Maximum Likelihood-based Automatic Lexicon Generation for AI Assistant-based Interaction with Mobile Devices

Donghyun Lee[1], Jae-Hyun Park[2], Kwang-Ho Kim[3], Jeong-Sik Park[4], Ji-Hwan Kim[1], Gil-Jin Jang[5], and Unsang Park[1]

[1] Department of Computer Science and Engineering, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul, 04107, Republic of Korea
[e-mail: {redizard, kimjihwan, unsangpark}@sogang.ac.kr]
[2] LG Electronics Institute of Technology
38, Baumoe-ro, Seocho-gu, Seoul, 06763, Republic of Korea
[e-mail: jaehyunn.park@lge.com]
[3] AIZEN Global Co., Inc.
200, Yeongdeungpo-ro, Yeongdeungpo-gu, Seoul, 07301, Republic of Korea
[e-mail: kwangho.kim@aizen.co]
[4] Department of Information and Communication Engineering, Yeungnam University
280 Daehak-Ro, Gyeongsan, Gyeongbuk, 38541, Republic of Korea
[e-mail: parkjs@yu.ac.kr]
[5] School of Electronics Engineering, Kyungpook National University
80 Daehakro, Bukgu, Daegu, 41566, Republic of Korea
[e-mail: gjang@knu.ac.kr]
*Corresponding author: Unsang Park

Received November 21, 2016; revised March 14, 2017; accepted March 29, 2017; published September 30, 2017

Abstract

In this paper, maximum likelihood-based automatic lexicon generation using mixed-syllables is proposed for unlimited vocabulary voice interface for East Asian languages (e.g. Korean, Chinese and Japanese) in AI-assistant based interaction with mobile devices. The conventional lexicon has two inevitable problems: 1) a tedious repetition of out-of-lexicon unit additions to the lexicon, and 2) the propagation of errors during a morpheme analysis and space segmentation.

The proposed method provides an automatic framework to solve the above problems. The proposed method produces a level of overall accuracy similar to one of previous methods in the presence of one out-of-lexicon word in a sentence, but the proposed method provides superior results with the absolute improvements of 1.62%, 5.58%, and 10.09% in terms of word accuracy when the number of out-of-lexicon words in a sentence was two, three and four, respectively.

Keywords: Maximum likelihood, automatic lexicon generation, intelligent personal assistant (IPA), out-of-lexicon (OOL), speech recognition

This material is based upon work supported by the Ministry of Trade, Industry & Energy(MOTIE, Korea) under Industrial Technology Innovation Program (No.10063424, 'development of distant speech recognition and multi-task dialog processing technologies for in-door conversational robots').

https://doi.org/10.3837/tiis.2017.09.005 **ISSN : 1976-7277**

1. Introduction

Intelligent Personal Assistants (IPAs) provide a new way for people to interact with mobile devices [1-2]. This type of interface has attracted worldwide interest, largely due to its improved unlimited vocabulary voice recognition [3]. This improvement is mainly attributed to an automatic re-training of the recognizer using accumulated number of spoken queries.

The goal of a voice interface is producing a word series from a human speech signal. The system computes a series of the most probable words, denoted as W from Y, which is an acoustic vector series. Because the number of possible Y is infinite, the following Bayes' theorem is applied as in Equation (1):

$$\widehat{W} = \operatorname{argmax}_W P(W|Y) = \operatorname{argmax}_W \frac{P(W)P(Y|W)}{P(Y)} \propto \operatorname{argmax}_W P(W)P(Y|W) \quad (1)$$

In Equation (1), $P(Y)$ is the probability of acoustic vector series Y, which is independent from W, hence, Y can be ignored when searching for the word series with the maximum probability. In the end, the voice recognition suggests a word series that yields the highest product of $P(W)$ and $P(Y|W)$. $P(W)$, the probability for W, is calculated from a Language Model (LM). $P(Y|W)$ is to be provided using an Acoustic Model (AM) for W. Before modeling AMs and LMs of a voice recognition, it is necessary to select a group of words depending on a domain of corpus. This set of words is called the lexicon of the voice interface. The decoder then searches for the word sequence with the highest probability.

Many unsupervised learning algorithms have been developed for the re-training of acoustic and language models [4-5]. In addition, the cloud computing community has reported many successful implementations of a large-scale commercial decoder [6]. However, previous studies have typically been based on fixed lexicons determined before the service launch or based on on-the-fly updates according to a new analysis of out-of-lexicon words. Large numbers of spoken queries are accumulated in an IPA. Because these queries reflect a person's interest at the time of communication, they frequently have out-of-lexicon words. As a result, out-of-lexicon words are perpetually observed.

In European languages, the best lexicon guarantees the lowest out-of-lexicon rate. This lexicon is composed based on a group of full-words that are frequently used and delimited by spaces [7]. However, lexicons for East Asian languages such as Korean, Chinese, and Japanese are often composed of frequently used morphemes. Morphemes have been used as the smallest linguistic unit of a word in previous works [8-9]. This choice is due to three characteristics of East Asian languages: 1) Full-words are typically formed through a linear concatenation of grammatical morphemes and lexical morphemes. **Table 1** shows examples of full-words and their corresponding morphological analysis results regarding the verb 'write' in Korean. In this table, a full-word is denoted as a sequence of its comprising syllables and dashes concatenating the syllables. The result of a morphological analysis is denoted as a sequence of composing morphemes and spaces separating the morphemes. If a morpheme has more than one syllable, its corresponding syllables, consonants, and vowels are concatenated by dashes. The root of [*sseu*] (meaning 'write' in Korean) consists of two different graphemes of [*sseo*] and [*sseu*] depending on the following ending graphemes. The morphemes [*b-ni-da*], [*si*], and [*ss*] are endings to denote normal ending, honorific style and past tense, respectively.

2) Space marks only partially exist or are generally absent. Spacing is inconsistent, particularly in Korean. For example, the past present verb 'have explained' can be denoted as [*seol-myeong-hae-wat-da*], [*seol-myeong-hae wat-da*], or [*seol-myeong hae-wat-da*]. 3) Substantial morphemes have only one syllable (e.g. [*sseu*], [*da*], [*si*], and [*sseo*] in **Table 1**) or even just one consonant or vowel (e.g. [*ss*] in **Table 1**).

Table 1. Examples of full-words and their corresponding morphological analysis results in Korean

Full-word	Morphological analysis result	Meaning
[*sseu-da*]	[*sseu da*]	write (base form)
[*sseub-ni-da*]	[*sseu b-ni-da*]	write (normal ending)
[*sseu-sib-ni-da*]	[*sseu si b-ni-da*]	write (honorific style)
[*ssoss-seum-ni-da*]	[*sseo ss seum-ni-da*]	wrote (past tense)

In many cases, dialogue with IPAs requires domain specific knowledge, reflecting the user's interest at the time of communication. If the number of words existing in the lexicon is defined, the number of words in the lexicon increases to enhance the coverage rate. However, whenever the number of words increases, the decoding process and the language model require a large memory capacity. There is a limitation in increasing the number of lexicon words for full word-based and morpheme-based lexicons.

For these reasons, it becomes tedious to manage both full word-based and morpheme-based lexicons. In addition, morpheme-based lexicons used in voice recognition have problems of errors propagated through a morpheme analysis and a space segmentation.

In this paper, a Maximum Likelihood (ML)-based automatically generated lexicon using mixed-syllables is proposed to improve the voice recognition for East Asian languages. The proposed method uses the ML algorithm. The ML algorithm does not require any prior knowledge of the target language, such as space segmentation or the results of a morphological analysis. This method is completely automatic. It begins with an initial lexicon consisting of approximately 3,900 syllables in Korean as defined by EUC-KR. The lexicon is expanded through the addition of pairs generated from lexicon words. A pair of lexicon words allows for the maximum increase in likelihood generated from the LM for a current corpus.

As a result, the proposed method solves the two previously stated two problems. In addition, the proposed method produces a level of overall accuracy similar to one of previous methods in the presence of one out-of-lexicon word in a sentence, but the proposed method provides superior results when a sentence has two or more out-of-lexicon words.

This paper is an extension of preliminary works [10-11]. Through a comparison with previous works, the details of the proposed method are further explained for an analysis of an ML-based automatically generated lexicon algorithm, and an implementation of the proposed method is described.

The rest of this paper is organized as follows. Section 2 reviews previous works on lexicons in continuous Korean speech recognition. Section 3 explains our implementation of the automatic generation of ML-based mixed-syllable lexicon units for an unlimited vocabulary voice interface for Korean. Section 4 describes our experimental setups and results, and Section 5 concludes this paper.

2. Previous Works

The Korean language is an agglutinative language, and has a large number of inflected forms. In particular, verbs are inflected heavily depending on their syntactic role. If full-words are used as a lexicon unit, the Korean speech recognition system will have a large number of lexicon words. Therefore, the system generates recognition results with a high out-of-lexicon rate if the size of the lexicon is limited (e.g. fewer than approximately 10,000 words) [12]. This becomes apparent in IPA, since the amount of speech and text queries collected in a single day are large order of magnitude. For this reason, at least 100,000 lexicon words are required when a full word-based lexicon is deployed in an IPA.

A morpheme-based lexicon has often been used to reduce the high out-of-lexicon rate of full-word based lexicons [13-14]. However, this approach does not correctly show co-articulation in short morphemes (e.g. suffixes and word-endings). In addition, this approach suffers from a short lexical coverage of LMs compared to those of longer lexicon units such as full-words when the same size of n-gram unit is applied.

Pseudo-morphemes have also been proposed as an extension of a morpheme to solve the above two problems of morpheme-based lexicons [15-17]. The general approach to pseudo-morpheme generation is to concatenate frequent and short morphemes. These merged morpheme pairs and the other morphemes compose a lexicon of pseudo-morphemes [18-19].

Statistical information is applied to morpheme combination with additionally imposed constraints [17]. This research compared a definition of two pseudo-morphemes, one of which concatenates these morphemes based on knowledge, and the other based on statistics.

According to knowledge-based definitions, morphemes are combined while maintaining part-of-speech (POS) tags assigned to them. Therefore, syntactic morphemes and semantic morphemes remain distinctive after concatenation. For an auxiliary predicative, multiple morphemes are combined to generate another new auxiliary predicative. This method generates recognition results with POS tags, but requires accurate morpheme identification and linguistic expertise to produce correct results. This work considers a semantic combination only in that the number of morpheme pairs becomes excessively large, which is controlled through a concatenation method with frequency.

Statistics-based morpheme combinations attempts to reduce the LM complexity by concatenating selected morpheme pairs automatically, although the size of the lexicon might increase. A statistics-based morpheme concatenation method used three criteria: mutual information, morpheme-pair frequency, and unigram log likelihood. Morpheme pairs with high points with regard to these three criteria were concatenated repeatedly using the given training corpus until the complexity no longer decreases in the given training corpus. Only frequently occurred morphemes were considered as candidates for combination, and all possible morpheme concatenations were paired while excluding pairs that cause unnecessary conflicts for an improvement in speed. At least one of candidates has to be a syllable, and resultant lengths of the combined morphemes were limited based on threshold values in the frequency-based morpheme combination method, which effectively reduces the complexity as much as possible. The lowest complexity is attained when the morpheme combination was based on the frequency of morpheme pairs with a syllable constraint, whereas the highest recognition rate of the voice recognition system was achieved when the morpheme combination was also based on the frequency of morpheme pairs while morpheme lengths were limited. This work analyzed different LMs and their successful voice recognition performance in accordance with different definitions of pseudo-morphemes.

