World-Wide Media Browser—Multilingual Audio-visual Content Retrieval and Browsing System

Takaaki Hori†, Katsuhito Sudoh, Hajime Tsukada, and Atsushi Nakamura

Abstract

This article describes the World-Wide Media Browser (WWMB) being developed at NTT Communication Science Laboratories. The aim of this system is to provide users with easy access to a large amount of multilingual audio/visual content via translingual techniques. To achieve this goal, we are integrating our latest research achievements in speech recognition, spoken language processing, utterance retrieval, and machine translation. In this article, we introduce a WWMB prototype that we have developed and review our speech and language technologies. We also report evaluation results for real lecture videos recorded at the Massachusetts Institute of Technology.

1. Introduction

In recent years, it has become much easier for people to listen to music and watch videos around the world through the Internet. Since many Internet video services provide movies and TV shows by using streaming technology, users can watch the videos they want to see anywhere and at any time. Furthermore, video-sharing sites such as YouTube™ not only provide a lot of videos, but also give end users the opportunity to publish their own videos. Such user-generated videos are being uploaded every day by a large but unspecified number of people, so the number of publicly available videos is increasing rapidly.

However, it is not easy for ordinary viewers to find a specific scene among these millions of videos. For example, in most video-sharing sites, the clue to finding a scene is only a few representative terms attached to each file by the file's contributor. When users try to find a desired scene, they type some keywords into the corresponding search engine, but they usually receive a long list of video files whose pre-attached terms match the keywords. Finally, they have to waste time playing each video to check if the target scene is included. At the same time, the wide range of languages is also a large problem. Although most people can access videos from all over the world, it is very difficult to find and watch videos in foreign languages even if the content is of interest because the keywords do not match any representative terms in foreign languages.

To solve these problems, we began to research multilingual audio-visual content retrieval and browsing technologies using simulated Japanese news programs [1] and developed a system called the World-Wide Media Browser (WWMB). In this system, we have integrated our research achievements in spontaneous speech recognition, discriminative language processing, utterance retrieval, and statistical machine translation, each of which is a cutting-edge technology in its research field. We think that the WWMB requires the following functions.

(1) Video content processing: accesses a lot of

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan

video content through the Internet, automatically transcribes the corresponding audio data, refines the resulting transcript, and translates it into other languages

(2) Scene retrieval: searches for specific video files and scenes in which a query word or phrase is spoken, represents the query in the user's preferred language, and plays back the videos starting from those scenes

(3) Subtitle generation: inserts subtitles in the user's preferred language while playing the video. The subtitles have already been generated in advance during video content processing.

If we successfully develop a full version of the WWMB, every user will be able to gain easy access to valuable content from around the world and also effectively deliver their own content to many more people, namely those who speak different languages.

Recently, the speech translation of audio/visual content has been actively investigated. For example, in the USA, the GALE project funded by the Defense Advanced Research Projects Agency (DARPA) began in 2005 as a national project. The aim of this project is to develop technologies for extracting various types of security information from Arabic and Chinese newspapers and broadcast news programs by using multilingual speech recognition and machine translation into English. Meanwhile, in Europe, the TC-Star project, which ran from 2004 to 2007, aimed to develop technology for translating between the languages used in the European Parliament Plenary Sessions and thus generate multilingual proceedings automatically.

Our research is similar to GALE and TC-Star. The difference is that we are trying to handle a wide range of real-world content on the web. Thus, our focus is on the more difficult tasks of recognition and translation. We also aim to study efficient methods specifically for Japanese, which is beyond the scope of other projects. In our work, we focus on Japanese use of the system, i.e., we want to enable users to find and watch non-Japanese language videos in Japanese, so we took approaches suitable for the Japanese language.

We have developed a prototype of the WWMB consisting of several subsystems that analyze speech and language content. These have been appropriately combined to perform the WWMB's functions. For all these subsystems, we used statistical approaches because they yield highly accurate results and are robust against the erroneous input that appears in real data.

