Learning Lexicons from Spoken Utterances Based on Statistical Model Selection

Ryo Taguchi 1,2, Naoto Iwahashi 1,3, Takashi Nose 1,4, Kotaro Funakoshi 5, Mikio Nakano 5

1 Advanced Telecommunications Research Institute International
2 Nagoya Institute of Technology
3 National Institute of Information and Communications Technology
4 Tokyo Institute of Technology
5 Honda Research Institute Japan Co., Ltd.
tag@nitech.ac.jp

Abstract
This paper proposes a method for the unsupervised learning of lexicons from pairs of a spoken utterance and an object as its meaning without any a priori linguistic knowledge other than a phoneme acoustic model. In order to obtain a lexicon, a statistical model of the joint probability of a spoken utterance and an object is learned based on the minimum description length principle. This model consists of a list of word phoneme sequences and three statistical models: the phoneme acoustic model, a word-bigram model, and a word meaning model. Experimental results show that the method can acquire acoustically, grammatically and semantically appropriate words with about 85% phoneme accuracy.

Index Terms: Lexical learning, language acquisition, model selection.

1. Introduction
Dialog systems such as communication robots need to have linguistic knowledge. However, the developers cannot describe all knowledge in advance because such systems can be used in situations other than that the developers assume. Therefore, it is desirable that systems automatically learn such knowledge through interactions with users. In particular, household robots have many opportunities to encounter unknown objects. In order to recognize and utter words indicating them, they have to be able to learn correct phoneme sequences and meanings of the words from user utterances.

It is very difficult to obtain the correct phoneme sequences of unknown words included in user utterances. In previous work [1], phoneme sequences are obtained using pre-defined fixed expressions in which unknown words are inserted, such as "my name is <name>", where any name can replace <name>. Gorin et al. [2], Alshawi[3] and Roy and Pentland [4] have conducted experiments to extract semantically useful phoneme sequences from natural utterances but they have not yet been able to acquire the correct phoneme sequences with high accuracy. On the other hand, in the field of speech recognition, several methods to extract out-of-vocabulary (OOV) words from utterances by learning and using acoustic and grammatical models of classes of OOV words, such as personal names or place names, have been proposed [5,6,7]. These studies, however, have not been able to improve the accuracy for each phoneme sequence of OOV words by integrating recognition results of multiple utterances.

This paper proposes a method for learning phoneme sequences of words and relationships between them and objects (hereafter meanings) from various speeches from a user, without any prior linguistic knowledge other than an acoustic model of Japanese phonemes.

2. Lexical learning task
This paper deals with the task where a robot learns the name of an object from a user’s instruction by voice while showing the object to the robot. We assume that objects can be visually identified without errors and a module for word acquisition can receive IDs of objects as the identification results. The user uses natural expressions. User utterances may include words other than names of objects. For example, the user might say "this is James." In this paper, names of objects are called keywords, and words (or phrases) other than keywords are called non-keyword expressions. We assume that the keywords and the non-keyword expressions are independent of each other. Therefore, the same non-keyword expressions can be used in instruction utterances for different keywords.

The robot has never been given linguistic knowledge other than an acoustic model of Japanese phonemes. By using the acoustic model, it can recognize user utterances as Japanese phoneme sequences, but cannot extract keywords from them. The robot must learn the correct phoneme sequences and the meanings of keywords from a set of pairs of an utterance and an object ID. After learning, we estimate the learning result by investigating whether it can output the correct phoneme sequence corresponding to each object ID.

3. Approach
The main problems of the above task are word boundary detection from the utterances, and phoneme sequence estimation of the words. Since the phoneme sequences obtained by recognizing utterances may contain errors, it is difficult to correctly identify the word boundaries. For example, Roy and Pentalnd [4] has extracted keywords by using similarities of both acoustic features and meanings, but 70% of the extracted words contained insertion or deletion errors at either or both ends of words. Moreover, this method did not have a mechanism to select the most correct phoneme sequence for each word from a lot of word candidates obtained through learning.

