CHAPTER 5

MANDARIN TEXT-TO-SPEECH SYNTHESIS

Ren-Hua Wang[†], Sin-Horng Chen[‡], Jianhua Tao[§], Min Chu[*]

[†]*USTC iFlytek Speech Lab, University of Science & Technology of China, Anhui,*
[‡]*Speech Processing Lab, National Chiao Tung University, Hsinchu,*
[§]*NLPR, Institute of Automation, Chinese Academy of Science, Beijing,*
[*]*Speech Group, Microsoft Research Asia, Beijing*
E-mail: {rhw@ustc.edu.cn, schen@cc.nctu.edu.tw,
jhtao@nlpr.ia.ac.cn, minchu@microsoft.com}

This chapter introduces Mandarin Text-To-Speech (MTTS) synthesis. Beginning with a brief review on the development history of MTTS and attributes of MTTS, three main constituents of the technology are presented: 1) Text processing: word segmentation, disambiguation of polyphones, and analysis of rhythm structure; 2) prosodic processing: features of Mandarin prosody, and prosody prediction, and; 3) speech synthesis: parametric synthesis and concatenative synthesis. Finally perspectives and applications for MTTS synthesis are discussed in the final sections.

1. Introduction

1.1. *Historical Review*

The development of Mandarin Text-To-Speech (MTTS) systems can be traced back to the seventies of the last century. Since the introduction of the first Mandarin speech synthesizer which used VOTRAX to generate speech from phonetic transcription in 1976,[1] we can roughly divide the development of MTTS technology into three stages.

In the early stage, the main focuses of research were on finding suitable speech synthesis techniques as well as in selecting proper synthesis units. The main concerns were the intelligibility of the synthetic speech and memory constraint. Formant synthesizers[2–5] and linear predictive coding (LPC)[6–8] were the two most popular techniques used in MTTS in the eighties. As for synthesis unit, both initial-final[6] and demisyllabic[7,9] schemes were often used. Systems

developed in this early stage usually adopted simple, rule-based prosody control techniques[3,8,10] with input text in the form of phonetic transcription.[4] Most of these systems could produce highly intelligible speech for isolated words. But they could not generate natural-sounding speech for unlimited input texts because of the use of relatively simple text analysis and prosody generation rules.

The trend of speech synthesis techniques switched to the PSOLA (pitch synchronous overlap and add) methods[11–14, 20–24] in the nineties, while the trend in synthesis unit selection, from the late eighties, favored syllable-based concatenative systems.[8,12–14,18–24] A few other schemes, such as word-based[15] and diphone-based[4,11,16] techniques, were also proposed in the nineties. For prosody control, although more sophisticated rule-based methods[18,19] were reported, the trend changed to the data-driven approach in the nineties. In a data-driven approach, prosody generation rules were implicitly included in a statistical model[16,17,22,24] or an artificial neural network (ANN)[23] with parameters trained from a large, well-annotated speech database. Meanwhile, more sophisticated text analysis methods with electronically coded Chinese text input[11,12,17–20] were proposed in mid-nineties. Hierarchical prosody structures were exploited and applied to prosody generation.[17,19,20,24] Due to the uses of more sophisticated prosody generation schemes and the larger speech inventory of waveform templates, many MTTS systems developed in late nineties could produce good quality, natural-sounding synthetic speech.[17-24] A good review of the progress of MTTS technology and system developments in the early and middle stages was made by Shih and Sproat in 1996.[17]

From the late nineties, in the modern stage of MTTS development, corpus-based MTTS approaches[25–30] became the mainstream. The new approach uses sophisticated unit selection schemes to choose long speech segments with appropriate prosody from a large, single-speaker speech corpus, and concatenates these segments directly with little or no prosody modifications. The main focus of this approach lies in its unit selection technique. Usually, a prosody model or a rule-based method is needed to generate proper prosody targets for guiding the unit selection.[25,26] In this stage an approach without explicit prosody model and prosody modification was also proposed.[27] The synthetic speech of a corpus-based MTTS system has always been reported to be of high quality and very natural.

1.2. *Attributes of Mandarin TTS*

A TTS system typically contains three main components shown in Figure 1: text processing, prosody processing and speech synthesis. First, the text processing

component converts any input text into corresponding phonetic notations with some other information needed for prosody processing. Then, the prosody processing component generates prosodic targets in symbolic format, such as ToBI,[31] or in numerical format, such as fundamental frequency (F0) curve and segment duration, or both. Lastly, the synthesis component outputs speech that matches the phonetic and prosodic specifications. The algorithms used in the prosody and synthesis components are normally language independent. However, the text processing component often contains language specific processes.

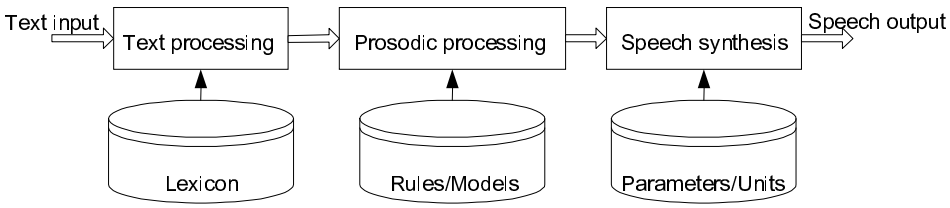

Fig. 1. Block-diagram of a TTS system.

1.2.1. *Characters and Syllables*

The base unit in written Chinese is the character. There are more than 70,000 different Chinese characters, but the 3,500 most frequently used characters can cover as much as 99.5% of actual occurrence.[32] For MTTS systems that do not aim to process ancient documents, handling up to 20,000 characters is enough.

Normally, one Chinese character corresponds to one syllable in speech. The only exception is in the case of a retroflexed syllable. For example, two characters, such as "玩儿" and "点儿", are converted to a single retroflexed syllable, pronounced as /wanr2/ and /dianr3/. The symbols between two slashes are modified Hanyu Pinyin (汉语拼音), which is the phonetic notation for Mandarin Chinese used in China and was adopted in 1979 by the International Organization for Standardization as the standard romanization system for modern Chinese. In Pinyin, 22 symbols are defined for initials (the consonants at the onset positions of Chinese syllables, one for zero-initial), 37 symbols for finals (the vowels plus coda consonants) and 5 symbols for the tones.

Theoretically, these 22 initials and 37 finals will generate 814 base syllables. Yet, in Mandarin, only about 410 syllables are valid. Also, since not all base syllables can go with all five tones, only about 1,338 tonal syllables can be found in an authoritative dictionary,[33] which keeps only the canonical pronunciations of each word. However, in continuous speech, more tonal syllables are used. There

are three main sources for these additional syllables. The first source is tone sandhi. When two falling-rising tone syllables occur adjacently, the first one will be spoken with a rising tone. For example, /che2/ is not a valid syllable in Mandarin dictionaries. However, in the word "扯谎", the canonical pronunciation /che3 huang3/ is to be spoken as /che2 huang3/. The second source is neutralization. Although, only about 40 neutralized syllables are listed in most dictionaries, between 200–300 will be used in actual continuous speech. The third source is retroflex. Today, about 200–300 retroflexed syllables are used by people living in northern China. If all these additional syllables are considered, the total number of tonal syllables in Mandarin is close to 2,000.

