IMPROVED LEXICON MODELING FOR CONTINUOUS SPEECH RECOGNITION

Seong Jin Yun, Yung Hwan Oh* and Gyung Chul Shin***

* Department of Computer Science,
Korea Advanced Institute of Science and Technology,
373-1 Kusong-dong, Yusong-gu,
Taejon 305-701, Korea
E-mail: sjyun@bulsai.kaist.ac.kr

** ATM M&A S/W Section,
Electronics & Telecommunications Research Institute,
Yusong-gu, Taejon, 305-350, Korea
E-mail: neuro@nice.etri.re.kr

ABSTRACT

We propose the stochastic lexicon model which represents the pronunciation variations to optimally cope with the continuous speech recognizer. In this lexicon model, the baseform of words are represented by subword states and probability distribution of subwords as hidden Markov model. Also, proposed approach can be applied to system employing non-linguistic recognition units and lexicon is automatically trained from a training utterances. In speaker independent speech recognition tests using a 3000 word continuous speech database, the proposed system improves the word accuracy by about 27.8% and the sentence accuracy by about 22.4%.

1. INTRODUCTION

Most large vocabulary speech recognizers employ subwords as basic modeling units. This implies that in order to obtain word(or sentence) recognition, a lexicon which defines the composition of the words in terms of basic units must be made available to the recognizer. In most cases, linguistically defined subwords are used as the basic recognition units, typically phonemes or phone-like units.

The lexicon is commonly created by the knowledge of human experts or by the use of standard pronunciation dictionaries. Theses approaches have particular problems, e.g., in pronunciation variations of many speakers of different dialects need to be represented by one or a multiple lexical entries. Also, traditional approaches cannot be readily applied to systems employing non-linguistic recognition units.

We describe a method for deriving a stochastic representation of a word baseform from sample utterances. This method result in a substantial decrease in the recognition error rate compared to methods based on standard phonetic representations of words.

This paper is organized as follows. Section 2 introduces the methods to obtain the stochastic lexicon form given subword units and sample utterances. Next, we describes the recognition procedures with the stochastic lexicon model. Experimental results are provided in section 3, and the conclusion is given in section 4.

2. STOCHASTIC LEXICON MODEL

The examination of pronunciation variants is an important problem, because standard pronunciation does not optimally describe real speech. For this reason word modeling in the form of powerful pronunciation lexicons is important. This pronunciation lexicon must be expanded by pronunciation variants to optimally cope with the tasks required by speech recognition as well as by language processing.

2.1 Generation of Stochastic Lexicon

Each word is represented by a sequence of phones, called the phonetic baseform of the word. A hidden Markov model (HMM) is established for each phone. The Markov model for a word is obtained by replacing each phone in the baseform for the word by its Markov model. Figure 1(a) shows an example of a Markov model for a phone, a phonetic baseform, and the resulting Markov model for the word form of a phonetics-based recognition system.

