IEICE TRANS. INF. & SYST., VOL.E91–D, NO.3 MARCH 2008

PAPER Special Section on Robust Speech Processing in Realistic Environments

Using Mutual Information Criterion to Design an Efficient Phoneme Set for Chinese Speech Recognition

Jin-Song ZHANG[†,††], Xin-Hui HU[†,††], *Nonmembers, and* Satoshi NAKAMURA[†,††], *Member*

SUMMARY Chinese is a representative tonal language, and it has been an attractive topic of how to process tone information in the state-of-the-art large vocabulary speech recognition system. This paper presents a novel way to derive an efficient phoneme set of tone-dependent units to build a recognition system, by iteratively merging a pair of tone-dependent units according to the principle of minimal loss of the Mutual Information (MI). The mutual information is measured between the word tokens and their phoneme transcriptions in a training text corpus, based on the system lexical and language model. The approach has a capability to keep discriminative tonal (and phoneme) contrasts that are most helpful for disambiguating homophone words due to lack of tones, and merge those tonal (and phoneme) contrasts that are not important for word disambiguation for the recognition task. This enables a flexible selection of phoneme set according to a balance between the MI information amount and the number of phonemes. We applied the method to traditional phoneme set of Initial/Finals, and derived several phoneme sets with different number of units. Speech recognition experiments using the derived sets showed its effectiveness.

key words: *mutual information, Chinese lexical tones, tone dependent units, speech recognition*

1. Introduction

Chinese is a tonal language, in which each syllable is associated with a kind of pitch tone. There are four basic tones and one neutral tone. The same syllables with different tones have different lexical meaning. It has been an interesting and important topic how to model the tone information to build a Chinese large vocabulary continuous speech recognition (LVCSR) system. Among a number of various kinds of approaches, the one using tone dependent sub-word units has the advantage of frame-synchronous consistency with the decoding strategy of the state-of-art LVCSR system, and has been widely adopted [1], [2], [4]. One common problem of these approaches is that the number of phoneme set of the LVCSR system will increase significantly after introducing tone dependencies. For example, in the case of widely used traditional Chinese phoneme set of Initials/Finals (IFs), the number of non-tonal IFs is 59, and that of tone-dependent ones is more than 200. As context dependent tri-phone HMMs are usually used in LVCSR systems, their number will explode from tens of thousands to millions when tone-dependency is used, making it very challenging how to train the tri-phone HMMs robustly. Also, the complexity of the phoneme hypotheses lattice will increase significantly, making the decoding much more computationally heavy.

The approaches to deal with the problem in the previous studies [1], [2], [4] are to hand-craft a small phoneme set containing tone-dependent phonemes, like tonemes [1], tonal main vowels [4], segmental tones [2] and etc.. Although they showed performance improvements in the recognition experiments, they still need to increase the phoneme set by several times due to a full expansion of non-tone units to tone dependent ones. However, we regard a full expansion of tone dependencies as unnecessary. On the one hand, speakers tend to reduce some tones from their lexical forms in daily speech [5] when the reductions do not obstacle speech communication. On the other hand, the lexical and language model (e.g., n-gram) information in an LVCSR system is usually very efficient to disambiguate most of homophone words due to a lack of tone information [6], as evidenced by the fact that an incorporation of several-times-big tonal phoneme set has led to only slight recognition improvements [1], [4].

By viewing the full expansion of tone dependencies as unnecessary, we propose that only those tone dependencies be incorporated that are necessary for disambiguating word confusions of an LVCSR system. Different from several previous studies on disambiguating word confusions which are based on the acoustic confusions of phonemes [7]–[10], our method focuses on the disambiguation power from the lexical and language model. In other words, a tone dependency is not incorporated when the lexical and language model can disambiguate those homophone words resulting from the lack of that tone.

The real approach is realized as compacting the redundancy of an initial full-tone-dependent unit set, according to the principle of minimal loss of the mutual information. The mutual information is measured between the word tokens and their phoneme transcriptions in a training text corpus. A greedy search is adopted to merge two units at a time to minimize the corresponding mutual information loss. The final phoneme set can be flexibly chosen according to a balance between the number of units and the information quantities. Speech recognition experiments have been carried out to testify the effectiveness of the deduced phoneme sets.

Manuscript received July 4, 2007.
Manuscript revised September 15, 2007.

[†]The authors are with the Spoken Language Communication Group, Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Kyoto-fu, 619–0288 Japan.
[††]The authors are also with ATR Spoken Language Translation Communication Research Laboratories.

DOI: 10.1093/ietisy/e91–d.3.508

2. Chinese Phonology

As illustrated by Fig. 1, a Chinese word is composed of one to several characters, and each character is pronounced as a monosyllable with a pitch tone. When tone information is discarded, a syllable is called as a base syllable. The total phonetically differentiable tonal syllables are about 1,400, and the number of base syllables is about 410 when pitch tones are discarded. Traditional Chinese phonology divides the syllables into demi-syllabic units: 21 *Initials* and 38 *Finals*, plus four basic lexical tones (Tone 1-4) and a neutral tone (0) (Table 1).

In the tone-dependent phoneme approach, all the Finals are expanded into tone dependent ones, like *a0, a1, a2, a3, a4* and etc.. Although not all the combinations of Finals and tones exist, the number of the tone dependent phoneme set is still more than 200. When isolate monosyllable words are considered, tone contrasts may play an important role in discriminating the words. For ex, the following words: *ma1* (mother), *ma2* (hemp), *m3* (horse) and *ma4* (scold), are only differentiated by the tones when in isolations. However, when in sentences, they will have very different context words. In other words, the lexical and language model (n-gram) has the power to disambiguate the four words even the tone information is ignored. Therefore, we regard that there are many redundancies in the original tone dependent IFs set for a recognition system, when given a lexical and language model.

Table 1 Pinyin symbols for Initials, Finals and lexical tones of Chinese syllables.

Initials	b, p, m, f, d, t, n, l, z, c, s, zh, ch, sh, r, j, q, x, g, k, h, *null initial*
Finals	a, ao, ai, an, ang, o, ou, ong, e, ei, en, eng, er ia, iao, ie, i1, i2, i3, iu, in, ian, ing, iang, iong, u, ua, uo, ui, uai, un, uan, uang, v, ve, vn, van
Tone	0, 1, 2, 3, 4

3. Theory for Phoneme Set Optimization

We formalize the phoneme set optimization problem into an information coding/decoding approach as illustrated in Fig. 2, where W stands for a language, appearing as a text corpus, Φ_1 and Φ_2 for two different phoneme sets, $F_{1,2}$ for the different phoneme transcriptions of the W using $\Phi_{1,2}$ based word lexicons respectively, $W_{1,2}$ for the decoded text from $F_{1,2}$ based on the same language model and the respective lexicons. When a coding method Φ_i is lossless, the decoded text W_i should satisfy $W = W_i$. However, when the coding is not uniquely decodable, a better coding Φ^* should be the one

$$\Phi^* = \arg\max_i I(W, F_i) \text{ where } i = 1, 2 \tag{1}$$

Meaning of the above equation is as follows: when a decoding from a phoneme sequence to text is not unique, it means that there is a problem of homophone words. Among two different phoneme sets, the one that has larger mutual information $I(W, F)$ with text W tends to trigger less homophone words.

The mutual information $I(W, F_i)$ is defined as

$$I(W, F_i) = H(W) - H(W|F_i) \tag{2}$$

$H(W)$ is the entropy of text corpus W, which depicts for sequences of words $\{w_1, w_2, \ldots, w_n\}$. $H(W)$ is usually calculated as the word based average entropy by:

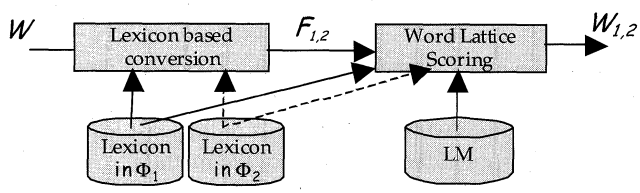


Fig. 2 Illustration of the problem formalization. Here, "W" stands for the original Chinese text, "$F_i, i = 1, 2$" for phoneme transcriptions for "W" based on two different phoneme sets $\Phi_i, i = 1, 2$, "$W_i, i = 1, 2$" for decoded text from "$F_i, i = 1, 2$" based on the language model "LM".

Units	Ex.1	Ex.2		Ex.3	Ex.4	Ex.5	Ex.6		Number of Types
Words	妈	妈妈		麻	骂	买	卖	...	10,000~100,000
Characters	妈	妈	妈	麻	骂	买	卖	...	6,000~10,000
Tonal syllables	ma1	ma1	ma0	ma2	ma4	mai3	mai4	...	1,400
Initial	M	M	M	M	M	M	M	...	21
Final	A1	A1	A0	A2	A4	AI3	AI4	...	185
Base syllables	ma	ma	ma	ma	ma	mai	mai	...	410
Initial	M	M	M	M	M	M	M	...	21
Final	A	A	A	A	A	AI	AI	...	38

Fig. 1 A brief illustration of Chinese Phonology: from word to character, pronunciations by tonal syllables and base syllables.

$$H(W) = \lim_{n\to\infty} \frac{1}{n} H(w_1, w_2, \ldots, w_n)$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{W} p(w_1, \ldots, w_n) \log p(w_1, \ldots, w_n) \quad (3)$$

According to [11], the Shannon-McMillan-Breiman theorem states that when the language is assumed both stationary and ergodic, the above equation becomes to,

$$H(W) = \lim_{n\to\infty} -\frac{1}{n} \log p(w_1, \ldots, w_n) \quad (4)$$

Based on Eq. (4), Eq. (2) can be rewritten as,

$$I(W, F_i) = \lim_{n\to\infty} \frac{1}{n} (\log P(W|F_i) - \log P(W))$$
$$= \lim_{n\to\infty} \frac{1}{n} \log \frac{P(F_i|W)}{P(F_i)}$$
$$= \lim_{n\to\infty} \frac{1}{n} \log \frac{P(F_i|W)}{\sum_{\text{all } j} P(F_i|W_j)P(W_j)} \quad (5)$$

Then Eq. (1) becomes as,

$$\Phi^* = \arg\max_i \lim_{n\to\infty} \frac{1}{n} \log \frac{P(F_i|W)}{\sum_{\text{all } j} P(F_i|W_j)P(W_j)}$$
$$= \arg\max_i \log \frac{P(F_i|W)}{\sum_{\text{all } j} P(F_i|W_j)P(W_j)} \quad (6)$$

$P(W)$ and $P(F_i|W)$ represent two main components in the current speech recognition system: i.e., language modeling and probabilistic pronunciation variation modeling.

4. Minimum MI Loss Based Phoneme Set Reduction

We have designed a greedy approach to compact the redundancies of an initial phoneme set by iteratively merging one pair of phonemes whose merge leads to the least loss of MI. Figure 3 illustrated the flow chart of the method.

- Initialization condition: the following resources are prepared.
  - Initial phoneme set Φ_0: it contains the full tone-dependent sub-word units.
  - Lexicon: the one for speech recognition task and is represented in the initial phoneme set.
  - Text corpus: the one standing for the speech recognition task.
  - Language model: the one of the speech recognition system.

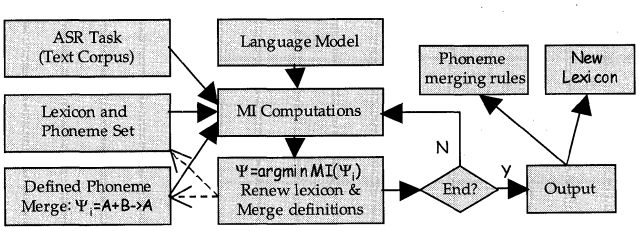


Fig. 3 Illustration of the minimum MI loss based phoneme set reduction.

- Optimization procedure:
  1. MI calculations: for each possible merge of two phonemes: $\Psi_i : A + B \rightarrow A$, uses Φ_{i-1} and Φ_i to depict the phoneme sets before and after merge. Then the reduced MI: $\Delta I(\Psi_i)$ is calculated by,
  $$\Delta I(\Psi_i) = I(W, \Phi_{i-1}) - I(W, \Phi_i)$$
  2. Merge decision: among all the possible merges, the one Ψ^* that has the smallest reduced MI is selected as the effective merge of this iteration.
  $$\Psi^* = \arg\min_{\text{all } i} \Delta I(\Psi_i)$$
  3. Renew the lexicon and phoneme set based on the effective merge Ψ^*.
  4. Check if the stop criterion is satisfied or not. If no, go to step 1 and do 1-3 once again. If yes, stop the optimization and output the phoneme merging rules and new lexicon. Here, the stop criterion is a specified least number of phonemes.

To avoid a computationally heavy exhaustive search through all possible phoneme merges, we limited the search to a constrained space of possible merges. It can be defined according to phonetic knowledge about acoustic similarities between pair of phonemes.

5. Experiments and Results

5.1 Phoneme Set Design Experiments

The text corpus (CBTEC) we used is the Chinese version of Basic Travel Conversation Text (BTEC) of ATR. It contains about 200,000 sentences with about one million words. The lexicon size is about 17,000, and the language model is a 2-gram model trained from CBTEC. The size of initial phoneme set is 206 with all tone-dependent units. The initially defined phoneme merges have 433 possibilities, which is designed based on the phonetic similarities of the tonal phonemes. There is one constraint: two different Finals can be merged only when all of them do not have tone-dependencies.

Figure 4 illustrates the MI variations with the increasing number of merged phonemes for the training text corpus, and Fig. 5 for the test corpus. The figures clearly show that:

- The MI gaps between the points T0 and T4 indicate that some information gets lost when non-tonal IF set is used as the phoneme set for the recognition system.
- There are flat periods of MI variations after T4s in both the training and test data, indicating that a significant number of phoneme merges including tone merges lead to no information loss. Hence, they are the redundancies in the initial full tone-dependent unit set, when given the lexical and language model.

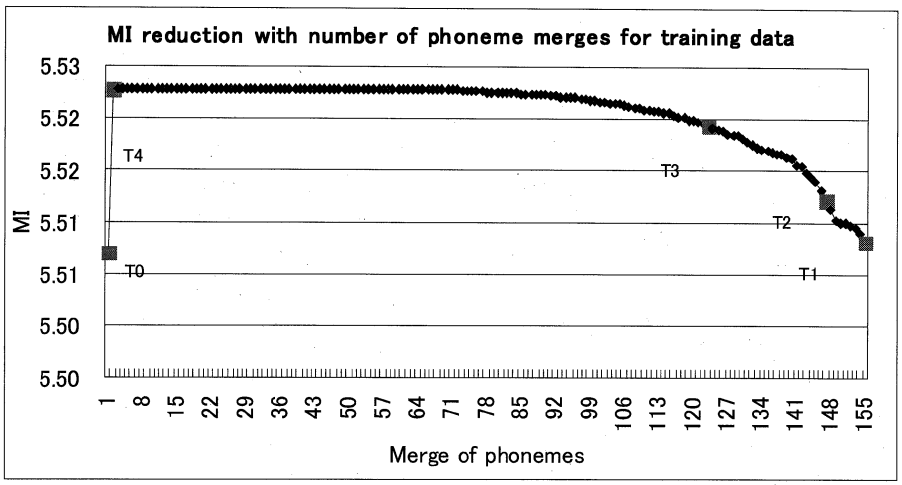


Fig. 4 Illustration of the MI variations with the increasing number merged phonemes for the training text corpus. Points "T0" and "T4" are special ones: the "T0" stands for the non-tonal IFs set with 59 phonemes, and the point "T4" for the initially full 206 phoneme set. "T1", "T2" and "T3" are other three studied points.

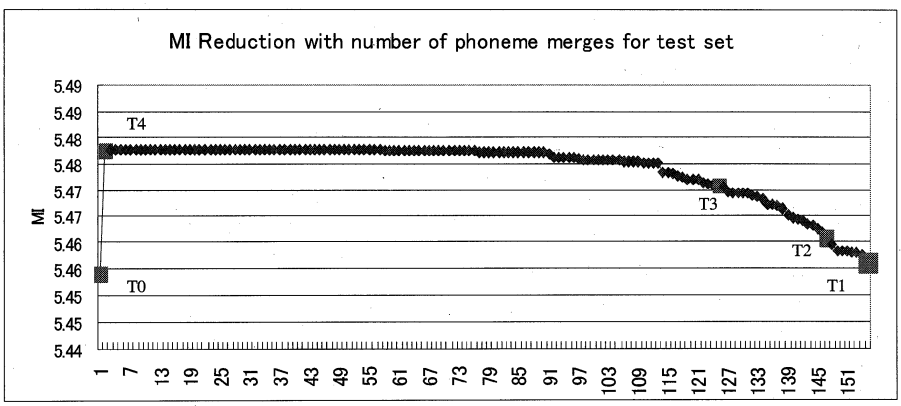


Fig. 5 Illustration of the MI variations with the increasing number of reduced phonemes for the test text corpus. "T0" to "T5" are the same corresponding points to those in Fig. 4.

- As the MI variations are monotonous in the above-mentioned greedy search algorithm, it is impossible to use MI to define an optimal point as the stop criterion currently.
- Although there is no optimal point, the merge process offers us a flexible way to define a phoneme set with the richest information in the given number of phonemes. Through considering a balance between the number of units and loss of MI, we selected several ones to build acoustic models for speech recognition experiments.

We heuristically selected five different unit sets to build our speech recognition systems. T0 is the conventional non-tonal IFs with 59 units; T1 has 50 units and showing a similar MI to that of T0; T2 has the same number of units as T0, but showing a better MI than T0; T3 has 80 units, and showing only slight MI loss from the initial phoneme set T4, which has the full tone-dependent set of 206 units. Table 2 lists the number of Initials, Finals and logical tri-phones in the five ASR systems respectively.

Table 2 Number of units in the selected sets and the number of their corresponding logical HMMs.

Set	Units	Initials	Finals	Tri-phones
T0	59	21	37†	107,441
T1	50	18	31	70,128
T2	59	18	40	114,945
T3	80	20	59	292,651
T4	206	21	184	3,022,775

5.2 Speech Recognition Experiments

The training speech data for acoustic models is the Beijing part of ATR Accented Speech (ATRAS). It contains more than 40,000 utterances with a total duration of 43 hours by 96 balanced male and female native Beijing speakers. The objective test speech data is a subset of CBTEC Putonghua test data. It contains 510 utterances by 5 male and 5 female

†One of original 38 Final units was merged at the beginning.

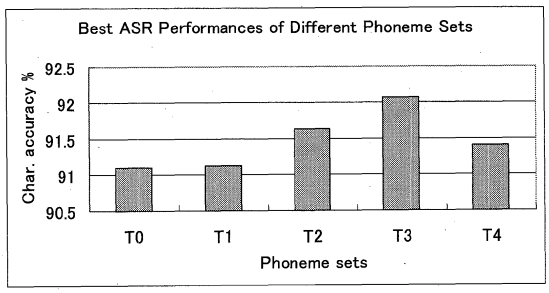


Fig. 6 Illustration of the speech recognition performances for the CBTEC test data in character accuracies using HMMs of different phoneme sets.

