GENERALIZED STATISTICAL MODELING OF PRONUNCIATION VARIATIONS USING VARIABLE-LENGTH PHONE CONTEXT

Yuya Akita   *Tatsuya Kawahara*

School of Informatics, Kyoto University, Kyoto 606-8501, Japan
PRESTO, Japan Science and Technology Agency

ABSTRACT

Pronunciation variation modeling is one of major issues in automatic transcription of spontaneous speech. We present statistical modeling of subword-based mapping between baseforms and surface forms using a large-scale spontaneous speech corpus (CSJ). Variation patterns of phone sequences are automatically extracted together with their contexts of up to two preceding and following phones, which are decided by their occurrence statistics. Then, we derive a set of rewrite rules with their probabilities and variable-length phone contexts. The model effectively predicts pronunciation variations depending on the phone context using a back-off scheme. Since it is based on phone sequences, the model is applicable to any lexicon to generate appropriate surface forms. The proposed method was evaluated on two transcription tasks whose domains are different from the training corpus (CSJ), and significant reduction of word error rates was achieved.

1. INTRODUCTION

Recently, the main target of automatic speech recognition (ASR) has shifted to spontaneous speech, which includes a variety of phenomena that degrade ASR. One of such phenomena is pronunciation variations, that is, multiple pronunciations are observed for a linguistically identical word. For ASR, the variations should be modeled by an acoustic model or a pronunciation lexicon. The former covers acoustic variations within a subword unit such as phoneme or syllable, while the latter covers variations that can be described with these units. The paper addresses the latter pronunciation modeling.

Design of a pronunciation lexicon is an empirical issue. Manual editing of lexicons[1], however, is extremely costly and not practical for large vocabulary ASR. Therefore, automatic generation of a lexicon is a desirable approach, which is based on prediction of possible pronunciations (surface forms) and their probabilities from orthodox ones (baseforms). Conventional studies include the knowledge-based approach such as application of phonological rules. But it is not able to assign probabilities to applied rules which are necessary to suppress false matching caused by increased entries. The data-driven approach has also been studied, for example, pattern extraction using automatic phone recognition. Most of the previous works, however, assume that the domain and lexicon of training data are same as those of the test-set.

In this paper, we present a generalized pronunciation modeling based on probabilistic mapping of phone sequences using a large-scale spontaneous speech corpus. Variation patterns of baseforms are extracted as a set of rewrite rules with their probabilities which have variable-length phone context. The model is flexibly applicable to any new lexicon, and their surface forms can be generated with appropriate probabilities. In this paper, the proposed method is tested on transcription tasks whose domains and lexicons are different from the training corpus.

2. STATISTICAL FRAMEWORK OF PRONUNCIATION MODELING

2.1. Role of pronunciation model

Statistical speech recognition is formulated as Equation (1).

$$w' = \arg\max_{w} P(x|w)P(w) \tag{1}$$

where $P(x|w)$ is an acoustic likelihood of input speech x for a word sequence w, and $P(w)$ is a linguistic likelihood of w. When multiple pronunciations of a word are considered, then the framework is extended to Equation (2).

$$w' = \arg\max_{w,p} P(x|p)P(p|w)P(w) \tag{2}$$

Here, $P(p|w)$ is a pronunciation probability of p for w. The pronunciation model should cover possible variants and give their appropriate probabilities $P(p|w)$.

2.2. Word-based pronunciation modeling

Previous works of pronunciation modeling such as [2] usually register pronunciation variants into a lexicon using automatically-derived phone sequences from the training data. In this study, we use “the Corpus of Spontaneous Japanese” (CSJ)[3] for extraction of pronunciation variations. The CSJ mainly consists of two kinds of live lectures: academic presentations and extemporaneous public speeches. Actual pronunciations of all speech materials are transcribed as well as orthographical transcriptions. So pronunciation variations observed in spontaneous speech can be extracted by matching these two kinds of transcriptions.

Pronunciation modeling using the CSJ was already addressed[4], where matching is performed word by word and pronunciation probability $P(p|w)$ is estimated for each possible pronunciation variant p of word w. Language modeling that separately handles pronunciation variants is also proposed. However, these word-based approaches are obviously limited to the vocabulary observed in the CSJ, and may not be applicable to different tasks.