In this article, we introduce the WWMB prototype and the technologies we used in this system. We also present evaluation results for the prototype that we obtained using real lecture videos recorded at the Massachusetts Institute of Technology (MIT). The lecture videos are publicly available on the MIT OpenCourseWare (OCW) web site and are used for speech and language research [2]. OCW is a worldwide consortium. The affiliated universities publish their lecture materials and recordings on corresponding web sites. Since the MIT lectures include a lot of speech and language content and their number is increasing rapidly, videos of these lectures are suitable for use in the development and evaluation of our system.

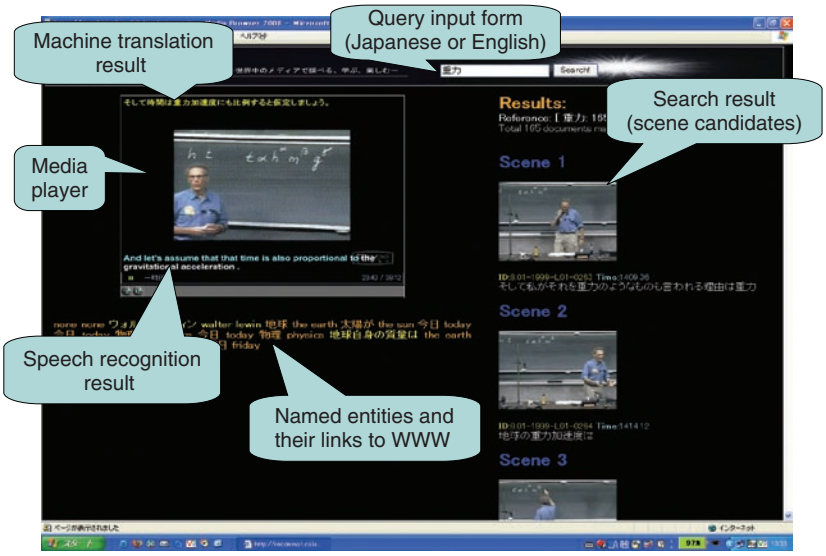
2. World-Wide Media Browser

A screen shot of our WWMB prototype is shown in **Fig. 1**. Some explanations (blue callouts) are superimposed on the picture to describe its interface. The prototype can currently handle Japanese and English content. By using the browser interface, for example, a Japanese user can find and watch videos in English with Japanese queries and subtitles. When the user enters a (Japanese) keyword or key phrase on the query input form, scene candidates that match the query are listed as the search results, showing where the query term or its (English) translation is spoken in each listed scene. The user can play a video that has a matching scene by selecting it from the list. The media player plays the video with multilingual (English and Japanese) subtitles, which have already been provided automatically by the speech recognition and machine translation subsystems. Thus, users can easily find, watch, and understand foreign-language videos in their own language.

The browser also provides a list of named entities (NEs) that are spoken in the video. NEs are keywords such as proper names. The user can check the meaning of the spoken NEs via their hyper links to a web search engine.

