Systems and Computers in Japan, Vol. 35, No. 7, 2004
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J86-D-II, No. 7, July 2003, pp. 974– 987

Estimating High-Confidence Portions Based on Agreement among Outputs of Multiple LVCSR Models

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

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

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

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

[4]Department of Information and Computer Sciences, Toyohashi University of Technology, Aic hi, 441-8580 Japan

SUMMARY

This paper experimentally evaluates 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. The results of experimental evaluation show that the agreement between the outputs with two LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. Furthermore, among various features of acoustic models based on Gaussian mixture HMMs, it is concluded that ones such as whether or not to have short pause models, as well as different units in HMMs are the most effective in achieving highly reliable confidence. © 2004 Wiley Periodicals, Inc. Syst Comp Jpn, 35(7): 33–40, 2004; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.10636

Key words: LVCSR models; confidence measures; combination of multiple models; acoustic models; recognition error detection.

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 [1], number of edges in word graphs [2], hypothesis density [1], likelihood of acoustic/language models [3], and posterior probabilities [4]—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 [5, 6]. Considering the success of a simple voting scheme such as ROVER, it also seems quite possible to improve reliability of previously studied features for confidence measures by simply exploiting more than one speech recognizer's outputs. From this observation, unlike those previous works on confidence measures, this paper studies features for confidence measures that are extracted from outputs of more than one LVCSR model.

For the purpose of estimating confidence for each hypothesized word, it is more important to examine which combination of existing LVCSR models can achieve high confidence and which combination cannot, although even simple voting schemes can achieve word error reduction. Therefore, in this paper, we experimentally evaluate 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. In this evaluation of existing Japanese LVCSR models, we concentrate on evaluating confidence of the agreement among outputs with different decoders and/or different acoustic models. The results of experimental evaluation show that the agreement between the outputs with two LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. Furthermore, among various features of acoustic models based on Gaussian mixture HMMs, it is concluded that ones such as whether or not to have short pause models, as well as different units in HMMs (e.g., triphone model or syllable model) are the most effective in achieving highly reliable confidence.

2. Specification of Japanese LVCSR Models

2.1. Decoders

As the decoders of Japanese LVCSR systems, we use the one named Julius, which is provided by the IPA Japanese dictation free software project [7], as well as the one named SPOJUS [8], which has been developed in the Spoken Language Processing Laboratory (Nakagawa Laboratory), Toyohashi University of Technology. 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 the decoder Julius

As the acoustic models used with the decoder Julius, we evaluate phoneme-based HMMs [7] as well as syllable-based HMMs [9]. 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. Triphone model
2. Phonetic tied mixture (PTM) triphone model
3. Monophone model
4. 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,[*] which amount to eight acoustic models in total.

2.2.2. Acoustic models with the decoder SPOJUS

The acoustic models used with the decoder SPOJUS are based on syllable HMMs, which have been developed in the Spoken Language Processing Laboratory, Toyohashi University of Technology [10]. The number of Japanese

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

newspaper sentence utterances		
decoder	word correct (%)	word accuracy (%)
Julius	93.9 (max) to 73.8 (min)	91.3 (max) to 70.3 (min)
SPOJUS	91.1 (max) to 79.5 (min)	86.2 (max) to 55.3 (min)
broadcast news speech		
decoder	word correct (%)	word accuracy (%)
Julius	72.4 (max) to 50.4 (min)	69.2 (max) to 40.8 (min)
SPOJUS	71.5 (max) to 55.6 (min)	63.9 (max) to 38.9 (min)

[*] 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).

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 the sets of feature parameters are evaluated:

- 2*m*-dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames, delta *m* dimensions calculated over 9 successive frames, delta delta *m* dimensions, and delta, delta delta powers ($m = 10, 12$).
- 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 have difference in only one feature. Then, as in the case of the acoustic models used with the decoder Julius, for each of the nine models, we evaluate HMMs both *with* and *without* the short pause states,[*] which amount to 18 acoustic models in total.

2.3. Language models

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

1. One trained using 45 months of Mainichi newspaper articles

2. One trained using 5 years of Japanese NHK[†] 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 [7] from the JNAS (Japanese Newspaper Article Sentences) speech data [11].

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

3. A Metric for Evaluating Confidence

This section gives the definition of our metric for evaluating confidence. In principle, the task of estimating confidence for each hypothesized word is to have an estimate of which words of the outputs of LVCSR models are likely to be correct and which are not reliable. In this paper, however, we focus on estimating correctly recognized words and evaluate confidence according to recall/precision rates of estimating correctly recognized words.

