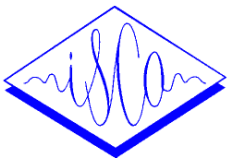

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

THE DESIGN AND APPLICATION OF A SPEECH DATABASE FOR CHINESE TTS SYSTEM

Muhua Lv *Lianhong Cai*

Department of Computer Science and Technology, Tsinghua University,
Beijing, P.R.China, 100084
Email: lmh@tts.cs.tsinghua.edu.cn

ABSTRACT

The design and application of a speech database for Mandarin TTS system is presented in this paper. To build a scientific, versatile speech database to meet the call for improving the quality of synthesis units and enhancing previous prosodic models, is the main point of the research. The database structure and contents and the methodology for creating similar database are described, and also some statistics and some research based on the database.

Key Words:

speech synthesis, speech database, chunk, prosodic characteristic

1. INTRODUCTION

With the development of the quality of synthesis speech, Text-to-Speech technology has become more reliant on speech database. Moreover, "data driven", the new approach that considered by more and more TTS researchers, has raised a higher requirement into the designing of speech database. Previous database cannot reach the increasing requirement.

The result comprises a well-designed speech database chiefly made up of continuous speech. It is the first time to get synthesis units directly from connected natural speech in Chinese Mainland. Most units used by synthesis system are chosen from prosodic chunks, the main part of the database, while corpus in other formats is designed mainly to get disciplines of prosodic character and to improve prosodic models. The whole database is also to serve as first training materials in a new "data driven" system.

In order to model prosody well, it is important to include many factors. The databases however do not contain all those factors. In the database designing, some factors are chiefly considered, such as covering most initial and final consonant combinations, covering different tone combinations, etc. Current information in the database is limited to the orthography, the pinyin data, sound recordings and corresponding prosodic labels.

2. DATABASE DESIGN

The database comprises various formats of Chinese text, each has its own emphasis and design goal

2.1 Prosodic chunks

This is the main part of the database. It is tentative to get synthesis units directly from continuous speech because units from single character or multi-syllable word appear too isolated to make up a whole natural sentence. However, it is quite creative to try "prosodic chunk" in the research. In continuous speech, there are a lot of short sentences (6 to 10 Chinese characters). The interaction between two syllables from different short sentences can be ignored in most cases. These short sentences, can be taken as one independent chunk each. And, according to the characteristic of human speech, when pronouncing a long sentence, it's quite similar to several chunks, while syllables from different chunks are of little influence to each other. When extracting synthesis units from chunks instead of from words, the environment information of the units will be kept; and comparing with extracting from whole sentences, it has remarkably reduced the scale of the database and improved the universality of those units. The prosodic research based on chunks will keep most useful information such as allophones, intonations, etc. and the research result will be easy to apply in synthesizing connected speech.

The prosodic chunk is often of six to eight words or so, varying with different context, usual larger than word

or phrase, but as part of a sentence. However, in some cases, the chunk is just a phrase or just equal to the sentence.

In this part, the chunks cover all phones, in its every possible position in Putonghua, which is part of good preparation to build a multi-lingual speech synthesis system and cover most tonic syllables in Putonghua.

2.2 Single syllable

This part comprises all tonic single syllables in Putonghua, each pronounced as separate word, designed to provide comparison to the same syllable in other contexts, and also as synthesis units in some single word cases.

2.3 Multi-syllable

This part comprises di-syllable, tri-syllable and four-syllable words, covering all possible tone combinations despite the actual initial or final. Since each tone combination has relatively steady pitch contours form, an important factor effecting the quality of synthesis, it is very necessary to study the inherent relation between them. This part helps to build tone model.

2.4 Neutral syllable groups

Neutral syllable plays an important role in natural speech, however, current research on neutral syllable is yet far from complete, and the quality of the neutral syllable produced from corresponding normal syllable is not satisfying. The database comprises nearly all most often used neutral syllables to study its relationship with corresponding normal one and to produce neutral syllable when synthesizing.

