Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling

Dau-Cheng Lyu [1,3]*, Min-siong Liang* [1]*, Yuang-chin Chiang*[3]*, Chun-Nan Hsu*[2]*, Ren-Yuan Lyu* [1,2]

[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: 886-3-2118800ext5967

Abstract

A large vocabulary Taiwanese (Min-nan) speech recognition system is described in this paper. Due to the severe multiple pronunciation phenomenon in Taiwanese partly caused by tone sandhi, a statistical pronunciation modeling technique based on tonal features is used. This system is speaker independent. It was trained by a bi-lingual Mandarin/Taiwanese speech corpus to alleviate the lack of pure Taiwanese speech corpus. The searching network is constructed based on nodes of Chinese characters and results in the direct output Chinese character string. Experiments show that by using the approaches proposed in this paper, the character error rate can decrease significantly from 21.50% to 11.97%.

1. Introduction

For the past decades, Mandarin has been the most widely studied Chinese dialects in speech recognition community due to its huge spoken population. This situation is easy to understand because Mandarin is the official language in major Chinese societies, including Mainland China and Taiwan. However, several important regional dialects other than Mandarin are still widely used in daily lives in all Chinese societies. These dialects, unlike dialects in any one of the other western languages, are mutually unintelligible to each other. In some linguistic viewpoints, these dialects can be even looked upon as different languages just as those in Europe.

In Taiwan, the most widely used dialect (or language) second to Mandarin is Taiwanese. It is the mother tongue of more than 75% of the population in Taiwan. It belongs to a larger Chinese dialectical family called Min-nan (or Southern-Min, Southern-Hokkian), which is also used by many overseas Chinese living in Singapore, Malaysia, Philippine, and other areas of Southern-East Asia. It was estimated that this language has more than 49 millions speakers and is ranked in the 21th place in the world [10].

In this paper, we are concerned about constructing a speech recognition system for Taiwanese. According to some linguists, the distinction between Mandarin and Taiwanese is much more than that between usual dialects in the same language. They can actually be looked upon as two different languages, just like French and English, from some linguistic viewpoints. However, we can still find much similarity between Mandarin and Taiwanese in phonetic, lexical or even syntactic level. Due to this similarity, it is natural for us to utilize speech or text corpus from both languages to help construct a speech recognizer for the Taiwanese speech.

Taiwanese, like Mandarin, is a tonal language. The pitch information, which is usually ignored in western languages, is very significant to help understand the meaning and discriminate homonyms. In this paper, we try to use the pitch information as another feature in addition to the widely used mel-frequency cepstrum (MFCC).

Another important issue in Taiwanese speech recognition is the severe problems of multiple pronunciations of each morpheme. Here a morpheme maybe a Chinese character (usually called hanzi in China, or kanji in Japan). A statistical pronunciation modeling technique was shown to be very helpful to conquer the issue of multiple pronunciations.

This paper is organized as follows: in section 2 we introduce background knowledge including Taiwanese phonetics/linguistics knowledge essential to speech recognition, and the speech corpus that was used to train acoustic models. In section 3, we build a baseline system by using a sub-syllabic CHMM modeling approach. Then we focus on issues of pitch tracking and multiple pronunciations in section 4. Finally, experimental results and conclusion are presented in section 5 and section 6.