The following gives a procedure for evaluating the agreement among the outputs of multiple LVCSR models as an estimate of correctly recognized words. First, let us suppose that we have two outputs Hyp_1 and Hyp_2 of two LVCSR models, each of which is represented as a sequence of hypothesized words. Next, two sequences Hyp_1 and Hyp_2 of hypothesized words are aligned by DP matching. Then, words that are aligned together and have an identical lexical form are collected into a list named *agreed word list*. Suppose that we have two sequences Hyp_1 and Hyp_2 of hypothesized words as below:

$$Hyp1 = w11,\ \ldots,\ w1i,\ \ldots,\ w1k$$

$$Hyp2 = w21,\ \ldots,\ w2j,\ \ldots,\ w2l$$

Then, the *agreed word list* is constructed by collecting those words w_{1i} ($= w_{2j}$) that satisfy the constraint: w_{1i} and w_{2j} are aligned together by DP matching, and w_{1i} and w_{2j} are lexically identical. Finally, the following recall/precision rates are calculated by comparing the agreed word list with the reference sentence considering both the lexical form and the position of each word.

[*] 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.

[†] Japan Broadcasting Corporation.

$$Recall = \frac{\text{\# of correct words in the agreed word list}}{\text{\# of words in the reference sentence}}$$

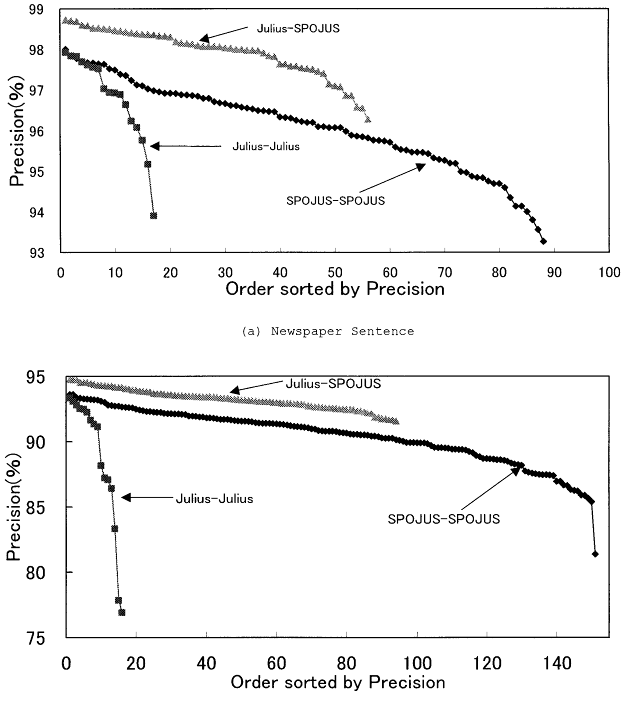
$$Precision = \frac{\text{\# of correct words in the agreed word list}}{\text{\# of words in the agreed word list}}$$

4. Experimental Results

This section describes the results of evaluating the agreement among the outputs of two LVCSR models as an estimate of confidence for each hypothesized word.

4.1. Agreement between two decoders

First, we evaluate correlation between difference of decoders and confidence. We classify 325 pairs of all 26 LVCSR models (8 for the decoder Julius and 18 for the decoder SPOJUS) according to the pairs of decoders, namely, Julius–SPOJUS, Julius–Julius, and SPOJUS–SPOJUS. Then, for each of the 325 LVCSR model pairs, we evaluate the precision/recall of the agreement between their outputs and plot their precision values in descending order. Figure 1 gives the plots for those with recall values above a threshold (80% for the newspaper sentence utterances and 50% for the broadcast news speech) for each of the decoder pairs.



(a) Newspaper Sentence

(b) Broadcast News

Fig. 1. Distribution of precision of agreement between two models (for each pair of decoders).

The maximum precision values achieved are almost 99% for the newspaper sentence utterances and 95% for the broadcast news speech, which indicate quite high confidence. It is also quite clear from this result that agreement between the outputs from two different decoders can achieve higher confidence than the same two decoders.* The maximum precision values for the same decoder pairs are, however, just 1% lower than those of the different decoder pairs, achieving sufficiently high confidence. As for the recall values for those maximum precision pairs, they are around 84% for the newspaper sentence utterances and 64% for the broadcast news speech. This means that, for the newspaper sentence utterances, nearly 99% precision is achieved by decreasing the word correct rate (= recall rate) by only 7%, and for the broadcast news speech, nearly 95% precision is achieved by decreasing the word correct rate (= recall rate) by only 8%.