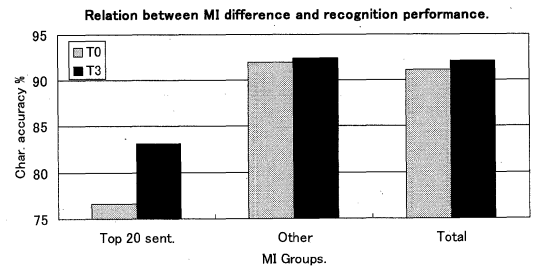


Fig. 7 Illustration of the relationship between MI differences and the recognition performances for the phoneme sets T0 and T3. Top 20 represents the 20 sentences with the maximum MI improvements when using T3 instead of T0.

speakers, each speaker uttering different sentences. The reason to use a slightly accent-mismatched training database is that the ATRAS database has manually annotated phonetic labels, including the tone labels that are different from their lexical forms.

We used the HTK toolkit [12] to build our speech recognition systems to test the five different phoneme sets. The feature vector contains 39 dimensions including standard MFCC features and log power, together with their first and second ordered derivatives. Cepstral mean subtraction is done at sentence level. We developed phonetic-decision-tree based state-tying tri-phone style HMMs for the different phoneme sets. Each HMM has left-to-right 3 states, the total number of tied states for each model has a similar number of 2,000, and each state has 20 Gaussian mixtures. The speech recognition experiments used the same lexical and language model as those used in phoneme set optimization procedure. The perplexity of the objective test set is about 40 for the 2-gram language model. The recognition performances are shown in Fig. 6 in Chinese character accuracies.

The results showed that

- Almost all the derived unit sets (T1 – T4) showed some better or similar performances compared with the non-tonal set T0, indicating that derived phoneme sets are efficient for the recognition task.
- Although T1 has 9 phonemes less than T0, it still got similar performance to that of T0, indicating the phoneme set more efficient.
- T3 achieved the highest performance, maybe due to its better balance between the number of units and MI information amount than others.

We also paid a close look into the relations between recognition performances and word-based average MI differences at sentence by sentence for the two phoneme sets T0 and T3. We separated the 510 test sentences into two groups: one included 20 sentences with the maximum MI improvements, and the other one with all left sentences. The first group showed more significant recognition improvements than the other one, as shown in Fig. 7, indicating that positive MI difference is correlated with recognition improvement.

6. Conclusion

We presented a novel method of deriving a compact and efficient tone-dependent phoneme set for building Chinese LVCSR system using MI based criterion between text corpus and phoneme set. The speech recognition experimental results showed that the derived phoneme set is efficient for Chinese speech recognition. However, there are still rooms to improve the study: First, the current study failed to provide an optimal stop criterion for the phoneme set design. We should study the question of how to design the phoneme set that achieve the best speech recognition performance. Second, it is necessary to further the study by using high order language models (3-gram) and more. Third, we might consider how to incorporate the acoustic confusability measurements into the phoneme design process. Fourth, we should also try the method to other phoneme sets and other languages.

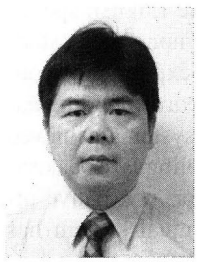
References

[1] C.-J. Chen, et al., "New methods in continuous Mandarin recognition," Proc. Eurospeech 1997, vol.3, pp.1543–1546, 1997.
[2] Ch. Huang and et al., "Segmental tonal modeling for phone set design in mandarin LVCSR," Proc. ICASSP2004, vol.1, pp.901–904, 2004.
[3] F. Seide and N. Wang, "Phonetic modeling in the Philips Chinese continuous-speech recognition system," Proc. Int. Symp. on Chinese Spoken Language Processing, 1998.
[4] C.-J. Chen and et al., "Recognize tone languages using pitch information on the main vowel of each syllable," Proc. ICASSP, 2001.
[5] Y. Xu, "Production and perception of coarticulated tones," J. Acoust. Soc. Am., vol.4, pp.2240–2253, 1994.
[6] J.-S. Zhang and et al., "Is tone recognition necessary for Chinese speech recognition?," Proc. ASJ, pp.5–6, Sept. 2002.
[7] M. Bacchiani and M. Ostendorf, "Using automatically-derived acoustic sub-word units in large vocabulary speech recognition," Proc. ICSLP, 1998.
[8] D.B. Roe and M.D. Riley, "Prediction of word confusabilities for speech recognition," Proc. ICSLP, pp.227–230, 1994.
[9] A. Simons, "Predictive assessment for speaker independent isolated word recognisers," Proc. Eurospeech, pp.1465–1467, 1995.
[10] D. Torre and et al., "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," Proc. ICASSP, pp.1463–1466, 1997.
[11] D. Jurafsky and J.H. Martin, Speech and language processing,