In order to solve these problems, in our method, a statistical model of the joint probability of an utterance and an object is learned based on the minimum description length (MDL) principle. This model consists of a word list, in which each word is represented by a phoneme sequence, and three statistical models: the phoneme acoustic model, a word-bigram model, and a word meaning model. We call this model the utterance-object joint probability model. The phoneme
acoustic model has been learned beforehand because it requires much more speech data. By alternating between the learning of the other two statistical models and the optimization of the word list, acoustically, grammatically and semantically appropriate phoneme sequences are acquired as words.

4. Utterance-object joint probability model

Figure 1 shows the graphical model that represents the joint probability \( P(A, O) \) of spoken utterance \( A \) and object \( O \). In this model, word sequence \( S=(W_0, W_1, \ldots, W_{L-1}) \) outputs spoken utterance \( A \), and each word \( W_j \) represents object \( O \). \( W_0 \) is the start point and \( W_{L-1} \) is the end point. Some of words included in \( S \) are keywords and the others are non-keyword expressions. We formulate this model as follows:

\[
\log P(A, O; \theta) = \log \sum_s P(A, O, S = s; \theta) = \log \sum_s [P(A | S = s; \theta)P(S = s; \theta)P(O | S = s; \theta)]
\]

\[
= \max_s \left\{ a_1 \log P(A | S = s; \theta) + a_2 \log \sum_{j=1}^{L'} P(W_{s+1} | W_s; \theta) + a_3 \log \sum_{j=1}^{L'} \gamma(W_s, s, \theta)P(O | W_s; \theta) \right\}
\]

where \( \theta \) is the parameter set of this model, \( s \) is a word sequence, \( L' \) is the number of words included in \( s \).

The three terms on the right-hand side of Eq. (1) respectively represent the acoustic score, the grammatical score and the semantic score. The acoustic score is the likelihood of the word sequence \( s \), which is calculated using the phoneme acoustic model as usual speech recognition systems do. The grammatical score is calculated from the word-bigram language model. Note that we use a class bigram model in which all keywords are treated as words in the keyword class. The semantic score is a weighted average of word meanings \( P(O | W) \) using a weighting function \( \gamma(W_s, s, \theta) \) defined as follows:

\[
\gamma(W_s, s, \theta) = \frac{N(W_s)}{N(s)}
\]

where \( N(W) \) is the total number of phonemes of word \( W \), \( N(s) \) is the total amount of phonemes of keywords included in a word sequence \( s \). \( \gamma(W_s, s, \theta) \) is assigned zero when word \( W_s \) is not a keyword. This means the meaning of an utterance is inferred from keywords in the utterance. A method for determining whether or not a word is a keyword is described in section 5.2.

Moreover, in order to adjust differences of the accuracies of the three statistical models, we multiply those scores by weighting parameters \( a_1, a_2, a_3 \).

5. Lexical learning method

Figure 2 gives an overview of our method for lexical learning. It consists of three steps, namely building the initial word list, the learning of the word-bigram language model and the word meaning model, and the optimization of the word list based on a model selection technique.

5.1. Step1: Building of the initial word list

At first, all user utterances are recognized as phoneme sequences by using the phoneme acoustic model. Next, a word list is built by extracting subsequences included in the phoneme sequences. The entropies of phonemes before or after each subsequence are calculated. If the entropies of a subsequence are not zero and the frequency of the subsequence is more than two, the subsequence is registered on the word list as a word candidate.

5.2. Step2: Learning of the word-bigram language model and the word meaning model

The utterances are re-recognized as word sequences using both the phoneme acoustic model and the word list. Note that N-best hypotheses are output as a recognition result for each utterance in this system. The word-bigram language model is learned from the word sequences included in the N-best hypotheses. The word meaning model is defined by the conditional probability distribution \( P(O | W) \) of an object \( O \) given a word \( W \). This model is learned from the co-occurrence frequencies between the recognized words and the shown objects.

