Text-to-Speech Synthesis for Mandarin Chinese

Yuan Yuan Li
Department of Computer & Information Sciences
Minnesota State University, Mankato
yuan.li@mnsu.edu

Steven Case
Department of Computer & Information Sciences
Minnesota State University, Mankato
steven.case@mnsu.edu

Abstract

A Text-To-Speech (TTS) synthesizer is a computer-based system that is able to automatically read text aloud, regardless whether the text is introduced by computer input stream or a scanned input that is submitted to an optical character recognition (OCR) engine. TTS synthesis can be used in many areas, such as telecommunication services, language education, vocal monitoring, multimedia, and as an aid to handicapped people. Generally speaking a TTS system can be divided into two major components: natural language processing (NLP) and digital signal processing (DSP). The major task of the NLP component is to gather as much linguistic information as it can and pass this information into the DSP component. The DSP component can function on its own. It may or may not use the linguistic information that was generating by the NLP component to produce output speech. In the past years, many studies have focused on Text-To-Speech (TTS) systems for different languages. In particular, Mandarin Chinese TTS systems have made significant progresses in the last two decades.

This paper gives a detailed overview of TTS systems for English and Chinese. The objective of this paper identifies the primary differences distinguishing Chinese TTS systems and English TTS systems, which mainly lie within the generation of synthesis units and prosody information. The paper will begin with a brief introduction to the major components in a text-to-speech system along with the main techniques that are used within each component. The paper will then explain how the synthesis units and prosody information are usually generated in a typical Mandarin Chinese TTS system.

Introduction

Dutoit [4] identified a Text-To-Speech (TTS) synthesizer as a computer-based system that should be able to read text aloud, regardless whether the text is introduced by computer input stream or a scanned input that is submitted to an optical character recognition (OCR) engine. This TTS synthesis should be intelligent enough to read "new" words/sentences and the speech it produces should be "natural" like human. Thus, a formal definition of text-to-speech is "the production of speech by machines, by way of the automatic phonetization of the sentences to utter".

The concept of high quality TTS synthesis appeared in the mid-eighties, as a result of important developments in speech synthesis and natural language processing techniques, mostly due to the emergence of new technologies like Digital Signal and Logical Inference Processors [4].

Text-to-speech synthesis can be used in many areas, such as telecommunications services, language education, vocal monitoring and, multimedia applications as well as an aid to handicapped people. Furthermore, the potential applications for such technology include teaching aids, text reading, and talking books/toys [4]. However, most TTS systems today only focus on a limited domain of applications, e.g. travel planning, weather services, and baggage lost-and-found [1].

TTS systems and synthesis technology for Chinese languages have been developed in the last two decades [15]. The main difference between general purpose TTS systems and Mandarin Chinese TTS systems are Mandarin Chinese TTS systems focus on the generation of synthesis units and prosodic information.

This paper is intended to provide a background on text-to-speech synthesis with an emphasis on text-to-speech synthesis for Mandarin Chinese. The paper begins with a brief overview of general text-to-speech systems and tradition English text-to-speech solutions. The paper then introduces the reader to current practices for TTS systems supporting Mandarin Chinese, with particular focus on synthesis unit selection and prosody generation. Finally, concluding remarks are presented.

An Overview of Text-to-Speech (TTS) Synthesis

Figure 1 is a simple functional diagram of a general TTS synthesizer. A TTS system is composed of two main parts, the Natural Language Processing (NLP) module and the Digital Signal Processing (DSP) module.

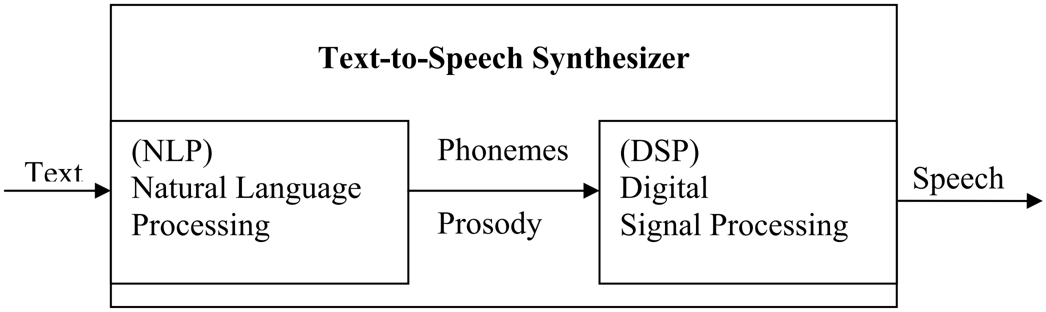

Figure 1. A General TTS Synthesizer

The NLP module takes a series of text input and produces a phonetic transcription together with the desired intonation and prosody (rhythm) that is ready to pass on the DSP module. There are three major components within the NLP module, the letter-to-sound component, the prosody generation component, and the morpho-syntactic analyzer component [4].

The DSP module takes the phonemes and prosody that were generated by the NLP module and transforms them into speech. There are two main approaches used by DSP module: rule-based-synthesis approach and concatenative-synthesis approach [4].

Many researchers, such as Edgington *et al.* [5,6], refer to the NLP module as the text-to-phoneme module and the DSP module as the phoneme-to-speech module.

Natural Language Processing Module

Figure 2 introduces the functional view of a NLP module of a general Text-to-Speech conversion system. The NLP module is composed of three major components: text-analyzer, letter-to-sound (LTS), and prosody generator.

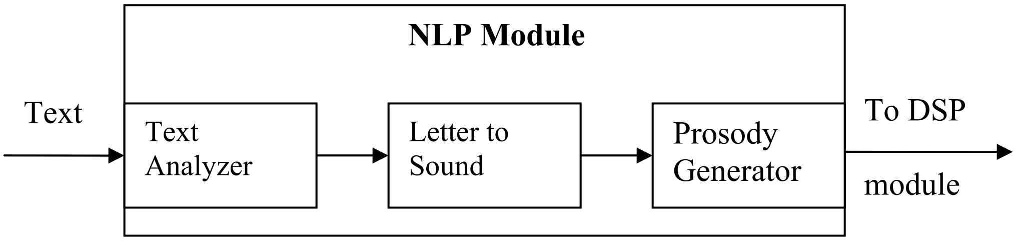

Figure 2. A Simple NLP Module

Besides the expected letter-to-sound and prosody generation blocks, the NLP module comprises a morpho-syntactic analyzer, underlying the need for some syntactic processing in a high quality TTS system. Being able to reduce a given sentence into something similar to the sequence of its parts-of-speech (POS), and further to describe it in the form of a syntax tree that unveils its internal structure is required for the following two reasons. First, accurate phonetic transcription can only be achieved by knowing the dependency relationship between the successive words and the words' part of speech category. Second, natural prosody relies heavily on syntax.

Text/Linguistic Analysis

Wu and Chen [15] suggested that text analysis is a language-dependent component in TTS system. It is invoked to analyze the input text. This process can be divided into three major steps:

- Pre-processing - At this stage, the main components of the input text are identified. Pre-processing also determines exact boundaries of the input, which is usually done by delimiting characters like white space, tab or carriage return. In addition, this task identifies abbreviations, numbers, and acronyms within the body of text and transfers them into a predefined format. Usually, pre-processing also segments the whole body of text into paragraphs and organizes these paragraphs into sentences. Finally, pre-processing divides the sentences into words. [4,5,6]
- Morphological analysis - The morphological analysis serves the purpose of generating pronunciations and syntactic information for every word (or lexical phrases) in the text. Most languages have a very large and ever-increasing number of words. Thus, it is impossible to produce an absolutely complete dictionary. Morphological analysis determines the 'root' form of every word, (i.e. love is a root form of loves), and allows the dictionary to store just headword entries, rather than all derived forms of a word. [4,5,6]
- Contextual analysis - This task considers words in their context and determines the part-of-speech (POS) for each word in the sentence. To aid this process we have to know the corresponding possible parts of speech of neighboring words. Context analysis is essential to solve problems like homographs (words that are spelled the same way but have different pronunciations). [4,5,6]

Letter-to-Sound (LTS)

Dutoit [4] indicates the Letter-To-Sound (LTS) module is responsible for automatically determining the incoming text's phonetic transcription. There are two different types of popular modules used within this task, a dictionary-based module or a rule-based module.

According to Dutoit, dictionary-based solutions are based on a large database of phonological knowledge. In order to keep the dictionary size reasonably small, entries are generally restricted to morphemes. Examples of morphemes are man, walk, words ending with 'ed', etc. The pronunciation of surface forms is accounted for by inflectional, derivational, and compounding morphophonemic rules [15].

On the other hand, Dutoit goes on to indicate that a different strategy is adopted in rule-based transcription systems, which transfer most of the phonological competence of dictionaries into a set of rules. This time, only those words that are pronounced in such a particular way that they constitute a rule on their own are stored in an exceptions dictionary. Since many exceptions are found in the most frequent words, a reasonably small exceptions dictionary can account for a large fraction of the words in a running text. For instance, in English, 2000 words typically suffice to cover 70% of the words in text.

In the early days powerful dictionary-based methods, which were inherently capable of achieving higher accuracy than letter-to-sound rules, were common given the availability of very large phonetic dictionaries on computers. Dutoit believes that recently, considerable efforts have been made towards designing sets of rules with a very wide coverage - start from computerized dictionaries and add rules and exceptions until all words are covered.

Prosody Generation

Dutoit [4] states that “Prosody refers to certain properties of the speech signal such as audible changes in pitch, loudness, and syllable length”. Bulyko [1] believed that one of the key problems with current TTS systems is poor prosody prediction, which gives information on the duration and the focus of the speech. He also believed that prosodic patterns are difficult to predict because they depend on high level factors such as syntax and discourse which have been less well studied in terms of their acoustic consequences.

Digital Signal Processing (DSP) Module

Generally, the DSP module takes the phonemes and prosodic information that were generated by the NLP module and turns them into speech signals. Sometimes, it might not use the phonemes and prosodic information that were generated by NLP module.

There are two main techniques used in the DSP module, a rule-based synthesizer or a concatenative-based synthesizer.

Rule-Based Synthesizers

A formal definition by Dutoit [4] for rule-based synthesizers is “the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements.” Rule-based synthesizers consist of a series of rules which formally describe the influence of phonemes on one another and are mostly in favor of phoneticians and philologists, as they constitute a cognitive, generative approach of the phonation mechanism.

Dutoit also stated that “For historical and practical reasons (mainly the need for a physical interpretability of the model), rule synthesizers always appear in the form of formant synthesizers. This approach describes speech as the interactions of formant and anti-formant frequencies and bandwidth, and glottal waveform”.

Dutoit goes on to clarify that rule-based synthesizers remained as a potentially powerful approach to speech synthesis. The advantage of this type of synthesizers is it allows speaker-dependent voice features so that switching from one synthetic voice into another can be achieved with the help of specialized rules in the rule database. However, the disadvantage of the rule-based synthesizer lies with the difficulty of collecting complete rules to describe the prosody diversity. Moreover, the derivation of the rule is labor-intensive and tedious. [4]

Concatenative Synthesizers

Concatenative synthesizers use real recorded speech as the synthesis units and concatenate the units together to produce speech. Dutoit [4] believed that the concatenative speech synthesis is the simplest and the most effective approach. In addition, Wu and Chen [15] indicate that this approach is adapted by most of the TTS systems today. Thus, by using concatenative approach unit selection becomes critical for producing high-quality speech. Speech units have to be chosen such that they minimize future concatenation problems such as disjoint speech. Usually, speech units are stored in a huge database.

In the past, phonemes have been adopted as the basic synthesis units. Using phonemes as the synthesis unit requires a small storage, but Wu and Chen indicate that it causes a lot of discontinuity between adjacent units. As the result, Dutoit suggests that other synthesis units, such as diphones and triphones are often chosen as speech units because they are involved in the most articulation while requiring affordable an amount of memory.

According to Dutoit, the models employed in concatenative synthesis are mostly based on signal processing tools and the most representative members are Linear Prediction Coding (LPC) synthesizers, Harmonic/Stochastic (H/S), and Time-Domain Pitch-Synchronous-OverLap-Add (TD-PSOLA). Dutoit states that TD-PSOLA is a time-domain algorithm and also believes it is the best concatenative method available today. However, in reality, the H/S model is more powerful than TD-PSOLA but it requires more computation.

Text-to-Speech for Mandarin Chinese

Text-to-speech for Mandarin Chinese is an active area of research. Sproat *et al.* [12], Wu and Chen [15], Hwang *et al.* [7], Shih and Kochanski [11], and Hwang *et al.* [7] have all contributed to a better understanding of the unique aspects of Mandarin Chinese and how it is uniquely different to English. Mandarin Chinese is different from English in that it is a tonal language based on monosyllables. Each syllable can be phonetically decomposed into a consonant initial and a vowel final. There are five basic tones in Mandarin Chinese, identified as tones 1 to 5, respectively. Tone 1 is the high-level tone, tone 2 is the mid-rising tone, tone 3 is the mid-falling–rising tone, tone 4 is the high-falling tone and tone 5 is the neutral tone. A syllable is usually used as the synthesis unit in a Mandarin Chinese TTS system because it is the basic rhythmical pronunciation unit. In addition, from the viewpoint of Mandarin Chinese phonology, the total number of phonologically allowed syllables in Mandarin speech is only about 1300, which are the set of all legal combinations of 411 base-syllables and 5 tones. Furthermore, the factors that affect prosodic properties of a syllable are tone combination, word length, POS of the word, and word position in a phrase.

Mandarin TTS systems' differences from English TTS systems lay mainly within the prosody information generation. In general English TTS systems lacking prosody information will produce output speech that is understandable; although it will likely sound unnatural or robotic. However, since Chinese is a tonal language, without proper prosody tags the output speech from a Mandarin TTS system may not be understandable.

There are two basic types of system structures used for Mandarin TTS systems: a conventional three-module structure and a two-module structure similar to English TTS systems.

Conventional Three-Module Mandarin TTS Structure

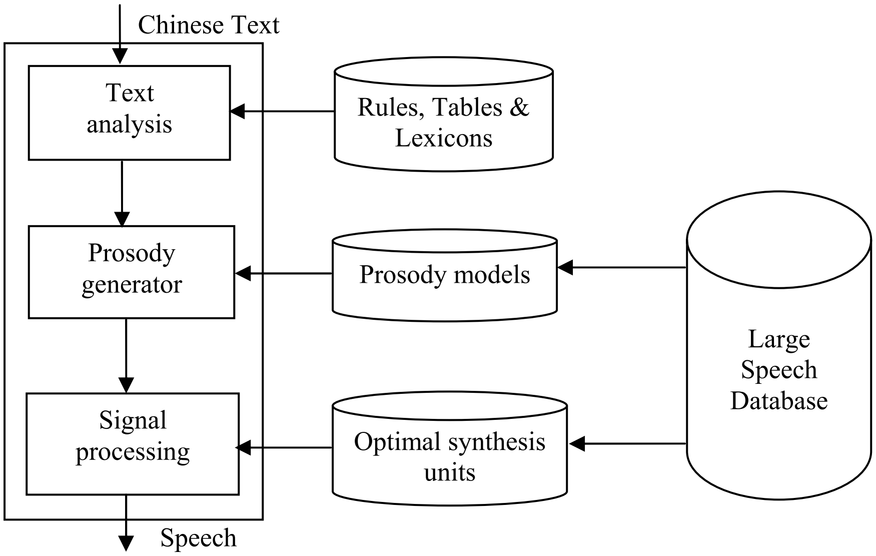

Figure 3. Overall View of Three-Module Mandarin TTS System [15]

Figure 3 shows a conventional three-module structure Mandarin TTS system. In this system, a large reading text corpus database is usually employed to generate prosody models and optimal synthesis units. Five procedures can be used to select a set of synthesis units from the speech database: pitch-period detection and smoothing, speech segment filtering, spectral feature extraction, unit selection, and manual examination. In addition, some kind of cost function needs to be used to minimize inter and intra- syllable distortion. Unlike English TTS systems, the Mandarin TTS system is functionally composed into three main parts: text analysis, prosodic information generation and signal processing. Notice that the Mandarin TTS system takes the prosody generator out of the text analysis module and treats the prosody generator as a module of equivalent importance as text analysis and signal processing, since prosody information is of critical significance to tonal languages like Chinese. Input Chinese text in the form of a character sequence is first segmented and tagged in the text analysis to obtain the best word sequence and the best part-of-speech (POS) sequence. The prosody generator then derives the prosody information from the linguistic features and sends them to signal processing to perform prosody modification [16]. Lastly, a signal processing algorithm is employed to modify optimized synthesis units to generate the output synthetic speech. POSOLA is used by most of Mandarin TTS system today to perform the signal processing.

