INVITED PAPER *Special Section on Recent Advances in Multimedia Signal Processing Techniques and Applications*

Selected Topics from LVCSR Research for Asian Languages at Tokyo Tech

Sadaoki FURUI[†a)], ***Fellow***

SUMMARY This paper presents our recent work in regard to building Large Vocabulary Continuous Speech Recognition (LVCSR) systems for the Thai, Indonesian, and Chinese languages. For Thai, since there is no word boundary in the written form, we have proposed a new method for automatically creating word-like units from a text corpus, and applied topic and speaking style adaptation to the language model to recognize spoken-style utterances. For Indonesian, we have applied proper noun-specific adaptation to acoustic modeling, and rule-based English-to-Indonesian phoneme mapping to solve the problem of large variation in proper noun and English word pronunciation in a spoken-query information retrieval system. In spoken Chinese, long organization names are frequently abbreviated, and abbreviated utterances cannot be recognized if the abbreviations are not included in the dictionary. We have proposed a new method for automatically generating Chinese abbreviations, and by expanding the vocabulary using the generated abbreviations, we have significantly improved the performance of spoken query-based search.

key words: *automatic speech recognition, LVCSR, Asian languages, spoken query*

1. Introduction

More than 6000 living languages are spoken in the world today, and the majority of them are concentrated in Asia. Every language has its own specific acoustic as well as linguistic characteristics that require special modeling techniques. Asian languages have significant variations, and therefore various acoustic as well as linguistic problems need to be solved to realize high-quality Large Vocabulary Continuous Speech Recognition (LVCSR) systems for them.

In our laboratory at Tokyo Tech, we have been conducting research on many languages, including Western and Asian languages. Since we have already published many papers on the Japanese language, this paper focuses on research on other Asian languages, specifically Thai, Indonesian and Chinese.

Written Thai is similar to Japanese and Chinese in the sense that it has no explicit word boundary marker (space), and thus no clear definition of a word [1]. This means that there are several options for defining lexical units that constitute the vocabulary of the Language Model (LM). The problem is thus to find the optimal solution that yields the best final result. This solution will not only have to be interesting from the point of view of language modeling (where one seeks high coverage with a reasonable number of units), but also from the point of view of acoustic confusability (where one prefers units which are acoustically distinct) and pronunciation dictionary development (where one prefers units having a context-independent pronunciation). For minority languages like Thai, it is difficult to collect a large-scale speech corpus, and it is therefore important to build techniques to adapt a baseline LM constructed using a limited size of corpus to various types of input utterances.

There are around 600 languages actively spoken in Indonesia, and only a small number of Indonesian people actually speak the national language, called "Bahasa Indonesia (BI)", as their mother tongue. Through a variety of religious, social, and cultural influences, BI has grown by borrowing words and terms from many languages, including Sanskrit, Arabic, Persian, Portuguese, Dutch, Chinese, and English. BI is written using the Latin alphabet consisting of 26 characters. Although the correspondence between sounds and their written forms is generally regular, there are still some exceptions regarding proper nouns, especially old written style proper nouns or proper nouns that came from regional languages or foreign words, typically English words. Proper nouns and English words are major sources of pronunciation variation in BI. In spoken query-based Information Retrieval (IR), proper nouns and English words play important roles as key words. Therefore, it is important to improve recognition accuracy of these words by adapting Acoustic Models (AMs) or pronunciation dictionaries to account for such pronunciation variations.

Chinese is the most widely spoken language in the world, thus Chinese speech recognition has been the subject of significant research efforts over the years. We have been investigating problems related to abbreviations for organization names, which are frequently used in Chinese spoken query-based IR. In such systems, if the abbreviation is not included in the vocabulary, it causes an OOV (out-of-vocabulary) problem, which means that the abbreviation utterance is mis-recognized as a phonetically similar word, since it is very difficult to detect that the input is an OOV. It is crucial to be able to automatically add abbreviations to the dictionary to avoid such errors. Partly because there is no clear word boundary or definition of words in Chinese, production rules for abbreviations are much more complex than Western languages.

Manuscript received April 8, 2011.
Manuscript revised July 4, 2011.

[†]The author is with Tokyo Institute of Technology, Tokyo, 152–8552 Japan.
a) E-mail: furui@cs.titech.ac.jp

DOI: 10.1587/transinf.E95.D.1182

2. Lexical Units and Language Model Adaptation for Thai LVCSR

2.1 Variation of Lexical Units

A Thai letter is a phonogram which roughly represents a phoneme or combination of phonemes. Thai is written from left to right the same as in English but without sentence or word boundary markers. A space is sometimes used to separate phrases and sentences for aesthetic reasons, but there is no rule or convention requiring this.

A. Word

A word is a unit that is typically adopted by Thai people. Native speakers should be able to determine word boundaries for any sentence. The word segmentation problem in Thai is however, highly ambiguous, and even native speakers will often come up with inconsistent segmentations of the same sentence. There are many words that can be constructed based on the grammar rules. Thus, too much data is required to make a high coverage, word-based LM. Thai natural language processing mostly relies on a word segmentation tool, named SWATH [2], that is based on a hand-coded lexicon containing about 35k words. It is quite effective at word segmentation but the quality of the segmentation results depends on how the lexicon was designed. With the inherently unclear definition of words, the construction of such a lexicon is difficult. Segmentation errors always occur when an unknown word appears in the sentence. In order to avoid the ambiguous definition of Thai words and the dependence of the word segmentation tool on a lexicon, we tried to find other, hopefully better lexical units.

B. Pseudo-morpheme (PM)

A PM is defined as a syllable-like unit of the written form. It must not be confused with an acoustic syllable that follows from the phonetic representation of an utterance. The problem of Thai word segmentation can be solved by PM segmentation because PMs are well-defined and can be more consistently analyzed than words. The most important is that the number of PM patterns is finite and all PM patterns can be listed. Therefore, every input string can be completely covered by the PM patterns. However, the PM pattern for each input string is not always unique, often achieving various PM segmentation results from a string. To find the most likely segmentation, 3-grams of PMs can be used. We use a PM segmentation tool developed by Aroonmanakun [3], in which around 200 PM patterns were defined and 3-grams of PMs were trained from a text corpus of about 553k PMs.

The coverage of a PM lexicon is high because the number of PM patterns is finite. Moreover, since PM segmentation is fairly deterministic, PMs may be effectively used as a lexical unit for Thai LVCSR. PMs however, are not suitable to be used as a lexical unit due to three basic problems.

(1) Higher acoustic confusion occurs when PM is used as a lexical unit, compared to the word unit. PM segmentation produces many short lexical units, resulting in many short utterance units. It is more difficult to acoustically distinguish short than long units.
(2) The span of an *n*-gram LM is significantly short since the lexical unit is short. Therefore, the LM cannot perform efficiently compared to a word-based LM.
(3) PMs may have variation in pronunciation depending on context. Pronunciations of any single PM cannot be generated reliably unless the preceding and following PMs are known.

As a simple solution, an *n*-gram LM with high order of *n* is used to solve problem (2). A special pronunciation dictionary that includes all possible PM pronunciations when composed with other PMs is developed to solve problem (3). Since the number of PMs is not that large, we just straightforwardly create these variations manually.

C. Compound PM (CPM)

Due to the three PM problems described in the previous subsection, it is preferable to merge PMs to form CPMs [4]. In this approach, the number of lexical units in an LM can be controlled. Since CPMs are longer than PMs, the PM problems can be solved in the following ways.

(1) Acoustic confusion among CPMs is lower than PMs.
(2) The span of an *n*-gram LM for CPMs is longer than that for PMs.
(3) The pronunciation variation of the lexical units is significantly reduced.

