A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese "Regionalects"

Ren-Yuan Lyu [1, 2], Dau-Cheng Lyu [1, 3], Min-Siong Liang[1], Min-Hong Wang [2], Yuang-Chin Chiang[4], Chun-Nan Hsu[3]

1. Dept. of Electrical Engineering, Chang Gung University, Taoyuan 333, Taiwan
2. Dept. of Computer Science and Information Engineering, Chang Gung University, Taoyuan 333, Taiwan
3. Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
4. Institute of Statistics, National Tsing Hua University, Hsin-chu 300, Taiwan
E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967

Abstract

In this paper, a new approach is proposed for recognizing speech of mutually unintelligible spoken Chinese regionalects based on a unified three-layer framework and a one-stage searching strategy. This framework includes (1) a unified acoustic model for all the considered regionalects; (2) a multiple pronunciation lexicon constructed by both a rule-based and a data-driven approaches; (3) a one-stage searching network, whose nodes represent the Chinese characters with their multiple pronunciations. Unlike the traditional approaches, the new approach avoids searching the intermediate local optimal syllable sequences or lattices. Instead, by using the Chinese characters as the searching nodes, the new approach can search to find the globally optimal character sequences directly. This paper reports the experiments on two of the Chinese regionalects, i.e., Taiwanese and Mandarin. Results show that the unified framework can efficiently deal with the issues of multiple pronunciations of the spoken Chinese regionalects. The character error reduction rate is 34.1%, which is achieved by using the new approach compared with the traditional two-stage scheme. Furthermore, the new approach is shown more robust when dealing with the poor uttered speech database.

1. Introduction

There are more than a billion people who are considered to be speakers of the Chinese language. Due to the historical and geographical reasons, there are vast varieties within the spoken Chinese language. The varieties of spoken Chinese are usually referred to as "dialects". However, unlike dialects in the other languages, which are usually mutually intelligible to each other, the "dialects" in spoken Chinese are almost completely mutually ***unintelligible***. These Chinese "dialects" include Mandarin (also called Putonghua, or Guoyu), Wu, Yue, Xiang, Gan, Hakka, Southern Min, and Northern Min. Due to their mutual unintelligibility, many linguists prefer to call all of these Chinese "dialects" as "languages". However, since these "languages" share the same written system, the Chinese characters (also called Hanzi), it seems that they are not so different as two distinct western languages, like English and French. So a linguistic term called "***regionalect***", whose meaning lies between dialect and language, has ever been proposed to call all these mutually unintelligible Chinese "dialects" and has been adopted in this paper. [1]

One of the purposes of speech recognition is to convert the acoustic representation of spoken language into the text representation of written language. As to the Chinese language, although Mandarin speech has been promoted to be the "Common Language" or "National Language", the other regionalects will still exist for a long time. In Taiwan, it is even possible to promote Taiwanese, which evolved from Southern-Min regionalect, to be another official language. It's necessary for a speech recognition system to be able to recognize all the regionalects and convert them to the Chinese written form encoded in machine readable codes, e.g., Big5 code, GB code, or Unicode, …, etc. In the past decades, most of the speech recognition effort in the Chinese societies has been made for Mandarin speech [2][3]. Relatively few researches were reported about the other regionalects [4] [5]. In this paper, we consider two regionalects of the Chinese language, i.e., Mandarin and Taiwanese, simultaneously within the same framework of speech recognition. Due to its success, we prospect this framework will still work well for combinations of all the other Chinese regionalects.

This paper is organized as follows. In section 2, we describe the problems in the multiple-regionalect Chinese speech recognition. In section 3, the new proposed framework is described in detail. In section 4, we describe the multiple-pronunciation lexicon and the acoustic models. After conducting experiments in section 5, the paper ends with a summary of the results and conclusions in section 6.