To develop an unlimited lexicon voice search, a greedy algorithm was proposed to learn the lexicon from large amounts of text corpus [20]. This algorithm generated a user-specified number of lexicon units chosen by the greedy method without focusing on semantics. As a result, any sentences can be generated as a series of units in this lexicon. However, the details of this lexicon generation method were not explained in detail and a comparison with other lexicons was not experimentally performed.

A sub-word-based automatic lexicon generation method using the greedy algorithm was proposed to develop an unlimited Korean lexicon [10-11]. This method generated an initial lexicon based on monosyllabic Korean characters, and sub-word pairs that maximize likelihood on the training corpus were then attached to the initial lexicon. The number of sub-word pairs was defined based on a user-specified threshold. However, a standard for the likelihood and a method for determining its likelihood were not described in detail and experiments were not performed. This method also required a high time-complexity because the LM was built using bi-gram counts and tri-gram counts for computing the likelihood whenever a sub-word pair was generated [10].

A compound-finding algorithm, previously used to find sequences of lexicon units in English, was used in an automatic process for the training lexicon units in Japanese [21]. A hybrid approach was proposed based on a concatenation of morpheme units and full-word units [22]. A one-pass algorithm and a lexicon generation method using cross-word phone variation were proposed [23]. It was reported that the cross-word phone variation lexicon was useful for morpheme-based lexicon.

3. Automatic Generation of ML-based Mixed-Syllable Lexicon

This chapter describes our proposed automatic generation algorithm using the ML-based lexicon with mixed-syllables. For East Asian languages, a bunch of syllables can be segmented through a variety of methods, and each segmented result has a different meaning. To produce spaces between words correctly, the same sequences of syllables must be distinguished depending on whether syllable sequences have spaces on their right or on their left. Thus, the lexicon generation method requires pre-processing for a space mark between words in the training corpus, and post-processing for voice recognition results.

3.1 Pre-processing for a Space between Words in Training Corpus

The same sequences of syllables in Korean language may have a different meaning depending on spacing between syllables. For example, a sentence [*a-beo-ji ga-bang-e deul-eo-ga-sin-da*] means “Father is going into the bag,” whereas the same syllable sentence [*a-beo-ji-ga bang-e deul-eo-ga-sin-da*] means “Father is going into the room.”

To learn where to put spaces, it is necessary to replace every space in the training corpus with a space marker. In this paper, an underscore is considered as a space marker. In addition, every lexicon unit is overlapped into four forms: an original lexicon unit, a unit with an underscore on its right, a unit with an underscore on its left, and a unit with underscores on both sides in the initial lexicon generation.

For example, in the training corpus, a sentence [*a-beo-ji ga-bang-e deul-eo-ga-sin-da*] is transformed into [*a-beo-ji_ga-bang-e_deul-eo-ga-sin-da*]. The lexicon unit [*ga-bang*] is overlapped into [*ga-bang*], [*ga-bang_*], [*_ga-bang*], and [*_ga-bang_*].

3.2 ML-Based Automatic Mixed-Syllable Lexicon Generation Algorithm

Suppose there are n sentences (s_1, s_2, $\cdots$, s_n) in the training corpus, coming from a distribution with an unknown probability density function of the language model, f_0. However, suppose that f_0 belongs to a specific family of distributions, called the parametric model, $\{f(*|\theta), \theta \in \alpha\}$, such that $f_0 = f(*|\theta_0)$. In our research, α is the set of all lexicon word combinations based on the current lexicon and θ is a combination in α. θ_0 is an unknown mixed-syllable lexicon word denoted as the true value of the parameter. It is thus necessary to find a mixed-syllable lexicon word $\hat{\theta}$ which is as close to θ_0 as possible. Therefore, $\hat{\theta}$ is added to the lexicon when our algorithm finds $\hat{\theta}$.

The joint density function is defined for all sentences in the training corpus as in Equation (2):

$$f(s_1, s_2, \cdots, s_n|\theta) = f(s_1|\theta) \times f(s_2|\theta) \times \cdots \times f(s_n|\theta) \tag{2}$$

In our algorithm, s_1, s_2, $\cdots$, s_n are treated as fixed parameters, but θ is treated as a variable parameter. The likelihood L is defined as in Equation (3):

$$L(\theta|s_1, \cdots, s_n) = f(s_1, s_2, \cdots, s_n|\theta) = \prod_{i=1}^{n} f(s_i|\theta) \tag{3}$$

In practice, it is convenient to work with a logarithm expression of L, which is called the log-likelihood. The log-likelihood is calculated as in Equation (4):

$$\ln L(\theta|s_1, \cdots, s_n) = \sum_{i=1}^{n} \ln f(s_i|\theta) \tag{4}$$

The proposed maximum likelihood estimation attempts to find $\hat{\theta}$ which maximizes the log-likelihood, as shown in Equation (5):

$$\hat{\theta} = \text{argmax}_{\theta \in \alpha} \sum_{i=1}^{n} \ln f(s_i|\theta) \tag{5}$$

Here, $\hat{\theta}$ for the current lexicon is added to the lexicon. As a result, α and f are updated. By using the updated α and f, the same procedures are repeated until the likelihood increases.

In this paper, the likelihood of s_i is estimated using a tri-gram language model probability of the sentence. If s_j is segmented as a sequence of J lexicon words, lw_1, ..., lw_J, the likelihood of s_i is estimated as in Equation (6):

$$\begin{aligned} f(s_i|\theta) = p(lw_1, lw_2, \ldots, lw_J) &= \prod_{j=1}^{J} p(lw_j|lw_1, \ldots, lw_{j-1}) \\ &\propto \prod_{j=1}^{J} p(lw_j|lw_{j-2}, lw_{j-1}) \end{aligned} \tag{6}$$

Here, $c(lw_j, lw_{j-1}, lw_{j-2})$ is the tri-gram count of three consecutive lexicon words in the training corpus, and $c(lw_{j-1}, lw_{j-2})$ is the bi-gram count of two consecutive lexicon words.

The pseudo algorithm of the proposed automatic lexicon generation based on ML is illustrated in **Fig. 1**. The detailed explanation of **Fig. 1** is as follows:

1. A initial lexicon is composed of Korean syllables that are defined in EUC-KR (the number of syllables is approximately 3,900). Syllables are overlapped into four cases as described in Section 3.1. The initial segmentation is operated in the training corpus at the syllabic level. A segmented syllable is substituted with its corresponding duplicated syllables

if syllables are attached to a space marker on its right or left. All spaces are eliminated from the training corpus. Based on the initial segmentation results in the training corpus, all counts of the initial bi-gram $c(lw_{j-1}, lw_{j-2})$ and initial tri-gram $c(lw_j, lw_{j-1}, lw_{j-2})$ are calculated.

2. Build a LM based on current bi-gram and tri-gram counts for the current lexicon.

3. Generate all lexicon unit pairs for the current lexicon. Choose the pair that maximizes likelihood L, when the pair is contained as a lexicon unit and the likelihood is calculated based on the updated lexicon.

4. Add the best pair to the current lexicon. In current segmentation results, consecutive lexicon units corresponding to the best pair are substituted with the best combination. All bi-gram and tri-gram counts are updated using the updated segmentation results.

5. Go to Step 2 until the number of lexicon units is less than the pre-defined limitation and the likelihood increases.

A lexicon with a pre-defined size limitation, and a LM based on this lexicon can be successfully produced using the proposed algorithm. A voice recognizer for East Asian languages can generate any sentences by using the LM and the lexicon generated through the proposed method. Because the proposed method is entirely automatic and does not require any prior knowledges of the target language (e.g. morphological analysis, space segmentation), two major problems of previous methods are solved using the proposed method: 1) the tedious repetition of out-of-lexicon unit additions to the lexicon and 2) error propagation from the space segmentation and morpheme analysis in the voice interface.

Because a lexicon based on the proposed method includes space markers, the voice recognizer based on this lexicon also generates space markers. These space markers reduce the readability of voice recognition transcriptions. As such, it is also necessary to post-process voice recognition results.

3.3 Post-Processing for Voice Recognition Result

A space marker in voice recognition results is produced with a lexicon unit attached to the space marker on its right or left side. In the case of successive double space markers, a first space marker is produced with the left lexicon unit, and the second space marker is produced with the right lexicon unit. In this case, successive double space markers are converted into a single space, whereas a single space marker is discarded. For example, a voice recognition result of [*_na-neun_ _chin-gu ga_ _jeok-da_*] is changed as a recognition result of [*na-neun chin-gu-ga jeok-da*].

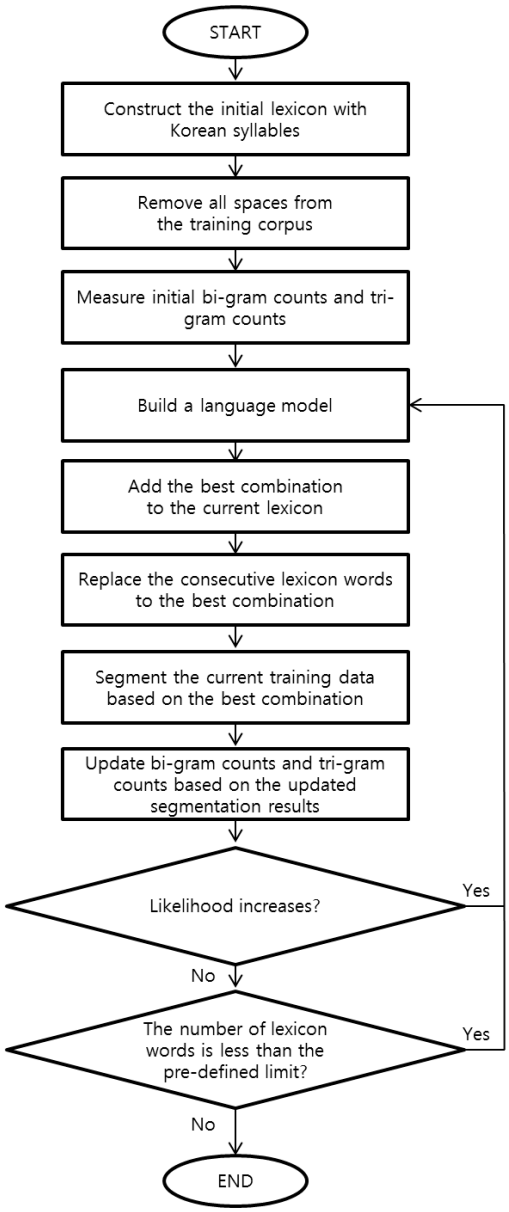


Fig. 1. Pseudo algorithm for the proposed ML-based automatic lexicon generation

4. Experiments and Results

For acoustic model training data, approximately 370K utterances were used, which were recorded at 16-bit resolution with a sampling rate of 16 KHz in a single channel. The acoustic model for this experiment uses 38 phonemes. Using mono-phones as basic recognition units requires the modeling of 38 phonemes in Hidden Markov Model (HMM) [24-25]. However, a phoneme is pronounced in a variety of ways depending on the preceding and following

phonemes [26]. Hence, a context-dependent phoneme model was chosen to accurately identify phonemes. **Table 2** shows the phoneme set used in this paper. This set is also used in [27-30].