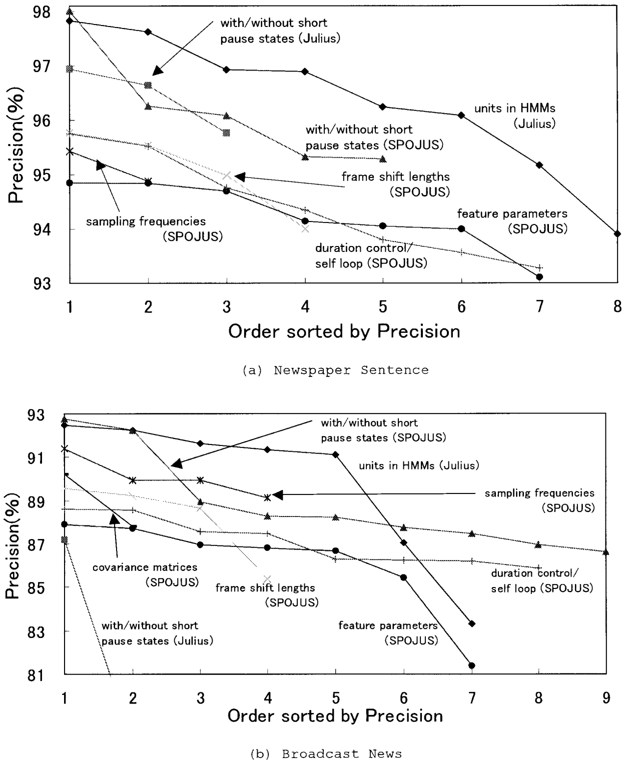
4.2. Agreement between two acoustic models

Next, among various features of acoustic models, this section examines which ones are the most effective in achieving high confidence. For this purpose, out of the 325 pairs of all 26 LVCSR models, we focus on pairs with the same decoders (i.e., Julius–Julius or SPOJUS–SPOJUS), which have differences only in their acoustic models.

4.2.1. Contribution of difference in a single feature of acoustic models

First, in order to evaluate the contribution of difference of a single feature of acoustic models to achieving high confidence, we examine the precision values for minimal pairs which differ in only one of those features. Features of acoustic models examined are as follows: for the decoder Julius, units in HMMs (i.e., the four types listed in Section 2.2.1), and with/without short pause states, while for the decoder SPOJUS, all the features described in Section 2.2.2, that is, sampling frequencies, frame shift lengths, feature parameters, covariance matrices, self-loop transition/duration control, and with/without short pause states. For each of those features, we evaluate the precision/recall of the agreement between minimal pairs and plot their precision values in descending order. Figure 2 gives this plot for those with recall values above the threshold. In terms of the maximum precision values, the best performing features are "with/without short pause states" for SPOJUS and "units in HMMs" for Julius. "With/without short pause states" for Julius performs slightly worse mainly because models without short pause states for Julius have word recognition rates much lower than those for SPOJUS.

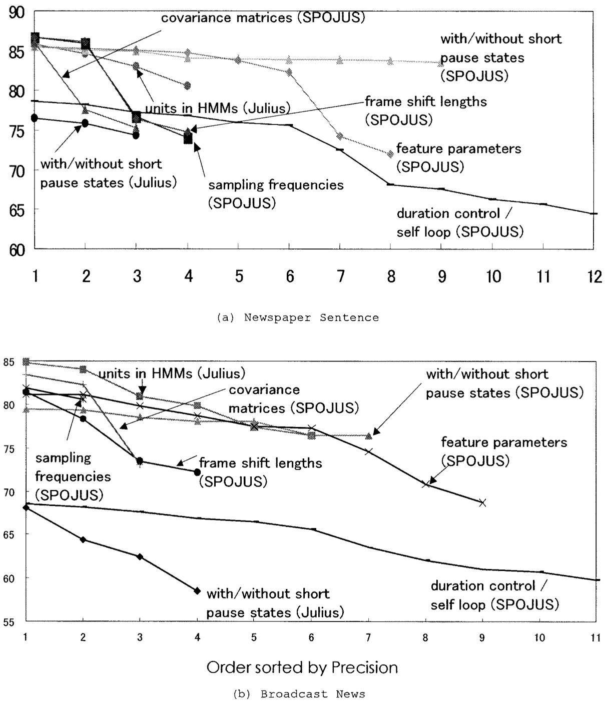
*We also have similar results when we add a third decoder ATRSPREC, evaluating against ATR travel conversational speech [12].