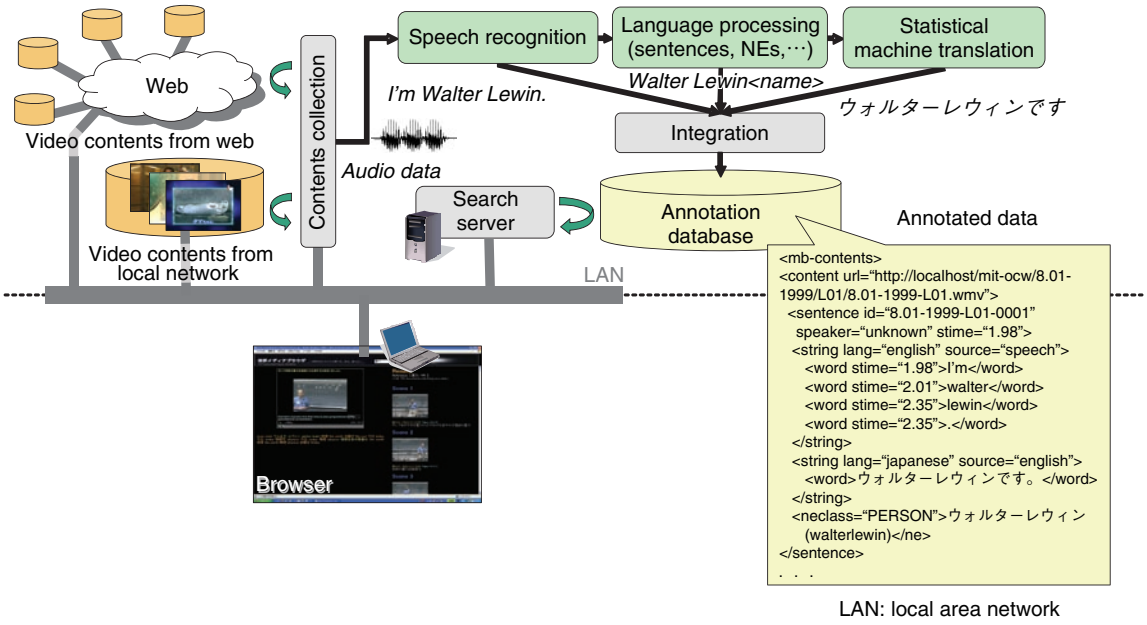
The WWMB assumes that all videos have been processed in advance by our technologies. Its content processing part, which consists of video content collection, speech recognition, language processing, and machine translation, is shown in **Fig. 2**. The language processing module includes sentence boundary detection and NE extraction from speech recognition results. The search server provides fast retrieval of video files and scenes.

Speech recognition and machine translation results



WWW: World Wide Web
Photos are from lectures by Walter Lewin, 8.01 Physics I: Classical Mechanics, Fall 1999. (Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu (Accessed Oct. 2007). License: Creative Commons BY-NC-SA

Fig. 1. Screen shot of World-Wide Media Browser.



LAN: local area network

Fig. 2. Content processing for World-Wide Media Browser.

are stored in the annotation database together with the timestamps for inserting subtitles during video playback. An index table is also constructed for the annotation data so that the content is searched for efficiently with complex queries. Important keywords (NEs, etc.) are also stored and used to characterize each video scene. These technologies are explained in more detail in the next section.

3. Core technologies in WWMB

The World-Wide Media Browser essentially requires highly accurate methods for speech recognition, language processing, and machine translation because a few speech recognition errors induce language processing errors, and these errors result in more errors in machine translation; that is, even a few errors have a big impact on the final system output. In addition, a large amount of data should be processed efficiently in a short computation time.

We have already studied some effective methods in related research fields. In this work, we built a WWMB prototype by integrating those methods. Below, we review our speech and language technologies used in the WWMB.

3.1 Fast and highly accurate speech recognition with extremely large vocabularies

Automatic speech recognition (ASR) technology has many applications including the control of consumer electronic devices, an interactive voice response interface in telephone services, and automatic generation of meeting minutes. To make ASR effective for various applications, we are working to improve speech analysis, model training, search, and backend processing algorithms.

In the WWMB, we use ASR to generate transcripts for subtitling, indexing, and translating the spoken content. In the prototype, we use the speech recognizer *SOLON* that we developed as a research platform in NTT Communication Science Laboratories.

Since the video content handled by the WWMB potentially includes a wide variety of topics from around the world, the ASR system needs to have a very large vocabulary. However, using a large vocabulary in ASR is computationally very expensive because an enormous number of hypotheses have to be evaluated.

ASR systems usually have a finite vocabulary. Speech recognition is a process that finds the most likely word sequence for the input speech by comparing it with reference speech models of all possible word sequences in the vocabulary. Accordingly, the ASR system inevitably misrecognizes some utterances that include out-of-vocabulary (OOV) words. To reduce the OOV words, the system should have a very large vocabulary.

We have developed a very fast algorithm that enables 10-million-word vocabulary realtime ASR using weighted finite-state transducers [3]. Our algorithm can greatly reduce the number of OOV words compared with conventional speech recognizers, which have at most 100,000-word vocabularies. Thus, our ASR has a great advantage over other systems.

On the other hand, we have already proposed efficient methods for estimating the parameters of speech models. Minimum classification error training is a method that directly improves the discrimination performance of models and thus reduces ASR errors [4]. Variational Bayesian estimation and clustering [5] is a robust parameter estimation method for speech models. It also yields automatic model-size selection.

