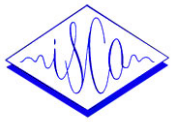

International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

RULE-BASED WORD PRONUNCIATION NETWORKS GENERATION FOR MANDARIN SPEECH RECOGNITION

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

ABSTRACT

Modeling pronunciation variation in spontaneous speech is very important for improving the recognition accuracy. One limitation of current recognition systems is their dictionaries for recognition only contain one standard pronunciation for each entry, so that the amount of variability that can be modeled is very limited. In this paper, we proposed to generate pronunciation networks based on rules to instead of traditional dictionary for decoder. The networks consider the special structure of Chinese and incorporate acceptable variants of each Chinese syllable . Also, an automatically learning algorithm is designed to get the variation rules. The proposed method was experimented on Hub4NE 1997 Mandarin Broadcast News Corpus and HLTC stack decoder. The syllable recognition error rate was reduced 3.20% absolutely with both intra- and inter-syllable variations are both modeled.

1.INTRODUCTION

In continuous speech recognition, the acoustic models are based on sub-word units. In order to recognize word, one need word models describing the pronunciation of the words in terms of these units [9]. For each word, if there are several acceptable pronunciation variants, they are considered as intra-word variants. Coarticulation between words can bring more pronunciation variation, which is regarded as inter-word variants. Due to phonological knowledge and some experimental results, in Mandarin spontaneous speech, the pronunciation variations can be classified into two types: one is variation within the same phoneme, such as *f* changes to *f_v*, *ts* to *ts_v* or *ts_h*; another variation is changed beyond phoneme which means changing to another quite different phonemes. The first variations can be modelled by using GIFs model [11], but GIFs cannot solve the second variations. Previous work on modeling the second variations is to include the possible variations of the phoneme to the lexicon, however it also increases the ambiguity between the phonemes and leads to actual performance decrease.

In this paper, we develop a new method for generating syllable pronunciation networks representing different acceptable pronunciations of each syllable. The pronunciation networks are generated from variation rules. We start from single standard traditional pronunciation of each syllable, the variations of each syllable are produced by applying pronunciation rules. Based on these rules learned from training data, the standard pronunciation can be rewrited, and the probability of each variation can be attached to the pronunciation networks. Although the variation rules can be generated from phonology knowledge, however, we proposed the rules trained from the data. 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, even in spontaneous Mandarin speech. It is extremely difficult to get a comprehensive set of phonological rules for all accents in Mandarin [7]. In our decoder, the generated pronunciation networks is efficiently incorporated to model both intra and inter-syllable pronunciation variations.

The paper is organized as follows: section 2 introduces pronunciation variations in spontaneous Mandarin speech. In section 3, we showed how to automatically trained the variation rules from the data. In section 4, we explain how to generate pronunciation networks based on the trained rules. In section 5, we will discuss the method about incorporate the pronunciation variations into decoder. The experimental results are given in section 6, we

conclude in section 7.

2. PRONUNCIATION VARIATIONS

In read speech, we often obtained high speech accuracy. However, in spontaneous speech, due to speaking rate, speaking style, accent and speaking mode, word/syllable are almost pronounced inconsistently and the speech accuracy is very low. Simply added the possible pronunciations of each word/syllable in the lexicon seems the easy way, but experiment showed that it decrease the recognition accuracy since it also increase the ambiguity between the words or syllables. Much works for modeling pronunciation variation have been done in western languages, however works related to Mandarin is rare. Chinese is monosyllabic and highly homophonic, Chinese character is ideographic and does not reflect the pronunciation of a word [10]. Each Chinese syllable will map to many Chinese characters and its structure is very simple, consisting only of an initial phone or 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 spontaneous speech. From [7] we know that inter-syllable variations cannot be ignored and we cannot just add pronunciation variations into the lexicon.

In the next sections, we describe how to automatically train the variation rules, generate pronunciation networks based on the special structure of Chinese syllables and incorporate the variation probabilities into the decoder.

3. TRAINING PRONUNCIATION RULES

The form for the pronunciation rules of western languages is like $LPR \rightarrow P^{'}$, P means focus phone, L and R are *left context* and *right context* condition, $P^{'}$ is modified phone [9]. However, this form is not suitable for Chinese. We have showed before, Chinese is monosyllabic and each syllable only most has initial and final, if we use $LPR \rightarrow P^{'}$ format, it must include both intra and inter syllable variations, and inter-syllable variation rules should applied to syllable pairs and not to pronunciation networks. So in this paper, we divided the pronunciation rules form as follows. For common syllables in the Chinese:

Initial: $LP \rightarrow P^{'}$ is inter-syllable variation. P is focus initial, L is left context, which is final part of previous syllable. $P^{'}$ is modified initial. $PR \rightarrow P^{'}$ is intra-syllable variation, R is the final part of its own syllable.

final: $LP \rightarrow P^{'}$ is intra-syllable variation. P is focus final, L is left context, which is initial part of the own syllable. $P^{'}$ is modified final. $PR \rightarrow P^{'}$ is inter-syllable variation, R is the initial part of the following syllables.