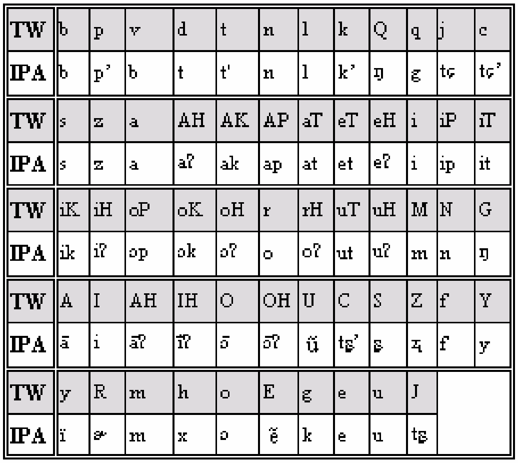
2. Background

2.1. Taiwanese Phonetics/Linguistics Essential to a Speech Recognition System

Taiwanese, like Mandarin as a member of Sino-Tibetan language family, is a tonal, syllabic language. Each Taiwanese sentence can be looked upon as a string of words. Most Taiwanese words can be written in the form of Chinese characters (usually called Hanzi in China or Kanji in Japan), but there are still many living words without commonly accepted written forms. For those without written forms of Chinese characters, each word is usually represented as a string of English characters to form one to several syllables. Phonetically speaking, each Taiwanese word can be composed of one to several syllables. Each syllable, carrying one particular lexical tone, can be further decomposed into an optional Initial and a Final. An Initial is just a consonant phoneme, while a Final may be a vowel, a vowel plus a nasal consonant, or a vowel plus stop consonant. There are about 18 Initials, about 90 Finals, and 7 lexical tones in Taiwanese. [2]. These basic phonetic units can be further combined to form a set of ***about 2000*** tonal syllables and ***about 800*** base-syllables, which were counted from ***58270 words*** of the electronic lexicon available to us. The set of 18-Initials/90-Finals can be transformed to a set of ***35 distinct phonemes***, if we divide each Final into one to several phonemes. All the phonemes are listed in <table.1>.

Furthermore, Taiwanese is a tonal language with more complex tonal structures than that of Mandarin. It has 7 lexical tones, two of which are carried in syllables ending with final /-p, -t, -k, -h/ (called entering-tone traditionally) and the other five are carried in those ending without final /-p, -t, -k, -h/ (called non-entering tone traditionally). The tone sandhi issue is relatively complex in Taiwanese. Usually each syllable (may corresponding to one Chinese character if it exists) has 2 kinds of tones called base-tone and sandhi-tone depending on the position it appears in words or sentences. One of the most frequently used sandhi rules is the so-called " tone sandhi cirle".[6] It says that if a syllable appears at the end of a sentence, or at the end of a word, then it is pronounced as its base-tone. In most of the other cases, however, it is pronounced as its sandhi tone.

Another linguistic/phonetic issue about Taiwanese is the phonetic variations of people whose ancestors came from different sub-dialectic areas in China. These main sub-dialects, however, are mutually intelligible and most people in Taiwan even don't realize which sub-dialect they are using. One more systematic phonetic variation occurs in the different situations of the language being used. When Taiwanese is used to read the classical literature like poetry, it uses the classical pronunciation system, otherwise it use the oral pronunciation system in daily lives.



TW	b	p	v	d	t	n	l	k	Q	q	j	c
IPA	b	p'	b	t	t'	n	l	k'	ŋ	ɡ	tɕ	tɕ'
TW	s	z	a	AH	AK	AP	aT	eT	eH	i	iP	iT
IPA	s	z	a	aʔ	ak	ap	at	et	eʔ	i	ip	it
TW	iK	iH	oP	oK	oH	r	rH	uT	uH	M	N	G
IPA	ik	iʔ	ɔp	ɔk	ɔʔ	o	oʔ	ut	uʔ	m	n	ŋ
TW	A	I	AH	IH	O	OH	U	C	S	Z	f	Y
IPA	ã	ĩ	ãʔ	ĩʔ	ɔ̃	ɔ̃ʔ	ũ	tʂ'	ʂ	ʐ	f	y
TW	y	R	m	h	o	E	g	e	u	J		
IPA	ɨ	ɚ	m	x	ɔ	ẽ	k	e	u	tʂ		

<table.1> The phone set for both Mandarin and Taiwanese. TW : the Twbet repersentation of phone. IPA : the Interational Phonetic Alphabet.

2.2. Bi-lingual Mandarin/Taiwanese speech Corpus

Due to the similarity of Taiwanese and Mandarin, it's natural for us to use speech corpus available for both languages. A bi-lingual Mandarin and Taiwanese microphone speech corpus consisting of 120 speakers, collected in the Multi-media Signal Processing Laboratory of Chang Gung University in Taiwan was used. The important statistics of the corpus are listed in <table.2>. It is divided into 2 sets, one for training and the other for testing. The set of training data composes of four subsets, including 11 hours of Taiwanese and 11.3 hours of Mandarin speech, uttered by 100 speakers. It totally includes 92160 utterances. We chose another 20 speakers' speech to form the testing data sets, including 34 minutes of speech, consisting of 2000 words.

	Training data (100)				Test data (20)	
	M		T		M (10)	T (10)
	Sub-Set 1	Sub-Set 2	Sub-Set 3	Sub-Set4		
Total # of syllables	1287	7790	2883	6064	2468	2547
Total # of phrases	1728	3157	2883	1913	1000	1000
Average word length	1	2.5	1	3.2	2.5	2.5
Total duration	2.3hr	9.0rh	5.2hr	5.8hr	17 min	17 min

<table.2> The statistics of the corpus, M for Mandarin and T for Taiwanese

2.3. Phonetic Inventory

In this paper, we use a single phonetic set to cover all the sounds of Taiwanese and Mandarin speech. The phonetic transcription system called Taiwan Phonetic Alphabet (TWbet) was designed to transcribe all languages spoken in Taiwan. Sounds which are represented by the same TWbet symbol share one common phoneme category. We have a set of 429 base syllables for Mandarin and a set of 825 base syllables for Taiwanese. There are 183 common base syllables for both languages, and thus a set of 1049 distinct base syllables are obtained when considering both languages simultaneously. Similarly, there are 34 phones for Mandarin and 52 phones for Taiwanese. A set of 58 phones are necessary to transcribe both languages. Despite their similarity in basic syllable structure, Mandarin and Taiwanese have fairly different phonological rules and phonetic composition. For example, in Taiwanese, there are additional Nasalized Vowels and Stop Vowels, which do not exist in Mandarin. For such phones, the postfixes -P, -T,-K, -H are tailed in normal vowels, e.g., "A", "aP", "aT", "aK" and "aH" [3]. All the phones are listed in <table.1>, where we also list the corresponding IPA synbol for each phone of TWbets.

3. The Baseline System

We adopted a CHMM based approach to construct a bi-lingual acoustic models for use in the current system. For the feature extraction, a 39-dimensional feature vector was used, including MFCCs, energy plus their first and second order derivatives. The acoustic models were intra-syllabic context dependent phone Hidden Markov model (HMM) trained by HTK3.0[8]. The number of Gaussian mixtures at each state is variable, which depends on the availability of training data for each model. The vocabulary size of the recognition task is 20 thousands. The searching net directly uses Chinese character (hanzi) as the node, therefore this is a one stage tactic to find the Chinese character string as the output in the decoding process. The diagram of the one-stage method recognizer was shown in <fig. 1>. The recognition performance of such a baseline system is shown as in <table.3>, where the character error rate (CER) and utterance error rate (UER) were shown.

Languages	CER	UER
Taiwanese	21.01%	30.80%

<table. 3>:Recognition performance of the baseline system, CER for character error rate, UER for utterance error rate

4. Adding Tone Feature and Multiple Pronunciation Modeling

4.1. Pitch Feature Extraction

As mentioned earlier, Taiwanese is a tonal language. To discriminate the tones could help discriminate the characters, words or phrases. However, MFCCs have no pitch information to characterize the tones of the language. So, extra tone feature like pitch contour should be extracted from the speech waveform.

In our system, we adopted an algorithm based on auto-correlation and harmonic-to-noise ratio to estimate pitch[1]. The normalized auto-correlation function $r(\tau)$ can be shown as follows:

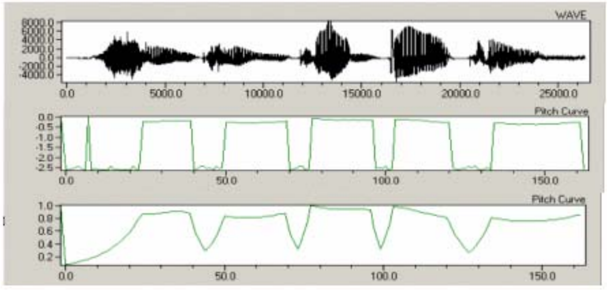
$$r(\tau) = r(-\tau) = \frac{\int_0^T a(t)a(t+\tau)dt}{\int_0^T a^2(t)dt} \qquad (1)$$

where *a(t)* is a windowed speech signal.

Pitch can only be correctly estimated in voiced region of a waveform. For unvoiced part, pitch is usually assigned zero for traditional pitch analysis program. However, for speech recognition purpose, such a zero padding strategy may not be appropriate because it will lead to problems of zero variances and undefined derivatives at voiced/unvoiced transitions. We used two other strategies to assign values to the unvoiced part of an utterance, where real pitch could not be estimated or even did not exist.