The voice interface performance is measured with recognition accuracy as in Equation (8), where *N*, *H*, and *I* denote the number of words in the text, correctly recognized 'hit' words, and incorrectly inserted words.

$$Accuracy = \frac{H-I}{N} \times 100\% \tag{8}$$

The number of lexicon words has to be determined before constructing a language model and developing voice interface. The size of the lexicon for our experiment was determined to be 200K because this size allows approximately 99% of coverage for the language model training data for the initial morphemes. The same sized lexicon was used throughout the experiment, and the coverage was shown to be 88% for a full-word based lexicon, 95% for lexicon based on pseudo-morpheme, and 100% for an automatically generated ML-based lexicon.

Section 4.1 describes the generation of three lexicons: full-word, pseudo-morpheme, and the proposed ML-based mixed-syllable. Section 4.2 shows experimental results.

Table 2. Phoneme set (no. of phoneme: 38)

Symbol	Pronunciation	Symbol	Pronunciation
AA	[a]	OI	[wae]
B	[b]	OO	[o]
BB	[pp]	P	[p]
C	[c]	R	[r] (first consonant)
D	[d]	S	[s]
DD	[tt]	SS	[ss]
EE	[ee]	T	[t]
G	[g]	UI	[wi]
GG	[kk]	UU	[u]
H	[h]	UV	[wo]
II	[ya]	VV	[eo]
J	[j]	XI	[ui]
JJ	[jj]	XX	[eu]
K	[k]	YA	[ya]
L	[r] (final consonant)	YE	[ye]
M	[m]	YO	[yo]
N	[i]	YU	[yu]
NG	[o] (final consonant)	YV	[yeo]
OA	[wa]	SIL	(silence)

4.1 Lexicon Generation

Two additional lexicon units were developed for a comparative analysis: one lexicon unit was based on full-words as frequently used in English, whereas the other unit was based on pseudo-morphemes. **Table 3** lists our proposed lexicon and the above two lexicon units.

Table 3. Lexicon unit

Lexicon units
Full-word
Pseudo-morpheme
ML-based mixed-syllable (the proposed method)

The full-word based lexicon units adopted the space-delimited units from the training data without any modification. Therefore, the full-word based lexicon is constructed using a set of frequently used full-words delimited by spaces.

The generation of the lexicon based on pseudo-morphemes starts from a morpheme analysis, which disassembles a full-word into morphemes. In this paper, morphemes are generated using UTagger, which won first prize at the Korean Information Processing System Competition. Each morpheme in the segmentation result of UTagger was attached with its corresponding POS tag (http://nlplab.ulsan.ac.kr/). However, the morpheme analysis result suffers from two problems: 1) the pronunciations of some of morphemes do not match that of corresponding syllables and 2) some of morphemes do not show a correct co-articulation in short morphemes (e.g. suffixes and word-endings).

The lexicon based on pseudo-morphemes is extended from a morpheme-based lexicon to include the combination of syllables whose pronunciations do not match those of their corresponding morphemes. **Table 4** shows two examples of full-words. In the first example [*do-wat-da*], pronunciations of two morphemes [*dop*] and [*at*] do not match with those corresponding syllables [*do-what*]. These two morphemes are combined to form a pseudo-morpheme [*do-wat*]. The merged pairs of short and frequent morphemes are included in the lexicon of pseudo-morphemes. Regarding the merging, a rule-based method and a statistical method were described in [4]. In this research, the rule-based method is used.

The proposed ML-based mixed-syllable lexicon is generated according to the descriptions in Section 3. For experiments, the number of mixed-syllable pairs was used as the standard for likelihood. The mixed-syllable pair was added to the initial lexicon if the number of pairs, which was counted from the training corpus, reached up to 1,000. The time-complexity of this method for implementation is less than the time-complexity of ML-based mixed-syllable lexicon algorithm because the LM based on bi-gram and tri-gram counts is not generated whenever the mixed-syllable pair is generated.

Table 4. Examples of generated pseudo-morphemes when the incorrectly pronounced morphemes are substituted with their corresponding syllables

Full-word	Morpheme analysis result	Generated pseudo-morpheme
[*do-wat-da*]	[*dop*]/VV + [*at*]/EP + [*da*]/EF	[*do-wat*] + [*da*]
[*mot-haet-seup-ni-da*]	[*mot-ha*]/VX + [*eot*]/EP + [*seup-ni-da*]/EF	[*mot-haet*] + [*seup-ni-da*]

4.2 Experimental Result

The experimental results were collected from lexicons based on different criteria: full-words, pseudo-morphemes, and ML-based mixed-syllables. Acoustic models were re-trained for each lexicon. The training corpus consists of 7,457 sentences that have between one and four out-of-lexicon words. An out-of-lexicon word is defined as a word not listed in the full word-based lexicon in this experiment. **Table 5** shows the number of sentences according to the number of out-of-lexicon words per sentence in the training corpus. The total number of sentences that have one out-of-lexicon word is counted as 3,746 in the training corpus. The total number of sentences which have two or more out-of-lexicon words is counted as 3,711 in the training corpus. Since the ratio in the range 2-4 is substantial (about 50%), our proposed method can improve the overall performance for the related applications.

