Advances in Multi-speaker Conversational Speech Recognition and Understanding

Takaaki Hori, Shoko Araki, Tomohiro Nakatani, and Atsushi Nakamura

Abstract

Opportunities have been increasing in recent years for ordinary people to use speech recognition technology. For example, we can easily operate smartphones using voice commands. However, attempts to construct a device that can recognize human conversation have produced unsatisfactory results in terms of accuracy and usability because current technology is not designed for this purpose. At NTT Communication Science Laboratories, our goal is to create a new technology for multi-speaker conversational speech recognition and understanding. In this article, we review the technology we have developed and present our meeting analysis system that can accurately recognize *who spoke when, what, to whom, and how* in meeting situations.

Keywords: multi-speaker, speech recognition, diarization

1. Introduction

A meeting is a basic human activity in which a group of people share information, present opinions, and make decisions. In formal meetings, it is standard for one person to take minutes. However, it often happens that certain important details are forgotten and therefore not recorded in the minutes. Moreover, meetings are not always easy to control, and this sometimes makes it difficult to achieve the objectives of the meeting. The participants may also be ill-informed, which can lead to misunderstandings or disagreements. Consequently, technology that is capable of automatically recognizing and understanding speech used in meetings has been attracting increasing attention [1], [2] as a way to overcome such problems.

Today, speech recognition technology is widely used in many applications such as the operation of smartphones using voice commands. If we speak clearly into such a device, the spoken words can be recognized correctly and the command executed as intended. However, when we try to construct a device that can be applied to recognize conversations in meetings, as many as half of the words are not recognized correctly. This is because speech signals are often degraded by background noise and the voices of other participants, and conversational speech itself involves a wide variety of acoustic and linguistic patterns compared with speech directed at a device. As a result, the speech recognition accuracy deteriorates significantly. At NTT Communication Science Laboratories, we are working hard to develop meeting speech recognition technology that can solve these problems.

However, another problem is that even if a speech recognizer achieves 100 percent accuracy for a meeting, no information about the meeting will be provided except for the spoken word sequence in a text format. This means that we can understand what words were spoken in the meeting but not who spoke when, to whom, and in what manner, which are all important pieces of information if we are to understand any meeting. In our research group, we are also

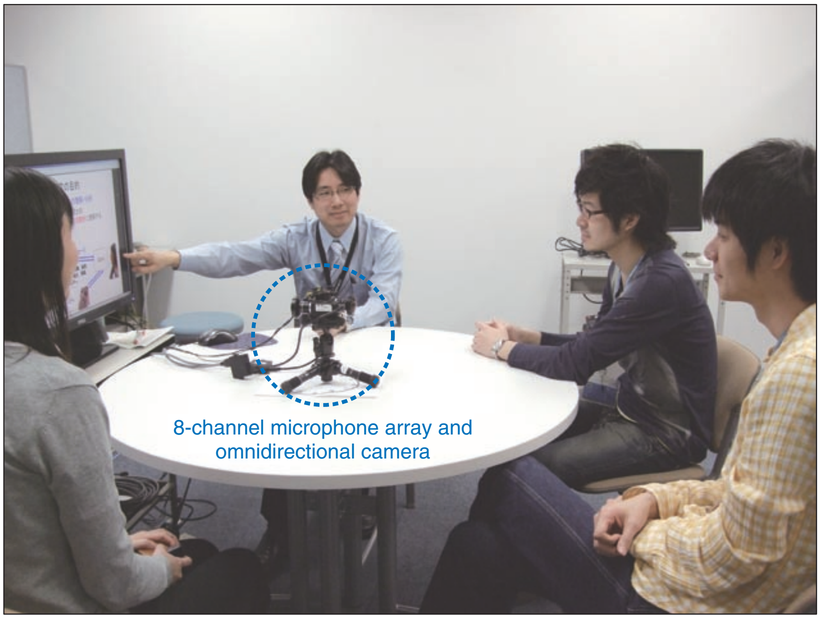


Fig. 1. Image of meeting captured by camera and microphone array.

studying meeting analysis technology that will enable us to understand a meeting in its entirety [2]. Our aim is to create a system that simultaneously obtains verbal information by speech recognition and nonverbal information by audio-visual scene analysis.