(a) Newspaper Sentence

(b) Broadcast News

Fig. 2. Distribution of precision of agreement between two models (difference in a single feature, word recognition).



(a) Newspaper Sentence

(b) Broadcast News

Fig. 3. Distribution of precision of agreement between two models (difference in a single feature, syllable recognition).

These results claim that the features of acoustic models most effective in achieving high confidence are "with/without short pause states" and "different units of HMMs." Note, however, that the feature "with/without short pause states" is not a pure acoustic one, but has its difference also in the language model side, that is, "with/without punctuation symbols." Therefore, for the purpose of evaluating contribution of the difference in pure acoustic features, we evaluate the confidence of continuous syllable recognition without language model. Evaluation procedures are the same as the case of Fig. 2 except that recall/precision values are measured against syllables. Evaluation results are given in Fig. 3, where it is quite clear that most features including "with/without short pause states" perform similarly in terms of the maximum precision values. It can be concluded that the contribution of "with/without punctuation symbols" in the language model is much greater than that of "with/without short pause states" in the acoustic model.[*]

4.2.2. Summary

As we showed in the previous section, the best performing features are "with/without short pause states" for SPOJUS and "units in HMMs" for Julius. Furthermore, pairs of models with more than one difference in their features tend to have higher precisions than those with difference in only one feature [13]. In order to evaluate contribution of difference of features of acoustic models in general cases, in this section, we classify difference of features of acoustic models as below.

First, for the decoder pair Julius–Julius, differences of features of acoustic models are classified as follows:

1. Pairs of models with differences both in units in HMMs and in with/without short pause states
2. Pairs of models with difference only in units in HMMs
3. Pairs of models with difference only in with/without short pause states

The left-hand side of Fig. 4 gives the precision value of a single model pair with the maximum precision value for each of the three categories. From these results, putting more emphasis on the performance against the harder

[*] This result suggests that the difference in the language model might help achieving high confidence and thus future works definitely include introducing various language models such as those with syntactic structures, trigger models, and topic categories into this task.

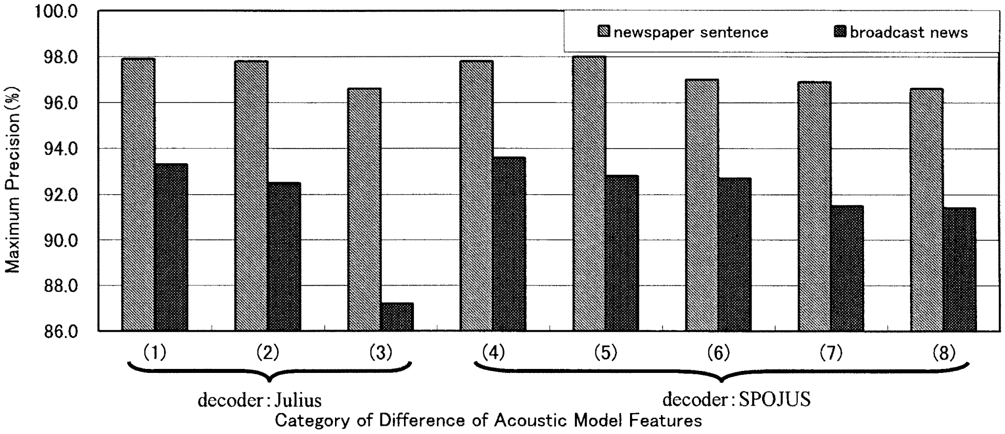


Fig. 4. Evaluation results of agreement between two acoustic models: summary.

speech, that is, the broadcast news speech, we conclude by representing the degree of the contribution of those acoustic features according to the inequalities below:

$$1 > 2 >> 3$$

Next, for the decoder pair SPOJUS–SPOJUS, differences of features of acoustic models are grouped into the following five classes:

4. Pairs of models with more than one difference in their features including with/without short pause states
5. Pairs of models with difference only in with/without short pause states
6. Pairs of models with short pause states which have more than one difference in their features
7. Pairs of models without short pause states which have more than one difference in their features
8. Pairs of models other than all of the above, with difference in only one feature

For each of the five categories, the right-hand side of Fig. 4 gives the precision value of a single model pair with the maximum precision value. Also from these results, putting more emphasis on the performance against the broadcast news speech, we conclude by representing the degree of contribution of those acoustic features according to the following inequalities:

$$4 > 5, 6 > 7 > 8$$

(The reason for “7 > 8” is that the pair achieving the maximum precision values for the broadcast news speech in the category does not achieve the maximum precision values for the newspaper sentence utterances.)

