A NEW METHOD FOR AUTOMATIC GENERATION OF SPEAKER-DEPENDENT PHONOLOGICAL RULES

Toru Imai, Akio Ando, and Eiichi Miyasaka

NHK (Japan Broadcasting Corp.) Science and Technical Research Laboratories
1-10-11, Kinuta, Setagaya, Tokyo, 157, Japan
E-mail: imai@strl.nhk.or.jp

ABSTRACT

This paper presents a new method for automatic generation of speaker-dependent phonological rules in order to decrease recognition errors caused by pronunciation variability. The proposed method generates phonological rules by using objective speaker's continuous speech and corresponding standard pronunciation, resulting in forming a multiple-pronunciation dictionary from a single-pronunciation dictionary. The method makes it possible to generate automatically speaker-dependent and recognizer-dependent phonological rules, and be applied to both a top-down recognizer and a bottom-up recognizer, while conventional methods are based on hand-derived general phonological rules such as coarticulation knowledge or are applied only to a bottom-up recognizer. Phrase recognition experiments with concatenated phoneme HMMs showed that the generated rules can decrease recognition errors and play a role of speaker adaptation at the phonological level.

1. INTRODUCTION

Pronunciation variability of words is a major problem in automatic speech recognition. Local acoustic variation in a phoneme is solved by a statistical acoustic model such as an HMM, but phoneme variation of pronunciation should be solved by a linguistic approach. In our application of Japanese speech recognition, a caption-superimposing system for TV programs[1], we designed a pronunciation dictionary with expected multiple entries for a word by hand. For example, a Japanese phrase "hatakikomide" is sometimes uttered as "hatakkomide" or "hatakikonde." The registration of multiple entries is troublesome because the pronunciation variability is dependent on speakers in almost cases.

Several researches to cope with the pronunciation variability have been reported. Methods based on linguistically heuristic knowledge or hand-derived general phonological rules such as coarticulation knowledge have led to performance improvement [2][3][4]. They have difficulties, however, to generate speaker-dependent and recognizer-dependent phonological rules, because the general phonological rules do not consider the speaking habit with each speaker. There is an automatic method by mappings between baseform pronunciations and hand-transcribed pronunciations [5], but it is troublesome to transcribe the actual pronunciations for a new speaker by hand. Methods to create pronunciation networks [5][6] require a large data set in order to obtain probabilities embedded in the networks. Another method is to correct phonetic recognition results by pre-examined phonetic recognition errors [7]. Though the method can generate speaker-dependent phonological rules, they are applied only to a bottom-up recognizer such as a segment-based phonetic recognizer and not applied to a top-down recognizer such as concatenated phoneme HMMs according to phonetic transcriptions of words or sentences.

To cope with the pronunciation variability without the problems of the conventional methods, we present here a new method to generate phonological rules. The proposed method generates phonological rules from unsuited phonetic symbols for objective speaker's continuous speech whose contents are independent of recognition tasks, and then it forms a multiple-pronunciation dictionary from a single-pronunciation dictionary. The method makes it possible to generate automatically speaker-dependent and recognizer-dependent phonological rules without giving hand-transcribed actual pronunciations. The method is applicable to both a top-down recognizer and a bottom-up recognizer.

2. GENERATION OF PHONOLOGICAL RULES

This section describes how to generate phonological rules automatically in the proposed method for a top-down recognizer. The rules are generated from three types of data: objective speaker's continuous speeches $\{S_k\}$ specially uttered for the rule generation, phonetic symbol sequences $\{P_k\}$ of standard pronunciation corresponding to $\{S_k\}$, and speaker-independent phoneme HMMs $\{Y_m\}$, as shown in Fig. 1. Though the contents of $\{S_k\}$ are not fixed, a phonetically balanced set is desired. In advance, the HMMs are trained by any training method such as a maximum likelihood estimation with phoneme data derived from a large number of speakers' speech. For each phoneme m, the average likelihood L_m and duration (the average μ_m, the standard deviation σ_m, and the minimum length τ_m) are stored after the HMM training. The Viterbi algorithm is applied to the continuous speech S_k with concatenated HMMs according to the standard pronunciation P_k, in order to obtain the total likelihood $L(S_k|P_k)$, the phoneme likelihood L^k_i, and the length t^k_i for the i-th phoneme X^k_i in P_k. In the Viterbi algorithm, the existence of pauses between every phrase in the sentence speech S_k is judged by using the pause HMM inserted between every phrase. From the detected phonemes with lower likelihood, shorter lengths, or longer lengths than the training phoneme data, tentative phonological rules are generated. If a new phonetic symbol sequence P_k' by the generated phonological rule improves the likelihood and the discriminative ability of $\{S_k\}$, the rule is accepted as one of the final phonological rules.

0-7803-2431-5/95 $4.00

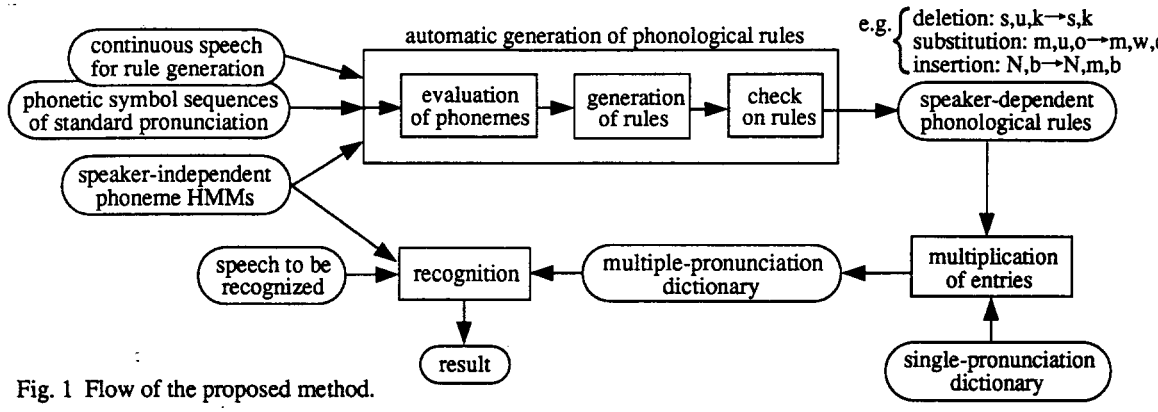

Fig. 1 Flow of the proposed method.

In the case of the application to a bottom-up recognizer, the likelihood under the standard pronunciation is calculated by a segment-based phonetic recognizer instead of the Viterbi algorithm, and unsuited phonetic symbols are detected in order to generate the tentative phonological rules. The other procedures are same as the method for the top-down recognizer.

2.1. Generation of Tentative Rules

We consider the four types of phonological rules from a biphone/triphone to a biphone/triphone:

(i) Deletion rules: ABC→AC (deleting phoneme B),
(ii) Substitution-1 rules: ABC→ADC (substituting phoneme B with phoneme D),
(iii) Substitution-2 rules: ABC→DC (substituting biphone AB with phoneme D),
(iv) Insertion rules: AB→ACB (inserting phoneme C).

Tentative phonological rules for (i)-(iv) are generated from all phonemes in $\{P_k\}$ as described below.

2.1.1. Deletion Rules

When a phoneme in a continuous speech is not uttered actually, there is a very fair possibility that the HMM for the expected phoneme shows shorter duration or lower likelihood by the Viterbi segmentation than those in the phoneme HMM training, as shown in Fig. 2. Therefore such an unsuited phonetic symbol is detected in order to generate a deletion rule. If i-th phoneme X^k_i in P_k is the phoneme m and one of the three conditions:

$$t^k_i < \mu_m - 2\sigma_m,\ t^k_i < \tau_m,\ L^k_i < L_m \qquad (1)$$

is satisfied, let P_k' be a new phonetic symbol sequence without the phoneme X^k_i. If the likelihood $L(S_k|P_k')$ by the concatenated HMMs according to P_k' satisfies the inequality:

$$L(S_k|P_k) < L(S_k|P_k'), \qquad (2)$$

let $X^k_{i-1}X^k_iX^k_{i+1} \rightarrow X^k_{i-1}X^k_{i+1}$ be a tentative deletion rule.

2.1.2. Substitution-1 rules

When a phoneme substitution occurs in an utterance, there is a very fair possibility that the HMM for the expected phoneme shows short duration or low likelihood by the Viterbi segmentation. Such an unsuited phonetic symbol is detected by the Viterbi algorithm in order to generate a substitution-1 rule. If X^k_i in P_k is the phoneme m and one of the inequalities (1) is satisfied, another phoneme is substituted in turn for the phoneme X^k_i in order to decide X_{new} which shows the highest likelihood for the speech S_k. When the inequality (2) is satisfied by a new sequence P_k' where X_{new} is substituted for X^k_i, let $X^k_{i-1}X^k_iX^k_{i+1} \rightarrow X^k_{i-1}X_{new}X^k_{i+1}$ be a tentative substitution-1 rule.

2.1.3. Substitution-2 Rule

When another phoneme is uttered instead of two consecutive phonemes in standard pronunciation, there is a very fair possibility that the HMMs for the consecutive phonemes show short duration or low likelihood by the Viterbi segmentation. If the sum of the duration for phoneme X^k_{i-1} and X^k_i in P_k or the sum of their likelihood is considered to be inferior to those in the phoneme HMM training, based on the similar inequalities as (1), another phoneme is substituted in turn for the phonemes $X^k_{i-1}X^k_i$ in order to decide X_{new} which shows the highest likelihood for the speech S_k. When the inequality (2) is satisfied by a new sequence P_k' where X_{new} is substituted for $X^k_{i-1}X^k_i$, let $X^k_{i-1}X^k_iX^k_{i+1} \rightarrow X_{new}X^k_{i+1}$ be a tentative substitution-2 rule.

2.1.4. Insertion Rules

When a continuous speech is uttered with any phoneme inserted between two phonemes in standard pronunciation, there is a very fair possibility that the HMMs for the two phonemes show

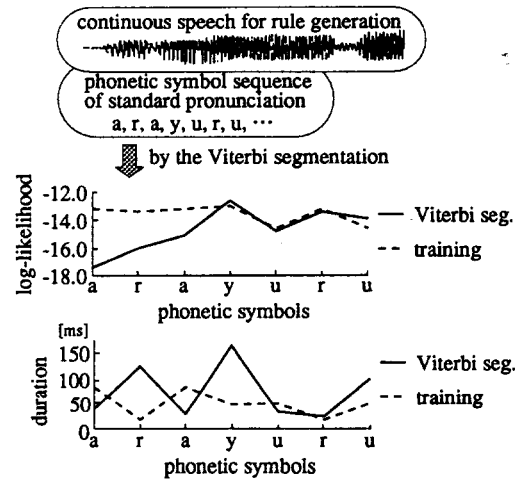

Fig. 2 Evaluation of phonemes.

long duration or low likelihood by the Viterbi segmentation. If the sum of the duration for phoneme X^k_{i-1} and X^k_i in P_k or the sum of their likelihood is considered to be inferior to those in the phoneme HMM training, based on the similar inequalities as (1), another phoneme is inserted in turn between the phonemes $X^k_{i-1}X^k_i$ in order to decide X_{new} which shows the highest likelihood for the speech S_k. When the inequality (2) is satisfied by a new sequence P_k' where X_{new} is inserted between $X^k_{i-1}X^k_i$, let $X^k_{i-1}X^k_i \rightarrow X^k_{i-1}X_{new}X^k_i$ be a tentative insertion rule.

2.2. Check on Rules

To select really effective rules, the generated tentative phonological rules are checked on, because the each tentative rule is generated by examining the effect on only one sentence in the continuous speeches $\{S_k\}$. Improvement of likelihood and discriminative ability by the tentative rules are examined on all the continuous speeches $\{S_k\}$.

2.2.1. Improvement of Likelihood

For each tentative phonological rule, phonetic symbol sequences of standard pronunciation including the left side of the rule are selected from $\{P_k\}$. The rule is applied to the selected sequences to obtain the average likelihood of the corresponding continuous speeches in $\{S_k\}$. If the average likelihood is lower than the likelihood without the phonological rule, the rule is deleted from the set of the tentative rules. If it is higher, the difference is stored as an effective value of the phonological rule. The values are used when forming a multi-pronunciation dictionary.

2.2.2. Improvement of Discriminative Ability

To examine the improvement of discriminative ability by the tentative rules, the speeches $\{S_k\}$ are divided into phrases at pauses by the Viterbi algorithm. The likelihood of the phrases is calculated by concatenated phoneme HMMs according to rule-applied multiple pronunciations of the phrases. If a phonological rule increases the difference of the likelihood between correct phrases and the nearest wrong phrases on average for all the phrases, the rule is accepted as one of the final phonological rules.