We have already developed a prototype system for meeting analysis, which we designed to evaluate and demonstrate our proposed techniques. The first version of the system visualized a meeting based on nonverbal information, where the system recognized *who spoke when and to whom* and estimated the visual focus of attention using a microphone array and an omnidirectional camera [1]. We then extended the system to recognize both verbal and nonverbal information by incorporating our meeting speech recognition technology [2]. We have already shown that the system can both create draft meeting minutes and assist meeting participants with functions for looking back at past utterances and accessing information related to the words spoken during the meeting.

In this article, we review the meeting speech recognition and understanding technology we have developed. In section 2, we describe our attempts to improve meeting speech recognition. In section 3, we present our meeting analysis system that accurately recognizes *who spoke when, what, to whom, and how*. We conclude the article and touch on future work in section 4.

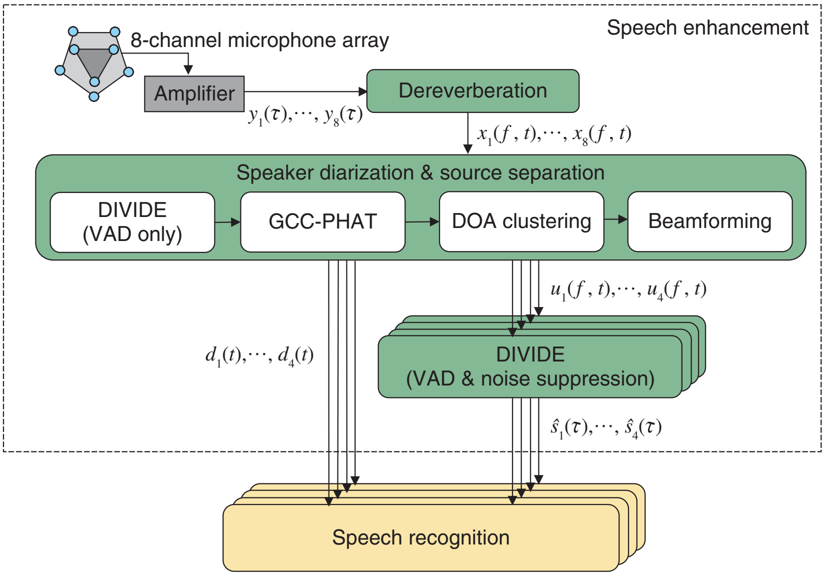
2. Recognition of meeting speech

2.1 Problems with meeting-speech recognition

We consider an ordinary meeting room as shown in **Fig. 1**, where four meeting participants freely discuss various topics, and all the utterances are recorded with a microphone placed at the center of the table. However, speech recognition is not easy in this situation, and the recognition result will include many errors. There are two reasons for this problem, as described below.

(1) Conversation oblivious to microphones

In face-to-face meetings, having participants wear a microphone or placing a microphone directly in front of each participant is not the preferred approach because it severely restricts the movement of the participants. Therefore, microphones should be located further away from each participant. However, this results in interference from acoustic noise and reverberation. Moreover, in settings where participants engage in informal and relaxed conversation, the utterances of two or more speakers often overlap. These factors significantly degrade speech recognition accuracy.


Speech enhancement
8-channel microphone array
Amplifier
$y_1(\tau),\cdots,y_8(\tau)$
Dereverberation
$x_1(f,t),\cdots,x_8(f,t)$
Speaker diarization & source separation
DIVIDE (VAD only)
GCC-PHAT
DOA clustering
Beamforming
$u_1(f,t),\cdots,u_4(f,t)$
$d_1(t),\cdots,d_4(t)$
DIVIDE (VAD & noise suppression)
$\hat{s}_1(\tau),\cdots,\hat{s}_4(\tau)$
Speech recognition

GCC-PHAT: generalized cross-correlation with phase transform
VAD: voice activity detection

Fig. 2. Speech enhancement system.

