Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese

Ming-yi Tsai[12], *Fu-chiang Chou*[2], *Lin-shan Lee*[1]

[1]Dept of Electrical Engineering, National Taiwan University
[2]Applied Speech Technology
Taipei, Taiwan, Republic of China
Email : pancho@speech.ee.ntu.edu.tw

Abstract

Chinese language has quite different characteristic structures from those of English. There are at least word, character, syllable, Initial-Final levels in Chinese, each carrying different levels of information with complicated correlations among them. In this paper, we investigate the dependency of pronunciation variation in conversational Mandarin speech on these different levels under various contextual conditions considering the structural features of the language. The influence of speaking rate and word frequency on such pronunciation variation is also analyzed. Different pruning methods, for including pronunciation variation in speech recognition were also evaluated, and the experimental results showed that improved accuracy is obtainable if the characteristics of the pronunciation variation found in the analysis can be properly taken into account. All discussions here are based on tests with the LDC Mandarin Call Home corpus.

1. Introduction

It has been well known that pronunciation variation very often seriously deteriorates the performance of ASR systems if not handled appropriately [1]. Since considerable pronunciation variation is usually present in conversational speech , in-depth analysis of such phenomenon becomes very important. In English and quite several other western languages, pronunciation variation in word, syllable or phone levels has been extensively studied. However, Chinese language has quite different monosyllabic structure and is a tonal language. There are at least Initial-Final, syllable, character, and word levels of linguistic units, each carrying different levels of information with complicated correlations among them. However, not too much analysis of Mandarin speech pronunciation variation with respect to the structural features of Chinese language has ever been reported in the literature. This is therefore the subject of this paper. Also, it has been pointed out that pronunciation variation is related to speaking rate and word frequency in English. Such a relationship is also investigated for Mandarin Chinese here. All results reported here are based on tests with the LDC Mandarin Call Home corpus.

In the following section 2 the framework of analysis and experiments reported in this paper is described. In section 3, the dependency of pronunciation variation for various linguistic levels is analyzed under different contextual conditions. The influence of speaking rate and word frequency on the pronunciation variation is then presented in section 4, and some recognition experiments examined in section 5. The conclusion is finally given in section 6.

2. Framework of analysis and experiments

The preliminary experiments were performed with the HTK tools on a part of the Mandarin Call Home corpus of about 6.7 hours of data of Putonghua accent, including 3.31 hours for male and 3.39 hours for female. Detailed statistics of the corpus is shown in Table 1. This corpus was used to train the gender-dependent acoustic models consisting of 58 three-state Initials and 22 four-state Finals regardless of the tone. All Initial/Final models are context independent. More explanations about the Initial/Final linguistic units will be given below. There are 24 Gaussian mixtures per state. The acoustic features are 13 MFCCs, 13 delta MFCCs and 13 acceleration MFCCs. The trained acoustic models are used both to acquire surface form and to perform recognition experiments as well.

The main steps to acquire the pronunciation confusion table are as follows:

1. Acquiring the canonic transcriptions in each level for the 6.7 hours of data
2. Using the unconstrained Initial-Final recognizer to acquire the surface form for the 6.7 hours of data
3. Aligning the canonic and surface forms with the dynamic programming algorithm
4. Generating the confusion table and obtaining the statistics for each level and each contextual condition

The pronunciation variation metrics used in this paper are the entropy (H_i and H are defined below), and the probabilities of model deletion and substitution (P_{del} and P_{sub}). Entropy has been widely used for pronunciation learning systems and found to be a good measure for the spread of pronunciations in a training set [2]. For the pronunciation set of a linguistic unit (Initial-final, syllable, character, word) i with probability distribution estimates $p_{i,r}$ for different pronunciations r, the entropy H_i is defined as

$$H_i = -\sum_r p_{i,r} \log_{10} p_{i,r} \quad , \qquad (1)$$

and the average entropy for a specific linguistic unit (Initial-Final, syllable, character or word) is then

$$H = \sum_i p_i H_i \quad , \qquad (2)$$

where p_i is the probability of observing the linguistic unit i.

