

# MODELLING PRONUNCIATION VARIATIONS IN SPONTANEOUS MANDARIN SPEECH

LIU Yi and Pascale FUNG  
Human Language Technology Center  
Department of Electrical and Electronic Engineering  
University of Science and Technology, Hong Kong  
{eelyx, pascale}@ee.ust.hk

## ABSTRACT

Pronunciation in spontaneous Mandarin speech tends to be much more variable than in read speech. In current recognition systems, pronunciation dictionaries usually only contain one standard pronunciation for each word, so that the amount of variability that can be modelled is very limited. Most recent research work for modelling variations in spontaneous speech focuses on the lexicon level, which can only solve intra-word variations. Inter-word variations cannot be modelled effectively. Chinese is monosyllabic and has simple syllable structure, giving rise to a high amount of pronunciation variations. In this paper, we propose two methods to model pronunciation variations in spontaneous Mandarin speech. First, we generate probability lexicon to model *intra-syllable* variations by using DP alignment algorithm between base form and surface strings. Second, we integrate variation probability into the decoder to model intra as well as *inter-syllable* variations. Experimental results show that modelling intra-syllable variation with a probability lexicon reduces syllable error rate by 0.85% (phone error rate reduction of 1.4%) while adding inter-syllable variation in addition reduces syllable error rate significantly by 4.76% (phone error rate reduction of 7.6%) compared to the baseline system.

## 1. INTRODUCTION

In large vocabulary speech recognition system, word, character or syllable is represented as a sequence of its constituent phonemes as defined by a hand-generated lexicon, or as several alternative sequences if multiple pronunciations of the unit are allowed. For each word, if there are several acceptable pronunciation variants, they are considered as *intra-word variants*. Coarticulation between words or syllables can bring more pronunciation variation. Such variation give rise to *inter-word variants*. A major factor in the performance of a speech recognizer is the consistency

between the way the speaker pronounces the words and the way that the pronunciation is specified in the dictionary. When there is a mismatch, recognition errors occur. There is a high variability in pronunciation in spontaneous and conversational speech. Users tend to speak more sloppily, which causes low recognition accuracy.

In the past few years, many researchers have been done to make the recognizers efficient independently from users and speaking style. However, each user has its own style to pronounce due to accent, speaking rate and speaking mode. That is why efforts are made to model pronunciation variations in ASR system. In Mandarin speech recognition, there is a large body of research work on connected and read speech and several commercially available systems. However, work related to spontaneous or conversational Mandarin is rare. Even fewer attempts have been made to model pronunciation variations. Applying acoustic models trained from read speech directly to spontaneous speech recognition always leads to high word error rate. It is well known that Mandarin has its own special monosyllabic structure. This structure makes pronunciation variations difficult to model. In this paper, we present two new methods to model intra-syllable and inter-syllable variations in spontaneous Mandarin speech—(1) we generate a probability lexicon to model intra-syllable variations and; (2) we incorporate the pronunciation variations in a recognizer to model both intra and inter-syllable variations.

The paper is organized as follows: section 2 introduces pronunciation variations in Mandarin speech. In section 3, we describe how to calculate the probability of pronunciation variations. In section 4, we explain how to integrate variation probability into a decoder. The experimental results are given in section 5. We conclude in section 6.

## 2. PRONUNCIATION VARIATIONS

If words were pronounced in a consistent manner, ASR would be relatively easy. However, due to different speakers, accent, speaking style, speaking rate, speaking mode, words/syllables are almost always pronounced inconsistently in spontaneous speech. Usually, studies on pronunciation variation consider such variation to be internal to a word [9]. A common way to model the variation is to introduce several phonetic entries per word in the lexicon or generate a pronunciation network. This type of variation modelling is however inadequate for Chinese. Chinese is monosyllabic and highly homophonic. Chinese words are very short, usually consisting of 2 to 3 syllables. Each Chinese syllable is also very simple, consisting only of an *initial* phone and a *final* or only the *final*. Almost all initials are very short in duration compared to the entire syllable and their pronunciations are very flexible in continuous speech [reference to CASS paper]. From our previous work [11], we found that a lot of initial variations are caused by cross syllable variations due to coarticulation and tone sandhi effect. For example, Initial “m, n” are always misrecognized when they are followed by “en,eng” finals. Initial “y” is sometimes deleted when following “i, in”. To solve such problems, we cannot ignore inter-syllable variations. Therefore, we cannot just add pronunciation variations into the lexicon.