(2) Acoustic and linguistic variety of spontaneous speech

In private or informal meetings, people rarely speak formally and clearly. From acoustic and linguistic points of view, the utterances are fully spontaneous and therefore tend to include ambiguous pronunciations, abbreviations, and dialectal and emotional expressions. Consequently, a wide variety of speech patterns exist even for words that have the same dictionary pronunciation. These patterns change greatly depending on the speaker, the speaking style, and the topic. This aspect of spontaneous speech is also a crucial problem that degrades recognition accuracy.

2.2 Solutions to problems

To solve these problems, we first worked on speech enhancement to improve the quality of speech signals in a meeting and proposed effective techniques based on a microphone array [2]. An overview of the speech enhancement method used in our meeting recognition system is shown in **Fig. 2**. The speech enhancement process consists of three phases: dereverberation, speaker diarization/source separation, and noise suppression.

(1) The dereverberation phase transforms the eight-channel microphone signals $y_1(\tau)\cdots y_8(\tau)$ into the time-frequency domain, removes reverberation components from the complex spectral sequences $y_1(f,\ t)\cdots y_8(f,\ t)$ of the microphone signals using multi-channel linear prediction, and outputs eight-channel dereverberated spectral sequences $x_1(f,\ t)\cdots x_8(f,\ t)$.

(2) The speaker diarization/source separation phase detects active speakers based on direction of arrival (DOA) information, which is estimated by applying the generalized cross-correlation method with phase transform (GCC-PHAT) [3] to the dereverberated speech spectral sequences $x_1(f,\ t)\cdots x_8(f,\ t)$. In our system, the diarization result is obtained by clustering the DOAs at each frame, and the utterance period for each speaker n is output as the binary speaker diarization results $d_1(t)\cdots d_4(t)$, where $d_n(t) = 1$ (or 0) indicates that speaker n is speaking (or silent) at frame t. Subsequently, source separation is performed to separate overlapping speech into speaker-dependent channels, where a null-beamforming approach is employed because it does not produce nonlinear artifacts that have detrimental effects on speech recognition. The beamformer coefficients for each speaker are estimated by leveraging the diarization result $d_n(t)$ [4].

(3) The noise suppression phase suppresses the noise components contained in each separated spectrum $u_n(f, t)$, where we use Dynamic Integration of Voice Identification and DE-noising (DIVIDE) [5], [6]. DIVIDE reduces only the noise component of the original signals by using the online estimation of the speech and noise components, and then it outputs time-domain enhanced speech signals $\hat{s}_n(\tau)$.

After the speech enhancement, speech recognition is performed by using the enhanced speech signals $\hat{s}_n(\tau)$, in which the diarization results $d_n(t)$ are also used to validate whether or not each recognized word is actually spoken by the participant associated with the separated channel.

We recently improved the speech enhancement further by using DOLPHIN [7], which extracts the target speech more clearly based on the acoustic patterns of speech in the time and frequency domains. Although this method does not currently work with online processing, we confirmed that it yielded a large gain in recognition accuracy of meeting speech.

2.3 Automatic speech recognition module

Next we present a brief overview of the automatic speech recognition (ASR) module that we designed for transcribing meeting speech. The module is based on SOLON [8], a speech recognizer that employs weighted finite-state transducers (WFSTs). SOLON employs an acoustic model consisting of a set of hidden Markov models (HMMs), a pronunciation lexicon, and language models represented as WFSTs that can be combined *on the fly* (i.e., as quickly as necessary) during decoding. The decoder efficiently finds the best scoring hypothesis in a search space organized with the given WFSTs.

The input signal to the ASR module is spontaneous speech uttered by meeting participants, recorded with distant microphones, and enhanced by the audio processing techniques shown in Fig. 2. In general, it is effective to use a large amount of meeting speech data and their transcriptions to train acoustic and language models. However, there are no available Japanese data recorded under similar conditions, and it is very costly to collect new meeting data. Therefore, we prepared only a small amount of matched-condition data and used them to adapt the acoustic and language models.

The acoustic model is a set of state-shared triphone HMMs, where each triphone (a sequence of three phonemes) is modeled as a left-to-right HMM with three states, and each shared state has a Gaussian mixture output distribution. First, initial HMMs are trained with a large corpus of clean speech data recorded via a close-talking microphone. The parameters of the initial HMMs are then estimated by discriminative training based on a differenced maximum mutual information (dMMI) criterion [9] to reduce recognition errors.

Next, the initial HMMs are adapted with a small amount of real meeting data, which were recorded and enhanced with our meeting recognition system in advance. That is, the data were recorded with an 8-channel microphone array, and then dereverberated, separated, and subjected to noise suppression using the techniques in Fig. 2. The adaptation was performed by using maximum likelihood linear regression with automatically obtained multiple regression matrices [10].

We employ two types of language models. One is a standard back-off *n*-gram model. As with acoustic modeling, it is difficult to obtain a meeting transcript that is large enough to estimate the *n*-gram model. We use several types of data sets including a large written-text corpus and a small meeting transcript, and combine them with different weights based on an Expectation-Maximization algorithm. The other is a discriminative language model (DLM) trained with the R2D2 criterion [11]. This criterion is effective for training a language model that directly reduces recognition errors in a baseline speech recognizer.

The decoder is based on efficient WFST-based one-pass decoding [8] in which fast on-the-fly composition can be used for combining WFSTs such as HCLG_1 and $\mathrm{G}_{3/1}$ during decoding, where HCLG_1 represents a WFST that transduces an HMM state sequence into a unigram-weighted word sequence, and $\mathrm{G}_{3/1}$ represents a WFST that weights a word sequence with the trigram probabilities divided by the unigram probabilities. This division is necessary to cancel out the unigram probabilities already contained in HCLG_1.

Since the algorithm can handle any number of WFSTs for composition on the fly, we combine four WFSTs, including two WFSTs that represent a DLM in the one-pass decoding. The first DLM WFST is based on word features and the second is based on part-of-speech (POS) features. Since a DLM can be represented as a set of word or POS *n*-grams with certain weights, it can be transformed into a WFST in the same way as a standard back-off *n*-gram model. The two DLM WFSTs for the word and POS features can be combined linearly by the composition operation during decoding. Thus, we can achieve one-pass real-time speech recognition using these DLM WFSTs unlike the conventional approaches that involve a rescoring step after the first-pass decoding.

As part of the recent advances made in speech recognition technology in our research group, we have proposed an all-in-one speech recognition model [12], which is a different approach from those in which acoustic and language models are separately trained. The all-in-one model is represented as a WFST (or a set of WFSTs) including all the acoustic, pronunciation, and language models, and it is effectively trained using a discriminative criterion to reduce the number of recognition errors. As a result, the model can capture not only the general characteristics of acoustic and linguistic patterns but also a wide variety of interdependencies between acoustic and linguistic patterns in conversational speech. With this model, we have improved the speech recognition accuracy for meeting speech, and have increased the accuracy substantially by integrating the model with deep learning techniques [13].

2.4 Diarization-based word filtering

Finally, we describe diarization-based word filtering (DWF), which is an important technique used with our meeting speech recognizer. As shown in Fig. 2, we employed speech recognition for each speaker's channel given by source separation. This approach is much more robust in the case of overlapping speech than that using only a single channel. However, even if we employ source separation, it is difficult to completely remove other speakers' voices from the target speaker's channel. Such remaining non-target speech signals often induce insertion errors in speech recognition. To solve this problem, we utilize frame-based speaker diarization results obtained using the method shown in Fig. 2 to reduce the number of insertion errors. Since the diarization result at each frame tends to be discontinuous on the time axis, we use the average value of the diarization results for each recognized word, which can be considered to represent the relevance of the word to the target speaker. With this measure, words with a low relevance can be effectively deleted from the recognition results.

The relevance of word w_n recognized for speaker n is computed as:

$$s(w_n) = \frac{1}{e(w_n) - b(w_n) + 1} \sum_{t = b(w_n)}^{e(w_n)} d_n(t)$$

where $d_n(t)$ is the frame-based diarization result at frame t, and $b(w_n)$ and $e(w_n)$ respectively indicate the beginning and ending frames of word w_n. If $s(w_n)$ is less than a predefined threshold, w_n is deleted from the speech recognition result. This method is effective for low-latency processing. Conventional methods detect a speech segment by speaker diarization, and then apply speech recognition for the segment. This requires a long time delay because speech recognition cannot start until the diarization step is finished. The DWF approach only requires a one-frame (32 msec) delay to obtain $d_n(t)$, and it can drive speech recognition in parallel.

3. Real-time meeting analysis system

Our meeting analyzer basically recognizes *who is speaking what* by using speech recognition and speaker diarization, and it detects the activity of each participant (e.g., speaking, laughing, looking at someone else) and the circumstances of the meeting (e.g., topic, activeness, casualness, and intelligibility (defined more specifically below)) by integrating results obtained from several processing modules. The detected results provide *speaking to whom and how* information. The results of analysis are continuously displayed on a browser running on an Android™* tablet, as shown in **Fig. 3**.

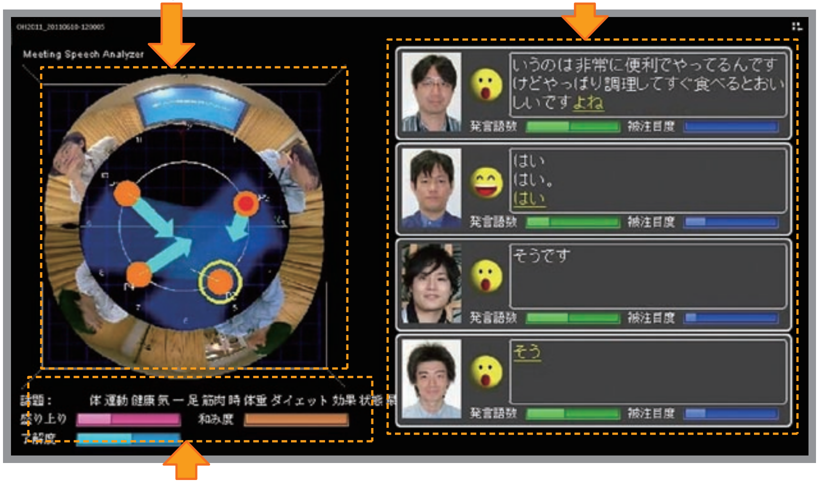
The panel on the left side of the browser displays live streaming video of a 360-degree view provided by a camera together with information about *who is speaking to whom* represented by orange circles and light blue arrows. The right side shows the real-time transcript for each participant, as well as his/her picture. The face icon beside the transcript indicates the participant's state (speaking, laughing, or silent), and the next two bars show the number of words spoken by the participant and how much he/she has been watched by others, i.e., the visual focus of attention for each speaker, based on the direction of each participant's face.

The lower left panel shows the current circumstances of the meeting in terms of topic words and the degrees of activeness, casualness, and intelligibility. Activeness is calculated as the number of words spoken multiplied by the entropy based on the relative frequencies of spoken words for all the speakers in a fixed time window. Casualness is estimated based on the frequency of laughter. Intelligibility is calculated based on the frequency of participants' nods. These

* Android™ is a trademark of Google.

Live streaming video of a 360-degree view from the camera together with information on *who is speaking to whom*

Real-time transcript for each participant and his/her current state



いうのは非常に便利でやってるんです
けどやっぱり調理してすぐ食べるとおい
しいですよね

はい
はい。
はい

そうです

そう

Current situation of the meeting
(topic words, activeness, casualness, and intelligibility)

Fig. 3. Real-time meeting browser working with Android™.

graphical representations help us to understand the current circumstances of the meeting both visually and objectively.

The system architecture is depicted in **Fig. 4**. In the speech enhancement block, the *m*-th microphone signal in the STFT domain $y_m(f, t)$ ($m = 1, \cdots, 8$) is dereverberated, separated, and denoised, and finally the enhanced signals $\hat{s}_n(\tau)$ are sent to the speech recognizer and the acoustic event detector. Here, f and t are frequency and time-frame indices, respectively. Speech recognition is used for each separated signal to transcribe utterances. The speech recognizer SOLON is also used to detect acoustic events including silence, speech, and laughter. Sentence boundary detection and topic tracking are applied to the word sequences from SOLON. In topic tracking, we use the Topic Tracking Language Model [14], which is an online extension of latent Dirichlet allocation that can adaptively track changes in topics by considering the information history of the meeting. In our system, we use conditional random fields for sentence boundary detection [15], where the transition features consist of bigrams of the labels that identify the presence or not of a sentence head. The other features consist of words and their POS tags in the scope of a 3-word context and the pause duration at each boundary candidate.

The camera captures a 360-degree view from the center of the table. During visual processing, the faces of the participants are detected, and face images are sent to the browser at the beginning of the meeting. Then the face pose tracker continues to work during the meeting. This incorporates a Sparse Template Condensation Tracker [1], which realizes the real-time robust tracking of multiple faces by utilizing GPUs (graphics processing units). With this tracking approach, the position of each participant and his/her face direction can be obtained continuously. This visual information is used to determine *who is speaking to whom* and to detect the visual focus of attention. The position information is also used to associate each utterance with the corresponding participant by combining it with the DOA information of the speech signal. Accordingly, a transcript and a face icon can be displayed on the right panel for the participant who is speaking.

The meeting analysis module calculates the activeness and casualness of the meeting based on speech recognition and acoustic event detection results. The intelligibility is obtained based on the frequency of participants' nods detected by face-pose tracking. All the analysis results are sent to the browser together with streaming video via a real-time messaging protocol (RTMP) server. Since the RTMP server can receive multiple requests, the analysis results can be broadcast to multiple browsers.

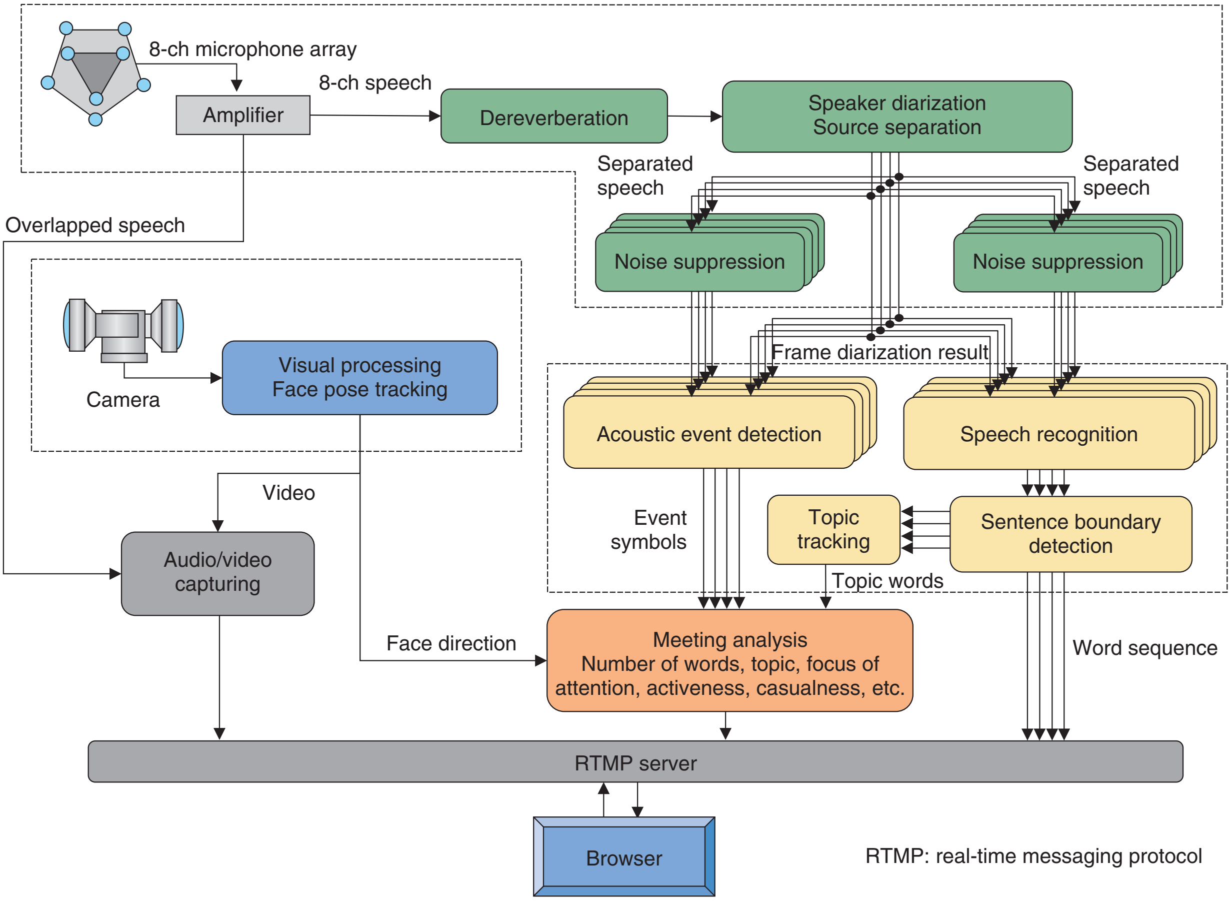


Fig. 4. Architecture of real-time meeting analysis system.

The current system runs on four computers: (1) an AMD Opteron 1389 2.9-GHz Quad Core for speaker diarization and source separation, (2) an Intel Xeon X5570 2.93 GHz 8-core dual processor for dereverberation and noise suppression, (3) the same Xeon model for speech recognition, acoustic event detection, and meeting analysis, and (4) an Intel Core 2 Extreme QX9650 3.0-GHz (+ NVIDIA GeForce-9800GX2/2 GPU cores) for visual processing.

4. Conclusion

In this article, we reviewed the technology we have developed in our research group and presented our meeting analysis system that provides accurate recognition of *who spoke when, what, to whom, and how* in a meeting situation. If conversational speech recognition and understanding based on audio-visual scene analysis becomes possible, many useful applications could be realized. In the future, it may be possible not only to generate meeting minutes automatically, but also to easily find past meeting scenes when required. We might have a virtual secretary who could answer our questions and register our plans in the scheduler autonomously. To realize such a system, it is important to improve speech recognition accuracy, and also to detect what is occurring in the surroundings, how the speaker is feeling, and why the meeting led us to a certain conclusion, etc. To enable such a deep understanding of human conversation, we need to extend the meeting analysis technology so that it can recognize higher-level concepts. To this end, we are continuing to address problems beyond the framework of speech recognition technology.

References

[1] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, “A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization,” Proc. of the 10th International Conference on Multimodal Interfaces (ICMI 2008),

pp. 257–264, Chania, Greece.

[2] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 2, pp. 499–513, 2012.

[3] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 24, No. 4, pp. 320–327, 1976.

[4] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "Speaker indexing and speech enhancement in real meeting/conversations," Proc. of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Vol. 1, pp. 93–96, Las Vegas, NV, USA.

[5] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual frontend processing method based on statistical model for noise robust speech recognition," Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009), pp. 1235–1238, Brighton, UK.

[6] H. Masataki, T. Asami, S. Yamahata, and M. Fujimoto, "Speech Recognition Technology That Can Adapt to Changes in Service and Environment," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa2.html

[7] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "LogMax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," Proc. of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 4029–4032, Kyoto, Japan.

[8] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 15, No. 4, pp. 1352–1365, 2007.

[9] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of MPE and MMI in margin and error space," Proc. of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), pp. 4894–4897, Dallas, TX, USA.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, Vol. 9, pp. 171–185, 1995.

[11] T. Oba, T. Hori, A. Nakamura, and A. Ito, "Round-robin duel discriminative language models," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1244–1255, 2012.

