

A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese

Minghui Dong¹, Kim-Teng Lua², Haizhou Li¹

¹Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

²Chinese and Oriental Languages Information Processing Society, Singapore
mhdong@i2r.a-star.edu.sg, luakt@colips.org, hli@i2r.a-star.edu.sg

Abstract

The paper presents a unit selection-based speech synthesis approach for mandarin Chinese. Unit selection-based approach generates speech by selecting proper units from a speech corpus and connecting them together. In this approach, a set of features are defined to describe the speech units in the corpus and the expected units in the synthesized utterance. Based on the features, cost function is defined to select a sequence of units that are able to generate high quality speech. The cost function describes the two aspects of the generated speech, ie. (1) the appropriateness level of each unit itself. (2) the smoothness level between two units to be concatenated. Viterbi search algorithm is used in this approach to find the best unit sequence that minimizes the two costs. The authors use a new prosody description to ensure the prosody quality of the generated speech. Experiment shows that this approach can generate very high quality speech.

Keywords

Text-to-speech; speech synthesis; unit selection; prosody parameter.

1 Introduction

Text-to-Speech system is a system that converts free text into speech. This is a process that a computer reads out the text for people. There is a wide range of application for text-to-speech system.

A typical text-to-speech system consists of three main parts, which are text analysis, prosody generation and speech synthesis. The text analysis part understands the text and determines the sound of each word. The prosody generation part generates some parameters that control the variability of the speech. The speech synthesis part generates the speech utterance based on the pronunciation and prosody requirements.

In the past decades, many approaches have been used to synthesize speech for Chinese (Lee et al., 1989,1993; Chan et al. 1992; Chu et al. 1995, Chen et al., 1998; Chou and Tseng, 1998, Shih and Sproat 1996) and other languages (Klatt 1987; Bigorgne et al., 1993; Kawai et al., 1995). The main approaches can be classified into two main categories, i.e. rule-based formant synthesis and concatenation synthesis (Dutoit 1997). Formant synthesis generates speech using a set of rules. The rules are usually accumulated during a long process of experiments. This approach needs small computer memory. But the speech

quality is not very good. Concatenation synthesis, however, uses some pre-recorded speech units as templates. During synthesis, the speech units are usually modified using signal processing techniques, such as PSOLA (Moulines et al. 1990, Bigorgne et al. 1993), and then concatenated together to form an utterance. This approach usually needs a larger memory. But the speech quality is relatively better. However, as the development of technology, people are not satisfied with the machine like speech utterances that are generated by using signal processing approach.

Normal concatenation synthesis works by keeping a small set of units in system. During synthesis, a unit is selected and then modified using signal processing techniques according to prosody features. Synthesis by this way can generate speech with relatively high quality. However, the synthetic speech is more or less distorted due to the signal processing process.

A simple idea of generating good speech is to store large quantities of speech segments of human speech in a database and, during synthesis, concatenate all the needed speech segments together without any modification to the units themselves. Of course the longer the stored segments selected for the concatenation, the more natural the generated speech. As each speech unit may have many variants in different contexts or prosodic situations, this approach needs a large memory to store a large number of speech segments. The approach was not practical some years ago because of the limitation of computer speed and memory storage. With the development of hardware, the use of large speech corpus as synthetic units for direct concatenation is possible.

Unit selection-based speech synthesis (or corpus-based synthesis) has been applied in English and other languages for some years. Some attempts (Liu, and Wang, 1998; Chu et al. 2001; Wang et al., 2000, Li et al, 2001) have been made for Chinese TTS using unit selection approach. A representative of the existing unit-selection based system is (Chu et al, 2001). In this approach, there was not prosody models defined. Wu et al. (2001) also proposed a scheme to select phonetically, linguistically best units and then apply prosodic modifications. However, the proposed approaches have limitations in the application of proper prosody. Without proper prosody consideration, the quality of the generated speech may be poor sometimes. This paper proposes a prosody model for unit selection based synthesis for Chinese Mandarin speech.