1.2.2. *Lexical Word and Prosodic Word*

The word, rather than the character, is used as the basic unit in most Chinese processing systems because the usages and meanings of words are far more consistent in Chinese. In written Chinese, characters run continuously without visual cues to indicate word boundaries. Word segmentation becomes a basic requirement for almost all Chinese processing systems. Another particularly Chinese language problem lies in the identification of proper names (including person names and organization names).

Since most Chinese characters are themselves single-character words, many mono-syllabic words are detected as a result of word segmentation. Yet, in spoken Mandarin, there exists a disyllabic rhythm. A successive string of single-character words are often grouped into disyllabic rhythmic units, and long spoken words are normally chunked into several units as well. This basic unit of rhythm is known as a foot or a prosodic word in the literature.[34] To distinguish words listed in a lexicon from prosodic words in speech, words in a lexicon are referred to as lexical words. A prosodic word may contain one or more lexical words, or may even be part of a lexical word. Since prosodic words are formed dynamically according to context, it is impossible to list all prosodic words in a lexicon, as has been done for lexical words. Therefore, in a Mandarin TTS system, segmenting text into lexical words is not enough. Prosodic word segmentation is also needed.[35,36]

1.2.3. *Homophone and Polyphone*

In Chinese, there are many homophones. On average, about 15 characters share a syllable. This feature causes some difficulties in speech recognition, yet, it is not a problem in TTS. There are about 800 polyphones in Mandarin. In many cases,

the pronunciation of a polyphonic character is fixed when it is used in a multi-character word. However, there are still some single-character words or multi-character words that have more than one pronunciation. Therefore, choosing the right pronunciation for a polyphone is a problem in TTS. If the most frequently used pronunciation of each polyphonic word is used, the error rate of character-to-syllable conversion is around 0.9%.[37] If a context model is used to disambiguate these possibilities, the error rate can be reduced to about 0.4%.[37–39]

Besides polyphonic words, there is no out-of-vocabulary problem in character-to-syllable conversion in Mandarin. Any new word can be converted simply by looking-up a character-to-syllable dictionary.

2. Text Processing in MTTS

The task of text processing is to process the input text and generate corresponding information for the later components, such as prosody modeling and speech synthesis. Text processing takes a very important role in TTS, because it not only determines the correctness of synthesized speech but also significantly influences the intelligibility and naturalness of the speech output.

To generate the Pinyin, rhythm, stress and other information corresponding to the input text, text processing in MTTS should include several processing steps to deal with the different problems in Chinese.

2.1. *Word Segmentation*

Automatic word segmentation is a fundamental and indispensable step in most Chinese processing tasks, including MTTS, due to the sole reason that there are no explicit word delimiters in Chinese text. By employing word segmentation, MTTS system can segment sentences (thereby paragraphs and articles) into words sequences, and so corresponding Pinyin, rhythm and part-of-speech information can be generated by looking up each word in a dictionary. The quality of word segmentation is critical for MTTS, but fortunately, Chinese word segmentation has been researched for tens of years, and significant progress has been achieved in each of the following four difficult problems.

Dictionary compilation is a preliminary but difficult task because there is no standard definition of what a word is in Chinese. In the last ten years, several high quality electronic dictionaries have been developed, with their sizes ranging from 60,000 to 270,000 entries. Research results show that higher segmentation performance can be achieved by incorporating more word entries.[40] To reduce

the huge effort of dictionary compilation, many effective algorithms have been proposed to identify new word candidates from Chinese text corpus.

Given a dictionary, the most serious problem in Chinese word segmentation is the problem of segmentation ambiguity. For example, the sequence "尚未来(still not come)" can be segmented as "尚(still) 未来(future)" and "尚未(still not) 来(come)". Segmentation algorithms based on maximizing the probabilities of word sequences, prove to be able to reduce this problem to some degree. Some researchers analyze the segmentation ambiguity and find that only one of the segmentation paths is reasonable for more than 90% of the cases of ambiguous text. So saving the correct segmentation path for this kind of ambiguous texts is also a practical approach to improve segmentation performance.

Proper names must be identified in MTTS because rhythm and Pinyin information often need special processing in proper names, such as "曾"(/ceng2/) in surnames should be read as /zeng1/. Automatic proper name identification has been studied deeply in relation to many other Chinese processing[41] tasks and can be integrated into MTTS rather easily with some additional work in the area of rhythm and Pinyin information processing.

Another problem in word segmentation is the derivative word problem, such as "着了火"(have caught fire) derived from "着火"(catch fire), and "开开心心"(/kai1 kai1 xin1 xin1/), from "开心"(/kai1 xin1/). Chinese derivative words need to be identified so that correct rhythm and Pinyin information can be generated for these words. Research shows that most of the Chinese derivative words can be identified by adopting some tens of derivation rules, and the relationship between the characters within these words (such as verb-object relationship in "着火") is useful in Chinese derivative word identification.

Based on the segmentation research described above, the accuracy of a state-of-the-art Chinese word segmentation system is about 96%-98%, which satisfies the basic need of most MTTS systems.

2.2. *Part-Of-Speech Tagging*

Part-Of-Speech (POS) is a type of linguistic information that is widely used. The POS of a word also plays an important role in both the disambiguation of polyphone pronunciations and the analysis of rhythm structure. Therefore POS tagging is always carried out after or during word segmentation. Several POS categories have been proposed for Chinese, and the most commonly-adopted categorization is the one proposed by Yu[42] which includes 26 POS categories. The trigram POS tagging algorithm has been proven to be an effective method

and the popular corpus annotated by Yu is often used to train POS trigram models. The tagging accuracy for Chinese is about 95%–97%.

2.3. *The Problem of Polyphones*

The polyphone problem is a particular issue of MTTS which needs to be handled almost exclusively by MTTS. To tackle this problem, listing as many words with polyphonic characters as possible, such as "中(zhong4)奖" and "行(hang2)长(zhang3)", into the dictionary is an approach commonly-employed in most MTTS systems. Detailed Pinyin information processing, after proper names and derivative words have been identified, can also help to solve part of the polyphone problem in MTTS.

Monosyllabic words, which are polyphones such as "长(chang2/zhang3)", "还(hai2/huan2)", and "干(gan1/gan4)", are often assigned with wrong pronunciations typically when the pronunciation selected is based only on the most frequent option after word segmentation. Several solutions have been proposed on this matter in recent years: summarizing pronunciation rules manually by human experts, applying machine learning methods on annotated corpus,[43] and introducing a hybrid method to integrate the strengths of both human and machine.[44]

As a result of the various efforts above, the accuracy of Pinyin generation on the whole is more than 99.8% in most MTTS systems.

2.4. *Rhythm Structure Analysis*

Beyond Pinyin, rhythm structure is another dimension of information which is introduced into MTTS, and this needs to be predicted from the input text too. And for Chinese, a widely applied rhythm structure definition consists of six layers:[35] syllable, prosodic word, minor prosodic phrase, major prosodic phrase breath group and sentence layers.

As discussed above, the lexical word sequence generated by word segmentation should be converted into a prosodic word sequence, and this conversion is often done through a series of human crafted chunking rules, such as the word "的(/de/)" should always be grouped with its preceding word, and so on.

From the minor prosodic phrase layer to the breath group layer, each phrase or group is defined by the length of the phrase, and also the obviousness of the break at the boundaries of the phrase. The prediction of the rhythm structure from a text is an interesting but difficult research task. Rule-based methods were used for phrase break prediction in the early days, but recently, many researchers have

been exploiting statistics-based methods[45,46] with part-of-speech and syllable number as the dominant features for this task and have achieved good performance. Syntactic parsing is assumed to contribute to rhythm structure prediction, but experimental results show that only shallow parsing, such as chunking, is effective.[47]

The accuracy of prosodic word generation is about 97%–99%, and about 82%–85% for predicted rhythm structure, which is regarded as acceptable.

