# Pronunciation Variation Speech Recognition Without New Dictionary Construction

*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, 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 dictionary 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. This paper presents efficient strategies for both training and decoding of a continuous speech recognition system: tree of knowledge-based pronunciation variations re-label training and state-level pronunciation variation model, respectively. These strategies can efficiently support the variations in pronunciation according to the rules without necessity to make pronunciation variation dictionary. The pronunciation variation training is modified from the re-label training to obtain the maximum likelihood pronunciation during training in order to reduce the error in an acoustic model. Although the database and rules used in this paper is Thai, this system can also be adapted to other languages easily as the variations are controlled by simple rules. The system shows better performance in the experiment.**

**Keyword** - Speech recognition, pronunciation variation, Hidden Markov Model (HMM)

## 1. Introduction

Hidden Markov Model (HMM) takes parts deeply in the field of speech technologies. This is because the properties of HMM which can normalize speech signal's time variation and represent the speech signal statistically. In the HMM training process, the Baum-Welch method [1] is unavoidably used to optimize the acoustic model. In the training process, an acoustic model is built from speech data and transcriptions according to the given speech. By this way, the acoustic model should be accurate if the speech data are correctly transcribed. However, the accuracy of acoustic model is usually limited by the precision of the pronunciation in the dictionary. As a result, many researchers [2][3][4] tried to construct a dictionary to contain all possible pronunciation variations automatically. By doing this, the variations of pronunciation are also constrained by the size of dictionary and the corpus used to construct the dictionary.

This variation problem becomes acute in the language having wide variation in speaking like Thai language. One of the problems of Thai language is there are too many pronunciations similar to each other such as /l/ and /r/. Therefore, in the real conversation, /l/ and /r/ can be pronounced incorrectly.

Our approach is similar to [4]. The differences are (i) there is no pronunciation variation dictionary constructed, (ii) each Tree is applied to each word, not each phoneme, (iii) instead of yes-no question; the question is "What is the most probably phone for this phoneme?" while each node represents phoneme next to its parent. The best pronunciation is chosen from the last node when there is no next phoneme left in the word's pronunciation.

By training acoustic model with this strategy, the model becomes weaker for the variation of testing data if there is no appropriate variation in the dictionary. Linking all possible variation phonemes together and retraining allows the real pronunciation to be searched during the decoding process. This pronunciation variation model can give the real variant during decoding without dictionary construction. These combination strategies greatly reduce a lot of time-consumption because it can vary the pronunciation while training and testing.

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. The experimental results and conclusion are described in Sections 4 and 5, respectively.

## 2. Pronunciation Variation Rules

Table 1 [10] 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 a longer 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).

There are 4 rules in this paper as follows:

(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 an initial consonant or a 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 [11]
Some pronunciations of loan words are 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/

*Table 1. Phonemes for Thai words in this paper.*

| | |
|---|---|
| **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 |

## 3. Pronunciation Variation Model

In order to achieve the 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 transcriptions during the training so that the high accuracy model can be obtained without manual process. This training strategy increases model accuracy by correcting the transcription according to the pronunciation list in the dictionary.

In this paper, this system is extended to cover large phoneme variations unable to be listed in the dictionary. As mention before, the real speech is varied according to many factors, such as noise, speaking style, etc. These variations are too large to be supported by any dictionary. Fig. 1 shows the tree-based pronunciation variation re-label training flowchart. The system starts from word transcriptions, dictionary, initial acoustic model and speech database as inputs. This acoustic model is trained from initial phoneme transcriptions and speech database with the re-estimation algorithm. The tree-based pronunciation variation then generates phoneme transcription according to the inputs and pronunciation variation rules. This phoneme transcription and speech data are then the inputs for re-estimation. After that, the re-estimation process updates the acoustic model to be the input for tree-based pronunciation variation. This process is continued until the log probability of update model is less than the last one.

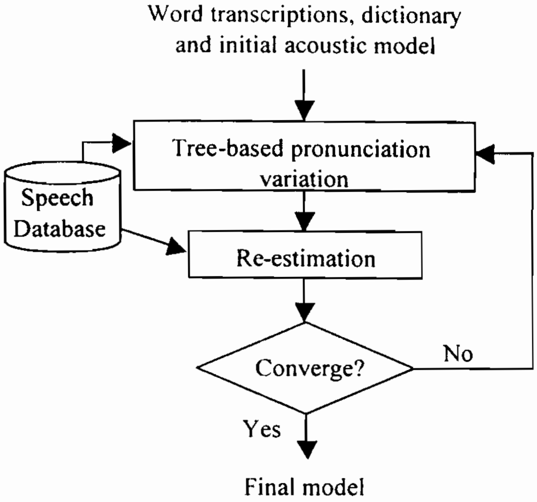

Word transcriptions, dictionary and initial acoustic model
Tree-based pronunciation variation
Speech Database
Re-estimation
Converge?
No
Yes
Final model

