Invited Keynote Speech, Oriental-COCOSDA, pp.9-23, Oct. 1-3, Sentosa, Singapore

Making Full Use of Chinese Speech Corpora

Thomas Fang Zheng

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Tsinghua University, Beijing, 100084, China,
fzheng@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn/~fzheng/
Beijing d-Ear Technologies Co., Ltd., Room 505, Building D, 2 Information Road, Shangdi, Haidian District, Beijing, 100085, China
fzheng@d-Ear.com, http://www.d-Ear.com

Abstract

It is well understood that the speech databases play a very important role for speech recognition. It is a dream for speech recognition researchers to create more useful databases with smaller efforts. To achieve this goal, the database should be well designed at first, and tools and more information should be provided so that the databases can be made full use of. This paper will illustrate the criteria according to which the Chinese speech databases will be created for different purposes. The way of transcription will also be discussed, which is the first thing to do after the data creation. Then examples on how to learn knowledge from the created database for other research purpose will be given.

1 Purpose of Speech Corpora

The speech corpora play a very important role in speech and language processing, and this has been aware of by the speech community for years. To make full of speech corpora efficiently, people in the speech community worldwide have established consortiums, such as LDC, ELRA, and so on. The purpose of speech corpora can be illustrated in Table 1 (Kuwabara 2002).

Table 1. Purpose of Speech Corpora

Item	Description	Percentage
1. Speech/speaker recognition	system development, evaluation, sentence comprehension and summarization, speech recognition, speaker recognition	73%
2. Speech synthesis	system development, prosodic analysis	11%
3. Acoustic analysis	acoustic analysis, speech coding	9%
4. Sentence analysis	syntactic and semantic analysis	5%
5. Speech/language education	speech and language education	2%

The purpose of a speech database determines the way how it will be created, including the database collection and database transcription. Normally the following factors will be considered before a speech corpus is to be created.

1. The language. For Chinese, for example, it could be standard Chinese, known as *Putonghua*[1], or it could be one of the nine main Chinese dialects, i.e., *Mandarin* (Northern China), *Wu*

[1] We don't use "*Mandarin*" to stand for the standard Chinese because in this paper it is used to represent a kind of Chinese dialect in north China, namely *Guan*1 *Hua*4 in Chinese Pinyin, or "官话" in Chinese character.

(Southern Jiangsu, Zhejiang, and Shanghai), *Yue* (Guangdong, Hong Kong, Nanning Guangxi), *Min* (Fujian, Shantou Guangdong, Haikou Hainan, Taipei Taiwan), *Hakka* (Meixian Guangdong, Hsin-Chu Taiwan), *Xiang* (Hunan), *Gan* (Jiangxi), *Hui* (Anhui), and *Jin* (Shanxi). It could be the simplified Chinese or the traditional Chinese.

2. Speaking style. Normally, there are two kinds of speaking styles for the database creation. In years ago, the database is often containing read speech only because of the research level of the speech processing or because it is for the text-to-speech purpose. Nowadays, more and more spontaneous speech corpora are collected and transcribed so as to meet the requirement of the spontaneous speech recognition or conversational speech recognition.
3. Recording channel. The factor often depends on the goal of the task or application or the application environment. For personal computers (PCs), close-talk microphones are often used as the device to collect speech, while for telephony applications the telephone, and/or the cellular (mobile) phone is often used. In other such embedded applications as personal digital assistant (PDA) or digital recorder, the channel will be different. It should be mentioned that normally the mono channel instead of the stereo channel will be used.
4. Sampling rate. This is often associated with the recording channel. For example, an 8 kHz sampling rate is often taken for the telephone/mobile-phone channel where the bandwidth is about 3.4 kHz, while often 16 kHz for the close-talk microphone PC channel though the bandwidth is higher than 8 kHz. Though in some specific applications other values of sampling rate could be chosen, 8 kHz and 16 kHz are two frequently used sampling rate values.
5. Sampling precision. The sampling precision will affect the speech quality directly. With the development of the analog-to-digit converter (ADC) technologies, people tend to use a higher precision, for example 16 bits. However, in telecommunication applications the 8-bit A-law or Miu-law bit width is often used which is 13-bit wide after decompression.
6. Signal-to-Noise Ratio (SNR) level. Years ago, the database was often collected in a good environment where the SNR was strictly controlled to be above an acceptable value, such a database is known as a clean speech database. When researchers want to do noise-related experiments, some noisy speech database such as NOISEX 92 (Varga 1992, NOISEX92 1992) will be mixed with a clean database in a certain way to obtain a noisy speech database. However, such a noisy database is quite different from a real noisy one, therefore conclusions and theories obtained after the experiments across such mixed databases are often not so effective as expected when applied practically. Based on this, databases are now being collected in really noisy environments.
7. Number of speakers and the speaker balance. For speech recognition, it is often required that a large number of balanced speakers be chosen so as to get a good speaker diversity. The more speakers we have the better. Often a fixed number of speakers will be determined at first. Then when balancing the speakers, what will be considered includes gender, age, education, birthplace, and so on.
8. Corpus size. The size of a corpus could be measured by either the number of speakers or the length of valid speech in hour, or both.

In this paper, those collection and transcription related issues for speech recognition purpose databases will be discussed.

