論文 / 著書情報
Article / Book Information

Title	Toward Robust Speech Recognition and Understanding
Author	Sadaoki Furui
Journal/Book name	Test Speech and Dialogue (TSD 2003), Vol. , No. , pp. 2-11
発行日 / Issue date	2003, 9
DOI	
権利情報 / Copyright	The original publication is available at www.springerlink.com.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Proc. TSD2003, V.Matousek, P.Mautner(Eds.), Springer-Verlag, pp.2-11 (2003)

Toward Robust Speech Recognition and Understanding

Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

Abstract

This paper overviews robust architecture and modeling techniques for automatic recognition and understanding. The topics include robust acoustic and language modeling for spontaneous speech recognition, unsupervised adaptation of acoustic and language models, robust architecture for spoken dialogue systems, multi-modal speech recognition, and speech understanding. This paper also discusses the most important research problems to be solved in order to achieve ultimate robust speech recognition and understanding systems.

1. Introduction

The field of automatic speech recognition has witnessed a number of significant advances in the past 10-20 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming-based search methods [1][2][3][4].

Read speech and similar types of speech, e.g. that from reading newspapers or from news broadcast, can be recognized with accuracy higher than 90% using the state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech. This decrease is due to the fact that the acoustic and linguistic models used have generally been built using written language or speech from written language. Unfortunately spontaneous speech and speech from written language are very different both acoustically and linguistically [5]. Broadening the application of speech recognition thus crucially depends on raising the recognition performance for spontaneous speech. In order to increase the recognition performance for spontaneous speech, it is crucial to build acoustic and language models suited to spontaneous speech.

The principal cause of speech recognition errors is a mismatch between trained acoustic/linguistic models and input speech due to the limited amount of training data in comparison with the vast variation of speech. Figure 1 shows the main causes of acoustic as well as linguistic variation in speech [6]. It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and other aspects. Also important is for the systems to impose few restrictions on tasks and vocabulary. Developing automatic adaptation techniques is essential to resolve these problems. Adaptation techniques can be classified into supervised and unsupervised methods. Since unsupervised methods can use recognition data itself for adaptation, they are more flexible than supervised methods. However, unsupervised methods are usually more difficult to develop than supervised methods, especially for spontaneous speech having a relatively high recognition error rate.

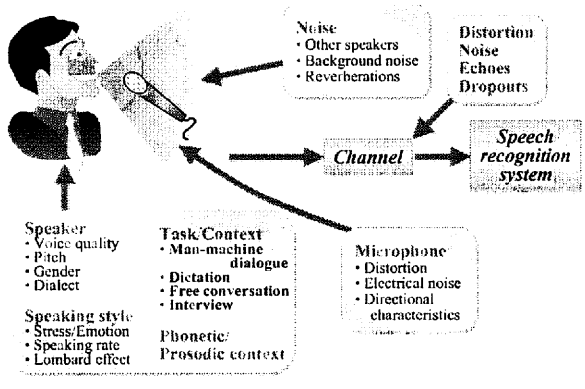

Fig. 1: Main causes of acoustic and linguistic variation in speech.

Providing spoken language interaction capability as a part of multimedia user interface is believed to add naturalness and efficiency to human-computer interaction. Most of the conventional dialogue systems are implemented by a system-initiative structure imposing constraints on the range and scope of allowed user inputs during an interaction. Since such systems are very troublesome for the users, mixed-initiative systems have also been investigated, in which the course of the dialogue can be changed by both the user and the system at any point [7]. These systems need to be able to accept and understand unrestricted utterances at any dialogue state. Such expansion automatically degrades not only the processing speed but also the performance of the system.

Multi-modal speech recognition, in which acoustic features and other information are used jointly, has been investigated and found to increase robustness and thus improve the accuracy of speech recognition. Most of the multi-modal speech recognition methods use visual features, typically lip information, in addition to the acoustic features [8].

Spontaneous speech is ill-formed and usually includes redundant information such as disfluencies, fillers, repetitions, repairs and word fragments. Therefore, recognition of spontaneous speech will require a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words [9].