A text corpus is first segmented into PMs. Then, consecutive PMs that have high co-occurrence probabilities are merged together to form CPMs. Since no hand-coded lexicon is used in the process of PM segmentation and merging, the process to generate CPMs does not face the problem of unknown word segmentation errors that occur in the traditional word segmentation approach.

D. PM-word combination (PMWORD)

To solve the problem of word segmentation errors occurring when an unknown word appears in the sentence, lexicon-based word segmentation results are used in regions where no unknown words appear in the text, and PM segmentation results are used otherwise.

2.2 Language Model Adaptation

A. Corpora for Thai broadcast news LVCSR

In addition to the first Thai broadcast news (BN) speech and language corpora we have developed [5], a collaborative work with NECTEC [6] was established to increase the size of the corpus. However, the amount of available BN transcript text is still less than 100 hours, while there are more than 1000 hours of transcribed text in other major languages. Since the construction of a BN corpus requires a lot of resources, labor, time and money, it would be favorable if an alternative text resource can be employed to train a LM for BN speech. Text resources of which the content seems to be similar to BN text are text in newspapers (NP). An NP text corpus can be constructed much more easily than a BN transcript text corpus. Hence, an NP text corpus is very attractive if it can improve the performance of an LVCSR system.

In fact, not only resource deficient languages suffer from the lack of training resources, but there are also general difficulties in constructing LMs matched to a target speech domain because the amount of well-matched data is usually limited. The focus of our work is to investigate LM topic and style adaptation for Thai BN LVCSR, using two text resources, BN and NP text corpora [7]. Styles here refer to the differences of text styles used in BN and NP, and specific speaking styles used in the Thai language. Based on the characteristics of the Thai language, a rule-based speaking style classification approach is used to classify text into spoken and written styles. LMs for different topics and styles are trained and then combined together using linear interpolation. CPMs are used in this work as the basic lexical unit.

B. *Thai speaking styles and BN speech*

One significant difference between Thai spoken and written style text is the level of politeness of a sentence. In a formal conversation as well as a BN report, a news announcer needs to speak politely to the other party. For a man, "ครับ" (khrap^3) is used and for a woman, "คะ" (kha3) or "ค่ะ" (kha1) is used. These words are added at the end of a sentence but sometimes they are also inserted within a sentence when the speaker tries to make a pause. Some other words are used together with the words indicating politeness to express additional meaning or feeling. For example, one of the most frequent words found in Thai BN is "นะ" (na3) which, in most cases, holds no special meaning but sometimes emphasizes the content of the sentence or is used in imperative sentences. Another word that appears occasionally is "ล่ะ" (la1) which is used mostly in questions. The above words are always placed in front of words indicating politeness, forming spoken style words such as "นะครับ", "นะคะ", "ล่ะครับ", and "ล่ะคะ". Since these spoken style words are often (but not always) put at the end of a sentence, we refer to these words as spoken style ending words (SSEW) for the rest of this paper.

C. *Rule-based speaking style classification*

Motivated by the different characteristics of Thai spoken and written style sentences, we propose a rule-based classification method to indicate the speaking style of a sentence. A sentence containing SSEWs is considered as a spoken style sentence, and a sentence without such SSEWs is classified as a written style sentence. In this work, 11 words were defined as SSEWs. Since CPMs are used as the lexical unit, a list of SSEWs must be constructed in the form of CPMs. A CPM is usually longer than the word unit and SSEWs can be combined with some other PMs. Therefore, CPMs including SSEWs are considered as spoken style CPMs which comprise 132 spoken style CPMs in our system.

Thai BN normally comprises written style speech and spoken style speech. Written style speech is often found when news announcers read news scripts narrating detailed news reports. On the other hand, spoken style speech is mostly found in introductions, transitions, and conclusions of news stories. In our transcribed corpus, written style and spoken style sentences represent 57% and 43%, respectively. On the other hand, NP normally employs written style text. For example, our NP text corpus contains mostly written style sentences (99.3%).

D. *Text clustering*

Training text is clustered based on three characteristics as follows:

(1) Text source: Text is grouped based on its source, which results in distinguishing BN and NP text.

(2) Speaking style clustering: Text is clustered based on spoken style (SP) and written style (WR). Here, the rule-based classification method described above is employed.

(3) Topic clustering: Text is clustered into topics. We use *tf-idf* (term frequency-inverse document frequency) vectors to represent sentences and the cosine function to measure the similarity between sentences. A cluster of text is constructed based on the similarity scores using a two-phase bisecting K-means algorithm. Around 350 function words (in CPM forms) are defined and excluded in the calculation of *tf-idf* vectors.

E. *Adaptation of n-gram models*

All training text is clustered by sources, topics and styles. Each specialized LM is trained from a text cluster. Interpolation weights of the models are optimized with EM algorithm on hypotheses derived from a previous pass of a multi-pass recognition system. The following model types are investigated for our LM adaptation scheme.

- Model Type I: A specialized model is trained from text from a specific source. The adapted model is obtained by interpolating a BN-based LM and an NP-based LM.
- Model Type II: Irrespective of text source, a specialized model is trained from text with a specific speaking style. The adapted model is obtained by interpolating an SP-based LM and a WR-based LM.
- Model Type III: A specialized model is trained from text from a specific source and speaking style. The adapted model is obtained by interpolating an NP-WR-based LM, an NP-SP-based LM, a BN-WR-based LM, and a BN-SP-based LM.

• Model Type IV: A specialized model is trained from each topic cluster without considering its source or style. The adapted model is obtained by interpolating topic-dependent LMs.
• Model Type V: A specialized model is trained from text in a topic cluster from a single source. The adapted model is obtained by interpolating NP-based topic-dependent LMs and BN-based topic-dependent LMs.
• Model Type VI: A specialized model is trained from text in a topic cluster with a particular speaking style. The adapted model is obtained by interpolating WR-based topic-dependent LMs and SP-based topic-dependent LMs.
• Model Type VII: A specialized model is trained from text in a topic cluster with a source and a particular style. The adapted model is obtained by interpolating NP-WR-based, NP-SP-based, BN-WR-based and BN-SP-based topic-dependent LMs.

In summary, Model Types I, II, and III can be considered LMs adapted to styles. Model Type IV is a topic adapted LM. Model Types V, VI, and VII are LMs adapted to both topics and styles.

2.3 Experimental Conditions

A. Acoustic modeling

Gender-dependent AMs were trained from newspaper read speech corpora (LOTUS [8] and a phonetically balanced sentence speech corpus collected by Tokyo Institute of Technology). The total amount of acoustic training data was 40.3 hours from 68 male and 68 female speakers. 25-dimensional feature vectors consisting of 12 Mel-Frequency Cepstral Coefficients (MFCCs), their delta, and delta energy were used for AM training. The HMM states were clustered using a phonetic decision tree. The number of leaves was 1,000. Each state of the HMM was modeled by a mixture of eight Gaussians. No special tone information was incorporated.

B. Text corpora

Two text corpora were used in our experiments. The first text corpus was collected from newspaper (NP) text. It covered approximately 5 years of news (2003–2007). Numbers and abbreviations were normalized throughout the corpus. The size of the newspaper text corpus was around 140M PMs. Another text corpus was compiled from Thai broadcast news (BN) transcripts [5]. The size of the BN text corpus was around 1M PMs in the experiments on lexical units and around 1.8M PMs in the experiments on LM adaptation. The CMU SLM Toolkit was used to train LMs. Unless stated otherwise, the trained models were 3-grams. All models were built using the Good-Turing smoothing technique.