Additionally, in order to determine whether or not a word is a keyword, the difference between the entropy \( H(O) \) and the conditional entropy \( H(O | W = w) \) is calculated as follows:

\[
I(O | W = w) = H(O) - H(O | W = w) = -\sum_o P(O = o) \log P(O = o) + \sum_o P(O = o | W = w) \log P(O = o | W = w)
\]

If the difference \( I(O | W = w) \) is higher than a certain threshold, the word \( w \) is considered as a keyword.
5.3. Step3: Optimization of the word list by model selection

The number of words in the word list is optimized based on the minimum description length principle [8,9] which is the basis for model selection. In this paper, we define the description length \( DL(\theta) \) as follows:

\[
DL(\theta) = -L(\theta, D) + \frac{f(\theta)}{2} \log M
\]  
\( (4) \)

\[
L(\theta, D) = \sum_{i=1}^{K} \log P(A = a_i, O = o_i; \theta)
\]  
\( (5) \)

\[
f(\theta) = K^2 + CK
\]  
\( (6) \)

where \( L(\theta, D) \) is a log likelihood of \( \theta \), \( D = \{d_i \mid 1 \leq i \leq M\} \) is the set of learning data \( d_i = (a_i, o_i) \). \( f(\theta) \) is the number of parameters of the word-bigram language model and the word meaning model, and represents the degrees of freedom of \( \theta \). \( K \) is the number of words and \( C \) is the number of objects.

The optimization of the word list requires calculating the log likelihoods in all combinations of possible word candidates. However, it is computationally expensive and not practical. Therefore, using the N-best hypotheses obtained in Step 2, we approximately calculate the difference of the description lengths of two models, one that includes word \( w \) and the other that does not. The process returns to Step 2 and re-learns the word bigram-language model and the word meaning model by using the new list. Through the iterations of these processes, acoustically, linguistically, and semantically useful words are acquired.

A new word list is built by merging both results. After that, the process returns to Step 2 and re-learns the word bigram-language model and the word meaning model by using the new list. Through the iteration of these processes, acoustically, linguistically, and semantically useful words are acquired. For example, Figure 3 shows the variation of the description length in the word deletion process (more than 50 words are omitted). In the first optimization ("1st" in this figure), word deletion halted at 32 words because DL of the 31 words was higher than DL of the 32 words. A new word list consisting of 46 words was built by integrating the 32 words and the 14 words that were made by the word concatenation. Through the iterations of Step 2 and 3, the number of newly added words gradually decreases and the number of words is convergent.

5.4. Keyword output

When the robot receives an object \( o_i \), it outputs the keyword \( \tilde{w} \) that is the best to represent the object \( o \) based on Eq. (7).

\[
\tilde{w} = \arg \max_{w \in \Omega} \left\{ \omega_1 \log P(W = w; \theta) + \omega_2 \log P(O = o \mid W = w; \theta) \right\}
\]  
\( (7) \)

where \( \Omega \) is the set of acquired keywords. In order to estimate the phoneme accuracy for the keyword corresponding to each object, we reduce Eq. (1) for multiple words to Eq. (7) for a single word. Note that, Eq. (7) does not include the acoustic score because the keywords are not converted into speech.

6. Experimental results and discussions

6.1. Experiment 1: The selection of weighting parameters

First, in order to determine the values of weighting parameters \( \omega_1, \omega_2, \omega_3 \), we investigated the influence of them using spoken utterance data from one person. The number of the objects is ten. The keywords corresponding to the objects are shown in Table 1. Six non-keyword expressions are used such as "kokowa <keyword> desu" and "konobayowa <keyword>", where each keyword can replace <keyword>. The sixty utterances which consist of all combinations of them were recorded. Some sets of the values of the weighting parameters were prepared heuristically as candidates. In the experiment, the constraint, \( \omega_3 = \omega_1 \), was imposed for simplicity in exploring the optimal value set. The release of this constraint is future work.

