VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences

Computer Science

A Speech Recognition System for Improving English Pronunciation of Mandarin Speakers

Mengjie Zhang

Technical Report CS-TR-01/14
October 2001

School of Mathematical and Computing Sciences
Victoria University
PO Box 600, Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341, Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

A Speech Recognition System for Improving English Pronunciation of Mandarin Speakers

Mengjie Zhang

Technical Report CS-TR-01/14
October 2001

Abstract

To improve the English pronunciation of Mandarin speakers, this paper describes an approach to the use of machine learning techniques for single English word recognition and pronunciation measure problems. After briefly presenting a simple pronunciation improvement system for user typed/selected English words, we mainly describe the design of an automatic speech recognition system based on pronounced words from Mandarin speakers.

Keywords Voice recognition, machine learning, pronunciation improvement.

Publishing Information

This work is supported by the NERF research project — *Novel interactive software tools for teaching English as a second language.*

Author Information

The author is a lecturer in computer science, Victoria University of Wellington, New Zealand.

1 Introduction

Since the early 1980s, English has become a major foreign language in China and other Asian countries. These countries wish to become more competent in English to communicate with "the West" for the need of cooperation and development in education, research, economy, etc. Accordingly, there is a clear demand for good quality teaching of English to students from these countries. However, the number of people requiring English tuition in the next 10-15 years by native English speakers is enormous, hopelessly out of proportion to the number of competent teachers available, if traditional methods continue to be used. In particular, the improvement of pronunciation and conversation skills has traditionally been very teacher-intensive. The development of interactive tools or systems which can be used by students working independently to improve pronunciation and conversation will greatly increase the productivity of the teaching work force.

The goal of this approach is to develop a speech recognition system for improving English pronunciation for foreign speakers, particularly those from China and other Asian countries. The focus is mainly on the design of a single/isolated word speech recognition system, but extend it to a continuous recognition system in the future. This approach consists of three stages, which correspond to three levels of speech recognition systems. These systems are designed for the use of different purposes and constructed with increasing difficulty. The three stages/systems are: (1) A simple pronunciation improvement system with user typed inputs; (2) An automatic speech recognition system for pronunciation improvement; and (3) A continuous speech recognition system.

In the remainder of this paper, the major differences between English and Mandarin pronunciation are firstly summarised. The main ideas, structures, methods and the key research issues of the first two stages/systems are then described. After briefly analysing the main techniques involved in the third stage/system, we conclude the work.

2 Characteristics of Pronunciation from Mandarin Speakers

In English, there are 44 phonetic alphabets. In Mandarin, there are 47 independent phones and 11 combined vowels in the internationally recognised, standard *Pinyin* system, as shown in table 1. The last two lines of the table show the combined vowels in the Mandarin *pinyin* system.

Mandarin Phones									
Consonants	b	p	m	f		d	t	n	l
	g	k	h			j	q	x	
	zh	ch	sh	r		z	c	s	
	y	w							
Vowels	a	o	e			i	u	ü	
	ie	üe	er						
	ai	ei	ao	ou	iu	ui			
	an	en	in	un	ün				
	ang	eng	ing	ong					
	ia	iao	ian	iang	iong				
	ua	uo	uai	uan	uang	üan			

Table 1: Chinese Mandarin Phones in the *Pining* system.

Comparing the the 44 English phones/sounds and the 58 Mandarin sounds, we can find some overlap between them. In other words, some common sounds are the same or similar in both languages, while some others are very different. There are two main levels of differences between English and Mandarin: phonetic level and word level.

- **Phonetic level.** In terms of the phonetic level, there are two main aspects of differences. one is that there are some Mandarin sounds which English does not have or are very different; the other is that there are some English sounds that Mandarin language does not have or are quite different. For our purpose, we will focus on the second case.

  For example, two similar friction sounds /s/ and /θ/ in English correspond to a single /s/ in the Mandarin *pinyin* system. In other words, we have to distinguish /s/ and /θ/ in English, which is a bit hard to Mandarin speakers.

- **Word level.** Regarding the word level, the major difference is consonant sequences. In Mandarin, each Chinese character (word) generally consists of a single consonant and a single vowel. In some special cases, a single vowel can form a Chinese character. However, it is quite common that two or more consonants come together to form a consonant sequence in English. A lot of Mandarin speaker can not pronounce these English words correctly or frequently, particularly for long consonant sequences.

  Stress is another difference between the two languages in the word level. However, this problem might not be as big as the two mentioned above for Mandarin speakers.

  Linking between different words is also a distinction between the two languages. Since this involves continuous speech recognition, we do not discuss it in details here.

Based on the distinction between the two languages, a single word database consisting of a set of difficult pronounced words for Mandarin speakers can be built depending on which we can develop a single word speech recognition system.

3 A Simple Pronunciation Improvement System

The proposed simple pronunciation system is presented in figure 1.

The main goal here is to develop a user-computer interaction system for the improvement of a non-English speaker's pronunciation. As shown in figure 1, at the beginning a user types a word from the keyboard or select a word from the screen. Then the typed word will be matched from a pre-defined, text, single word database. If the match is successful, the system will search the word from a pre-defined, voice database (single words), and give the correct pronunciation through a speaker. The user now follows the speaker to improve his/her pronunciation. In this system, the user should be able to control the repetitions of the correct pronunciation and appropriate feedback should be provided to the user.

The main tasks here are:

1. Create a single word, text database (SWTDB) based on the analysis of the Mandarin speakers' pronunciation characteristics as presented in section 2.

2. Select/develop a searching/matching algorithm to find the word in the SWTDB which matches the user typed single/isolated word.

3. Create a single word voice/pronunciation database (SWVDB) which corresponds to the SWTDB. In the SWVDB, the pronunciation of the difficult words should be correct and be made by a native English speaker/teacher.

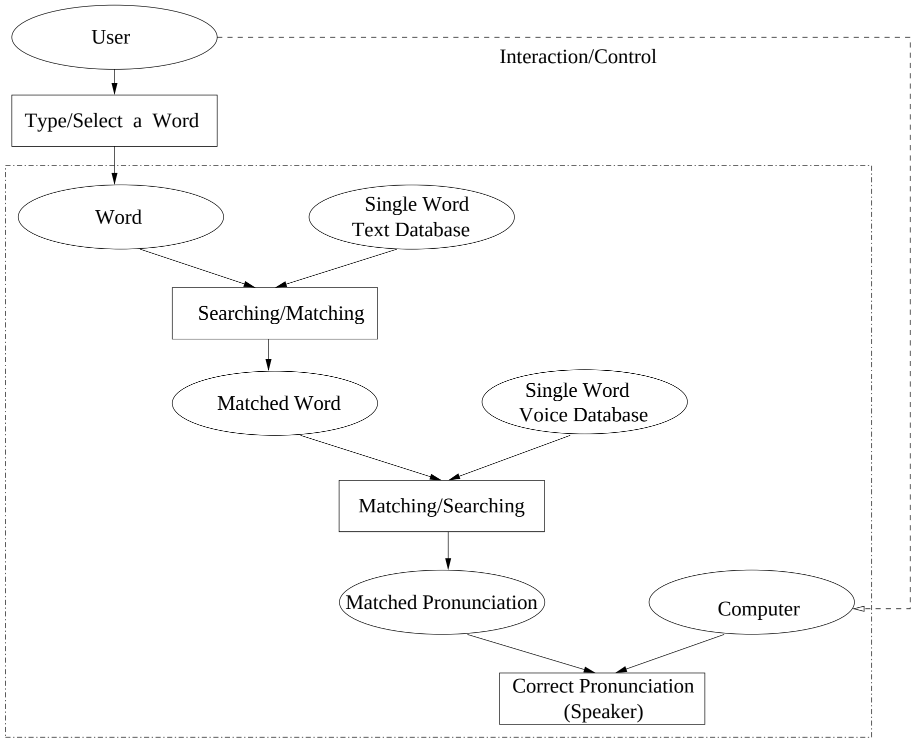

Figure 1: A pronunciation improvement system

4. Select/develop a matching/searching algorithm to find the right pronunciation of the matched, user typed word in the SWVDB.

5. Determine what kinds of feedback should be given to the user.

The single word text database and the single word voice/pronunciation database might be able to be merged into a single database.

The key techniques include the word searching/matching algorithm and the conversion between the analog signal and digital signal. In general, for word searching, we could select appropriate techniques from existing ones, such as B-trees, B+ trees and extendible hashing.

4 An Automatic Speech Recognition System

The main goal of this system is to develop a single/isolated word voice recognition system for automatic recognition of single word (English) pronunciation from Mandarin speakers, then to use the simple pronunciation system in stage one to improve the user's pronunciation. The proposed single/isolated word voice recognition system is presented in figure 2.

As shown in figure 2, there are three parts in this system: Part I — data preprocessing; Part II — a learning system; and Part III — an online recognition system.

4.1 Part I — Data Collection and Data Preprocessing

This part consists of two main tasks:

1. Create a word pronunciation source database (WPSDB) containing a set of typical, relatively difficult pronounced words. These words are stored in analogous form (voice)

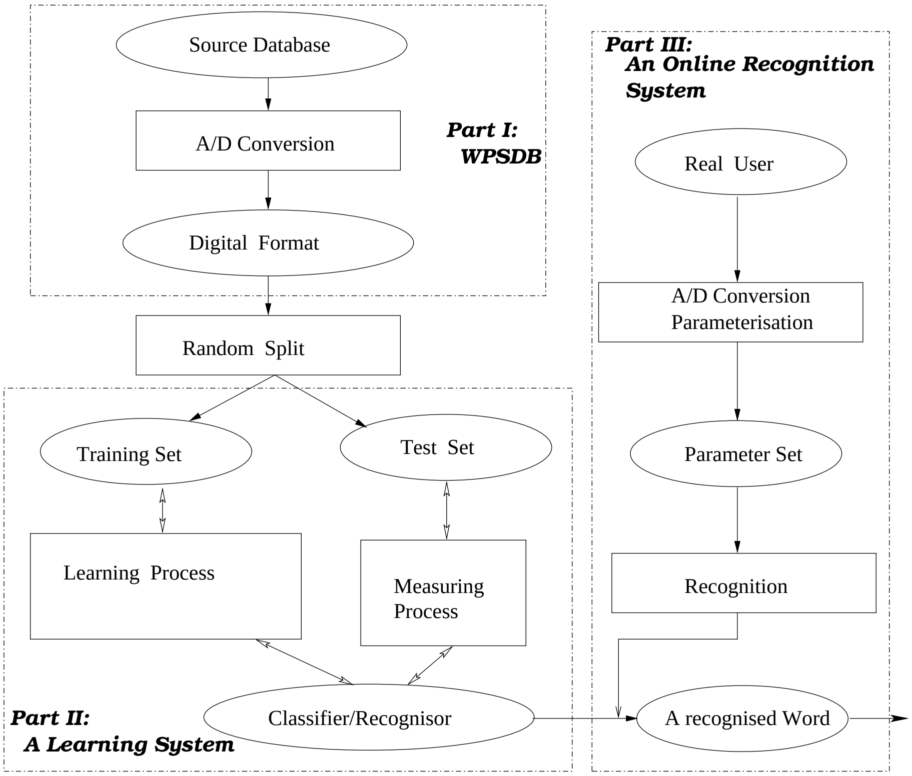

Figure 2: A speech recognition system

on a large secondary memory, such as a CDROM or a ZIP disk. This database should have the following characteristics:

- It should be obtained from a "local area", such as north China Mandarin speaking region.
- It should contain only relatively difficult pronounced words, rather than a whole dictionary.
- Each word is pronounced by different people with several repetitions. For example, for each of the words `think, sink, zink`, 10–20 people make pronunciation and each person repeats several times. As shown in table 2, each word is pronounced by 10 people and each person repeats 10 times.

Word Class Label	Pronunciation of the word
Class 1 (word 1)	$\mathrm{W}_{p0,0}$, $\mathrm{W}_{p0,1}$, $\cdots$, $\mathrm{W}_{p0,9}$,
	$\mathrm{W}_{p1,0}$, $\mathrm{W}_{p1,1}$, $\cdots$, $\mathrm{W}_{p1,9}$,
	$\cdots$, $\cdots$, $\cdots$
	$\mathrm{W}_{p9,0}$, $\mathrm{W}_{p9,1}$, $\cdots$, $\mathrm{W}_{p9,9}$,
Class 2 (word 2)	$\mathrm{W}_{p0,0}$, $\mathrm{W}_{p1,1}$, $\cdots$, $\mathrm{W}_{p1,9}$,
	$\mathrm{W}_{p1,0}$, $\mathrm{W}_{p1,1}$, $\cdots$, $\mathrm{W}_{p1,9}$,
	$\cdots$, $\cdots$, $\cdots$
	$\mathrm{W}_{p9,0}$, $\mathrm{W}_{p9,1}$, $\cdots$, $\mathrm{W}_{p9,9}$,
$\cdots$	$\cdots$, $\cdots$, $\cdots$

Table 2: Data collection.

2. Preprocess this database to make it accepted by computers. This needs to convert these words from the analogous format to digital format.

Note that this word pronunciation source database (WPSDB) should be consistent with the single word database (SWTDB/SWVDB) in stage one.

4.2 Part II — A Learning System

This is the key part of our automatic voice recognition system of English pronunciation. The main goal of this part is to develop a voice/speech recogniser or classifier based on the preprocessed digital WPSDB in Part I.

To achieve this goal, we need to develop a learning/adaptive system. This will involve the following steps:

1. Parameter selection and extraction (*Parameterisation*) based on the preprocessed data (WPSDB). The following research tasks need to be considered:
   - Determine a set of good parameters/features appropriate for this problem.
   - Develop an algorithm to extract these parameters/features.
   - Form a parameter/feature based data set by applying the parameter/feature extraction algorithm to the digital WPSDB.
2. Data splitting. To deal with the learning process, the parameterised data needs to be randomly split into two data sets: *a training set*, which is used for the learning process, and *a test set*, which is used for measuring the performance of the learned recogniser. Alternatively, a third set, *validation set* can be used for guiding the learning procedure to avoid the overfitting problem.
3. Development of the learning method. This will involve the following tasks:
   - Construct patterns. This is a special format of the training set and the test set. An example of a pattern file is shown in table 3. In this table, *p1, p2, ..., p5* are extracted parameters/features of the words, and *class1, class2* are desired word labels.

Pattern No.	Input pattern					Output pattern
	p1	p2	p3	p4	p5	
1	0.10	0.4	0.2	0.8	0.5	class1
2	0.13	0.38	0.19	0.82	0.51	class1
3	0.32	0.23	0.42	0.34	0.29	class2
... ...	...	...	...	...	...	...

Table 3: A sample pattern file.

   - Determine the learning strategy. What learning paradigm should be used? Neural networks? Evolutionary process? Rule based learning? Decision tree? Supervised learning and/or reinforcement learning should be used rather than unsupervised learning.
   - Determine the learning system architecture.

- Develop a learning algorithm. This should be consistent with the system architecture.
- Design experiments. The recogniser/classifier would be obtained from here.

4. Performance Measurement. This involves selecting appropriate evaluation criteria to measure the learned recogniser. The *word error rate*, which reflects the recognition accuracy, can be used.