C. Grapheme-to-phoneme (G2P) conversion

In order to avoid manual work, the pronunciation of each entry was created by means of the NECTEC G2P converter [9]. The tool was implemented on the Probabilistic Generalized Left-to-right Rightmost (PGLR) approach. A sequence of Thai graphemes is parsed by the GLR parser with the context-free grammar (CFG) for syllable construction. The most likely parsed tree is selected and is used to generate the phonetic transcription by table look-up.

D. Recognition experiments

All recognition experiments were performed on a test set that was extracted from the Thai broadcast news speech corpus [5]. Only clean speech with the planned speaking style was selected. The test set consisted of 1,033 spoken utterances (626 male and 407 female utterances), and there was no overlap between the training and test sets.

E. Evaluation method

Since several LMs based on various lexical units were used in the experiments, the lexical units for different LVCSR systems were not the same. Therefore, the test-set perplexities (PP) as well as word error rates (WER) from different experiments cannot be compared. In order to compare LM PP of the different versions of the test set with different text lengths (the number of units in the text), perplexities were normalized. PM error rates (PER) were used as a measurement for the comparison of LVCSR performance. Character error rate (CER) would have been too optimistic since the Thai grapheme is a phonogram, and correctly recognized character inside a word does not directly relate to the correctness of that word. On the other hand, a PM can be thought of as an ideogram like a Chinese character. Therefore, PER reflects a better assessment of recognition accuracies for Thai.

2.4 Experimental Results on Lexical Units

Several experiments were performed to evaluate LVCSR systems based on various lexical units.

A. PM-based LVCSR system

The newspaper text corpus was segmented into PMs. The number of unique PMs after segmentation was around 45k. We have investigated the segmented corpus and listed all PMs with their frequencies. The relationship between the size of PMs that were mostly used and the coverage on the text corpus is plotted in Fig. 1. We found that the PM size of 15k covered around 99.0% of the text corpus. The rest of PMs occurred only once or twice in the segmented corpus. Further analysis showed that these PMs were generated by misspelled words, foreign words, very rare proper nouns, and segmentation errors. We therefore extracted sentences that contained only the 15k most likely PMs to be used for training LMs, although our test set was not restricted to these 15k PMs. The corresponding text corpus contained 139M PMs in total.

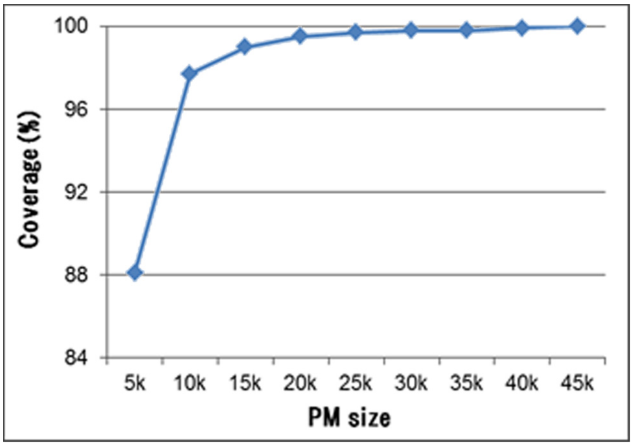


Fig. 1 PM size and coverage.

B. Word-based LVCSR system

Language modeling was performed on sentences that were fully covered by the 15k most frequent PMs. In the word sequences resulting from the SWATH segmentation of these sentences, 139k different words were found. Since the recognizer is limited to 65k vocabulary entries, all sentences not fully covered by the 65k most frequent words were excluded. They represent 17.1% of all the sentences in the corpus. The PMWORD language modeling technique described in Sect. 2.1-D was applied to train a 3-gram LM. There were around 7.4M SWATH-unknown words (out of 106M words) in the segmented text corpus. The PMWORD technique reduced the vocabulary size from 139k to about 43k words. The percentage of sentences excluded from the training text corpus due to unsuccessful G2P conversions was 2.3%.

C. CPM-based LVCSR system

The text corpus covered by the 15k most frequent PMs was used to generate CPMs. In order to keep the number of CPMs restricted to 65k, dynamic thresholds were employed. Instead of using a fixed threshold value, the threshold value was dynamically raised following each iteration, until the number of units being generated reached 65k. It was found that pronunciations of CPMs can largely be determined without context.

D. Comparison of the different systems

Comparison of the PER of the different systems using various lexical units is shown in Fig. 2. The OOV rate of the SWATH system was 2.2%, which was considerably higher than that of all other systems using PMs which were 0.3–0.5%. This is due to a high coverage of PM units. Due to the fact that PMs are short lexical units, the span of PM-based n-gram LM had to be 5 whereas a span of 3 seemed to work well for the word and CPM units. Note that using 5-grams for PM units can be thought of as using approximate 2-grams for traditional word units because the average number of PMs per word unit was about 2.1 PMs. Hence, the performance of the 5-gram LM was not really competitive to other systems. Instead of directly increasing the n-gram span as in the PM system, CPM lexical units can be constructed to resolve this problem. CPMs can also resolve two other problems of PMs regarding acoustic confusability and pronunciation generation. The PER was reduced by 4.5% compared to that of the PM system. Compared to the system based on the traditional word unit (SWATH), the CPM system yielded a 2.8% lower PER. This improvement was statistically significant according to the matched-pair sentence segment test ($p < 0.001$). This shows that the CPM is the most suitable lexical unit for Thai LVCSR systems.

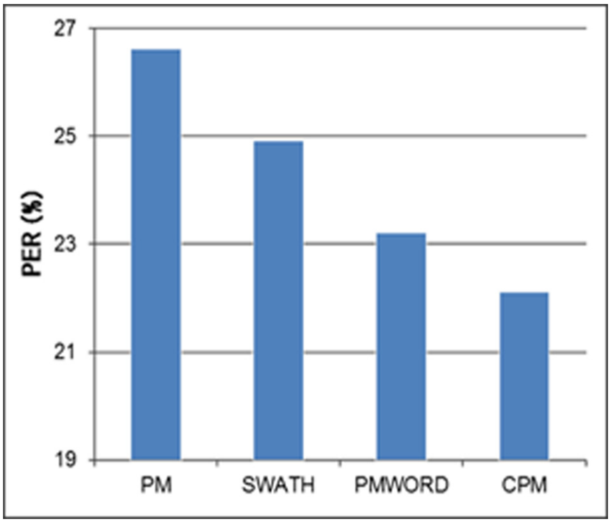


Fig. 2 PER (PM error rate) for each system.

2.5 Experimental Results on Language Model Adaptation

A. Number of topic clusters

LVCSR experiments were performed based on the source, topic and style LM adaptation schemes proposed in Sect. 2.2-E. For Model Types IV, V, VI, and VII, the number of topics in the text corpora was decided first in order to reduce the required computation time by varying the number of topic clusters for each model type. The number of topic clusters was varied from 2 to 20, and Model Type IV was used to test the performance of adapted models. PERs of the systems ranged from 19.5% to 20.7%, and the system with 8 topic clusters performed the best. Therefore, the rest of the experiments were conducted with 8 topic clusters.

B. Unsupervised adaptation experiments

Performance for each of the model types was evaluated following the application of unsupervised adaptation. PPs and

Table 1 PP and PER (%) obtained from unsupervised adaptation.

