Systems and Computers in Japan, Vol. 36, No. 10, 2005
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J87-D-II, No. 7, July 2004, pp. 1428–1440

Combining Outputs of Multiple LVCSR Models by Machine Learning

Takehito Utsuro,[1] Yasuhiro Kodama,[2] Tomohiro Watanabe,[3] Hiromitsu Nishizaki,[4]
and Seiichi Nakagawa[3]

[1]Department of Intelligence Science and Technology, Graduate School of Informatics,
Kyoto University, Kyoto, 606-8501 Japan

[2]Sony Corporation, Tokyo, 141-0001 Japan

[3]Department of Information and Computer Sciences, Toyohashi University of Technology, Aichi, 441-8580 Japan

[4]Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, 400-8511 Japan

SUMMARY

This paper proposes to apply machine learning techniques to the task of combining outputs of multiple LVCSR models, where, as features of machine learning, information such as the models which output the hypothesized word, its part-of-speech, and its syllable length are useful for improving the word recognition rate. Experimental results show that the combination result outperforms several baselines including model combination by voting such as ROVER in the word recognition rate. Furthermore, unlike model combination by voting, word recognition rate of model combination by machine learning is not damaged even in the case where only the minority of the participating models perform well in the word recognition task. © 2005 Wiley Periodicals, Inc. Syst Comp Jpn, 36(10): 9–15, 2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20340

Key words: LVCSR models; machine learning; combination of multiple models; SVM; confidence measures.

1. Introduction

Since current speech recognizers' outputs are far from perfect and always include a certain amount of recognition errors, it is quite desirable to have an estimate of confidence for each hypothesized word. This is especially true for many practical applications of speech recognition systems such as word selection for unsupervised adaptation schemes, automatic weighting of additional, nonspeech knowledge sources, keyword-based speech understanding, and recognition error rejection–confirmation in spoken dialogue systems.

Most previous works on confidence measures for LVCSR (large vocabulary continuous speech recognition)—such as those based on acoustic stability [6], number of edges in word graphs [13], hypothesis density [6], likelihood of acoustic/language models [11], and posterior probabilities [19]—are based on features available in a single LVCSR model. However, it is well known that a voting scheme such as ROVER (Recognizer Output Voting Error Reduction) for combining multiple speech recognizers' outputs can achieve word error reduction [2, 14]. Considering the success of a simple voting scheme such as

© 2005 Wiley Periodicals, Inc.

ROVER, it also seems quite possible to improve the reliability of previously studied features for confidence measures by simply exploiting more than one speech recognizer's outputs. From this observation, we experimentally evaluated the agreement among the outputs of multiple Japanese LVCSR models, with respect to whether it is effective as an estimate of confidence for each hypothesized word.

Our previous study [7] reported that the agreement between the outputs with two different acoustic models can achieve quite reliable confidence, and also showed that the proposed measure of confidence outperforms previously studied features for confidence measures such as the *acoustic stability* and the *hypothesis density* [6]. We also reported evaluation results with 26 distinct acoustic models and identified the features of acoustic models most effective in achieving high confidence [15]. The most remarkable results are as follows: for the newspaper sentence utterances, nearly 99% precision is achieved by decreasing 94% word correct rate of the best performing single model by only 7%. For the broadcast news speech, nearly 95% precision is achieved by decreasing 72% word correct rate of the best performing single model by only 8%. It is also shown that the confidence measure is useful in an unsupervised speaker adaptation framework [18].

Based on those results of our previous studies, this paper proposes to apply machine learning techniques to the task of combining outputs of multiple LVCSR models. As a machine learning technique, the Support Vector Machine (SVM) [17] learning technique is employed. A Support Vector Machine is trained for choosing the most confident one among several hypothesized words, where, as features of SVM learning, information such as the model IDs which output the hypothesized word, its part-of-speech, and the number of syllables are useful for improving the word recognition rate.

Model combination by high-performance machine learning techniques such as SVM learning has advantages over that by voting schemes such as ROVER and others [1, 2], especially when the majority of participating models are not reliable. In the model combination techniques based on voting schemes, outputs of multiple LVCSR models are combined according to simple majority vote or weighted majority vote based on confidence of each hypothesized word such as its likelihood. The results of model combination by those voting techniques can be harmed when the majority of participating models have quite low performance and output word recognition errors with high confidence. On the other hand, in the model combination by high-performance machine learning techniques such as SVM learning, among those participating models, reliable ones and unreliable ones are easily discriminated through the training process of machine learning framework. Furthermore, depending on the features of hypothesized words such as its part-of-speech and the number of syllables, outputs of multiple models are combined in an optimal fashion so as to minimize word recognition errors in the combination results.