In the first one, we add very small random numbers to the unvoiced regions. We call it pitch smoothing by random numbers. In the second one, we fill the pitch gap by exponential decay/grow functions to connect the pitch contours of two voiced regions we call it pitch smoothing by exponential functions.

The results of the two smoothing strategies can be shown by an example as <fig. 1>.



<fig. 1> An example to demonstrate the two pitch smoothing strategies.

4.2. Multiple Pronunciation Modeling

As mentioned earlier, Taiwanese has severe pronunciation variations. Although native speakers could easily understand the other one speaking Taiwanese with all kinds of pronunciation variations, it is not easy for a speech recognizer to do so, unless special methods were used to address this issue. Here, we proposed a multiple pronunciation modeling technique to alleviate such a difficulty.

To incorporate pronunciation modeling at lexicon level, one approach is to use a pronunciation model to build a probabilistic lexicon to include alternative pronunciations. For a given observation X, the target function to be maximized in the searching process becomes:

$$\hat{w} = \arg\max_{w}\{p(w) \cdot \max_{i}[P(v_i \mid w)p(x \mid v_i)]\} \qquad (2)$$

where v_i is the i^{th} multiple pronunciation of the word *w*. $P(x \mid v_i)$ is the acoustic likelihood of pronunciation v_i. The unigram probability distribution of the pronunciations of *w* is given by $P(v_i \mid w)$, subjected to the normalization constraint:

$$\sum_{i}^{N} P(v_i \mid W) = 1$$

where N: is the total number of pronunciations of word W

According to the above technique, our practical implementation is sketched as in <fig.2>.

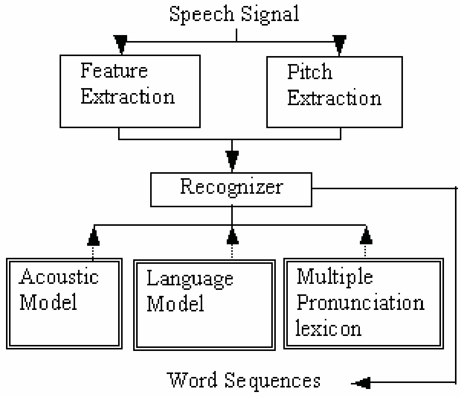
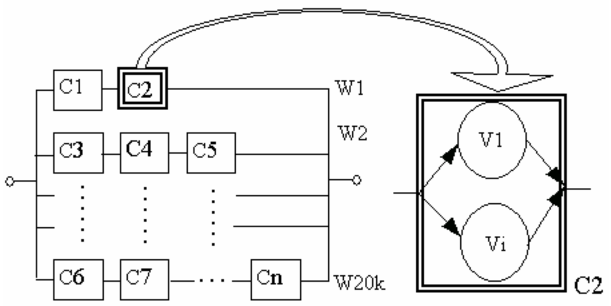


Figure 2. The diagram of the speech recognition.

5. Experimental Results

All the experimental results are presented in this section after introducing the techniques of pitch feature extraction and multiple pronunciation modeling. Our goal is to build a Taiwanese speech recognition system which can transcribe Taiwanese speech into a text string composed of Chinese characters. The vocabulary size is set to be 20 thousand. In this system, we construct a linear character searching net as shown in <fig.3>, where each branch represents a Taiwanese word, composing of nodes cascaded linearly. Each node represents a Chinese character, which is made of a subnet representing all the multiple pronunciations of the character.

First of all, in addition to the original 39 dimensional feature vector used in the baseline system, a 3-dimensional vector, representing pitch and its first and second order derivatives was added to be parts of the speech features. Secondly, right context dependent Initial-Final models with 3 states per Initial and 4 states per Finial are adopted as the acoustic models. For each model, there are two streams of feature vectors[4]; one stream is for the 39-dimensional vector; and the other is for the 3-dimensional vector. Both streams are combined with the same weighing factors. On the other hands, there are 16 Initials, and 297 tonal Finials in the Taiwanese language. After contextual expanding the number of Initials increases to 146, and the number of tonal Finials keeps the same. Furthermore, the lexicon with single pronunciation per word at baseline system is changed by adding probabilistic multiple pronunciations. The probability of each word's pronunciation is estimated from the available electronic Taiwanese lexicon by adding phone level confusing sets [7].