For zero initial-syllables in the Chinese, both $LP \rightarrow P^{'}$ and $PR \rightarrow P^{'}$ are inter-syllable variations.

The learning algorithm of obtaining pronunciation rules is showed as follows:

1. For each utterance in database, generate canonical transcription T_{true} (base form) from standard lexicon (one entry one standard pronunciation), it will be represented by an initial/final string.
2. Perform phone recognition and obtained observed transcription T_{ob} (surface form).
3. Define pronunciation rules structure as showed above.
4. Align T_{true} and T_{ob} with dynamic programming search algorithm.
5. Given the syllable boundary information in T_{true}, with forced alignment Viterbi algorithm, generate both intra-syllable and inter-syllable pronunciation rules.

From the algorithm, pronunciation rules can be learned automatically, the format can be showed as:

$(eng)m \rightarrow n$ 0.35 (inter-syllable rules)
$(eng)m \rightarrow m$ 0.45 (inter-syllable rules)
$(y)an \rightarrow i, an$ 0.45 (intra-syllable rules)

After processing the training set in the above algorithm, a large number of pronunciation rules are obtained. Generally this set is too large, the occurrences of some rules are very small and cannot represent the common variation. In order to

guarantee the trained rules robust, we clustering the different L and R conditions into classes. The classes are decided by phonologic rules of Mandarin, more details showed in [8]. After clustering, pronunciation rules with low coverage and low probabilities are still pruned. The threshold set for pruning is 5% for probability.

4. PRONUNCIATION NETWORKS

Based on section.3, the set of pronunciation rules contains two types of rules: intra-syllable rules and inter-syllable variation rules. The latter one showed pronunciation changes at the syllable boundaries and the syllable internal rules are occurred in syllables, they are taken into account for generating pronunciation networks. At the entrance of the networks, we include the inter-syllable rules, so it will have different conditional entries and exits. In the next section, we will show to combine inter-syllable rules with syllable bigram LM.

These are examples of the part of pronunciation networks:

$$\begin{aligned} &R1 \rightarrow \text{zhong}\ \ 0.7\ \text{zh} \quad 0.7\ \text{ong} \rightarrow \text{Prob1} \\ &R2 \rightarrow \text{zhong}\ \ 0.18\ \text{z} \quad 0.7\ \text{ong} \rightarrow \text{Prob2} \\ &R3 \rightarrow \text{zhong}\ \ 0.1\ \text{ch} \quad 0.2\ \text{eng} \rightarrow \text{Prob3} \end{aligned}$$

R1, R2 and R3 are different previous finals of previous syllable, Prob1, Prob2 and Prob3 give different emissions.

5. INCORPORATE PRONUNCIATION NETWORKS IN DECODER

Define: $S = s_1, s_2, \cdots s_n$ is the syllable sequence, A is the observed acoustic sequence. The speech recognition target is to find the most possible syllable sequence S^* given A.

The formula is: $S^* = \arg\max_S P(S \mid A)$, by Bayes equation:

$$\begin{aligned} S^* &= \arg\max_S \frac{P(A \mid S) \cdot P(S)}{P(A)} \\ &= \arg\max_S P(S) P(A \mid S) \end{aligned}$$

From definition, we got:

$$S^* = \arg\max_S P(s_1, s_2, \cdots s_n) P(A \mid s_1, s_2, \cdots s_n) \quad (1)$$

If we do not use pronunciation networks, which means each syllable is pronounced in a consistent manner, Eq.1 can present the speech recognition job. However, when pronunciation network applied and each syllable has several variations, for example, syllable s_i has the variations $v_{i1}, v_{i2}, \cdots v_{in}$, Eq.1 changes to:

$$\begin{aligned} S^* = \arg\max_S P(s_1, s_2, \cdots s_n) \Big[\sum P(A \mid v_{1k1}, \cdots v_{nkn}, s_1, \cdots s_n) \\ \cdot P(v_{1k1}, \cdots v_{nkn} \mid s_1, \cdots s_n) \Big] \end{aligned} \quad (2)$$

