

# MODULAR DESIGN FOR MANDARIN TEXT-TO-SPEECH SYNTHESIS

Jilei Tian, Jani Nurminen, Feng Ding and Imre Kiss

Multimedia Technologies Laboratory  
Nokia Research Center  
Finland

{ Jilei.tian, jani.k.nurminen, feng.f.ding, imre.kiss }@nokia.com

## Abstract

In the European Union funded project Technology and Corpora for Speech-to-Speech Translation (TC-STAR) [3], we have developed a modular concatenative TTS system for Mandarin Chinese. A common architecture has been introduced based on well-defined modules and interfaces. Three main modules, text processing, prosody processing and acoustic synthesis modules, are used following a commonly employed approach. A large number of applications can benefit from using the high-quality TTS design. Preliminary evaluations have shown promising performance and flexibility.

## 1. Introduction

The term speech synthesis refers to the artificial production of human speech. A system used for this purpose is termed a speech synthesizer. It takes as input a sequence of words and converts them to speech. Speech synthesis systems are often called text-to-speech (TTS) systems in reference to their ability to convert text into speech. The ultimate goal is to have the best human-like speech quality from the overall system. TTS system can be applied whenever a computerized application needs to communicate with a human user. Some typical use cases for the TTS technology are:

- telecommunication services to access textual information by a phone;
- enhanced mobile device user interface to support voice prompt, confirmative response in voice dialing, SMS/Email reader, guidance or instruction;
- customer support dialog systems, e.g. help desks;
- language education system;
- speech-to-speech translation;
- talking books and gaming.

Though recorded material still provides the highest quality, recordings are often impractical due to cost or time constraints.

Before modern computing technique was invented, scientists tried to build machines to create human speech. In the 1930s, Bell lab developed the VOCODER, a keyboard-operated electronic speech synthesizer that was clearly intelligible. Later Dudley refined this device into the VODER exhibited at the New York World's Fair. The first computer-based speech synthesis systems were developed in the 1950s and the first complete TTS system was completed in 1968. Since then, there have been many advances in the speech synthesis field [1].

TTS systems perform a range of processes, from text normalization, pronunciation, and several aspects on symbolic and acoustic prosody, finally generating speech at the last step. There are two main alternative technologies used for speech synthesis: concatenative synthesis and formant synthesis. Formant synthesis does not require any pre-recorded speech samples. Instead, the synthetic speech is created using an acoustic model. Parameters such as pitch, voicing, and noise levels are varied over time to

synthesize speech waveforms. This method is also called parametric based synthesis. Concatenative synthesis, on the other hand, is based on the concatenation of segments of pre-recorded speech. Generally, concatenative synthesis gives the most natural sounding synthesized speech.

In general, concatenative TTS can be further classified into two sub-categories, traditional single instance diphone synthesis and unit selection based synthesis. The diphone synthesis approach takes use of only one representative acoustic unit for each diphone; and the pitch, duration and amplitude of the diphones are modified according to some prosody prediction model. Nevertheless, prosody prediction is not error-free and signal processing methods for carrying out the modification introduce speech distortion. Unit selection based synthesis uses a larger speech corpus selecting the best matching units based on a pre-defined distance measure. Prosodic modifications and the related digital signal processing techniques are not necessarily required because the selected units already contain appropriate prosodic properties. Unit selection gives the greatest naturalness due to the fact that it does not apply a large amount of signal processing to the pre-recorded speech, although some systems may perform some processing at the boundaries of the concatenated segments to smooth the waveform.

Our TTS system was implemented based on the unit selection approach. The remainder of the paper is organized as follows. We first introduce the modular design of our TTS system in Section 2. Then, the individual modules are briefly described in Sections 3, 4 and 5. In Section 6, the interface design for the modules is presented. Finally, evaluation result is presented in Section 7, and conclusions are drawn in section 8.

## 2. Architecture design

Our objective was to design architecture for TTS system based on well-defined modules and interfaces. By this approach the modules are interchangeable and are evaluated using a common set of evaluation criteria within TC-STAR. It makes the system design more transparent and easier to maintain.