The following chapters describe recent progress in increasing robustness of spontaneous speech recognition focusing on the major results of experiments that have been conducted at Tokyo Institute of Technology. The paper also discusses the most important research problems to be solved in order to achieve ultimate spontaneous speech recognition systems.

2. Spontaneous speech modeling

For building language models for spontaneous speech, large spontaneous speech corpora are indispensable. In this context, a Science and Technology Agency Priority Program entitled "Spontaneous Speech: Corpus and Processing Technology" started in Japan in 1999 [5]. The project will be conducted over a 5-year period under the following three major themes as shown in Fig. 2.

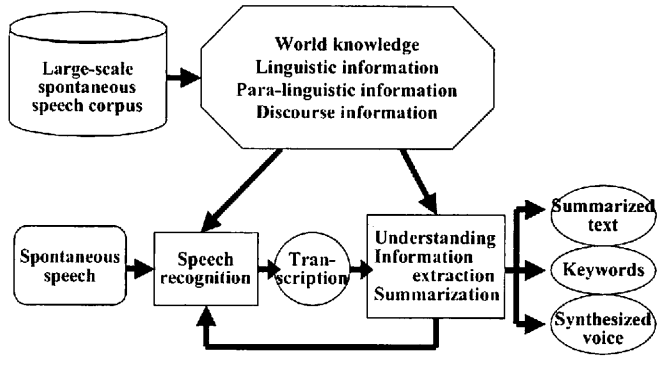

Fig. 2: Overview of the Japanese national project on spontaneous speech corpus and processing technology.

1) Building a large-scale spontaneous speech corpus. Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7M words with the total speech length of 700 hours. Mainly recorded will be monologues such as lectures, presentations and news commentaries. The recordings will be manually given orthographic and phonetic transcription. One-tenth of the utterances, hereafter referred to as the Core, will be tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program for automatically analyzing all of the 700-hour utterances. The Core will also be tagged with para-linguistic information including intonation.
2) Acoustic and linguistic modeling for spontaneous speech understanding using linguistic as well as para-linguistic information in speech.
3) Investigating spontaneous speech summarization technology.

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

Experimental results show that the mean recognition error for the spontaneous presentation utterances with the vocabulary size of 30k has become approximately 1/2 by replacing the acoustic as well as language models trained using read speech and presentation transcript with written language by the models made using the CSJ corpus [10].

Individual differences in spontaneous presentation speech recognition performances have been analyzed using 10 minutes from each presentation given by 51 male speakers, for a total of 510 minutes [11]. Seven kinds of speaker attributes have been considered in the analysis. They are word accuracy, averaged acoustic frame likelihood, speaking rate, word perplexity, out of vocabulary rate, filled pause rate and repair rate. It was found that, although all these attributes had correlation with the recognition performance, the attributes having real correlation with the accuracy were speaking rate, out of vocabulary rate, and repair rate.

3. Unsupervised adaptation of acoustic models

In many applications such as broadcast news and meeting speech transcription, speakers change frequently and each of them utters a series of several sentences. For these applications, we have proposed an incremental speaker adaptation method combined with automatic speaker-change detection [12]. In this method, the speaker change is detected using speaker-independent (SI) and speaker-adaptive (SA) Gaussian mixture models (GMMs). Both phone HMMs and GMMs are incrementally adapted to each speaker by the combination of Maximum Likelihood Linear Regression (MLLR), Maximum A Posteriori (MAP) and Vector Field Smoothing (VFS) methods using SI models as initial models. By selecting an initial model for speaker adaptation from a set of models made by speaker clustering, the adaptation performance can be improved [13]. This method corresponds to the piecewise linear approximation of the nonlinear effects of speaker-to-speaker variation in the cepstral domain.

Although the effect of additive noise on speech is linear in the waveform and spectral domain, it is nonlinear in the cepstral domain where speech is usually modeled for speech recognition. Therefore, nonlinear adaptation techniques, such as Parallel Model Combination (PMC, also called HMM composition) [14][15] and neural network-based mapping [16], have been investigated. Although these methods have been confirmed to be effective, they have a disadvantage in that they require a large amount of computation including non-linear conversion. In addition, they cannot guarantee the

likelihood maximization for each input speech, and therefore they cannot be used when noise is time varying and the noise effect needs to be compensated for each utterance.

