Pronunciation Variation Modeling for Mandarin with Accent

ZHANG Chi, WU Ji, XIAO Xi, WANG Zuoying

Department of Electronics Engineering
Tsinghua University , Beijing, P.R.China
zclzc99@mails.tsinghua.edu.cn wuji@thsp.ee.tsinghua.edu.cn

Abstract

In order to solve the problem of the performance decrease when state-of-art automatic speech recognition (ASR) system facing accent speech, we propose the Pronunciation Variation Model (PVM). Two approaches are proposed to construct the PVM in this paper. 6.38% and 7.78% relative error rate reduction is achieved for Shanghai and Wuhan accent mandarin, respectively. The experiment on these two typical accent mandarin shows it is a possible way to deal with accent speech.
Index Terms: speech recognition, accent, pronunciation variation Mode

1. Introduction

Nowadays, the performance of state-of-art ASR system is fairly good in the quiet laboratory environment, but degrades drastically in real applications. Many factors can cause the increase of recognition error, such as background noise, channel difference, rate of speech, and so on. In this paper, we mainly focus on an important problem: accent of speech.

There are many dialect areas in Chinese; some dialects are quite different from others. All the people would like to speak in Mandarin to communicate easily with each other. But the people from different dialect areas would like to have different accents. Usually, the ASR systems are trained with Mandarin speech. So the performance will degrade dramatically when such system faces accent speech.

One method to solve this problem is to train a specific ASR system for each accent. Then we have two choices, one is the input speech is decided to belong to some accent speech with accent identification model and sent to the corresponding recognizer; the other is the input speech is sent to every recognizer, and the recognition result is chosen from the output of these recognizers. For the first scheme, the decision error may affect the recognition accuracy significantly. For the second scheme, the time-complexity will be very high.

Fortunately, most people in the same dialect area still have similar accent. In this paper, we try to set up a framework of acoustic model, which can describe different accent speech at the same time.

In Section 2, we introduce the concept of PVM, and how to build PVM model and apply the model in ASR system. Section 3 presents some experimental result on PVM as compare to traditional system. Finally, Section 4 concludes the paper.

2. Pronunciation Variation model

2.1. Framework of ASR system

In Mandarin speech, all Chinese words are monosyllables, and each one made up of one initial (the first half of syllable: consonant or zero-consonant) and one final (the second half of syllable: vowel). Let $W=\{W_1, W_2, \cdots, W_M\}$be the word sequence, M is the length of the sequence. W_i is made up of inital Ci, and final V_i, and the pronunciation of W_i is A_i (i = 1,2,…,M). The corresponding observing sequence of acoustic vectors can be denoted as $X=\{X_1, X_2,\ldots X_T\}$, which are derived from speech signal in the preprocessing step of acoustic analysis, T is the frame number.

Speech recognition is to determine the word sequence W to maximize the posterior probability $P(W/X)$, can be written as:

$$\hat{W} = \arg\{\max_{W} P(W/X)\} \tag{1}$$

Because more knowledge can leads to better performance, many kinds of knowledge has been integrated into the advanced ASR system, such as pitch and tone information, language Model, part of speech (POS), semantics and so on.

Let A, P, S be the pronunciation sequence, POS sequence, semantics sequence. From (1), we have:

$$\hat{W}\hat{A}\hat{P}\hat{S} = \arg\{\max_{(S,P,W,A)} P(WAPS/X)\} \tag{2}$$

By applying Bayes' theorem on conditional probabilities, equation (2) can be written as:

$$P(WAPS/X) = P(W)P(XAPS/W)/P(X) \tag{3}$$

Since the observing probability on acoustic vector X is only depend on pronunciation A, so we obtain:

$$\begin{aligned} P(XAPS/W) &= P(A/W)P(XPS/WA) \\ &= P(A/W)P(P/W)P(S/WP)P(X/A) \end{aligned} \tag{4}$$

The recognition can defined as an optimization problem:

$$\hat{W} = \arg\{\max_{(S,P,W,A)} [P(A/W)P(P/W)P(S/WP)P(X/A)P(W)]\} \tag{5}$$

In (5), $P(X/A)$ stands for acoustic model, $P(W)$ stands for word-based language model, and $P(A/W)$, $P(P/W)$, $P(S/WP)$, are transcription model, POS model and Semantic Model respectively.

Formula (5) is a uniform framework for speech recognition, which provides an effective way to utilize and combine all kinds of knowledge in speech recognition.