In the following sections, we describe how to generate a probability lexicon with intra-syllable variations and how to model inter-syllable variations in addition by incorporating left-context dependent I/F probabilities into the decoder.

### 3. MODELING INTRA-SYLLABLE VARIATIONS IN A PROBABILITY LEXICON

Generally, there are two ways to extract information in pronunciation variation: data-driven and by phonological rules. It is clear that the choice depends on the amount of transcription label files we have and the reliability of phonological rules for different users. For Mandarin speech, we tend to adopt data-driven method. This is because although the characters or canonical syllables are the same in Mandarin Chinese, actual speaker pronunciation can be very different due to regional accents<sup>1</sup>. It is extremely difficult to generate a comprehensive set of phonological rules for all accents in Mandarin Chinese. In our work, we focus

<sup>1</sup> There are four major languages in China and 8 major accent groups. Mandarin is not the first language of most Chinese speakers.

on the Broadcast News Corpus (HUB4-NE), obtaining pronunciation variations in Mandarin using data-driven methods.

We can obtain intra-syllable variations using the following algorithm:

Define a set  $B = \{q_1, q_2 \dots q_N, insertion\}$

Where  $B$  contains all initials and finals of Mandarin, as well as.  $q_i$  is one of the phones in the set,  $N$  is total number of initial, finals and insertion.

Given utterances in the database, for each utterance:

1. Let  $y = y_1 y_2 \dots y_m$  be the *canonical* transcription (base form) generated from standard lexicon (one entry one standard pronunciation), which are represented by an initial/final string.
2. Generate HMM of each Initial/Final by using  $y$  as label files with EM training on the original database.
3. Let  $x = x_1 x_2 \dots x_n$  be the *observed* initial/final string of the utterance (surface form), which can be obtained by phone recognition.
4. Since  $m$  is not always equal to  $n$ , align strings  $x$  and  $y$  by a DP algorithm.
5. Use all the training utterances in an iteration of steps 1--4, to generate a phone confusion matrix.
6. Calculate variation probability  $P(q_j^o | q_i^T)$  from the confusion matrix.  $q_j^o$  is the observed phoneme and  $q_i^T$  is the canonical phoneme,  $i, j = 1, 2, \dots, N$ .

$$P(q_j^o | q_i^T) = \frac{\text{count}(q_j^o | q_i^T)}{\sum_j \text{count}(q_j^o | q_i^T)}$$

7. Obtain  $P(q_i^T | q_j^o)$  by Bayes equation,  $P(q_i^T | q_j^o) = P(q_j^o | q_i^T) * P(q_i^T) / P(q_j^o)$

$P(q_j^o | q_i^T)$  is then used to generate the probability lexicon. For each  $P(q_j^o | q_i^T)$ , the five highest variation probability candidates are selected for the probability lexicon. An example is as follows:

$p$  0.557692  $p$  0.134615  $t$   $p$  0.057692  $f$

*m* 0.565476 *m* 0.113095 *Del* *m* 0.053571 *n*  
 The first phone is  $q_i^T$ , the second phone is  $q_j^O$ .