There are also some other issues that should be handled in the text processing stage of MTTS, including text normalization,[48] and stress prediction. All these problems have been researched on but not explored here due to space constraints.

3. Prosody Processing in MTTS

3.1. *Features of Mandarin Prosody*

Prosody is an inherent supra-segmental feature of human speech. It carries stress, intonation patterns and timing structures of continuous speech which, in turn, determine the naturalness and intelligibility of an utterance. Prosody is even more important for Mandarin Chinese because Chinese is a tonal language. As the syllable is the basic pronunciation unit, the prosodic features in Chinese are known to include syllable pitch contours, syllable energy contours, syllable or initial/final durations, and inter-syllable durations. Hierarchical prosody structure can be formed by taking these elements as its basic building blocks.[49,50]

3.2. *Prosody Prediction*

The general approach of prosody prediction includes the following two steps: (1) extract some linguistic features from the input text, and (2) generate prosodic features from those linguistic features. Methods of prosodic feature generation can be classified into two general categories: rule-based and data-driven. The former approach is the more conventional one, involving the use of linguistic expertise to manually infer the phonological rules of prosody generation from a large set of utterances.[18–20] The main disadvantage of this approach lies in the difficulty of collecting enough rules without long-term dedication to the task. The data-driven approach tries to construct a prosody model from a large speech corpus, usually by statistical methods or artificial neural network (ANN) techniques.[21–23,51–53] The primary advantage of this approach is that it can be automatically realized from the training data set without the help of linguistic experts. As a result, the data-driven approach has gained popularity in recent years.

3.2.1. *F0 Prediction*

Although there are only five lexical tones, syllable pitch contour patterns in continuous Mandarin Chinese speech are highly varied and can deviate dramatically from their canonical forms. The factors that have major influences on pitch contours include the effects of neighboring tones, referred to as *sandhi* rules,[54] coarticulation, stress, intonation type, semantics, emotional status, and so on. F0 prediction therefore needs to consider not only the basic tone patterns, but also high-level factors from the hierarchical prosody structure.

(a) The rule-based approach

The method proposed by Wang[18] extends the four canonical tones of H, R, L, and F (i.e., high, rising, low, and falling) to $H^1 - H^3, R^1 - R^3, L^1 - L^3, F^1 - F^5$, and adds two kinds of light tones. Then, tone patterns of multi-syllable words are formed by the combinations of basic units of the proposed *monotonemes* with tone sandhi rules. Sentence intonation is realized by applying global modification rules. A formal evaluation confirmed that the resulting synthetic speech sounds very natural. In an improved method,[19] pitch generation is realized by building up a stable template at the prosodic word level and generating a base intonation contour. A subjective MOS test confirmed that the KD2000 Mandarin TTS system, which employed the F0 prediction method, performed better than KD863 which was ranked No.1 in the 1998 national assessment on the naturalness of synthesized speech.

Chou[20] assigns a pitch contour pattern to each word, and then superimposes the pattern with an intonation pattern of the major prosodic phrase level. At the word-level, the pitch contour patterns of all tone combinations are used and extracted from an isolated-word speech database. In the major prosodic phrase level, four global intonation patterns are applied depending on the type of punctuation mark, including sentence middle, comma, period and question mark. These four intonation pitch contour patterns are extracted from a sentence database. An informal test confirmed that the synthesized speech sounds indeed natural.

(b) The data-driven approach

Wu[22] built a word prosody tree to store both prosodic features and linguistic features of each word in a speech database. The tree contains two levels: word-length level and tone-combination level. In synthesis, it first calculates the sentence intonation to find the target pitch period of the first syllable. It then

traverses the word prosody template tree by using word length and tone combination to extract some appropriate word template candidates. Lastly, cost functions considering the matching of linguistic features between the input word and these word template candidates are calculated and used to determine the word template. Experimental results showed that most synthesized pitch parameter sequences match quite well with their original counterparts.

In Yu's approach,[51] the syllable pitch contour pattern is generated by a linear regression method based on a 4-level hierarchical prosody structure containing the syllable, word, prosodic phrase, and utterance levels. Moreover, the predicted basic syllable pitch contour pattern is further refined by finding a syllable pitch contour pattern of real speech. A subjective test using a 10-scale MOS shows the results of 6.87 and 7.08 for the basic and modified methods, respectively.

In Chen's article,[23] an RNN-based method is proposed. It employs a three-layer recurrent neural network (RNN) to generate some prosodic features including the syllable pitch contour, syllable energy level, initial/final durations, and inter-syllable pause duration. The inputs of the RNN are syllable- and word-level linguistic features extracted from the input text. As one of the merits of RNN, the dynamic variations of syllable pitch contour can be automatically learned using only lower-level linguistic features without explicit information from high-level prosodic structural elements. Experimental results showed that all synthesized pitch contours resemble their original counterparts quite well. Moreover, many phonological rules, including the well-known tone sandhi rule for the 3-3 tone pair,[54] were automatic learned by the RNN.

The quantitative model has also been introduced to generate the pitch contour, such as the modified Fujisaki model.[18,46] A quantitative model allows us to analyze and represent the acoustic features of pitch contour more effectively. The model parameters can be extracted automatically based on a speech database. Experimental results have shown its feasibility in MTTS.

3.2.2. *Duration Prediction*

For MTTS, the task of duration prediction is to determine the duration of the whole syllable or the duration of its initial/final. A general approach to tackle this problem is to first identify important and relevant linguistic features, and then exploit rules or computational models to describe their relationships with syllable duration.

A two-level, rule-based syllable duration prediction method is proposed by Wang.[18] In both the lower word level and upper sentence level, a set of rules are derived.

Chou's method[20] generates syllable duration using a multiplicative model with the following four affecting factors considered: average syllable duration, tone, position, and break index.

An alternative RNN-based method for generating initial and final durations is also proposed[21] as discussed previously.

In Sun's method,[53] a duration prediction method based on a polynomial regression model is proposed. The method consists of three steps: linguistic features selection, polynomial model determination, and duration generation by nonlinear regression. The Eta-squared statistical concept is used to determine the most forceful linguistic features.

3.2.3. *Pause Duration Prediction*

A general rule-based approach of inter-syllable pause duration prediction is to first determine the break indices from high-level hierarchical prosody structure and then assign pause durations according to the estimated break indices. The pause duration of each inter-syllable location can be simply assigned a constant value according to its break index, and then added with a random perturbation.[20] An alternative data-driven approach[23] uses a 3-layer RNN to predict pause duration directly from some syllable and word level linguistic features.

4. Parametric Synthesis

The technology of speech synthesis dates back to the parametric techniques introduced by Homer Dudley in the late 1930's and early 1940's.[55] These methods are "parametric" in the sense that they construct a computational model of the acoustic properties of the human vocal tract, and then analyze speech by determining the values of the parameters of the model. Speech is then generated from the model controlled by time-varying parameter trajectories.

