The development of a waveform synthesis system based on phonetic rules

Li Aijun, Cai Dehe, Hua Wu and Chen Xiaoxia

基于语音规则的普通话半编辑式语音合成系统的研制报告

李爱军　蔡德和　华武　陈肖霞

[摘要]

1．问题的提出:

语音合成是利用电脑或数字信号处理技术来产生人类言语（如音素、音节和句子）的一种技术．它涉及语言学、语音学、数字信号处理、计算机科学及心理学的知识，是一门新兴的边缘学科．它不但可以用于人机通讯系统，而且也是语音学研究的一种手段。

语音合成技术可以分为波形编码合成、参数分析合成、规则合成和文语转换等四类。其中波形编辑合成的音质最好，但存储量最大；规则合成的存储量最小，但音质在目前来说不太好．

我们的目的是开发一种性能价格比高，即音质尚好，占用内存很小的合成系统，适用于掌上机的语音输出，如校对、机器翻译的语句输出等．本文第一作者一直从事共振峰的规则合成的研究，从这里面受到一点启发，提出一种基于语音规则的普通话半编辑式语音合成方法．“半编辑”指合成单位选用普通话声母和韵母的变体，它是相对于下面的“全编辑”，即全音节编辑合成而提出的；“语音规则”是指合成单元的产生、音节合成的拼接规则。简言之，就是以声母、韵母的变体做合成单元，依据一定的语音学规则，将音节对应的声母和韵母拚接在一起．这种方法，可以将音节内的声学过渡很好地考虑进来。

2．系统实现的几个步骤

（1）确定合成的声韵单元

这一步是本系统的难点与重点．包括录音、切音、合成试听。

录音工作在语言所的录音室里进行，发音人是一名经过挑选的讲标准普通话的男性．材料包括1275个全部单音节，以及特殊字母、字符的读音。

切音是在5500语图仪和ASL上进行的。首先，用5500语图仪分析某一类声母开头的音节（CVN，C代表声母辅音，V代表韵母，N代表韵尾），再按声韵过渡不同细分为几类，根据声韵过渡的特点，选择一部分音节做声、韵母切分，过渡部分切给声母(C(V))．切分的这类声母单元经ASL存入计算机（采样率10KHZ，精度为了16bit），再通过ASL将切分的声母单元与韵母单元拼接起来，送到5500中回放，与原始的单节比较，若不满意重新选择切音音节，直到切到满意的单元为止．对于一些“顽固的”音节，只好用全音节了。

必读轻声音节的合成，采用韵母时长缩短一半，能量减半的语音规则。一些特殊字母、字符，不做切分．

最后我们得到的合成音库，有89个声母变体，149个韵母变体，加上特殊的字母、字符共有352个合成单元．

（2）合成系统的软、硬件的设计

图 1 给出了系统的原理模块图。输入文本通过分词模块分词，得到停顿标志，本系统设置了段落、语句、词组三级停顿，时长分别为 500ms 、 300ms 和 50ms 。通过多音字词库检索出多音字的读音，通过两级检索得到声母、韵母合成单元，拼接以后由 D/A 板（ 8bit ）放音。

（3） 全编辑、半编辑和规则合成系统的评价

一般对语音合成系统的评价可以从不同的层次进行，即可以是音段的清晰度，也可以是一个字的可懂度。对合成系统的评价可以是不同的系统间的各种性能的比较，也可以是对某个系统采用不同技术的比较。这里给出的全编辑、半编辑和规则合成系统的评价属于前一种情况。

评测的语料用声学手册的 3 个音节清晰度测试表和 3 个单词可懂度测试表，共有 225 个音节， 300 个词。并用伊索寓言“北风和太阳”的故事测试句子的自然度。共有 4 位听音人参加了听音，他们将听到的内容写在答卷上，最后统计出出错率和标准差：半编辑合成系统的音节和词的出错率分别是 0.04%、 0.02%全音节合成系统的音节和词的出错率分别是 0.007%、 0.018%，共振峰式规则合成系统的音节和词的出错率分别是 0.14%、 0.04%。从结果可以看到，半编辑合成系统的音节和单词的清晰度与全编辑系统相差无几，而句子的自然度都不如规则合成系统。这也许是这种方法的天生的缺点。所以目前又有了新的基频同步的拼接合成法，如果将这种方法用进来也许会是一种不错的方法。

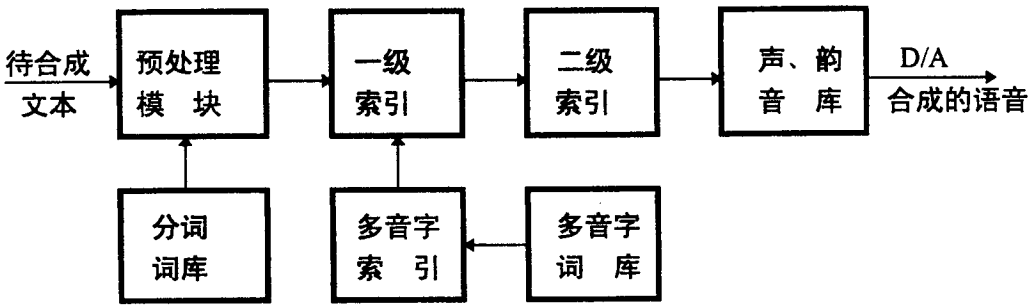


图 1. 半编辑合成系统原理框图

Abstract

It is well known that speech synthesis technology is devided into four types, waveform synthesis, parametric analysis synthesis, rule based synthesis and text-to-speech synthesis. Waveform synthesis generates the speech-sound with the best quality ,but the largest memory capacity is required. Then rule based synthesis needs the smallest memory capacity but the quality is not satisfactory so far.

The aim of this paper is to establish a proper performance-cost ratio system, which can be used in PDA as the speech output system, such as a checking system or a teaching system like tutor or speech out of a machinery interpretation system. As the first author is currently working on a rule-based TTS, she drew a great deal of inspiration from her experience and presented a waveform synthesis based on phonetic rules which adopt initial unit C(V) and final unit V(N) as the synthetic unit. The transitional information within the syllable is contained in C(V). When a

syllable is synthesized, the corresponding C(V) and V(N) are concatenated via phonetic rules.

It is an important process to get the synthetic inventories by segmenting the syllables using phonetic rules. This is carried on KAY SONOGRAPH MODEL 5500 and ALS. Finally we get 352 inventories containing 89 initials C(V) and 149 finals V(N) and other special symbols. Total memory capacity is 500kB which is 10 percent of that of the waveform synthesis in which syllable is used as the synthetic unit.

The first section will decribe how to segment the syllables prerecorded in Beijing Mandarin. The second part will illustrate the software and hardware. The last section will give an overview of the evaluation of the system.

1. Getting the synthetic inventories

1.1. *Recording*

Our informants selected here is a male who can speak Beijing Mandarin fairly fluently. The speech materials are 1275 syllables (CV(N)) and some special letters and alphabets. The recording is done in the recording room in our laboratory.

1.2. *Segmenting using phonetic knowledges and getting phonetic rules*

Here segmentation is meant to segment a syllable into initial part C(V) and finial part V(N). First all the utterances are analyzed by SONOGRAPH 5500 with 10KHz sample rate and 16bit precision. From the wideband spectrum on 5500 screen, the initial and final of a syllable are clearly presented. Any part of the spectrum between cursors can be heard and also can be stored and analyzed by ASL.

Take as an example the syllable whose initial is /p/ to illustrate this processing. First it is hypothesized that all of the finial V(N) inventories have been got. The next step is to solve the problem of obtaining initial inventories. As /a,i,u/ are three extreme articulatory positions,the transition from /p/ to /a-,i-,u-/ are quite different, so /p(a)/ /p(u)/ /p(i)/ are three inventories we must get.But for syllable /pe-/ or /po-/, is it necessary to segment another two initial inventories /p(e)/ and /p(o)/? From the phonetic point of view, when you speak /po/, the lip's movement is /p/--> /u/ --> /o/, so the transitional information in /p(u)/ is enough for /p(o)/, and therefore /p(o)/ is omitted. After observing the spectrum of the syllables /pe-/ and making a synthetic test, we think /p(e)/ can also be omitted by using /p(a)/ instead.