2.5 Retro syllable groups

Similar to above, the database comprises hundreds of retro syllables to cover different retro finals. This part is prepared to produce retro syllables in output speech.

2.6 Sentences

The database also has hundreds of complete sentences, comprises sentences of different mood, length and sentence structure. This is for study on intonation and sentence mood and also the further research on producing continuous sentences with better naturalness. And these sentences are also the base of most chunks, and this provides comparison of similar context in sentences and in chunks when necessary.

The database comprises these six parts as above according to different formats of text, each has sound recording files, prosodic labels and corresponding text and pinyin data. Data were processed in the following order:

- Text selection

Six formats of text as mentioned above are selected to create the database.

- recordings

A male and a female speaker are chosen to make the recordings. The speaker was asked to read the text clearly, smoothly and impassively. The text were recorded into digital cassettes and then transferred to PCM wave files into a computer.

- phonetic transcription

The pinyin and text data of the sentences were checked against the corresponding recordings. This step was done by listening to the recordings and editing the data.

- Phonetic segmentation and labeling

First a program did phonetic segmentation to the wave files automatically, marked the start and end point of each syllable. And this was then checked by manual. Then the program draw out the pitch contours.

- Prosodic transcription

Only part of the database has prosodic transcription now. Current transcription is according to a published labeling text design.[1] The final transcription standard is still in developing.

3. CHUNK PRODUCTION

Prosodic chunks constitute the main part of the database. The production procedure of these chunks is as follows:

3.1 Text preprocessing

The chunks was produced from Renmin Daily. The text of the newspaper is first splitted into separate sentences,

and those have special punctuation that is difficult to deal with are discarded then. With the text analysis model of our TTS system, corresponding pinyin and word segmentation information is automatically generated.

3.2 Improved greedy algorithm

According to some conditions such as initials and finals distribution, improved greedy algorithm is applied to extract required sentences from large quantity of text.

Greedy Algorithm scans the input text sequentially and adds a sentence as long as it has a new unit. It's easy to implement and maximizes the coverage. However, this produces a lot of redundancy and is greatly influenced by the sequence of input material.

Improved Greedy Algorithm used here, made some modifications on previous algorithm to avoid these shortcomings. First, it sorts the candidate sentences according to their value, i.e., how many new units each sentence will bring if selected. It is to ensure that the choice is of high efficiency. Second, after a period of processing, scan all the chosen sentences and get rid of those of little value. This helps to reduce redundancy and to produce a better distribution.

3.3 chunk production

After the pinyin and word segmentation are collated by manual, the sentences and relevant pinyin are marked with word margin. Then those skilled in the art marked the sentences with chunk margin where short intervals may be added according to usual pronunciation.

From these new generated chunks, Improved Greedy Algorithm is again applied to select final chunks. And there were some chunks directly composed to fulfill better coverage.

A major priority of this method is the automation of text selection. Automation remarkably reduced the time to produce a new prosodic database. And, Improved Greedy Algorithm, comparing with previous one, reduced redundancy, i.e., reduced the proportion of those often used units but raised that of those scarce units.

4. DATABASE STATISTICS

Here some basic statistics is presented. As the main part of the database, the prosodic chunks, in order to provide synthesis units, it should provide a full-scale coverage in a manner; and, to get reasonable prosodic model, especially using "data driven", the distribution should be similar to that of natural speech.