2 Database Collection

Roughly speaking, a corpus designed for speech recognition has two kinds of purposes, acoustic training or speech recognition evaluation (testing). Therefore it is often that a database is divided into two sets, a training set and a testing set. On the other hand, there are two kinds of databases, one is the read speech and the other is the spontaneous speech.

No matter what kind of database is to be created, it should be well designed before creation.

The design of a speech database covers several aspects as mentioned in last section. Among these factors, the language, speaking style, recording channel, sampling rate and precision, and corpus size,

are often determined according to the application/task background, while the SNR levels, number of speakers and the speaker balance are often for diversity purpose. Additionally, another important factor should also be paid attention to, the speaking content balance.

Such a speaking content balance is for the content diversity purpose. For read speech, it is important for acoustic training, so such a balance could be on the phone level, the di-phone level, tri-phone level, and so on. For Chinese, it could be the IF[2], di-IF, tri-IF, syllable, di-syllable, tri-syllable, and so on, additionally. For spontaneous speech, topics for speakers to talk about should be well defined.

In this section, the speaking content balance design for the speech recognition database collection will be discussed.

2.1 Read Speech Database

Though the spontaneous speech recognition is becoming one of the research focuses recently, the read speech database collection is still necessary. A high quality read speech corpus is helpful to train a good acoustic model which can be used as an initial model in spontaneous speech recognition where pronunciation modelling techniques as well as pronunciation lexicons are adopted to get a finally practical acoustic model (Fung 2000, Zheng 2001, and Zheng & Song 2002).

Due to the huge collection and transcription efforts, the design of the prompting texts in training purpose corpus collection becomes important. The goal of such a design is often to balance the phones, mono-phones, di-phones, and tri-phones, so as to cover as many co-articulations as possible using a set of sentences as small as possible (Wang 1999, Li 1999, Zu 1997, Zu 1999). Such a minimal sentence set can be used for not only the training of acoustic models but also the speaker adaptation.

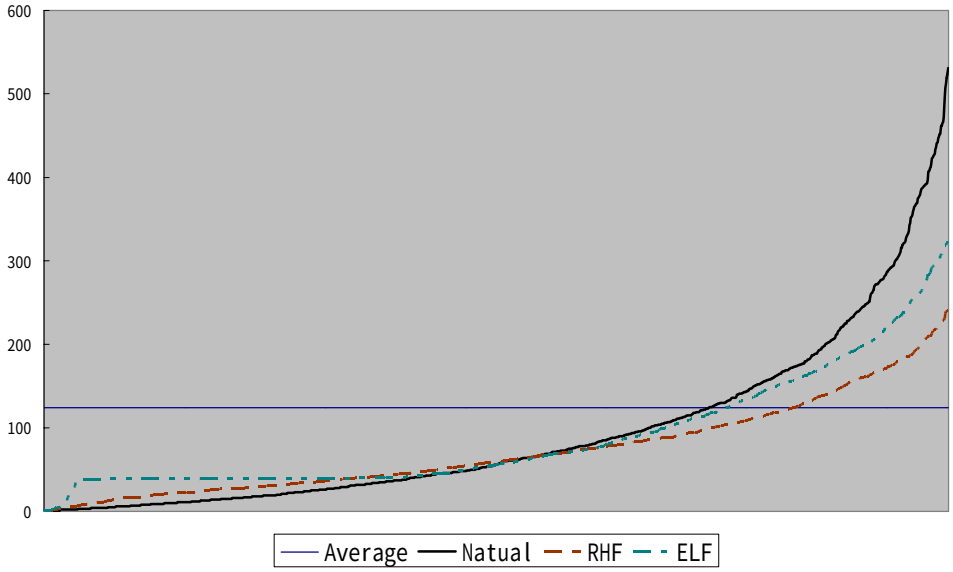
Here an example will be given on how to choose the balanced sentences. The goal here is to choose 6,000 sentences (about 0.75%) from 800,000 sentences (the mother set) taken from the *People's Daily* with a balanced di-IF distribution. There are several alternative criteria:

1. Natural selection. 6,000 sentences are selected from the mother set randomly. It is supposed that the di-IF distribution is similar to its natural distribution.
2. Restraining high-frequency di-IFs (RHF). It is ideal that each di-IF occurs almost equally in the selected sentence set so that the acoustic model of each di-IF can be well trained. However it is almost impossible. To reach this goal as close as possible, a strategy in such a sentence selection algorithm is to restrain those high-frequency di-IFs from occurring more frequently.
3. Encouraging low-frequency di-IFs (ELF). As an alternative, such a strategy should be adopted to encourage those low-frequency di-IFs to occur in the selected sentence set as frequently as possible. In addition, the number of occurring times of any di-IF should be greater than a doable threshold pre-defined according to the raw big sentence set - the mother set.