2 Unit Selection Model

A unit selection model has a well-organized unit database. The database contains the speech units from a large corpus, which is carefully designed to have a large coverage of all phonetic and prosodic variants of each unit. In the database, each speech unit has a number of possible variants, which are suitable to appear in different phonetic and prosodic contexts. The large speech corpus is analyzed offline and all the calculated features are stored in a unit database. In the database, each instance of a unit is described by a vector of features. Each feature may be a discrete or continuous value. The features describe the unit itself and the context of the unit. The features of the unit itself are used for selecting the correct unit that meets the segmental requirement, while the features of context are used for selecting the contextually best unit, which help minimize the discontinuity between the selected units.

The corpus-based concatenation synthesis is actually a pattern matching process. During the synthesis, the work need to do is to select the best units that phonetically and prosodically best match the target units. Meanwhile, the discontinuity between selected units should be kept as small as possible. To meet these requirements, two costs should be

defined in synthesis. One is unit cost, which describes how close a selected unit to the desired unit. The other is connection cost, which describes the degree of continuity between the selected units. The whole cost is a weighted sum of the two costs.

3 Unit Selection Process

The speech synthesis part accepts information from prosody generation part, retrieves the speech unit database to find a proper template for every target speech unit. During the selection process, the phonetic and prosodic constraints are applied. The smoothness of the concatenation is also concerned.

The unit selection process can be illustrated as Figure 1. In the figure, the target sentence is “今天很热 (it is very hot today)”, which consists of 4 syllables (jin1, tian1, hen3, re4). Each syllable has a set of candidate units. The thick line and thick edge box indicate the selected unit sequence. In unit selection process, to get the best speech, we have to consider (1) the appropriateness of the candidate unit compared with target unit, (2) the smoothness between the selected units to be connected. Therefore, the selection process is to find a best path among all the possible paths in the connection lattice. The search process is guided by a cost function, which describes the degree of appropriateness of a unit and degree of smoothness between two units.

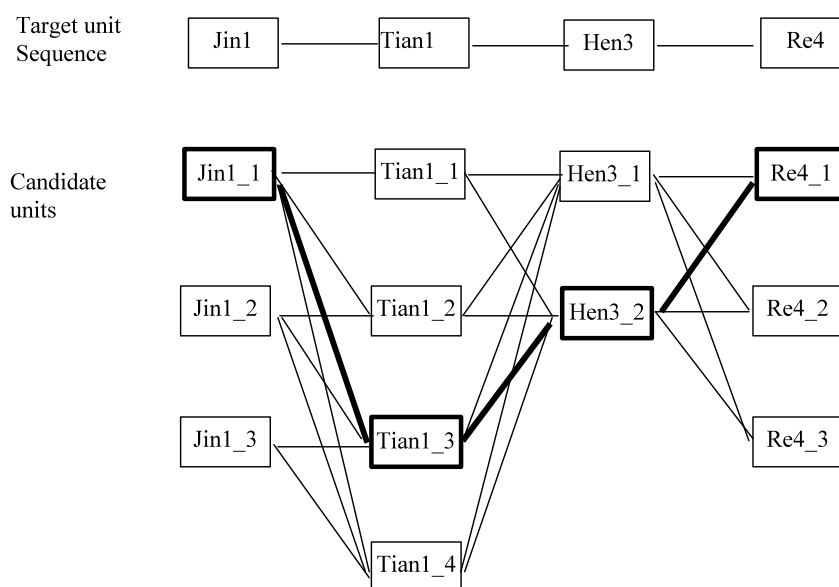


Figure 1. Illustration of unit selection

4 Corpus

As we have mentioned, a large speech corpus is used in unit selection based synthesis. The speech corpus consists of a large collection of utterances. The unit for the synthesis will be extracted from the corpus. It is ideal to cover context dependent units and prosody variants as much as possible. However, building a very large speech corpus is not easy. As the cost

of constructing a large corpus with high quality is very expensive, balance is usually made between coverage of unit variants and corpus size.

In this research, we built a corpus of around 38000 syllables. The script of this speech corpus is selected from a large text corpus (around 300M Chinese characters). The corpus is designed to cover the frequently used syllable and context as much as possible. We use PKU People's Daily text corpus (Yu et al, 2002) as a reference of real word text to evaluate the script of the corpus. This corpus consists of the text of six months articles from People's Daily. There are about 11.4 million Chinese characters in this corpus. We use our text-to-pronunciation programs to convert the text into pinyin representation. Totally, there are 1,373 distinct syllables in this corpus. We calculated that the built speech corpus covers 99.8% of syllable occurrences in the PKU corpus. When context of unit is grouped by Initial and Final class (we defined 11 Initial classes and 10 Final classes), the speech corpus covers 76.8% of the classes of unit occurrences in PKU text corpus. With such coverage, we think the corpus is proper for unit selection based synthesis.