2.2. Pronunciation Variation Model

In Mandarin, the transcription module is much simpler than in western language, the reason is all Chinese words are monosyllable. We should only do with polyphone of some certain Chinese words in this module.

In this paper, we try to extend transcription model $P(A/W)$ to deal with accent speech. And we call this extend model as PVM: Pronunciation Variation Model.

Neglecting the influence of two nearby single word, we have:

$$P(A/W) = \prod_{k=1}^{M} P(A_k / W_k) \tag{6}$$

Usually in recognition system for Mandarin speech, we consider the word and pronunciation in a standard way. That means, from the word sequence W to the pronunciation sequence A is almost a one-to-one mapping. So we can take $P(X/W)$ as classical acoustic model $P(X/A)$ directly.

If we consider pronunciation variation caused by accent, $P(X/W)$ is different from the acoustic model $P(X/A)$, the recognition problem can be written as:

$$\hat{W}\hat{A} = \arg\{\max_{WA}[P(X/A) + P(A/W)]\} \tag{7}$$

$P(A/W)$ is PVM, the model proposed to describe the different pronunciation of word sequence A in different accent speech.

For Mandarin, there are 100 initials and 164 finals. Since each syllable is made up of one initial and one final. So the number of total possible syllable is 16400. But there are only 1254 syllables exists in Mandarin, more than 99 percent connections are illegal in Mandarin.

We find that most of the pronunciation in accent speech can be expressed by the syllables within the 16400 possible connections, although they cannot be found in 1254 standard syllables. So it becomes a natural method to transcribe the pronunciation in accent speech with an extended syllable list, which has more syllables than the Mandarin syllable list. And all these syllables are from 16400 possible connections.

Then for each word W_k, we have different pronunciations in PVM. In order to integrate the PVM into the ASR system, we should depict the mapping from word W_k to pronunciations $A_{k_i}, i = 1,2,\cdots n$ statistically. The formula we use to estimate the probability $P(A_k|W_k)$ is:

$$P(A_{k_i} / W_k) = \frac{Number\ of\ A_{k_i}}{Total\ Number\ of\ W_k} \tag{8}$$

In order to find if such method can describe the pronunciation in accent speech, we test it in two typical accents in Chinese speech, accent speech from Shanghai and Wuhan dialect areas. In our implementation of PVM, we have 2378 syllables in the extended syllable list, which covers 97.85% pronunciations of the Shanghai accent speech, and 99.36% pronunciations of the Wuhan accent speech.

The results show this extend syllable list can describe accent speech quite good.

2.3. Recognition algorithm with PVM

The proposed PVM has been integrated in THEESP Mandarin speech recognition system. The flow chart of traditional recognition system is in Figure1. For the acoustic level recognition, we first get the word candidates, which is word graph or word lattice. Then with the word-based language model, the final recognition result is determined.

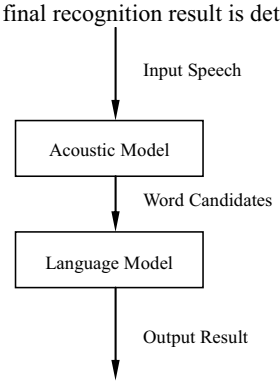

Figure 1 *Traditional ASR System*

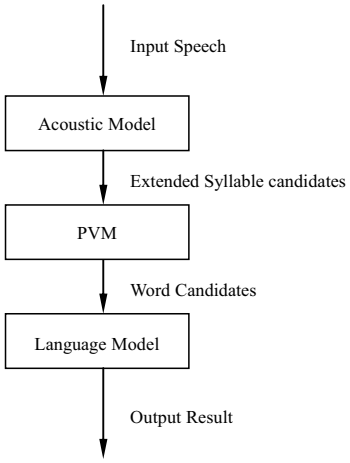

Figure 2 *ASR System with PVM*

Figure 2 is the flow chart of the ASR system with PVM. The first step is still the acoustic level recognition, but with the extended syllable list including pronunciation in accent speech, what we derived is a sequence with extended syllables.

$$\hat{A} = \arg\{\max_{A}[P(X/A)]\} \tag{9}$$

Ignore the correlation between syllables, we have

$$(\hat{A}_1\hat{A}_2\cdots\hat{A}_N) = \arg\{\max_{A_1A_2\cdots A_N} \prod_{i=1}^{N} P(X_{i_0}\cdots X_{i_{N_i}} / A_i)\} \tag{10}$$

$X_{i_0} \cdots X_{i_{N_i}}$ is the observing sequence on acoustic vectors corresponding to syllable A_{ib}.