Adaptation type	Model type	PP	PER (%)
No adaptation	Baseline	207.8	20.2
Source and style adaptation	I	157.5	19.0
	II	168.1	19.2
	III	138.7	18.2
Topic adaptation	IV	171.6	19.5
Source, topic and style adaptation	V	127.1	18.0
	VI	146.1	18.5
	VII	129.5	18.3

PERs of all model types are shown in Table 1. Recognition hypotheses from the baseline system using one LM trained with combined BN and NP text with 20.2% PER were used for LM adaptation. Adaptation to the text source and speaking styles performed in Model Type I and II successfully reduced PP and PER. Moreover, Model Type III, which performed adaptation to both text source and speaking styles, further lowered PP and PER significantly. Topic adaptation achieved by Model Type IV was able to decrease PP and PER but not as significantly as Model Types I to III. The rest of the model types utilizing source, topic and style adaptation gave similar results to Model Type III. The best result was obtained using Model Type V.

2.6 Summary

Since there is no word boundary in Thai, we have investigated various lexical units for a Thai LVCSR system. The pseudo-morpheme (PM) was introduced, and the compound pseudo-morpheme (CPM) constructed by PM merging was proposed. The construction of CPMs can circumvent the problem of the ambiguous definition of Thai words, word segmentation errors, and the ambiguity of PM pronunciations. The experiments showed that the CPM is the best lexical units for Thai LVCSR.

We have proposed a simple rule-based speaking style classification to categorize spoken and written style text, based on the existence of specific spoken style words. Various kinds of *n*-gram models adapted to topics and styles were investigated, and could successfully reduce test-set perplexity and recognition error rates. An analysis of experimental results showed that we could employ written style text from newspapers to alleviate the sparseness of the broadcast news transcript text. However, spoken style text from the broadcast news corpus was still essential for building a reliable LM. Therefore, for a resource deficient language like Thai, a broadcast news corpus including a large number of spoken style utterances should be constructed with the highest priority, to which newspaper text can be added to model written style speech and increase the topic coverage in the broadcast news speech recognition.

3. Adaptation to Pronunciation Variations in Indonesian Spoken Query-Based Information Retrieval

3.1 Pronunciation Problems in Bahasa Indonesia

A. *Proper noun problems*

Developing accurate pronunciations for proper nouns is difficult in many languages. The most commonly cited reason is that names are derived from many source languages from many regions and countries. As described in the introduction, names in Bahasa Indonesia (BI) are influenced by hundreds of regional languages and certain foreign languages. The pronunciations often do not follow the rules of the standard language. There is variability due to regional influences and even personal preferences. There are many variations in writing proper nouns with similar sounds that tend to confuse people. For example, “Khairul”, “Koirul”, “Khoirul”, “Chairul”, and “Choirul” are proper nouns derived from foreign proper nouns that consist of sounds not existing in BI. Furthermore, proper nouns grow in number through the process of human creativity in making names and the process of language assimilation. Therefore, no authoritative pronunciation lexicon has been developed for proper nouns in BI. People cannot correctly read the names of unfamiliar persons. Even when the names are familiar, people are sometimes unaware of the correct pronunciations of other people’s names. When proper nouns are less familiar, people tend to pronounce them more carefully than other words; they pronounce them less confidently, less fluently, too softly, or slowly.

B. *English word problems*

Despite the fact that the Indonesian government has defined rules on how to transform foreign words into Indonesian words, people tend to use the original foreign words, especially English words, on both formal and informal occasions. This is mainly because the official translated Indonesian words are not familiar to Indonesian people. This phenomenon frequently appears in news articles, technical books, and conversations. Some famous politicians, actors and teachers use English terms in public speeches, and these terms together with pronunciation variations depending on the fluency of the English speaker tend to become popular. It is also common for Indonesian people to input queries to search engines by mixing Indonesian and English words.

3.2 Infinite Network-Based IR

An Inference Network (IN) model is used in our spoken query-based IR system. The IN model is basically a directed acyclic graph (DAG) of a Bayesian Network [10]. The network is used to model documents and their content (document sub-network, DN) and to model queries (query sub-network, QN), as shown in Fig. 3.

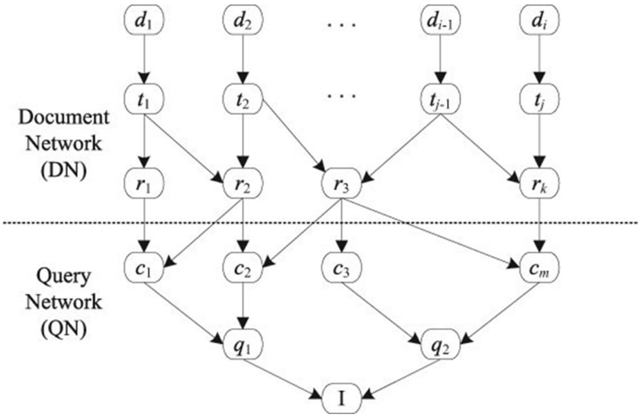


Fig. 3 Inference network.

The DN consists of three layers of nodes: document nodes (d_i nodes) that represent the events for which the documents are observed, text representation nodes (t_j nodes), and representation nodes (r_k nodes) that represent the concepts in the collection. They can be used as indexing features for the document. A causal link represented as a down arrow between nodes indicates that the parent nodes are related to the child nodes. The child nodes inherit information from the parent nodes. Each causal link contains a conditional probability or a weight to indicate the strength of the relationship. Each node is evaluated using the value of the parent nodes and the conditional probabilities/weight. This evaluation basically relies on an indexing weight, such as the *tf-idf* weighting.

The QN consists of three layers of nodes: query concept nodes (c_m nodes), query nodes (q nodes), and a user information-need node (I node). Each query node contains a specification in the form of link matrices to describe the dependency of the query on its parent query concepts. In the retrieval process, to form the complete IN, the query sub-network is attached to the document sub-network when the concepts in both networks are the same. After the attachment phase, the complete IN is evaluated for each document node to compute the probability of the relevance to the query. The evaluation is initialized by setting the output of one document node to true (1), and the output of all the other document nodes to false (0). This procedure is applied to each document node in turn. The probability of a document's relevance is taken from the final node I and this is what is used in the ranking.

3.3 LVCSR and IR Systems

A. Corpus for acoustic modeling

An ideal Indonesian speech corpus should cover not only all phones in Bahasa Indonesia, but also those of other Indonesian dialects. Since it was difficult to collect data on all Indonesian dialects, we collected Bahasa Indonesia (BI) speech data from 20 Indonesian speakers (11 males and 9 females) belonging to the five largest Indonesian tribes: Javanese, Sundanese, Madurese, Minang, and Batak. Each speaker was asked to read 328 phonetically balanced sentences selected from the Information and Language System (ILPS) document collections [11]. Those document collections were taken from an Indonesian national newspaper and a magazine. Speech was recorded in a quiet room, and digitized at a 16 kHz sampling rate. The total size of the speech corpus after manual sentence segmentation was 14.5 hours.

B. Corpus for language modeling

A document collection developed by the ILPS group [11] was used for building the LM. The articles in the corpus were taken from popular Indonesian newspaper and magazine sites. Some text processing was conducted on the corpus. The text corpus consists of 615k sentences, 9,853k words, 130k word vocabulary, and 16.02 words per sentence on average. Half of the articles in the newspaper corpus were used to build the LM. The total number of words used to build the LM was 3,125k.

C. Lexicon

We developed a lexicon from the ILPS corpus by selecting 26,581 words that occur in the text corpus more than three times. An Indonesian grapheme-to-phoneme tool developed in our laboratory was then employed to add word pronunciations to the lexicon.

D. Baseline system

HMM-based AMs and *n*-gram LMs were used to develop the LVCSR system [12]. The 1st through 12th order MFCCs were extracted every 10 ms and delta features of MFCCs and energy were also incorporated. We used 32 Gaussian mixtures per state to train context-dependent HMMs. The total number of shared states was 1,746. The 2-gram and 3-gram LMs were smoothed using the Good-Turing back-off technique. The 3-gram LM had a test-set perplexity of 61.04 and an OOV rate of 1.75% for the spoken queries described in the next subsection.