5. Concluding Remarks

This paper 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. The results of experimental evaluation showed that the agreement between the outputs with two LVCSR models with different decoders and acoustic models can achieve quite reliable confidence. Along with the results presented in this paper, we [14] showed that the proposed measure of confidence outperforms previously studied features for confidence measures such as the *acoustic stability* and the *hypothesis density* [1]. We also applied SVM learning technique to the task of combining outputs of multiple LVCSR models, where highly confident portions of hypothesized words are combined, achieving relative word error reduction of up to 53% against the best performing single model [15]. Future work includes employing LVCSR models with various language models such as structural language models [e.g., 16, 17] and topic-based language models [e.g., 17, 18] as participating individual models.

Acknowledgments. The authors thank members of the Spoken Language Processing Laboratory (Nakagawa Laboratory), Department of Information and Computer Sciences, Toyohashi University of Technology. They also thank members of NHK Science Technical Research Laboratories for providing us with the broadcast news scripts and the broadcast news speech data.

REFERENCES

1. Kemp T, Schaaf T. Estimating confidence using word lattices. Proc 5th Eurospeech, p 827–830, 1997.
2. Ogata J, Ariki Y. Improved speech recognition using iterative decoding based on confidence measures. Proc 7th Eurospeech, p 2577–2580, 2001.
3. 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)
4. 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.
5. 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.
6. Schwenk H, Gauvain J-L. Combining multiple speech recognizers using voting and language model information. Proc 6th ICSLP, p 915–918, 2000.
7. 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.
8. 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, 1998.
9. Moroto M, Matsumoto H. Evaluation of Mel-LPC analysis by a large vocabulary Japanese dictation system. Proc 7th Western Pacific Regional Acoustics Conference, 2000.
10. Nakagawa S, Yamamoto K. Evaluation of segmental unit input HMM. Proc 21st ICASSP, p 439–442, 1996.
11. 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.
12. Watanabe T, Yamamoto H, Kokubo H, Kikui G, Nishizaki H, Kodama Y, Utsuro T, Nakagawa S. Combining outputs of multiple LVCSR models—Evaluation on travel conversational speech. Proc 2003 Spring Meeting of the Acoustical Society of Japan, Vol. I, p 209–210. (in Japanese)
13. Utsuro T, Nishizaki H, Harada T, Kodama Y, Nakagawa S. Performance analysis of confidence of agreement among multiple LVCSR models. Tech Rep IEICE 2002;SP2001-128:25–32. (in Japanese)
14. 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.
15. 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.
16. Chelba C, Jelinek F. Structured language modeling. Comput Speech Lang 2000;14:283–332.
17. Khudanpur S, Wu J. Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. Comput Speech Lang 2000;14:355–372.
18. Florian R, Yarowsky D. Dynamic non-local language modeling via hierarchical topic-based adaptation. Proc 37th Annual Meeting of ACL, p 167–174, 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. After serving at Nara Institute of Science and Technology and Toyohashi University of Technology, he has been a lecturer in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, since 2003. He was a visiting scholar in the Department of Computer Science at Johns Hopkins University in 1999–2000. His professional interests lie in natural language processing, spoken language processing, machine learning, and artificial intelligence.

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. He is now a research associate in the Interdisciplinary Graduate School of Medicine and Engineering at the University of Yamanashi. His research interests include spoken language processing.

Yasuhiro Kodama received his B.E. and M.E. degrees in information and computer sciences from Toyohashi University of Technology in 2001 and 2003, and is now involved in research and development of information technologies at Sony Corporation.

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. He has been a professor in the Department of Information and Computer Sciences at Toyohashi University of Technology since 1990. He was a visiting scientist in the Department of Computer Science at Carnegie-Mellon University in 1985–1986. He received 1997 and 2001 Paper Awards from IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronics Telecommunication Engineers. His major research interests include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence.