Design of Speech Corpus for Mandarin Text to Speech

Jianhua Tao Fangzhou Liu Meng Zhang Huibin Jia

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

{jhtao,fzliu,mzhang,hbjia}@nlpr.ia.ac.cn

Abstract

This paper introduces the CASIA Mandarin corpus designed for Mandarin speech synthesis research. It has been carefully recorded by a professional female speaker under studio conditions. The corpus contains 5000 phonetic context balanced sentences with about 7 hours. The text transcription with word boundaries, POS tags and pronunciation are also involved. The final corpus has been delivered to Blizzard Challenge 2008 as the common corpus for Mandarin speech synthesis evaluation among all participants.

1. Introduction

There has been lots of progress in speech synthesis research during last ten years. Both the voice quality and naturalness of synthesized speech have been greatly improved by the efforts with corpus based unit selection systems and HMM based speech systems [1]. As a basic and important part in speech synthesis research, the design of the speech synthesis corpus has been introduced by several works [2][3][4][5][6]. However, for Mandarin research, it's still hard for us to find a common Mandarin speech synthesis corpus which can be shared with others in public. In this paper, we are trying to summary our method on the construction of the Mandarin speech synthesis corpus.

The paper is organized in two parts. In section 2, the paper introduces the method of prompt selection from a large raw text resource. We have designed several criterions for prompt selection from a huge raw text, speaker selection and data annotation. Professional equipments and staff are employed to ensure the high quality of recorded speech. Section 3 introduces the work for the database recording and annotation. The recording environment is also introduced in this part. We use "Expressiveness Control," "Easy to segment," "Speaking rate control," "Prosody Structure Control," and "Voice beauty" as the basic features to select speakers. The section 4 is the conclusion of the work.

2. Prompt Selection

2.1 Basic units

Chinese is a tonal syllabic language in which the tone of a syllable is described by its pitch contour. There are five different tones (including a neutral tone) and each tonal syllable is composed of one initial and one final which are denoted by the Pinyin system.

There are totally 21 initials, which are "b", "z", "p", "c", "m", "s", "f", "zh", "d", "ch", "t", "sh", "n", "r", "l", "j", "g", "q", "k", "x", "h".

The finals include 8 single vowels (shown in Table 1) and 29 compound vowels (shown in Table 2).

Table 1. The single vowels and samples

Pinyin	Samples
a	da, fa, sa, ...
o	fo, …
e	de, ye, he, ...
i	Yi, di, ti, …
u	du, fu, hu, ...
ü (v)	lü, nü, ...
-i	si, shi, zhi, chi, ...
er	er,

Table 2. The compound vowels and samples

Pinyin	Samples	Pinyin	Samples
ai	ai, sai, shai,…	iong	yong, xiong, qiong, …
ei	fei, hei, …	ua	wa, kua, hua, …
ao	ao, hao, kao,…	uo	wo, huo, guo, luo, duo,…
ou	ou, hou, fou,…	uai	wai, kuai, …
an	an, fan, kan, san,…	ui	wei, gui,…
en	en,	uan	wan, duan, suan,…
ang	ang, pang, bang,…	uen	wen, fen, cen, men,…
eng	beng, peng, feng,…	uang	wang, huang, chuang,..,
ia	Ya, qia, xia, …	ong	hong, gong, dong,…
ie	Ye, qie, jie, pie,…	ueng	weng, feng, seng,…
iao	Yao, qiao, xiao, …	üe	yue, que, xue,…
iou	You	üan	yan, xuan, quan, …
ian	Yan, qian, xian,…	ün	yun,
in	Yin, jin, qin, xin,…	ing	Ying, xing, qing, …
iang	Yang, xiang, qiang,…		

The coverage of all syllables is the basic requirement for the construction of Mandarin corpus. There are about 417 syllables without tone, or about 1500 syllable with tone.

2.2 Text Analysis

Written Chinese characters, known as Hanzi, generally have the same meaning in all dialects. Normally, each character corresponds to one syllable (except retroflex syllables). There

are no spaces or other word boundary markers between the words in Chinese text. This problem places special requirements for the text processing module.

The text analysis module is necessary for prompt selection, while we want the selected prompts covering all basic units and most phonetic context information for TTS system. For the Mandarin TTS system, the text analysis part includes the digital processing, symbol processing, word segmentation, POS tagging, prosodic boundaries prediction, homograph disambiguity of phonetic transcription, etc..

Among all functions, the prosodic boundaries prediction is the most important part for the corpus design. In our Mandarin corpus, the prosodic breaks are classified into four types,

B3: major prosodic phrase boundary with strong intonational marking with/without lengthening or change in speech tempo.

B2: minor prosodic phrase boundary with rather weak intonational marking.

B1: prosodic word boundary. The minimum prosodic unit

B0: phone foot boundary (default as syllable boundaries, not marked explicitly).