2. Problem Description

As shown in <fig.1>, the problem of Chinese speech recognition can be looked upon as to decode the acoustic feature sequence X to be the optimal Chinese character sequence C*. Traditionally, the optimal syllable sequence or lattice S* was chosen as an intermediate output and the decoding problem can be described as the following equations:

$$S^*(X) = \arg\max_S P(S \mid X) \quad (1)$$

$$C^*(S^*) = \arg\max_C P(C \mid S^*) \quad (2)$$

In such a conventional Mandarin Chinese speech recognition system, the syllable decoding can be usually implemented by searching in a 3-layer network composed of an acoustic model layer, a lexical layer, and a grammar layer as shown in <fig.2>. When the optimal syllable sequence or the syllable lattice was obtained after syllable decoding, another syllable-to-character converter should be followed to deal with the issues of homonyms to get the Chinese character string as the final output, which can be shown as in <fig.3>.The above framework worked well and has been adopted for Mandarin Chinese speech recognition for a long time. However, it is hard to generalize to incorporate the other Chinese regionalects, e.g., Taiwanese or Hakka. In Taiwan, most people can speak both Mandarin and Taiwanese. In many situations, they speak a sentence in mixing both languages. This is also the case for people living in Southern China, where people speak both Mandarin and their own specific regionalects. For a speech recognizer to incorporate at least 2 regionalects, it seems obvious that one can still adopt the above framework by only extending the system by incorporating more acoustic models, more items in the lexicon and more paths in the grammar. Such a straightforward approach will suffer from the following difficulties:

1. It is hard to generate all instances for the syllable network when we want to recognize two or more regionalects in

spoken Chinese. If we really do this, it will not only increase the unnecessary searching space, but also cost the decoding time too much.

2. It is not trivial to generate the multiple pronunciation lexicons efficiently.
3. The language model for mixed language is hard to estimate.
4. When new acoustic features like tones were considered to be added to the system, all 3 layers in the syllable decoding and the syllable-to-character converter should be modified. This is not a trivial task at all.

To tackle with all the above difficulties, a unified framework for multi-regionalect Chinese speech recognition was proposed and described in the following section.

3. A Unified Framework for multi-regionalect Chinese Speech Recognition

Unlike the conventional approach, which divides the recognition task as syllable decoding and character decoding, the new proposed approach adopts a one-stage searching strategy as shown in <fig.4>, which decodes the acoustic feature sequence X directly to the desired character sequence C*, no matter what regionalects are spoken. The decoding equation can thus be shown as follows:

$$C^*(X) = \arg\max_C P(C \mid X) \qquad (3)$$

In such a new proposed framework, the character decoding can be implemented by searching in a 3-layer network composed of an acoustic model layer, a lexical layer, and a grammar layer as shown in <fig.5>. In this framework, there are at least 2 critical differences from the conventional one. One is in the lexicon layer, where the new framework adopts the character-to-pronunciation mapping which can easily incorporate the multiple pronunciations caused by multiple regionalects or even multiple "languages", including Japanese, Korean and even Vietnam, which also use the Chinese characters more or less. Another one is in the grammar layer, where the character is adopted as the nodes of the searching net. This makes it be regionalect ***independent***! By the way, tone is a common feature for all Chinese regionalects. In the unified framework described above, it is easy to incorporate the tone feature into the system. There is no need to change the grammar layer.

4. Acoustic Modeling & Multiple Pronunciation Lexicon

For validating the new proposed frame work, a large-vocabulary bi-regionalect speech recognition system was constructed. First of all, a 42-dimension feature vector, consisting of MFCCs, their derivatives, and tonal parameters, was used as the feature vector [6]. In the acoustic modeling, we adopted the Initials and tonal context-independent Finals as the HMM models. The same phoneme (including tone) symbols in the transcription in both regionalects share the same training data.