These techniques are all implemented in *SOLON*. They have enabled *SOLON* to record the best performance in the benchmark test of the Corpus of Spontaneous Japanese.

3.2 Discriminative language processing

The language processing module in the WWMB detects sentence boundaries and extracts NEs. Sentence boundary detection (SBD) is indispensable for handling speech recognition results in the following language processing, such as machine translation, because most natural language processing methods are assumed to handle input text sentence by sentence. Since the punctuation that appears in written text does not exist in speech recognition results, the boundaries must be detected automatically. However, since sentence boundaries are usually ambiguous in spoken language, it is not easy to find the correct ones. We have already proposed an efficient SBD method that utilizes linguistic features around sentence boundaries and the consistency of the dependency structure of each sentence and the boundaries [6].

In this work, we used discriminative methods to improve the accuracy of SBD. A sentence boundary model was trained on the basis of conditional random fields and the dependency structure as a relative preference model was trained based on the maximum entropy principle. As a result, we achieved a great improvement in SBD accuracy compared with conventional SBD methods.

NE extraction identifies keywords from a document, such as proper names and expressions for dates, times, and quantities. NEs hold important information that is used for metadata, information extraction, and question answering. NE extraction can be regarded as the problem of classifying a word or a compound word into an NE class (person, location, date, time, etc.) or a *not-NE* class.

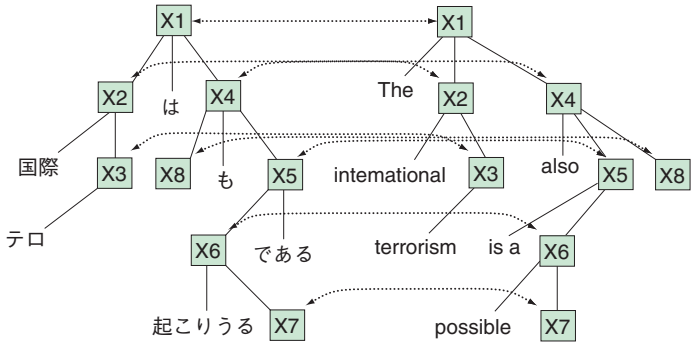


Fig. 3. Hierarchical phrase-based machine translation.

In the WWMB, we extract NEs from speech on the basis of speech recognition results, which may have errors. Since speech recognition errors degrade NE extraction performance, we predict speech recognition errors using the confidence of the speech recognition results and incorporate this prediction into the NE extraction process as a feature of support vector machines [7]. We can precisely extract NEs by ignoring NE-like words that are suspected to be speech recognition errors.

3.3 Hierarchical phrase-based statistical machine translation with a large number of features

The WWMB uses statistical machine translation (SMT). SMT enables us to build a translation system automatically from a bilingual corpus in which the same content is described in two different languages. SMT does not rely on language experts and enables us to build robust translation systems quickly and inexpensively if bilingual corpora are available. The SMT algorithm does not depend on the language, so translation systems for any language pairs can be constructed from the corresponding bilingual corpora using the same algorithm. This nature is highly suitable for the WWMB requirements because the WWMB has the potential to deal with various languages in the world.

Our SMT method uses a large number (several millions) of features based on surface forms as well as translation and language models [8]. The large number of features meant we could achieve fine-grained translations. As the translation model, we use a hierarchical phrase-based model, which is represented by weighted synchronous context-free grammar obtained automatically from bilingual corpora. An example of a translation obtained with this translation model is shown in **Fig. 3**. The hierarchical phrase-based model is especially effective for language pairs where the word order is very different such as English and Japanese because the translation model considers syntax and models the likelihood of word re-ordering as well as translation. Our method restricts the form of the weighted synchronous context-free grammar and enables us to achieve an effective translation process (decoding) with n-gram language models [9].

3.4 Open-vocabulary spoken utterance retrieval using confusion networks

The search server in the WWMB uses an inverted index to find relevant utterances quickly. An inverted index is a data structure in which each word is associated with a list of relevant files and the positions at which the word appears in each file. Once this index data has been constructed, the search result can be obtained very quickly by returning the list associated with the query word. If the query is a word sequence, i.e., a phrase, the results can also be obtained efficiently by first enumerating the utterances that include all the words in the query and then checking the proximities of the words in each enumerated utterance based on the word positions.