2.3. Phone-based pronunciation modeling

To realize portability to other domains, phone-based modeling is considered. Pronunciation variation is described as a transformation of one phone to another. Surface forms are obtained by applying such a model to phone sequences of baseforms. As the modeling framework, decision tree[5], neural network[6] and confusion matrix[7] were proposed. Although pronunciation variations depend on preceding and following contexts, most of the methods did not consider the context or only count neighboring phones. Moreover, the methods mentioned above do not necessarily give appropriate pronunciation probabilities $P(p|w)$, because they do not estimate in the maximum likelihood manner with reliable and sufficient data.

In this paper, we adopt probabilistic rewrite rules[8] with variable-length phone context, which is a kind of statistical language model. Pronunciation variations are detected by the alignment of phonetic and orthographic transcriptions, and variation patterns including neighboring phone contexts are extracted. Furthermore, variation probabilities are derived from occurrence counts of baseforms and surface forms. Appropriate length of phone contexts is determined based on the statistics, and context back-off mechanism is introduced for robust estimation and matching of the model.

3. TRAINING AND APPLICATION OF THE PROPOSED MODEL

The proposed modeling method consists of three steps. First, patterns of pronunciation variations are detected, and necessary statistics of variation patterns and their phone contexts are estimated. Next, a set of rewrite rules are derived with appropriate contexts and probabilities. Finally, these rules are applied to baseforms to generate new pronunciation entries (surface forms).

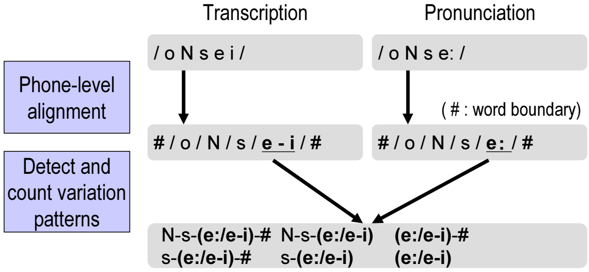

Fig. 1. Training of pronunciation variation patterns

3.1. Extraction of pronunciation variations

The training corpus (CSJ) contains 2,540 spontaneous speeches by distinct 1,362 speakers. First, morphological analysis is performed on transcriptions in order to insert word boundaries and generate baseforms. Total number of words is 6.3 million. Second, word-level alignment between baseforms and phonetic transcriptions (surface forms) is performed. Japanese words often have multiple distinct baseforms, thus the most likely baseform is also determined by the alignment process. As shown in Figure 1, if a mismatched pair of baseform and surface form (i.e., variant) is found, their phone sequences are identified. Each variation is extracted together with its preceding and following phone context, and the number of its occurrence is counted. We consider up to two phones in both directions as the phone context. Note that the word boundary is also considered as a context, because it provides useful information for pronunciation variations.

3.2. Generation of probabilistic rewrite rules

Next, probabilistic rewrite rules are generated based on the statistics of variations obtained by the previous step. Let q be a certain phone (sequence) with phone context c, and q' be a variant of q. And $C(q|c)$ and $C(q \rightarrow q'|c)$ denote occurrence counts of baseform q and surface form q' with context c, respectively. A threshold θ_1 is introduced for $C(q|c)$ to determine the adequate length of context c so that the model has reliable statistics. Namely, patterns that are more frequent than θ_1 (i.e., $C(q|c) \geq \theta_1$) are adopted as rules, and their probabilities are computed by Equation (3):

$$P(q \rightarrow q'|c) = \frac{C(q \rightarrow q'|c)}{C(q|c)} \qquad (3)$$

The contextual patterns eliminated by the threshold θ_1 are backed-off to shorter-context rules.

We use at most two phones as preceding and following contexts, respectively. Let i and j be the length of the preceding and following contexts, respectively, and R_{ij} be a set of rules whose context length is i and j. Rules are defined

Table 1. Examples of rewrite rules

Variation pattern	Context		Probability
	preceding	following	
e i → e:	#-t	r-i	0.9647
	#-t	t	0.8077
	—	r-i	0.6531
k-u → q	g-a	k-a	0.5385
	a	k	0.1818
	—	k	0.1549
a-w-a → a:	#-m	r-i	0.2770
	#-g	#	0.1408
	a-z	—	0.4286

"#" denotes word boundary.

in a descending order, from the longest context set R_{22} to a context-independent rule R_{00}. Those once adopted should be excluded from the back-off computation. For example, the adjusted frequency of variation pattern "$a\ b - q + d$" which has preceding context "$a\ b$" and following one "d" is computed by Equation (4).