There are two approaches to deal with pronunciation variations, i.e., the knowledge-based approach and the data-driven approach.[7][8] The former consists of generating variants by using phonological rules, and the later consists of performing phone recognition to obtain information on the pronunciation variation in the data. We adopt the rule-based approach to deal with the issues of Taiwanese tone sandhi, and the pronunciation variation between spontaneous and read speech. In another way, using data-driven approach based on confusion matrix for finding syllable mapping between the real pronunciations and canonical pronunciations. In knowledge-based approach, we use a statistical technique to build a mapping from each character to its multiple pronunciations by using the Formosa Lexicon [9]. Then every character has a reliable probability mapping to possible pronunciations. In the data-driven approach, the syllable confusion matrix is constructed by a dynamic-programming technique to align the recognition results of an evaluation data set. We chose the most variational pronunciations by the confidence measure which is the occurrence possibility, and eliminate the relatively small counts. The sum of all the occurred possibility of each character is then normalized to unity for fair competition in the Viterbi search.

5. Experiment Results

5.1 Corpus and Task Description

Some information about the speech corpus used in the experiments is listed in Table 1. All the speech data were recorded using close-talk microphones in normal office environments.

The corpus was divided into training and testing sets. The training data set and one of the two testing data set were designed to be phonetically abundant. The other testing data set is a specially designed task with several thousand words derived from poems of the Tang Dynasty. All the statistical information for the whole speech database were also listed in Table1. The transcription levels for the training data sets and one of the two testing data sets are syllables. This means that for each utterance, there is a syllable sequence associated with it. Speakers were required to speak each utterance following the prompt of the syllable sequence. The other testing data set, i.e. the specially designed task, was further divided into 2 subsets for each of both Mandarin and Taiwanese according to different transcription levels. For those data sets with character level transcription, only Chinese character sequences were provided to the speakers when they were recording the speech data. Without the syllable level transcriptions, speakers spontaneously utter each sentence as correct as their literacy can achieve. In the worst case, the code switching phenomenon may occur, which means speakers change the language during uttering a sentence. Take the sentence as an example "美國總統布希先生 ". When pronounced in Mandarin, it will be /*mei3 guo2 zong2 tong3 bu4 si1 sian1 sheng1*/; if pronounced in so-called "standard" Taiwanese, it will be /**bhi1 gok2 zong1 tong4 bo4 hi2 sen2 sinn1**/. In the Testing data 2, it would be very probably pronounced as / **bhi1 gok2 zong1 tong4** *bu4 si1 sian1 sheng1*/**,** where the first 4 syllables are Taiwanese and the last 4 syllables are Mandarin.

For convenience of reference in this paper, each of all data sets was given a name, e.g., M100-PA represents the Mandarin training data set of 100 speakers, designed as phonetically abundant.

5.2 Experiment Setup

HMM acoustic models:

HMM-based acoustic models for two regionalects, namely Mandarin (M) and Taiwanese (T), and the mixing bi-regionalect (B) are trained by HTK tool [10]; using training data set M100-PA, T100-PA and B100-PA, respectively. The acoustic units chosen here were right-context dependent mono-initials and tonal-finals. Since all the transcriptions are based on the Formosa Phonetic Alphabet (ForPA)[9], the speech data labeled as the same acoustic unit in both regionalects are shared to train the Gaussian mixtures for that unit. The HMM for each acoustic unit has 3 states and each state has 2 to 8 Gaussian mixtures dependent on the occurrence of the training data for that state.

Pronunciation lexicons:

In this paper, each character in the lexicon contains not only the multiple pronunciations but also their associated probabilities. The probability model was estimated in two ways. One is based on the equally probable distribution of pronunciations for each character in the Formosa Lexicon. The other was based on the tonal syllable confusion matrix which was constructed from the recognition results of the testing data set B20-PA using a traditional

syllable decoding. By combining the two models with well adjusted weights, the final probability model was obtained. The lexicon densities for Mandarin, Taiwanese and Bi-regionalect are 1.08, 1.52 and 2.5, which represent the expected numbers of pronunciations that a Chinese character could have.

Searching nets:

There are 2 types of searching nets used in the experiments. One is the free-syllable net which was built by connecting each syllable to all syllables in a language as shown in fig<6>. The perplexity of the free-syllable net equals to the number of the distinct monosyllables in the language. For the cases in Mandarin, Taiwanese, and mixing bi-regionalect, the perplexities are 408,709 and 925, respectively. We named those 3 searching nets as M408, T709 and B925 for convenient reference. The other searching net used in this paper is the lexicon tree, which was constructed by concatenating all the nodes, which may represent syllables or characters, to a tree data structure like the syllable tree net shown in fig<2>.

Baseline Experimental Results:

By using the experimental setup described above, we achieved the syllable error rates (SER) listed in Table 2, where there are 3 acoustic models(AM), namely Mandarin (M), Taiwanese (T) and Bi-regionalect (B), trained by M100-PA, T100-PA and B100-PA respectively. The free-syllable nets(FSN) M408, T709, and B925 were obtained for Mandarin with 408 base syllables, for Taiwanese with 709 base syllables and for Bi-regionalect with 925 base syllables. Their perplexities are also shown as 408, 709, 925 respectively. The SER can be shown as the last row in Table2, where we see that the SER for testing data set M10-PA is 37.1% and for M10-ST is 60.3% by using Mandarin(M) acoustic model. There are some points worth to point out as follows:

First of all, B acoustic model is always better than M or T acoustic model. This is because B acoustic model used more training data. For examples, under the same perplexity(709), using the same testing data (T10-PA), the SER of B acoustic model is 36.9%, which is better than the SER 37.9% of the M acoustic model, and so as the others in Table2.

Secondly, the SER increase insignificantly when testing datasets were changed from M10-PA, T10-PA to B20-PA as long as the B acoustic model was used, even when the perplexity is increasing significantly from 408, 709 to 925. The reason might be that the B acoustic model uses the training data which has the same phoneme symbols in the transcription in both regionalects, and is thus more robust.

Thirdly, the SER increases very fast when testing data sets are changed to M10-ST, T10-ST and B10-ST, which consists of spontaneous speech data without syllable level transcription prompts provided to speakers. Thus the degree of consistency of the utterance and its associated syllable transcription is relatively low. This leads to the high SER. As shown in Table 2, the worst case is 77.3% when T10-ST data set was used.

5.3 The Two-stage Scheme v.s. The Unified Framework

We have already noted that the performance by using B acoustic model is better than that by using M or T acoustic model. Therefore we just used the B acoustic model to do the following experiments. In order to justify the unified framework proposed in this paper, several experiments of large vocabulary speech recognition have been performed to compare the recognition performance with that of the traditional two-stage scheme. Instead of the free syllable net, here we use a 15-thousand-word tree net for each of the two regionalects and a 30-thousand-word tree net for the Bi-regionalect, namely M15k, T15k, and B30k respectively. Furthermore, for a typical special domain task, where several thousand words are enough, we chose sentences from Tang poems where there are 3223 short sentences, each of which contains 5 or 7 Chinese characters. We have searching nets M3223, T3223, and B6446 for Mandarin, Taiwanese, and Bi-regionalect respectively. The SER and character error rate (CER) were obtained for the traditional Two-stage scheme. Only CER were available for the new proposed unified framework. All the recognition results were listed in Table 3, where we have to pay hood to some points as follows.

First of all, when the searching nets were changed from free syllable nets to lexicon trees, the SER decreases very significantly. For example, the SER decreases from 39.8% (Table 2, B925) to 16.1% (Table 3, B30k) for the same testing data set B20-PA. This error reduction is mainly due to perplexity reduction, so as the decrease from 70.8% (Table 2, B925) to 24.6% (Table 3, B6446) for the same testing data set B10-ST.

Secondly, when the Two-stage scheme was used, the second stage, i.e., the character decoder, will probably increase the error rate when the vocabulary size is large. This is because of the increase of the number of homonyms. However, when the vocabulary size is small, the second stage usually decreases the error rate. For example: for B20-PA testing data in Table 3, error rate increased from 16.1% (SER) to 23.9% (CER); on the other hand, for B10-ST testing data set, error rate decrease from 24.6%(SER) to 21.7% (CER).