Experimental results show that model combination by SVM achieves the following: for the newspaper sentence utterances, a relative word error reduction of 39% against the best performing single model and that of 23% against ROVER; for the broadcast news speech, a relative word error reduction of 13% against the best performing single model and that of 8% against ROVER. We further empirically show that it performs better when LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates.[1]

2. Specification of Japanese LVCSR Models

2.1. Decoders

As the decoders of Japanese LVCSR systems, we use one named Julius, which is provided by the IPA Japanese dictation free software project [5], as well as one named SPOJUS [4], which has been developed in our laboratory. Both decoders are composed of two decoding passes, where the first pass uses the word bigram, and the second pass uses the word trigram. Julius is with word-trellis searches and hence has much broader search space than SPOJUS, which is with N-best searches.

2.2. Acoustic models

The acoustic models of Japanese LVCSR systems are based on Gaussian mixture HMM. We evaluate phoneme-based HMMs as well as syllable-based HMMs.

2.2.1. Acoustic models with decoder Julius

As the acoustic models used with Julius, we evaluate phoneme-based HMMs [5] as well as syllable-based HMMs [10]. The number of Japanese phonemes for the phoneme HMMs is 43, while the number of Japanese syllables for the syllable HMMs is 124. The speech data are sampled at 16 kHz and 16 bits. The feature parameters consist of 25 dimensions: 12-dimensional mel frequency cepstrum coefficients (MFCC), the cepstrum difference coefficients (delta MFCC), and delta power are calculated every 10 ms. The following four types of HMMs are evaluated:

[1] Compared with our previous report [16], the major achievement of the paper is this empirical result. Reference 16 examined the correlation between each word's confidence and the word's features, and then introduced the framework of combining outputs of multiple LVCSR models by SVM learning.

Triphone model
Phonetic tied mixture (PTM) triphone model
Monophone model
Syllable model

Every HMM phoneme model consists of three states and is gender-dependent (male). The number of Gaussian mixtures of an HMM state with diagonal covariance matrices is 16 for the monophone, triphone, and syllable models, and 64 for the PTM triphone model. For each of the four models above, we evaluate both HMMs *with* and *without* the short pause state,[2] amounting to eight acoustic models in total.

2.2.2. Acoustic models with decoder SPOJUS

The acoustic models used with SPOJUS are based on syllable HMMs, which have been developed in Nakagawa's laboratory at Toyohashi University of Technology [12]. The number of Japanese syllables for the syllable HMMs is 116. The sampling frequencies are 12 kHz/16 kHz and the frame shift lengths are 8 ms/10 ms. The following three types of sets of feature parameters are evaluated:

- $2m$ dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames, delta $\boldsymbol{m}$ dimensions calculated over 9 successive frames, delta delta $\boldsymbol{m}$ dimensions, and delta, delta delta powers ($\boldsymbol{m} = \mathbf{10}, \mathbf{12}$).
- 12-dimensional mel frequency cepstrum coefficients (MFCC), delta 12 dimensions, delta delta 12 dimensions, and delta, delta delta powers.

The acoustic models are gender-dependent (male) syllable unit HMMs that have 5 states 4 densities, 4 Gaussian mixture models per density with full covariance/diagonal covariance matrices. We also switch between conventional HMM with self-loop transition and HMM with duration control, and evaluate both of them.

Among various combinations of features such as the sampling frequencies, frame shift lengths, feature parameters, covariance matrices, and self-loop transition/duration control, we carefully choose nine acoustic models so that they include the best performing ones as well as a sufficient number of minimal pairs which differ in only one feature. Then, as in the case of the acoustic models used with Julius, for each of the nine models, we evaluate HMMs both *with* and *without* the short pause states,[3] amounting to 18 acoustic models in total.