We have already proposed a novel indexing method using confusion networks (CNs) [10], which is robust against speech recognition errors. This method utilizes word CNs, which efficiently represent multiple ASR candidates. In our indexing, each arc of a CN is stored as a component of the associated list. The use of such multiple candidates means that more target utterances can be found than when only one best candidate (ASR result) is used because multiple candidates often include the correct word sequence even though the best candidate is not correct.

Furthermore, we have expanded our method to handle OOV words. Queries including OOV words

Table 1. Word error rate (WER) and BLEU score.

Lecture ID	WER%	BLEU	BLEU (TRS)
6.001-1986-L01	24.2	0.16	0.26
6.001-1986-L02	40.9	0.13	0.31
8.01-1999-L01	18.8	0.27	0.41

6.001: Computer Science, 8.01: Physics I

(OOV queries) do not match any utterances even if we use multiple candidates. To overcome this problem, we have incorporated word-phoneme combined CNs. This method can find the OOV words represented as a phoneme sequence in the combined CNs.

Currently, a CN cannot be generated from an SMT output. We plan to expand our retrieval method to handle multiple SMT candidates in the future.

4. Evaluation

We have evaluated the WWMB prototype from the viewpoint of ASR and SMT using MIT OpenCourseWare. We used 141 lectures for training the speech models and 39 lectures for training the translation models. The word error rate (WER) in ASR and the BLEU score in SMT for each test lecture are shown in **Table 1**. BLEU in the table denotes the scores for the translations of ASR results and BLEU (TRS) denotes the scores for those of correct transcripts, i.e., no ASR errors. The WER is the ratio of misrecognized words to spoken words, i.e., a smaller WER indicates a better recognition result. The BLEU score is a measure of the similarity between system translations and multiple reference translations created by human translators. The score is calculated simply by considering the matches of 1-N consecutive words. Usually N=4 is used. The BLEU score ranges from 0 to 1 and a higher score indicates a better translation result. In principle, even a correct translation cannot achieve a score of 1 since an identical translation may not be included in the references.

We have achieved good ASR results that are beyond the standard level of current ASR for real spontaneous lectures. Although lecture 6.001-1986-L02 had a WER of around 40%, this was caused by a lot of noisy talking by the audience. The error rate excluding such utterances was about 20%. Subtitles with a word-error level of 20% are acceptable in terms of enabling users to roughly understand what the lecturer said.

We have obtained standard BLEU scores comparable to those reported for other tasks. Although we have not achieved a sufficiently high quality translation for a person to understand the lecture, it could be used to help clarify the meaning of English captions. We can see that ASR errors have a significant impact on SMT performance by comparing the BLEU scores with the BLEU (TRS) scores. To achieve the translation level that we consider necessary with the WWMB, we should improve both ASR and SMT.

The BLEU scores in Table 1 were obtained by assuming that each sentence was correctly segmented in the ASR results because it is impossible to calculate the BLEU score unless the sentence boundaries are the same as those in human translations. Translation accuracy is degraded by errors in SBD as well as ASR. The accuracy of SBD was relatively low (68%) in this task and we must consider this problem further.

ASR errors cause SBD errors, and SBD errors cause SMT errors. This error chain can be frequently observed in the results. We need to enhance the ASR, SBD, and SMT methods and integrate them more closely so that they assist each other.

5. Conclusion

This article introduced the World-Wide Media Browser (WWMB) being developed at NTT Communication Science Laboratories. This system provides users with easy access to a large amount of multilingual audio/visual content by using translingual techniques. We evaluated the system using real lecture videos recorded at MIT and obtained good ASR and SMT results that could be used to help Japanese viewers understand the gist of English lectures. Future work involves improving ASR, SBD, and SMT, and closely coupling these techniques to further improve the WWMB functions.

Acknowledgment

We thank Professor Jim Glass at MIT for providing us with the lecture corpus.

References

[1] T. Hori, K. Sudoh, T. Oba, S. Watanabe, T. Watanabe, H. Tsukada, and A. Nakamura, "World-Wide Media Browser—the multilingual audio/visual contents retrieval and browsing system based on speech recognition and statistical machine translation," Proc. of 2nd Workshop on Spoken Document Processing, pp. 59–66, 2008 (in Japanese).

[2] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," Proc. of Interspeech'07, pp. 2553–2556, Antwerp, Belgium, 2007.

[3] 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. ASLP, Vol. 15, No. 4, pp. 1352–1365, 2007.
[4] E. McDermott and A. Nakamura, "String and Lattice Based Discriminative Training for the Corpus of Spontaneous Japanese Lecture Transcription Task," Proc. of Interspeech'07, pp. 2081–2084, Antwerp, Belgium, 2007.
[5] S. Watanabe, A. Sako, and A. Nakamura, "Automatic determination of acoustic model topology using variational Bayesian estimation and clustering for large vocabulary continuous speech recognition," IEEE Trans. ASLP, Vol. 14, No. 3, pp. 855–872, 2006.
[6] T. Oba, T. Hori, and A. Nakamura, "Sentence Boundary Detection Using Sequential Dependency Analysis Combined with CRF-based Chunking," Proc. of Interspeech'06, pp. 1153–1156, Pittsburgh, USA, 2006.
[7] K. Sudoh, H. Tsukada, and H. Isozaki, "Incorporating speech recognition confidence into discriminative named entity recognition of speech data," Proc. of COLING-ACL'06, pp. 617–624, Sydney, Australia, 2006.
[8] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "Online Large-margin Training for Statistical Machine Translation," Proc. of EMNLP-CoNLL'07, pp. 764–773, Prague, 2007.
[9] T. Watanabe, H. Tsukada, and H. Isozaki, "Left-to-right target generation for hierarchical phrase-based translation," Proc. of COLING-ACL'06, pp. 777–784, Sydney, Australia, 2006.
[10] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary Spoken Utterance Retrieval Using Confusion Networks," Proc. of ICASSP '07, Vol. IV, pp. 73–76, Honolulu, USA, 2007.

Takaaki Hori
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, Yamagata, in 1994, 1996, and 1999, respectively. Since 1999, he has been engaged in research on spoken language processing at NTT Cyber Space Laboratories. During 2006–2007, he was a visiting scientist at the Massachusetts Institute of Technology, Cambridge, USA. He received the 22nd Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2005. He is a member of IEEE, the Institute of Electronics, Information, and Communication Engineers (IEICE) of Japan, and ASJ.

Katsuhito Sudoh
Research Scientist, Knowledge Processing Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.
He received the B.E. and M.I. (Master of Informatics) degrees from Kyoto University, Kyoto, in 2000 and 2002, respectively. His research interests include spoken language processing and machine translation. He is a member of the Association for Computational Linguistics (ACL), ASJ, and the Association for Natural Language Processing.

Hajime Tsukada
Senior Research Scientist, Knowledge Processing Research Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.
He received the B.S. and M.S. degrees in information science from Tokyo Institute of Technology, Tokyo, in 1987 and 1989, respectively. He joined NTT Human Interface Laboratories in 1989. He joined ATR Interpreting Telecommunications Research Laboratories (ATR-ITL) in 1997. During 1998–1999, he stayed AT&T Laboratories Research as a visiting researcher. Since 2003, he has been with NTT Communication Science Laboratories. His research interests include statistical machine translation as well as speech and language processing. He is a member of ACL, IEICE, the Information Processing Society of Japan, and ASJ.

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. He joined NTT in 1987. During 1994–2000, he was with Advanced Telecommunications Research Institute International. Since April 2000, he has been with NTT Communication Science Laboratories. His research interests include speech signal processing, spoken language processing, and the application of learning theories to signal analysis and modeling. He is a senior member of IEEE and a member of IEICE and ASJ. He serves as a Vice Chair of the IEEE Signal Processing Society Kansai Chapter. He received the IEICE Paper Award and the Telecom-technology Award from the Telecommunications Advancement Foundation in 2004 and 2006, respectively.