PRONUNCIATION VARIATION SPEECH RECOGNITION WITHOUT DICTIONARY MODIFICATION ON SPARSE DATABASE

Supphanat Kanokphara, Virongrong Tesprasit, Rachod Thongprasirt

Information R&D Division,
National Electronics and Computer Technology Center (NECTEC)
112 Paholyothin Rd., Klong 1, Klong Luang, Pathumthani 12120 Thailand
supphanat_k@notes.nectec.or.th, {virong, rachod)@nectec.or.th

ABSTRACT

Generally, a speech recognition system uses a fixed set of pronunciations according to the dictionary for training and decoding. However, even a well-defined lexicon cannot be used to support all variations in human's pronunciation. Besides, in order to cover all possible pronunciations, the size of the dictionary would be too large to implement. Sharing gaussian densities across phonetic models and decision tree for pronunciation variation are proved to be efficient for pronunciation variation system without dictionary modification. This paper presents the alternative methods that can be used even in the sparse database situation. Re-label training is modified to have rule-based pronunciation variation in order to obtain real phonetic acoustic models. Phonemic acoustic models are then retrained from the tying HMM states across phonetic models. These new phonemic models allow alternative search path during recognition. The system shows better performance in the experiment.

1. INTRODUCTION

Sharing gaussian densities across phonetic models and decision tree [1] have successfully shown the great improvement in pronunciation variation speech recognition without dictionary modification. That paper starts from showing that training acoustic model from phonetic transcriptions is better than training from phonemic transcriptions. Then the various phonemic-to-phonetic transcriptions techniques are proposed. The phonetic transcriptions generated from hand-labeled-trained acoustic models gave the best result in the experiment. Finally, the pronunciation variability at the level of HMM states are explained and the experimental result was observed comparing with the pronunciation variation system requiring dictionary modification [2].

By using tree-base pronunciation variation, the system required a large corpus for training in order to cover all variation, as the weak point of all corpus-based pronunciation system is unable to observe variation beyond the corpus. This problem becomes serious for a sparse database like Thai. As the result, hybrid of corpus- and knowledge-based models [3] is used instead in this paper. The hybrid method varies pronunciations according to the rule from linguistic knowledge and observed from the corpus. This rule-based model also has the advantage of reducing in time-consumption, as there is no need to calculate possible alternate pronunciations as rules are already fixed. The tying HMM states across phonetic models are used instead of sharing HMM states across phonetic models in this paper.

This paper is organized as follows. The next section describes all the Thai pronunciation variation rules used in this paper. Then, the training strategy is described in Sections 3 and 4. Section 3 contains the overview of the training system, while Section 4 emphasizes on rule-based pronunciation variation algorithm. Section 5 explains how phonemic models are constructed for decoding. The experimental results and conclusion are described in Sections 6 and 7, respectively.

2. PRONUNCIATION VARIATION RULES

Table 1 [4] demonstrates all 76 phonemes used in this paper. "sp" and "sil" are short pause and silence symbols, respectively. A double character means long vowel such as /@@/ is long vowel version of /@/. Some vowels are not included in Table 1 because they have fewer occurrences in Thai speech such as /ia/, /ua/, etc. A character with "^" symbol indicates the final consonant. A character combined with "h" is the aspirated version of that sound such as /kh/ is the aspirated version of /k/. Character with /w/, /r/ and /l/ are called cluster /w/, /r/ and /l/, respectively (cluster is pronounced two phonemes together).

All of the rules in this paper are from linguistic knowledge [5] and observed from the corpus by our linguistic expert. There are 4 rules as follows:

Initial consonants	/k, kh, ng, c, ch, s, j, d, t, th, n, b, p, ph, f, m, r, l, w, h, z/
Cluster consonants	/pr, pl, tr, kr, kl, kw, phr, phl, thr, khr, khl, khw, br, bl, fr, fl, dr/
Final consonants	/k^, ng^, j^, t^, n^, p^, m^, w^, z^,ch^, f^, l^, s^, jf^, ks/
Vowels	/a, aa, i, ii, v, vv, u, uu, e, ee, x, xx, o, oo, @, @@, q, qq, iia, vva, uua/
Special symbols	sil, sp

Table 1: Phonemes for Thai words in this paper.