*Figure 1. Illustrates Tree-based pronunciation variation re-label training process .*

### 3.1. Tree-Based Pronunciation Variation

This section clarifies the tree-based pronunciation variation, one of the processes mentioned above. This process uses word transcriptions, dictionary, acoustic model and speech database as inputs. The process starts from placing the first phoneme of each word of each transcription at the root node of the variation decision tree. Then, as the tree path goes down, the phoneme in the question is the phoneme next to the last phoneme from its parent node. This iteration continues until there is no phoneme next to the phoneme in the question. For example, the word "rak^" (=love) in Fig. 2, the question at the root node is "What is the possible variation of /r/ following by /a/ of the word?" The candidate for this question is /l/ according to the rule (b) in Section 2. The second level node is associated with the question "What is the possible variation of /a/ preceding by /l/ and following by /k/?" The possible variant is /aa/. The question is continued until there is no next phoneme left. As this is a three-phoneme word, there are only four tree levels in this example. The darker line shows the best tree path and the leaf node is the best pronunciation for this word which is "lak^" as shown in Fig. 2. The question of child node depends on the answer from the parent node because some rules can be applied only at the specific preceding and following phonemes such as "sp" can be inserted after vowel if the following phoneme is an initial or cluster consonants, not a final consonant.

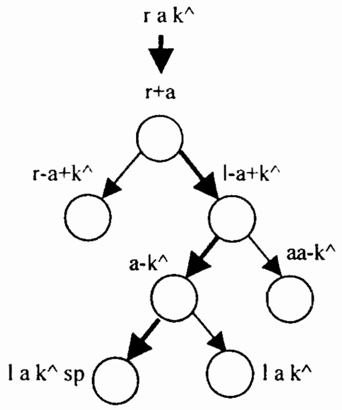


*Figure 2. Illustrates Variation decision tree of the word "rak^".*

The question at each node is answered by using Viterbi algorithm giving the acoustic model. By this way, the best pronunciation of each word in each transcription is chosen according to the speech database in the maximum likelihood sense. In common sense, searching by using tree eliminates unnecessary path in the process comparing with searching on every possible pronunciation in the pronunciation lists. In practice, this reduces the number of possible candidates without any degradation in performance.

### 3.2. Pronunciation Variation Model

By using pronunciation variation training, the model is trained from the real transcriptions. For example, the word "rak" can be trained as the word "lak" if the speaker really read "lak" in the recording process. With this training strategy, the model "l" or "r" becomes the real "l" or "r", respectively. Consequence, these models cannot well recognize "l" pronounced as "r" or "r" pronounced as "l" in the testing data. Of course, it is impossible to construct such a gigantic-size dictionary covering all possible pronunciations of every word. The tree-based pronunciation variation also cannot be used at this point because there is no given word in the decoding process. The problem can be solved by allowing the alternative path of pronunciation in each model during recognition. This can be done by tying the start and end states of the all models in the same variation group according to the rules presented in Section 2. The transitions from the first state to each individual sub model are all the same in order to allow the fair pronunciation variation. These prototype models are all retrained to obtain maximum likelihood models. 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 pronunciation variation model /l/ or /r/ is shown in Fig. 3. The transitions to /r/ and /l/ are equally divided to be 0.5.

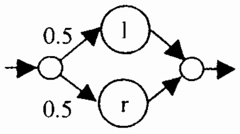


*Figure 3. Illustrates Prototype for "l" or "r" pronunciation variation model*.

## 4. Experiment

HTK Toolkit [5] is used as the base system for this experiment. The experiment procedure starts from data preparation, wave to MFCC conversion, 5-state 1-mixture left-right model prototype building, transcription labeling and dictionary construction (in HTK format), model training, and testing finally.

There are two points concerned in this experiment, i.e. effect of initial phoneme transcriptions and training strategy. There are three types of initial phoneme transcriptions: the transcriptions generated automatically by using Thai Grapheme-to-Phoneme (G2P) developed by NECTEC [7] (I), the transcriptions edited by expert labelers (II), and the transcriptions generated from re-label training processes (III). Training strategies are: training without re-label re-estimation (IV), training with re-label re-estimation (V), training with pronunciation variation (VI).