In our text analysis part, Maximum Entropy (ME) methods are used for word segmentation, POS tagging and phrase boundaries detection [8].

2.3 Automatic prompt selection

2.3.1 Context information

The prosody of speech is greatly impacted by the context information. For instance, in unit selection system, the selection of a syllable which is in the start of prosodic phrase will cause an unnatural perception if the TTS engine needs a syllable which is in the end of prosodic phrase.

Table 3 shows the factors which have been in our focus.

Table 3. The factors used for corpus design

Factor Index	Description
1	Identity of the current syllable
2	Identity of the current tone
3	Identity of the previous syllable
4	Identity of the previous tone
5	Identity of the next syllable
6	Identity of the next tone
7	Number of preceding syllables in the word
8	Number of following syllables in the word
9	Number of preceding syllables in the phrase
10	Number of following syllables in the phrase
11	Number of preceding syllables in the utterance
12	Number of following syllables in the utterance

Factor 1 has 1472 different syllables in our corpus. Factors 2, 4, and 6 each have 5 values that correspond to the 4 full tones and the neutral tone (5). Each of factors 3 and 5 groups sounds into 10 categories, with the definition of initial and final types.

Factors 7 through 10 have three values each, 0, 1 and 2, where 0 means that the segment in question lies at the boundary, 1 means that it is one syllable away, and 2 means that it is 2 or more syllables away from the boundary. Factors 11 and 12 have two values each, 0 and 1, where 0 means that the segment lies at the boundary and 1 means that it is 1 syllable or more away from the boundary.

We don't use the stress information, because the stress in Chinese is not as clearly defined acoustically or perceptually as in a stress language such as English. During the text selection phase, phrasing is coded on the basis of text analysis results. Although this might give some errors, it can be ignored if we collect data large enough. The syllable number of the sentences is limited from 5 to 30 in order to make speaker easier to control his/her speaking style.

2.3.2 Searching and Selection

The searching method is used for the corpus design.

- Preprocess the text. Reject the sentences with some non-Chinese character sequences or number of syllable more than limitation (we constrained the length of candidate sentences to be less than 30).
- Selection from the candidate sentences with the phonetic and prosodic criterions.
- Manual design of special sentences for including the uncovered syllables and phonemes.
- Second selection for covering all the syllables and initial-finial groups.
- Detailed manual correction of the text corpus.

The search algorithm selects the sentence which matches at least one of the criterions above, removes the sentence from the large candidate set and updates target cover context information.

Among all factors in 2.3.1, Factor 1 (identity of all syllables) is forced to be fully covered. For others, they should be covered as much as possible. The searching steps will not be stopped until the above criterion is reached.

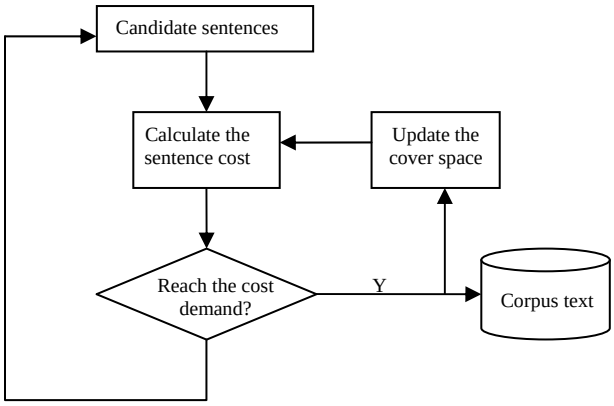

Fig. 1 The diagram of greedy search algorithm

Based on above criterions, greedy algorithm is adopted. The raw text is coming from the 10 years' newspaper of "people's daily". Finally, 10000 sentences which contain 178,614 syllables are selected. That corpus covers all Mandarin syllables,

and most phonetic and prosodic context.

3. Database Recording and Labeling

3.1 Speaker Comparison

Before we recorded the corpus, we asked 8 professional Mandarin speakers (4 males and 4 females) with the age from 25 to 45 years old for testing. All of them come from "Communication University of China". We record their speech with the same prompts of 200 sentences. Then, their speech was randomly presented to 20 subjects who were 22 and 45 years old native speakers with no hearing problem. For each sentence, subjects are asked to judge the following question by using a 5-point scale.

Expressiveness Control:
Although most of current TTS systems can only process the neutral speech, some systems can use a little bit more expressive data. In this factor, we are trying to see if the speaker can easily control his/her voice with a defined expressiveness as we wished. Normally, a well trained speaker can make a good controlling for this.

Easy to segment:
This factor is trying to find if syllables in recorded speech can be clearly and easily perceived. Some voices seem to be very good, but most of the syllables are hardly segmented from the speech to be used as the basic units for TTS systems.

Speaking rate control:
This factor is trying to find if the speaker can easily control his/her speaking rate. This is very important for the prosodic stability of the corpus while we record the speech in a long period.