(a) "sp" insertion

Thai language does not have punctuation marker to pause within a sentence, a short pause can occur anywhere after syllable. As a result, the short pause is selectively inserted at the end of each syllable in each word. In Thai language, the end of the syllable can be either vowel or final consonant. Additionally, the beginning of the syllable must be initial consonant or cluster consonant.

(b) /r/ sound ←→ /l/ sound (nonstandard pronunciation)

/r/ sound is difficult to pronounce in the real speech. For convenience, sometimes /r/ sound is pronounced as /l/ sound. Contrarily, some over-accented Thai speakers would produce /l/ sound as /r/ sound. The phonemes following this rule are listed below.

- /pr/ ←→ /pl/ ←→ /p/
- /tr/ ←→ /t/
- /kr/ ←→ /kl/ ←→ /k/
- /phr/ ←→ /phl/ ←→ /ph/
- /thr/ ←→ /thl/ ←→ /th/
- /khr/ ←→ /khl/ ←→ /kh/
- /br/ ←→ /bl/ ←→ /b/
- /fr/ ←→ /fl/ ←→ /f/
- /dr/ ←→ /d/
- /r/ ←→ /l/

(c) Loan word error

Some pronunciation of loan words is hard to pronounce in Thai. Some speakers pronounce those words in English accent while some pronounce in Thai accent. The phonemes following this rule are listed below.

- /s/ ←→ /ch/
- /l^/ ←→ /n^/ ←→ /w^/
- /s^/ ←→ /t^/
- /f^/ ←→ /p^/
- /ch^/ ←→ /t^/
- /t/ ←→ /th/
- /p/ ←→ /ph/
- /k/ ←→ /kh/

(d) "Short vowel" ←→ "long vowel"

In conversation, a fast speaking rate would shorten some Thai vowels. In the same way, a slow speaking rate would lengthen a vowel. The phonemes following this rule are listed below.

- /i/ ←→ /ii/
- /e/ ←→ /ee/
- /a/ ←→ /aa/
- /@/ ←→ /@@/
- /x/ ←→ /xx/

3. RELABEL TRAINING WITH PRONUNCAITION VARIATION

In order to achieve high accuracy acoustic model, speech data should be correctly marked. Nevertheless, for many reasons, the transcriptions are not perfectly marked. Traditionally, transcriptions are generated from automatic segmentation and rechecked by human. However, in a large database, manual-checking process is time-consuming and usually ignored. The re-label training strategy [6] is designed to update labels during the training so that high accuracy model can be obtained and even initialized by automatically-generated transcriptions. In addition, some ambiguities among the phonemes are hard to classify by human.

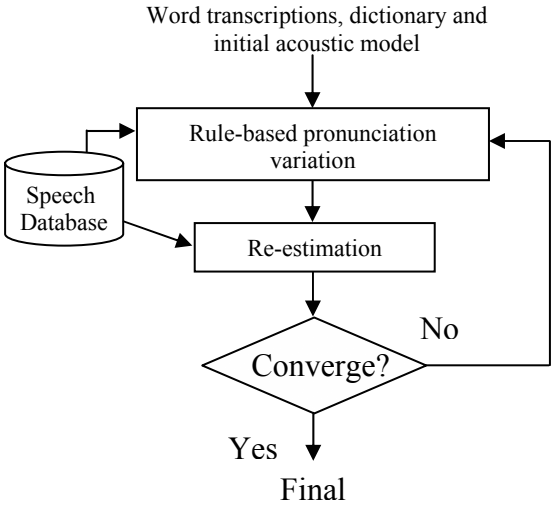

Figure 1: Re-label training process.