For syllable sequence $A_1 A_2 \cdots\cdots A_N$, there are M_j different pronunciations of syllable A_j. We have to calculate M times to determine the word sequence candidates.

$$M = M_1 * M_2 * \cdots * M_N \tag{11}$$

When the length of syllable sequence keep growing, and most of the words can be pronounced in several ways. It will be over-loaded if we want to find the best word sequence candidates. Obviously, it cannot be acceptable in real applications.

To simplify the recognition process, we provide the algorithm below. At first we use viterbi algorithm to decide the segmentation the whole sequence and get a series of segment point pairs. Then we directly calculate the distance for the pronunciations to acoustic vectors between these pairs of segment point. After such simplifying, the calculation times reduce dramatically, like that:

$$M = M_1 + M_2 + \cdots + M_N \tag{12}$$

Because manpower is used to decide the pronunciation of each word in training data, we called this algorithm as knowledge-based PVM.

2.4. Data Driven PVM

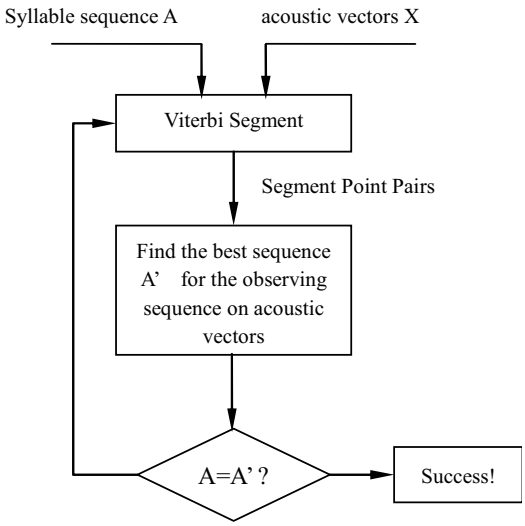

Figure 3 *Data Driven PVM*

In knowledge-base PVM, we provide the pronunciation label for each word in training data according to the extended syllable list with manpower. It is unavoidable that must be some error in decision rule and man-made label. So we try to use data-driven method to give the optimization result. The flow chart is in Figure 3.

The first step is dynamic programming for observing sequence and standard syllable sequence with viterbi algorithm. In the next step, the best syllables are chose from different pronunciations of certain word candidates. If the best syllables are the same as the standard syllable sequence, the iterating stops, otherwise, changing the standard syllable sequence according to the best sequence, do the dynamic programming for another time. Keep the process, until the iteration stop.

2.5. Probability Smoothing

Since the training data is always limited, there maybe only a few samples can be used to train the PVM for some unfamiliar pronunciations. So we must do smoothing for such words in a particular way. For the words with N(N>2) different pronunciations, we set $P(A_k|W_k)$ a fixed value a for those standard pronunciation, and divide the rest probability to other pronunciations on average.

$$P(A_{k_i} / W_k) = \begin{cases} a & A_{k_i} = W_{k_i} \\ (1-a)/(N-1) & Others \end{cases} \tag{13}$$

3. Experiment Results

The database for experiment is the INTEL Mandarin Corpus with Beijiing, Shanghai, Wuhan accent. All the speech data is pre-emphasized, windowed to 25-ms frames with 10-ms frame shift, and parameterized into 45-order MFCC, consisting of 14 cepstral coefficients, energy, and their first and second order differences. Cpestral mean subtraction (CMS) is performed within each utterance to remove the effect of channels. The acoustic model is trained by Gaussian mixture model, with 24 mixtures in it.

The test corpus comprise of about 400 sentences each from 22 speakers, 10 from Beijng, 6 from Shanghai and 6 from Wuhan. Half of the speakers are men and the other half are women.

In order to build a contrast group, we trained three different acoustic models for Beijing, Shanghai and Wuhan, respectively. These three acoustic models present the ideal performance of our experiment.

3.1. Knowledge-based PVM

SH	Corr	Del	Ins	Sub	WER
Baseline	72.40%	1.03%	3.88%	26.57%	31.48%
KB PVM	73.16%	1.10%	3.95%	25.74%	30.80%
WH	Corr	Del	Ins	Sub	WER
Baseline	57.73%	0.91%	6.61%	41.35%	48.87%
KB PVM	60.51%	0.82%	6.47%	38.67%	45.96%

Table 1. *Result of knowledge PVM*

In Table 1, Baseline represents the baseline system without PVM. KB PVM represents the ASR system with Knowledge-based PVM. SH and WH represent Shanghai and Wuhan accent mandarin speech, respectively. The results show the WER (word error rate) reduces with PVM.