Prosody Structure Control:
A good speaker can make a good balance in prosody structure and syntactic structure. The good prosody structure control can release lots of pressure for corpus labeling.

Voice beauty:
The beauty of the voice is another important for the success of the corpus design. It makes the listener feel pleasant to accept the voice.

With the comparison from all participants, we finally select a female speaker with the age of 30 years old.

3.2 Recording

The selected text have been recorded in a professional recording studio which equipped with several device such as, Fireface 800 (sound card), RODE K2 (large membrane microphone), etc.. The SNA is more than 60dB. During the recording, a staff member will assist the recording procedure to maintain the consistence of the speech corpus. The criterions used for speaker selection are also used to control the recording in this part.

Till now, 5000 sentences which include 84,839 syllables and are about 7 hours have been finished for recording and labeling. All rest data will be finished in the near future. The Laryngograph will be used in paralell with the speech recording in the later work.

Fig. 2 The recording studio

3.3 Annotation and manual checking

Before the data recording, all of the sentences have been automatically labeled with word boundaries, POS tagging, and phonetic transcription by the text analysis module of Wiston TTS system which is developed by us. While the recording has been finished, the manual checking of the annotation is based on the following rules:

- All speech recorded is transcribed with the correct normalized text, POS tags, word boundaries.
- For baseline voices the speech recorded is completely phonetically transcribed with manual check. Some problems in polyphone, digital reading, tone sandhi, retroflex syllables, name reading and address are paid special attention for correction.

All of the above annotation has been involved into the current released corpus which was supplied to Blizzard Challenge 2008 as the basic corpus of Mandarin speech synthesis evaluation for all participants.

However, some more annotation has been automatically labeled without manual checking. They include,

- Automatically segment the speech in phone and syllable levels.
- Mark the recorded speech with prosody word, prosody phrase boundaries by the method described in [7].Both linguistic features and acoustic features are used to label the prosodic structure.
- Get the pitch marks for all recorded speech with a pitch tracking method. In the later data recording, we will use Laryngograph signal to improve the pitch tracking results.

We will try to finish all manual checking for these three parts and make a new release of the corpus together with the rest data recording.

4. Conclusion

This paper introduces the design of Mandarin speech synthesis corpus. 12 criterions which include Mandarin speech features are used to for prompt selection from raw text. 5 factors have been used for the speaker comparison and selection. The speech

data was recorded in a well equipped studio. The current released corpus has been successfully used for Mandarin speech synthesis evaluation in Blizzard Challenge 2008. The final corpus will be shared via ChineseLDC (http:///www.chineseldc.org).

5. Acknowledgments

This work was partially supported by Hi-Tech Research and Development Program of China (Grant No.: 2006AA01Z138) and National Natural Science Foundation of China (Grand No.: 60575032).

6. Reference

[1]. Alan W Black, Keiichi Tokuda, The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets, Interspeech 2005, Lisbon

[2]. John Kominek and Alan W Black, CMU ARCTIC databases for speech synthesis, http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

[3]. Jindrich Matousek, Josef Psutka, Jirı Kruta, Design of Speech Corpus for Text-to-Speech Synthesis, Eurospeech 2001, Alborg, 2001

[4]. Tania Ellbogen, Florian Schiel, Alexander Steffen, The BITS Speech Synthesis Corpus for German, 4th Conference on Language Resources and Evaluation (LREC) (pp. 2091–2094). Lisbon, Portugal.

[5]. Anja Elsner, Maria Wolters, Thomas Portele, Monika Rauth, Gerit Sonntag, Designing and Labelling A Prosodic Database For American English, In Proc. First Conf. on Language Resources, 1998

[6]. Jindrich Matousek and Jan Romportl, Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis, 10th International Conference on Text, Speech and Dialogue, TSD 2007, Pilsen, September, 2007

[7]. Jianhua Tao, Acoustic and Linguistic information Based Chinese Prosodic Boundary Labelling, International Conference on Tonal Language Aspects, 2004, Beijing

[8]. Jianhua Tao, Shen Zhao (2003). Syntactic Structure to Prosodic Structure Mapping with Inductive Learning Method, ICPhS2003.

[9]. Shen Zhao, Jianhua Tao, Danling Jiang (2003). Chinese prosodic phrasing with extended features, ICASSP2003

[10]. Li Aijun, Lin Maocan (2000). Speech corpus of Chinese discourse and the phonetic research. ICSLP2000

[11]. Li Aijun (1999). A national database design for speech synthesis and prosodic labelling of standard Chinese, Proc. of oriental COCDA'99, TaiPei, TaiWan, 1999.

[12]. Wang Pei, Yang Yufang (2001). Prosodic Structure and Syntactic Structure, Proc. of the third international Conference on Cognitive Science, PP491-496 2001