PAPER *Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction*

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

Dessi Puji LESTARI[†a)], ***Nonmember and*** **Sadaoki FURUI**[†b)], ***Fellow, Honorary Member***

SUMMARY Recognition errors of proper nouns and foreign words significantly decrease the performance of ASR-based speech applications such as voice dialing systems, speech summarization, spoken document retrieval, and spoken query-based information retrieval (IR). The reason is that proper nouns and words that come from other languages are usually the most important key words. The loss of such words due to misrecognition in turn leads to a loss of significant information from the speech source. This paper focuses on how to improve the performance of Indonesian ASR by alleviating the problem of pronunciation variation of proper nouns and foreign words (English words in particular). To improve the proper noun recognition accuracy, proper-noun specific acoustic models are created by supervised adaptation using maximum likelihood linear regression (MLLR). To improve English word recognition, the pronunciation of English words contained in the lexicon is fixed by using rule-based English-to-Indonesian phoneme mapping. The effectiveness of the proposed method was confirmed through spoken query based Indonesian IR. We used Inference Network-based (IN-based) IR and compared its results with those of the classical Vector Space Model (VSM) IR, both using a tf-idf weighting schema. Experimental results show that IN-based IR outperforms VSM IR.

key words: *proper noun, foreign word, Indonesian ASR, spoken query-based IR*

1. Introduction

Advances in speech processing technology, especially in speech recognition and speech synthesis, are providing opportunities to use speech as a means of natural interaction between humans and machines. There are cases in which it is impossible or inconvenient to use a keyboard as an input device, such as when driving a car. Moreover, widespread access to the Internet has made information resources easily accessible from everywhere. However, a very large part of the world's population does not have access to computers or the Internet; in many rural areas, the most convenient means of communication is telephone or mobile phone. This means if we want to enable all users to take advantage of the information stored in digital repositories, we need to devise ways to enable voice input rather than text input for queries.

Sometimes the error rate of a speech recognition system severely hampers the information retrieval (IR) system's effectiveness. If spoken key terms are incorrectly recognized, relevant documents may not be able to be retrieved, and irrelevant documents containing error terms may be retrieved instead. The ranking of the retrieved documents becomes lower as the number of misrecognized terms in the transcribed query increases.

This paper focuses on Indonesian spoken-query IR. Indonesian LVCSR (Large Vocabulary Continuous Speech Recognition) systems have much more difficulty in recognizing proper nouns and English words than other words of the language. Proper nouns are usually the keywords of queries in most IR systems. Foreign terms, especially English terms, are also often keywords, especially in IR systems for domains such as science, engineering, economy, and politics. Thus, a high recognition error rate for proper nouns and English words can significantly impair IR performance. To solve this problem, we propose a proper noun adaptation method based on the maximum likelihood linear regression (MLLR) and rule-based English-to-Indonesian phoneme mapping. Our experiments detailed below proved that these two techniques improve the recognition accuracy of spoken queries; hence, they improve IR performance.

2. Related Work

A classical method of text query-based IR is to use the tf-idf (term frequency-inverse document frequency) weighting schema in the vector space model [1]. Some researchers have also used this technique for spoken query-based IR [2], [3]. Reference [2] reported that word errors reduce the performance of spoken query-based IR. Reference [3] evaluated the effectiveness of an IR system by varying the error rates of 35 TREC queries. The results show that the use of classical IR techniques for spoken query-based IR is quite robust to considerably high levels of WER (up to about 47%). However, the experiments used artificial long spoken queries (the average length of the queries was 58 words) instead of real spoken queries; hence the results might not be representative of practical situations.

References [4], [5] suggested a new type of IR system integrated with a speech input interface. The document collection from the first pass search is utilized to adapt the ASR language model, and the lexicon is adapted with the same framework to alleviate the out-of-vocabulary (OOV) problem. Language model adaptation and lexicon adaptation were shown to decrease the word error rate.

Some researchers have worked to improve proper noun recognition. Several automated methods, such as Boltzmann machines [6], [7] and Decision Trees [8], have

Manuscript received November 30, 2009.
Manuscript revised February 26, 2010.

[†]The authors are with the Tokyo Institute of Technology, Tokyo, 152–8552 Japan.
a) E-mail: dessi@furui.cs.titech.ac.jp
b) E-mail: furui@cs.titech.ac.jp

DOI: 10.1587/transinf.E93.D.2388

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers

been proposed. References [6] and [7] show that a Boltzmann machine neural network is well-suited to the task of generating the most likely pronunciations for each proper noun based on an analysis of its spelling. However, their experiments with various architectures and different lexical domains indicated that the basic neural network fails to effectively learn the grapheme-to-phoneme distribution in the training data. Thus, they have not yet reached an acceptable error rate. In [8], a decision tree combining both heuristic and statistical methods was used to generate the most likely pronunciations for each proper noun.