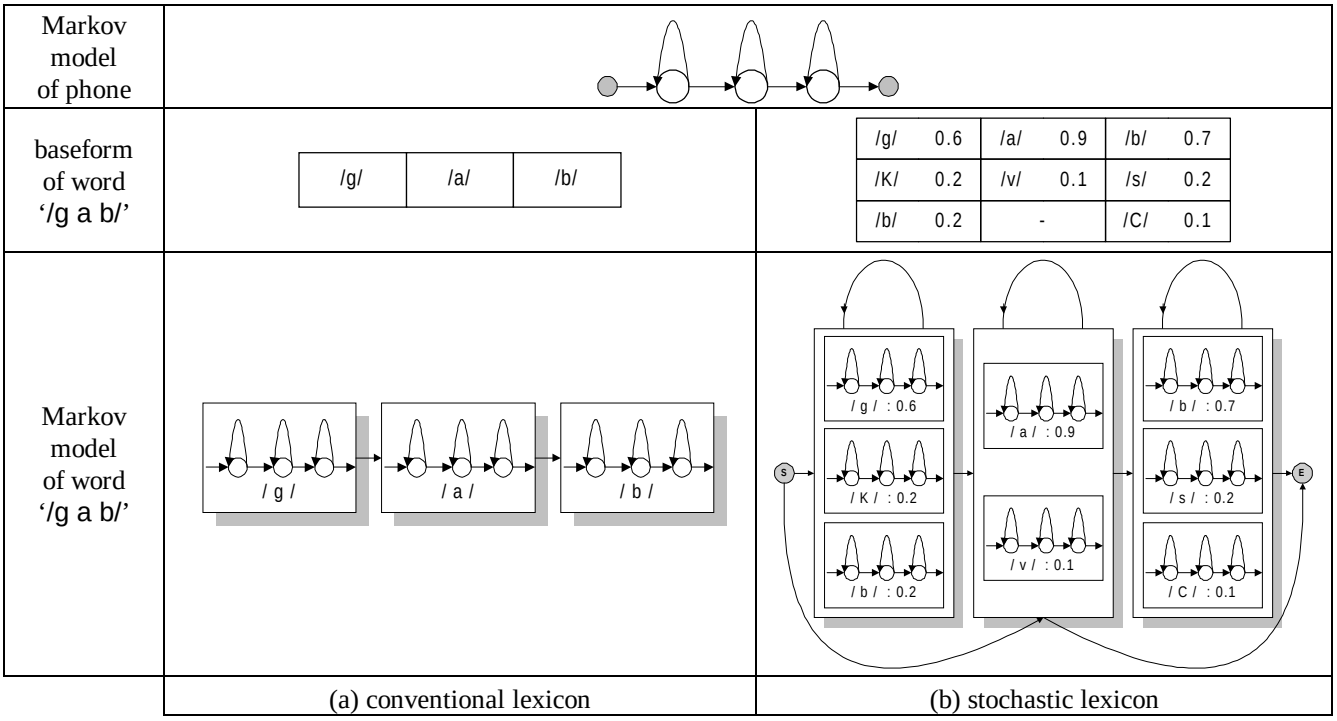

Figure 1. Example of phone model, baseform and word model

A phonetic baseform of a word is deterministic in the case of traditional approaches. However, the proposed stochastic lexicon has stochastic baseforms that can effectively represent pronunciation variations. The stochastic lexicon is similar to a multiple-baseform based lexicon from the viewpoint of pronunciation variations modeling. Compared to the multiple baseform-based lexicon which expands the baseform into a graph deterministically representing the various pronunciations of a word, the stochastic lexicon uses a probability distribution of subword units.

As shown in Figure 1(b), We can regard stochastic lexicon as HMM. HMM is a stochastic finite state automata which consisting of a Markov chain of subword states with a probabilistic function for each of the states, modeling the emission and observation of subword units. Each subword state in the baseform has a probability distribution of subword units. These probability distributions can be obtained from the likelihood of the subword model and the subword segment. In this method, an acoustic representation of a word can be derived automatically from sample sentence utterances. Additionally, the stochastic baseform is further optimized to the subword model and recognizer. The stochastic lexicon generation algorithm is as follows:

1. Subword unit segmentation.

Train the phone model with ***segmental K-means***[5] training method and collect the subword segments S_{ij} from each training utterance. where, S_{ij} is a subword segment of the *j-th* subword unit and the *i-th* word.

$$S_{ij} = \{ S_{ij}^1, S_{ij}^2, ..., S_{ij}^K \}$$

K ; number of subword segment

2. Subword unit recognition.

For each baseform unit (subword unit) W_{ij}, perform the phone recognition with the subword segments S_{ij}^k and compute the contribution of subword model $P_{ijv}(k)$. We obtain $P_{ijv}(k)$ through fuzzy distance in equation (2).

$$P_{ijv}(k) = p(\lambda_v | W_{ij}, S_{ij}^k) \approx F_{ijv}^d(k) \Big/ \sum_p F_{ijp}^d(k) \quad (1)$$

$$F_{ijv}(k) = \left[\sum_{p=1}^{P} [D(S_{ij}{}^k | \lambda_v) / D(S_{ij}{}^k | \lambda_p)]^{1/(d-1)} \right]^{-1} \quad (2)$$

$$D(S_{ij}^k | \lambda_v) = -\log p(S_{ij}^k | \lambda_v) \quad (3)$$

where, $L = \{ W_1, W_2, ..., W_N \}$; lexicon,
$W_i = \{ W_{i1}, W_{i2}, ..., W_{iM} \}$; baseform of word,
$\lambda = \{ \lambda_1, \lambda_2, ..., \lambda_P \}$; subword HMM,