5 Unit Specifications

In this work, we choose syllable as the synthesis unit. The reason to choose syllable is that syllable is a relatively stable unit. The coarticulation between syllables is relatively loose, while the coarticulation between sub-syllable units is very tight.

Each unit is specified by a feature vector, which is used for matching in a unit selection process. Both the target units and units in inventory are described using this feature vector. The features describe the phonetic identity, phonetic context, break types around the unit, and prosody parameters of each unit. The detailed features are as the following:

- **Phonetic identity of the unit:** Using the pronunciation of the unit is to ensure that the candidate unit will have the same sound as the expected one. The pronunciation includes the initial, final and tone. There are 22 initials, 38 finals, and 5 tones defined in this work.
- **Phonetic context:** The coarticulation between two units is determined by the phonetic identity of its neighbours. The context of the unit will help find the unit with similar context of a unit. The phonetic context consists of the initials, finals, and tones of the previous and next units.
- **Breaks around the unit:** The break types before and after the unit. The prosodic properties of a unit before a break and after a break are quite different. The break type information is an important index to evaluate the similarity of two units. The types we defined include syllable break, word break, and prosodic word break.
- **Prosody parameters:** The prosody parameters are a collection of parameters that describe the duration, pitch contour and energy of a unit. We defined 10 parameters to describe the duration, pitch contour and energy of each unit (Dong et al 2005).

6 Costs for Unit Selection

Cost function describes to what degree that the selected units deviate from perfect ones. The cost function mainly consists of unit cost and connection cost. Unit cost mainly concerns quality of the unit, while connection concerns the coarticulation effect between the two selected units.

6.1 Unit Cost (C_{Unit})

Unit cost expresses the distance between the unit to select and the unit that we expect. In the selection of units, we first look for the units of the same syllable identity (Initial, Final and tone) as the expected units. Ideally, we expect to find the syllable that has same context as that in the target speech. Unit cost is calculated by comparing the corresponding features of a unit or a sequence of units, as illustrated in Figure 2. In the figure, T_i is the target unit, U_i is the unit

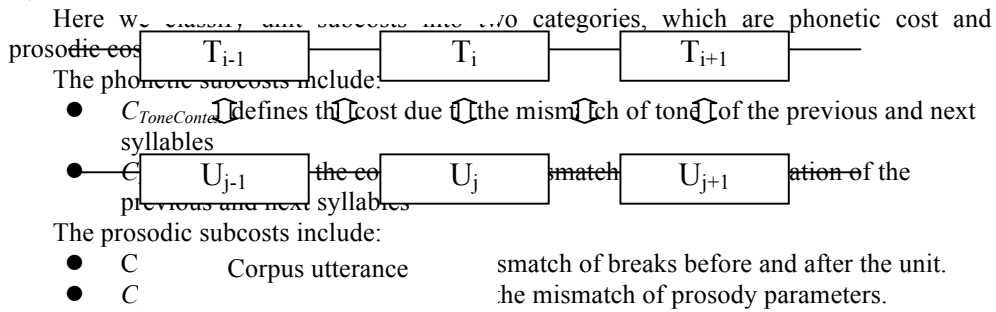


Figure 2. Illustration of unit cost calculation

6.2 Connection Cost (C_{Conn})

When two selected best units from separate places are connected together, they do not necessarily match each other. Two successive units with sub-optimal unit cost may be preferable over two non-adjacent units with optimal unit cost.

The connection cost consists of two measures: coarticulatory continuity measure and prosodic continuity measure (Yi 1997). The First is inspired by the fact that certain phones spoken in succession exhibit a significant amount of coarticulation. Phone pairs with more perfect continuity in formants are more preferable to be concatenated. Prosodic continuity compares the prosodic information of two connected syllables.

When two syllables are to be connected, if they were not spoken in succession, a connection cost must occur. The connection cost measures how much degrading in the connection is caused when two speech units come from non-contiguous syllable constituents. The connection cost can be calculated in two ways:

- Directly calculated by calculating the spectrum continuity or prosody continuity between two units to be connected (as in Figure 3, in which units U_i and V_j are to be

connected). This usually involves calculation of mismatch of acoustic or prosodic parameters.