[12] Y. Kubo, S. Watanabe, T. Hori, and A. Nakamura, "Structural classification methods based on weighted finite-state transducers for automatic speech recognition," IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 8, pp. 2240–2251, 2012.

[13] Y. Kubo, T. Hori, and A. Nakamura, "Large vocabulary continuous speech recognition based on WFST structured classifiers and deep bottleneck features," Proc. of the 38th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 7629–7632, Vancouver, Canada.

[14] S. Watanabe, T. Iwata, T. Hori, A. Sako, and Y. Ariki, "Topic tracking language model for speech recognition," Computer Speech and Language, Vol. 25, No. 2, pp. 440–461, 2011.

[15] T. Oba, T. Hori, and A. Nakamura, "Sentence boundary detection using sequential dependency analysis combined with CRF-based chunking," Proc. of the 9th International Conference on Spoken Language Processing (INTERSPEECH 2006), pp. 1153–1156, Pittsburgh, PA, USA.

Takaaki Hori

Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and information engineering and the Ph.D. degree in system and information engineering from Yamagata University in 1994, 1996, and 1999, respectively. He joined NTT in 1999 and began researching spoken language processing at NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). He moved to NTT Communication Science Laboratories in 2002. He was a visiting scientist at the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2006 to 2007. He received the 22nd Awaya Prize Young Researcher Award from the Acoustical Society of Japan (ASJ) in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2012, and the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2013. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE.

Shoko Araki

Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.

She received the B.E. and M.E. degrees from the University of Tokyo in 1998 and 2000, respectively, and the Ph.D. degree from Hokkaido University in 2007. Since joining NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation (BSS) applied to speech signals, meeting diarization, and auditory scene analysis. She has been a member or had chairing roles in various committees and conferences, including ICA 2003, IWAENC 2003, EUSIPCO 2006, WASPAA 2007, and SiSEC 2008, 2010, and 2011. She received the 19th Awaya Prize Young Researcher Award from ASJ in 2001, the Best Paper Award of IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from IEICE in 2006, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2008. She is a member of ASJ, IEICE, and IEEE.

Tomohiro Nakatani

Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees from Kyoto University in 1989, 1991, and 2002, respectively. Since joining NTT as a researcher in 1991, he has been investigating speech enhancement technologies for developing intelligent human-machine interfaces. During 2005–2006, he was a Visiting Scholar at the Georgia Institute of Technology, Atlanta, GA, USA. Since 2008, he has been a Visiting Assistant Professor in the Department of Media Science, Nagoya University. He received the 1997 the Japanese Society for Artificial Intelligence Conference Best Paper Award, the 2002 ASJ Poster Award, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award. He has been a member of IEEE SP Society Audio and Acoustics Technical Committee (AASP-TC) since 2009 and a chair of the AASP-TC Review Subcommittee since 2013. He served as an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing during 2008–2011 and has chaired or co-chaired several committees and conferences, including the IEEE Kansai Section Technical Program Committee, IEEE WASPAA-2007, and the IEEE CAS Society Blind Signal Processing Technical Committee. He is a senior member of IEEE, and a member of ASJ and IEICE.

Atsushi Nakamura

Senior Research Scientist, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Dr.Eng. degrees from Kyushu University, Fukuoka, in 1985, 1987, and 2001, respectively. In 1987, he joined NTT, where he engaged in R&D of network service platforms, including studies on application of speech processing technologies to network services at Musashino Electrical Communication Laboratories. From 1994 to 2000, he was a Senior Researcher at Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan, where he was engaged in spontaneous speech recognition research, construction of a spoken language database, and development of speech translation systems. Since April 2000, he has been with NTT Communication Science Laboratories. His research interests include acoustic modeling of speech, speech recognition and synthesis, spoken language processing systems, speech production and perception, computational phonetics and phonology, and application of learning theories to signal analysis and modeling. He received the IEICE Paper Award in 2004, and twice received the Telecom-technology Award of The Telecommunications Advancement Foundation, in 2006 and 2009. He is a senior member of IEEE and serves as a member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee, and as the Chair of the IEEE Signal Processing Society Kansai Chapter. He is also a member of ASJ and IEICE.