E. Text and speech corpus for IR experiments

Since there is no standard evaluation corpus for spoken query IR in BI, we created a test set of spoken queries for the experiments. The queries were derived from the BI collection developed by the ILPS [11] and from the collection developed by the School of Computer Science and Information Technology, RMIT University, Australia [13]. There are 35 query topics available for the magazine corpus ("magazine A") and for the newspaper corpus in the ILPS corpus ("newspaper corpus"), respectively. In [13], there are 20 query topics ("magazine B"). In total, there are 90 query topics. Both IR collections, which contain documents, queries, and exhaustive relevance judgments, are stored in the TREC format. They can be used in the TREC-like ad hoc evaluations with standard TREC retrieval and evaluation tools.

For each topic of the query, we recorded spoken queries by 20 native Indonesian speakers (11 males, 9 females), each uttering 270 queries, consisting of three different query lengths: short (2–4 words), medium-length (4–8 words), and long (8–16 words), on 90 different topics. These speakers were different from those used for training the AM. There were 5,400 spoken queries in total.

The Indonesian newspaper text corpus provided by ILPS was divided into two parts; the first part was used to train the LM for LVCSR, and the second part was used as the document collection for the newspaper IR system, while the whole collection of magazines A and B were used for the magazine IR system. None of the articles was used to train the LM.

3.4 Solutions to the Proper Noun and English Word Problems

A. *Proper noun adaptation (PNA)*

The average word accuracy of the baseline system was 75.1%, and it was found that the majority of the misrecognized words came from proper nouns (23% error from regular proper nouns, and 14% error from abbreviated proper nouns). There were 10,720 proper nouns in the test data. In order to improve the proper noun recognition accuracy, we decided to adapt the general AM to create a proper noun specific model [14]. Although other methods, such as knowledge-based ones, could have been used to resolve the pronunciation variation problem, such a method could not be used in the Indonesian case because of the difficulties in developing accurate pronunciation rules for proper nouns, as described in Sect. 3.1-A.

The procedure to build the proper noun specific models is as follows:

(1) Extract the proper noun words from the speech corpus used to train the baseline AM as the adaptation data (14,840 words).
(2) Make proper noun specific phone HMMs by applying supervised adaptation based on MLLR using eight regression classes to the baseline acoustic HMMs.
(3) Combine the baseline HMMs and the proper noun-specific HMMs. Thus, the number of HMMs in the proper-noun-adapted system is twice the number of HMMs in the baseline system. However, the proper noun-specific HMMs are used only for proper nouns.
(4) Add the proper noun pronunciation to the baseline lexicon. Using the proper noun dictionary provided in the Indonesian Standard Dictionary (Kamus Besar Bahasa Indonesia), we found 3,216 proper nouns in the lexicon, and these pronunciations were added.

B. *English to Indonesian phoneme mapping (EIPM)*

The second largest set of misrecognized words in the baseline system consisted of foreign words (24% error). In the testing data, most of the foreign words were English words and the rest were from languages such as Arabic. Thus, we focused on English words. Of the 20 speakers in the testing data, eight speakers produced pronunciations close to the original English sounds, while 12 speakers produced pronunciations at English levels varying from beginner and advanced.

Several techniques to handle foreign words in general speech recognition have been investigated as a multi-dimensional problem [15]. There are several ways to map the phoneme symbols across languages: either knowledge-based or data-driven approaches. In the data driven approach, AMs are trained based on a phonetic transcription for each foreign word, which needs a large amount of data. Unfortunately, a large database of the foreign words pronounced by Indonesian speakers is not available for BI. Therefore, we have decided to use a knowledge-based approach using phonetic mappings based on similarities between languages [16]. In [17], an Indonesian-to-English phoneme mapping was developed to rapidly build an Indonesian speech recognizer using an English corpus to train an AM. We have used the English-to-Indonesian phoneme mapping rules reported in [17], modifying several rules. The phonemes "er", "ey", "ii", "ng-k", and "sha" are not available in [17] but are available in the CMU (Carnegie Mellon University) phoneme set that we used in the experiment; thus we added these phonemes to our rules.

English word pronunciations were added to the lexicon. We used the CMU lexicon, consisting of 39 English phonemes, as the reference for the English words. The resulting English word list was filtered, using the most standard Indonesian dictionary made by the Indonesian government, the Kamus Besar Bahasa Indonesia (KBBI), as a reference. We consulted the dictionary to delete some words that were recognized as English words but also existed in the KBBI to avoid ambiguities in the recognition. Some English words having the same pronunciation as the Indonesian words were also deleted from the list to avoid redundancy.

By using the CMU lexicon as the reference, 4,050 words were recognized as English words. After filtering the resulting English word list using the KBBI and removing the words with the same pronunciation as Indonesian, the number of English words was reduced to 1,939.

C. *LVCSR and IR results*

Table 2 shows word accuracy and MRR (mean reciprocal rank) results for LVCSR and IR, respectively, for the three systems: baseline system, the system with proper noun adaptation (PNA), and the system with English to Indonesian phoneme mapping (EIPM). PNA and EIPM processes were applied incrementally to the baseline system. LVCSR results were fed to the IR system after removing the stop words in BI [18]. Correct queries with no LVCSR errors were also given to the IR in order to compare their results

Table 2 Accuracies for LVCSR and MRR of IR for the baseline system (Baseline), the system with proper noun adaptation (PNA), and the system with English to Indonesian phoneme mapping (EIPM). IR results with text queries are also shown. VSM: standard vector space method, IN: IN-based method.

	LVCSR accuracy (%)	MRR by VSM	MRR by IN
Baseline	75.1	62.2	71.1
+PNA	76.8	63.6	72.0
+EIPM	77.8	64.9	73.2
Text query	(100.0)	77.3	80.5

with those obtained from LVCSR. IN-based IR was compared with the classical VSM-based IR. For all the test data, the PNA and the EIPM processes improved both the LVCSR and IR performance. Based on a detailed analysis, it was found that 27% of proper noun recognition errors were eliminated by the PNA process, and 53% of the foreign word recognition errors were removed by the EIPM process. IN-based IR outperformed the classical VSM approach for both spoken queries and text queries. IN-based IR gave a larger improvement for spoken queries than for text queries, which means that IN-based IR is more robust than traditional VSM-based IR for spoken queries.

3.5 Summary

We have proposed methods to reduce the LVCSR errors caused by proper noun and foreign word pronunciation variations in order to improve the performance of spoken query-based Indonesian IR. To improve the proper noun recognition accuracy, we have added proper noun-specific AMs made by adapting the baseline AM using MLLR. To increase the English word recognition accuracy, rule-based English-to-Indonesian phoneme mapping was applied to the English words in the lexicon. Both techniques significantly reduced the word error rate of the spoken queries and there was no negative effect.

We have found that IN-based IR is more robust to the LVCSR errors than the IR approach based on the vector space model.

We are currently investigating several query expansion methods to improve the IR performance.

4. Vocabulary Expansion through Automatic Abbreviation Generation for Chinese Spoken Query-Based Information Retrieval

4.1 Chinese Abbreviations

In Chinese spoken query-based IR, official names of organizations are frequently abbreviated for efficiency and convenience, since the full-names are sometimes very long and difficult to remember. While English abbreviations are usually formed as acronyms, Chinese abbreviations are much more complex. Chinese abbreviations are generated by three methods [19]: reduction, elimination, and generalization. In both the reduction and the elimination methods, characters are selected from the full-name, but they are not necessarily the first character of a word, and their order is sometimes changed. Note that our research does not cover the case when the order is changed. Elimination means that one or more words in the full-name are ignored completely, while a reduction requires that at least one character is selected from each word. Figure 4 shows two examples produced by elimination, where at least one word is skipped. Generalization, which is used to abbreviate a list of similar terms, usually produces a word which is composed of the number of the terms and a shared character across the terms. An example is "三军" (three forces) for "陆军, 海军, 空军" (army, navy, air force). This is the most difficult scenario for abbreviations and is not considered in our research.