Thirdly, when the unified framework was used, the CER decreased from 23.9% to 22.6% for B20-PA and from 21.7% to 14.3% for B10-ST. The CER reduction is from 5.9% and 34.1% respectively. Those results justify the newly proposed Unified framework is indeed better in recognition performance than the traditional Two-stage scheme, especially in the case when the speech recognition system is used for the non-mainstream regionalects.

By observing Table3 in more details, we can find some interesting facts which reveals some phenomena of the language reality in Taiwanese and further justify the new approach: The CER of T10-ST(for Taiwanese) is more than 6 times of the M10-ST(for Mandarin) in both the Two-stage scheme and the Unified framework. This is because many people could not correctly utter the character strings for Taiwanese due to the lack of Taiwanese literacy, which is because of the ignorance of Taiwanese and over emphasis of Mandarin speech in elementary education for past several decades. In addition, it's reasonable that the recognition performance in read speech data set (B10-ST-Syl) is better than that in the spontaneous data set (B10-ST-Chr) for both frameworks. Nevertheless, the increase of CER from read speech to spontaneous speech in the Unified framework (1.6%) is lower than that in the Two-stage scheme (2.1%). This evidence proves again that the new approach is more robust when dealing with poor uttered speech database.

6 Conclusion and Future Work

In this paper, we have reported a new framework based on a unified framework with one stage searching strategy by using a multiple pronunciation lexicon and a large vocabulary searching net with Chinese characters as its nodes. The lexicon is generalized by both data-driven and knowledge-based statistical approaches. This framework shows its validity and efficiency to deal with the Mandarin/Taiwanese bi-regionalect speech recognition at a unified framework. Experiments to prove its validity for the other regionalects of spoken Chinese are under the going in our lab. The results are expected to be as good as the one reported in this paper.

7. References

[1]. John DeFrancis, "The Chinese Language - Fact and Fantasy", 1984, ISBN 0-8248-1068-6, p. 53~p.67

[2]. Mingkuan Liu, Bo Xu, Taiyi Huang, Yonggang Deng, Chengrong Li, "Mandrain Accent Adaptation Based on

Contest-Independent/Context-Depent Pronunciation Modeling," In Proc.ICASSP, 2000
[3]. Guoliang Zhang, Fang Zheng and Wenhu Wu, "A Two-Layer Lexical Tree Based Beam Search in Continuous Chinese Speech Recognition," In Proc. EuroSpeech, Denmark, September 2001
[4]. Tan Lee, Wai Lau, Y. W. Wong and P.C. Ching, "Using tone Information In Cantonese Continuous Speech Recognition," ACM Transactions on Asian Language Information Processing, Vol. 1, pp. 83 - 102, 2002
[5]. Dau-Cheng Lyu, et al., "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Twaiwanse/Mandrain Continuous Speech Recognition," In Proc. SST 02, Melbourne, December 2002
[6]. Dau-Cheng Lyu, et al, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling" In Proc. EuroSpeech, Switzerland, 2003.
[7]. Mirjam Wester, "Pronunciation Modeling for ASR-knowledge-based and Data-drived Methods," Journal of Computer Speech and Language 17(2003), pp. 69-85, 2003
[8]. Liu, Yi and Pascale Fung, "Partial change accent models for accented Mandarin speech recognition." In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, December, 2003.
[9]. Liang M.S., R.Y. Lyu, Y.C. Chiang "An efficient algorithm to select phonetically balanced scripts for constructing corpus" NLP-KE, Beijing 2003
[10]. Steve Yang et al. Hidden Markov Model Toolkit V3.1, Cambridge University Engineering Department, 2002