The training strategy starts from word transcriptions, dictionary, initial acoustic models and speech database as inputs of rule-based pronunciation variation. These acoustic models are trained from initial phonemic transcriptions and speech database with the re-estimation algorithm. The rule-based pronunciation variation then generates phonetic transcriptions according to the speech data. These phonetic transcriptions and speech data are then the inputs for re-estimation. After that the re-estimation process updates the acoustic models to be the inputs for rule-based pronunciation variation. This process is continued until the log probability of updated models is less than the last one. Fig. 1 shows the re-label training with pronunciation variation.

4. RULE-BASED PRONUNCIATION VARIATION

This section clarifies the rule-based pronunciation variation, one of the processes from Section 3. This process uses word transcriptions, dictionary, acoustic model and speech database as inputs. In fact, this process is just the force recognition with the phonetic network representing possible variation according to the rules. For example, the sentence "rak^ kun^" (=love you) can be constructed as shown in Fig. 2. According to the rule (b) in Section 2, /r/ can be altered to /l/. Rule (d), short vowel /a/ can be changed to long vowel /aa/. Rule (a), end of syllable /k^/ can be followed by "sp". The phoneme /k/ is influenced by two rules, (b) and (c). The alternative paths are the combination of two rules as shown in Fig. 2. There are 4 possible paths in the picture. /u/ remains the same, as there is no matched rule. Finally, /n^/ with the combination of rules (a) and (c) have 6 possible paths in the network. The most likely selected paths are the phonetic transcriptions using for next training

We believe that the phonemic transcriptions affect the acoustic models' accuracy as well as the phonetic transcriptions. Therefore, in order to obtain real optimum phonetic transcriptions, various methods are investigated.

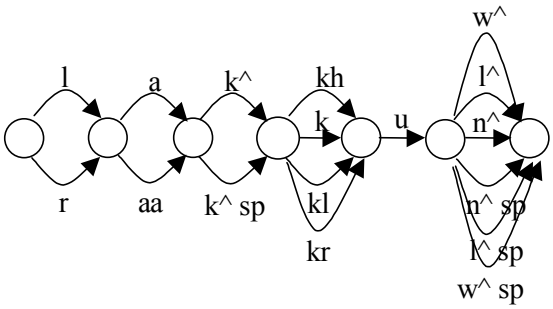

Figure 2: Phone-level network for "rak^ kun^" according to the pronunciation variation rules.

There are three ways to generate phonemic transcriptions: automatically by using Thai Grapheme-to-Phoneme (G2P) developed by NECTEC [7] (I), manually by expert labelers (II), and from re-label training (III). Also three types of training process are used in this experiment; training without re-label re-estimation (IV), training with re-label re-estimation (V), training with pronunciation variation (VI). In (II) phonemic transcriptions are generated from G2P and edited by our expert labelers. The transcriptions were examined by using Wavesurfer 1.0.4 [8]. There are only 2 expert labelers for the correction process in order to preserve the consistency. Complicated points in transcriptions and boundaries alignment are discussed and adopted during the process. Note that the various phonemic transcriptions are also investigated. Therefore, the log probability of training process is used to measure the quality of phonetic transcriptions, as there is no appropriate reference answer for phone error rate measurement.

5. PRONUNCIATION VARIATION MODEL

The models trained by rule-based pronunciation variation re-label training can degrade the system as the acoustic models are trained to be phonetic models while the dictionary is still in phonemic form. Of course, it is also impossible to construct such a gigantic-size phonetic-form lexicon. The rule-based pronunciation variation also cannot be used at this point because there is no given phoneme network in the decoding process. To construct a phonemic model, we tie the start and end states of every phonetic model in the same variation group together according to the rules presented in Section 2. The transition probabilities from the first state to each individual phonetic model are starting with the same value in order to allow fair pronunciation variation. These phonemic model prototypes are then retrained by phonemic transcriptions to obtain maximum likelihood phonemic models. Therefore, each phonemic model composes of general phonetic model and phonetic models that varied from the phonemic model according to the rules. With this model, the best matching path can be obtained even when no variation presents in the dictionary. For example, phonemes /l/ and /r/ can be varied according to the rule in Section 2. The prototype of phonemic model /l/ or /r/ is shown in Fig. 3.