The Maximum-likelihood linear regression (MLLR)-based acoustic model adaptation technique has been widely used in automatic speech recognition, especially for speaker variations [17] and accents [9]. However, the MLLR method has not been used for proper noun recognition.

A group of researchers have tried to handle the foreign word recognition problem [10]. They made acoustic models of English words in a German speech recognition system by using transcription results based on the shortest entropy distance between English and German phonemes.

3. Indonesian Language

3.1 Bahasa Indonesia

The Indonesian national language, called "Bahasa Indonesia," is a variant of the Malay language and categorized as an Austronesian or Malayo-Polynesian language. Originally, it was mainly spoken on the Riau Islands of Indonesia. It spread throughout the country and became the lingua franca after its vocabulary and idioms had become enriched by a great number of local languages. Through a variety of religious, social, and cultural influences, the Indonesian language has grown by borrowing words and terms from many languages, including Sanskrit, Arabic, Persian, Portuguese, Dutch, Chinese, and English.

Although Bahasa Indonesia is the lingua franca and the Indonesian government's official language, local languages and dialects continue to be spoken by different ethnic groups of the population. There are around 583 languages and dialects spoken in the Indonesian archipelago, and they usually belong to different ethnic groups. Some of the distinctly different local languages are Acehnese, Batak, Sundanese, Javanese, Sasak, Tetum of Timor, Dayak, Minahasa, Toraja, Buginese, Halmahera, Ambonese, Ceramese, and several Irianese languages. Such diversity is one of the sources of Indonesian pronunciation variation. The variations are even richer, since the local languages themselves may have different dialects.

The Indonesian government has defined rules about the correct pronunciation of each phoneme in the Bahasa Indonesia in correspondence with the written text. Bahasa Indonesia is written using the Latin alphabet consisting of 26 characters from A to Z, and the correspondence between sounds and their written forms is generally regular. The Indonesian standard phoneme set described by Darjowidjojo [11] is in Table A· 1. It contains 6 vowels, 3 diphthongs, and 22 consonants. In addition, there is also an informal diphthong.

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 and foreign words. Proper nouns and foreign words are major sources of pronunciation variations in the Bahasa Indonesia. The problem with these words are explained in the subsections below.

In Bahasa Indonesia, the space symbol is used to separate words and punctuation symbols; e.g., ".", ",", "!", and "?", are used to separate sentences as in English. The basic word order in sentences is Subject-Verb-Object. The adjective, demonstrative pronoun, and possessive pronoun are written following the modified noun.

3.2 Proper Noun Problems

Proper nouns in Bahasa Indonesia are more difficult to recognize than general words. Some of the reasons that we found in our Indonesian speech database are as follows:

- 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 Sect. 3.1, names in Bahasa Indonesia are influenced by hundreds of regional languages and certain foreign languages. The pronunciations often do not follow the rules of the native 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 Bahasa Indonesia. Furthermore, proper nouns grow in number through the process of human creativity in making names and the process of language assimilation. This has meant that no authoritative pronunciation lexicon has been developed for proper nouns in Bahasa Indonesia.
- Unfamiliar proper nouns may be incorrectly pronounced. Sometimes people are unaware of the correct pronunciations of other people's names, even when the names are familiar. More often, people cannot correctly read the names of unfamiliar persons. When proper nouns are less familiar, people tend to pronounce them more carefully than other words; they pronounce them less confidently, less fluently, or even too softly, hence, lengthening their duration.

3.3 Foreign 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, even when the official translated Indonesian words exist. This is mainly because the official translated Indonesian words are still not familiar to Indonesian people. This phenomenon frequently appears in news articles, technical books, and conversations. Moreover, although foreign words are supposed to be written in the italic-style in the written text, some authors do not follow this rule.

Some famous politicians, actors, teachers use English terms in public speeches. 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.

4. Information Retrieval System

4.1 Inference Network-Based IR

The Inference Network (IN) model is basically a directed acyclic graph (DAG) of a Bayesian Network [12]. The network is used to model documents and their content (the document sub-network DN) and to model queries (the query sub-network QN), as shown in Fig. 1.

The document sub-network 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 children nodes. The children 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 query sub-network 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 form the probability of the relevance to the query. The evaluation is initialized by setting the output of one document node to true (1) and all the other document nodes to false (0). This procedure is applied to each document node in turn. The probability of document relevance is taken from the final node I and is used in the ranking.