After various analysis of the pronunciation variation on the 6.7 hours of Mandarin Call Home corpus as will be presented in the sections below, another set of about 30 minutes of data with the same accent from the Mandarin Call Home corpus was tested in recognition experiments with bigram language model based on a lexicon of roughly 8K words. The baseline

experiment was based on a static lexicon with only one canonic pronunciation per word. Because simply adding all the pronunciation alternatives to the lexicon may result in degradation of recognition performance, appropriate pruning is therefore crucial. Both the probability-based and count-based pruning methods were investigated. Afterwards, the same 30 minutes of data was tested again on the retrained acoustic models.

Call Home Training data						
accent	Total		Male		Female	
Putonghua	Syl/sec	hour	Syl/sec	hour	Syl/sec	hour
	max=13.45		max=13.19		max=13.45	
	min=0.53		min=0.6		min=0.53	
	avg=5.29	6.7	avg=5.57	3.31	agv=5.12	3.39

Table 1. Detailed statistics of the 6.7 hours of data of Putonghua accent used in training.

3. Pronunciation variation at various levels of linguistic units and under different contextual dependency conditions

Because the pronunciation variation apparently has to do with the characteristic structures of the different levels of linguistic units (Initial-Final, syllable, character, word). Such characteristic structures of Mandarin Chinese are first summarized here. Analysis of pronunciation variation with respect to these characteristic structures will then follow.

Chinese is a monosyllabic-structure language. Most morphemes (i.e., the smallest meaningful units [3]) are represented by single syllables. The basic graphic unit, or the character, on the other hand, is always pronounced as a monosyllable. As a result, the overwhelming majority of characters represent single morphemes in Chinese. A Chinese word is then composed of one to several characters. Most characters can also be a mono-character word. In the Chinese writing system, the words (and characters) are connected together one after another in a sentence without any word boundaries (such as the blanks serving as word boundaries in western languages); the reader supplies necessary boundaries as he reading along [4]. In addition, almost every character has its own meaning and can play quite independent role linguistically (since it is a morpheme and can be a mono-character word). As a result, although it seems easy to identify words by native speakers of Chinese, it is in fact difficult to define the words rigorously, even if in principle a word is defined as a unit "which has specific meaning and can be used freely" [4]. For example, there doesn't exist a commonly accepted lexicon including all the commonly accepted words in Chinese, and the segmentation of a sentence into words may be different for different readers. In general, there are more than ninety thousand frequently used words in Chinese and the number of frequently used characters is at least ten thousand. However, there are only about 418 base-syllables (disregarding the tones) or 1360 tonal syllables (including the differences in tones) in Chinese. Consequently, very often several, if not many, different characters share the same syllable, the so-called homophones. There are also many cases of the reverse of homophones, i.e., the homographs, which is the same character but with several pronunciations, for which a character may have different meanings with different pronunciations. Traditional Chinese phonology decomposed the syllable into an Initial and a Final. The Initial is the way a syllable begins, usually with a consonant. A small number of syllables do not begin with a consonant. They are said to begin with the zero Initial. The Final of a syllable is the syllable minus the Initial. The longest form of a Final consists of three parts: an optional medial, or semivowel; a main vowel, or head vowel; and an optional ending [3]. There are about 21 Initials and 36 Finals in Mandarin. Generally speaking, from Initial-Final, syllable, character to word level, the level is shifted from phonological to semantic significance.

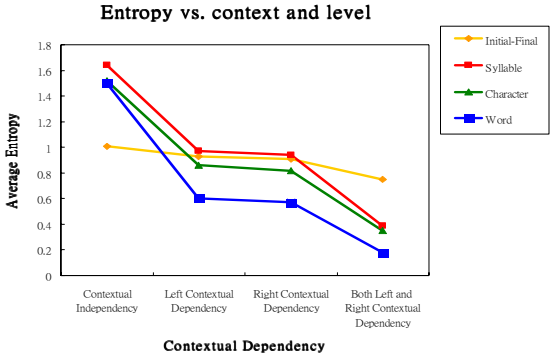

Fig. 1. Average entropy for different levels of linguistic units under different conditions of contextual dependency.

In order to analyze the pronunciation variation with respect to the different levels of linguistic units and contextual conditions, we plotted the average entropy H in equation (2) for each level of linguistic unites with respect to each case of contextual conditions in Fig. 1. Quite many interesting observations can be made in Fig. 1. For example, since each word, character, or syllable is a combination of Initial(s) and Final(s), the average entropy for the Initial-Final level is the lowest compared to those for other levels, under the condition of contextual independency. However, by specifying more contextual dependency conditions, the average entropy is reduced enormously in all other levels except only slightly in Initial-Final level. From an information-theoretic perspective, the reduction of average entropy may arise from the reduction of possible pronunciation variation, which is reasonable when more contextual dependency is specified in each level. In other words, the pronunciation distribution becomes more focused with additional contextual conditions. The reason for the exception for the Initial-Final case, on the other hand, is probably that the pronunciation of most Initials or Finals is not dependent on their neighboring Initial or Final only, but very often cross the character or word boundaries. That may be why the phenomenon is not significant for the Initial-Final case. It is interesting that for the both left and right contextual dependency condition on the right of Fig. 1, the average entropy of the syllable level is very close to that of the character level. This implies that the information (or entropy) carried by a syllable is approached to that of a character if enough contextual information is specified. This is in good agreement with the experiences of processing Mandarin Chinese. For example, even if each syllable is often shared by many homophone characters, the native speakers know exactly which character is referred to by listening to the syllable due to the contextual information. This is also the reason that character-based N-gram language models are able to clarify the ambiguity caused by homophone characters. In Table 2, the syllable ***shi*** may corresponding with several characters. However, by specifying both left and right neighboring syllables, the syllable ***shi*** is then constrained to a specific character. In addition, the average entropies of character and word are comparable in contextual-independent case. This may be due to the fact that many frequently used mono-characters

appear much more frequently than other multi-character words and therefore the former actually dominates in word level entropy under the contextual independency condition.

ㄕ		ㄐㄧㄡ-ㄕ+ㄕㄨㄛ		ㄉㄜ-ㄕ+ㄏㄡ	
LogUnigram	1.36	logUnigram	-2.59	logUnigram	-3.03
Entropy :	1.94	Entropy :	1.49	Entropy :	1.29
canonic :	S Y	canonic :	S Y	canonic :	S Y
S Y	0.24		0.28	S Y	0.23
S	0.08	S	0.10	S	0.16
	0.07	S Y	0.08		0.15
Y	0.03	Y	0.07	Y	0.03
Z Y	0.02	E	0.02	R Y	0.02

Table 2. Examples of pruned pronunciation variation in syllable level with contextual independency condition (left column) and both left and right contextual dependency condition (middle and right columns).

4. Pronunciation variation for different speaking rates and different word frequencies

In order to analyze the pronunciation variation with respect to the speaking rates, the speaking rate used here is determined by dividing the time length between transcribed silences by the corresponding number of syllables in the transcriptions. In addition, for analysis purposes we divide the speech data into four groups based on four different ranges of speaking rates, i.e., less than 4.65 syllables/sec, above 6.15 syllables/sec, etc. The ranges for speaking rates are determined in such a way that almost equal number of syllables are included in each group. The average entropy for different levels of linguistic units at different ranges of speaking rates is plotted in Fig. 2. From Fig. 2, it is obvious that the entropy is higher at higher speaking rates, which is natural because more pronunciation variation occurs at higher speaking rates. However, for the group with the slowest speaking rate (<4.65 syllables/sec), the entropy is comparable to those with middle speaking rates. We found that most hesitating words and utterances for the Call Home conversation are in this group. The pronunciation of these hesitating words and utterances is more uncertain due to their inherent characteristics and the inconsistent transcriptions. This situation is more evident in character and word levels, as can be found in Fig. 2. In addition, it is interesting to note that the average entropy for all the data (not dividing into four groups) as dots at the left of Fig. 2 is higher than those of any of the four groups of speaking rates at either the syllable, character or word levels. In other words, simply merging the four different groups with different statistical properties into one group will result in higher entropy, which is reasonable. The difference becomes limited for the Initial-final level, since in this case the four different groups behave quite similarly.

In Fig. 3, the probabilities of model substitution or deletion (P_{sub} and P_{del}) are plotted for the four groups of different speaking rates. It can be found that both P_{sub} and P_{del} increase as the speaking rate increases, which is also intuitively reasonable. The tendency is more significant for P_{del}, which again agrees with the fact that fast speaking rates tend to coincide with significant phonological reduction [2].