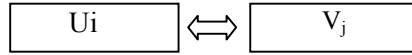


Figure 3. Direct calculation of connection cost

- Indirectly calculated by comparing the connected unit with its original neighbor in speech (as in Figure 4, in which units U_i and V_j are to be connected). This can be done by considering phonetic information. This work uses this way to calculate connection cost.

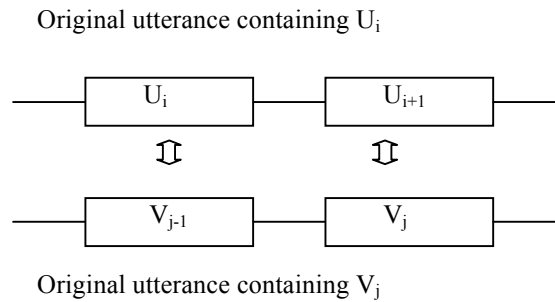


Figure 4. Indirect calculation of connection cost

The subcosts that we defined include:

- C_{Succ} defines the cost due to the units to be concatenated are not originally in succession in the same utterance.
- $C_{ToneConn}$ defines the cost due to the mismatch of the tones of the previous and next syllables.
- $C_{EdgeConn}$ defines the cost due to the mismatch of the pronunciations of the previous and next syllables

Because some of the connections are more important (because of tight coarticulation or prosodic coherence) than the others are, failing in satisfying the connection continuity may lead obvious degrading in speech quality. Therefore, we defined an importance factor for connection I_{Conn} . 3 levels for connection coherence degree between units are defined in this work.

6.3 Prosody Consideration

The prosody is implemented in unit selection by selecting units with proper prosody properties. This is done by using prosody related subcosts in cost function. The selected units will be concatenated together to form a speech utterance. The speech of connected

units itself exhibits prosody. Break is implemented by selecting proper boundary units. No silence is inserted into speech to create a prosodic break in utterance. Tone is implemented by selecting units with proper pitch contour.

We define a subcost for prosody to account for all the prosody parameters. In this research, 10 parameters are used to describe the prosody of a unit (Dong et al 2005). We use CART approach for prosody prediction. In the prediction of prosody parameters, we obtain not only the values of predicted prosody parameters but also a value of standard deviation for this prediction. This is done by doing a statistics for the samples falling into the corresponding leaf nodes of the regression tree. The prosodic value gives the expected parameters, while the standard deviation reflects the accuracy of the value. The two values together give an accurate prediction of prosody parameters.

Suppose the predicted prosody parameters are represented using vector T .

$$T = (t_1, t_2, \dots, t_{10})$$

The corresponding standard deviations are presented using vector D .

$$D = (d_1, d_2, \dots, d_{10})$$

The prosody parameters of a unit from inventory are represented using vector S .

$$S = (s_1, s_2, \dots, s_8)$$

The cost is calculated using

$$c = \sum_{i=1}^{10} (w_i |t_i - s_i| / d_i)$$

$$C_{\text{ProsodyParam}} = \begin{cases} 5c, & \text{if } c < 20 \\ 100, & \text{if } c \geq 20 \end{cases}$$

where w_i is the weight for each parameter.

6.4 Total Cost

Total unit cost is calculated as:

$$C_{\text{Phonetic}} = W_{\text{ToneContext}} C_{\text{ToneContext}} + W_{\text{PronContext}} C_{\text{PronContext}}$$

$$C_{\text{Prosodic}} = W_{\text{Break}} C_{\text{Break}} + W_{\text{ProsodyParam}} C_{\text{ProsodyParam}}$$

$$C_{\text{Unit}} = C_{\text{Phonetic}} + C_{\text{Prosodic}}$$

where $W_{\text{ToneContext}}$, $W_{\text{PronContext}}$, W_{Break} , and $W_{\text{ProsodyParam}}$ are weights for the corresponding subcosts respectively.