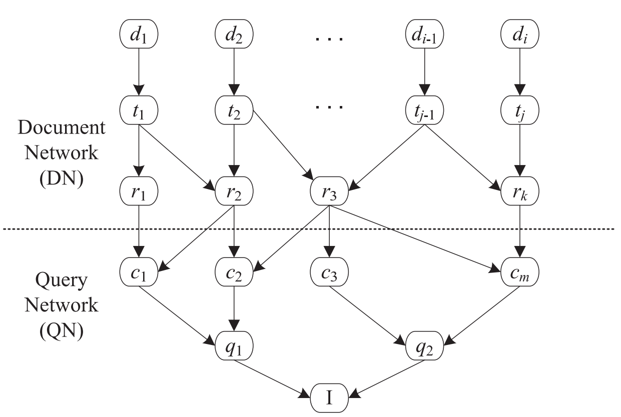


Fig. 1 Inference network.

The probability of each node except the root in the inference network needs to be calculated from its parent values. That is, if a node A has a set of parents $\pi = \{p_1, \ldots, p_n\}$, we need to estimate $P(A|p_1, \ldots, p_n)$. This usually requires a link matrix to provide diagnostic information to the set of parents based on belief with A. In practice, a canonical link matrix form is implemented. This link matrix can be used to implement a variety of weighting schemes, including familiar term weighting schemes based on the frequency of a term in a single document (tf), the inverse value of the frequency of documents including the term (idf), and their combination (tf-idf) [1].

4.2 tf-idf Weighting Schema

The tf-idf weighting method [1] is often used in IR. It is a statistical technique to evaluate how important a term is in a document. The importance increases proportionally to the number of times a word appears in the document but is offset by how common the word is in the document collection.

The belief of a representation node r_k (bel(r_k)) is computed when the concept r_k is true given a single document d_i, i.e., $bel(r_k) = P(r_k = true|d_i = true, d_{j\neq i} = false)$. The belief of node r_k can be computed using the tf-idf weight, as follows:

$$bel(r_k) = \lambda + (1 - \lambda)\overline{tf_{rk,di}}idf_{rk} \tag{1}$$

where λ is an arbitrary default belief. This ensures that every representation is allocated a non-zero belief for the observed document, even if it is not present in the document. The $\overline{tf_{rk,di}}idf_{rk}$ value can be calculated using any standard method for estimating tf-idf weights for the representation r_k.

Here, we use the Okapi tf score [13]. The Okapi tf score for a representation r_k and document d_i and idf score can be written as follows:

$$\overline{tf_{rk,di}} = \frac{tf_{rk,di}}{tf_{rk,di} + 0.5 + 1.5\frac{|d_i|}{|D_{avg}|}} \tag{2}$$

$$idf_{rk} = \frac{\log(\frac{|C|+0.5}{tf_{rk,di}})}{\log(|C| + 1)} \tag{3}$$

where $tf_{rk,di}$ is the number of times that representation r_k is matched in a document d_i, $|d_i|$ is the length of document i, $|D_{avg}|$ is the average document length in the collection, and $|C|$ is the number of words in the collection.

The retrieval status value (RSV) is evaluated by forming the dot product of the document and query representations obtained using the tf-idf weighting schema. The score for each document is calculated by summing the tf-tdf weights of all query terms found in the document.

5. Indonesian LVCSR

5.1 Baseline System

Hidden Markov Model (HMM) based acoustic models and n-gram language models were used to develop the LVCSR system for the Indonesian language [14]. The speech corpus described in Sect. 5.2 was used in the experiments. The first through 12th order Mel-Frequency Cepstral Coefficients (MFCC) were computed every 10 ms by using a 25-ms-wide window. Temporal differences of MFCC coefficients and energy were also incorporated. The LVCSR parameters were optimized using the training data described in the next section. We used 32 Gaussian mixtures per state to train context-dependent HMMs. The total number of states was 1,746, and the number of context-dependant models was 6,088.

The training text (see Sect. 5.3) was used for training the 2-gram and 3-gram language models. Both bigrams and trigrams were smoothed using the Good-Turing backoff technique. The 3-gram language model had a test-set perplexity of 61.04 and an OOV rate of 1.75% for the spoken queries described in the Sect. 5.4.

5.2 Acoustic Model Speech Corpus

An ideal Indonesian speech corpus should cover not only all phones in Bahasa Indonesia, but also those of other Indonesian dialects. However, it would have been too difficult for us to collect data on all Indonesian dialects, so we collected Bahasa Indonesia 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 [15]. Those document collections were taken from an Indonesian national newspaper and a magazine. Speech was recorded in a quiet room on DAT tapes. Then, it was digitized at a 16 kHz sampling rate. The total size of the speech corpus after manual sentence segmentation was 14.5 hours.