3. MULTIPLE-PRONUNCIATION DICTIONARY

The automatically generated phonological rules are applied to a single-pronunciation dictionary in order to obtain a multiple-pronunciation dictionary. For example, when the insertion rule "k,a→k,w,a" or the substitution-1 rule "i,g,a→i,ky,a" is applied to the single-pronunciation "s,e,k,a,i,g,a", the pronunciations "s,e,k,w,a,i,g,a" or "s,e,k,a,i,ky,a" are added to the multiple-pronunciation dictionary. For a word or a phrase in the single-pronunciation dictionary, the multiple-pronunciation dictionary adopts n sequences showing high effective values of likelihood among generated sequences by the phonological rules, where n is the number of phonemes in each word or phrase in the single-pronunciation dictionary. In the recognition experiments described below, we used the average likelihood for each pronunciation entry of phonetic symbol sequences, without weighting them.

4. RECOGNITION EXPERIMENTS

Recognition experiments on phrase speeches or Japanese "bunsetsu" speeches were performed by the proposed method. Discrete HMMs with three types of output symbols (LPC cepstral coefficients, differenced LPC cepstral coefficients, and differenced logarithmic power) were used for 41 Japanese phonemes. They were left-to-right models with two states for vowels and pauses, and four states for consonants. Training of the phoneme HMMs and the design of codebooks were carried out with lots of labeled phoneme data in the ATR Japanese database [8] by 15 male speakers to obtain speaker-independent acoustic models. Generation of the phonological rules was carried out with 50 phonetically balanced sentences (A-set) by a male speaker in the continuous speech database by the Acoustical Society of Japan (ASJ). Phonetic symbol sequences of standard pronunciation corresponding to the A-set sentences were transcribed by hand, independent of actual pronunciations. The test data were approximately 140 phrases extracted from the same speaker's 25 sentences whose contents were different from the data for the rule generation. The single-pronunciation dictionary for the test data consisted of phonetic symbol sequences of standard pronunciation corresponding only to the test phrases. A multiple-pronunciation dictionary was formed from the single-pronunciation dictionary by the generated phonological rules by the A-set sentences. The recognition was carried out by selecting the phrase showing the highest likelihood among the phrases in the dictionaries, with no grammar like word recognition.

The recognition experiments were performed for five speakers. In the five speakers' experiments, the common data set was used for the rule generation, but different data sets were used as the speech data to be recognized. Table 1 shows examples of the generated phonological rules for speaker *can0001* in the ASJ database. In the case, 1026 tentative rules were extracted from the 50 sentences, and the number was finally reduced to 599 by the check on the rules. Table 2 shows examples of entries in the multiple-pronunciation dictionaries. Fig. 3 shows the results of recognition rates with the single-pronunciation dictionaries or the multiple-pronunciation dictionaries for every speaker. The multiple-

Table 1 Examples of generated phonological rules.

type of rules	number	examples	effective values of log-likelihood
deletion	105	sh,i,t→sh,t	7.0
substitution-1	240	t,e,i→t,ii,i	25.7
substitution-2	78	u,h,e→y,e	22.9
insertion	176	a,o→a,w,o	9.3

(speaker: can0001)

Table 2 Examples of multiple pronunciations.

Japanese characters	世界が	限られた
single-pronunciation	s,e,k,a,i,g,a	k,a,g,i,r,a,r,e,t,a
multiple-pronunciation	s,e,k,a,i,g,a s,e,k,a,i,g,u,a s,e,k,a,i,ky,a s,e,k,r,a,i,g,a	k,a,g,i,r,a,r,e,t,a k,a,u,g,i,r,a,r,e,t,a k,a,g,i,r,y,r,e,t,a k,a,g,i,b,r,e,t,a

(speaker: can0001)

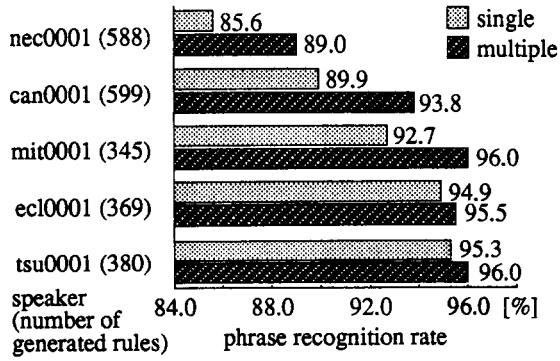

Fig. 3 Phrase recognition rates.

pronunciation dictionaries raised recognition rates by 2.4% on average and 3.9% at maximum from the rates by the single-pronunciation dictionaries. The multiple-pronunciation dictionaries gave good performance especially for the speakers showing relatively lower recognition rates by the single-pronunciation dictionaries. Fig. 4 shows examples of the likelihood for a phrase which was mis-recognized by the single-pronunciation dictionary but correctly recognized by the multiple-pronunciation dictionary. The figure makes it clear that the pronunciation entries giving higher likelihood than standard pronunciations were added to the multiple-pronunciation dictionaries. It is considered that these results show the effectiveness of the multiple-pronunciation dictionaries based on the automatically generated phonological rules.