5. The above steps should be repeated and the relevant parameters or algorithms should be improved until a "good" classifier/recogniser is obtained.

4.3 Part 3 — An Online Recognition System

The goal of this part is to create an online voice recognition system for automatic recognition of a totally "new" pronunciation. This consists of two steps:

1. Parameterise the given new pronunciation by applying the same methods in Part 2 and form a new input pattern.
2. Feed the input pattern into the learned recogniser to produce the output pattern — the recognised word (class label associated with a desired word) as the result.

4.4 A Hybrid System

By combining stage 1 in section 3, the *simple pronunciation improvement system*, and stage 2 in this section, the *automatic speech recognition system*, a hybrid voice recognition and pronunciation improvement system can be created, as shown in figure 3.

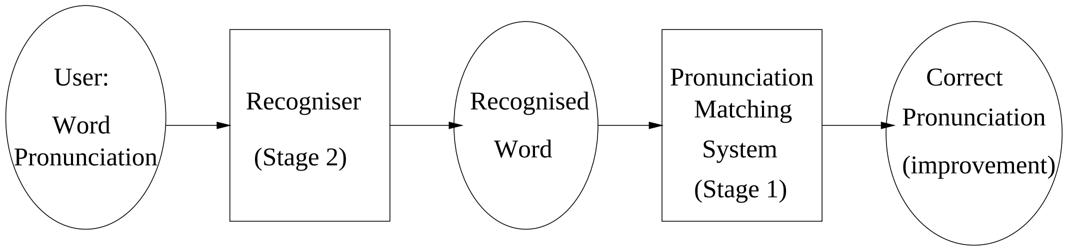

Figure 3: An automatic voice recognition and pronunciation improvement system

As can be seen from figure 3, this hybrid system can automatically recognise a new word pronunciation then present the correct pronunciation (from a native speaker) of the word. By following and repeating the correct pronunciation, the user can directly improve his/her pronunciation without a native English teacher.

5 A Continuous Speech Recognition System

The goal of the continuous speech recognition system is to automatically recognise words in continuous sentences. When speech is produced in a sequence of words, language models and artificial grammars will be used. Thus the words in such a sequence is not just the simple combination of words. To achieve this goal, machine learning and natural language processing and understanding techniques need to be applied. Instead of discussing this system in details here, we consider it a future work.

6 Conclusions

The goal of this paper is to design a learning approach to automatic speech recognition, particularly for improving Mandarin speakers' English pronunciation. Based on the analysis of the major differences between English and Mandarin phonetic alphabets, a simple pronunciation improvement system with word matching though computer keyboards and an automatic single word speech recognition system using machine learning techniques were proposed. These two systems not only can be used independently, but also can be integrated into an automatic voice recognition and pronunciation improvement tool for Mandarin speakers. We expect that the development of this approach can help Mandarin speakers improve their pronunciation using the online system by themselves without native English teachers.

7 Acknowledgement

I would like to thank David Crabbe, Mike Doig, Irina Elgort, Peter.Andreae, Paul Warren and Neil Leslie for useful discussions.

References

[1] Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue. *Survey of the State of the Art in Human Language Technology.* Cambridge University Press, 1996. ISBN 0-521-59277-1.

[2] Norman Foo. *Advanced Topics in Artificial Inteeligence – Proceedings of Australian Joint Conference on Artificial Intelligence (AI'99).* Springer-verlag, Berlin Heidelberg New York, 1999. Lecture Notes in Artificial Intelligence 1747, ISBN 3-540-66822-5.

[3] Wolfgang Menzel, Daniel Herron, Rachel Morton, Datio Pezzotta, Patrizia Bonaventura, and Peter Howarth. Interactive pronunciation training. *ReCALL*, 13(1):67–78, 2001. Cambridge University Press.

[4] T. Mitchell. *Machine Learning.* McGraw-Hill, New York, 1997.