5.3 Language Model Text Corpus

A document collection developed by the ILPS group [15] was used for building the language model. The articles in the corpus were taken from two popular Indonesian newspaper[†] and magazine[††] sites. Some text processing were conducted on the corpus; for example, all numbers appearing in the articles were changed into words, e.g., “103” to “seratus tiga”, and all punctuation symbols, except “,”, “.”, “!” and “?” were changed into “.”. Manual corrections were also made to split long sentences into several sentences or to merge two short grammatically incorrect sentences that appeared in the document into one correct grammatical sentence. Table 1 summarizes the resulting text corpus. Half of the articles in the newspaper corpus were used to build the language model. The total number of words for building the language model was 3,125,431.

Table 1 Text corpus statistics.

Attribute	Training set
Number of Sentences	615,248
Number of words	9,853,517
Vocabulary sizes	129,919
Average sent. length (words)	16.02

5.4 Indonesian Lexicon

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

5.5 Text and Speech Corpus for IR Experiments

Since there is no standard evaluation corpus for spoken query IR in Bahasa Indonesia, we had to create the test set of spoken queries for the experiments by ourselves. The queries were derived from the Bahasa Indonesia IR collection developed by the ILPS [15] and from the Bahasa Indonesia IR collection developed by the School of Computer Science and Information Technology, RMIT University, Australia [16]. There are 35 query topics available for the magazine corpus (called “magazine A” in what follows) and 35 query topics available for the newspaper corpus in the ILPS corpus (called “newspaper corpus”). In [16], there are 20 query topics (called “magazine B”). In total, there are 90 query topics. Both IR collections are stored in the TREC format and contain documents, queries, and exhaustive relevance judgments. 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 developed three kinds of spoken queries in terms of length: short query (2–4 words), medium-length query (4–8 words), and long query (8–16 words). The aim was to analyze the effect of varying the query length on the retrieval performance. Table 2 shows examples of short, medium and long queries. We recorded

[†] http://www.kompas.com

[††] http://www.tempointeraktif.com

Table 2 Spoken query examples.

Query Type	Example (translation)
Short	perjanjian IMF Indonesia (IMF Indonesia agreement)
Medium	perjanjian kontrak antara IMF dan Indonesia (contract agreement between IMF and Indonesia)
Long	perundingan dan perjanjian kontrak antara International Monetary Fund atau IMF dengan Indonesia (discussion and contract agreement between International Monetary Fund or IMF with Indonesia)

Table 3 Number of speakers and number of queries for each source.

Source	Speakers	Short	Medium	Long	Total
Newspaper	20	35	35	35	2100
Magazine A	20	35	35	35	2100
Magazine B	20	20	20	20	1200

these queries spoken by 20 native Indonesian speakers (11 males, 9 females), each uttering 90 queries on different topics. These speakers were different from those used for training the acoustic model. There were 5400 Indonesian spoken queries in total. The spoken queries are described in Table 3.

The Indonesian newspaper text corpus provided by ILPS was divided into two parts. The first part was used to train the language model of Bahasa Indonesia 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 each magazine IR system. None of the articles was used to train the language model.

6. Experiments

6.1 Proper Noun Adaptation

The average accuracy of the baseline system was 75.1%. Figure 2 shows the error analysis of the transcribed queries. The results indicate 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. Figure 3 shows the proportion of those proper nouns in the experimental data. From those 10,720 proper nouns in the test data, words from Bahasa Indonesia and Java caused minor problems, while the rest caused major problems. Most of the foreign proper nouns from English were unfamiliar, while those from Arabic were more familiar for most of the people. The familiarity of the regional proper nouns depends on the speakers' origin.

Assuming that the difficulties of recognizing proper nouns came from acoustic variation, we tried to resolve them by enhancing the acoustic model. To model the acoustic variations of Indonesian speakers uttering proper nouns, we used an adaptation technique to create proper noun specific acoustic models. 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 in Bahasa Indonesia, as described in Sect. 3.2.

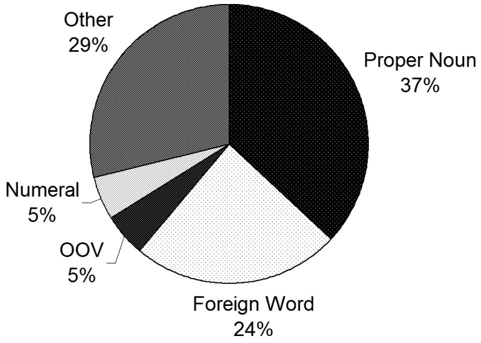


Fig. 2 Error analysis of the baseline ASR system.