A TTS system is composed of two parts: the front-end and the back-end. Broadly speaking, the front-end takes its input in the form of text and outputs a symbolic linguistic

representation. The back-end takes the symbolic linguistic representation as its input and outputs the synthesized speech waveform. The back-end can further be divided into prosodic processing and acoustic synthesis parts to emphasize the critical role of prosodic information. From this perspective, the TTS system is designed to consist of three main modules as shown in Figure 1: text processing, prosodic processing and acoustic synthesis.

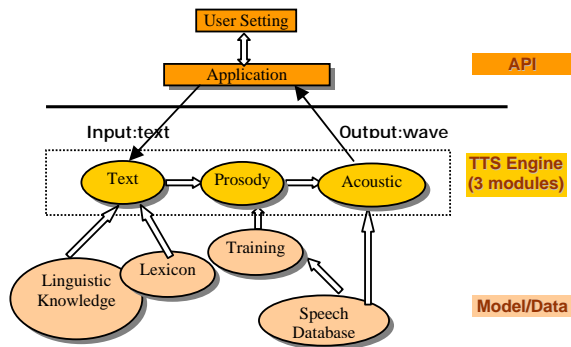


Figure 1. Block diagram of TTS system.

The text processing module has two major tasks. First it takes the raw text and converts entities such as numbers and abbreviations into their written-out word equivalents. This process is often called text normalization, or tokenization. Then it assigns Part-of-speech (POS) tags and phonetic transcriptions to each word. Furthermore, it divides and marks the text into various prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme (TTP) conversion. The prosodic processing module takes as its input the symbolic linguistic representation parsed in the text processing module and determines the prosodic features such as silences, duration, energy and fundamental frequency of the acoustic units. The combination of the phonetic transcriptions and the prosodic information make up the symbolic linguistic representation input to the acoustic synthesis module that converts it into actual sound output.

The text input to the TTS system is formatted according to the speech synthesis markup language (SSML). SSML defines tags to control the text structure and to give information about the desired prosody and style. Since all the tags defined in SSML are optional, plain text can easily be transformed into an SSML document by embracing it between the < speak > and < /speak > SSML definition tags. Extensible markup language (XML) is introduced to describe the interface between the text processing and the prosody processing modules, and between the prosody processing and acoustic synthesis modules. Since each module performs a complementary task, only one Document Type Definition (DTD) is necessary. Each module adds information to the corresponding part of the XML document while maintaining the information previously added by any other module. The cascaded process is depicted in Figure 2.

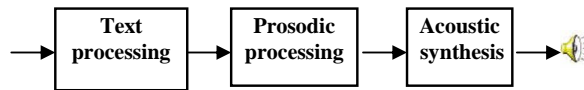


Figure 2. Modular TTS design and cascade processing.

### 3. Text Processing

As the text processing is at least partially language specific, the peculiarities of the Chinese language had to be taken into account in the design of the text processing module [2]. Mandarin Chinese is a tonal syllabic language in which the tone of a syllable is described by its pitch contour. There are five different tones and each tonal syllable is composed of one initial (21, consonant part) and one final (35, vowel part). The Chinese Phonetic System (CPS) is the standard Romanization scheme used in Mainland China. It consists of five parts: the alphabet, the initials, the finals, the tone marks and a syllable-dividing mark. Most people in mainland China speak Mandarin. Most people in Hong Kong speak Cantonese. Most people in Taiwan speak Mandarin and Taiwanese. Written Chinese characters, known as Hanzi in Chinese, generally have the same meaning in all dialects. The Chinese characters include the simplified characters used in mainland China and Singapore as well as the traditional characters used in Taiwan, Hong Kong and most other Chinese-speaking communities. Each character corresponds to one syllable. There are no spaces or other word boundary markers between the words in Chinese text, placing special requirements for the text processing module.

As the phonetic alphabet, we use the SAMPA phonetic alphabet with syllable boundary marker, stress marker and tone markers for tonal languages. Part-of-speech (POS) coding is partly based on the formal definition specified by the LC-STAR project, among others: NOM (name), ADJ (adjective), ADV (adverb), PRE (preposition), DET (determinant).