4.1 Units coverage

Phones(initials, tonic finals) at possible different position (beginning, middle, end) : 100%

Existing tonic syllable(neutral and retro syllable not taken into consideration) : 93.9%

4.2 Units distribution

The percentage of initials, finals and tones are presented here. In the following tables, the lines marked with 'C' mean 'current database', show the statistics of this database, while the lines marked with 'R'(Reference), show the result of Putonghua natural speech as reference. [2]

Table1. Percentage of tones

声调		1	2	3	4	5
概率 (%)	C	20.25	21.61	17.63	32.56	7.95
	R	18.71	19.37	17.51	35.78	8.63

Table 2. Percentage of initials

声母		零声母	b	p	m	f	d
概率 (%)	C	13.34	4.23	1.62	3.42	2.88	10.03
	R	12.45	5.15	0.98	3.74	2.45	12.0
声母		t	n	l	z	c	s
概率 (%)	C	3.61	2.14	5.97	3.33	1.44	1.63
	R	3.53	2.53	5.69	3.01	1.15	1.08
声母		g	k	h	j	q	x
概率 (%)	C	5.14	1.98	4.39	7.11	3.47	5.04
	R	5.50	1.98	4.42	6.98	3.11	4.86
声母		zh	ch	sh	r		
概率 (%)	C	6.69	3.34	7.06	2.13		
	R	7.18	2.74	7.66	1.94		

Table3. Percentage of finals

韵母		a	ai	an	ang	ao	o
概率 (%)	C	3.14	4.21	3.76	3.13	3.02	0.37
	R	**3.89**	2.83	3.41	2.87	3.10	0.54
韵母		ou	e	ei	en	eng	er
概率 (%)	C	2.04	9.85	1.70	2.97	3.28	0.38
	R	1.88	12.38	1.28	3.62	3.09	0.28
韵母		E	-i	i	ia	ie	iao
概率 (%)	C	~0	5.87	8.83	**1.06**	2.11	**2.21**
	R	~0	6.41	8.80	1.09	**2.42**	2.06
韵母		iou	ian	in	iang	ing	u
概率	C	2.53	**4.44**	**2.49**	**2.09**	3.67	**6.82**
	R	2.60	**4.10**	1.95	1.80	**3.05**	7.11
韵母		ua	uo	uai	uei	uan	uen
概率	C	**0.75**	3.25	0.54	2.77	1.80	1.15
	R	**0.44**	**4.40**	0.32	**2.75**	**0.85**	0.89
韵母		uang	ueng	ong	v	ve	van
概率	C	**0.91**	**0.006**	**3.68**	**2.16**	0.92	1.06
	R	**0.65**	0.003	0.65	1.80	1.01	0.89
韵母		vn	iong				
概率	C	**0.50**	0.50				
	R	**0.52**	0.42				

From the tables above, it is not difficult to find that the distribution of units in the database is quite similar to previous result given as in common Putonghua speech.

However, as mentioned above, for those scarce units, the percentage will be raised in this database because of the principle used in Improved Greedy Algorithm to generate these chunks.

Similar stat about other factors has been done on the database.

Finally, it is proved in some manner that the database may be used as an epitome of Putonghua natural speech and results from training on these material are reasonable and practical.

5. SOME APPLICATION

5.1 A new TTS system based on this database

A new TTS system using this database as its synthesis units database has been created.

Comparing with the previous system which based on isolated words, the system appears much better naturalness. It has on the whole overcome the obvious shortcoming of last system that the conjunction between words sounds inarticulate. And more intonation seems to be embodied in the new system. However, some units didn't appear to act well when put into other context, which may need more transcription.

5.2 Prosodic models

Various prosodic characters, such as duration, pitch and range, are been in study on this database. Some pilot study shows, the distribution of these characters do have some rules and even an approximate goal value can be draw out.

6. CONCLUSION

We are currently concentrating on developing prosodic model using the data in these databases and will test these models using synthesized speech. With those results in hand, we will be able to confirm in a more definitive way the adequacy of our database design methods. However, more detailed prosodic transcription is needed for further research.

7. REFERENCES

[1] Wu, Zongji, 1997, Toward a projiect of All-Phonetic-Labelling-Text(APLT) for TTS synthesis of spoken Chinese. In Proceedings of the first China-Japan workshop on spoken language processing (CJSLP '97), pp.26-36.

[2] Chen Yongbin, Wang Renhua, Language signal processing, pp. 54-55.