The piecewise linear transformation (PLT)-based adaptation method described above has recently been successfully applied to solve these problems [17]. The PLT method consists of two parts: best-matching HMM selection and linear transformation of the selected HMM based on the maximum likelihood criterion. In order to reduce the cost of model selection for input speech, two methods are used. First, tree-structured noise-adapted HMMs are made by clustering noises or noisy speech, and model selection is performed by tracing the tree from the root to the leaves. Second, GMMs that correspond to the HMMs in the tree structure are made and used to select the best model instead of the HMMs. The HMM corresponding to the selected GMM is further adapted to match the input speech.

The proposed method has been evaluated using a dialogue system, in which two kinds of real noise were added to speech at three different SNR levels (5, 10 and 15dB). The noises differed from those used for creating noise-adapted HMMs and GMMs. Experimental results show that the proposed method with HMM-based and GMM-based model selection achieved error rate reductions of 36.1% and 33.0%, respectively.

As described in the previous section, one of the most important issues in spontaneous speech recognition is how to cope with the degradation of recognition accuracy due to speaking rate fluctuation. We have recently proposed an acoustic modeling for adjusting mixture weights and transition probabilities of an HMM for each frame according to the local speaking rate [18]. The proposed model implemented using the Bayesian network framework has a hidden variable representing variation of the "mode" of the speaking rate and its value controls the parameters of underlying HMM. Model training and maximum probability assignment of the variables were conducted using the EM/GEM and inference algorithms for the Bayesian networks. Utterances from meetings were used for evaluation in which the Bayesian network-based acoustic models were used to rescore the likelihood of the N-best hypotheses. In the experiments, the proposed model indicated consistently higher performance than conventional models.

4. Unsupervised adaptation of language models

An unsupervised, batch-type, class-based language model adaptation method for spontaneous speech recognition has been proposed. Figure 3 shows the overview of the proposed method [19]. Using many transcriptions in the training data set, a general language model (G-LM) consisting of word-based n-grams is built. Word classes approximately maximizing the average mutual information between classes are also made by applying a clustering algorithm, the "incremental greedy merging algorithm", to the training data set. The proposed adaptation method consists of the following three steps.

(1) Recognizing whole utterances using the G-LM,
(2) Training a class-based language model (C-LM) using the recognition results and the word-class information, and
(3) Obtaining an adapted language model (A-LM) by linearly interpolating the G-LM and the C-LM.

The proposed language model adaptation method was combined with an unsupervised acoustic model adaptation method as follows.

(1) Recognizing all utterances using the G-LM and a general speaker-independent acoustic model (G-AM),
(2) Building a speaker-adapted acoustic model (A-AM) by adapting the G-AM by the MLLR technique using the recognition results obtained in (1),
(3) Obtaining improved recognition hypothesis by re-recognizing the utterances using G-LM and A-AM,
(4) Building an A-LM by the language model adaptation method described above using the recognition hypotheses obtained in (3), and
(5) Re-recognizing the utterances using A-LM and A-AM.

Experimental results using spontaneous presentations show that this method is effective in improving the word accuracy and that the effects of acoustic and language model adaptation are additive.

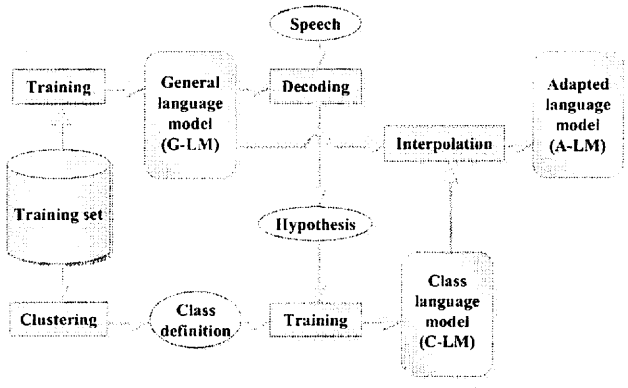

Fig. 3 : An overview of the unsupervised class-based language model adaptation method.