Table 1. Word recognition rates of LVCSR models (%)

newspaper sentence utterances		
decoder	word correct	word accuracy
Julius	93.0 (max) to 72.7 (min)	90.4 (max) to 69.4 (min)
SPOJUS	90.2 (max) to 78.1 (min)	85.3 (max) to 51.0 (min)
broadcast news speech		
decoder	word correct	word accuracy
Julius	71.7 (max) to 49.0 (min)	68.8 (max) to 39.7 (min)
SPOJUS	70.7 (max) to 55.4 (min)	62.8 (max) to 36.2 (min)

2.3. Language models

As the language models, the following two types of word bigram/trigram language models for 20k vocabulary size are evaluated:

- One trained using 45 months of Mainichi newspaper articles
- One trained using 5 years of Japanese NHK[4] broadcast news scripts (about 120,000 sentences)

2.4. Evaluation data sets

The evaluation data sets consist of newspaper sentence utterance, which is relatively easier for speech recognizers, and rather harder broadcast news speech:

1. 100 newspaper sentence utterances from 10 male speakers consisting of 1565 words, selected by IPA Japanese dictation free software project [5] from the JNAS (Japanese Newspaper Article Sentences) speech data [3].

2. 175 Japanese NHK broadcast news (June 1, 1996) speech sentences consisting of 6813 words, uttered by 14 male speakers (six announcers and eight reporters).

2.5. Word recognition rates

Word correct and accuracy rates of the individual LVCSR models for the above two evaluation data sets are measured, where for the recognition of the newspaper sentence utterances, the language model used is the one trained using newspaper articles, and for the recognition of the broadcast news speech, the language model used is the one

[2] When running the decoder with HMMs *without* the short pause state, we remove the powerless frames from the input speech, and use the language model trained without punctuation symbols (i.e., comma and period).

[3] The reason why we evaluate acoustic models both *with* and *without* the short pause states is that, from the preliminary evaluation result, this difference proved to be among those most effective in achieving high confidence.

[4] Japan Broadcasting Corporation.

trained using broadcast news scripts. Word recognition rates for the above two evaluation data sets are summarized in Table 1.

3. Combining Outputs of Multiple LVCSR Models by SVM

This section describes how to apply the SVM learning technique to the task of combining outputs of multiple LVCSR models considering the confidence of each word. We divide each of the data sets described in Section 2.4 into two halves, where one half is used for training and the other half for testing. Here, it is guaranteed that the two halves do not share speakers. An SVM is trained for choosing the most confident one among several hypothesized words from the outputs of the 26 LVCSR models. We used *SVM*^{light} (http://svmlight.joachims.org) as a tool for SVM learning. We compared linear and quadratic kernels and the linear kernel performs better. As features of the SVM learning, we use the model IDs which output the word, the part-of-speech of the word, and the number of syllables. Contribution of the parts-of-speech and the numbers of syllables was slight. We also evaluated the effect of acoustic and language scores of each hypothesized word as features of SVM, where their contribution to improving the overall performance was very little. As classes of the SVM learning, we use whether each hypothesized word is correct or incorrect. Since SVMs are binary classifiers, we regard the distance from the separating hyperplane to each hypothesized word as the word's confidence. The outputs of the 26 LVCSR models are aligned by Dynamic Time Warping, and the most confident one among those competing hypothesized words is chosen as the result of model combination. We also require the confidence of hypothesized words to be higher than a certain threshold, and choose those having confidence above this threshold as the result of model combination.

4. Experimental Results

The results of the performance evaluation against the test data are shown in Fig. 1. All the results plotted are the best performing ones among those for combining outputs of n ($3 \leq n \leq 26$) models. The results of model combination by SVM are indicated as "SVM." As a baseline performance, that of the best performing single model with respect to word correct rate ("Individual Model with Max Cor") is shown. (Note that their word recognition rates are those for half of the whole data set, and thus differ from those in Section 2.5.) For both speech data, model combination by SVM significantly outperforms the best performing single model. In terms of word accuracy rate, relative word error reduction is 39% for the newspaper sentence utterances and 13% for the broadcast news speech. Figure 1 also shows the performance of ROVER [2] as another baseline, where "Majority Vote" shows the performance of the strategy of outputting no word at a tie, while "Weighted Majority Vote" shows the performance when, for each individual model, word correct rate for each sentence is estimated and used as the weight of hypothesized words. Model combination by SVM mostly outperforms ROVER for both speech data. In terms of word accuracy rate, relative word error rate reduction is 23% for the newspaper sentence utterances and 8% for the broadcast news speech. Remarkable improvements are achieved especially in word accuracy rates. This is due to the strategy of requiring the confidence of hypothesized words to be higher than a certain threshold, where insertion error words tend to be discarded.