For Shanghai accent speech, the WER reduces from 31.48% down to 30.80%, 2.16% relative error rate reduction achieved. For the Wuhan accent speech, 2.91% relative error rate reduction achieved, which reduces from 48.87% to 45.96%.

3.2. Data Driven PVM

SH	Corr	Del	Ins	Sub	WER
Baseline	72.40%	1.03%	3.88%	26.57%	31.48%
DD PVM	74.65%	0.98%	4.11%	24.38%	29.47%
WH	Corr	Del	Ins	Sub	WER
Baseline	57.73%	0.91%	6.61%	41.35%	48.87%
DD PVM	61.78%	0.80%	6.84%	37.43%	45.07%

Table 2. Result of Data Driven PVM

In Table 2, Baseline presents the baseline system without PVM. DD PVM presents the ASR system with data driven PVM.

6.38% and 7.78% decline of relative error rate reduction is achieved for Shanghai and Wuhan accent speech, respectively.

We can find out that the performance of data driven PVM is better than knowledge-based PVM.

3.3. mulit-accent PVM

Test Data	Train Data	WER
SH	baseline	31.48%
	SH	29.47%
	SH WH	29.55%
	SH WH BJ	30.02%
WH	baseline	48.87%
	WH	45.07%
	SH WH	45.28%
	SH WH BJ	46.20%
BJ	baseline	28.48%
	WH	31.31%
	SH	31.55%
	SH WH	31.55%
	SH WH BJ	29.27%

Table 3. Result of mix-training with multi-accent data

The above experiment is the situation that only one kind of accent speech exists. But in the real application, we should face different accent speech at the same time.

Because the PVM can model the pronunciation of different accent speech, we can produce mix-training acoustic model to face multi-accent speech.

Table 3 is the result for the recognition for multi-accent data with mix-training model. Column 1 is the test corpus; Column 2 is the training data we derive the PVM from. Here “SH WH” represents the training data of PVM include Shanghai and Wuhan accent speech, so the model hope to provide better performance both for Shanghai and Wuhan accent speech.

The result in Table 3 shows, for the certain accent exists in the training data, the multi-accent PVM brings a little increase in WER compared with the single-accent PVM. But the multi-accent has better ability to deal with more accents at the same time. We can rewrite the result in Table 4.Then we can find out obviously that the multi-accent model has the best general performance.

WER	SH	WH	BJ	average
baseline	31.48%	48.87%	28.48%	36.28%
SH WH	29.55%	45.28%	31.55%	35.46%
SH WH BJ	30.02%	46.20%	29.27%	35.16%

Table 4. Result of mix-training with multi-accent data

4. Conclusions

This paper proposes a new model, the Pronunciation Variation Model (PVM), to deal with the accent speech. The PVM can be obtained by either knowledge-base approach or the data-driven ones. The latter approach has better performance.

Pronunciation Variation Model (PVM) is to solve the problem that the decrease of recognition rate when the recognizer processing speech data with accent or other pronunciation variation. To solve this problem, first, we use the knowledge-based PVM. By using this model, compared with the baseline system, the Word Error Rate (WER) of the recognizer decreases by 2.16% and 2.91%, recognizing the Mandarin speech with the Shanghai and Wuhan accent respectively. Furthermore, we propose the data driven PVM. Because this approach is appropriate for the statistical framework, and can optimize the acoustic model according to the training data, we get even better performance.

And then, we use mix-training method to deal with multi-accent at the same time, the result shows the multi-accent PVM has the best general performance, and will be helpful in real applications.

5. References

[1] Helmer S，Catia C。 Modeling pronunciation variation for ASR: A survey of the literature[J] 。 Speech Communication，1999，29，225-246。

[2] Philipp S， Ronald A C， Mark F 。 Automatically Generated Word Pronunciations From Phoneme Classifier Output[A]。Proc. ICASSP93[C]。Minneapolis：1993。2223-2226。

[3] Goronzy S，Eisele K。Automatic pronunciation modeling for multiple non-native accents[A] 。 ASRU03 IEEE Workshop on[C]。Virgin Islands：2003。123-128。

[4] WANG Zuoying, FrameWork of the Automatic Speech Recognition, 1999.