Then a phonetic concatenation rule is formed :

if syllable /po-/ or /pu-/ is to be synthesized , then /p(u)/ is used as the initial;

if syllable /pa-/ or /pe-/ is to be synthesized , then /p(a)/ is used as the initial;

if syllable /pi-/ is to be synthesized ,the /p(i)/ is used as the initial.

As for the initial /p(a)/ ,whether it is got from the syllable /pa/ or from the syllable /pa-/, it can only be decided by synthesis test and resegmenting of the syllables. So we say it is a very sophisticated work. The waveform of the special symbols, e.g. English letters, Chinese punctuation marks are stored directly by ASL.

Finaly 352 inventories are obtained containing 89 initials,149 finals and other special symbols.

When a neutral tone syllable is synthesized the phonetic rule is to half of the final duration and the amplitude.

2. Software and hardware

2.1. *Creating index database and speech database*

The record structures of the index and speech database are given in Fig.2, in which v0.idxs, v1.idx, v2.idx and pwidx1.idx are four index files, voice.dat and pw.dat are speech database and word database for polytonal syllables respectively.

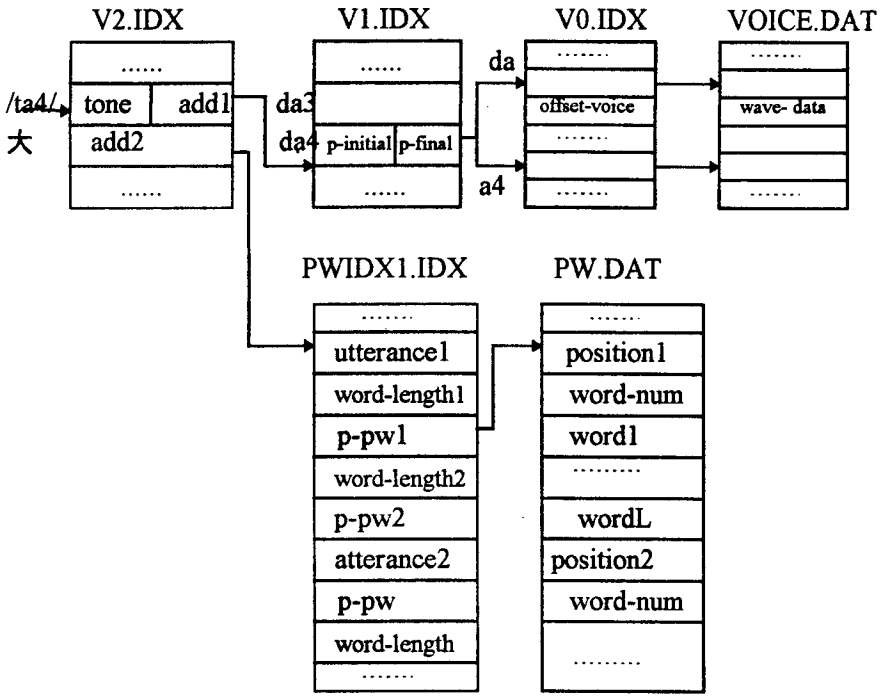


Fig.2 The structure of the index and speech database
图 2. 索引文件和语音数据库的结构

B0, B1, B2 and dyc.dat given in Fig.3 are four ASCII files to create the index files and database files. B0 is a table containing all the 352 inventories. There are 1273 tonal monosyllables of Beijing Mandarin listed in B1 with the first column being the Chinese PINYIN and the 2nd and the 3rd column being the corresponding synthetic inventory code number of the initial and final respectively. B2 is a table containing 6763+89+15 records which has four items. The 1st item is the Chinese characters exchanged code (GB2312-80) from the 16th zone 啊 to the 87th zone 齄, and the special symbols are arranged from the 88 zone to the 89 zone. The 2nd item is the tone number(0-4,0 for neutral tone). If the syllable is polytone, the tone number is the most often used one. The 3rd item is the number of polytone, and the last item is the corresponding utterance

number in B1. For a polytonal syllable this item is corresponding the most often used utterance. Other utterances are listed in DYC.DAT. The organization of DYC.DAT is illustrated in Fig 3(d). The number stands for the number of utterances of the polytonal syllable next to this number except in B2. The utterance of the syllable next to * is the alternative utterance of this polytonal syllable. For example there are 3 utterances of 差 in SC, the most often used utterance is /tʂʻa1/ which is listed in the B2, and the other two utterances /tʂʻa4/ and /tʂʻai1/ are listed in dyc.dat. When using this strategy, we can reduce the capacity of DYC.DAT and retrieve it faster.