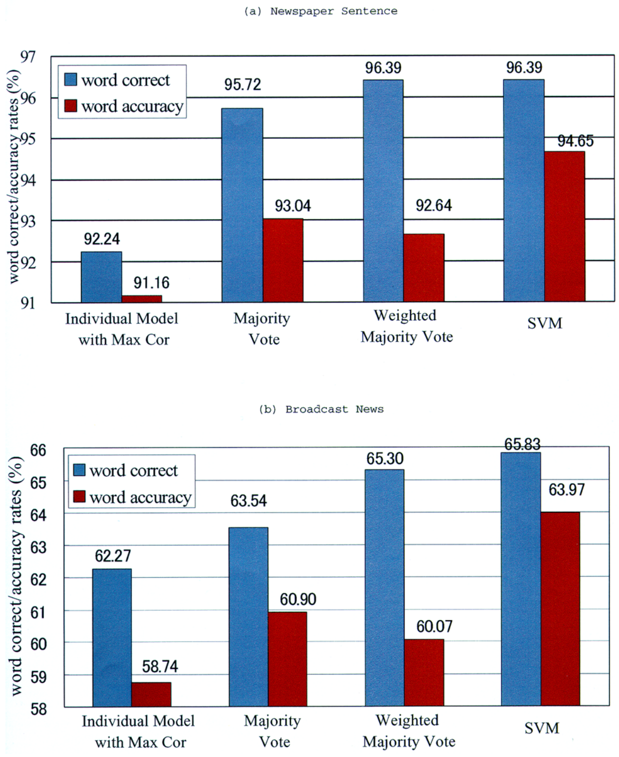

Fig. 1. Comparison among combination by SVM/(weighted) majority votes/individual models.

Figure 2 plots the changes of word accuracy rates against the increasing number of models which participate in LVCSR model combination. Here, LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates. (As we show later, the former outperforms the latter.) It is quite clear from this result that the difference of model combination by SVM and (weighted) majority votes becomes much larger

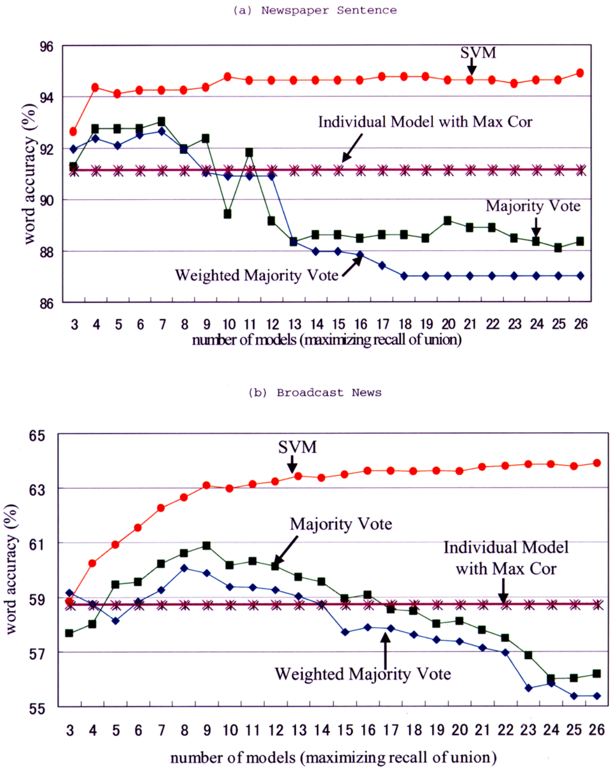

Fig. 2. Comparing methods for combining outputs of n ($3 \le n \le 26$) models.