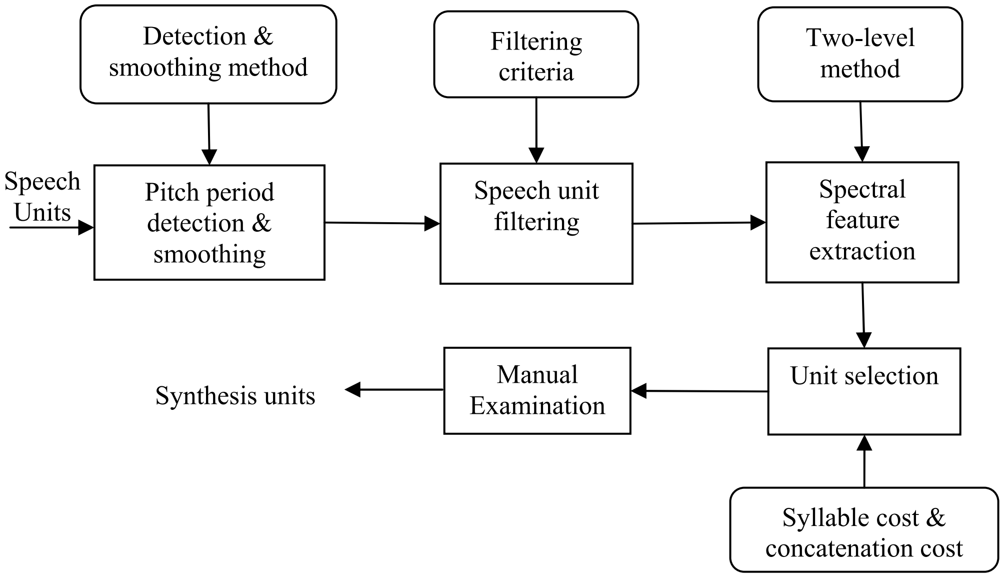

Figure 4. Block Diagram of the Synthesis Unit Selection by Wu and Chen [15]

Text Analysis

The text analysis component of a Mandarin TTS system first takes the Chinese input text string, identifies Chinese characters, numbers, and punctuations within body of text, and then finds the lexicon word. The text analysis component then extracts the syntactic structure to deal with homograph disambiguation. Unlike English, Mandarin TTS systems take either Big5 for traditional Chinese or GB for simplified Chinese as input code. In addition, there are no white spaces between Chinese lexical words. Therefore, Sproat [13] suggests finding the word boundaries within Chinese sentences is required to "reconstruct" rather than delimit white spaces. Different approaches have been proposed for exploring the Chinese lexicon words including dictionary lookup, statistical models, and stochastic finite-state word-segmentation algorithm [12].

Wu and Chen [15] proposed to use a Mandarin Chinese word dictionary of 80,000 entries for word identification and word pronunciation. However, they used rules for identifying homonym disambiguation and a phonetic table to reference phonetic transcription for each character.

Shi *et al.* [10] and Hwang, Chen & Wang [7], see figure 5, proposed grouping and tagging techniques to determine lexicon words based on statistical models. The statistic model is based on the frequency of the words in the vocabulary. Furthermore, Shi enhanced the approach by using dynamic programming (DP) to find an optimum path for segmentations.

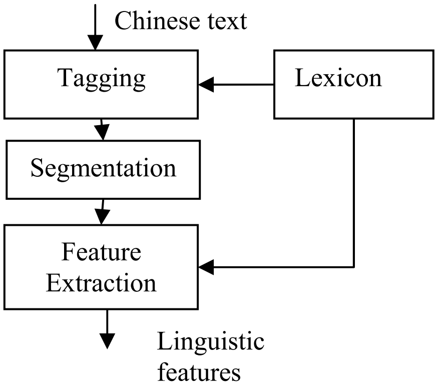

Figure 5. Block Diagram of Text Analysis by Hwang, Chen and Wang [7]

Identifying lexicon words is important to Mandarin TTS systems. However, Qian, Chu, & Peng [8] claimed that they have used a statistical model to find prosodic words directly without first determine the lexicon words while still maintaining good results.

Prosody Generator

Prosody generation of Mandarin is complicated. As identified by Chu *et al.* [3], the prosody module is responsible for calculating an "appropriate" set of prosodic contours, such as fundamental frequency, duration and amplitude. Hwan, Chen, & Wang [7] identify many factors that may affect the generation of prosodic information, such as linguistic features of all levels of syntactical structure, the semantics, the speaking habit and the emotional status of the speaker, as well as the pronunciation environment. Chu, Peng, & Chang [2] summarize that prosody generators for Mandarin TTS systems can be a set of rules, a word-prosody template tree, a statistical model or a neural network trained from a speech corpus.

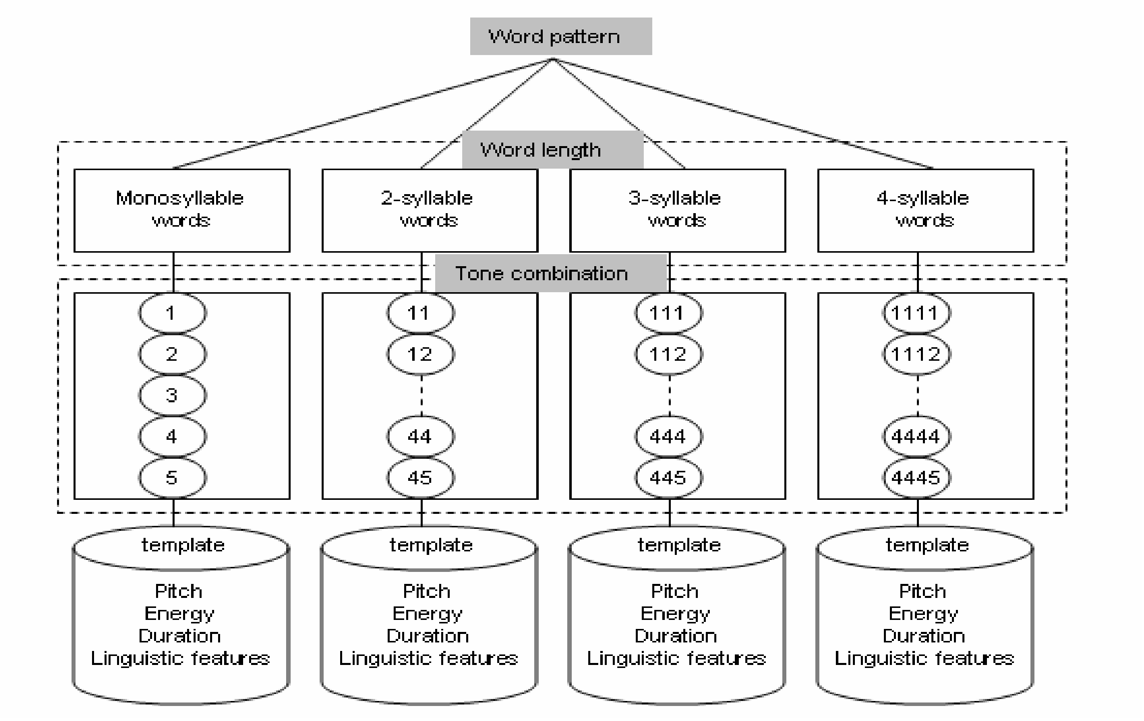

Figure 6. Structure of the Word-Prosody Template Tree [15]

Wu & Chen [15] introduced a word-prosody template tree (figure 6) to record the relationship between the linguistic features and the word-prosody templates in the speech database. Each word-prosody template contains the syllable duration, average energy, and pitch contour of a word. The pitch contour in the word-prosody template records sandhi for the syllables in the word. For each word in a sentence/phrase, word length is first determined and used to traverse the template tree. Tone combination is then used to retrieve the stored templates. Finally, a sentence intonation module and a template selection module are proposed to select the target prosodic templates.

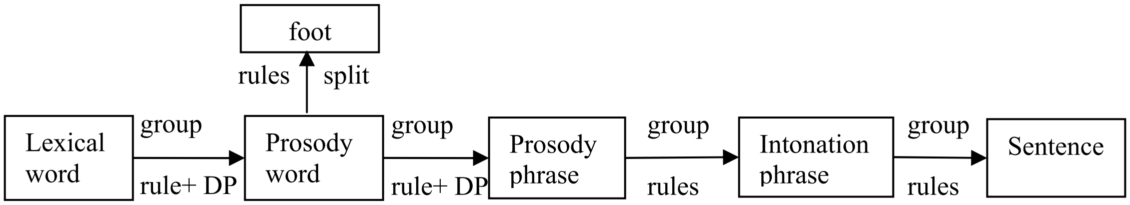

Figure 7. Processing of Generating Prosody Structure by Shi *et al*. [10]

The IBM Mandarin TTS Corpus, developed and documented by Shi *et* al. [10], uses a statistical method that predicts the prosody structure by combining dynamic program methods with the rules. Their prosody structure (see figure 7) is generated by manually labeling a corpus, since there is a tight relationship between the syntactic structure and prosodic structure. The statistical features are based on the POS and length of the lexical word. Furthermore, the statistical models for generating prosody structure are trained based on the corpus. After analyzing the prosody structure for an input text, the position of a lexical word in the prosody structure and its context information can be easily extracted. Their experimental results reached 91.2% accuracy for predicting prosodic structure.

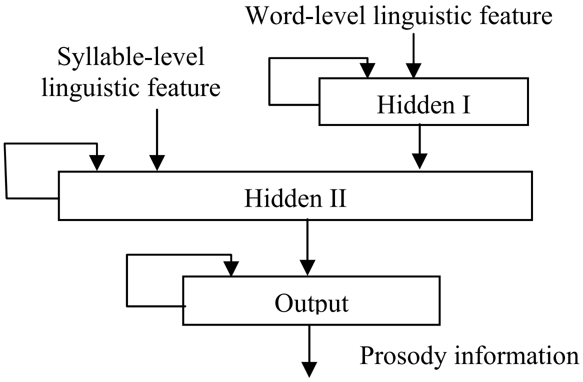
Word-level linguistic feature
Syllable-level linguistic feature
Hidden I
Hidden II
Output
Prosody information

Figure 8. Block Diagram of RNN by Hwang, Chen & Wang [7]

Hwang, Chen & Wang [7] employed a recurrent neural network (RNN) to generate prosodic information including pitch contour, energy level, initial duration, final duration of syllable and inter-syllable pause duration. The RNN has four-layer structure with two hidden layers to simulate human's prosody pronunciation mechanism for generating all prosodic information required in their system. Figure 8 shows the block diagram of the RNN. It can be functionally divided into two parts: The first part consists of the input layer and the first hidden layer and is responsible for finding the prosodic phrase structure by using the word-level linguistic features. It operates on a clock synchronized with words to generate outputs representing the phonologic state of the prosodic phrase structure at the current word. The second part consists of the second hidden layer and the output layer. It operates on a clock synchronized with syllables to generate the prosodic information by using the prosodic state fed-in from the first part and the input linguistic features. The RNN prosody synthesizer can automatically learn many prosody phonologic rules of human beings such as the sandhi rule of Tone 3 change. It can therefore be used to generate proper prosodic information required for synthesizing natural and fluent speech. In addition, they also employed a waveform table to provide the basic primitive wave forms of the synthetic speech. All waveform templates of syllables are selected from the speech database semi-automatically. Each selected waveform template is further processed to normalize its energy contour to the average of all energy contours of the same base syllables in the speech database before being stored in the wave table. A segmental k-mean algorithm is employed to obtain the average energy contours of all base-syllables.

Signal Processing

Hwang, Chen, & Wang [7] indicate that recently the PSOLA algorithm has been widely adopted by Mandarin TTS systems for the prosody modification phase. The benefit of the PSOLA algorithms is it can generate high-quality synthetic speech with low computational complexity. Chu *et al.* [3] indicate it is applied to optimized synthesis units to guarantee that the prosodic features of synthetic speech meet the predicted target values. Hwang clarifies that the modifications include changing the pitch contour for each syllable, adjusting the durations of the initial consonant and the final vowel of each syllable, scaling the energy level of syllable, and setting the inter-syllable pause duration. Finally, the output synthetic speech is generated.

Two-Module Mandarin TTS Structure

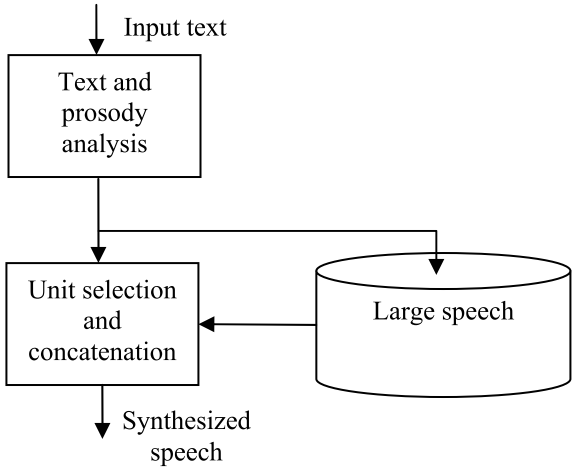

Figure 9. Two-Module TTS Structure by Chu *et al.* [3]

Figure 9 is the two-module Mandarin TTS system proposed by Microsoft research group members Chu *et al.* [2, 3]. This TTS structure more closely resembles the general organization of English TTS systems than Mandarin TTS systems. It bypasses the prosody generation layer that predicts numerical prosodic parameters for synthetic speech. Chu *et al*. believe that although traditional three-module Mandarin TTS systems have the advantages of flexibility in controlling of prosody, these systems often suffer from significant quality decreasing in timbre. Instead, their system classifies many instances of each basic unit from a large speech corpus into categories by a CART in which the expectation of the weighted sum of square regression error of prosodic features is used as a splitting criterion. Better prosody can be achieved by keeping slender diversity in prosodic features of instances belonging to the same class. Furthermore, Chen *et al*. present a multi-tier non-uniform unit selection method to make final decisions of choices under the condition of minimizing the concatenated cost of the synthesized utterance. The system can make the best decision on unit selection by minimizing the concatenated cost of a whole utterance. However, this approach is based on an ultimate assumption that having a very large speech corpus containing enough prosodic and spectral varieties for all synthetic units is practical. This assumption is valid under a constraint that the whole corpus retains the same speaking style. They claim that their model has a very natural and fluent speech according to informal listening tests.

Conclusion

This paper has briefly introduced different techniques used in each component of English TTS systems and summarized different approaches used in different Mandarin TTS systems with particular implementation examples. The paper identified the primary differences distinguishing Chinese TTS systems and English TTS systems, which mainly lie within prosody information generation.

Although TTS has come long way since it first invented, as suggested by Schroeter *et al.* [12], research still has a long way to go before delivering natural speech output for any input text with any intended emotions.

References

1. Bulyko, Ivan. (2002). *Flexible speech synthesis using weighted finite state transducers*. Ph.D Thesis. University of Washington.

2. Chu, Min., Peng, Hu., & Chang, Eric. (2001). *A concatenative Mandarin TTS system without prosody model and prosody modification*. Proceedings of 4th ISCA workshop on speech synthesis. Scotland.

3. Chu, Min., Peng, Hu., Yang, Hong-yun., & Chang, Eric. (2001). *Selecting non-uniform units from a very large corpus for concationative speech synthesizer*. Proceeding of ICASSP2001, Salt Lake City.

4. Dutoit, Thierry. (1997). *An Introduction to Text-To-Speech Synthesis*. Boston: Kluwer Academic Publishers.

5. Edgington, M., Lowry, A., Jackson, P., Breen, A.P., & Minnis, S. (1996). *Overview of current text-to-speech techniques: Part I – text and linguistic analysis*. BT Technology Journal. Vol.14 No.1.

6. Edgington, M., Lowry, A., Jackson, P., Breen, A.P., & Minnis, S. (1996). *Overview of current text-to-speech techniques: Part II –prosody and speech generation*. BT Technology Journal. Vol.14 No.1.

7. Hwang, Shaw-Hwa., Chen, Sin-Horng., & Wang, Yih-Ru. (1996). *A Mandarin Text-To-Speech System*. Proceedings of ICSLP '96. Philadelphia, PA. Vol. 3, 1421 - 1424.

8. Qian, Yao., Chu, Min., & Hu, Peng. (2001). *Segmenting unrestricted Chinese text into prosodic words instead of lexical words*. Proceeding of ICASSP2001, Salt Lake City.

9. Schroeter, J., Conkie, A., Syrdal, A., Beutnagel, M., Jilka, M., Strom, V., Kim, J.K., Kang, H.G., & Kapilow, D. (2002). *A perspective on the next challenges for TTS research*. IEEE 2002 Workshop on Speech Synthesis. 11-13 September 2002. Santa Monica, CA.

10. Shi, Qin., Ma, XiHun., Zhu, WeiBin., Zhang, Wei., & Shen, LiQin. (2002). *Statistic Prosody Structure Prediction*. IEEE 2002 Workshop on Speech Synthesis. 11-13 September 2002. Santa Monica, CA.

11. Shih, Chilin., & Kochansi, Greg. (2000). *Chinese tone modeling with Stem-ML*. Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, China.

12. Sproat, Richard., Shih, Chilin., Gale, William., & Chang, Nancy. (1996). *A Stochastic Finite-State Word-Segmentation Algorithm for Chinese*. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. pp. 66-73.

13. Sproat, Richard. (1996). *Multilingual text analysis for text-to-speech synthesis*. Journal of Natural Language Engineering, 2(4):369--380

14. Wang, Bei., Zheng, Bo., Lu, Shinan., Cao, Jianfen., & Yang, Yufang. (2000). *The Pitch Movement of Word Stress in Chinese*. Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, China.

15. Wu, Chung-Hsien., & Chen, Jau-Hung. (2001). *Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis*. *Speech Comuunication, 35,* 219-237.

16. Xu, Jun., Guan, Cuntai., & Li, Haizhou. (2002). *An objective measure for assessment of a corpus-based text-to-speech system*. IEEE 2002 Workshop on Speech Synthesis. 11-13 September 2002. Santa Monica, CA.