#### 4. MODELLING BOTH INTRA AND INTER-SYLLABLE VARIATIONS IN A DECODER

Using a probability lexicon instead of the canonical lexicon enables the recogniser to handle intra-syllable variations, accommodating more flexible pronunciations. However, as we mentioned before, many pronunciation variations in spontaneous Mandarin speech are across syllables. Simply incorporating the probability  $P(q_j^T | q_i^O)$  into a decoder cannot handle inter-syllable pronunciation variations since this probability comes from the confusion matrix and has no context information. Due to this problem, we propose using left-context dependent variation probability to model inter-syllable variants. It is defined as  $P(q_j^T | q_i^O, q_{i-1}^O)$ , where  $q_{i-1}^O$  is the previous phone of  $q_i^O$  in the initial/final string of  $x$ , the surface form. Actual spontaneous speech data shows that the final part of the previous syllable heavily influence the initial part of the following syllable. Finals are much more stable than initials and Chinese syllable must end with a final. This means that left-context dependent variation modeling is very important for Mandarin speech.

The algorithm of obtaining  $P(q_j^T | q_i^O, q_{i-1}^O)$  is the same as described in the previous section. In order to train  $P(q_j^T | q_i^O, q_{i-1}^O)$  robustly,  $q_i^O$  and  $q_{i-1}^O$  can be clustered into a phone class according to phonological rules, this rule-based clustering technique is described in [10].

##### 4.1 Incorporating variation in a decoder

In this section, we demonstrate how  $P(q_i^T | q_j^O)$  and  $P(q_j^T | q_i^O, q_{i-1}^O)$  are incorporated into the decoding part for intra and inter-syllable pronunciation variations. Our baseline recogniser is a stack decoder with standard lexicon (one pronunciation per entry) and N-best lists. Assuming that N-best phone sequences have been obtained from the recognizer. For each phone sequence,

Define:

$$Q^{(n)} = \{q_1^{(n)}, q_2^{(n)} \dots q_{m_n}^{(n)}\}$$

$Q^{(n)}$  is the  $n$ th phone sequence in the N-best lists.  $q_i$  is one of the recognized (observation) phone,  $m_n$  is the number of phones in this sequence. We save the log likelihood of each recognized phone.

Then define:

$$R^{(n)} = \sum_{i=1}^{m_n} \max_j \log P(q_j^T | q_i^{(n)})$$

For left-context dependent phone probability, we have

$$R^{(n)} = \sum_{i=1}^{m_n} \max_j \log P(q_j^T | q_i^{(n)}, q_{i-1}^{(n)})$$

$q_j^T$  is the canonical phoneme,  $q_i^{(n)}$  and  $q_{i-1}^{(n)}$  are the observed phones in  $Q^{(n)}$ .

Combining with  $R^{(n)}$ , the log likelihood of each recognized phone in the phone sequences  $Q^{(n)}$  is re-scored. The re-scored phone sequences are:

$$Q'^{(n)} = \{q_1'^{(n)}, q_2'^{(n)} \dots q_{m_n}'^{(n)}\}$$

$m_n'$  is the number of phones in the phone sequence after re-scored. Due to deletions and insertions,  $m_n'$  is not always equal to  $m_n$ . Each  $q_i'^{(n)}$  in the phone sequence can be obtained by

$$q_i'^{(n)} = \arg \max_{q_j^T} (\lambda \log P(q_j^T | q_i^{(n)}) + \log P(q_i^{(n)}))$$

or

$$q_i'^{(n)} = \arg \max_{q_j^T} (\lambda P(q_j^T | q_i^{(n)}, q_{i-1}^{(n)}) + \log P(q_i^{(n)}))$$

where  $\lambda$  is the weight.

Depending on the pronunciation lexicon used in the decoder, new phone sequence  $Q'^{(n)}$  can be represented either in Chinese characters, Chinese syllable or initial/final sequence.

##### 4.2 Modelling insertion and deletion errors

We suggest that some substitution errors can be corrected by re-scoring the phone sequence with error models. Both insertion and deletion errors can be modelled in a similar manner as the phones. We regard insertion as a phone in set **B**. So its probability can also be written as  $P(q_j^T | q_i^{(n)})$  and  $P(q_j^T | q_i^{(n)}, q_{i-1}^{(n)})$ .