$$C'(q|ab:d) = C(q|ab:d) - \sum_{\substack{(ab:dx) \\ \in R_{22}}} C(q|ab:dx) \qquad (4)$$

Rewrite rules for variation $q \rightarrow q'$ consist of context sets R_{ij} $(0 \leq i, j \leq 2)$, and individual rule entries have their own probabilities $P(q \rightarrow q'|c)$. Finally, we also introduce a threshold θ_2 for the probabilities, and rules that have larger probabilities than θ_2 (i.e., $P(q \rightarrow q'|c) \geq \theta_2$) are adopted.

We made preliminary experiments on threshold θ_1 and θ_2, and determined as θ_1=20 and θ_2=0.1. As a result, 265 kinds of variation patterns and 1,381 rules were obtained. The derived rule set includes typical cases of pronunciation variations that are phonologically predictable, for example, "/e i/ → /e:/" (diphthong to long vowel) and "/k u/ → /q/" (vanishing vowel). However, our result attaches appropriate probabilities to them. Moreover, a number of variants that are characteristic to spontaneous speech and unpredictable by the phonology are also found.

3.3. Application of variation rules

Then, new surface forms are generated by applying the set of rules to baseforms in a lexicon. Rules with longer context are applied with higher priority, and then backed-off to shorter contexts if necessary. Probabilities of resulting new pronunciation entry p' and original one p are updated as (5) and (6), respectively.

$$P(p'|w) \leftarrow P(p|w) \cdot P(q \rightarrow q'|c) \qquad (5)$$

$$P(p|w) \leftarrow P(p|w) \cdot \{1 - P(q \rightarrow q'|c)\} \qquad (6)$$

where initial probabilities of $P(p|w)$ are equal (i.e., 1 divided by the number of baseforms). The rules are applied to every possible position in the baseform, and the probability of new entry is calculated by multiplying all probabilities of the applied rules. However, a new entry is discarded if its probability is smaller than a threshold θ_2.

3.4. Use of pronunciation model weight in decoding

Conventionally, decoding in ASR is performed by computing the log likelihood given by (7):

$$\log P(x|p) + w_l \{\log P(w) + \log P(p|w)\} \qquad (7)$$

where w_l is a language model weight. Since dynamic ranges of $\log P(w)$ and $\log P(p|w)$ are different, the pronunciation probability may have little effect. Therefore, we introduce pronunciation model weight w_p. In this case, the log likelihood is formulated as (8):

$$\log P(w|p) + w_l \log P(w) + w_p \log P(p|w) \qquad (8)$$

4. EXPERIMENTAL EVALUATION

4.1. Experimental setup

The proposed method is evaluated on transcription of spontaneous speeches that are different from the training corpus (CSJ). We used two kinds of discussions as test-sets: televised panel discussions ("Sunday discussion") and meetings in the National Diet (Congress) of Japan ("Diet"). "Sunday discussion" is a one-hour session in which five to eight persons take part, and ten sessions are compiled as a test-set. "Diet" data is chosen from a single session whose duration is about 5.5 hours, and the number of speakers is 23.

The language model is a mixture of topic-oriented and spontaneity-oriented models trained with the minutes of the National Diet and the CSJ, respectively. The acoustic model is a triphone HMM trained with the CSJ, and adapted to individual speakers by the unsupervised MLLR method. As the decoder, our Julius rev.3.4.2 is used. The vocabulary for ASR is determined as a union of those of the two language models. Its size is 29,720, and the baseline pronunciation lexicon has 31,571 baseforms. Then, the proposed method was applied to generate surface form entries with their occurrence probabilities. The size of the new lexicon is 38,207.