5. Robust architecture of spoken dialogue systems

We have proposed a new method of implementing mixed-initiative spoken dialogue systems based on parallel computing architecture [20]. In a mixed-initiative dialogue, the user as well as the system needs to be capable of controlling the dialogue sequence. In our implementation, various language models corresponding to different dialogue contents, such as requests for information or replies to the system, are built and multiple recognizers using these language models are driven under a parallel computing architecture. The dialogue content of the user is automatically detected based on likelihood scores given by the recognizers, and the content is used to build the dialogue. A transitional probability from one dialogue state uttering a kind of content to another state uttering a different content is incorporated into the likelihood score. A flexible dialogue structure that gives users the initiative to control the dialogue was implemented by this architecture. Real-time dialogue systems for retrieving information about restaurants and food shops were built and evaluated in terms of dialogue content identification rate and

keyword accuracy. The proposed architecture has the advantage that the dialogue system can be easily modified without remaking the whole language model.

6. Multi-modal speech recognition

We have proposed a new multi-modal speech recognition method using optical-flow analysis, as shown in Fig. 4, and evaluated its robustness to acoustic and visual noises [21]. Optical flow is defined as the distribution of apparent velocities in the movement of brightness patterns in an image. Since the optical flow is computed without extracting speaker's lip contours and location, robust visual features can be obtained for lip movements. Our method calculates a visual feature set in each frame consisting of maximum and minimum values of integral of the optical flow. This feature set has not only silence information but also open/close status of the speaker's mouth. The visual feature set is combined with an acoustic feature set in the framework of HMM-based recognition. Triphone HMMs were trained using the combined parameter set extracted from clean speech data.

Two multi-modal speech recognition experiments were carried out. First, acoustic white noise was added to speech waveforms, and a speech recognition experiment was conducted using audio-visual data from 11 male speakers uttering connected Japanese digits. The following improvements of relative reduction of digit error rate over the audio-only recognition scheme were achieved, when the visual information was incorporated into silence HMM: 32% at SNR=10dB and 47% at SNR=15dB. Second, a real-world data distorted both acoustically and visually was recorded in a driving car from six male speakers and recognized. We achieved approximately 17% and 11% relative error reduction compared with audio-only results on batch and incremental MLLR-based adaptation, respectively.

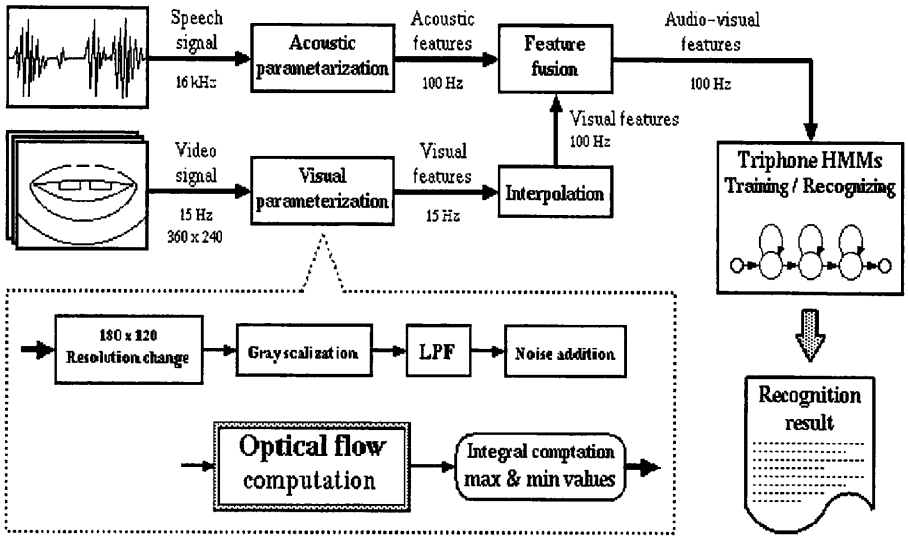

Fig. 4: Block diagram of the multi-modal speech recognition system using optical flow analysis.

When using this method in mobile environments, users need to hold a handset with a camera in front of their mouth at some distance, which is not only unnatural but also inconvenient for talking. Furthermore, the recognition accuracy may worsen due to the decreasing SNR. If the lip information can be taken by using a handset held in the usual way of telephone conversation, this would greatly improve its desirability. From this point of view, we have proposed an audio-visual speech recognition method using side-face images, assuming that a small camera is installed near the microphone of the mobile device [22]. This method captures the images of lips located of a small distance from the microphone. Visual features are extracted by optical-flow analysis and combined with audio features in the same way as the above method. Experiments conducted using Japanese connected digit speech contaminated with white noise in various SNR conditions show effectiveness of the proposed method.

7. Speech understanding

We have investigated techniques of automatic speech summarization as methods for realizing speech understanding, since a good summary can be considered as one of the representations of the essential meanings of the input utterance. We have proposed techniques for speech-to-text and speech-to-speech automatic summarization based on speech unit extraction and concatenation [23]. For the former case, a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction has been investigated. For the purpose of creating readable summaries, preserving as much important information as possible, removing speech recognition errors, and maintaining the meanings of the original sentences, sentence and word units which maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units are extracted from the speech recognition results and concatenated.

Figure 5 shows the two-stage speech-to-text summarization method consisting of important sentence extraction and sentence compaction [24]. Using speech recognition results, the score for important sentence extraction is calculated for each sentence. After removing all the fillers, a set of relatively important sentences is extracted, and sentence compaction using our proposed method [25] is applied to the set of extracted sentences. The ratios of sentence extraction and compaction are controlled according to a summarization ratio initially determined by the user.

These methods have been applied to summarization of unrestricted-domain spontaneous presentations and evaluated by objective and subjective measures. It was confirmed that proposed methods are effective in spontaneous speech summarization.

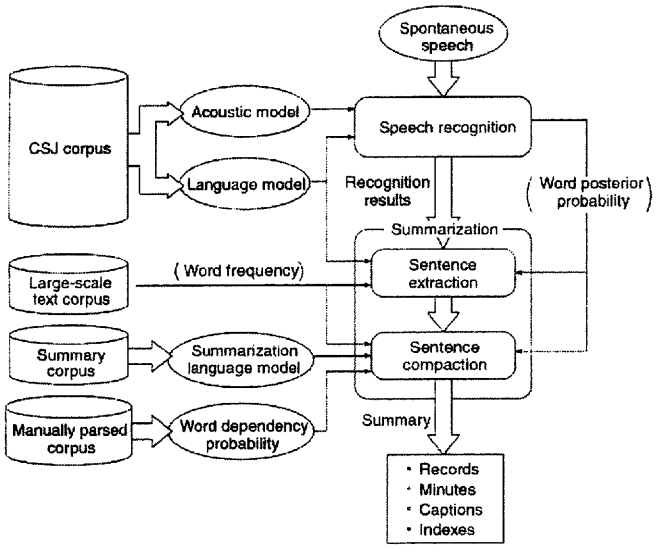

Fig. 5: Automatic speech-to-text summarization system.

8. Conclusion

The remarkable progress recently made in automatic speech recognition has enabled various application systems to be developed using transcription and spoken dialogue technology. While we are still far from having a machine that converses with a human like a human, many important scientific advances have taken place, bringing us closer to the "Holy Grail" of automatic speech recognition and understanding by machine [1]. Speech recognition and understanding will become one of the key techniques for human computer interaction in the multimodal/ubiquitous/wearable computing environment. To successfully use speech recognition in such an environment, every process such as start-stop control of recognition and adaptation to individuals and the surrounding environment must be performed without being noticed. Speech recognition should not be as it is in popular science fiction; instead it should be used unobtrusively, unconsciously and effortlessly. It is also necessary to operate in a consistent manner no matter where the user goes.

The most important issue is how to make the speech recognition systems robust against acoustic and linguistic variation in spontaneous speech. In this context, a paradigm shift from speech recognition to understanding, where underlying messages of the speaker, that is, meaning/context that the speaker intended to convey, are extracted, instead of transcribing all the spoken words, will be indispensable. To reach such a goal, we need to have an efficient way of representing, storing, retrieving, and utilizing world knowledge.