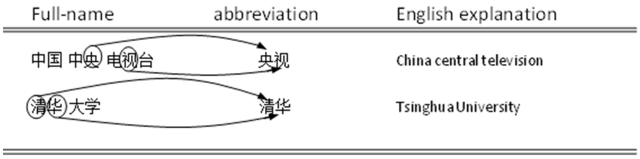


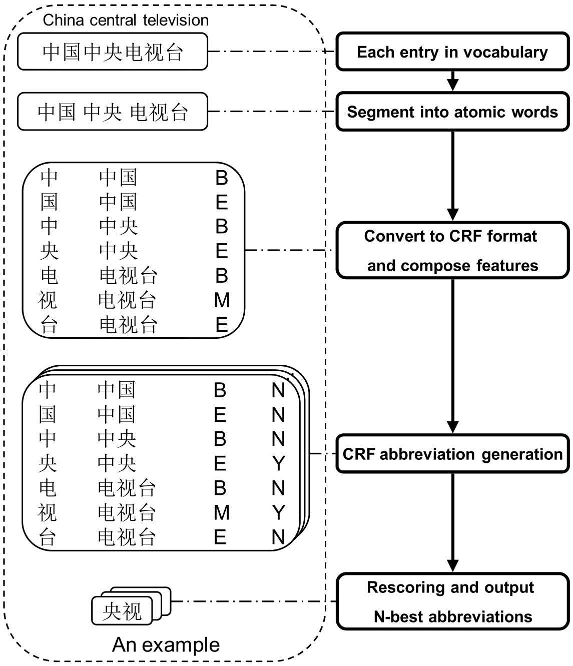
Fig. 4 Chinese abbreviation examples.

Although the abbreviation problem occurs in both text and spoken query-based search applications, the abbreviation problem is more serious in the latter case, since it causes an OOV problem. A simple abbreviation dictionary cannot solve the problem. This is because no such dictionary exists, and new organization names keep on coming into use. Although there has been a considerable amount of research on extracting full-name and abbreviation pairs in the same document for obtaining abbreviations [20]–[22], their performance is not yet satisfactory. We have therefore investigated a method to automatically generate abbreviations given a full-name. Chang and Lai [23] have proposed using HMMs to generate abbreviations from full-names. However, their method assumes that there is no word-to-null mapping, which means that every word in the full-name has to contribute at least one character to the abbreviation. This assumption does not hold for named entities that have word skips in the abbreviation generation.

4.2 Chinese Abbreviation Modeling

The structure of our abbreviation generation system is shown in Fig. 5 [24], [25]. We propose a new hybrid abbreviation generation method for Chinese, formalizing the Chinese abbreviation generation problem into a character tagging problem, and using conditional random fields (CRF) [26] as the tagging tool. Each character in a full-name is tagged using a binary variable with the values of either *Y* or *N*. *Y* stands for a character used in the abbreviation and *N* means the character was not used. By using CRFs, a list of abbreviation candidates with associated probability scores is obtained. An example is shown in Fig. 5.

We also use the prior conditional probability of the length of the abbreviations given the length of the full-


China central television
中国中央电视台
Each entry in vocabulary
中国 中央 电视台
Segment into atomic words
Convert to CRF format and compose features
CRF abbreviation generation
央视
Rescoring and output N-best abbreviations
An example

中	中国	B
国	中国	E
中	中央	B
央	中央	E
电	电视台	B
视	电视台	M
台	电视台	E

中	中国	B	N
国	中国	E	N
中	中央	B	N
央	中央	E	Y
电	电视台	B	N
视	电视台	M	Y
台	电视台	E	N

Fig. 5 Abbreviation generation procedures.

names to complement the CRF probability scores. In addition, we apply the full-name and abbreviation candidate cooccurrence statistics obtained on the web to increase the correctness of the abbreviation candidates. Although we need a corpus of full-name-abbreviation pairs, it does not have to be very large. We can predict variations of full-names that are not covered by the full-name-abbreviation training corpus. Finally we make use of the abbreviation output through vocabulary expansion in spoken query-based search applications by adding top-N generated abbreviation candidates to the vocabulary. Details are given in the following subsections.

A. *Segmenting organization names into atomic words*

Chinese is written continuously with no word boundaries in text; as a result, a segmenter is usually needed before any further processing can be done. Almost all Chinese segmenters treat a named entity as a single word. In our abbreviation modeling, instead of treating an organization name as a word, we segment each organization name into a composite list of atomic words using a 2-tag CRF segmenter as shown in Fig. 5. The segmenter was trained using the "Penn Chinese Treebank" corpus [27], in which Chinese text are manually segmented into atomic words.

B. *CRF for abbreviation modeling*

A CRF is an undirected graphical model and assigns the following probability to a tag sequence $T = t_1 t_2 \ldots t_N$, given an input sequence $F = c_1 c_2 \ldots c_N$,

$$P(T|F) = \frac{1}{Z(F)} \exp\left(\sum_{n=1}^{N} \sum_{k} \lambda_k f_k(t_n, t_{n-1}, F, n) \right) \qquad (1)$$

Here, f_k is the feature function for the k-th feature, λ_k is the parameter which controls the weight of the k-th feature in the model, and $Z(F)$ is the normalization term that ensures that the probability of all label sequences sum to 1. CRF training is usually performed through the typical L-BFGS algorithm and decoding is performed with the Viterbi algorithm. In this research, we use an open source toolkit "crf++" [28].

C. *Feature selection for the CRF*

In the CRF method, a feature function describes a cooccurrence relation, and it is defined as $f_k(t_n, t_{n-1}, F, t)$ (Eq. (1)). f_k is usually a binary function, and takes the value 1 when both observation F and the transition t_{n-1} to t_n are observed. In our abbreviation generation model, we use the following features:

(1) Current character: the 1st column in our CRF format in Fig. 5.
(2) Current word: the 2nd column in the CRF format in Fig. 5.
(3) Position of the current character in the current word: the 3rd column in CRF format in Fig. 5, where B, M, E stand for the beginning, middle and end of a word, respectively.
(4) Combination of the features (2) and (3).

In addition to the features above, we have examined other contextual information, such as previous word, previous character, next character, and other local features like the length of the word, but these features did not improve the performance. This is probably due to the sparseness of the training data.

D. *Improvement via incorporating a length model*

From the CRF we can obtain a list of abbreviation candidates with their conditional probabilities, written as $P'(A|F)$, where variable A is the abbreviation and F is the full-name. There is a strong correlation between the lengths of organizations' full-names and their abbreviations. Therefore, $P'(A|F)$ is weighted (multiplied) by the length model based on the discrete probability, $P(M|L)$, where variables M and L are lengths of the abbreviation and the full-name, respectively. Since there is a data sparseness problem in the modeling of $P(M|L)$, we use a simple smoothing method to avoid zero probabilities for unseen length mapping relations, which are mainly for full-names longer than 10 characters. The probability $P'(A|F)$ is normalized over the same length abbreviations before the weighting.