The results of the experiment, in which the iterations of Step 2 and Step 3 were carried out ten times, are shown in Figure 4 and 5. Figure 4 shows the number of the words registered in the word list. We can see that more than one hundred words remained in the case where weighting was not adopted (\( \omega_1 = \omega_2 = \omega_3 = 1 \)), while the number of the words decreased when using the other sets of values. Figure 5 shows the phoneme accuracy for the keywords, which were extracted according to Eq. (7), for ten objects. While the accuracy was less than 70% without weighting, approximately 90% accuracy, which was the highest in all weight combinations, was obtained using values \( \omega_1 = 0.0001, \omega_2 = \omega_3 = 10 \).

6.2. Experiment 2: Effectiveness of the optimization of the word list

Setting the weighting values \( \omega_1 = 0.0001, \omega_2 = \omega_3 = 10 \) which led to the highest phoneme accuracy in the above selection process, the effectiveness of the optimization of the word list
Table 1: Keywords used in the experiments.

<table>
<thead>
<tr>
<th>O</th>
<th>Keyword</th>
<th>O</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/kaigishitsu/</td>
<td>6</td>
<td>/takeuchisainobuusunominami/</td>
</tr>
<tr>
<td>2</td>
<td>/tsuzinosa/</td>
<td>7</td>
<td>/koosakushitsu/</td>
</tr>
<tr>
<td>3</td>
<td>/furoana/</td>
<td>8</td>
<td>/ashimonohaya/</td>
</tr>
<tr>
<td>4</td>
<td>/gakuseebye/</td>
<td>9</td>
<td>/sumatorumuu/</td>
</tr>
<tr>
<td>5</td>
<td>/ochanomiba/</td>
<td>10</td>
<td>/sumatorumunoigichii/</td>
</tr>
</tbody>
</table>

Table 2: The examples of obtained keywords before and after the optimization of the word list.

<table>
<thead>
<tr>
<th>O</th>
<th>before the optimization</th>
<th>O</th>
<th>after the tenth optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/kaigishitsu/</td>
<td>O</td>
<td>/kaigishitsu/</td>
</tr>
<tr>
<td>2</td>
<td>/tsuzinosa/</td>
<td>7</td>
<td>/tsuzinsuanobusu/</td>
</tr>
<tr>
<td>3</td>
<td>/naka/</td>
<td>8</td>
<td>/furoana/</td>
</tr>
<tr>
<td>4</td>
<td>/gakuseebye/</td>
<td>9</td>
<td>/gakuseebye/</td>
</tr>
<tr>
<td>5</td>
<td>/ochi/</td>
<td>10</td>
<td>/ochinoma/</td>
</tr>
<tr>
<td>6</td>
<td>/naka/</td>
<td></td>
<td>/furoana/</td>
</tr>
</tbody>
</table>

was evaluated in Experiment 2. The spoken utterance sets of seventeen speakers were recorded, each of which contains the same sentences as the set used in Experiment 1. For each speaker, the performance of the proposed method was evaluated. Figure 6 shows the average results among all speakers. The horizontal axis represents the number of learning iterations. In the figure, the histogram indicates the number of acquired words and keywords included in the word list. We can see that the iterations made the number of the words decrease. Finally, thirteen keywords were obtained on average. This number is close to ten, which is the real number of the keywords in the training utterance set.

Before lexical learning, the phoneme accuracy in all the utterances was 82%. The keywords were extracted according to Eq. (7), and the phoneme accuracy for the keywords was evaluated. The phoneme accuracy was less than 50% without optimization. In contrast, by iterating Step 2 and Step 3 the accuracy increased up to 85%. Table 2 shows the examples of obtained keywords before and after iterating them ten times. We can see that word segmentation errors were fixed. These results prove that the proposed method makes it possible to decide the boundary of keywords appropriately.

### 7. Conclusions

This paper proposed a method to learn phoneme sequences, meanings and bigrams of words from spontaneous speeches. Experimental results show that, without a priori word knowledge, it can acquire phoneme sequences of object names with about 85 percent accuracy. We expect the basic principle presented in this study will provide us with a clue to resolve the general language acquisition problem in which morphemes of speech language are extracted by using only non-linguistic semantic information related to each utterance.

### 8. References