After these four files having been established, the index file and the database can be created easily then. V0.IDX and VOICE.DAT are created from B0 and the waveform data. V1.IDX is produced from B1. V2.IDX is formed from B2 and modified as we establish PWIDX1.IDX by DYC.DAT. Word database for polytonal syllable PW.DAT is produced from DYC.DAT.

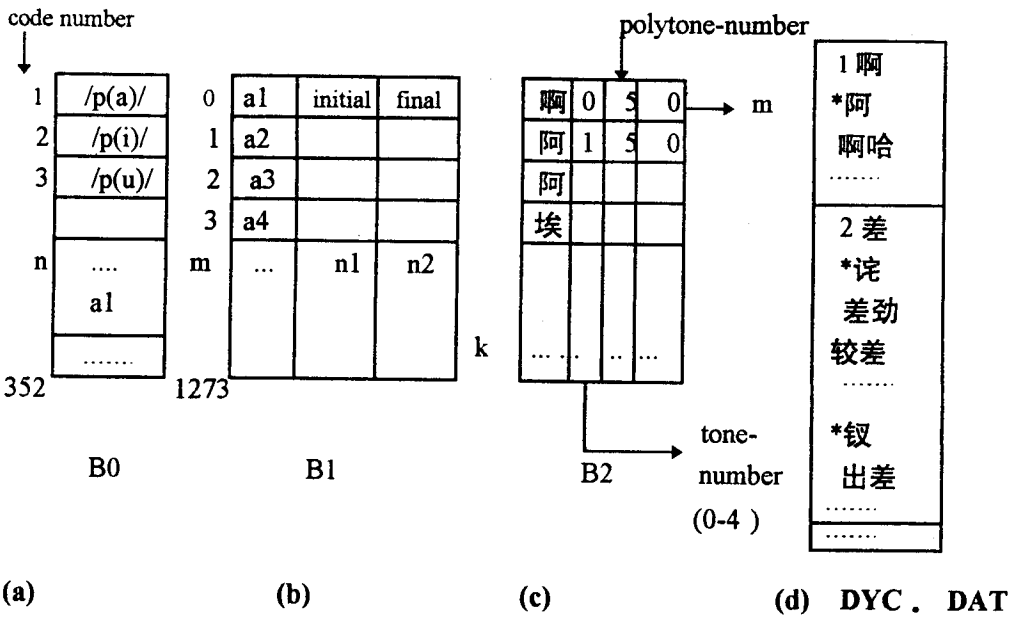
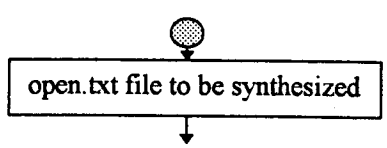


Fig.3 The structure of B0, B1, B2 and DYC.DAT

图 3. 文件 B0, B1, B2 和 DYC. DAT 的结构

2.2. *Programming main program*

Fig.4 is the flowchart of the retrieve program which has an internal code of a Chinese character as its input and the waveform speech as its output. The word parsing method does not describe here, because it adopts the most ordinary one as often use in TTS.