E. *Improvement via a web search engine*

Co-occurrence of a full-name and an abbreviation candidate in the same document or web page can be a clue as to the correctness of the abbreviation. We decided to use web information to assist our abbreviation re-ranking. We use the "abbreviation candidate + full-name" as queries and input them to the most popular Chinese search engine "www.baidu.com", and then we use the number of hits as the metric to perform re-ranking. The number of hits is related to the number of pages that contain both the full-name and the abbreviation. The bigger the number of hits is, the higher the probability that the abbreviation is correct. We simply multiply the probability score weighted by the length model with a hit ratio to the total number of hits for top-30 candidates, and re-rank the top-30 candidates. The reasons why we decided to re-rank the top-30 candidates are as follows: the coverage of the top-30 candidates after incorporating the length model reached 92%, and we also wanted to keep the access load to the web search engine as small as possible.

F. Vocabulary expansion

Top-*N* abbreviation candidates generated using the method described above are added to the original vocabulary. There is a tradeoff between the coverage of the abbreviation and the automatic speech recognition (ASR) ambiguity caused by an enlarged vocabulary. If *N* is too small, the possibility that correct abbreviation is not included in the vocabulary is high, and if *N* is too large, the confusability within words in the vocabulary increases and the ASR performance decreases.

4.3 Experiments

A. Abbreviation generation experiment

The corpus we used for abbreviation training and evaluation came from two sources: the book "modern Chinese abbreviation dictionary" [29] and Wikipedia. The first source has around 4,500 entries of full-name and abbreviation pairs, but most of them are not organization names. We selected all the entries of organization names. Another source was Wikipedia; we used the keyword "简称" (abbreviation or abbreviated as) as a query to search Wikipedia, and manually extracted the organization full-name and abbreviation pairs from the articles returned by the search. Then we merged the two sources, and altogether we collected 1,945 pairs of organization full-names and their abbreviations. The data was randomly divided into two parts, a training set with 1,298 pairs and a test set with 647 pairs, in which there are 1,202 and 622 unique full-names, respectively. There is no overlap between training and testing pairs in terms of full-names, but there are overlaps in terms of words constituting the full-names.

We added up to 10 abbreviation candidates into the vocabulary of our spoken query-based search application for each organization name, and hence we mainly evaluated top-10 coverage of the abbreviation modeling, which is defined as the ratio of abbreviations covered by the top-10 candidates (similar to the recall in IR). For comparison, we also measured top-*N* coverage for different *N*'s ($n = 1, 2, \ldots, 10$). Figure 6 shows the coverage obtained using the CRF method with all the features, and the improvements achieved by including the length feature and the web search hits. The CRF gives top-10 coverage of 79.9%. Both the length model and the web search engine show significant improvement over the CRF baseline and the coverage increases to 88.3%.

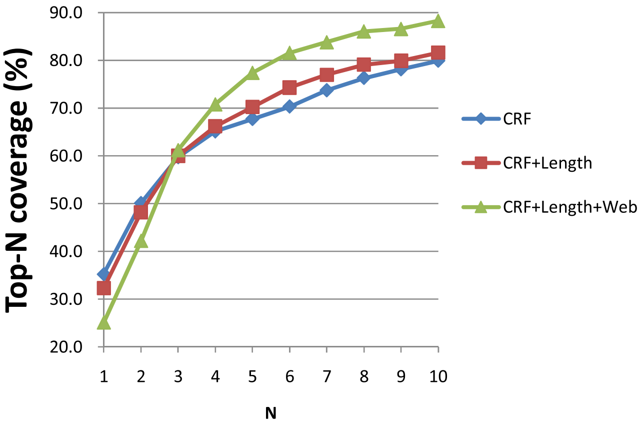


Fig. 6 Top-*N* coverage for different methods.

B. Spoken query-based full-name search experiment

The training corpus for our ASR AM was "King-ASR-018" [30]. It was recorded from 850 speakers (430 females and 420 males) with various ages, accents and education levels. The corpus contains 150 hours of manually transcribed clean speech data. Our AM uses a toneless phoneme set and a typical 3-state HMM structure with 25 dimension features, composed of 12 MFCC features plus their delta features and delta energy. We set the number of Gaussian mixtures to 16.

We selected 400 organization full-names from the test set used in the previous subsection and collected abbreviations along with speech data for them from 20 human subjects. Each subject was requested to process 80 organization names and each organization name was allocated to 4 subjects. As a result, the data were able to cover wide variations of the abbreviations. Since we noticed that multiple abbreviations for one full-name were very common, we collected multiple abbreviations for reference along with the speech data from each subject. In total, we collected 2,200 abbreviation utterances from the 20 subjects, in which there were 783 unique abbreviations and the average number for each full-name was 1.96. The top-10 coverage by our proposed method for the collected abbreviations was 91.8%.

In our spoken query-based search experiment, inputs are utterances of organization names (full-names or abbreviations) and the outputs are corresponding full-names in the vocabulary. We measure the search performance by the accuracy of output full-names, and we call it full-name search

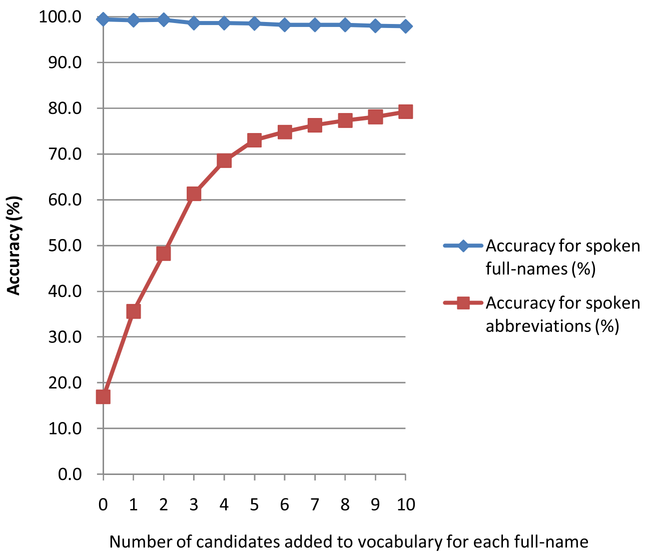


Fig. 7 Full-name search accuracy.

accuracy. If the abbreviation is not included in the vocabulary, the search will fail theoretically for the abbreviated utterance. When the abbreviation is long enough to be very similar to the full-name however, it is also possible to recognize it correctly and get the correct search result.

In our experiment, we started from the vocabulary containing full-names only, and then added the abbreviation candidates to the vocabulary, one by one from top-1 to top-10. Figure 7 shows that, as the added abbreviation candidates increase, the search accuracy by abbreviation utterances keeps increasing from 16.9% with no abbreviation in the vocabulary to 79.2% with 10 abbreviation candidates in the vocabulary for each entry. We can also see that the accuracy for full-name utterances slightly decreases from 99.4% to 97.9%.

4.4 Summary

We have proposed a novel hybrid method for automatic Chinese abbreviation generation and successfully applied abbreviation modeling to spoken query-based search through vocabulary expansion. Our abbreviation model formalized the task into a character tagging problem first and used the CRF as the tagging tool; then we made use of length mapping relation and web search engine to re-score the abbreviation candidates from the tagging task. We achieved 88.3% top-10 coverage on our abbreviation test data. In our search experiment, we collected a test corpus using human subjects, for which the top-10 coverage of abbreviations generated by our method was 91.8%. After adding top-10 abbreviation candidates to the vocabulary in the search experiment, the full-name search accuracy for abbreviation utterances was increased from 16.9% to 79.2%.

Although our experiments were performed on Chinese, our method can be applied to other languages that exhibit similar abbreviation phenomena, such as Japanese. In regard to abbreviation modeling, we were only able to use 1,298 full-name-abbreviation pairs for training. Future work will include making better use of web data, for example, to extract the pairs from a very large amount of web text. Rule-based methods could be compared or combined, too. We would also like to investigate whether it is possible to set a threshold for the probability values to select candidates.