4.1. *Parametric Representation of Speech Signal*

4.1.1. *Formant Synthesizer*

The formant synthesizer uses a number of formant resonances which can be realized with a second-order IIR filter to represent functions of the vocal tract in speech production. A filter with several resonances can be constructed by cascading several second-order sections (cascade model) or by adding several such sections together (parallel model). Formants which are used in formant synthesizers can be generated by rules or be data-driven. Due to the complicated

formant structures of compound vowels in Chinese, such as /iang, ian, iong/, etc., it is an even bigger challenge to develop a formant synthesizer for Chinese, compared to some other foreign languages. But despite that, there are still several very well-performing Chinese formant speech synthesizers.[56–60] In 1993, Lee[61] presented a set of improved tone concatenation rules to be used in a formant-based Chinese TTS system. A total of 14 representative tone patterns were defined for the five tones, and different rules about which pattern should be used under what kind of tone concatenation conditions were organized in detail. Preliminary subjective tests indicate that these rules actually produce better synthesized speech for a formant-based Chinese TTS system.

4.1.2. *Linear Predictive Coding*

Linear Predictive Coding (LPC) is a powerful method for speech analysis and synthesis. In recent years, several variations of linear prediction have been developed to improve the quality of the basic method.[62,63] These variations include Multi-pulse Linear Prediction Coding (MLPC), Residual Excited Linear Prediction (RELP), and Code Excited Linear Prediction (CELP).[64] A number of typical Chinese LPC synthesizers have been developed in the Institute of Acoustics, Chinese Academy of Sciences and the University of Science and Technology of China, where the speech code book method was integrated into their systems.[65,66]

Line Spectral Frequencies (LSF) serve as an equivalent representation of predictor coefficients. In practice, they are frequently used for their better quantization and interpolation properties. LSF can also be derived from the spectrum generated by other analysis methods, for example the STRAIGHT model to be described later, for parametric synthesis application.[67]

The Sinusoidal Model and the Harmonics Plus Noise Model have been proved effective for parametric synthesis due to the improvement of excitation.[68,69] But few reports are published regarding their application in MTTS.

4.1.3. *STRAIGHT Model*

The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) model[70] is a very high quality speech analysis-modification-synthesis method to represent and manipulate speech signals. It uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time and frequency region and consists of an F0 extraction using instantaneous frequency calculation based on a new concept

called *fundamentalness*. The proposed procedures preserve the details of time and frequency surfaces while almost perfectly removing the fine structures resulting from signal periodicity and allow for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, while maintaining high reproductive quality. The flexibility of STRAIGHT may help promote research on the relation between physical parameters and perceptual correlates. The F0 extraction procedure also provides a versatile method for investigating quasiperiodic structures in arbitrary signals. This method has been successfully applied to Chinese TTS systems, such as database compression,[71] HMM-based parametric synthesis[67] and voice conversion.[72]

5. Concatenative Speech Synthesis

In concatenative speech synthesis, there is always a unit inventory that stores pre-recorded speech segments. During synthesis, suitable segments are selected and concatenated with or without signal processing. Four key problem areas include: defining a base unit set; choosing a prosody strategy; designing a unit selection scheme and collecting and annotating a speech corpus.

5.1. *Base Unit Set*

A base unit in a concatenative speech synthesizer is the lowest constituent in the unit selection process. There are many possible basic unit choices, such as phoneme, diphone, semi-syllable, syllable or even word. In order to obtain natural prosody and smooth concatenation in synthetic speech, for each base unit, rich prosodic and phonetic variations are often expected. This is easy to achieve when smaller base units are used. However, smaller units mean more units per utterance and more instances per unit, and this implies a larger search space for unit selection and thus a longer search time. Besides, smaller units do cause more difficulties in precise unit segmentation. It is found that longer base units are useful as long as enough instances are guaranteed to appear in the database.[73]

Mandarin has only about 410 base syllables and about 2,000 tonal syllables. Therefore, tonal syllable is a natural choice for the base unit in Mandarin and it is used in most state-of-the-art Mandarin TTS systems.[74-77] However, when the speech database is as small as 1 to 2 hours of speech, initials plus tonal finals are the alternative unit choice. When a moderate-sized speech corpus is available, defining a base unit set that contains all initials and finals as well as some frequently used syllables is a good, balanced solution.

5.2. *Prosody Strategy*

There are three typical choices for prosody control in a concatenative TTS:

Fully controlled: In many TTS systems, numerical targets for prosodic features like F0, segmental duration and energy, are predicted first. These targets are fully realized in the synthesized speech by adjusting prosodic features with signal processing algorithms, such as PSOLA[78] or HNM.[69] Such a prosody strategy has been widely adopted in late 80's and early 90's, when only one or a few instances of a unit are allowed to be stored in the unit inventory. On the one hand, this prosody model draws out such a limited range of prosodic variation, that there can be few unexpected or bad prosodic outputs in the synthetic speech. However, on the other hand, the generated prosody often has a rather flat intonation pattern and the speech sounds monotonous. Furthermore, signal processing used for pitch and time scaling often have side-effects that distort the speech quality. To reduce the extent of signal processing, a semi-controlled strategy is used.

Semi-controlled: Numerical prosody targets are predicted and embedded into the target cost for unit selection. If the prosodic features of a selected unit are close enough to the predicted targets, no signal processing is needed. Otherwise, signal processing is performed. The advantage of this strategy is that the degree of pitch and time scaling are constrained to small ranges in most of the cases. However, such an advantage can be achieved only when a large enough speech corpus is used. The resulting synthetic prosody is still rather monotonous.

Soft controlled: There is neither a numerical prosody model, nor any signal processing involved in this strategy. Instead, contextual features used in traditional prosody models are used to predict a cluster of speech instances that share similar prosodic features. Therefore, under the soft controlled strategy, acceptable regions of prosodic features, rather than the best path in the features space, are predicted by minimizing the probability of violating the invariant property in prosody. Normally, more than one valid path is kept in the acceptable regions. The final choice can be either random among all candidates, or the pattern that has the closest counterpart in the corpus. The advantage of the soft control strategy is that synthetic speech will have richer variations in prosody, sounding close to the original speaker. When there is a large enough speech corpus available, the soft control strategy works well in most cases and it has been successfully applied in many Mandarin TTS systems.[79] However, it still has the disadvantage that some unnatural utterances will be generated when improper units are selected, especially when the speech corpus is not large enough.

Among the three prosody strategies, there is no universal good or bad choice. It depends on the target task and the size of speech corpus available.

5.3. *Unit Selection*

Normally, the suitability of the instance sequence for a given text is described by two types of costs: the target cost and the smoothness cost.[80,81] Target cost describes the local goodness of an instance in relation to its target, and smoothness cost measures how smooth the synthesized utterance will be by concatenating these instances one by one.

Target cost: After text processing and prosody processing, each unit in the text to be synthesized has been assigned a specification that describes its target phonetic features and prosodic features. The same set of features has been derived for all speech segments in the unit inventory. If a speech segment in the unit inventory matches the specification of a target unit exactly, the target cost for selecting this speech segment is zero. Otherwise, there will be a positive penalty. In some TTS systems, especially those adopting fully-controlled or semi-controlled prosody strategies, numerical targets are predicted for prosodic features such as pitch, duration and intensity. While, in other systems, especially those adopting the soft-controlled prosodic strategy, categorical features, such as position in phrases, position in words, presence or absence of stress, are used and penalties for mismatch in these categories are decided experientially. Besides, penalties can be calculated from the segment likelihood to a target HMM[81] or from the similarity of the left and right phones of a segment in the unit inventory to the target unit.[80] Since Chinese is a tonal language, penalties are also derived from the similarity of the left and right tones of a unit.