$q_j^T$  is insertion. We add another model  $P(q_j^D | q_i^{(n)}, q_{i-1}^{(n)})$  to deal with deletion, where  $q_j^D$  represents all possible phones.  $P(q_j^D | q_i^{(n)}, q_{i-1}^{(n)})$  is the probability of having a deletion of  $q_j^D$  between  $q_i^n$  and  $q_{i-1}^n$ .

## 5. EXPERIMENTAL RESULTS

We use Hub4NE 1997 Mandarin Broadcast News database provided by LDC to evaluate the effectiveness of our approach. There are 23 initials and 37 finals. The syllable number is 415. We use three-state, left-to-right HMMs and 32 Gaussian mixtures. The acoustic features are 13MFCCs, 13 delta MFCCs and 13 acceleration MFCCs.

Two CDs (7 hours) of Hub4NE data is used for training acoustic model and variation probability. The data includes planned, spontaneous and conversational speech, speech with music and background noise. The testing data is about 1 hour spontaneous speech selected by hand from the database. The syllable error rate of the baseline system is 36.3%. Using our probability lexicon (selected top 5 variations), it decreased to 35.45%. When we incorporate both intra and inter-syllable variation probability into the decoder, syllable error rate is reduced significantly to 31.54%

## 6. CONCLUSION

We present a new approach to model both intra-syllable and inter-syllable pronunciation variations by data-driven methods. Preliminary results show a significant increase in the performance of predicting the correct pronunciation variations as well as major improvement in recognition accuracy. Although Chinese syllable is much simpler than western languages, we can still model the pronunciation variations effectively. In addition, to the best of our knowledge, this is the first time that inter-syllable pronunciation variations in Mandarin Chinese is modelled. Experiment results show that our method is also effective for modelling insertion and deletion errors.

Our future work includes generation of robust lexical level models, using a hand transcribed data for bootstrapping, and incorporating supra-segmental information into the data-driven algorithm.

## 7. REFERENCE

- [1] Simon Downey and Richard Wiseman, "Dynamic and Static Improvements to Lexical Baseforms". Proceedings of Eurospeech'97
- [2] Seong-Jin Yun et.al., "Stochastic Lexicon Modeling for Speech Recognition". IEEE Signal Processing Letters. Vol.6 No.2 Feb.1999
- [3] Toshiaki Fukada et.al., "Automatic Generation of a Pronunciation Dictionary Based on a Pronunciation Network". Proceedings of Eurospeech'97
- [4] Judith Kessens and Mirjam Wester., "Improving Recognition Performance by Modelling Pronunciation Variation".
- [5] John Eric Fosler-Lussier "Dynamic Pronunciation Models for Automatic Speech Recognition". Ph.D. thesis, International Computer Science Institute, 1999
- [6] Ellen Eide. "Automatic Modeling of Pronunciation Variations". Eurospeech'99
- [7] Murat Saraclar, et.al, "Pronunciation modeling by sharing Gaussian densities across phonetic models" Computer Speech and Language (2000) 14, 137-160
- [8] Yumi Wakita, et.al, "Multiple pronunciation dictionary using HMM-based confusion characteristics" Computer Speech and Language (1999) 13, 143-153
- [9] Laure Brueussel-Pousse et.al, "Language model level VS. Lexical Level for Modeling Pronunciation Variation in a French CSR" Eurospeech'99 Vol.4, pp1771-1774
- [10] LIU Yi, Pascale FUNG "Ruled-based word pronunciation networks generation for Mandarin speech recognition" Appear in ISCSLP2000
- [11] LIU Yi, Pascale FUNG "Decision tree-based triphones are robust and practical for Mandarin speech recognition" Eurospeech'99 pp 895-898