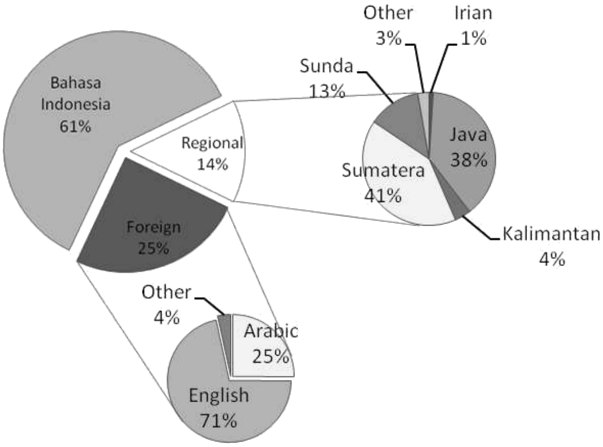


Fig. 3 Breakdown of proper nouns in the experimental data.

The procedure to build the proper noun specific models is:

- Extract the proper noun word from the speech corpus used to train the baseline acoustic model for the adaptation data (14,840 words).
- Make phone-based proper noun specific HMMs by conducting supervised adaptation based on MLLR [17] using eight regression classes to the baseline acoustic HMMs.
- 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.
- 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 base-

line lexicon, and these pronunciations were added.

6.2 English to Indonesian Phoneme Mapping (EIPM)

As shown in Fig. 2, the second largest number of misrecognized words were of foreign origin (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 pronounced closely to the original English phoneme, while 12 speakers pronounced at an English level between beginner and advanced. None of the speakers were beginners.

Several techniques to handle foreign words in general speech recognition are described as a multidimensional problem in [18]. There are several ways to map the phoneme symbols across languages: they are either knowledge-based or data-driven approaches. A data-driven approach means that acoustic models are trained using a phonetic transcription for each foreign word. The problem is that the amount of data is usually very limited. A large number of samples of the foreign language pronounced by Indonesian speakers would be required to avoid mismatches between training and testing. To our knowledge, this kind of database is not available for resource-deficient languages such as Bahasa Indonesia. The most intuitive and straightforward approach is using linguistic knowledge-based phonetic mappings based on similarities between languages [19]. One way to recognize foreign words is by mapping the phoneme symbols of the foreign words that exist in the lexicon to the phoneme symbols of the target language. In [20], an Indonesian-to-English phoneme mapping was developed to make a rapid Indonesian speech recognizer by using an English corpus to train an acoustic model. In our experiment, we used the English-to-Indonesian phoneme mapping rules found in [20], and we modified several rules (in Table 4, the "*" mark indicates the modified rule). The phonemes "er", "ey", "ii", "ng-k", and "sha" are not available in [20] but are available in the CMU phoneme set that we used in the experiment; thus, we add those phonemes to our rules.

English word pronunciations were then added to the lexicon by following the rules described in Table 4. We used the CMU (Carnegie Mellon University) lexicon as the reference for the English words. The CMU lexicon has 39 English phonemes. To re-filter the resulting English word list, we used the most standard Indonesian dictionary managed 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 that also existed in the KBBI to avoid ambiguities in the recognition. Some English words that had the same pronunciation as the Indonesian words were also deleted from the list to avoid redundancy.

Table 4 English to Indonesian phoneme mapping (EIPM).

Eng	Ind	Eng	Ind	Eng	Ind
aa	a	f	f	oy	oy
ae, eh	e	g	g	p	p
ah	e2	hh	h	r	r
ao	o	ih, iy	i	s	s
aw	aw	jh	j	sh	sy
ay	ay	k	k	t, th	t
b	b	l	l	uh, uw	u
ch	c	m	m	v	f
d, dh	d	n	n	w	w
er	e r*	ng	ng	y	y
ey	e y*	ao, ow	o	z, zh	z
n+y	ny	k+h	kh	ii	i*
ng-k	ng*	sha	sie2*		

By using the CMU lexicon as the reference, we found 4,050 words that 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.

6.3 LVCSR Evaluation

Tables 5, 6, and 7 list the average accuracies for each query length for the baseline system, proper noun-adapted (PNA) system, and English word corrected lexicon (ECL) system, for the three query corpora respectively. The English to Indonesian phoneme mapping was applied to the PNA system. For all test data, the PNA system outperformed the baseline system, and the ECL system outperformed the PNA system. The test data using the newspaper corpus gave the best results since the language model was trained using the newspaper corpus and none of the articles from the magazine corpora was used to train the language model. Table 8 summarizes the results.

Table 5 ASR accuracies of the baseline system (Base), proper-noun-adapted system (PNA), and English word corrected lexicon (ECL) for each type of query for Newspaper Corpus.