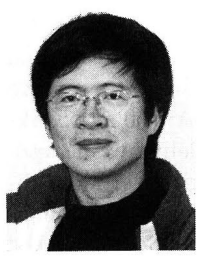
pp.223–232, Prentice-Hall, 2000.

[12] S. Young, et al., HTK Speech Recognition Toolkit ver. 3.2, Cambridge Univ.

[13] J.-S. Zhang, X.-H. Hu, and S. Nakamura, "Automatic derivation of a phoneme set with tone information for Chinese speech recognition based on mutual information criterion," Proc. ICASSP, 2006.



Jin-Song Zhang was born in China on October 4, 1968. He received his B.E. degree in electronic engineering from Hefei University of Technoloy, China in 1989, the M.E. degree from the University of Science and Technology of China (USTC) in 1992, and the Ph.D. degree from the University of Tokyo, Japan in 2000. From 1992 to 1996 he worked as a teaching assistant and lecturer in the department of electronic engineering of USTC. Since 2000, he joined ATR Spoken Language Translation Research Laboratories as an invited researcher. Dr. Zhang is a member of Acoustic Society of Japan. His main research interests include speech recognition, prosody information processing, and speech synthesis.



Xin-Hui Hu was born in China on April 4, 1963. He received his B.E. degree in electronic engineering and M.E. degree in communication and system from Harbin Institute of Technology in 1983, and 1988 respectively. He received his Ph.D. degree from the University of Tokyo in 1995. From 1995 to 2000, he joined the Fujisoft Incorporated as a system engineer. From 2001 to 2003, he joined the Research and Development Center of Toshiba as a researcher. Since 2003, he joined ATR Spoken Language Translation Research Laboratories as an invited researcher. Dr. Hu is a member of Information Processing Society of Japan. His main interests include language model, speech recognition and information retrieval.

Satoshi Nakamura was born in Japan on August 4, 1958. He received his B.S. in Electronic Engineering from the Kyoto Institute of Technology in 1981 and his Ph.D. in Information Science from Kyoto University in 1992. From 1981 to 1993, he worked for Sharps Central Research Laboratory in Nara. From 1986 to 1989, he worked for ATR Interpreting Telephony Research Laboratories. From 1994 to 2000, he was an associate professor of the graduate school of Information Science at the Nara Institute of Science and Technology. In 1996, he was a visiting research professor of the CAIP center at Rutgers University in New Jersey. He is currently the vice president of ATR, the director of ATR Spoken Language Communication Research Laboratories, the head of the Acoustics and Speech Research Department. He is also the group leader of Spoken Language Communication Group, Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology. He also serves as an honorary professor of University Karlsruhe, Germany since 2004. He received the Awaya Award from the Acoustical Society of Japan in 1992, the Interaction 2001 Best Paper Award in 2001, Yamashita Research Award from the Information Processing Society of Japan in 2001, and Telecom System Award and AAMT Nagao Award in 2007. He served as an associate editor for the Journal of the IEICE Information from 2000 to 2002, a member of the Speech Technical Committee of the IEEE Signal Processing Society in 2001-2004, a general chair of International Workshop of Spoken Language Translation (IWSLT2006), Program Chairs of Oriental Cocosda, and IEEE ASRU2007.