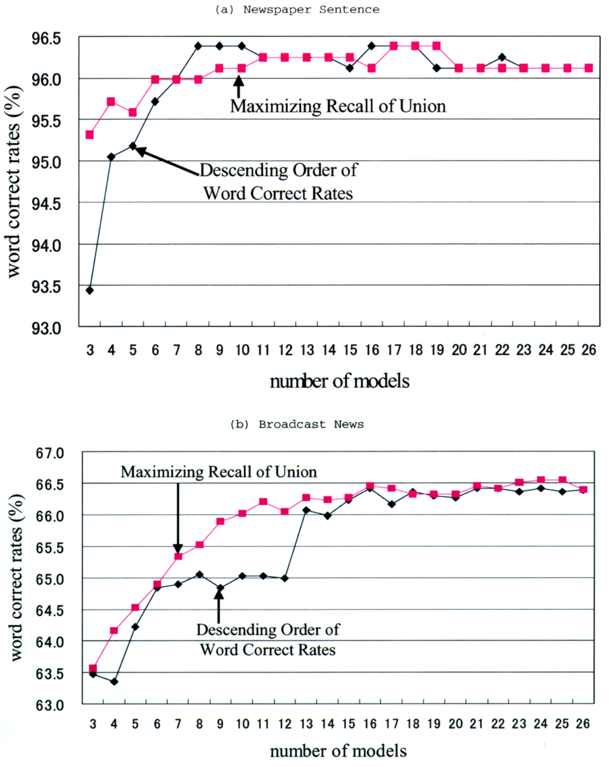

Fig. 3. Comparison between maximizing recall of union/descending order of word correct rates.

as more and more models participate in model combination. This is because the majority of participating models become unreliable in the second half of the curves in Fig. 2.

Figure 3 compares the model selection procedures, that is, choosing models so as to cover as many correctly recognized words as possible (indicated as "Maximizing Recall of Union"), and choosing models in descending order of their word correct rates (indicated as "Descending Order of Word Correct Rates"). The former performs better in the first half of the curves. This result indicates that, even if recognition error words increase in the outputs of models participating in LVCSR model combination, it is better to cover as many correctly recognized words as possible. This is because, in the model combination by high-performance machine learning techniques such as SVM learning, reliable and unreliable hypothesized words are easily discriminated through the training process.

5. Concluding Remarks

This paper proposed to apply machine learning techniques to the task of combining outputs of multiple LVCSR models. The proposed technique has advantages over those by voting schemes such as ROVER, especially when the majority of participating models are not reliable. In this machine learning framework, as features of machine learning, information such as the model IDs which output the hypothesized word are useful for improving the word recognition rate. Experimental results showed that the combination results achieved a relative word error reduction of up to 39% against the best performing single model and that of up to 23% against ROVER. We further empirically showed that it performed better when LVCSR models to be combined are chosen so as to cover as many correctly recognized words as possible, rather than choosing models in descending order of their word correct rates. The proposed technique has been proved to be effective in improving the performance of speech-driven WEB retrieval task [8, 9].

Acknowledgments. The authors thank members of Nakagawa's laboratory of the Department of Information and Computer Sciences, Toyohashi University of Technology. They also thank members of NHK Science Technical Research Laboratories for providing the broadcast news scripts and the broadcast news speech data.

REFERENCES

1. Evermann G, Woodland P. Posterior probability decoding, confidence estimation and system combina-

tion. Proc NIST Speech Transcription Workshop, 2000.