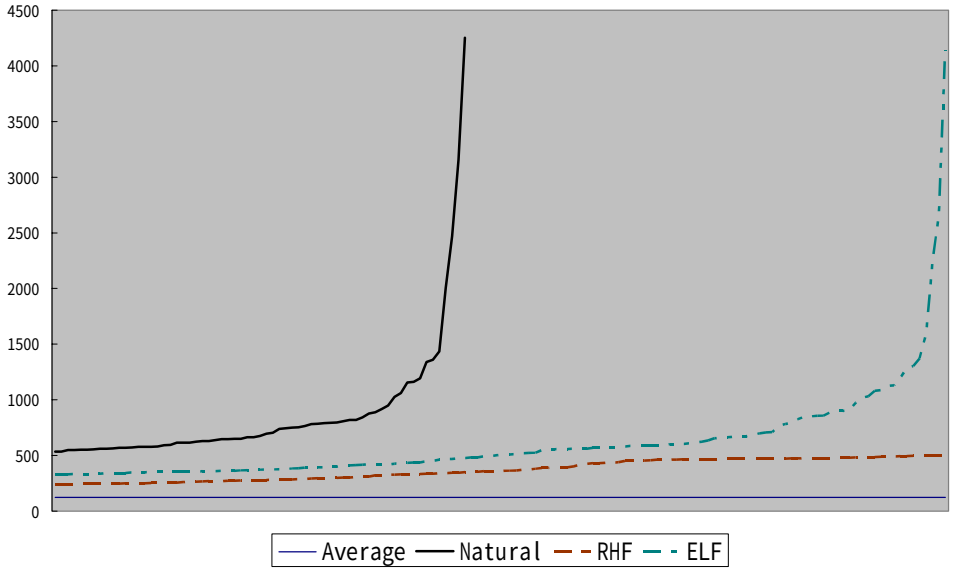
The performance comparison of the above criteria is illustrated in Figure 1. In this example, all seen di-IFs are sorted in an ascending order of their occurring counts in the mother set. The occurring count as a function of the di-IF order is shown in Figure 1 (a), the first 90% di-IFs, and Figure 1 (b), the next 10% di-IFs. For the sake of comparison, the average di-IF occurring count is also drawn.

The RHF strategy is helpful to make the distribution curve as close to the average curve as possible, however it cannot guarantee a minimal occurring count for those less frequently seen di-IFs. On the contrary, the ELF can guarantee a minimal occurring count but will lead to relatively big occurring counts for those most frequently seen d-IFs though the distribution is close to natural. It is expected that a good result can be achieved by combining both the RHF and the ELF strategies.

[2] An IF (声韵母) means either a Chinese Initial (声母) or a Chinese Final (韵母).



(a)



(b)

Figure 1. Comparison of different sentence selection methods

2.2 Spontaneous Speech Database Design

Research on spontaneous speech recognition becomes popular nowadays. A spontaneous speech corpus should contain rich spontaneous speech phenomena (phonetically and acoustically) and rich spoken language phenomena (linguistically). One way is to record the dialogues "by stealth"[3], for example Zheng & Yan 2002. Utterances in such a database will be really spontaneous. However there are disadvantages: (1) the content scope cannot be pre-defined; and (2) if the speakers do not agree the collected data cannot be used so it costs a lot of efforts to get such permissions for use.

Another way is to design a spontaneous speech corpus. Different from the read speech database collection, there is not any prompting sentence for speakers to read. Therefore the design is different.

[3] Permission of recording utterances and legal use of the recorded utterances should be obtained from the speakers when recording.

Unlike the sentence design in the read speech corpus collection, what should be designed for the spontaneous speech corpus collection includes topics and questions according to the application/task requirements. This is to guarantee the content diversity as well as the content requirements. Examples can be found in Table 2.

Table 2. Examples of topics and sub-topics for spontaneous speech corpus collection

Topic	**Sub-topics**
Sports	general basketball football the game of go table tennis tennis Olympic games ...
Politics & Economy	general international relations terrorism economic globalization ...
Entertainment	general film review movie and teleplay stars Chinese Spring Festival get-togethers comic dialogue travel and tourism ...
Lifestyle	family work and study education traffic telecommunication going abroad drinking and smoking habits and customs language and dialects friendship and love shopping for and buying houses ...
Technologies	computer networking company cloning ...

3 Database Transcription

A database without the transcription information is not a complete one. In this section, what will be transcribed for a database and how it will be transcribed will be introduced. Additionally what can be learned from a transcribed database will also be discussed.

3.1 Symbols

Roughly speaking, there are two major groups of information in need of transcription, the phonetic information and the linguistic information. For the former, there are two forms, the base form and the surface form. The base form gives the canonical pronunciation of the utterance, while the surface form gives the observed (actual) pronunciation. For any language, there is an alphabet set for its basic units which means the base form transcription is easy to give in a machine-readable style. However for the surface form, it is more difficult. The International Phonetic Alphabet (IPA) can be used to accurately represent the actual pronunciation of the utterance, but it is not machine readable. Taking it into consideration, people define a new kind of alphabet named Speech Assessment Methods Phonetic Alphabet (SAMPA). Accordingly for Chinese, we have a labeling set of machine-readable IPA symbols adapted for Chinese languages from SAMPA, and it is named as SAMPA-C (Chen & Li 2000, Li & Chen 2000, and Fung 2000). In SAMPA-C, there are 23 phonologic consonants, 9 phonologic vowels and 10 kinds of sound change marks (nasalization, centralization, voiced, voiceless, rounding, syllabic, pharyngrealization, aspiration, insertion, and deletion), by which 21 initials, 38 finals, 38 retroflexed finals as well as their corresponding sound variability forms can be represented. Tones after tone sandhi, or tonal variation, are attached to the finals. See Tables 2-6.

Table 2. Chinese Consonants and Vowels in SAMPA-C

Consonants				Vowels	
Pinyin [IPA]	**SAMPA-C**	**Pinyin [IPA]**	**SAMPA-C**	**Pinyin [IPA]**	**SAMPA-C**
b [p]	p	z [ts]	ts	a [§]	A
p [pH]	p_h	c [tsH]	ts_h	o [o]	o
m [m]	m	s [s]	s	e [°]	7
f [f]	f	zh [t©]	ts`	i [i]	i
d [t]	t	ch [t©H]	ts_h`	u [u]	u
t [tH]	t_h	sh [©]	s`	ü [y]	y
n [n]	n	r [˛]	z`	ii[4] [¡]	i\
(a)n [n]	_n	j [t»]	ts\	iii[5] [Ÿ]	i`
l [l]	l	q [t»H]	ts_h	er [Ä]	@`
g [k]	k	x [»]	s\		
k [kH]	k_h	?	?		
h [x]	x				
ng [Ð]	N				

Table 3. Phonological Chinese Vowels in SAMPA-C [6]

Phoneme		**Allophone (IPA)**	**SAMPA-C**	**Phoneme**		**Allophone (IPA)**	**SAMPA-C**
Pinyin	**IPA**			**Pinyin**	**IPA**		
a	**A**	**a**	**a**	**i**	i	**i**	**i**
		A	**a_"**			**I**	**I**
		§	**A**			**j**	**j**
		Q	**E**		**ï**	**Ÿ**	**i`**
		P	**{**			**ị**	**i**
o	**o**	**o**	**o**	**u**	**u**	**U**	**U**

[4] In this paper, "ii" stands for the Final in Pinyins "zi", "ci", and "si".

[5] In this paper, "iii" stands for the Final in Pinyins "zhi", "chi", "shi", and "ri".

[6] The Phonological Chinese consonant set is the same as the Chinese consonant set.

		U	U			u	u
		u	u			w	w
		È	@			U	v\
e	°	°	7	ü	Y	y	y
		e	e			á	H
		E	E_r	er	Ä	Ä	@`
		È	@				

Table 4. Chinese Initials

Method of Articulatory	Place of Articulatory	Pinyin	IPA	SAMPA-C
塞音 Stops	唇　音 Labial	b	p	p
	唇齿音 Labiodental			
	舌尖前音 Dentalveolar	d	t	t
	舌尖后音 Retroflex			
	舌面音 Dorsal			
	舌根音 Velar	g	k	k
塞送气 Aspirated Stops	唇　音 Labial	p	pH	p_h
	唇齿音 Labiodental			
	舌尖音 Alveolar	t	tH	t_h
	舌尖后音 Retroflex			
	舌面音 Dorsal			
	舌根音 Velar	k	kH	k_h
塞擦音 Affricates	唇　音 Labial			
	唇齿音 Labiodental			
	舌尖前音 Dentalveolar	z	ts	ts
	舌尖后音 Retroflex	zh	t©	ts`
	舌面音 Dorsal	j	t»	ts\
	舌根音 Velar			
塞擦送气 Aspirated Affricates	唇　音 Labial			
	唇齿音 Labiodental			
	舌尖前音 Dentalveolar	c	tsH	ts_h
	舌尖后音 Retroflex	ch	t©H	ts`_h
	舌面音 Dorsal	q	t»H	ts_h
	舌根音 Velar			
鼻音 Nasals	唇　音 Labial	m	m	m
	唇齿音 Labiodental			
	舌尖音 Alveolar	n	n	n
	舌尖后音 Retroflex			
	舌面音 Dorsal			
	舌根音 Velar			
擦音 Fricatives	唇　音 Labial			
	唇齿音 Labiodental	f	f	f
	舌尖前音Dentalveolar	s	s	s

	舌尖后音 Retroflex	sh	©	s`
		r	¸	z`
	舌面音 Dorsal	x	»	s\
	舌根音 Velar	h	x	x
边音 Laterals	唇　音 Labial			
	唇齿音 Labiodental			
	舌尖音 Alveolar	l	l	l
	舌尖后音 Retroflex			
	舌面音 Dorsal			
	舌根音 Velar			

Table 5. Chinese Finals

Articulatory Method of the first vowel	**Pinyin**	**IPA**	**SAMPA-C**	**Retroflexed**		
				Pinyin	IPA	SAMPA-C
开 Open	a	A	a_"	ar	ar	a`
	o	o	o	or	or	o`
	e	°	7	er	°r	7`
	ai	aI	aI	air	ar	a`
	ei	ei	ei	eir	Èr	@`
	er	Ä	@`			
	ao	§U	AU	aor	§or	Ao`
	ou	Èu	@u	our	our	ou`
	an	an	a_n	anr	ar	a`
	en	Èn	@_n	enr	Èr	@`
	ang	§Ð	AN	angr	ar	a~`
	eng	ÈÐ	@N	engr	Èr	@~`
	ong	uÐ	uN	ongr	ur	u~`
	ii	¡	i\	iir	Èr	@`
	iii	Ÿ	i`	iiir	Èr	@`
齐 Stretched	i	i	i	ir	iÈr	i@`
	ia	iA	ia_"	iar	iar	ia`
	ie	iE	iE_r	ier	iEr	iE_r`
	iao	i§U	iAU	iaor	i§or	iAo`
	iou	iÈu	i@u	iour	iour	iou`
	ian	iQn	iE_n	ianr	iar	ia`
	in	in	i_n	inr	iÈr	i@`
	iang	i§Ð	iAN	iangr	iar	ia~`
	ing	iÐ	iN	ingr	iÈr	i@~`
	iong	iuÐ	iUN	iongr	iur	iu~`
合 Round	u	u	u	ur	ur	u`
	ua	uA	ua_"	uar	ur	u`
	uo	uo	uo	uor	uor	uo`
	uai	uaI	uaI	uair	uar	ua`
	ui	uei	uei	ueir	uÈr	u@`
	uan	uan	ua_n	uan	uar	ua`

	uen	uÈn	u@_n	uenr	uÈr	u@`
	uang	u§Ð	uAN	uangr	uar	ua~`
	ueng	uÈÐ	u@N	uengr	uÈr	u@`
撮 Protruded	ü	y	y	ür	yÈr	y@`
	üe	yE	yE_r	üer	yEr	yE_r`
	üan	y P n	y{_n	üanr	yar	ya`
	ün	yn	y_n	ünr	yÈr	y@`

Table 6. Sound Changes Representation in SAMPA-C

Classifications	SAMPA-C	Examples		
		IPA	SAMPA-C	Explanation
鼻　化 **Nasalized**	~	a	§~	'§' is *nasalized*.
央　化 **Centralized**	_"	e	e_"	'e' is *centeralized*.
清　化 **Voiceless**	_u	n	n_u	'n' is *voiceless*.
浊　化 **Voiced**	_v	d	d_v	'd' is *voiced*.
圆唇化 **More Rounded**	_O	f w	f_O	'f' is more *rounded*.
成音节 **Syllabic**	=	m l	M=	'M' is *syllabic*.
喉　化 **Pharyngealized**	_?\	t/	A_?\	'A' is *pharyngealized*.
送　气 **Aspirated/Breathy**	_h	a	a_h	'a' is more breathy
增　音 **Inserted**	(+)		(+N)	'N' is *inserted*.
减　音 **Deleted**	(-)		(-i)	'i' is *deleted*.

3.2 Transcription Layers

In addition to those mentioned in Section 1, the database transcription layers should cover the base form and the surface form, as well as some linguistic and non-linguistic information. There could be all or some of the following transcription layers for a Chinese speech corpus.

3.2.1 Chinese character layer

In this layer, the transcription information includes the sentences in Chinese character (without word boundary information), or in Chinese word (with word boundary information). The Chinese character could be either traditional or simplified.

Non-Chinese words should also be labeled, for example, they can be enclosed in {}. Additionally, paralinguistic and non-linguistic phenomena could also be transcribed (See Table 7 for more details).

Here is an example.

{hi} 你怎么也 LA< 也来了 LA>　　(*without word boundary information*)

or

{hi} 你/怎么/也/ LA< 也/来/了 LA>　　(*with word boundary information*)

3.2.2 Canonical Chinese Pinyin layer

In this layer, the pronunciation of the corresponding sentence(s) is given in canonical Chinese Pinyin (or Chinese syllable). Because each Chinese Pinyin consists of a unique Chinese Initial[7] and unique Chinese Final, alternatively the pronunciation can be given in Chinese IF. The Pinyin (or Final) here can be either toned or toneless. In Chinese, there are four tones (represented by 1~4) and a neutral tone (by 0).

[7] If there is no consonant in a Chinese syllable, its Initial is called a *null Initial* (零声母).

Similarly, non-Chinese words, paralinguistic phenomena and non-linguistic phenomena could also be transcribed (See Table 7 for more details).

The example is as follows:

{hi} ni3 zen3 me0 ye2 LA< ye2 lai2 la0 LA> (*toned Pinyin string*)

or

{hi} n i3 z en3 m e0 ie2 LA< ie2 l ai2 l a0 LA> (*toned IF string*)

3.2.3 Surface form Chinese IF layer

This layer provides the surface form pronunciation of the sentence(s) in Chinese IF roughly. To make the surface form more accurate, a super set of the canonical Chinese IF set should be defined, for example Zheng and Song (2001) define the generalized IF (GIF) set for this purpose.

For Chinese, the tones are often changed in spoken language, known as *tone-sandhi*. This should also be transcribed. There are some rules for tone-sandhi. For example, two adjacent 3rd tones (33) will mostly change into a 2nd tone followed by a 3rd tone (23). Depending on the tone of the following character, the tone for character "一/yi1/", "七/qi1/", "八/ba1/", or "不/bu4/" will also change according to the tone-sandhi rules, for example "一个/yi**2** ge4/". Such a kind of tone change is different from the ordinary tone change because the changed tone can be regarded as the quasi-canonical tone. To reflect such tone-sandhi rules, following the Pinyin of such a character is a two-digit tone string where the first digit is the original canonical tone while the second the tone-sandhi. For example, "一个/yi**12** ge4/".

3.2.4 Surface form SAMPA-C layer

Alternatively, the SAMPA-C symbols can be used to provide more accurate surface form transcription than the surface form IF layer, even though a generalized IF set is defined and used. In this layer, the observed tone is mostly attached to the SAMPA-C sequence of each Final.

3.2.5 Miscellaneous layer

Non-speech information is provided in this layer. Those spontaneous phonetic phenomena and spoken language phenomena (paralinguistic or non-linguistic) are also given in this layer.

Some symbols are listed in Table 7.