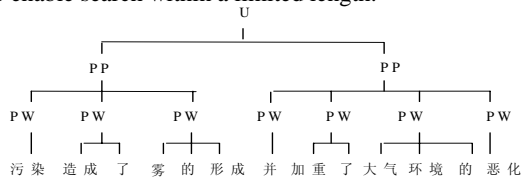
In general, symbolic pre-processing module performs the tokenization, POS tagging and phonetic transcription of the input text. The phonetic transcription information is processed for each word in the way that the word is spoken in isolation. For the Mandarin TTS system, it is aiming to convert digit, phone number, control symbol, punctuation, non-Chinese-character to corresponding pronounceable Mandarin words, word segmentation, POS tagging, word/phrase prediction, break prediction, homograph disambiguity of phonetic transcription of the input text. PinYin (initial + final) is used for each monosyllable character. The backward (right-to-left) longest matching algorithm is used for word segmentation. POS is tagged based on N-gram model using dynamic programming to find the best POS sequence.

The prosodic structure, as demonstrated in Figure 3, is predicted by

$$P(J_i = status | POS_i, POS_{i+1}) \cdot P(J_i = status | nLen_i, nLen_{i+1}) \tag{1}$$

where POS and nLen denote the part of speech and the length of the phrase on both sides of the boundary to be predicted. Prosodic breaks are predicted using decisions based on the POS and character length N-gram

probabilities. Additional length constraints are introduced to enable search within a limited length.

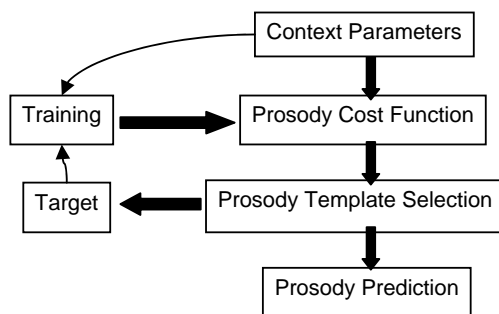


**Figure 3.** Utterance (U) prosodic tree structure consisting of prosodic word (PW) and prosodic phrase (PP).

#### 4. Prosodic Processing

Prosody is an inherent supra-segmental feature of human's speech that can express attitude, emotion, intent and attention, etc. The prosodic processing module used in the system presented in this paper utilizes a technique based on syllabic templates. This approach can be considered particularly suitable for Mandarin Chinese due to the fact that it is a syllabic tonal language. The prosodic templates are chosen during the training of the system in such a manner that sufficient coverage is achieved on both tonal syllables and on the contexts. Each entry in the prosodic template inventory contains information on the context in which the syllable occurred and information on the syllable itself, including the pitch contour and the duration of that particular instance of the syllable. The storage of the prosodic templates naturally increases the memory usage of the system but we have achieved significant memory savings using e.g. the duration modeling technique presented in [5].

The process of syllabic prosody template selection is depicted in Figure 4. First, context parameters are retrieved from the output of the text processing module for each syllable. Then, a cost function is used to measure the distance between the context parameters extracted from the text and the context parameters of the syllable data stored in the prosody model. The prosodic template offering the best matching context is selected from the template inventory stored in the prosody model data. The prosodic data corresponding to the retrieved template is used as the predicted prosody.



**Figure 4.** Syllabic prosody template selection.

For the representation of the syllabic pitch contour data, we have proposed the concept of syllable based eigenpitch [4]. We have shown that the dimension of the pitch vectors can be efficiently reduced in the eigen space while

minimizing the energy loss. The eigenpitch representation also preserves well the tonal features and offers a high classification capability.

#### 5. Acoustic synthesis

In concatenative acoustic synthesis, unit selection plays a critical role in reaching high quality of synthetic speech. In our system, the unit selection operates on syllable units and is based on the outputs of the text processing and the prosody modules. The information on the syllables and their contexts, as derived by the text processing module for a given input text, are used for limiting the unit selection to a set of potential unit candidates. Then, both the context information and the prosodic features predicted by the prosody module are used for selecting the best-matching acoustic unit instance of the syllable from the acoustic database inventory. The usage of the predicted prosody in the unit selection is assumed to enhance the quality of the output. However, the predicted prosody is only used as one of the metrics to avoid or limit severe quality degradations in cases in which the prosody prediction fails to produce correct prosodic information.

It would also be possible to incorporate an additional metric for enhancing the smoothness at the unit boundaries but in the particular case of syllable based Mandarin synthesis this is not very crucial. The reason for this is the fact that there can often be natural discontinuities at the syllable boundaries also in natural speech. Due to the same reason, the signal processing at unit boundaries can be kept reasonably simple.

#### 6. Modular interface design

As mentioned in Section 2, the text input to the TTS system is formatted into the SSML format and XML based interfaces are defined between the three main modules. Since each module performs a complementary task, only one DTD is necessary. The text processing module is dealing with tokenization, POS tagging and phonetic transcription of the input text. A token is an individually distinguishable element of the input text. We have introduced tokens in order to handle special input elements that can potentially be handled differently by the text processing module. Each token is divided into words in such a way that each word is associated with transcription and POS tag. The phonetic transcription information will be coded for each word in the way the word is spoken in isolation. The SAMPA phonetic alphabet is adopted with syllable boundary marker, stress marker, and tone markers for tonal languages. The LC-STAR POS tagset is also taken into consideration. The prosodic processing module fills in the information regarding the syllabic structure of the words, corresponding to phonemes, and their associated pauses, accents, fundamental frequency and voice-quality attributes. Each phoneme has a reference fundamental frequency (pitch), duration and intensity.

To make the modular XML interfaces more useful with a wider language support, non-English languages, especially Chinese should also be taken into account [6]. The peculiarities of the Chinese language must be taken into account when designing the XML interface. For Chinese, it is natural to use tonal syllables as the basic unit

of a TTS system. Word segmentation is a very crucial issue in Chinese since the written form does not contain word boundaries. Thus `<word>` element is defined to enhance the word segmentation in cases when the automatic word segmentation does not work reliably or the user forces the system to take a certain word segment. Its attribute is the segmented word. Inside `<word>` element, `<pos>` is used for determining the pronunciation of given word in the cases when the POS tagging is not reliable or the user forces the system to use a certain POS tag. `<break>` can be used for defining the break strength at the boundary, such as character boundary, word boundary, prosodic phrase boundary, sentence boundary, etc.

For Chinese, the pitch contours play a very important role in rendering TTS speech. The same phoneme sequence or the same baseform syllable with a different tone usually leads to completely different meanings. Therefore, it is recommended to enhance the descriptions on prosodic features, particularly on pitch. We describe the prosody features in (time, value) format. This approach gives the possibility to cover any prosodic needs. The element `<syllable>` is introduced to define the given character. The elements `<frequency>` and `<energy>` are taken into use to describe prosodic features, pitch and volume, in the (time, value) format in order to provide a better representation capability for prosodic features.

Figure 5 shows an example case demonstrating the XML interface used for the Mandarin TTS system. Original input is highlighted by *italic*, the text processing and prosodic processing module results are added and highlighted as **bold** and normal font, respectively.

```
<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xml:lang="cn">
  <s>
    <TOKEN token=="下午">
      <word word="下午">
        <pos>
          <NOUN />
        </pos>
        <syllable syl="下午">
          <frequency>
            <pair time="0" value="380" />
            <pair time="80" value="363" />
            <pair time="160" value="340" />
            <pair time="240" value="301" />
          </frequency>
          <energy>
            <pair time="267" value="74" />
          </energy>
          <break strength="none" />
        </syllable>
        <syllable syl="午">
          <frequency>
            <pair time="0" value="290" />
            <pair time="54" value="285" />
            <pair time="108" value="285" />
            <pair time="162" value="290" />
          </frequency>
          <energy>
            <pair time="181" value="71" />
          </energy>
          <break strength="weak" />
        </syllable>
      </word>
    </TOKEN>
  </s>
</speak>
```

```
</syllable>
</word>
</TOKEN>
</s>
</speak>
```

**Figure 5.** Example of XML interface used for Mandarin TTS system. Original input, text processing and prosodic processing module outputs are highlighted to be italic, bold and normal fonts, respectively.

## 7. Evaluation

The Mandarin Chinese TTS system was evaluated in the TC-STAR project. The newly recorded TC-STAR voice (~14h) was used for building up the Mandarin TTS system including the lexicon (~110K entries), the text processing models, the prosodic models and the acoustic inventory. The testing text contained 12 paragraphs taken from the 863 Chinese national project and they were given in the SSML format. The mean opinion score (MOS) test was arranged by an independent evaluation agency. The MOS test results are given in Table 1. It shows the comparable results for the baseline speaker and the TTS system. Clearly, the quality of TTS speech is far below quality of human speech indicating some improvement needs, as summarized below:

- System tuning on TTS data: Since the TC\_STAR data was collected shortly before evaluation, there was no time to tune the system properly. Better tuning and improvements on the annotations will bring quality gains. This is certain since we have achieved clearly better quality using a better tuned internal database;
- The text processing module could be improved by taking more linguistic information into use, e.g. intonational boundary, etc. It is beneficial for "long" sentences containing complex syntax structure.
- TTS system can be made more robust against annotation errors. E.g. in the pitch model training, the contextual and linguistic knowledge could be used to protect against annotation errors.

Evaluation items:	MOS (speaker)	MOS (TTS)
Overall quality	4.44	2.77
Listening effort	4.34	3.03
Pronunciation	4.59	2.65
Comprehension	4.49	3.81
Articulation	4.63	3.22
Speaking rate	4.46	3.80
Naturalness	4.09	2.69
Ease listening	3.93	2.52
Pleasantness	3.98	2.43
Audio flow	4.35	2.61

**Table 1.** Evaluation results given in MOS scores.

## 8. Conclusions

In this paper, we focused on system-level architecture design and applied a modular structure to develop a Mandarin Chinese TTS system. A module-wise XML interface design was also proposed for the Mandarin Chinese TTS system. The TTS system consists of three main modules: text processing, prosodic processing and acoustic synthesis modules. The main algorithms used in the modules were briefly outlined. Then, XML based

interfaces between the module pairs were proposed especially for the Chinese language.

We have carried out experiments with Mandarin Chinese speech databases. The experiments show that it is possible to exploit the proposed structure and interfaces. The resulting system architecture is very flexible. The system described in this paper was also evaluated in the TC-STAR/ECESS public evaluation. It can be concluded that this paper provides a very promising method that can be extended in several ways and used with other languages.

Even though the proposed TTS system design provides good performance and flexible development platform, several improvements can still be made to achieve truly high-quality TTS synthesis and enhanced portability in the future. As shown in the paper, the development of a successful TTS system is a highly time-consuming and cost expensive task that requires taking robustness issues into consideration to protect against annotation errors. For example, it was noticed that imperfect annotations caused the trained models to perform in a non-optimal way. Instead, linguistic rules could be applied to make the system more tolerant on obvious annotation errors. Since the reliable annotations require large amounts of manual work, better robustness could speed up the development process and improve the performance.

Currently, the symbolic linguistic processing works reasonably well as long as the sentences are medium-sized and the sentence structures are not very complex. The system could utilize richer linguistic information to enhance the linguistic analysis to make the synthesis speeches more natural and intelligible. For example, the intonational phrase break can be incorporated in the prosodic structure as shown in Figure 3.

## 9. References

- [1] Dutoid, T., *An introduction to text-to-speech synthesis*. Kluwer, Dordrecht, 2001
- [2] Li, A., “Chinese prosody and prosodic labelling of spontaneous speech”, In *Proceedings of International Workshop on Speech Prosody*, Aix-en-Provence, France, 2002.
- [3] TC-STAR project website: <http://www.tc-star.org/>
- [4] Tian, J. and Nurminen, J., “On analysis of eigenpitch in Mandarin Chinese”, In *Proceedings of 4<sup>th</sup> International Symposium on Chinese Spoken Language Processing*, HongKong, China, 2004
- [5] Tian, J., Nurminen, J. and Kiss, I., “Duration modeling and memory optimizations in a Mandarin TTS system”, In *Proceedings of European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005
- [6] Tian, J., Wang, X. and Nurminen, J., “SSML extensions aimed to improve Asian language TTS rendering”, In *W3C Workshop on Internationalizing the Speech Synthesis Markup Language*. Beijing, China, 2005.