There are two reasons that sharing HMM states phonetic model is not employed in this experiment. First, the database is not enough to train multi-gaussian and context-dependent model. Second, sharing HMM states and tying HMM states phonetic model showed almost the same WER in [1] if there is no future improvement (merged or merged, further training).

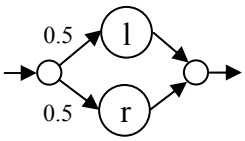

Figure 3: Prototype for "l" or "r" pronunciation variation model.

The merit of tying HMM states phonetic model is that it can be expanded to appropriate number of guassian system easier than train multi-guassian system and reduce the number of guassian after the problem of data sparseness is found after sharing the gaussian.

Some might say that these phonemic models may result in word confusions, for instance, the word “r a k^” (=love) can be confused with the word “l a k^” (=steal). We don’t say that this system can solve this problem perfectly but it is better than the traditional one. Let’s assume that “r a k^” is the correct word and the influence from language model is ignored. In the traditional system, “l a k^” can be chosen if phoneme “r” is pronounced as phoneme “l”. In our system, phoneme “r” is the combination of phonetic “r” and phonetic “r pronounced as l”. Therefore, there is more chance that phoneme “r” can win if phonetic “r pronounced as l” in phoneme “r” is better than phonetic “l” in phoneme “l”.

6. EXPERIMENT

HTK Toolkit [9] is used as the base system for this experiment. The experiment procedure starts from data preparation, wave to MFCC conversion, making topology prototype, label and dictionary construction (in HTK format), training acoustic and language models, and testing finally. In the decoding process, a back-off bi-gram language model is constructed and Viterbi algorithm is applied for speech recognition process.

6.1. Database

As Thai corpus project is just starting, our corpus is still relative small comparing with the other language corpus. In 3,097 words database, 1,246 utterances are used as a training set. 140 utterances having less error in language model are selected as a testing set. As this experiment aims at improving of acoustic model, we designed the experiment to have less effect from language model error. This can be done by selecting the most-occurrence-words sentences as a test set. The algorithm of selecting test sentences is somewhat similar to [10]. A female professional speaker is set to record all speech utterances in order to avoid any error occurring from speaker.

The language model is constructed from 1,246 sentences according to the utterances. Back-off bi-gram’s perplexity is 73.68 and entropy is 6.20. Dictionary is generated from G2P.

Speech utterances (16 kHz sampling frequency with 16 bits quantization) are parameterized into 12 dimensional vectors, energy, and their delta and acceleration (39 length front-end parameters).

6.2. Results

There are many training types in this experiment according to initial phonemic transcriptions and training strategies. As

Training type	Training log probability	% Correct	% Accuracy
I + IV	-59.58	70.36	67.87
II + IV	-58.63	78.52	73.63
III + IV	-58.73	74.01	71.56

Table 2: Training without re-label re-estimation.

Training type	Training log probability	% Correct	% Accuracy
I + V	-58.73	77.87	72.77
II + V	-58.63	78.52	73.63
III + V	-58.71	78.11	72.91

Table 3: Training with re-label re-estimation.

Training type	Training log probability	% Correct	% Accuracy
I + VI	-58.38	77.66	72.46
II + VI	-57.60	79.42	74.11
III + VI	-57.68	80.46	75.42

Table 4: Training with pronunciation variation.

mentioned in Section 4, three types of initial phonemic transcriptions are listed as I, II and III, and three types of training strategy are as IV, V and VI. For example, in Table 1, I + IV means training without re-label re-estimation and the system is initialized by phonemic transcriptions generated from G2P. Training log probability tells us the quality of phonetic transcriptions while percentage correct and accuracy informs us how good the system can recognize word sequences.