Query Type	Base	PNA	ECL
Short	79.4	82.1	84.5
Medium	81.8	84.2	84.6
Long	79.1	83.7	83.8
Average	80.1	83.3	84.3

Table 6 ASR accuracies of the baseline system (Base), proper-noun-adapted system (PNA), and English word corrected lexicon (ECL) for each type of query for Magazine Corpus A.

Query Type	Base	PNA	ECL
Short	71.3	72.9	73.9
Medium	70.6	71.2	72.9
Long	70.5	72.0	73.4
Average	70.8	72	73.3

Table 7 ASR accuracies of the baseline system (Base), the proper-noun-adapted system (PNA), and English word corrected lexicon (ECL) for each type of query for Magazine Corpus B.

Query Type	Base	PNA	ECL
Short	70.6	72.0	73.4
Medium	76.9	77.5	77.9
Long	75.8	75.9	76.4
Average	74.5	75.1	75.9

Table 8 ASR accuracies of baseline system (Base), proper-noun-adapted system (PNA), and English word corrected lexicon (ECL) for all test data.

Base	PNA	ECL
75.1	76.8	77.8

Table 9 MRR scores for spoken queries using baseline ASR (Base), proper-noun-adapted ASR (PNA), English word-corrected lexicon (ECL) ASR, and text queries. VSM: standard tf-idf vector space model, IN: IN-based tf-idf method.

	VSM	IN
Newspaper Corpus		
Base	69.8	80.8
PNA	72.3	82.2
ECL	72.70	82.7
Text query	82.4	86.7
Magazine Corpus A		
Base	61.7	72.6
PNA	63.0	73.5
ECL	65.3	75.5
Text query	84.7	85.5
Magazine Corpus B		
Base	55.1	59.8
PNA	55.4	60.4
ECL	56.8	61.4
Text query	64.7	69.2
All Data		
Base	62.2	71.1
PNA	63.6	72.0
ECL	64.9	73.2
Text query	77.3	80.5

6.4 IR Evaluation

The transcribed query was fed to the IR system after removing the stop words in Bahasa Indonesia [21]. The correct query text was also given to the IR in order to compare them with the results obtained from ASR. IN-based IR was compared with the classical VSM-based IR (Table 9). We used the mean reciprocal rank (MRR) as the evaluation measure. 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. This result shows that IN-based IR is more suitable than traditional VSM-based IR for spoken queries.

6.5 Discussion

6.5.1 Proper Noun Adaptation

Proper noun adaptation increased the recognition accuracy by 1.7% on average (Table 8). The MRR score for VSM-based IR increased 1.4% on average, whereas the score for IN-based IR increased 0.9% on average (Table 9).

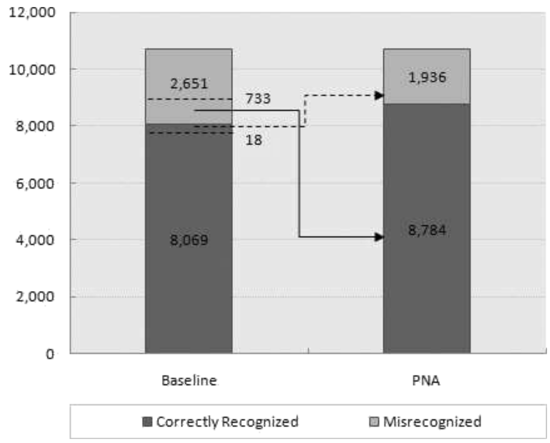


Fig. 4 Number of correctly and incorrectly recognized proper nouns by the baseline and the proper noun adapted (PNA) systems.

As shown in Fig. 4, the baseline system correctly recognized 8,069 words out of the 10,720 proper nouns in the testing data and misrecognized 2,651 proper nouns. The proper noun adapted (PNA) system recognized 8,784 proper nouns and misrecognized 1,936 proper nouns. 733 words misrecognized by the baseline system were correctly recognized by the PNA system, and 18 words correctly recognized by the baseline system were misrecognized by the PNA system.

The effectiveness of the adaptation to proper noun recognition was calculated as the ratio of the number of proper nouns that were recognized correctly by the PNA system but misrecognized by the baseline system to the number of proper nouns that were misrecognized by the baseline system: i.e., (8,784 – 8,069)/2,651 = 0.27. This means that the ability of the Indonesian LVCSR system in recognizing proper nouns increased 27% as a result of using the proposed method.

The effect of conducting proper noun adaptation to non-proper noun recognition was also analyzed. The proper-noun-adapted system misrecognized eight non-proper nouns that were recognized correctly by the baseline system, and correctly recognized two non-proper nouns that were misrecognized by the baseline system. This shows that proper noun adaptation had a slightly negative effect on non-proper noun recognition. However, the negative effect was minor since the number of non-proper nouns in the test data was quite large (32,634 words).