Table 7. Paralinguistic phenomena and non-linguistic phenomena

No	Phenomenon		Label	
			Start	End
1	Para-linguistic	Lengthening (拉长)	LE<	LE>
2		Breathing (吸气、喘气)	BR<	BR>
3		Laughing (笑)	LA<	LA>
4		Crying (哭)	CR<	CR>
5		Coughing (咳嗽)	CO<	CO>
6		Disfluency (不联贯)	DS<	DS >
7		Error (口误)	ER<	ER>
8		Silence (long) (长时间静音)	SI<	SI>
9		Murmur/Uncertain segment (不清发音)	UC<	UC>
10		Modal/Exclamation (语其次/感叹词)	MO<	MO>
11		Smack (咂嘴音)	SM<	SM>
12		Non-Chinese (非汉语)	NC<	NC>
13		Sniffle (吸鼻子)	SN<	SN>
14		Yawn (打哈欠)	YA<	YA>

15		Overlap (重叠发音)	OV<	OV>
16		Interjection (插话)	IN<	IN>
17		Deglutition (吞咽音)	DE<	DE>
18		Hawk (清嗓子)	HA<	HA>
19		Sneezes (打喷嚏)	SE<	SE>
20		Filled Pause (填充停顿)	FP<	FP>
21		Trill (颤音)	TR<	TR>
22		Whisper (耳语)	WH<	WH>
23	Non-linguistics	Noise (噪音)	NS<	NS>
24		Steady Noise (平稳噪音)	TN<	TN>
25		Beep (电话忙音)	BP<	BP>

3.3 Transcription Tools

There are many tools that can be used for transcription. Most of these tools have a lot of features in addition to the transcription function, such as analysis of pitch, formant, spectrum, cepstrum, and etc. In Table 8, three popular transcription tools are listed and compared.

Table 8. Transcription tools comparison

Tool Name / Item	Praat	SFS	X-Waves+
URL	http://www.fon.hum.uva.nl/praat/	http://www.phon.ucl.ac.uk/resource/sfs/	
Platform	PC, Mac, WS	PC	PC, WS
OS	Win, MacOS, Unix, Linux	Win, Unix, Linux	Unix, Linux (X-win)
Price	Free	Free	
Supported file formats	WAV, RIFF, AU, Binary raw, ASCII, etc	WAV, AU, AIFF, ILS, HTK, etc	AU, Binary Raw, ASCII, etc
Analysis	Pitch track, Epoch, Pitch modification, Spectrum, Formant, LPC, PSOLA LPC, Wavelet, Cepstrum, Excitation, Cochlea-gram, Vocal track analysis, ...	Resampling and speed/pitch changing, Fundamental frequency estimation (from SP or from LX), Spectrographic analysis, Formant frequency estimation & formant synthesis, ...	Pitch track, Epoch, Pitch modification, Spectrum, Format, LPC, etc
Labelling layers	Multiple	Single	Multiple

3.4 Learning from Databases

Databases exist not only for acoustic model training and system assessment. Knowledge can be learned from the speech corpus, especially for the spontaneous speech corpus. What can be learned from the corpus includes the phonetic knowledge that can be used to improve the acoustic model, and the linguistic knowledge that can be used for the natural language processing and understanding. In this section, examples will be given on this.

3.4.1 Knowledge for pronunciation modelling

In spontaneous speech, there are two kinds of differences between the canonical IFs and their surface forms if the deletion and insertion are not considered. One is the sound change from one IF to a SAMPA-C sequence close to its canonical IF, such as nasalization, centeralization, voiceless, voiced,

rounding, syllabic, pharyngealization, and aspiration. Such a surface form of an IF was called as its generalized IF (GIF) (Zheng & Song 2001). Obviously, the IFs can be regarded as special GIFs. The other is the phone change directly from one IF to another quite different IF or GIF, for example, initial /*zh*/ may be changed into /*z*/ or voiced /*z*/.

How to define the GIF set is a problem. However basically the first step should be to find all possible SAMPA-C sequences of all IFs.

The canonical IF set consists of 21 initials and 38 finals, totally 59 IFs. By searching in the CASS corpus (Zheng & Song 2001), we initially obtain a GIF set containing over 140 possible SAMPA-C sequences (pronunciations) of IFs; two examples are given in Table 9. However, some of them are least frequently observed sound variability forms therefore they are merged into the most similar canonical IFs because introducing least frequently observed GIFs will increase the pronunciation lexicon intrinsic confusion. Finally we have 86 GIFs.

Table 9. Examples of IFs and their possible pronunciations

IF (Pinyin)	SAMPA-C	COMMENTS
z	/ts/	Canonical
z	/ts_v/	Voiced
z	/ts`/	Changed to '*zh*'
z	/ts`_v/	Changed to voiced '*zh*'
e	/7/	Canonical
e	/7`/	Retroflexed, or changed to '*er*'
e	/@/	Changed to /@/ (a GIF)

In addition, probabilistic GIFs can also be learned so as to provide the GIF output probability distribution given an IF, i.e., $P(GIF \mid IF)$. The GIF N-Grams, including unigram $P(GIF)$, bigram $P(GIF_2 \mid GIF_1)$ and trigram $P(GIF_3 \mid GIF_1, GIF_2)$, can also be trained to give the GIF transition probabilities.