We have so far evaluated our proposed method by recognizing abbreviations or full-names uttered in isolation. Since these words are usually spoken in a query sentence or a phrase, our future work includes evaluation in the context of continuous speech recognition.

5. Conclusion

This paper has presented selected topics from our recent LVCSR research for Asian languages. One of the typical features of several Asian languages is that there is no spacing between words in the written form of the language. Since it is effective to use statistical LMs based on word-like units in LVCSR, such as 2-grams and 3-grams, it is crucial to define the word-like units even for these languages. In order to avoid troublesome manual work, these units need to be automatically defined and extracted from text. For Thai, we have proposed using compound pseudo-morphemes as word-like units.

Another feature of several Asian languages is significant difference between written and spoken sentences. For Thai, we have proposed an automatic spoken-style adaptation method for an LM.

Asian languages include considerable variation in pronunciations caused by dialects and foreign words. For Indonesian, we have proposed constructing proper noun-specific AMs and rule-based English-to-Indonesian phoneme mapping to improve the performance of a spoken query-based IR system.

Asian languages include many abbreviation words, especially for long organization names. If the abbreviations are not included in the dictionary for LVCSR, they cause an OOV problem. For Chinese, we have proposed a method to automatically generate abbreviations from full-names and expand the dictionary by adding them.

Experimental evaluation results have confirmed that all these methods are effective in improving speech recognition as well as spoken query-based IR. Although some of these methods are specific to the languages that we have investigated, most of them are expected to be applicable to not only other Asian languages but also various other languages in the world.

References

[1] C. Wutiwiwatcai and S. Furui, "Thai speech processing technology: A review," Speech Commun., vol.49, no.1, pp.8–27, 2007.

[2] NECTEC, n.d., "SWATH (Smart Word Analysis for THai)," http://www.links.nectec.or.th/download.php

[3] W. Aroonmanakun, "Collocation and Thai word segmentation," Proc. SNLP and Oriental COCOSDA Workshop, pp.68–75, 2002.

[4] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai, and S. Furui, "Lexical units for Thai LVCSR," Speech Commun., vol.51, no.4, pp.369–389, 2009.

[5] M. Jongtaveesataporn, C. Wutiwiwatchai, K. Iwano, and S. Furui, "Thai broadcast news corpus construction and evaluation," Proc. LREC, 2008.

[6] A. Chotimongkol, K. Saykhum, P. Chootrakool, N. Thatphithakkul, and C. Wutiwiwatchai, "LOTUS-BN: A Thai broadcast news corpus and its research applications," Proc. Oriental COCOSDA, pp.44–50, 2009.

[7] M. Jongtaveesataporn and S. Furui, "Topic and style-adapted language modeling for Thai broadcast news ASR," Proc. INTERSPEECH 2010, 2010.

[8] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara, and N. Thatphithakkul, "Thai speech corpus for Thai speech recognition," Proc. Int. Conf. Speech Databases and Assessments (Oriental-COCOSDA) 2003, pp.54–61, 2003.

[9] P. Tarsaku, V. Sornlertlamvanich, and R. Thongprasirt, "Thai grapheme-to-phoneme using probabilistic GLR parser," Proc. EUROSPEECH 2001, pp.1057–1060, 2001.

[10] H.R. Turtle and W.B. Croft, "Inference networks for document retrieval," ACM Trans. Information Systems, vol.9, no.3, pp.187–222, 1991.

[11] F.Z. Tala, J. Kamps, K. Muller, and M. de Rijke, "The impact of stemming on information retrieval in Bahasa Indonesia," Proc. CLIN, the Netherlands, 2003.

[12] D.P. Lestari, K. Iwano, and S. Furui, "A large vocabulary continuous speech recognition system for Indonesian language," Proc. 15th Indonesian Scientific Conference in Japan (ISA-Japan), pp.17–22, Hiroshima, Japan, 2006.

[13] J. Asian, H.E. Williams, and S.M.M. Tahaghoghi, "A testbed for Indonesian text retrieval," Proc. 9th Australasian Document Computing Symposium (ADCS 2004), pp.55–58, Melbourne, Australia, 2004.

[14] D.P. Lestari and S. Furui, "Adaptation to pronunciation variations in Indonesian spoken query-based information retrieval," IEICE Trans. Inf. & Syst., vol.E93-D, no.9, pp.2388–2396, Sept. 2010

[15] R. Eklund and A. Lindstrom, "Pronunciation in an internationalized society: A multi-dimensional problem considered," Proc. FONETIK 96, Swedish Phonetics Conference, TMH-QPSR 2/1996, pp.123–126, Nasslingen, 1996.

[16] C. Nieuwondt and E.C. Botha, "Cross-language use of acoustic information for automatic speech recognition," Speech Commun., vol.38, pp.101–113, 2002.

[17] S. Sakti, K. Markov, and S. Nakamura, "Rapid development of initial Indonesian phoneme-based speech recognition using the cross-language approach," Proc. Oriental COCOSDA, pp.38–43, Jakarta, Indonesia, 2005.

[18] F.Z. Tala, A study of stemming effects on information retrieval in Bahasa Indonesia, M.Sc. Thesis, Appendix D, pp.39–46, University of Amsterdam, 2003.

[19] H. Wing and D. Lee, A study of automatic expansion of Chinese abbreviations, MA Thesis, The University of Hong Kong, 2005.

[20] Z. Li and D. Yarowsky, "Unsupervised translation induction for Chinese abbreviations using monolingual corpora," Proc. Annual Meeting of the Association for Computational Linguistics, pp.425–433, 2008.

[21] J.-S. Chang and W.-I. Teng, "Mining atomic Chinese abbreviation pairs: A probabilistic model for single character word recovery," Proc. Annual Meeting of the Association for Computational Linguistics SIGHAN Workshop, pp.17–24, 2006.

[22] G. Fu, K.-K. Luke, G. Zhou, and R. Xu, "Automatic expansion of abbreviations in Chinese news text," Lect. Notes Comput. Sci., vol.4182, pp.530–536, 2006.

[23] J.-S. Chang and Y.-T. Lai, "A preliminary study on probabilistic models for Chinese abbreviations," Proc. Annual Meeting of the Association for Computational Linguistics SIGHAN Workshop, pp.9–16, 2004.

[24] D. Yang, Y.-C. Pan, and S. Furui, "Vocabulary expansion through automatic abbreviation generation for Chinese voice search," Proc. INTERSPEECH 2009, pp.728–731, 2009.

[25] D. Yang, Y.-C. Pan, and S. Furui, "Chinese abbreviation generation using conditional random field," Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp.273–276, 2009.

[26] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. International Conference on Machine Learning, pp.282–289, 2001.

[27] F. Xia, M. Palmer, N. Xue, M.E. Okurowski, J. Kovarik, F.-D. Chiou, S. Huang, T. Kroch, and M. Marcus, "Developing guidelines and ensuring consistency for Chinese text annotation," Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 2000.

[28] T. Kudo, http://crfpp.sourceforge.net/

[29] H. Yuan and X. Ruan, Modern Chinese abbreviation dictionary, Yuwen Press, 2002.

[30] Kingline-Data-Center, http://www.speechocean.com/productdetail(e).asp?id=king-asr-018

Sadaoki Furui received B.S., M.S. and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan in 1968, 1970 and 1978, respectively. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 900 published articles. He has received Paper Awards and Achievement Awards from the IEEE, the IEICE, the ASJ, the ISCA, the Minister of Science and Technology, and the Minister of Education, and the Purple Ribbon Medal from the Japanese Emperor.