Smoothness cost: A smoothness cost is needed to measure how smooth it will be if two speech segments are concatenated together. In some systems, the spectral distance across the concatenating boundary is used as the smoothness penalty. However, such a measurement is mostly suitable for systems that use diphone as the base unit, in which the concatenating point is at the stable part of a speech phoneme. For Mandarin TTS systems that use syllables as base units, concatenating boundaries are often at the parts with rapid changes. Therefore, spectral distance is not a good measure of the smoothness of concatenation. A very simple smoothness cost has been introduced.[27] If two segments are continuous in the original recording, the smoothness cost between them is zero. Otherwise, a non-zero value is assigned according to the type of concatenation.

With such a constraint, the continuous segments in the unit inventory tend to be selected and the unvoiced-unvoiced concatenation is preferred.

The final concatenation cost of an utterance is the weighted sum of target cost and smoothness cost. The weighting of the importance of all these penalties is still an open problem. They are often decided experientially. When subjective evaluation results are available, weights can be tuned to maximize the correlation between the subjective score and the penalty.[82]

5.4. *Resources Needed for Creating a TTS Voice*

The different TTS systems do share some common resource requirements. The functions and existing issues in the process of generating these resources are described below:

Script generation: The goal is to maximize the coverage of prosodic and phonetic variations of the base units in a limited amount of text script. Thus, at least three parameters, including the base unit set, the function for calculating coverage and the total amount of script to be recorded, are to be decided according to the characteristics of the target language and the target scenario. Normally, script generation is performed as a sentence selection problem with a weighted greedy algorithm. The relationship between the size of a speech database and voice quality has been studied.[83]

Speech recording: The recording process is normally carried out by a professional team in a sound-proof studio. The voice talent is carefully selected and well-trained. With such considerations, the recorded speech, generally, has good quality. Yet, it often has some mismatched words between the speech and the script. These mismatches are mostly caused by reading errors and the idiosyncratic pronunciation of the speaker. Detecting these mismatches automatically is still an unsolved problem.[84]

Text processing: When generating recording script and the corresponding phonetic transcription, many text processing functions, such as text normalization and grapheme-to-phoneme conversion are needed. These processes typically do generate more errors and again will cause the mismatch between speech and phonetic transcription, and therefore the generated transcription should be checked manually.

Unit segmentation: To make a speech corpus usable to a concatenative TTS, the phonetic transcriptions has to be aligned with the corresponding speech waveforms. The HMM-based forced alignment has been widely adopted for automatic boundary alignment. Post-refining is often performed to guide the

boundaries moving toward the optimal locations for speech synthesis.[85] Besides, there are some improved approaches, such as discriminative training and explicit duration modeling, which have been introduced into HMM-based segmentation for Chinese speech.[86]

Prosody annotation: Prosody annotation is often performed on the speech corpus, either manually or automatically.[87] For Mandarin, the most important annotation is the break index.

6. Summary and Conclusion

6.1. *Perspective*

In the past years, there have been significant achievements in the field of speech synthesis research. Now that the intelligibility of synthetic speech is close approaching that of human speech, more diverse and more attractive research areas are possible, which will bring about more innovation and propagation to speech research. Among them, personalized speech synthesis (including speaker simulation/adaptation, expressive speech synthesis), HMM-based speech synthesis, and articulatory speech synthesis are the most active domains.

6.1.1. *Personalized Speech Synthesis*

A personalized TTS system is more expressive and valuable than a universal single voice/style and more appropriate for practical applications. It includes two parts: speaker simulation, which tries to synthesize a range of different voices that users can choose from, and Expressive Speech Synthesis (ESS), which tries to synthesize voices that contain more human expressions.

6.1.1.1. Speaker Simulation

Starting from a speech signal uttered by a speaker, speaker simulation, also called voice transformation, voice conversion (VC), or voice morphing, aims at transforming the characteristics of the speech signal in such a way that a person naturally perceives the target speaker's own characteristics in the transformed speech.[88] Most VC systems to date have been focusing on transforming the spectral envelope. Mapping codebooks,[90] linear regression and dynamic frequency warping (DFW),[91] and Gaussian mixture modeling (GMM)[89,92] are three popular mapping methods of spectral conversion. Because of the difficulty to extract and manipulate higher-level information with present speech technologies, prosodic features such as F0 contour, energy contour and speaking

rate of the source speaker are often trivially adjusted to match the target speaker's average prosody.[89] At present, simulating F0 contour is the emphasis of prosody conversion, and the statistical model, the deterministic/stochastic model, piecewise linear mapping,[93] the CART model[72] and pitch target model[94] can be employed in the simulation of F0 contours.

Some attempts on Chinese have been made[72] and effective conversions achieved. However, VC is a complex task involving speech analysis, time alignment, mapping algorithm, speech synthesis, and other speech technologies. Based on that, a perfect VC system is still an unrealized application.

6.1.1.2. Expressive Speech Synthesis

ESS can offer a much more human-like scenario in human-machine interactions. Traditional methods on ESS consist of formant synthesis with rule based prosodic control, diphone concatenation and so on. But none of these methods can generate satisfying results. A number of new methods have been explored recently. These are unit selection based ESS and voice conversion based ESS.

(1) Expressive speech synthesis based on unit selection

The unit selection method is the most popular method in normal speech synthesis, which has been applied to ESS. Iida[95] constructed an emotional speech synthesis system based on unit selection, by recording three unit selection databases using the same speaker for three kinds of emotions: anger, joy and sadness. When synthesizing speech with these given emotions, only units from the corresponding database are selected. The evaluation experiment shows that 50-80% of the synthesized speech can be easily recognized. Another approach is to select the appropriate unit for the given emotion from only one database. This has been attempted by Marumoto and Campbell,[96] who used parameters related to voice quality and prosody as emotion-specific selection criteria. The results indicated a partial success: anger and sadness were recognized with up to 60% accuracy, while joy was not recognized above chance level.

(2) Expressive speech synthesis based on voice conversion

In this method, a neutral speech is converted to an emotional speech using mapping functions of a spectrum, F0 and other prosodic features. However, there is one big problem that needs to be resolved. Among these features, which ones are most important and which ones can be neglected.

Some researchers focus on voice quality. Kawanami[97] constructed an emotional speech synthesis system based on voice conversion, which used GMM and DFW to construct the mapping function for Mel-cesptrum derived from the STRAIGHT spectrum. As to the fundamental frequency, it was simply converted using the standard linear mean-variance transformation. However, other researchers believe that fundamental frequencies play the most important role. Kang *et al.*[94] used a parametric F0 model to explore underlying relations between source and target F0 contours for Mandarin ESS, and the pitch target model was selected for its capability of describing Mandarin F0 contour and its convenience for parametric alignment. The GMM and CART methods were used to build mapping functions for well-chosen pitch target parameters.

Although the systems mentioned above achieve good results, there is no conclusion about which feature is most important. It is possible that in different kinds of emotional speech, or in different languages, the same parameter plays different roles.

6.1.2. *HMM-based Speech Synthesis*

Hidden Markov processes are a powerful and tractable method of modeling non-stationary signals, which have been frequently used in speech recognition. Recently, a HMM-based synthesis system has been developed, where spectral and excitation parameters are extracted from a speech database and modeled by context-dependent HMMs. In the synthesis part of the system, spectral and excitation parameters are generated from the HMMs themselves. Then waveforms are generated based on a decoding process.[98] Compared with traditional method based on concatenation, this new system has many benefits:

(1) HMM-based systems can generate smooth and natural sounding speech, while there are always some inconsistencies at the concatenation points of the synthesized speech of concatenative systems.