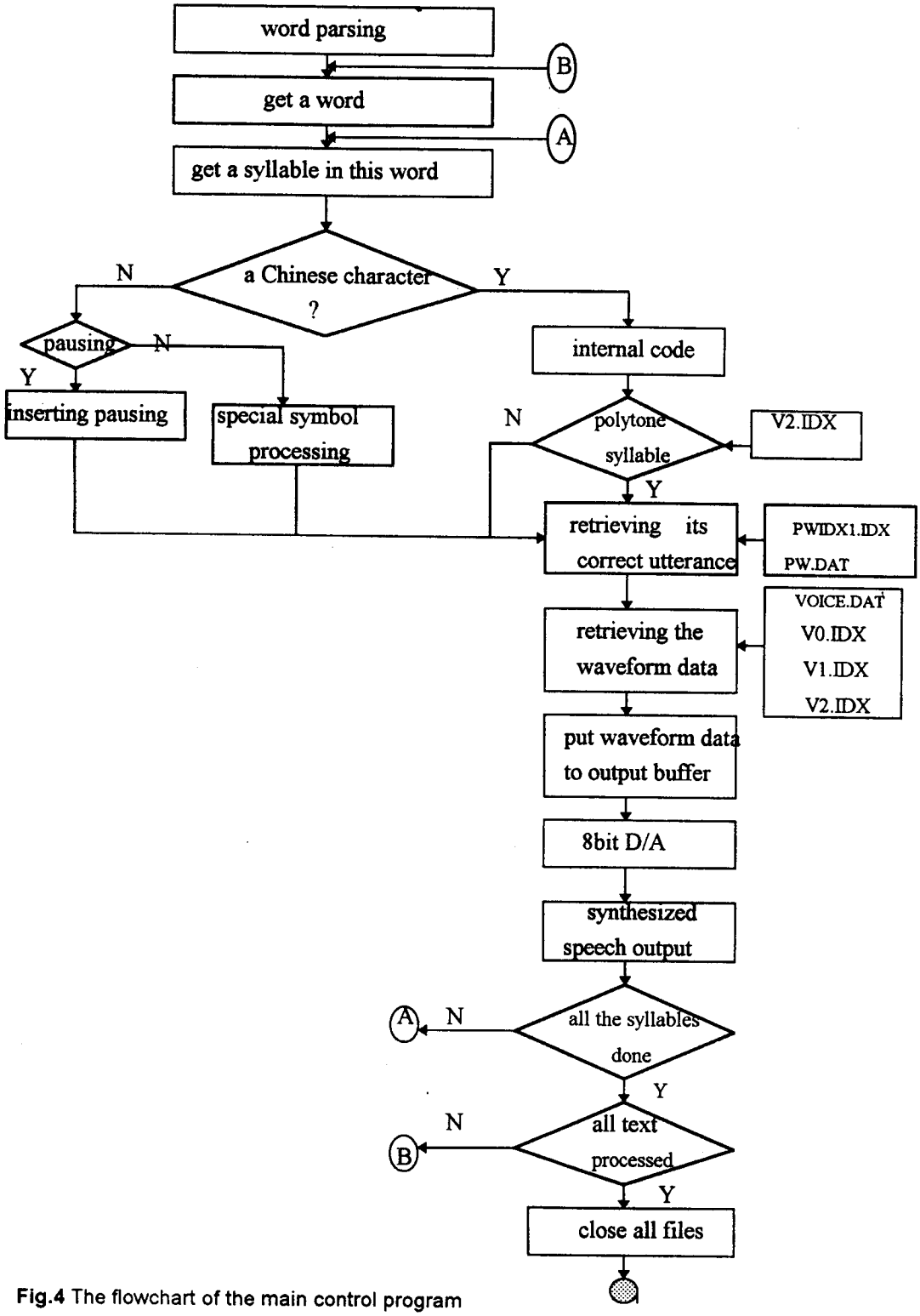


Fig.4 The flowchart of the main control program

图 4. 主控程序流程图

2.3. *Designing Speech output board(D/A)*

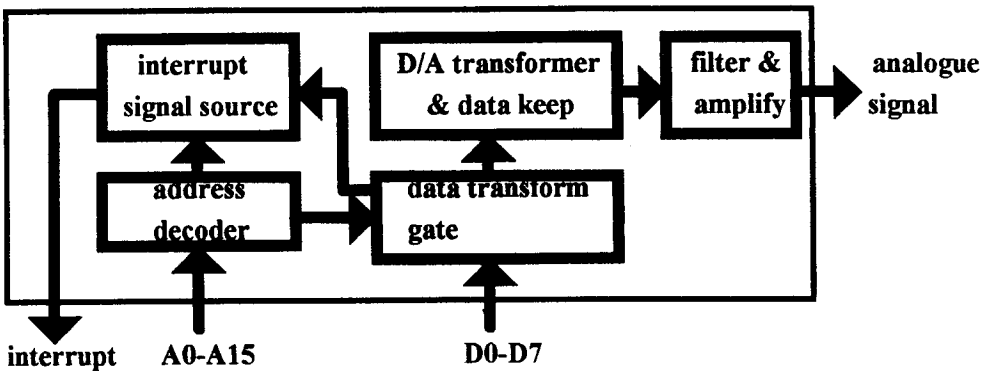


Fig. 5 The principle of D/A board

图 5. D/A 板的原理框图

The principle of D/A print board is given in Fig. 5, which contains three main part:

(a) setting-up the board. The interrupting signal is setup through the specific address and data from computer and make an interrupting request of 10KHz to the computer.

(b) working process of D/A board. The data transform gate is opened by the working address code from computer. And then the data is transferred through D/A and filtered and amplified before being outputted.

(c)closing D/A board. It is the computer that controls the closing of the interrupting signal source and D/A working.

The main parameters used in programming the D/A control module are

(a) port address . data port 0000???? ???? 0010 ; control port 0000???? ???? 0000

here ? stands for the DIP switch value which can be set by the user on D/A board. ?=1 if it is ON; ?=0 if it is OFF.

e.g. address of data port 3B2 corresponds the DIP state: 0000 **0011 1011** 0010

(b) sample rate:10Khz

(c) D/A presion:8 bit

(d) output voltage:+-0.6v

3. Evaluating three systems

Speech synthesis system can be evaluated on different levers ranging from segmental intelligibility to sentence acceptability and naturalness. The evaluation can be carried on between different synthesis systems to obtain a comparative scoring of the system. It can also be carried on in an individual system to evaluate a change.

We have three systems with different synthetic methods i.e. half-waveform synthesis, waveform synthesis and rule based formant synthesis. So the evaluation given here is belong to the latter one.

The test materials cover three tables for testing syllabic articulation, three tables for testing word intelligibility, and a short story in CORPUS FABULARUM AESOPICARM called "Northwind and the Sun" is used to test the sentence naturalness. The intervals between syllables and words are 3 seconds.