For spontaneous speech recognition, a multi-pronunciation lexicon is necessary to reflect the pronunciation variation. By learning form the database, we can get a probabilistic multi-pronunciation lexicon where the output probability is calculated as $P(GIF_1, GIF_2 \mid Syllable)$. Such a lexicon is called a probabilistic multi-entry syllable-to-GIF lexicon. In Table 10, probabilistic multiple pronunciations for Pinyin /*chang*/ are listed.

Table 10. Probabilistic multiple pronunciation for /*chang*/

Syllable (Pinyin)	Initial (SAMPA-C)	Final (SAMPA-C)	Output Probability
chang	ts`_h	AN	0.7850
chang	ts`_h_v	AN	0.1215
chang	ts`_v	AN	0.0280
chang	<deletion>	AN	0.0187
chang	z`	AN	0.0187
chang	<deletion>	iAN	0.0093
chang	ts_h/	AN	0.0093
chang	ts`_h	UN	0.0093

3.4.2 Knowledge for natural language processing

A spontaneous speech corpus often contains rich spoken language phenomena. For example, in the Chinese Spontaneous Telephone Speech Corpus in the flight enquiry and reservation domain (*CSTSC-Flight Corpus*), we summarize the following kinds of sentence templates (Zheng & Yan 2002):

- The courtesy items / sentences inessential for semantic analysis.
  C: 喂，你好，请问是中关村航空客运代理处么?
  Hi, hello, could you tell me ...?
- Simple repetitions because of the pondering or thinking when speaking.
  C: 我问一下那个四月三十呃四月三十号北京到...
  ... 30th April ... 30th April ...
- Semantic repetitions for emphasis.
  C: 请问那个周四就是四月三十号北京到...
  ... Thursday ... 30th April ...
- Speech corrections or repairs.
  C: 呃，那个什么星期三，呃，星期四的去南京的机票还有吗？
  ... Wednesday ... Thursday ...
- Ellipsis in the context.
  C: 我问一下那个四月三十呃四月三十号北京到福州的机票最后一班还有么? (Asking if there are tickets available for the last flight from the *departure city* to the *arrival city*, on the *departure date*.)
  O: 只有一班有。("Only one flight with tickets available")
  C: 那个那五月一号的下午三点有么? (How about the flight at a *departure time* on another *departure date*?)
- Constituents appearing in any order (as long as the sufficient information is given).
  C: ...五点二十五国航飞深圳的... (Time, airline code, location and some other items can appear in any order)
- Constituents appearing in reverse order.
  C: ...的机票 多少钱 得?
  How much cost
  (Normal order: "得" "多少钱")
- Parol (verbal idioms) or unnecessary terms.
  C: 那，那个八点二十那个是去什么机场的呀？("那"/"那个" is somewhat similar to "*uhm*")
- And long sentences with all required information. Or additional explanation following the previous sentence.
  C: 哎，您好，这样那个我订一张(one)那个明天(tomorrow)下午(afternoon)五点(5 o'clock)四十五(45)去北京(from Beijing)到上海(to Shanghai)的那个机票(ticket)的。

  C: 您给我看一下有没有(is there)北京到湛江(from Beijing to Zhanjiang)的(ticket)？二十九(29^{th})、三十号(30^{th})?

In the above dialogues, "C:" leads the customer's utterances while "O:" operator's.

Based on the above information, four types of parsing rules are proposed for robust spoken language understanding (Yan & Zheng 2001, Yan & Zheng 2002). They are:

1. ***Up-tying rules***. The up-tying rules are needed in at least one case when the customer's ID card number is to be parsed where the ID card number is taken as a crucial piece of information forbidden to be inserted by or mixed with other terms.
2. ***By-passing rules***. A large number of rules are of this type, which is based on the assumption that the input keyword string contains recognized fillers/rejections, speech fragments or some other nonsense parts. E.g., "星期啊三嗯星期四" ("week *ah* three *en* week four"/Wedn-*ah*-esday and *eh* Thursday) is admitted if the by-passing rules exist.

3. ***Up-messing rules***. The up-messing rules are required in case no matter what order sub-constituents may follow. In the flight domain, constituents of time, location, and plane type can be grouped by up-messing rules since the user can tell them in any order.
4. ***Over-crossing rules***. Some concepts, which can be defined as the task-relative minimal elements, may be derived from several different by-passing rules and can be used to form other constituents. Over-crossing rules are used to avoid the definition of many similar rules in this case. E.g., "是不是 (be or not) *confirm_c*?" , "*confirm_c*是不是?", "*confirm_c*是吗(be or not)?", and "是(be) *confirm_c*吗?" can be described by a single over-crossing rule.

4 Chinese Corpus Consortium (CCC)

It is obvious and well known that the speech and text databases are very important to the research and development in the areas of automatic speech recognition, text-to-speech, natural language processing and etc. And there have been several consortiums or associations devoting themselves in collecting and distributing such data resources, such as the Linguistic Data Consortium (LDC 1992), the European Language Resources Association (ELRA 1995), and The Gengo Shigen Kyouyuukikou Language Resource Consortium (GSK 1999). However, there is not such a consortium or association for the Chinese language that about one fifth of the world population speaks in.

After communications and discussions for a long time, now the Chinese Corpus Consortium (CCC 2003) comes out.