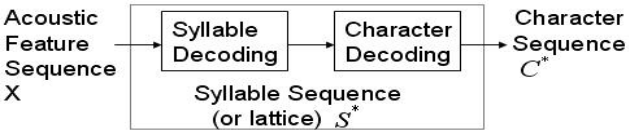

Figure 1: the decoding problem of Chinese speech recognition

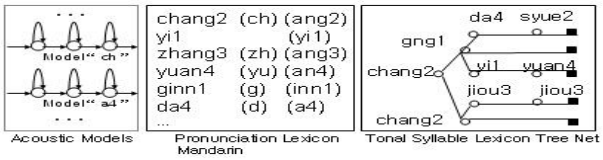

Figure 2: a 3-layer grammar searching net for syllable decoding

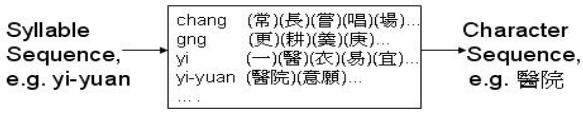

Figure 3: the syllable-to-character converter

Figure 4: one-stage searching strategy for Chinese speech recognition

Figure 5: a Unified 3-layer framework for multi-regionalect Chinese speech recognition

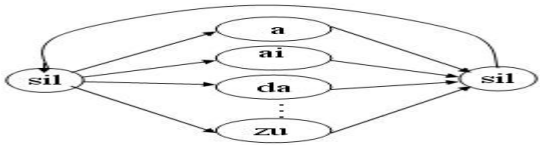

Figure 6: a Free syllable net

Testing Data	M10-PA / M10-ST		T10-PA / T10-ST		B20-PA / B10-ST
AM	M	B	T	B	B
FSN	M408	M408	T709	T709	B925
Perplexity	408	408	709	709	925
SER [%]	37.1 / 60.3	36.9 / 58.6	37.9 / 77.3	36.9 / 75.2	39.8 / 70.8

Table 2: Syllable Error Rate(SER) [%] Results for Free Syllable Net(FSN) Decoding, using different Acoustic Model (AM)

Testing data	B20-PA	B10-ST	M10-ST	T10-ST	B10-ST-Syl	B10-ST-Chr
AM	B	B	B	B	B	
Lexicon Tree	B30k	B6446	M3223	T3223	B6446	
Voc. Size	30k	6446	3223	3223	6446	
SER[%]of Two-Stage	16.1	24.6	5.6	37.1	26.9	22.2
CER[%]of Two-Stage	23.9	21.7	5.6	40.7	20.6	22.7
CER[%]of Unified	22.6	14.3	5.0	34.8	13.5	15.1
CER-R[%]	5.9	34.1	10.7	14.5	34.5	33.4

Table 3: Comparison of recognition performance for the two-stage and the unified framework in Syllable Error Rate (SER), Character Error Rate (CER) and CER reduction (CER-R) from the two-stage to the unified framework.

	Training Data Phonetically Abundant		Testing Data 1 Phonetically Abundant		Testing Data2 Special Task (Tang Poems)			
Regionalect	Mandarin	Taiwanese	Mandarin	Taiwanese	Mandarin	Taiwanese	Mandarin	Taiwanese
No. of Speakers	100	100	10	10	10		10	
No. of Utterances	43078	46086	1000	1000	250	250	250	250
No. of Hours	11.3	11.2	0.28	0.28	0.14	0.14	0.13	0.14
No. of Syllables per Utterance	2.7	1.9	2.5	2.6	5.9	5.9	5.9	5.9
Transcription Level	Syllable	Syllable	Syllable	Syllable	Syllable	Syllable	Character	Character
Speaking Style	Read	Read	Read	Read	Read	Read	Spontaneous	Spontaneous
Names of data sets for reference in this paper	M100-PA	T100-PA	M10-PA	T10-PA	M10-ST-Syl	T10-ST-Syl	M10-ST-Chr	T10-ST-Chr
	B100-PA		B20-PA		B10-ST-Syl		B10-ST-Chr	

Table 1: Information about the speech corpus