6.5.2 English-to-Indonesian Phoneme Mapping (EIPM)

By employing the rule-based EIPM, the recognition result increased in accuracy by 1.0% on average compared with the proper noun adapted system (Table 8). The MRR score for VSM-based IR increased 1.3% on average, whereas the score for IN-based IR increased 1.2% (Table 9). The test data contained 2,440 foreign words, most of which were English words. The proper-noun-adapted system recognized 888 foreign words and misrecognized 1,522 foreign words. By using the rule-based EIPM system, 1,678 foreign words were correctly recognized and 762 foreign words were misrecognized. The rule-based EIPM had no negative effect on non-foreign word recognition. The effectiveness of this method regarding the foreign word recognition could thus be measured as the ratio of the number of foreign words recognized correctly by the rule-based EIPM system but misrecognized by the proper noun adapted system to the number of foreign words misrecognized by the proper noun adapted system: i.e., 762/1,522 = 0.53. This means that 53% of the foreign words misrecognized by the baseline system were correctly recognized by employing the rule-based EIPM.

6.5.3 Confidence Measure

To further improve the IR performance in our experiments, we tried to incorporate confidence measures into the information retrieval process. IN-based information retrieval has an advantage that it can employ explicit term weightings to directly make a term more or less important in comparison to other terms in the query. This technique can be used to give specific information to the query [12]. We used a confidence score based on the word posterior probability to explicitly weight each transcribed term. The aim was to give additional information to the query on how certain the recognized words are in the query as correct terms and reduce irrelevant documents. However, our experimental results showed that this technique could not achieve any improvement.

7. Conclusion and Future Work

This paper presented our investigation on spoken query-based Indonesian information retrieval. Pronunciation variations in Bahasa Indonesia mainly come from three sources: dialects, proper nouns, and foreign words. This paper investigated proper noun and foreign word pronunciation variations. To increase the proper noun recognition rate in Indonesian LVCSR, we proposed a proper noun adaptation based on MLLR. The pronunciation of proper nouns in the baseline lexicon was adapted using a proper noun pronunciation model. To increase the English word recognition rate, rule-based English-to-Indonesian phoneme mapping was applied to the English words in the lexicon. It is sometimes a problem to add variants to the baseline pronunciation lexicon, since they could increase the confusability and increase the word error rate. However, in our experiments, the negative effect of adding pronunciation variations was not significant. Hence, both techniques could reduce the word error rate of the spoken queries.

We also compared the vector space model and the inference network (IN)-based IR with the tf-idf scores for Indonesian spoken queries. We found that IN-based IR outperforms the vector space model IR both for text queries and spoken queries. The improvement for spoken queries is larger than that for text queries. This shows that for the spoken queries, it is better to employ IN-based IR rather than the traditional VSM-based IR.

Regarding future work, we plan to use several query expansion methods to improve the IR performance using spoken queries. Since words in spoken queries are often misrecognized, we need to find a suitable combination of methods to minimize their effect in expanded queries.

We did not consider the OOV problem in this paper since the OOV rate of the test data was relatively low, 1.75%. However, we plan to handle the OOV problem in our future work by adapting the lexicon using new words contained in the retrieved documents.

Acknowledgments

The authors would like to thank the ILPS, University of Amsterdam, for giving us the Kompas and Majalah Tempo collections.

References

[1] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Proc. Information Processing and Management, vol.24, no.5, pp.513–523, 1988.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S.W. Kuo, "Experiments in spoken queries for documents retrieval," Proc. Eurospeech, vol.3, pp.1323–1326, 1997.

[3] F. Crestani, "Spoken query processing for interactive information retrieval," Proc. Data & Knowledge Engineering, vol.41, no.1, pp.105–124, April 2002.

[4] A. Fujii, K. Itou, and T. Ishikawa, "A method for open-vocabulary speech-driven text retrieval," Proc. 2002 Conference on Empirical Methods in Natural Language Processing, pp.188–195, 2002.

[5] A. Fujii, K. Itou, and T. Ishikawa, "Speech driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition," in Lecture Notes of Information Retrieval Techniques for Speech Application (LNCS 2273), pp.94–104, 2002.

[6] D. Neeraj, A. Le, J. Ngan, J. Hamaker, and J. Picone, "An advanced system to generate multiple pronunciations of proper nouns," Proc. IEEE ICASSP, pp.1467–1470, Munich, Germany, April I997.