2. Fiscus JG. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). Proc IEEE Workshop on Automatic Speech Recognition and Understanding, p 347–354, 1997.
3. Itou K, Yamamoto M, Takeda K, Takezawa T, Matsuoka T, Kobayashi T, Shikano K, Itahashi S. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. Proc 5th ICSLP, p 3261–3264, 1998.
4. Kai A, Hirose Y, Nakagawa S. Dealing with out-of-vocabulary words and speech disfluencies in an N-gram based speech understanding system. Proc 5th ICSLP, p 2427–2430, 1999.
5. Kawahara T, Lee A, Kobayashi T, Takeda K, Minematsu N, Sagayama S, Itou K, Ito A, Yamamoto M, Yamada A, Utsuro T, Shikano K. Free software toolkit for Japanese large vocabulary continuous speech recognition. Proc 6th ICSLP, p 476–489, 2000.
6. Kemp T, Schaaf T. Estimating confidence using word lattices. Proc 5th Eurospeech, p 827–830, 1997.
7. Kodama Y, Utsuro T, Nishizaki H, Nakagawa S. Experimental evaluation on confidence of agreement among multiple Japanese LVCSR models. Proc 7th Eurospeech, p 2549–2552, 2001.
8. Matsushita M, Nishizaki H, Utsuro T, Kodama Y, Nakagawa S. Evaluating multiple LVCSR model combination in NTCIR-3 speech-driven WEB retrieval task. Proc 8th Eurospeech, p 1205–1208, 2003.
9. Matsushita M, Nishizaki H, Nakagawa S, Utsuro T. Keyword recognition and extraction by multiple-LVCSRs with 60,000 words in speech-driven WEB retrieval task. Proc 8th ICSLP, p 1625–1628, 2004.
10. Moroto M, Matsumoto H. Evaluation of Mel-LPC analysis by a large vocabulary Japanese dictation system. Proc 7th Western Pacific Regional Acoustics Conference, 2000.
11. Nakagawa S, Horibe Y. Confidence measures for speech recognition by using likelihood of acoustic model and language model. IPSJ SIG Notes, 2001-SLP-36, p 87–92. (in Japanese)
12. Nakagawa S, Yamamoto K. Evaluation of segmental unit input HMM. Proc 21st ICASSP, p 439–442, 1996.
13. Ogata J, Ariki Y. Improved speech recognition using iterative decoding based on confidence measures. Proc 7th Eurospeech, p 2577–2580, 2001.
14. Schwenk H, Gauvain J-L. Combining multiple speech recognizers using voting and language model information. Proc 6th ICSLP, p 915–918, 2000.
15. Utsuro T, Harada T, Nishizaki H, Nakagawa S. A confidence measure based on agreement among multiple LVCSR models—Correlation between pair of acoustic models and confidence. Proc 7th ICSLP, p 701–704, 2002.
16. Utsuro T, Kodama Y, Watanabe T, Nishizaki H, Nakagawa S. Confidence of agreement among multiple LVCSR models and model combination by SVM. Proc 28th ICASSP, p 16–19, 2003.
17. Vapnik VN. The nature of statistical learning theory. Springer-Verlag; 1995.
18. Watanabe T, Nishizaki H, Utsuro T, Nakagawa S. Unsupervised speaker adaptation using high confidence portion recognition results by multiple recognition systems. Proc 8th ICSLP, p 1989–1992, 2004.
19. Wessel F, Macherey K, Ney H. A comparison of word graph and N-best list based confidence measures. Proc 6th Eurospeech, p 315–318, 1999.

AUTHORS (from left to right)

Takehito Utsuro received his B.E., M.E., and D.Eng. degrees in electrical engineering from Kyoto University in 1989, 1991, and 1994. He was an instructor at the Graduate School of Information Science of Nara Institute of Science and Technology from 1994 to 2000, and a lecturer at the Department of Information and Computer Sciences of Toyohashi University of Technology from 2000 to 2002. He has been a lecturer in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, since 2003. From 1999 to 2000, he was a visiting scholar at the Department of Computer Science of Johns Hopkins University. His professional interests lie in natural language processing, spoken language processing, machine learning, and artificial intelligence.

Yasuhiro Kodama received his B.E. and M.E. degrees in information and computer sciences from Toyohashi University of Technology in 2001 and 2003. Since then, he has been involved in research and development activities of information technologies at Sony Corporation.

Tomohiro Watanabe received his B.E. degree in information and computer sciences from Toyohashi University of Technology in 2002 and is now in the master's program there. His research interests include spoken language processing.

Hiromitsu Nishizaki received his B.E., M.E., and D.Eng. degrees in information and computer sciences from Toyohashi University of Technology in 1998, 2000, and 2003 and joined the Interdisciplinary Graduate School of Medicine and Engineering of the University of Yamanashi as a research assistant. His research interests include spoken language processing.

Seiichi Nakagawa received his B.E. and M.E. degrees from Kyoto Institute of Technology in 1971 and 1973, and D.Eng. degree from Kyoto University in 1977. After serving as a research associate and an assistant professor at Kyoto University, he moved to Toyohashi University of Technology in 1983, and has been a professor in the Department of Information and Computer Sciences since 1990. From 1985 to 1986, he was a visiting scientist in the Department of Computer Science, Carnegie-Mellon University. He received 1997 and 2001 paper awards from IEICE and the 1988 JC Bose Memorial Award from the Institution of Electro. Telecommunication Engineers. His major research interests include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence.