directions.

The above methods can reduce the directional noise effects. Thus, they are often used as front-end processing in Automatic Speech Recognition (ASR). In this paper, we use the delay-and-sum beamformer to realize robust high-quality sound capture of distant-talking speech in various environments and achieve real-time processing. Also, it is necessary to localize the speaker direction to realize high-quality sound capture of distant-talking speech. Until now, much research on Direction of Arrival (DOA) estimation has been conducted. CSP (Cross-power Spectrum Phase)[5] analysis, which can be done by simple calculation, is an effective method for estimating DOA. However, CSP analysis degrades localization performance in noisy reverberant environments. To overcome this problem, we proposed the CSP coefficient addition method[6] based on CSP analysis. In this paper, DOA is also estimated by the CSP coefficient addition method, and a video camera is also steered automatically to estimated DOA.

Figure 1: Speech-to-speech translation system of distant-talking speech.

By using the above methods, the speaker speech and image can be captured robustly and accurately. However, these methods can not translate the speech, although they can capture it automatically with high quality. To cope with this problem, we translate beamformed speech using ATR's Multi-lingual Automatic Translation System for Information Exchange (ATR-MATRIX)[7]. It is an ideal candidate as an effective tool for translating speech-to-speech. It consists of a speech recognition sub-system (ATR-SPREC), a language translation sub-system (TDMT), and a speech synthesis sub-system (CHATR). Currently, it has achieved a performance of TOEIC 550 points[8]. Furthermore, it not only immediately shows the text of translated speech for multi-lingual tele-conferencing but also synthesizes translated speech using CHATR.

In this paper, we propose a system with automatic steering of the microphone array and the video camera as a step toward achieving multi-lingual tele-conferencing through speech-to-speech translation. Figure 1 shows the setup of this system.

2. KEY TECHNOLOGY FOR PROPOSED SYSTEM

Figure 2 shows an overview of the proposed system. In this system, a video camera and a microphone array are first automati-

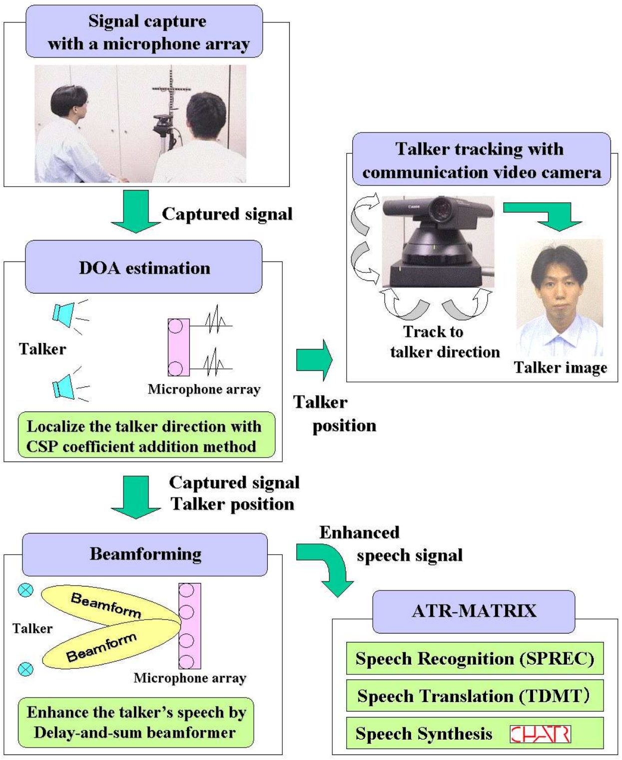

Figure 2: Proposed system overview.

Figure 3: Microphone array and video camera.

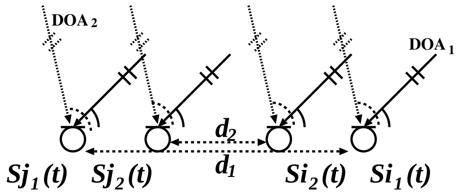

Figure 4: DOA estimation with CSP coefficient addition method.

cally steered to the DOA estimated by the CSP coefficient addition method after capturing speech with a microphone array. Next, speech beamformed by the steering directivity of the microphone array is translated and synthesized using ATR-MATRIX for multi-lingual tele-conferencing. The speaker image is captured by a video camera and shown at the same time. A natural multi-lingual tele-conference can be realized with this system. Figure 3 shows the microphone array and the video camera used in the proposed system. Next, the key technologies for the proposed system are explained in detail.

2.1. DOA estimation by CSP coefficient addition method

DOA must be estimated to automatically steer the microphone array and the video camera. Thus, we use the CSP coefficient addition method[6] to estimate DOA. In the environment of Figure 4, the CSP coefficients are derived from Equation (1).

$$\mathrm{CSP}_{i_n,j_n}(k) = \mathrm{IDFT}\left[\frac{\mathrm{DFT}\left[s_{i_n}(t)\right]\mathrm{DFT}\left[s_{j_n}(t)\right]^*}{\left|\mathrm{DFT}\left[s_{i_n}(t)\right]\right|\left|\mathrm{DFT}\left[s_{j_n}(t)\right]\right|}\right], \quad (1)$$

where t and k are the time index, DFT $[\cdot]$ (or IDFT$[\cdot]$) is the discrete Fourier transform (or the inverse discrete Fourier transform), and the symbol * is the complex conjugate. Then, CSP coefficients are added as shown in Equation (2).

$$\begin{aligned}\mathrm{CSP}_{i,j}(\theta) &= \sum_{n=1}^{N}\mathrm{CSP}_{i_n,j_n}(\theta), \\ subject \quad to \quad & \theta = \cos^{-1}\left(\frac{c\cdot k/F_s}{d_n}\right), \end{aligned} \quad (2)$$

where N is the number of additions, d_n is the distance between two adjacent transducers, c is the sound propagation speed, and F_s is the sampling frequency. The DOA can be accurately estimated by finding the maximum values of the added CSP coefficients by Equation (3).

$$\mathrm{DOA_n} = \underset{\theta}{\mathrm{argmax}}(\mathrm{CSP}_{ij}(\theta)). \quad (3)$$

The CSP coefficient addition method can estimate multiple DOAs. However, as this system has only one video camera, we selected the desired DOA based on signal energy of estimated DOAs. The CSP coefficient addition method is suitable for real-time processing because it can accurately estimate DOAs by simple calculation.

2.2. Automatic video camera steering for capturing speaker image

The video camera is automatically steered to the DOA estimated by the CSP coefficient addition method in order to automatically capture the speaker image and to facilitate multi-lingual tele-conferencing. In this system, the speaker image is automatically captured with a video camera as shown in Figure 3. It can move not only in the horizontal direction but also in the vertical direction. Video camera steering is controlled through an RS232C port by a server computer. The video image is shown immediately on a monitor and translated speech is uttered from a loud speaker.

2.3. Microphone array steering for speech enhancement

Beamforming is necessary to capture distant-talking speech at high quality with a microphone array. In this paper, a delay-and-sum beamformer[1, 2] is used to form the directivity to the desired sound direction. As shown in Figure 5, to capture the signal by microphone array, we assume that the plane wave of the desired sound signal comes from direction θ, the number of transducers is M, and the spacing between the transducers is d. In beamforming, the captured signals $x_1(t), x_2(t), \cdots, x_M(t)$ are shown as time

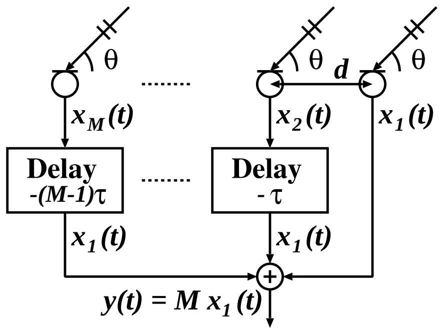

Figure 5: Delay-and-sum beamformer.

delays of $x_1(t)$ in Equation (4).

$$x_m(t) = x_1\left(t - (m-1)\tau\right),\ \tau = \frac{d\cos\theta}{c}, \tag{4}$$

where m ($m = 1, 2, \cdots, M$) is number of transducers and c is the sound propagation speed. Output signal $y(t)$ of the delay-and-sum beamformer is shown in Equation (5).

$$y(t) = \sum_{m=1}^{M} x_m\left(t + (m-1)\tau\right),\ \tau = \frac{d\cos\theta}{c}. \tag{5}$$

In Equation (5), the desired sound signal from direction θ is emphasized M times because the sound signals captured with multiple transducers are added after synchronizing them. On the other hand, no other sound signal is M times as large as the desired sound signal because the directions of the other signals are different from the direction of the desired sound signal. Thus, the directivity of the delay-and-sum beamformer can only be formed to direction θ. Therefore, the delay-and-sum beamformer can form directivity to the estimated DOA.

2.4. Speech-to-Speech Translation with ATR-MATRIX

ATR-MATRIX[7] consists of a speech recognition sub-system (ATR-SPREC), a language translation sub-system (TDMT), and a speech synthesis sub-system (CHATR). The current implementation of our system deals with a hotel room reservation task/domain. The speech recognition sub-system recognizes speech that is beamformed with the microphone array. Then, the language translation sub-system translates the recognized speech. Finally, the translated speech is synthesized by CHATR.

2.4.1. Speech recognition with ATR-SPREC

The speech recognizer module was built based on ATR-SPREC, a speech recognition software toolkit developed at ATR. ATR-SPREC has the following settings:

- Acoustic model: Shared-state context-dependent (triphone) HMMs produced by the ML-SSS algorithm.
- Language model: Multi-class composite N-gram.
- Search engine: A decoder featuring multi-pass search and word graph output.

Table 1: System components

AD converter	Thinknet DF4448
Microphone	Hoshiden KUC1333
Microphone array	Onkyo Sokki OMA520
	29 transducers,
	(horizontal:15, vertical:15)
	2.125 cm spacing
Microphone amplifier	Thinknet MA2016
Server computer	COMPAQ XP-1000 $\times$ 2
	CPU: 500 MHz, Memory: 512 MB
OS	COMPAQ Tru64 UNIX V5.0A
Video camera	CANON VC-C3

Table 2: System algorithms

DOA estimation	CSP coef. additional method[6]
Beamformer	Delay-and-sum beamformer[1, 2]
Speech-to-speech translation	ATR-MATRIX[7]

2.4.2. Language translation with TDMT

The language translator module uses Transfer Driven Machine Translation (TDMT) technology and can deal with various expressions in spoken languages because it uses not only sentence structures but also translation examples. The basic mechanisms of TDMT are as follows:

- Extraction of partial linguistic structures (patterns) from an input sentence.
- Example-based and pattern-by-pattern transfer to a target language.
- A search for the most likely combination of transferred patterns.

2.4.3. Speech synthesis with CHATR

Speech synthesis is essential for realistic multi-lingual teleconferencing through speech-to-speech translation. CHATR produces natural synthetic speech by selecting and re-sequencing wave units from a CHATR-specific speech database. We used CHATR as a speech synthesizer for realizing the proposed system. Since the current configuration of our system has male and female acoustic modules, the CHATR speech synthesis sub-system can output either male or female voices.

3. SYSTEM SPECIFICATIONS

Tables 1, 2, and 3 show the proposed system's specifications. This system uses two workstations (server computers). One is for multi-channel signal capture, DOA estimation, beamforming, and video camera steering. The other is for speech-to-speech translation. The two computers are connected by a LAN (Local Area Network) and communicate with each other by socket protocol. If we conduct multi-lingual tele-conferencing, two sets of this system will be needed. Although the video camera can move in both horizontal and vertical directions, movement in the vertical direction is slower than that in the horizontal direction because of the video camera performance. An AD converter is connected to the server computer by SCSI, and the video camera is connected through an RS232C port.

Table 3: System conditions

AD conversion	
Sampling frequency	16 kHz
Quantization	16 bit
DOA estimation and Beamforming	
Frame length	128 msec. (interval: 64 msec.)
Window	Hamming window
ATR-MATRIX	
Frame length	25 msec. (interval: 10 msec.)
Pre-emphasis	$1 - 0.95z^{-1}$
Feature vector	MFCC 12 orders, ΔMFCC 12 orders, Δlog-power 1 order
Window	Hamming window
Video camera	
CCD pixel size	1/4 inch
Moving performance	Pan: (speed: $1^\circ \sim 76^\circ$/s , angle: 180°)
	Tilt: (speed: $1^\circ \sim 70^\circ$/s , angle: 55°)

4. SYSTEM PERFORMANCE

The proposed system was evaluated in an acoustic experimental room. Reverberation time of this room is $T_{[60]} = 0.27$ seconds and ambient noise level is 24.3 dBA. Two speakers engage in mutual talk to as in a tele-conference using one system. Also, we evaluate the proposed system by assuming that one speaker speaks Japanese and the other speaker listens in English because we can only realize Japanese to English translation at this time. Two speakers are located at positions along $30^\circ, 60^\circ, 90^\circ, 120^\circ$, and 150° directions and 2 meters distance from the microphone array as shown in Figure 1. Table 4 shows our experimental results. These results are averaged from 4 subjects (1 female and 3 males). A hotel room reservation task consisting of 42 dialogues was used as test data. With the proposed system, DOA estimation rate (i.e., speaker image capturing rate) was 97.8%, and speech recognition rate was 90.0%, compared to 91.4% with a closed talking microphone. We also evaluated speech translation performance with TOEIC score according to reference [8]. As a result, we confirmed that the proposed system may achieve about 530∼540 points while the system with a closed talking microphone is 546 points. Next, we investigated the system response speed. As a result, we confirmed that DOA can be estimated within 0.192 seconds, the video camera is steered automatically with about 0.2 seconds delay after capturing speech, and beamforming still takes more than about 0.064 seconds after estimating DOA. Thus, we can conclude that it will take about 0.256 seconds delay after capturing speech to ATR-MATRIX. We could confirm that the proposed system achieves high speech-to-speech translation performance, although it is slightly less effective than a system with a closed talking microphone. We could also confirm a system response speed of within 0.256 seconds for the automatic steering of the microphone array and video camera.

5. CONCLUSIONS

In this paper, we proposed automatic steering of a microphone array and video camera as a step toward achieving multi-lingual teleconferencing through speech-to-speech translation. First, DOA is estimated by the CSP coefficient addition method after capturing speech with the microphone array. Then, the microphone array and the video camera are automatically steered to the estimated DOA, and the speaker image is captured by the video camera. Speech beamformed by the microphone array is translated and then synthesized by ATR-MATRIX. Finally, the translated speech and speaker image are shown at the same time. We could realize a system that can process in real-time.

Table 4: System performances

DOA estimation rate (i.e., speaker image capturing rate)	97.8%
Speech recognition rate	90.0% (91.4%)
TOEIC score	530∼540 pts. (546 pts.)

( ) shows performance with closed talking microphone

In the future, we have to consider how to translate multi-lingual speech for multi-lingual tele-conferencing and how to estimate the speaker directions among the estimated DOAs. In addition, we also have to consider barge-in and high quality capture of speech in noisy reverberant environments. In this experiment, although speech recognition rate achieves 90.0%, these results are gained in a non-noisy environment. Therefore, we need to consider making the system more robust against noise. To increase robustness, we may have to use the image data captured by steering the video camera, and use a sharper beamformer like the multiple beamformer [3].

6. REFERENCES

[1] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms," J. Acoust. Soc. Am., Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.

[2] S.U. Pillai, "Array Signal Processing," Springer-Verlag, New York, 1989.

[3] J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," Speech Communication, Vol. 13, pp. 207–222, 1993.

[4] L.J. Griffiths, and C.W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beam-forming," IEEE Trans. AP, Vol. AP-30, No. 1, pp. 27–34, 1982.

[5] C.H. Knapp, and G.C. Carter, "The Generalized Correlation Method for Estimation of Time Delay, " IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-24, pp. 320–327, 1976.

[6] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of Multiple Sound Sources Based on a CSP Analysis with a Microphone Array," Proc. ICASSP2000, pp. 1053–1056, Jun. 2000.

[7] A. Nakamura, et. al., "A Speech Translation System Applied to a Real-World Task/Domain and Its Evaluation Using Real-World Speech Data," IEICE Trans. Inf & Syst., Vol. E84-D, No. 1, pp. 142–154, Jan. 2001.

[8] F. Sugaya, et. al., "Evaluation of the ATR-MATRIX Speech Translation System With a Pair Comparison Method Between the System and Humans," Proc. ICSLP2000, pp. 1105–1108, Oct. 2000.