[7] D. Neeraj, M. Weber, and J. Picone, "Automated generation of N-best pronunciations of proper nouns," Proc. IEEE ICASSP, pp.283–286, Atlanta, Georgia, May 1996.

[8] J. Ngan, A. Ganapathiraju, and J. Picone, "Improved surname pronunciations using decision trees," Proc. ICSLP, pp.3285–3288, Sydney, Australia, Nov. 1998.

[9] C. Huang, E. Chang, J. Zhou, and K. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition," Proc. ICSLP, Beijing, China, Oct. 2000.

[10] G. Stemmer, E. Noth, and H. Niemann, "Acoustic modeling of foreign words in a German speech recognition system," Proc. 7th Eurospeech, Scandinavia, 2001.

[11] S. Darjowidjojo, Indonesian syntax, Ph.D. dissertation, Georgetown, University, Washington, 1966.

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

[13] S.E. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," Proc. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.232–241, New York, 1994.

[14] 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.

[15] 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.

[16] 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, Dec. 2004.

[17] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Proc. Computer Speech and Language, vol.9, no.2, pp.171–185, April 1995.

[18] 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.

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

[20] 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.

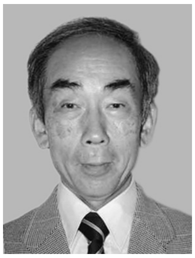
[21] F.Z. Tala, A Study of stemming effects on information retrieval in Bahasa Indonesia, M.Sc. Thesis, Appendix D, pp.39–46, University of Amsterdam, 2003.

Appendix

Table A· 1 Indonesian phoneme set.

Category	Phoneme	Word	Ph. sequence
Vowels	/a/	saya	/s *a* y *a*/
	/e/	enak	/*e* n a k/
	/E/	kEmana	/k *E* m a n a/
	/i/	ingin	/*i* n g *i* n/
	/o/	orang	/*o* r a ng/
	/u/	untuk	/*u* n t *u* k/
Diphthongs	/ai/	sungai	/s u ng *ai*/
	/au/	danau	/d a n *au*/
	/ei/	amboi	/a m b *oi*/
Informal Diphthongs	/ei/	hei	/h *ei*/
Semi-vowels	/w/	wanita	/*w* a n i t a/
	/y/	saya	/s a *y* a/
Cons. Plosives	/b/	berapa	/*b* e r a p a/
	/p/	petani	/*p* e t a n i/
	/d/	dia	/*d* i a/
	/t/	teman	/*t* e m a n/
	/g/	giat	/*g* i a t/
	/k/	kamu	/*k* a m u/
	/kh/	khairul	/*kh* a i r u l/
Cons. Africates	/j/	juga	/*j* u g a/
	/c/	cinta	/*c* i n t a/
Cons. Fricatives	/f/	maaf video	/m a a *f*/ /*f* i d e o/
	/z/	jenazah	/j e n a *z* a h/
	/s/	saya	/*s* a y a/
	/sy/	syahdu	/*sy* a h d u/
	/h/	hujan	/*h* u j a n/
Cons. Liquids	/r/	ramai	/*r* a m a i/
	/l/	lambat	/*l* a m b a t/
Cons. Nasals	/m/	mana	/*m* a n a/
	/n/	mana	/m a *n* a/
	/ny/	nyanyian	/*ny* a *ny* i an/
	/ng/	lambang	/l a m b a *ng*/

Dessi Puji Lestari received the B.E. degree in Informatics Engineering from Bandung Institute of Technology, Bandung, Indonesia in 2002. She received the M.E. degree in Computer Science from the Tokyo Institute of Technology, Tokyo, Japan, in 2007. She is currently pursuing a Ph.D. degree at the Tokyo Institute of Technology.



Sadaoki Furui is currently a Professor at the Tokyo Institute of Technology, Department of Computer Science. 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 800 published articles. He is a Fellow of the IEEE, the International Speech Communication Association (ISCA), and the Acoustical Society of America. He has served as President of the Acoustical Society of Japan (ASJ) and the ISCA. He has served as a member of the Board of Governors of the IEEE Signal Processing (SP) Society and Editor-in-Chief of both the Transactions of the IEICE and the Journal of Speech Communication. He has received the Yonezawa Prize, the Paper Award and the Achievement Award from the IEICE (1975, 88, 93, 2003, 2003, 2008), and the Sato Paper Award from the ASJ (1985, 87). He has received the Senior Award and Society Award from the IEEE SP Society (1989, 2006), the Achievement Award from the Minister of Science and Technology and the Minister of Education, Japan (1989, 2006), the Purple Ribbon Medal from the Japanese Emperor (2006), and the ISCA Medal for Scientific Achievement (2009). In 1993 he served as an IEEE SPS Distinguished Lecturer.