In Table 2, training by using manual phonemic transcriptions result in highest score in both training log probability and percentage correct and accuracy. These show that phonemic transcriptions edited by our labelers are good in quality. They also reveal that the phonemic transcriptions generated from our re-label training system give better result than the one from G2P.

Table 3 shows the results from training with re-label re-estimation. The system trained by manual phonemic transcriptions (II) is the same as in Table 2. This is because it is already satisfied and need no re-label in the maximum likelihood sense. The effect of re-label training can be clearly seen from the training initialized by G2P phonemic transcriptions. The accuracy is increased by 4.9%.

The results of pronunciation variation approach to the system are demonstrated in Table 4. Surprisingly, (I+V) has better recognition rate than (I+VI) but is worse in log probability. This can be concluded that even phonetic transcriptions are better, worse phonemic transcriptions can also degrade the system in pronunciation variation system. The higher percentage correct of (III+VI) than (II+VI) illustrates that the accuracy of phonemic transcriptions generated automatically are better than the manual one. This is because the automatic system sometimes can solve the ambiguous phonemes that human cannot solve.

7. CONCLUSION

In this paper, we have proposed an efficient way of pronunciation variation approach to the speech recognition. Rule-base pronunciation variation and phonemic models are used for training and decoding, respectively. Various techniques to find best phonemic and phonetic transcriptions have been investigated. The experimental results demonstrate that the accuracy of both phonemic and phonetic transcriptions greatly affect the accuracy of the system. Pronunciation variation system initialized from phonemic transcriptions generated from re-label training shows the best performance in the experiment.

More rules pronunciation variation, speaker-independent, multi-gaussian and context-dependent system will be experienced in the future work if we can obtain a larger corpus.

8. ACKNOWLEDGMENTS

Thank you to Treepop Sunpetchnivom for the motivation of pronunciation variation in speech.

9. REFERENCES

[1] M. Saraclar, H. Nock, S. Khudanpur, “Pronunciation modeling by sharing Gaussian densities across phonetic models”, *Computer Speech and Language*, (14): 137-160, 2000.

[2] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, G. Zavaliagkos, “Stochastic pronunciation modeling from hand-labelled phonetic corpora”, *Speech Communication* 29: 209-224 (1999).

[3] H. Nakajima, Y. Sagisaka, H. Yamamoto, “Pronunciation Variants Description using Recognition error Modeling with Phonetic Derivation Hypotheses", *Proc. ICSLP2000*, (3): 1093-1096, 2000.

[4] Sirivisoot, S., *Variation of Final (l) in English Loanwords in Thai According to Style and Educational Background*, Master Thesis, Department of Linguistics, Chulalongkorn University, 1994. (in Thai).

[5] Khanitthanan, W., *Phasa lae Phasasart*, Thammasat University Press, 1990. (in Thai).

[6] P. Tarsaku, S. Kanokphara, “A Study of HMM-based automatic segmentations for Thai Continuous Speech Recognition System”, *Proc. SNLP2002*.

[7] P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, “Thai Grapheme-to-Phoneme using Probabilistic GLR Parser”, *Proc. Eurospeech*, (2): 1057-1060, 2001.

[8] K. Sjölander, J. Beskow, “Wavesurfer – An Open Source Speech Tool”, *Proc. ICSLP*, (4):, 464 - 467, 2000.

[9] Young, S., Jansen, J., Odell, J., Ollasen, D., Woodland, P., 1995. *The HTK Book* (Version 3.0), Entropic Cambridge Research Laboratory, Cambridge, England.

[10] J. Shen, H. Wang, R. Lyu and L. Lee, “Automatic Selection of Phonetically Distributed Sentence sets for Speaker Adaptation with Application to Large Vocabulary Mandarin Speech Recognition”, *Computer Speech and Language*, (13): 79-98, 1999.