Furthermore, Linguists have recognized that word frequency affects the perception and production of phones. From an information-theoretic perspective, speakers would tend to preserve the most information in speech by limiting the pronunciation variation in the least informative words – that is, the "words" most predictable from context [2]. However, different from English, as mentioned previously the "words" in Chinese are not clearly identifiable. Conventionally, the word predictability is determined by the N-gram parameters for the words. We therefore divide all the data into seven categories according to the log unigram probabilities for each level of linguistic units, and plot the entropy for pronunciation variation for word, character, syllable and Initial-Final for these seven categories in Fig. 4. It can be found from Fig. 4 that more frequently spoken Initials, Finals, syllables and characters are more uncertain relatively, which is in good parallel with the principles mentioned above. As for the word level, there is a similar trend when the log unigram is below –1.75. Whereas, as can be seen in Fig. 4, the entropy decreases contrariwise in the category of the most frequently spoken words. A possible explanation for this phenomenon is given below. Generally speaking, speakers tend to preserve the most information by limiting the pronunciation variation in the least informative units, which are most predictable from the context. However, in the word level, a character(syllable) in multiple-character(syllable) words is more predictable by referring to its neighboring characters(syllables) in the same words. Consequently, in the word level, more frequently spoken multiple-character words may tend to suffer more pronunciation variation. However, in Chinese conversation the category of most frequently used words include quite many mono-character words, while many of the frequently spoken multiple-character words are in the second most frequently used category of Fig. 4, i.e., with log unigram probabilities from –2.25 to –1.75. This may be why the entropy decreases contrariwise in the category of most frequently spoken words in Fig. 4. In this regard Chinese language is again different from English [2].

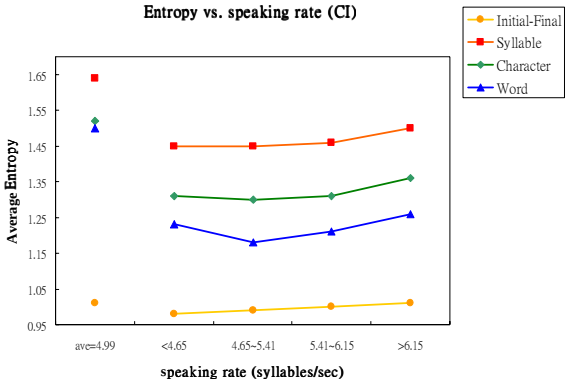

Fig 2. Average entropy for the different levels of linguistic units (Initial-Final, syllable, character and word) for all the data (the most left dots) and four different speaking rates under the contextual independency condition.

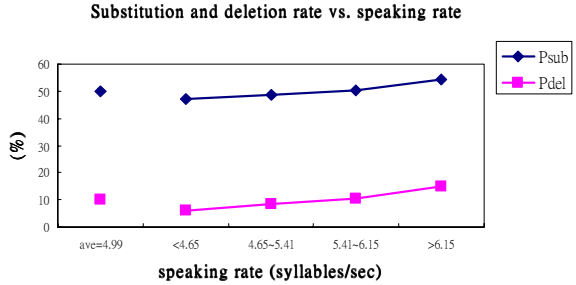

Fig 3. Probabilities of model substitution and deletion (P_{sub} and P_{del}) for four different speaking rates and for all data (the most left dots) under the contextual independency condition.

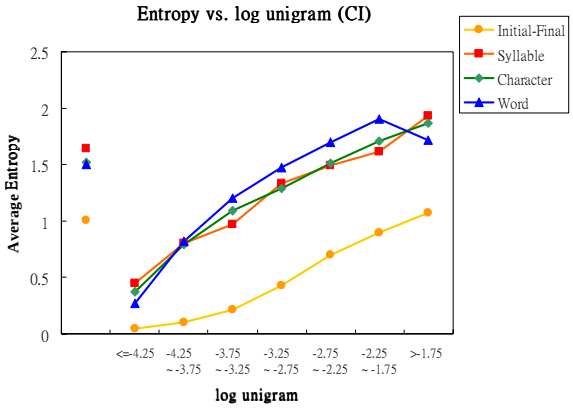

Fig 4. Average entropy for different levels of linguistic units (Initial-Final, syllable, character and word) for all data (the most left dots) and seven categories of different linguistic unit frequency (or log unigram probability) under contextual independency condition.