Assume $s_1, \cdots s_n \subset v_{1k1}, \cdots v_{nkn}$, Eq.2 is simplified:

$$\begin{aligned} S^* = \arg\max_S P(s_1, s_2, \cdots s_n) \Big[\sum P(A \mid v_{1k1}, \cdots v_{nkn}) \\ \cdot P(v_{1k1}, \cdots v_{nkn} \mid s_1, \cdots s_n) \Big] \end{aligned} \quad (3)$$

The first part of Eq.3 is independent of A, in recognition it determined by language model, in the follows we will show inter-syllable rule variation can be joined with $P(s_1, s_2, \cdots s_n)$. The second part is the probability of observation given the syllable pronunciation variations sequence. From Eq.3, the probability is determined by the acoustic likelihood of all acceptable pronunciation syllable sequences and their variation probabilities. If we use pronunciation networks, consider only the intra-syllable variations:

$$\begin{aligned} S^* = \arg\max_S P(s_1, s_2, \cdots s_n) \Big[\sum P(A \mid v_{1k1}, \cdots v_{nkn}) \\ \cdot P(v_{1k1} \mid s_1) \cdots P(v_{nkn} \mid s_n) \Big] \end{aligned} \quad (4)$$

Eq.4 is true since in section.3 the networks are generated from intra-syllable rules. We focus on $P(v_{v1k1} \mid s_1) \cdots P(v_{nkn} \mid s_n)$ since it is the variation information. To each v_{ikn}, we write it in initial/final sequence $v_{ikn} = (I_{i,j}, F_{i,j})$, $I_{i,j}$ is for possible initials and $F_{i,j}$ is for finals. So that:

$$\begin{aligned} &P(v_{v1k1} \mid s_1) \cdots P(v_{nkn} \mid s_n) \\ &\quad = \sum P(I_{1,j}, F_{1,j} \mid s_1) \cdots P(I_{n,j}, F_{n,j} \mid s_n) \\ &\quad = \sum P(F_{1,j} \mid s_1) \cdot P(I_{1,j} \mid F_{1,j}, s_1) \\ &\qquad \cdots \sum P(F_{n,j} \mid s_n) \cdot P(I_{n,j} \mid F_{n,j}, s_n) \end{aligned} \quad (5)$$

It is obvious that Eq.5 can be obtained from pronunciation networks, which means intra-syllable variations have been integrated into the decoder.

The inter-syllable rules are applied to syllable pairs, which means at enter or exit part of pronunciation

networks. In our decoder, only back-off syllable bigram is considered. With inter-syllable rules, the new transition probability can be:

$$P_{new} = P(Rule, s_2 \mid s_1)$$
$$= P(s_2 \mid s_1) \cdot P(Rule \mid s_2, s_1) \quad (6)$$

The first part is original syllable bigram language model, the second part is obtained from the corresponding inter-syllable rules.

Combined Eq.4, Eq.5 and Eq.6, both the intra-syllable and inter-syllable variations have been incorporated into the decoder.

6. EXPERIMENTAL RESULTS

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

Two CDs (about 7 hours) of Hub4NE data is used for training acoustic model and pronunciation rules. 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%. After clustering and pruning, about 600 pronunciation rules are remained. Using pronunciation networks generated from internal rules, the syllable error rate decreased to 35.24%. When we incorporate both intra and inter-syllable rules to the decoder, syllable error rate is reduced significantly to 33.10%.

7. CONCLUSION

In this paper, we proposed a new method for generating pronunciation networks for Mandarin speech recognition. The networks consider the special structure of Chinese and incorporate different acceptable pronunciation variations. Also, we designed an automatically learning algorithm to generate a set of both intra and inter-syllable rules from the training data. Preliminary results show a significant increase in the performance of predicting the correct pronunciation variations as well as major improvement in recognition accuracy.

Our further work includes detailed designing the clustering classes, using hand transcribed data for bootstrapping and working on Mandarin accent spontaneous speech.

8. 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] LIU Yi, Pascale Fung, “Modelling pronunciation variations in spontaneous Mandarin speech” To appear in ICSLP2000

[8] LIU Yi, Pascale Fung, “Decision tree-based triphones are robust and practical for Mandarin speech recognition” Eurospeech’99 pp895-898

[9] Nick Cremelie, et.al, “Automatic rule-based generation of word pronunciation networks” Eurospeech’97

[10] Mingkuan Liu, et.al, “Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling” ICASSP’2000

[11] Li Aijun et.al., “CASS: A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech” To appear in ICSLP2000