N ; number of words

M ; number of subword unit in the *i-th* word

P ; number of subword unit

d ; degree of fuzziness ($d > 1$)

3. Confusion probability of subword units

Compute the confusion matrix of subword units. This confusion matrix is applied to backing-off scheme.

$$C_j^i = p(S_i|\lambda_j) = \frac{1}{K}\sum_{k=1}^{K}\left[F_{ij}^d(k)\Big/\sum_v F_{iv}^d(k)\right] \qquad (4)$$

4. Stochastic lexicon construction.

Compute the subword observation probability of each baseform unit W_{ij}.

$$P_{ijv} = p(\lambda_v|W_{ij}) = G_{ijv}\Big/\sum_p G_{ijp} \qquad (5)$$

$$G_{ijv} = \frac{1}{K+1}\left[\sum_{k=1}^{K} P_{ijv}(k) + C_v^{phone(ij)}\right] \qquad (6)$$

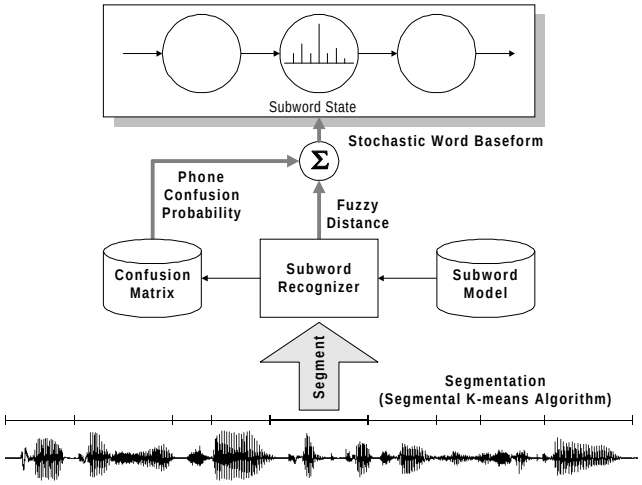

Figure 2. Block diagram of stochastic lexicon generation algorithm

2.2 Recognition with Stochastic Lexicon

The recognition procedure is based on the time synchronous beam search method as described in [1]. The search space can be described as a huge finite state network, consisting of nodes representing a certain state in the lexicon and the subword unit model. If a word has multiple pronunciations, the phonological rules can be used to expand the baseform into a graph representing the various pronunciations of the word. Thus there are many more search nodes in the case of multiple baseforms compared to single baseforms. However, the stochastic lexicon has the same number of search nodes as the case of the single baseform based lexicon.

In recognition, the state sequence is unknown, and all combinations of state and time must be hypothesized. The Viterbi algorithm is an efficient method for computing the probability $p(X_1 \ldots X_t|W_{ij})$, i.e. that, given subword model W_{ij}, the acoustic vectors $X_1 \ldots X_t$, are produced and cover the time interval *1...t*. Using an auxiliary quantity Q with the argument k for state and t for time, the best probability was obtained for the subword W_{ij}.

$$p(X_1 \ldots X_t|W_{ij}) = Q_{ijS(W_{ij})}(t)$$

$$Q_{ijk}(t) = \max_{k'}\left[a(s_k|s_{k'})\cdot Q_{ijk'}(t-1)\right] D_{ijk}(X_t) \qquad (7)$$

where the terminal state of subword W_{ij} is $S(W_{ij})$, and $a(s_k|s_{k'})$ is the state transition probability. If a single baseform is used in the pronunciation lexicon, the probability D is only obtained from the emission probability of the corresponding subword model as in equation (8). When the stochastic lexicon is used, the summation is carried out over all product of subword observation probability and emission probability of subword models as in equation (9). Therefore the computation increases slightly compared to the single baseform-based recognition.

$$D_{ijk}(X_t) = p(X_t|W_{ij}, s_k) = \begin{cases} p(X_t|s_k, \lambda_v) & \text{if subword of } W_{ij} \text{ is } \lambda_v \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

$$D_{ijk}(X_t) = p(X_t|W_{ij}, s_k) = \sum_{v=1}^{P} P_{ijv}\cdot p(X_t|s_k, \lambda_v) \qquad (9)$$

3. EXPERIMENTAL RESULTS

The proposed method has been tested on the Korean continuous speech recognition system which has a vocabulary of 3000 words. The system was tested on a total of 50 speakers (25 males, 25 females) and in the speaker independent mode. Feature vectors containing 30 parameters (14 MFCCs, normalized energy and differenced parameters) were computed every 10 ms using a 20 ms window. 36 context independent phone models or 3017 context dependent triphone models were used. Each subword model is represented by a left-to-right HMVQM(hidden Markov VQ model)[3] and its configurations are show in Table 1. Word bigram grammar is used for the recognition

experiments.