For comparison, word-based pronunciation modeling[4] which was described in Section 2.2 was conducted. The pronunciation variation was extracted for each lexical entry of the CSJ using the alignment between baseforms and surface forms. Then, those entries were added to the baseline lexicon with their probabilities. It has coverage of 57.8% of the baseform entries of the test vocabulary. For the remaining words, only baseforms are retained. The resulting new

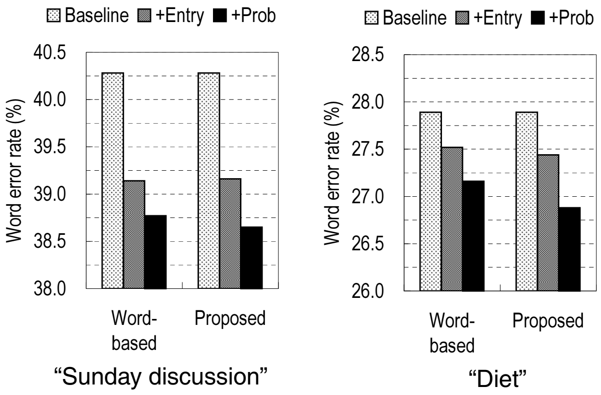

Fig. 2. Comparison of word error rates between word-based modeling and proposed method

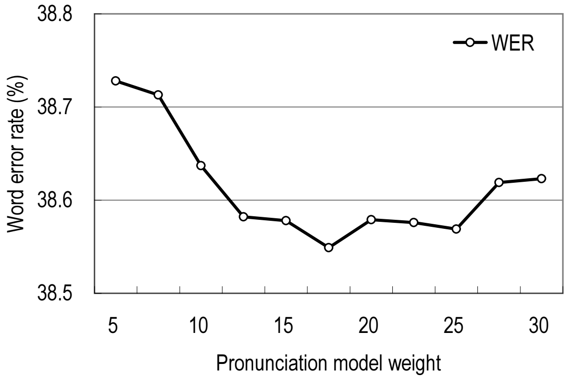

Fig. 3. Effect of pronunciation model weight on word error rate

lexicon has 33,508 entries. In summary, it has 4,227 new surface forms, while the proposed method generated 8,271. Note that 2,290 and 1,635 baseforms are eliminated by the same threshold θ_2 of word-based and the proposed method, respectively.

4.2. Experimental results

Figure 2 compares word error rates (WER) by the word-based and the proposed methods for the two test-sets. By "+Entry", new surface form entries were added without probabilities, while probabilities were also used by "+Prob." WER was reduced by addition of the surface forms, and further improvement was obtained by introducing the pronunciation probabilities. In all cases, the proposed method achieved lower WER than the word-based method. While comparable accuracy was obtained for "Sunday discussion", the proposed method outperformed the word-based model in "Diet" data. Relative improvements by the proposed method over the baseline are 4.0% and 3.6% for "Sunday discussion" and "Diet", respectively, and these improvements are statistically significant.

Then, the effect of the pronunciation model weight w_p is investigated. The ASR experiment was made on "Sunday discussion" by changing the value of w_p in Function (8). The language model weight w_l is 7.0. Figure 3 shows average WER. WER was improved when the pronunciation model weight w_p was two or three times larger than the language model weight w_l. The result shows that weighting the pronunciation model has actual effect.

5. CONCLUSIONS

We have presented a method of statistical pronunciation modeling applicable for any vocabulary. Pronunciation variations between baseform and surface form are extracted from a large-scale spontaneous speech corpus (CSJ), and phone context dependent variation patterns and their occurrence probabilities are trained. Probabilistic rewrite rules with variable-length phone context are then constructed based on this statistics. Since the probabilistic model is generalized, it can be applied to any lexicon of new domains to generate appropriate surface forms with their probabilities, which match the framework of statistical speech recognition.

6. REFERENCES

[1] L. Lamel and G. Adda, "On designing pronunciation lexicons for large vocabulary, continuous speech recognition," in *Proc. ICSLP*, 1996, vol. 1, pp. 6–9.

[2] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP*, 1996, vol. 4, pp. 2328–2331.

[3] S. Furui, K. Maekawa, and H. Isahara, "Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –," in *Proc. ICSLP*, 2000, vol. 3, pp. 518–521.

[4] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech & Audio Process.*, vol. 12, no. 4, pp. 391–400, 2004.

[5] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[6] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciations based on neural networks," *Speech Communication*, vol. 27, pp. 63–73, 1999.

[7] D. Torre, L. Villarrubia, J.M. Elvira, and L. Hernandez-Gomez, "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," in *Proc. ICASSP*, 1997, vol. 2, pp. 1463–1466.

[8] Q. Yang, J.-P. Martens, P.-J. Ghesquiere, and D.V. Compernolle, "Pronunciation variation modeling for ASR: Large improvements are possible but small ones are likely to achieve," in *Proc. PMLA*, 2002, pp. 123–128.