<fig.3> The 20K multiple pronunciation searching net.

The results of this experiment are listed in <table.4>. First, we compare two different smoothing scheme for pitch contour, as shown in 2nd and 3rd rows, pitch smoothing by exponential functions (denoted as SmoothE) is better than pitch smoothing by random numbers (denoted as SmoothR). The reason is because the pitch contour should be as smooth as possible to fit the rhythm while human being speaking. However, the recognition errors are more than that of the baseline system. This is because Taiwanese has a severe tone sandhi problem, if we only assign one character with one pronunciation with tone, it is very probable that the tone assignment is incorrect, unless some rules of tone sandhi being considered carefully. Actually, tone sandhi problems can be considered one kind of multiple pronunciations. So we leave it to the next stage by using the multiple pronunciation modeling.

Taiwanese	CER	UER
Single Pronunciations	21.5%	28.7%
SmoothR, Single Pronunciation	26.73%	37.94%
SmoothE, Single Pronunciations	23.03%	34.54%
SmoothE, Multiple Pronunciations	11.97%	18.00%

<*table 4*:> Taiwanese tonal features and multiple pronunciations results. SmoothR: pitch smoothing by random numbers; SmoothE: pitch smoothing by exponential functions.

When multiple pronunciations modeling as shown in <fig.3> was used, with the same pitch smoothing scheme, the character error rate (CER) decreases from 23.03% to 11.97%. Compared with the baseline system, such a CER is also lower. We can see the great improvement by using multiple pronunciations modeling, and it is especially important in Taiwanese speech recognition.

At last, we add a sub corpus of Mandarin speech to double the quantity of training data, the total 22 hours acoustic modeling is build. Under 20k Taiwanese character net, the result is listed at the last row, CER decreases 2.73%, 3.5% in UER. The results tell us using more data will give a better performance.

6. Conclusions

In this paper we described two improvements in Taiwanese LVCSR. One is the adding pitch features, and the other is adding the probabilistic multiple pronunciation modeling. Normalized autocorrelation function is as the main pitch tracking algorithm, by this approach, we have achieved 14% reduction of character error rate on average. The later using prior knowledge as the pronunciation probability to multiply the acoustic modeling score. It decreases the error rate very significantly and thus considered an indispensable in a Taiwanese speech recognition system.

7. References

[1] Paul Boersma "Accurate Short-Term analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratip of A sampled Sound", *1993*.

[2] Hank, Huang C.H., Frank Seide "Pitch Tracking and Tone Features for Mandarin Speech Recognition". In Proc. ICASSP, 2000.

[3] SharLene Liu, Sean Doyle, Allen Morris, Farzad Ehsani "The Effect of Fundamental Frequency on Mandarin Speech Recognition", ICSLP 98, vol. 6, Sydny, 1998, pp. 1543-1546

[4] Frank Seide, Nick J.C. Wang "Two-Stream Modeling of Mandarin Tones", ICSLP, 2000

[5] J.J. Humphries, P.C.Woodland, D.Pearce "Using Accent-Specific Pronunciation Modeling For Robust Speech Recognition", ISCA ITR-Workshop on Pronunciation Modeling and Lexicon Adaptation, Colorado, September 2002

[6] R. Y. Lyu, Z. H. Fu, Y. C. Chiang, H. M. Liu "A Taiwanese(Min-nan) Text-to-Speech(TTS) system Based on Automatically Generated Synthetic Units", the 6th International Conference on Spoken Language Processing (ICSLP2000), Oct. 2000, Beijing, China

[7] Dau-Cheng Lyu, Bo-hou Yang, Min-Siong Liang, Ren-Yuan Lyu, Chun-Nan Hsu "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition", SST, Melbourne, 2002

[8] Steve Young "HTK Book.3.1"

[9] Patgi KAM, Tan LEE "Modeling Pronunciation Variation For Cantonese Speech Recognition", ISCA ITR-Workshop on Pronunciation Modeling and Lexicon Adaptation, Colorado, September 2002

[10] http://www.sil.org/ethnologue/top100.htm