9. References

[1] B.-H. Juang and S. Furui, "Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication," *Proc. IEEE*, 88, 8, pp. 1142-1165, 2000

[2] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993

[3] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, 2nd Edition*, Marcel Dekker, 2000

[4] H. Ney, "Corpus-based statistical methods in speech and language processing," in *Corpus-based Methods in Language and Speech Processing, S. Young and G. Bloothooft Ed.*, Kluwer, pp. 1-26, 1997

[5] S. Furui, "Recent advances in spontaneous speech recognition and understanding," *Proc. IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, pp. 1-6, 2003

[6] S. Furui, "Steps toward natural human-machine communication in the 21st century," Proc. ISCA Workshop on Voice Operated Telecom Services, Ghent, pp. 17-24, 2000

[7] E. Levin et al., "The AT&T-DARPA COMMUNICATOR mixed-initiative spoken dialogue system," *Proc. ICSLP*, Beijing, pp. II-122-125, 2000

[8] S. Basu et al., "Audio-visual large vocabulary continuous speech recognition in the broadcast domain," Proc. IEEE Multimedia Signal Processing (MMSP), Copenhagen, pp. 475-481, 1999

[9] S. Furui, "Toward spontaneous speech recognition and understanding," in *Pattern Recognition in Speech and language Processing, W. Chou and B.-H. Juang Ed.*, CRC Press, pp. 191-227, 2003

[10] T. Shinozaki et al., "Towards automatic transcription of spontaneous presentations." *Proc. Eurospeech*, Aalborg, 1, pp.491-494, 2001

[11] T. Shinozaki and S, Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," *Proc. ICASSP*, Orlando, pp. I-729-732, 2002

[12] Z. Zhang et al., "On-line incremental speaker adaptation for broadcast news transcription," *Speech Communication*, 37, pp. 271-281, 2002

[13] Z. Zhang et al., "An online incremental speaker adaptation method using speaker-clustered initial models," Proc. ICSLP, Beijing, pp. III-694-697, 2000

[14] M. J. F. Gales et al., "An improved approach to the hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, San Francisco, pp. 233-236, 1992

[15] F. Martin et al., "Recognition of noisy speech by composition of hidden Markov models," *Proc. Eurospeech*, Berlin, pp. 1031-1034, 1993

[16] S. Furui et al., "Noise adaptation of HMMs using neural networks," *Proc. ISCA Workshop on Automatic Speech Recognition*, Paris, pp. 160-167, 2000

[17] Z. Zhang et al., "Tree-structured noise-adapted HMM modeling for piecewise linear-transformation-based adaptation," *Proc. Eurospeech*, Geneva, 2003

[18] T. Shinozaki and S. Furui, "Time adjustable mixture weights for speaking rate fluctuation," *Proc. Eurospeech*, Geneva, 2003

[19] Y. Yokoyama et al., "Unsupervised language model adaptation using word classes for spontaneous speech recognition," *Proc. IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, pp. 71-74, 2003

[20] R. Taguma et al., "Parallel computing-based architecture for mixed-initiative spoken dialogue," *Proc. IEEE Int. Conf. on Multimodal Interfaces (ICMI)*, Pittsburgh, pp. 53-58, 2002

[21] S. Tamura et al., "A robust multi-modal speech recognition method using optical-flow analysis," *Proc. ISCA Workshop on Multi-modal Dialogue in Mobile Environments*, Kloster Irsee, 2002

[22] T. Yoshinaga et al., "Audio-visual speech recognition using lip movement extracted from side-face images," *Proc. Eurospeech*, Geneva, 2003

[23] S. Furui et al., "Speech-to-speech and speech-to-text summarization," *Proc. Int. Workshop on Language Understanding and Agents for Real World Interaction*, Sapporo, 2003

[24] T. Kikuchi et al., "Two-stage automatic speech summarization by sentence extraction and compaction," *Proc. IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, pp. 207-210, 2003

[25] C. Hori et al., "A statistical approach to automatic speech summarization," *EURASIP Journal on Applied Signal Processing*, pp. 128-139, 2003