CCC will provide corpora for, but not limited to, Chinese ASR, TTS, NLP, perception analysis, phonetics analysis, linguistic analysis, and so on. Corpora could be speech and text, read and spontaneous, wideband and narrowband, Putonghua, dialectal Chinese and Chinese dialects, clean and with noise, and of any other kinds which are helpful to the foresaid purposes.

More details can be found in the CCC web site (http://www.d-Ear.com/CCC/).

Acknowledgements

Thanks go to Dr. Jing Li and Dr. Zhenyu Xiong for their implementing the algorithm of sentence selection for read speech database collection and their design of topics for spontaneous speech database collection. Thanks also go to Dr. William Byrne (the Johns Hopkins University), Prof. Lianhong Cai (Tsinghua University), Prof. Aijun Li (Chinese Academy of Social Sciences), Dr. Helen Meng (Chinese University of Hong Kong), Dr. Satoshi Nakamura (ATR), Prof. Lua Kim Teng (Chinese and Oriental Language Information Processing Society), Prof. Hsiao-Chuan Wang (Association for Computational Linguistics and Chinese Language Processing), Prof. Wenhu Wu (Tsinghua University), Miss Xia Wang (Nokia), and Dr. Mingxing Xu (Tsinghua University), for their effort in providing the database information and in the formation of the CCC.

References

CCC (2003). The Chinese Corpus Consortium, founded in 2003 in China. http://www.d-Ear.com/CCC/

Chen X.-X., Li A.-J., *et al* (2000). *An application of SAMPA-C for standard Chinese*. International Conference on Spoken Language Processing (ICSLP'2000), Oct. 16-20, 2000, Beijing

ELRA (1995). The European Language Resources Association, founded in Feb., 1995 in Europe. http://www.icp.inpg.fr/ELRA/

Fung P., Byrne W., Zheng F., Kamm T., Liu Y., Song Z.-J., Venkataramani V., and Ruhi U. (2000). *Pronunciation modeling of Mandarin casual speech*. Final Report for Workshop 2000 for Language Engineering for Students and Professionals Integrating Research and Education, http://www.clsp.jhu.edu/ws2000/final_reports/mpm/

GSK (1999). The Gengo Shigen Kyouyuukikou Language Resource Consortium, founded in May 1999 in Japan. http://tanaka-www.cs.titech.ac.jp/gsk/gsk-eng.htm

LDC (1992). The Linguistic Data Consortium, founded in 1992 in USA, http://www.ldc.upenn.edu/

Li A.-J., Zu Y.-Q., Li Z.-Q. (1999). *A national database design and prosodic labeling for speech synthesis*, Oriental COCOSDA '1999, pp. 13-16, May 13-14, 1999, Taipei, Taiwan, China

Li A.-J., Chen X.-X., *et al* (2000). *The phonetic labeling on read and spontaneous discourse corpora*. International Conference on Spoken Language Processing (ICSLP'2000), Oct. 16-20, 2000, Beijing

NOISEX92 (1992). http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

Kuwabara H., Nakamura S., Itahashi S., Lee Y.-J., Zheng F. ,Li A.-J., Wang R.-H., Dawa I., Wang H.-C. (2002). *Overview of Recent Activities of Corpus Development in East Asia*. COCOSDA 2002

Varga A., Steenneken H. J. M, Tomlinson M., and Jones D. (1992). *The NOISEX-92 study on the effect of additive noise on automatic speech recognition*, 1992. Documentation included in the NOISEX-92 CD-ROMs.

Wang Z.-Y., Sun J.-S., Xiao X., Wang X., (1999). *A minimum corpus designed for training the acoustic model*, Oriental COCOSDA '1999, pp. 77-80, May 13-14, 1999, Taipei, Taiwan, China

Yan P.-J., Zheng F., and Xu M.-X. (2001). *Robust parsing in spoken dialogue systems*. EuroSpeech, 3:2149-2152, Sept. 3-7, 2001, Aalborg, Denmark

Yan P.-J., Zheng F., Sun H., and Xu M.-X. (2002). *Spontaneous speech parsing in travel information inquiring and booking systems*. Journal of Computer Science and Technology, pp. 924-932, Vol.17, No.6, November 2002

Zheng F., Song Z.-J., Fung P., and Byrne W. (2001). *Modeling pronunciation variation using context-dependent weighting and B/S refined acoustic modeling*. EuroSpeech, 1:57-60, Sept. 3-7, 2001, Aalborg, Denmark

Zheng F., Yan P.-J., Sun H., Xu M.-X., and Wu W.-H. (2002). *Collection of a Chinese Spontaneous Telephone Speech Corpus and Proposal of Robust Rules for Robust Natural Language Parsing*. Joint International Conference of SNLP-O-COCOSDA 2002, pp. 60-67, 09-11 May 2002, Hua Hin, Thailand

Zheng F., Song Z.-J., Fung P., and Byrne W. (2002). *Reducing pronunciation lexicon confusion and using more data without phonetic transcription for pronunciation modelling*. International Conference on Spoken Language Processing (ICSLP) 2002, pp. 2461-2464, Sep. 16-20, 2002, Colorado, USA

Zu Y.-Q. (1997). *Sentence design for speech synthesis and speech recognition by phonetic rules*, EuroSpeech, 2: 743-746, 1997

Zu Y.-Q. (1999). *The text design for continuous speech database of standard Chinese*. Chinese Journal of Acoustics, 18(1): 56-59, Jan. 1999