5. Applications with improved recognition accuracy

Here we performed the recognition tests with the other 30 minutes of data from LDC Mandarin Call Home corpus with the same accent as in the 6.7 hours of training data. The canonic word lexicon used in the baseline consisting of 7619 words, most of which were present in the 6.7 hours of training data. In this canonic lexicon each word has only one pronunciation. In order to compare the performance between two different pruning methods, the average numbers of pronunciation per word are tuned to be comparable in the two methods. In the count-based pruning method, the number of pronunciations for each word to be added into the lexicon was 0.75 (this number had been tuned empirically) times of the log Unigram of each word. In probability-based pruning method, on the other hand, for each word the pronunciations with priori probabilities less than 0.5 (this number had been tuned empirically) times of P_{max} , the probability for the most probable pronunciation, were not be added into the lexicon. The results are listed in Table 3.

	Average number of pronunciation per word	**character accuracy**
Baseline	1	*24.2*
Count-based pruning	1.009	*24.87*
probability-based pruning	1.007	*25.81*
probability-based pruning plus retrained acoustic models	1.007	*26.27*

Table 3. Character accuracy for various recognition tests considering pronunciation variation

As can be observed in Table 3, the count-based pruning did not perform as good as the probability-based method. This is in fact in good agreement with the results in Fig. 4, i.e., unlike in English, to determine the allowed pronunciation variation for a word simply by its frequency (or log unigram) may not be appropriate, because the entropy for word does not always increase with the log unigram. Finally, the augmented lexicon with pronunciation variation selected by the probability-based pruning method was used to align the 6.7 hours of training data, and the acoustic models are retrained on this new transcription. Afterwards, the same 30 minutes of evaluation data was again tested on the retrained acoustic models with the same augmented lexicon. The results in the last row of Table 3 show improved recognition accuracy when the pronunciation variation was handled in this way.

6. Conclusion

This paper presents pronunciation variation analysis with respect to various linguistic units and contextual conditions for Mandarin Chinese. There are at least word, character, syllable, Initial-Final levels of linguistic units in Chinese, each carrying different levels of information with complicated correlations among them. It was found that by specifying more contextual conditions the average entropy for pronunciation variation is reduced enormously almost in each level of linguistic units. It is also found that higher entropy of pronunciation variation is in accordance with higher speaking rate, except for the most hesitation words which are spoken with relatively slow speed. Also, more frequently spoken units usually appear with more pronunciation variation, but this is not always the case for frequently used words. This is different from English. The recognition tests indicated reasonable improvements in accuracy if the pronunciation variation was properly handled, especially when the various characteristics for Mandarin Chinese were well considered.

References

[1] Helmer Strick and Catia Cucchiarini, “Modeling pronunciation variation for ASR: A survey of the literature”, page 225-246, Speech Communication 1999.

[2] Eric Fosler-Lussier and Nelson Morgan, “Effects of speaking rate and word frequency on prounciations in conversational speech”, page 137-158, Speech Communication 1999.

[3] Yuen Ren Chao, “ A grammar of spoken Chinese”, University of California Press 1965.

[4] Jerry Norman, “Chinese”, Cambridge University Press 1988.

[5] Eric Fosler-Lussier and Gethin Williams, “Not just what, but also when: Guided automatic pronunciation modeling for Broadcast News”, DARPA Broadcast News Workshop 1999, Herndon, VA.

[6] Lui Yi and Pascale Fung, “Rule-based word pronunciation networks generation for Mandarin speech recognition”, ISCSLP 2000.

[7] Lui Yi and Pascale Fung, “Modeling pronunciation variations in spontaneous Mandarin speech”, ICSLP 2000.

[8] Chao Huang and Eric Chang, “Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition”, ICSLP 2000.

[9] Fu-Hua Liu, Michael Picheny,... “Speech recognition on Mandarin Call Home: a large-vocabulary, conversational, and telephone speech corpus”, page 157-160, ICASSP 96.