We performed another experiment after speaker adaptation of the codebooks and the HMMs instead of speaker-independent models. They are adapted in a discriminative way [1][9] with the continuous speech for the rule generation and their standard pronunciations regardless of the pronunciation variability. The generated phonological rules improved recognition rates by 1.3% on average, though the effect was lower than the experiments without the adaptation of the codebooks and the HMMs. The reason is considered that the pronunciation variability was embedded in the acoustic models to some extent, in spite of the adaptation under wrong labels of phonetic symbols.

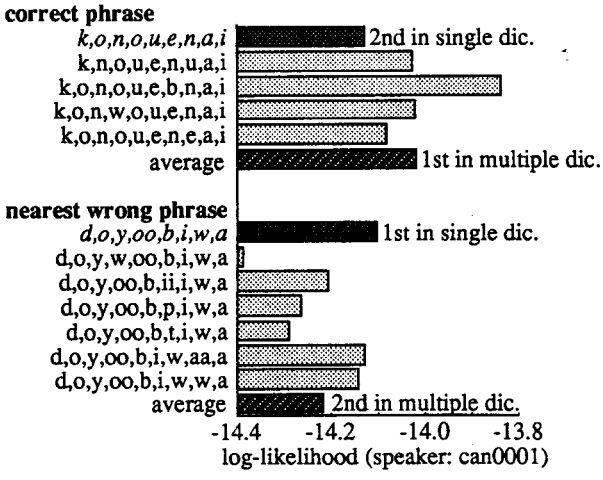

Fig. 4 Examples of effective multiple pronunciations.

5. CONCLUDING REMARKS

In this paper, we have presented a new method for automatic generation of speaker-dependent phonological rules in order to decrease recognition errors caused by pronunciation variability dependent on speakers. In the recognition experiments by the multiple-pronunciation dictionaries based on the phonological rules generated from 50 sentences, the phrase recognition errors have decreased by 2.4% on average and 3.9% at the maximum. The multiple-pronunciation dictionaries gave good performance especially for the speakers showing relatively lower recognition rates by the single-pronunciation dictionaries. The result suggests that the proposed method can play a role of speaker adaptation at the phonological level.

In the proposed method, the following should be noted. Since the phonological rules are generated from the objective speaker's continuous speech specially uttered for the rule generation, any rules are not generated from a biphone/triphone which does not exist in the continuous speech. Data sets of the continuous speech should be appropriately designed to cover objective speakers' pronunciation variability and apply the phonological rules to any recognition tasks. Another point to be noted is the design of the acoustic models. The generated phonological rules have significant relation to speaker adaptation and subword units of the acoustic models. It is intended to examine these points.

REFERENCES

[1] T. Imai, A. Ando, and E. Miyasaka, " An Automatic Caption-Superimposing System with a New Continuous Speech Recognizer," IEEE Trans. Broadcast., Vol. 40, No. 3, pp. 184-189, 1994.

[2] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," Proc. ICASSP-89, pp. 699-702, 1989.

[3] V. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT SUMMIT Speech Recognition System: a Progress Report," Proc. DARPA Speech and Natural Language Workshop, pp. 179-189, 1989.

[4] Aikawa, M. Sugiyama, and K. Shikano, "Spoken Word Recognition Based on Top-Down Phoneme Segmentation," Proc. ICASSP-85, pp.33-36, 1985.

[5] F. R. Chen, "Identification of Contextual Factors for Pronunciation Networks," Proc. ICASSP-90, pp. 753-756, 1990.

[6] P. Schmid, R. Cole, and M. Fanty, "Automatically Generated Word Pronunciations from Phoneme Classifier Output," Proc. ICASSP-93, vol. II, pp. 223-226, 1993.

[7] A. Ito, and S. Makino, "A New Word Pre-Selection Method Based on an Extended Redundant Hash Addressing for Continuous Speech Recognition," Proc. ICASSP-93, vol. II, pp.299-302, 1993.

[8] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a Large-Scale Japanese Speech Database and Its Management System," Proc. ICASSP-89, pp. 560-563, 1989.

[9] T. Imai and A. Ando, "An HMM Learning Algorithm for Minimizing an Error Function on All Training Data," J. Acoust. Soc. Jpn. (E), 13, 6, pp. 369-378, 1992.