The proposed method is compared to a conventional lexicon with a single baseform. Table 2 shows the experimental results. In this experiment, the use of stochastic lexicon reduced the word error rate by 23.6(phone), 27.8(triphone)% and the sentence error rate by 6.7(phone), 22.4(triphone)%.

Table 1. Configurations of the subword model

	No. of model	No. of state	No. of codeword
phone model	36	3	20
triphone model	3017	3	10
silence model	2	1	8

Table 2. Recognition results

lexicon	subword unit	word accuracy (%)	sentence accuracy (%)
conventional lexicon	phone	58.13	26
	triphone	89.18	60.6
stochastic lexicon	phone	68.00	31
	triphone	92.19	67.8

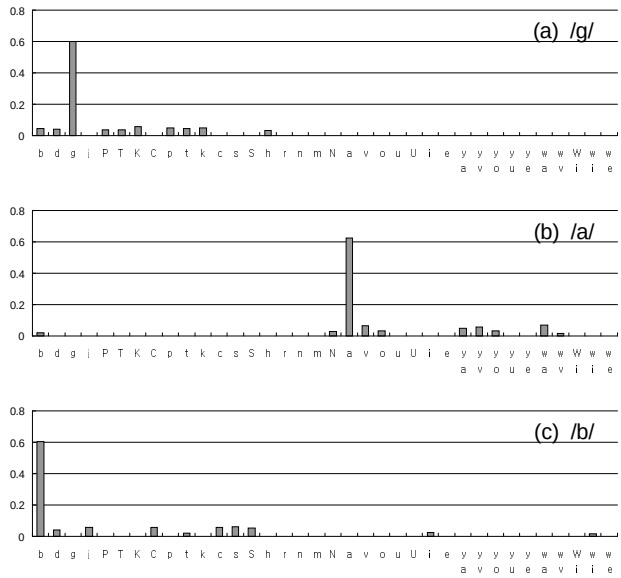

Figure 3. Example of stochastic baseform
(word : 값 / g a b /)

An example of resulting pronunciation for Korean word "/g a b/" are shown in Figure 3. In this figure, we see pronunciation variations which can often found in spontaneously spoken speech.

4. CONCLUSION

We have described the stochastic lexicon model allowing for pronunciation variations in speech recognition. In this lexicon model, the baseform of words are represented by hidden Markov model with subword probability distributions. Also, the stochastic lexicon automatically trained from a training utterances and further optimized to the subword model. From the experiments, the effectiveness of the proposed method has been confirmed. In the future, we will test the tree structured stochastic lexicon on the continuous speech recognition system.

REFERENCES

[1] H. Ney, D. Mergel, A. Noll, A.Paeseler, "Data Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition", IEEE Trans. on Signal Processing, Vol.40, No.2, pp. 272-281, Feb. 1992

[2] L.R.Bahl, P.F.Brown, P.V.de Souza, R.L. Mercer, M.A. Picheny, "A method for the Construction of Acoustic Markov Models for Words", IEEE Tans. Speech and Audio Processing, Vol.1, No.4, pp. 443-452, Oct. 1993.

[3] Seong Jin Yun and Yung Hwan Oh, "Performance Improvement of Speaker Recognition System for Small Training Data", Proc. ICSLP'94, pp. 1863-1866, Yokohama, 1994

[4] Torbjϕrn Svendsen, Frank K. Soong, Heiko Purnhagen, "Optimizing Baseforms for HMM-Based Speech Recognition", Proc. Eurospeech'95, pp. 783-785, Madrid, 1995

[5] Lawrence Rabiner, Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993

[6] Youngjik Lee, Kyu-Woong Hwang, "Selecting Good Speech Features for Recognition", ETRI journal, Vol. 18, No. 1, pp. 29-40, Apr. 1996