Table 1 Error rate and standard deviation(SD) of the three systems(%),s1 s2 s3 stand for half-waveform synthesis, waveform synthesis and rule based synthesis respectively.
表 1. 三个系统的错误率和标准差，s1, s2, s3 分别代表半编辑系统、全编辑系统以及规则合成系统.

	monosyllable		word	
	mean	SD	mean	SD
S1	4	18	2	14
S2	0.7	4.3	1.8	15
S3	14.2	87	4.1	15

Four subjects listened to the speech sounds generated by the three systems in a soundproof cubicle, each containing three sessions: monosyllable test(225 syllables), word test(300 words) and a story test. The subjects are requested to write down what they heard on the answer sheet. The error rate and the standard deviation were given in table 1. The naturalness of S3 is the best, and the others are almost the same. From this preliminary test it can be seen that the coarticulation of the half-waveform system is better than the rule based one and is very closer to the waveform one. So we have come closer to the goal of the smaller memory capacity and higher syllable and word intelligibility.

References

Eric, keller (1994) *Fundamentals of speech synthesis and speech recognition.*Chichester: John, Wiley & Sons.

Luo, Nian-sheng et.al. complain (1981) *Corpus fabulatum aesopicarum.* Beijing: Ren Min Wen Xue Press.

Ma, Dayou & Shen, Hao (1983) *The handbook of acoustics.* Beijing: Science Press.

Rolf, Carlson, Bjorn, Granstеrom & Lennart, Nord(1990) Evaulation and development of the KTH text-to-speech system on the segmental level, *Speech Communication*, 9(4), 271-277.

Yan, jingzhu (1993) A study of vowel formant pattern and coarticulation in the voiceless stop initial monosyllable of Beijing mandarin. *Proeceedings of the 6th national Conference on Speech Image Communition and Signal Processing,* pp.127-130. NanPing, China.

Yang, Xingjun & Chi, Huisheng (1995) *Digital Processing of Speech Signal.* Beijing: Electric Industry Press.