Total connection cost is calculated as:

$$C_{\text{Smooth}} = W_{\text{SuccUnit}} C_{\text{SuccUnit}} + W_{\text{CToneConn}} C_{\text{CToneConn}}$$

$$+ W_{\text{CEdgeConn}} C_{\text{CEdgeConn}}$$

$$C_{\text{Connection}} = C_{\text{Smooth}} + C_{\text{Conn}}$$

where W_{SuccUnit} , W_{ToneConn} and W_{EdgeConn} are weights for the corresponding subcosts respectively.

Suppose a sequence of n units is selected for a target sequence of n units. The total cost is calculated with the following function.

$$C_{Total} = \sum_{i=1}^n C_{Unit}(i) + \sum_{i=0}^n C_{Connection}(i)$$

where the C_{Total} is total cost for the selected unit sequence, $C_{Unit}(i)$ is the unit cost of unit i , $C_{Connection}(i)$ is the connection cost between unit i and unit $i+1$. Unit 0 and $n+1$ are defined as start and end symbol to indicate start and end of the utterance.

6.5 Weight Determination

As our definition, the total cost of a sequence of units is a weighted sum of the unit cost and connection cost. The unit cost and connection cost are both weighted sum of sub-costs. Determining the weights is important for the general performance of the whole system. Unfortunately, it is hard to find an objective way to compare the quality of speech utterances generated by using different weight settings. Therefore, we need to have some alternatives to determine the weights. In this research, the weights are mainly determined experimentally based on knowledge and informal perception test.

7 Search Algorithm for Unit Selection

During unit selection process, for each unit of the target speech, there are multiple candidate speech units. The candidate units of all target units form a lattice. To find the path that has the lowest cost, a dynamic programming approach is needed. In this research, Viterbi algorithm is used to find the best path. The Viterbi search progress works in the following steps:

1. Initialize $C(0,1) = 0$;
2. For $i = 1$ to $N_{SeqUnit}$ do
 - a. For $j = 1$ to N_{Cand}

Calculate unit cost $C_{Unit}(j)$
 - b. Sort units in ascending order of $C_{Unit}(j)$, and keep the best M ones.
 - c. For $j = 1$ to N_{Path} do
 - For $k = 1$ to M do

$$C(i, jM+k) = C(i-1, j) + W_{Unit} C_{Unit}(V_k) + W_{Con} C_{Con}(U_{i-1,j}, V_k)$$
 - d. Sort the paths in ascending order of $C(i, 1:jM+k)$, keep the best N ones.
3. Back trace to find the best sequence that has a minimal cost value.
4. Output the sequence of units.

where the meanings of the notations are as following:

- $N_{SeqUnit}$: number of units in the sequence;
- N_{Cand} : number of candidate units in current step;
- N_{Path} : number of paths in previous step;
- M : number of candidate units for further calculation in current step;
- N : number of paths to keep in this step;
- $C(i,j)$: accumulative cost of the j th path in the i th step;
- V_k : the k th candidate in current step;
- U_{ij} : the j th selected unit in the i th step;
- $C_{Unit}(V)$: the unit cost of unit V ;
- $C_{Con}(U, V)$: the connection cost between U and V ;
- W_{Unit} : weight for unit cost;
- W_{con} : weight for connection cost.

8 Experiment

Testing text of the listening test is selected from PKU People's Daily corpus (Yu et al, 2002). First, we select the sentences with 8 to 12 characters as our candidate sentence set. Then, a greedy algorithm is used to select 1000 sentences that have a largest coverage of the unit variants. Finally, we select 100 sentences randomly from the first 1000 sentences as our testing sentence set. The 100 sentences consist of 1091 characters. The testing sentences will be used in some of the following experiments.

In this experiment, we evaluate the performance of the proposed prosody application. We compare the quality of speech utterances that are generated using two cost function scheme: (1) simple prosody consideration (without parameters only break considered) (2) the proposed prosody consideration (prosody parameters used in cost function).

Naturalness test is done using MOS testing in this work. The MOS approach in the work is to ask the listener to score each sentence based on a 5-level scale. A best speech is marked as 5 and worst is marked as 1. 30 native speakers participated in the listening test. The listeners are asked to compare utterances generated by 2 approaches and score them.

The MOS testing result is shown in Table 1. In the table, we see that when using this prosody scheme, we achieve a MOS score of 4.21, which is better than using a simple prosody consideration (3.12). The result shows using this prosody description improves the quality of speech.