Table 5. No. of sentences according to the number of out-of-lexicon words per sentence (total no. of sentences: 7,457)

No. of out-of-lexicon words per sentence	No. of sentences
1	3,746
2	2,474
3	982
4	255

The 3,746 sentences that have only one out-of-lexicon words per sentence were tested first. For sentences with one out-of-lexicon word, the pseudo-morpheme lexicon units showed the highest accuracy, followed by the proposed ML-based mixed-syllable unit and the full-word units. It is obvious that full-word units are very weak against the out-of-lexicon word. **Table 6** shows the recognition accuracies of lexicon units for sentences with one out-of-lexicon word per sentence.

Table 6. Recognition accuracy for lexicon unit (no. of out-of-lexicon words per sentence: 1, no. of test sentences: 3,746)

Lexicon unit	Accuracy (%)
Full-word	69.53
Pseudo-morpheme	75.05
ML-based mixed-syllable	72.52

Next, the 2,474 sentences with two out-of-lexicon words per sentence were tested. For these sentences, the proposed ML-based mixed-syllable unit showed the highest recognition accuracy, followed by the pseudo-morpheme unit and the full-word unit. The full-word unit also suffered drastic performance degradation with these two out-of-lexicon words sentences, similar to that seen with the one out-of-lexicon word sentences. As the number of out-of-lexicon words per sentence increases from 1 to 2, the decreases in recognition accuracy were measured as 17.5%, 12.57% and 8.42% for lexicons based on full word, pseudo-morpheme, and ML-based mixed-syllable, respectively. **Table 7** shows the recognition accuracies of lexicon units for sentences with two out-of-lexicon words per sentence.

Table 7. Recognition accuracy for lexicon unit
(no. of out-of-lexicon words per sentence: 2, no. of test sentences: 2,474)

Lexicon unit	Accuracy (%)
Full-word	52.03
Pseudo-morpheme	62.48
ML-based mixed-syllable	64.10

The 982 sentences with three out-of-lexicon words per sentence were tested. For these sentences, the proposed ML-based mixed-syllable unit again showed the highest recognition accuracy, followed by the pseudo-morpheme unit and the full-word unit. The proposed ML-based mixed-syllable unit showed a significantly better performance than the pseudo-morpheme unit. As the number of out-of-lexicon words per sentence increases from 2 to 3, the decreases in recognition accuracy were measured as 13.9%, 9.94% and 5.98% for lexicons based on full word, pseudo-morpheme, and ML-based mixed-syllable, respectively. **Table 8** shows the recognition accuracies of lexicon units for sentences with three out-of-lexicon words per sentence.

Table 8. Recognition accuracy for lexicon unit
(no. of out-of-lexicon words per sentence: 3, no. of test sentences: 982)

Lexicon unit	Accuracy (%)
Full-word	38.13
Pseudo-morpheme	52.54
ML-based mixed-syllable	58.12

Finally, the 255 sentences with four out-of-lexicon words per sentence were tested. For these sentences, the proposed ML-based mixed-syllable unit once more showed the highest recognition accuracy, followed by the pseudo-morpheme unit and the full-word unit. As the number of out-of-lexicon words per sentence increases from 3 to 4, the decreases in recognition accuracy were measured as 14.68%, 9.64% and 5.13% for lexicons based on full word, pseudo-morpheme, and ML-based mixed-syllable, respectively. **Table 9** shows the recognition accuracies of lexicon units for sentences with four out-of-lexicon words per sentence.

Table 9. Recognition accuracy for lexicon unit
(no. of out-of-lexicon words per sentence: 4, no. of test sentences: 982)

Lexicon unit	Accuracy (%)
Full-word	38.13
Pseudo-morpheme	52.54
ML-based mixed-syllable	58.12

Fig. 2 summarizes the recognition accuracies of the lexicon units according to the number of out-of-lexicon words per sentence. The proposed lexicon generation method using the ML-based mixed-syllable unit provides superior results to conventional units in the presence of two or more out-of-lexicon words in a sentence. The proposed method produces a level of overall accuracy similar to one of conventional methods in the presence of one out-of-lexicon word in a sentence. Therefore, it is concluded that the proposed lexicon unit is robust in the presence of out-of-lexicon words which are frequently observed in dialogue with IPAs.

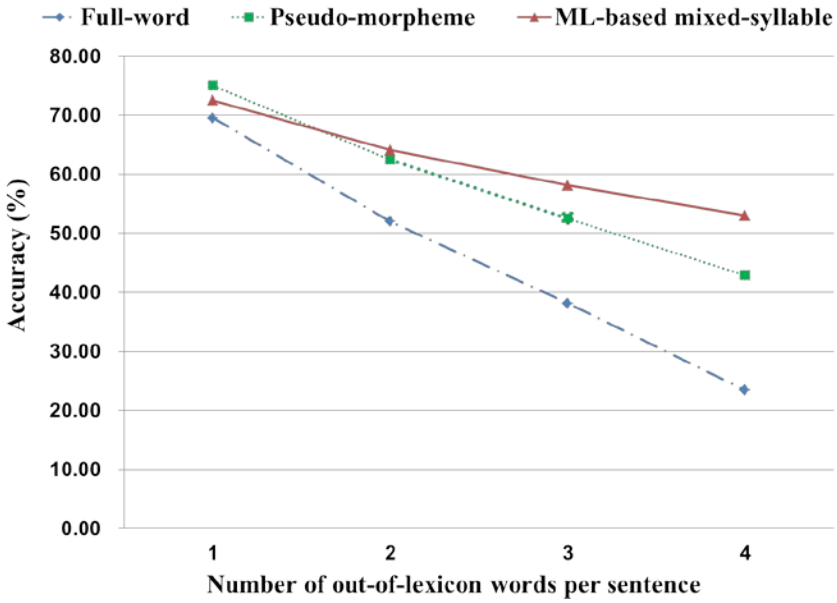


Fig. 2. Recognition accuracy of lexicon unit according to number of out-of-lexicon words per sentence.