A back-off bi-gram language model is constructed as the speech data in this experiment is continuous and Viterbi algorithm is applied for speech recognition process.

### 4.1. Database

In 3,097 words database, 1,246 continuous read speech utterances are used as a training set and 140 continuous read speech utterances having less error in language model are selected as a testing set. Each utterance has approximately 10 words. As this experiment aims at improving of acoustic model, we designed the experiment to have less effect from language model error, such as out-of-vocabulary problem, insufficient number of tri-gram for training, etc. 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 [8]. A female professional speaker is set to record all speech utterances in order to avoid any error occurring from speaker's specific characteristic.

The manual phoneme transcriptions are generated from G2P and edited by our expert labelers. The transcriptions were examined by using Wavesurfer 1.0.4 [9]. There are only 2 expert labelers for the correction process in order to preserve the consistency. Complicated points in transcription and boundary alignment are discussed and adopted during the process.

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.

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

| Training type | Training log probability | % Correction | % 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 3. Training with re-label re-estimation.*

| Training type | Training log probability | % Correction | % 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 4. Training with pronunciation variation.*

| Training type | Training log probability | % Correction | % 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 |

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).

### 4.2. Results

There are many training types in this experiment according to initial phoneme transcriptions and training strategies. As mentioned above, three types of initial phoneme transcriptions are listed as I, II and III, and three types of training strategy are as IV, V and VI. For example, in Table 2, I + IV means training without re-label re-estimation, and the system is initialized by automatic generated phoneme transcriptions. Training log probability tells us how the acoustic model is close to the training data. Percentage of accuracy and correction declares how the recognition result matches the testing data. Therefore, observing all three values gives both the effect of the system for training and testing data.

In Table 2, training by using manual phoneme transcriptions is the best method. This shows that phoneme transcriptions edited by our labeller are good in quality. It also shows that the phoneme transcriptions generated from our re-label training system give better result than the one from G2P. Moreover, the accuracy is only 0.59% less than the system initialized by manual phoneme transcriptions.

Table 3 shows the results from training with re-label re-estimation. The systems trained by manual and re-label phoneme transcription (II and III) are the same as in Table 2. This is because they are already satified and need no re-label in the maximum likelihood sense. The effect of re-label training can be seen from the training initialized by G2P phoneme transcriptions. The accuracy is increased by 3.31%.

The result of pronunciation variation approach to the system is demonstrated in Table 4. The result from this Table shows the superior result to the Table 2 and 3. The higher percentage correction of III than II also illustrates that the efficiency of transcription generated automatically can reach the level of the manual one if our pronunciation variation processes are applied.

## 5. Conclusion

In this paper, we have proposed an efficient way of pronunciation variation approach to the speech recognition. Tree-base pronunciation variation and pronunciation variation model are used for training and decoding, respectively. The merits of these techniques are: (1) they can be implemented easily and can be applied to any languages. (2) Tree base search is faster comparing with the direct pronunciation search. (3) The system results in better performance.

There is no tonal experiment performed in this paper. The tonal experiment will be done in the future work.

## 6. References

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Appplications in Speech Recognition", *Proc. IEEE*, 77 (2): 257-286, 1989.

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

[3] J. Jeon, S. Cha, M. Chung, J. Park, K. Hwang, "Automatic Generation of Korean Pronunciation Variants by Multistage Applications of Phonological Rules", *Proc. ICSLP 98.*

[4] 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).

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

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

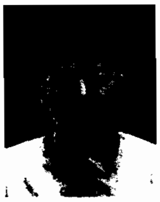
[7] P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, "Thai Grapheme-to-Phoneme using Probabilistic GLR Parser", *Proc. Eurospeech*, (2): 1057-1060, 2001.

[8] 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.

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

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

[11] 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)



**Supphanat Kanokphara** was born in Bangkok, Thailand on May 30, 1975. He received B.S. degree in electrical engineering from Sirindhorn Institute of Technology, Thammasat University, Thailand (SIIT) in 1994 and M.S. degree in computer and electrical engineering from Waseda University, Japan in 2000. During the summer and spring of 2000, he was a training student at Advanced Telecommunications Research Institute International, Japan (ATR). In 2001, he joined NECTEC as a research assistant. His research interests are speech recognition and related technologies.

**Virongrong Tesprasit** has joined NECTEC since April 1996 after receiving her BA. (Linguistics) degree from Thammasat University. She joined both Royal Institute Dictionary Development Network Project and Development of Thai Corpus Base Project. At present, she researches on Thai Text-to-Speech synthesis Project. Her research interests are Phonetics and Speech Technology.