Cost scheme	MOS	Standard deviation
Simple Prosody	3.12	0.38
Prosody parameters	4.21	0.23

Table 1. Result for naturalness test

9 Conclusion

The paper proposed a unit selection based approach for mandarin Chinese speech synthesis. We applied a set of prosody parameters in the cost function to guide the unit selection process. Experiment shows that this approach generates very good speech.

10 References

- Bigorgne, D., Boe.ard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J.L., Metayer, I., Sorin, C., and White, S., 1993. Multilingual PSOLA text-to-speech system. In: Proc. ICASSP, pp. II.187-190.
- Chan, N. C. and Chan, C. Prosodic Rules for Connected Mandarin Synthesis. *J. Inform. Sci. Eng.* 8, 261-281. 1992.
- Chen, S. H.; Hwang, S. H. and Wang, Y. R., An RNN-Based Prosodic Information Synthesizer For Mandarin Text-To-Speech. *IEEE Trans. Speech Audio Processing.* 6(3), 226-239. 1998.
- Chou, F. C. and Tseng, CV. Y. Corpus-Based Mandarin Text-To-Speech Synthesis with Contextual Syllabic Units Based on Phonetic Properties. In: Proc. ICASSP, pp. 893-896, 1998.

- Chu, Min and Lv, Shinan. High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules. In Proceedings of the XIII International Congress of Phonetic Sciences, (Stockholm), pp. 334--337, 1995.
- Chu, Min; Peng, Hu; Yang, Hongyun and Chang, Eric. Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer. ICASSP2001, Salt Lake City, May 7-11, 2001.
- Dong, Minghui; Lua, Kim-Teng; Xu, Jun. Selecting Prosody Parameters for Unit Selection Based Chinese TTS. Natural Language Processing - IJCNLP 2004: 272-279, LNCS, Springer, 2005.
- Dutoit, Thierry. An Introduction to Text-To-Speech Synthesis, volume 3 of Text, Speech and Language Technology. Kluwer Academic Publishers, P.O. Box 322, 3300 AH Dordrecht, The Netherlands, 1997.
- Kawai, H., Higuchi, H., Simuzi, T. and Yamamoto, S., 1995. Development of a text-to-speech for Japanese based on waveform splicing. In: Proc. ICASSP, pp. 1569-1572.
- Klatt, D. H. Review of Text to Speech Conversion for English, Journal of the Acoustical Society of America, vol.82, no.3, pp.737-793, 1987.
- Lee, Lin-Shan; Tseng, Chiu-Yu and Hsieh, Ching-Jiang. Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System. IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 3, pp. 287--294, 1993.
- Lee, Lin-Shan; Tseng, Chiu-Yu and Ouh-young, Ming. The Synthesis Rules in a Chinese Text-to-Speech System. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 9, pp. 1309--1320, 1989.
- Li, Wei; Lin, Zhenhua; Hu, Yu; Wang, Renhua. A Statistical Method for Computing Candidate Unit Cost in Corpus Based Chinese Speech Synthesis System. In proceeding of International Conference on Chinese Computing, Singapore, 2001.
- Liu, Qingfeng; Wang, Ren-hua; Ma, Zhongke and Yin, Bo. Design and Realization of a Chinese Speech Platform, Tianyin Huwang System. Communications of Chinese and Oriental Languages Information Processing Society 8 (2), pp. 211-220, 1998.
- Moulines, E. and Charpentier, F. Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication 9,453-467, 1990.
- Shih, Chilin and Sproat, Richard. Issues in Text-to-Speech Conversion for Mandarin. Computational Linguistics and Chinese Language Processing, 1(1), 37-86, 1996.
- Wang, Ren-Hua, Ma, Zhongke. Li, Wei, and Zhu, Donglai, A Corpus-Based Chinese Speech Synthesis with Contextual-Dependent Unit Selection. ICSLP, 2000.
- Wu, Chung-Hsien; Chen, Jau-Hung. Automatic generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis, Speech Communication, vol. 35, 219-237, 2001.
- Yi, Jon. Natural Sounding Speech Synthesis Using Variable-length Units, Master's thesis. MIT, 1997.
- Yu, Shiwen, et al. The Specification of Basic Processing of Contemporary Chinese Corpus. Journal of Chinese Information Processing, Issue 5 & 6. 2002.