(2) HMM-based systems can freely change the target voice characteristics by changing the parameters of the HMMs, while concatenation systems can only synthesize and generate the voice of only one speaker.

(3) HMM-based systems need comparatively smaller corpora compared to concatenative speech synthesis systems.

Although the speech quality from HMM-based systems is not as good as that of concatenative systems, this can be much improved by a high quality decoder, such as the STRAIGHT algorithm. Besides, there are some new criteria, such as the Minimum Generation Error,[99] which have been introduced into model

training, achieving satisfactory improvements in the performance of Mandarin TTS.

6.1.3. *Articulatory Speech Synthesis*

Articulatory models can be divided into two and three dimensional models on one hand, and into geometric, statistical and biomechanical models on the other hand. Statistical 3D models have the advantage of having relatively few uncorrelated parameters. However, these models require huge amounts of MRI or CT (Computed Tomography) data for their construction and they are usually specific for a particular speaker. Biomechanical vocal tract models simulate the behavior of the articulators by means of finite element methods. They are especially suited to facilitate new insights into the relation between muscle activation and articulatory movements. On the other hand, they have many degrees of freedom, are difficult to control and require much computational power. Geometric vocal tract models are similar to statistical models in that their parameters define the vocal tract shape directly in geometrical terms, but the kinds and number of parameters are chosen *a priori* and fitted to particular data *a posteriori*. Therefore, the geometric vocal tract model has become the most popular method, achieving good results.[100]

6.2. *Applications*

With the rapid development of speech communication technology, TTS systems are being more widely applied in our daily lives.

(1) Call Centers

Since 2000, MTTS has been introduced to call centers on a large scale, to synthesize prompt sentences or queried information in interactive systems, especially dynamic information. InterPhonic,[101] a bilingual (Mandarin and English) TTS engine, now runs in most of the call centers in mainland China, including in the 168 information line, the voice portal of Unicom, the PICC countrywide call center, etc. Recently, China Telecom upgraded its 114 information service platform to the "Best Tone" (号码百事通) service which is powered by the InterPhonic engine. Mandarin TTS can significantly relieve the workload of human operators, or even substitute operators by being integrated with telephone-keyboard input or speech recognition technologies, along with simple (or even complex) dialogue management technologies.

(2) Mobile Phone Utility

With the advent of the information age, the use of mobile phones has become pervasive. Very natural speech can now be generated by state-of-the-art HMM-based MTTS with very limited resources. This means that MTTS systems can also be ported into mobile phones or other small devices. The phone can then read out short messages to you while you listen. Whenever there is a new call or scheduled item due, the caller's name or event can be read out to you, even before you see it on screen. In fact, the mobile phone may even serve as an online language learning device, or a translator.

(3) GPS Car Navigation Systems

The intelligent car has become a worldwide focus, and intelligent speech synthesis technology plays an important role in this technological pursuit. This navigation application involves the broadcasting of locations and directions of target destinations or stopovers, such as gas stations, parks, hotels, and so on. Integrated with wireless communication, the system can also broadcast real-time traffic, news, and weather forecast. In China, the ratio of speech-interfaced car navigation systems is only 2% of all car navigation systems, while it is 50% in Japan and 25% in America and Europe. The potential for wider MTTS deployment in this application area is therefore tremendous.

(4) Entertainment Industry

Recently, the application of speech synthesis system has entered the field of entertainment. For example, in games that involve voices, game characters can speak using the player's own voice. For users of e-books, listening to the book, instead of visually reading it, is possible with the integration of TTS, and even desirable when reading is an inconvenience or a hazard, such as in moving vehicles. Also, having your personal voice heard while text-chatting in cyberspace can add another dimension and color to your chat-identity.

With the integration of other speech technologies such as speech recognition and machine translation, the application of MTTS can be expanded to a much broader range than ever before.

References

1. Ching Y. Suen, "Computer Synthesis of Mandarin", *IEEE ICASSP*, p.698-700, (1976).
2. Samuel C. Lee, Shilin Xu and Bailing Guo, "Microcomputer-generated Chinese Speech", *Computer Processing of Chinese and Oriental Languages*, vol.1, no.2, p.87-103, (1983).

3. Jialu Zhang, "Acoustic Parameters and Phonological Rules of a Text-to-Speech Systems for Chinese", *IEEE ICASSP*, p.2023-2026, (1986).
4. Chilin Shih and Mark Y. Liberman, "A Chinese Tone Synthesizer", Technical Report, *AT&T Bell Laboratories* (1987).
5. Shun-an Yang and Yi Xu, "An Acoustic-phonetic Oriented System for Synthesizing Chinese", *Speech Communication*, vol.7, p.317-325, (1988).
6. Tai-Yi Huang, Cai-Fei Wang and Yoh-Han Pao, "A Chinese Text-to-Speech Synthesis System Based on an Initial-Final Model", *Computer Processing of Chinese and Oriental Languages*, vol.1, no.1, p.59-70, (1983).
7. Wen C. Lin and Tzien-Tsai Luo, "Synthesis of Mandarin by Means of Chinese Phonemes and Phonemes-pairs (JIFH)", *Computer Processing of Chinese and Oriental Languages*, vol.2, no.1, p.23-35, (1985).
8. Ming Ouh-Young, Chin-Jiang Shie, Chiu-Yu Tseng and Lin-Shan Lee, "A Chinese Text-to-Speech System Based upon a Syllable Concatenation Model", *IEEE ICASSP*, p.2439-2442, (1986).
9. Hong-Bin Chiou, Hsiao-Chuan Wang and Yueh-Chin Chang, "Synthesis of Mandarin Speech Based on Hybrid Concatenation", *Computer Processing of Chinese and Oriental Languages*, vol.5, no.3/4, p.217-231, (1991).
10. Lin-Shan Lee, Chiu-Yu Tseng and Ming Ouh-Young, "The Synthesis Rules in a Chinese Text-to-Speech System", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.37, no.9, p.1309-1320, (1989).
11. John Choi, Hsiao-Wuen Hon, Jean-Luc Lebrun, Sun-Pin Lee, Gareth Loudon, Viet-Hoang Phan and Yogananthan S. Yanhui, "A Software Based High Performance Mandarin Text-to-Speech System", *ROCLING VII*, p.35-50, (1994).
12. Lianhong Cai, Hao Liu and Qiaofeng Zhou, "Design and Achievement of a Chinese Text-to-Speech System under Windows", *Microcomputer*, vol.3 (1995).
13. Min Chu and Shinan Lu, "High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules", *the XIII Intenational Congress of Phonetic Sciences*, p.334-337, (1995).
14. Shaw-Hwa Hwang, Yih-Ru Wang and Sin-Horng Chen, "A Mandarin Text-to-Speech System", *ICSLP* (1996).
15. Jun Xu and Baozong Yuan, "New Generation of Chinese Text-to-Speech System", *IEEE TENCON*, p.1078-1081, (1993).
16. Benjamin Ao, Chilin Shih and Richard Sproat, "A Corpus-Based Mandarin Text-to-Speech Synthesizer", *ICSLP*, p.1771-1774, (1994).
17. Chilin Shih and Richard Sproat, "Issues in Text-to-Speech Conversion for Mandarin", *Computational Linguistics and Chinese Language Processing*, vol.1, no.1, p.37-86, (1996).
18. Ren-Hua Wang, Qinfeng Liu and Difei Tang, "A New Chinese Text-to-Speech System with High Naturalness", *ICSLP*, p.1441-1444, (1996).
19. Ren-Hua Wang, Qinfeng Liu, Yu Hu, Bo Yin and Xiaoru Wu, "KD2000 Chinese Text-to-Speech System", *ICMI*, p.300-307, (2000).
20. Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, "Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis", *ICSLP*, p.1624-1627, (1996).
21. ShaoHuang Pin, Yehlin Lee, Yong-cheng Chen, Hsin-min Wang and Chiu-yu Tseng, "A Mandarin TTS system with an Integrated Prosodic Model", *ISCSLP*, p.169-172, (2004).
22. Chung-Hsien Wu and Jau-Hung Chen, "Template-Driven Generation of Prosodic Information for Chinese Concatenative Synthesis", *IEEE ICASSP*, vol.1, p.65-68, (1999).
23. Sin-Horng Chen, Shaw-Hwa Hwang and Yih-Ru Wang, "An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-Speech", *IEEE Trans. on Speech and Audio Processing*, vol.6, no.3, p.226-239, (1998).

24. Ming-Shing Yu and Neng-Huang Pan, "A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-To-Speech System", *Journal of the Chinese Institute of Engineers*, vol.28, no.5, p. 385-399, (2005).
25. Fu-chiang Chou and Chiu-yu Tseng, "Corpus-based Mandarin Speech Synthesis with Contextual Syllabic Units Based on Phonetic Properties", *IEEE ICASSP*, p. 893-896, (1998).
26. Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, "A set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, vol.10, no.7, p.481-494, (2002).
27. Min Chu, Hu Peng, Yong Zhao, Zhengyu Niu and Eric Wang, Microsoft Mulan – "A Bilingual TTS System", *IEEE ICASSP* (2003).
28. Minghui Dong, Kim-Teng Lua and Haizhou Li, "A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese", *Journal of Chinese Language and Computing*, vol.16, no.1 (2006).
29. Ren-Hua Wang, Zhongke Ma, Wei Li and Donglai Zhu, "A Corpus-based Chinese Speech Synthesis with Contextual Dependent Unit Selection", *ICSLP*, vol.2, p.391-394, (2000).
30. Zhen-Hua Ling, Yu Hu, Zhi-Wei Shuang and Ren-Hua Wang, "Decision Tree Based Unit Pre-selection in Mandarin Chinese Synthesis", *ISCSLP* (2002).
31. M. Beckman and G. Ayers Elam, Guidelines for ToBI Labeling, Version 3 (1997).
32. The list of frequently used characters in modern Chinese (《现代汉语常用字表》)[in Chinese] http://www.gmw.cn/content/2004-07/29/content_67735.htm.
33. Lexicography and Chinese dictionary compilation group in Institute of Linguistics, CASS, Ed (中国社会科学院语言所词典室词典编辑室编), The contemporary Chinese dictionary, the 5th edition(《现代汉语词典》) [in Chinese], *the commercial press* (2002).
34. Y. Qian, M. Chu and H. Peng, "Segmenting unrestricted Chinese text into prosodic words instead of lexical words", *IEEE ICASSP* (2001).
35. M. Chu and Y. Qian, "Locating boundaries for prosodic constituent in unrestricted Mandarin texts", *Computational Linguistics and Chinese Language Processing*, vol.6, no.1, p.61-82, (2001).
36. S. Zhao, J.H. Tao and L.H. Cai, "Prosodic phrasing with inductive learning", *ICSLP* (2002).
37. Z. R. Zhang, M. Chu and E. Chang, "An efficient way to learn rules for grapheme-to-phoneme conversion in Chinese", *ISCSLP*, Taipei (2002).
38. W. J. Wang, S. H. Hwang and S. H. Chen, "The broad study of homograph disambiguity for Mandarin speech synthesis", *ICSLP* (1996).
39. M. Zheng and L. H. Cai, "A new rule-based method of automatic phonetic notation on polyphones", *ICSP'04* (2004).
40. Ben-Feng Chen, Guo-Ping Hu and Ren-Hua Wang, "Large lexicon construction for TTS system", *ISCSLP* (2002).
41. Hua-Ping Zhang, Qun Liu, Hong-Kui Yu, Xue-Qi Cheng and Shuo Bai, "Chinese Named Entity Recognition Using Role Model", *Computational Linguistics and Chinese Language Processing*, vol. 8, no. 2 (2003).
42. Shiwen Yu, "Annotation for the Dictionary of Modern Chinese Grammar Information", *TsingHua University Press* (1998).
43. Zi-rong Zhang, Min Chu and Eric Chang, "An Efficient Way to Learn Rules for Grapheme-to-Phoneme Conversion in Chinese", *ISCSLP* (2002).
44. Guo-ping Hu, Zhi-Gang Chen and Ren-Hua Wang, "A Rule-Based Approach with SVM-Based Weight Estimation for Phoneme Disambiguation of Polyphone", *ICCPOL*, p.599-605, (2003).
45. Xipeng Shen and Bo Xu, "A CART-based Hierarchical Stochastic Model for Prosodic Phrasing in Chinese", *ISCSLP* (2000).

46. Jian-Feng Li, Guo-ping Hu and Ren-hua Wang, "Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model". *ICSLP* (2004).
47. Jian-Feng Li, Ming Fan, Guo-Ping Hu and Ren-Hua Wang, "Text Chunking for Intonational Phrase Prediction in Chinese", *NLP-KE*, p.231-237, (2003).
48. Zhi-Gang Chen, Guo-Ping Hu and Xu-Fa Wang, "Text Normalization In Chinese Text-To-Speech System", *Journal of Chinese information processing*, vol.17, no.4, p.45-51, (2003).
49. Yuan Ren Chao, A Grammar of Spoken Chinese, *University of California Press* (1968).
50. Chiu-yu Tseng, Shao-huang Pin, Yeh-lin Lee, Hsin-min Wang and Yong-cheng Chen, "Fluent speech prosody: framework and modeling", *Speech Communication*, vol.46, issues 3-4, *Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation*, p.284-309, (2005).
51. Ming-Shing Yu, Neng-Huang Pan and Ming-Jer Wu, "A Intonation Prediction Model that can Outputs Real Pitch Pattern", *the Seventh Conference on Artificial Intelligence and Applications*, p.784-788, (2002).
52. Gao-peng Chen, Yu Hu, Ren-Hua Wang, "A Concatenative-Tone Model With Its Parameters' Extraction", *Speech Prosody 2004* (2004).
53. Sun Lu, Yu Hu and Ren-Hua Wang, "Polynomial Regression Model for Duration Prediction in Mandarin", *Journal of Chinese Information Processing* (2005).
54. Zong-Ji Wu, "Can Poly-Syllabic Tone-Sandhi Patterns be the Invariant Units of Intonation in Spoken Standard Chinese", *ICSLP*, p.12.10.1-12.10.4, (1990).
55. Sami Lemmetty, "Review of Speech Synthesis Technology", *Master thesis, Helsinki Univ. of Technology*
56. Ziying Li, "合成无限词汇汉语语言的初步研究 " [in Chinese], *Chinese Journal of Acoustics*, vol.5, p.291-298.
57. Jialu Zhang, "汉语文语转换系统的语音规则和声学参数 " [in Chinese], *Chinese Journal of Acoustics*,vol.15, no.22, p. 113-120.
58. Shun-an Yang, "面向声学语音学的普通话语音合成技术" [in Chinese], *Social Sciences Academic Press*.
59. Shinan Lu and A. Almeida, "The Effects of Voice Disguise Upon Formant Transition", *IEEE ICASSP*, p.885-888, (1986).
60. Shinan Lu, Jialu Zhang and Shiqian Qi, "Chinese text-to-speech system based on parallel formant synthesizer", *14th International Congress on Acoustics* (1992).
61. L. Lee, C. Tseng and C. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system", *IEEE Trans. on Speech and Audio Processing*, vol.1, no.3, p.287-294, (1993).
62. D. Childers and H. Hu, "Speech Synthesis by Glottal Excited Linear Prediction", *Journal of the Acoustical Society of America*, vol. 96 (4), p.2026-2036, (1994).
63. R. Donovan, Trainable Speech Synthesis, *PhD. Thesis. Cambridge University Engineering Department*, England (1996).
64. G. Campos and E. Gouvea, "Speech Synthesis Using the CELP Algorithm", *ICSLP* (1996).
65. Fuyuan Mo, Changli Li, Hong Ni, Jingchen Sun and Tong Li, "Chinese All-Syllable Real Time Synthesis System", *International Conference on Signal Processing*, p.369-372, (1990).
66. Qingfeng Liu and Ren-Hua Wang, "A new synthesis method based on the LMA vocal tract model", *Chinese Journal of Acoustics*, vol.17, no.2, p153-162, (1998).
67. Yi-Jian Wu and Ren-Hua Wang, "HMM-based trainable speech synthesis for Chinese" [in Chinese], *Journal of Chinese Information Processing*, accepted
68. R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, p.744–745, (1986).

69. Yannis Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis", *IEEE Trans. on Speech and Audio Processing*, vol. 9, p.21–29, (2001).
70. H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication*, 27(3-4), p.187-207, (1999).
71. Zhen-Hua Ling, Yu Hu, Zhi-Wei Shuang and Ren-Hua Wang, "Compression of Speech Database by Feature Separation and Pattern Clustering Using STRAIGHT", *ICSLP* (2004).
72. Zhi-Wei Shuang, Zi-Xiang Wang, Zhen-Hua Ling and Ren-Hua Wang, "A Novel Voice Conversion System based on Codebook Mapping with Phoneme-tied Weighting", *ICSLP* (2004).
73. Y. N. Chen, Y. Zhao and M. Chu, "Customizing Base Unit Set with Speech Database in TTS Systems", *Eurospeech* (2005).
74. M. Chu and S. N. Lu, "A Text-to-Speech System with High Intelligibility and High Naturalness for Chinese", *Chinese Journal of Acoustics*, vol.15, no.1, p.81-90, (1996).
75. S. H. Hwang, S. H. Chen and Y. R. Wang, "A Mandarin Text-to-Speech System", *ICSLP* (1996).
76. Min Chu, Hu Peng, Hong-Yun Yang and Eric Chang, "Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer", *IEEE ICASSP* (2001).
77. F. C. Chou, C. Y. Tseng and L. S. Lee, "A Set of Corpus-Based Text-to-Speech Technologies for Mandarin Chinese", *IEEE Trans. on Speech and Audio Processing*, vol. 10, issue 7, p.481-494, (2002).
78. E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphone", *Speech Communication*, vol. 9, p.453-467, (1990).
79. M. Chu, Y. Zhao and E. Chang, "Modeling Stylized Invariance and Local Variability of Prosody in Text-to-Speech Synthesis", *Speech Communication*, vol. 48, p.716-726, (2006).
80. A. Black and N. Campbell, "Optimizing Selection of Units from Speech Database for Concatenative Synthesis", *IEEE ICASSP*, p.373-376, (1996).
81. H. Hon, A. Acero, S. Huang, J. Liu and M. Plumpe, "Automatic Generation of Synthesis Units for Trainable Text-to-Speech System", *IEEE ICASSP*, vol.1, p.293-296, (1998).
82. H. Peng, Y. Zhao and M. Chu, "Perpetually Optimizing the Cost Function for Unit Selection in a TTS System with One Single Run of MOS Evaluation", *ICSLP* (2002).
83. Y. Zhao, M. Chu, H. Peng and E. Chang, "Custom-Tailoring TTS Voice Font – Keeping the Naturalness When Reducing Database Size", *Eurospeech* (2003).
84. L. J. Wang, Y. Zhao, M. Chu, F. K. Soong and Z. G. Cao, "Phonetic Transcription Verification with Generalized Posterior Probability", *Eurospeech* (2005).
85. L. J. Wang, Y. Zhao, M. Chu, F. K. Soong, J. L. Zhou and Z. G. Cao, "Context-Dependent Boundary Model for Refining Boundaries Segmentation of TTS Units", *IEICE Trans. on Information and System*, vol. E89-D, no. 3, p.1082-1091, (2006).
86. Yi-Jian Wu, Hisashi Kawai, Jinfu Ni and Ren-Hua Wang, "Discriminative training and explicit duration modeling for HMM-based automatic segmentation", *Speech Communication*, vol. 47 (2005).
87. Y. N. Chen, M. Lai, M. Chu, F. K. Soong, Y. Zhao and F. Y. Hu, "Automatic Accent Annotation with Limited Manually Labeled Data", *Speech Prosody* (2006).
88. E. Moulines and Y. Sagisaka, "Voice conversion: State of the art and perspectives", *Speech Communication*, vol. 16, no. 2, p.125–126, (1995).
89. Alexander Blouke Kain, High Resolution Voice Transformation, *Ph.D. thesis, Oregon Health and Science Unitversity* (2001).

90. M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization", *IEEE ICASSP*, p.655-658, (1988).
91. H. Valbret and et al, "Voice transformation using psola technique", *Speech Communication*, vol. 11, no. 2-3, p. 175–187, (1992).
92. Y. Stylianou and et al, "Continuous probabilistic transform for voice conversion", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, p.131– 142, (1998).
93. T. Ceyssens and et al, "On the construction of a pitch conversion system", *EUSIPCO*, (2002).
94. Yongguo Kang, Jianhua Tao and Bo Xu, "Applying Pitch Target Model to Convert F0 Contour for Expressive Mandarin Speech Synthesis", *IEEE ICASSP* (2006).
95. A. Iida, N. Campbell, S. Iga, F. Higuchi and M. Yasumura, "A Speech Synthesis System for Assisting Communication", *ISCA Workshop on Speech & Emotion*, p.167-172, (2000).
96. T. Marumoto and N. Campbell, "Control of speaking types for emotion in a speech re-sequencing system" [in Japanese], *the Acoustic Society of Japan, Spring meeting*, p.213-214, (2000).
97. H. Kawanami, Y. Iwami and T. Toda, "Gmm-based voice conversion applied to emotional speech synthesis", *Eurospeech*, p.2401-2404, (2003).
98. T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", *Eurospeech*, p. 2347-2350, (1999).
99. Yi-Jian Wu and Ren-Hua Wang, "Minimum Generation Error Training for HMM-based Speech Synthesis", *IEEE ICASSP* (2006).
100. Peter Birkholz and Dietmar Jackèl, "Construction and Control of a Three-Dimensional Vocal Tract Model", *IEEE ICASSP* (2006).
101. http://www.iflytek.com