5. Conclusions

IPA has attracted interest worldwide, largely due to the improvement of voice interface components such as acoustic and language models. Such improvement is mainly contributed to automatic re-training of components using many unsupervised learning algorithms and the accumulated number of spoken queries. However, previous works on lexicon were typically based on fixed lexicons determined before the service launch or manual updates according to the analysis of out-of-lexicon words.

This paper proposes a maximum likelihood-based automatic lexicon generation for an unlimited vocabulary voice interface for East Asian languages in an IPA. The conventional lexicon for these languages have two inevitable problems: 1) A tedious repetition of out-of-lexicon unit additions to the lexicon. The order of magnitude of spoken and text queries are accumulated in IPA. Because such queries show the user's interest at the time of communication, they frequently have out-of-lexicon units. 2) Errors are propagated from space segmentation and the morpheme analysis. The proposed method provides an automatic framework to solve these problems, which does not require any prior knowledge of target languages (e.g. space segmentation or morphological analysis). The proposed method produces a level of overall accuracy similar to one of previous methods in the presence of one out-of-lexicon word in a sentence, but the proposed method provides superior results when a sentence has two or more out-of-lexicon words.

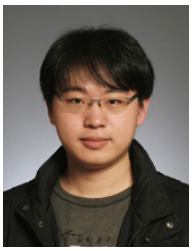
References

[1] T. Bosse, R. Duell, M. Hoogendoorn et al., "A multi-agent system architecture for personal support during demanding tasks," *Opportunities and Challenges for Next-Generation Applied Intelligence*, vol. 214, pp. 285-290, 2009. Article (CrossRef Link)

[2] J. Jiang, A. H. Awadallah, R. Jones et al., "Automatic online evaluation of intelligent assistants," in *Proc. of the 24th International Conference on World Wide Web*, pp. 506-516, May 18-22, 2015. Article (CrossRef Link)
[3] H. Chung, J. Park, Y. Lee, and I. Chung, "Fast speech recognition to access a very large list of items on embedded devices," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 803-807, July, 2008. Article (CrossRef Link)
[4] G. Riccardi and D. Hakkani-Tur, "Active and unsupervised learning for automatic speech recognition," in *Proc. of the Eurospeech*, pp. 1825-1828, September 1-4, 2003. Article (CrossRef Link)
[5] G. Riccardi, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504-511, June, 2005. Article (CrossRef Link)
[6] Y. Chang, S. Hung, N. Wang, and B. Lin, "CSR: A cloud-assisted speech recognition service for personal mobile device," in *Proc. of the International Conference on Parallel Processing*, September 13-16, pp. 305-314, 2011. Article (CrossRef Link)
[7] D. Kwon, H. Lim, W. Lee, H. Kim, S. Jung, T. Suh, and K. Nam, "A personalized English vocabulary learning system based on cognitive abilities related to foreign language proficiency," *KSII Transactions on Internet and Information Systems*, vol. 4, no. 4, pp. 595-617, August, 2010. Article (CrossRef Link)
[8] H. Liao, Y. Guan, J. Tu, and J. Chen, "A prototype of an adaptive Chinese pronunciation training system," *System*, vol. 45, pp. 52-66, August, 2014. Article (CrossRef Link)
[9] R. Sproat, W. Gale, C. Shih, and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese," *Computational Linguistics*, vol. 22, no. 3, pp. 377-404, September, 1996. Article (CrossRef Link)
[10] K. Kim, "Recurrent neural network for discrete input and its use in the implementation of Korean syllable-based language model," *Ph.D. Thesis*, Sogang University, 2016. Article (CrossRef Link)
[11] D. Lee, U. Park, J. Kim, J. Park, J. Park, and G. Jang, "Automatic generation of sub-word recognition units for unlimited Korean lexicon in continuous speech recognition system," in *Proc. of the 2nd International Conference on Electrical, Engineering, Computer Science*, pp. 42-43, Aug., 2016. Article (CrossRef Link)
[12] O. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, vol. 39, pp. 287-300, February, 2003. Article (CrossRef Link)
[13] S. Lee, J. Seo, and Y. Oh, "A Korean part of speech tagging system with handling unknown words," in *Proc. of the International Conference on Computer Processing of Oriental Languages*, pp. 164-171, November, 1995. Article (CrossRef Link)
[14] Y. Park, D. Ahn, and M. Chung, "Morpheme-based lexical modeling for Korean broadcast news transcription," in *Proc. of the Eurospeech*, pp.1129-1132, September 1-4, 2003. Article (CrossRef Link)
[15] O. Kwon, K. Hwang, and J. Park, "Korean large vocabulary continuous speech recognition using pseudomorpheme units," in *Proc. of the Eurospeech*, pp.483-486, September 5-9, 1999. Article (CrossRef Link)
[16] K. Lee and M. Chung, "Pseudo-morpheme-based continuous speech recognition," in *Proc. of the Speech Communication and Signal Processing Workshop*, pp. 309-314, August, 1998.
[17] K. Lee and M. Chung, "Morphological analysis of spoken Korean based on pseudo-morphemes", in *Proc. of the Conference of Korean Information Processing*, pp. 396-404, October 9-10, 1998. Article (CrossRef Link)
[18] M. Chung and K. Lee, "Modeling cross-morpheme pronunciation variations for Korean large vocabulary continuous speech recognition," in *Proc. of the Eurospeech*, pp. 261-264, September 1-4, 2003. Article (CrossRef Link)
[19] Y. Park and M. Chung, "Automatic Generation of Concatenate Morphemes for Korean LVCSR," *Journal of the Acoustical Society of Korea*, vol. 21, no. 4, pp. 407-414, 2002. Article (CrossRef Link)

[20] M. Schuster, "Japanese and Korean voice search," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 5149-5152, March 25-30, 2012. Article (CrossRef Link)
[21] L. Tomokiyo and K. Ries, "What makes a word: Learning base units in Japanese for speech recognition," in *Proc. of the ACL Special Interest Group in Natural Language Learning*, pp. 60-69, August 18-20, 1997. Article (CrossRef Link)
[22] C. Lee, B. Juang, F. Soong, and L. Rabiner, "Word recognition using whole word and subword models," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 683-686, May 23-26, 1989. Article (CrossRef Link)
[23] H. Yu, H. Kim, J. Hong, M. Kim, and J. Lee, "Large vocabulary Korean continuous speech recognition using a one-pass algorithm," in *Proc. of the International Conference of Spoken Language Processing*, pp. 278-281, October 16-20, 2000. Article (CrossRef Link)
[24] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251-272, August, 1991. Article (CrossRef Link)
[25] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4-16, January, 1986. Article (CrossRef Link)
[26] J. Odell, "The use of context in large vocabulary speech recognition," *Ph.D. Thesis*, Cambridge University, 1995. Article (CrossRef Link)
[27] H. Kim, C. Seon, and J. Seo, "Review of Korean speech act classification: Machine learning method," *Journal of Computing Science and Engineering*, vol. 5, no. 4, pp.288-293, December, 2011. Article (CrossRef Link)
[28] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 845-848, May 13-17, 2002. Article (CrossRef Link)
[29] R. Singh, B. Raj, and R.M. Stern, "Automatic generation of phone sets and lexical transcriptions," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1691-1694, June 5-9, 2000. Article (CrossRef Link)
[30] C. Schillo, G. A. Fink, and F. Kummert, "Grapheme based speech recognition for large vocabularies," in *Proc. of the International Conference on Spoken Language Processing*, pp. 129-132, October 16-20, 2000. Article (CrossRef Link)

Donghyun Lee received his B.E. degree in Computer Science and Engineering from Sogang University in 2013. He is currently pursuing a Ph.D. degree in Computer Science and Engineering at Sogang University. His research interests include speech recognition and spoken multimedia content search.



Jae-Hyun Park received his B.E. and M.E degrees in Computer Science and Engineering from Sogang University in 2011 and 2013, repectively. His research interests include speech recognition and spoken multimedia content search. From 2013, he is a research engineer in LG Electronics Institute of Technology, where he is engaged in development of speech recognizers for mobile devices.

Kwang-Ho Kim received his B.E. degree in Computer and Communications Engineering from Kangwon National University in 2008 and his M.E. and Ph.D. degree in Computer Science and Engineering from Sogang University in 2010 and 2016, respectively. From 2016, he is a research engineer in AIZEN Global Inc. His research interests include speech recognition, spoken multimedia content search and natural language processing.

Jeong-sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. From 2010 to 2011, he was a Post-Doc. researcher in the Computer Science Department, KAIST. He had been a faculty member in Mokwon University. He is now an assistant professor in the Information and Communication Engineering, Yeungnam University. His research interests include speech emotion recognition, speech recognition, speech enhancement, and voice interface for human-computer interaction.

Ji-Hwan Kim received the B.E. and M.E. degrees in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 1996 and 1998 respectively and Ph.D. degree in Engineering from the University of Cambridge in 2001. From 2001 to 2007, he was a chief research engineer and a senior research engineer in LG Electronics Institute of Technology, where he was engaged in development of speech recognizers for mobile devices. In 2004, he was a visiting scientist in MIT Media Lab. Since 2007, he has been a faculty member in the Department of Computer Science and Engineering, Sogang University. Currently, he is a full professor. His research interests include spoken multimedia content search, speech recognition for embedded systems and dialogue understanding.

Gil-Jin Jang is an assistant professor at Kyungpook National University, South Korea. He received his B.S., M.S., and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea in 1997, 1999, and 2004, respectively. From 2004 to 2006 he was a research staff at Samsung Advanced Institute of Technology and from 2006 to 2007 he worked as a research engineer at Softmax, Inc. in San Diego. From 2008 to 2009, he joined Hamilton Glaucoma center at University of California, San Diego as a postdoctoral employee. He had been a faculty member in UNIST. His research interests include acoustic signal processing, pattern recognition, speech recognition and enhancement, and biomedical signal engineering.

Unsang Park received the B.S. and M.S. degrees from the Department of Materials Engineering, Hanyang University, Seoul, Korea, in 1998 and 2000, respectively. He received the M.S. and Ph.D. degrees from the Department of Computer Science and Engineering, Michigan State University, MI, USA in 2004 and 2009, respectively. He has been an assistant professor in the Department of Computer Science and Engineering at Sogang University since 2012. His research interests include pattern recognition, image processing, computer vision, and machine learning.