Modular Text-to-Speech Synthesis Evaluation for Mandarin Chinese

Jilei Tian, Jani Nurminen, and Imre Kiss
Multimedia Technologies Laboratory, Nokia Research Center
P.O. Box 100, FIN-33721 Tampere, FINLAND
{jilei.tian, jani.k.nurminen, imre.kiss}@nokia.com

Abstract. Proper evaluation can efficiently drive the development of text-to-speech (TTS) systems. The assessment is needed to determine how well a system or technique compares to others or how it compares with the previous version of the system. In order to obtain more useful feedback for the development, we do not only evaluate the whole system but also each module of the TTS system separately. Based on the evaluation results, new progress can be achieved on individual modules. Furthermore, the evaluation allows identifying the best techniques in the different modules and comparing different modules for which we have defined a common evaluation specification. To enable the evaluation, the Mandarin TTS system is separated into three modules: text, prosody and acoustic processing modules. Furthermore, common interfaces have been designed for the communication between the TTS modules. The objective and subjective metrics have been proposed for the module-wise evaluation. This paper also presents the evaluation results that our Mandarin TTS modules achieved in the evaluation carried out in the European project on technology and corpora for speech-to-speech translation (TC-STAR).

Keywords: Text-to-speech, speech synthesis, SSML, prosodic modeling, evaluation

1 Introduction

The term speech synthesis refers to the artificial production of human speech. The speech synthesis system, or a speech synthesizer, takes as its input a sequence of words and converts it 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 possible, even human-like, speech quality from the overall system. The TTS technology can be applied whenever a computerized application needs to communicate with a human user. Typical examples of such applications include e-mail and SMS reading, audio books and gaming.

There are two main technologies used for the generating synthetic speech waveforms: concatenative synthesis and formant synthesis [2][7]. The formant based synthesis approach does not use any human speech samples during the runtime processing. Instead, the output speech is created using an acoustic model. The synthesis parameters such as fundamental frequency, voicing, and noise levels are

varied over time to create a waveform of artificial speech. This method is also sometimes called rule-based synthesis. The concatenative synthesis approach, on the other hand, is based on the concatenation of segments of recorded speech. Usually, concatenative synthesis gives the most natural sounding synthesized speech. In general, the concatenation based TTS systems can be further classified into two sub-categories, traditional diphone synthesizers and unit selection synthesizers. The traditional diphone synthesis takes use of only one representative acoustic unit for each diphone but the pitch level, the duration and the amplitude level of each diphone can be modified according to some prosody prediction model. Nevertheless, the prosody prediction is not error-free and the signal processing methods needed for carrying out the prosodic modifications introduce audible distortions to the output speech. In contrast, the unit selection based synthesis approach uses a larger speech corpus and tries to select the best-matching units based on pre-defined distance measures. Prosodic modification and digital signal processing (DSP) techniques are not necessarily required because the selected units already contain appropriate prosodic properties. The current speech synthesis efforts, both in research and in applications, are dominated by methods based on the concatenation of spoken units as shown in Figure 1.

TTS systems perform a range of processes, from text normalization and pronunciation generation, to several aspects on symbolic and acoustic prosody, and finally to speech signal generation. The system design, the development processes and the corresponding evaluation methods should be jointly investigated to achieve the largest possible improvements in the synthesized speech quality, in terms of all aspects of intelligibility and naturalness. The conventional black box evaluation of the entire system, however, does not allow identifying detailed issues or problems in different parts or sub-parts of the system. In addition, this method does not allow for small teams of researchers whose specialty of research is related to one specific technique to participate in common evaluations. The evaluation of individual modules/tasks can certainly result in more useful information for improving the system performance and drive more valid conclusions about the performance of different algorithms. Thus, a glass box type of evaluation is brought up to assess separate TTS modules with well defined common input and output interfaces. This evaluation approach used also in the TC-STAR project allows meaningful comparisons and the pinpointing of problems in the algorithms used in the modules.

There are many processes involved in the whole TTS procedure. However, the number of modules needs to be limited when designing a general evaluation framework in order to limit the number of tests that have to be carried out in the evaluation. Consequently, a compromise was made by defining three broad modules: text processing, prosody processing and acoustic processing modules. The modules have been defined through their interfaces.

The remainder of the paper is organized as follows. We first very briefly describe our Mandarin TTS system that has been designed and reorganized for easy evaluation in Section 2. Then, the modular design and the XML based interfaces for Mandarin TTS are further discussed in Section 3. In Section 4, we summarize the experimental evaluation results achieved in the TC-STAR evaluations on the Mandarin TTS modules. Finally, conclusions are drawn in Section 5.

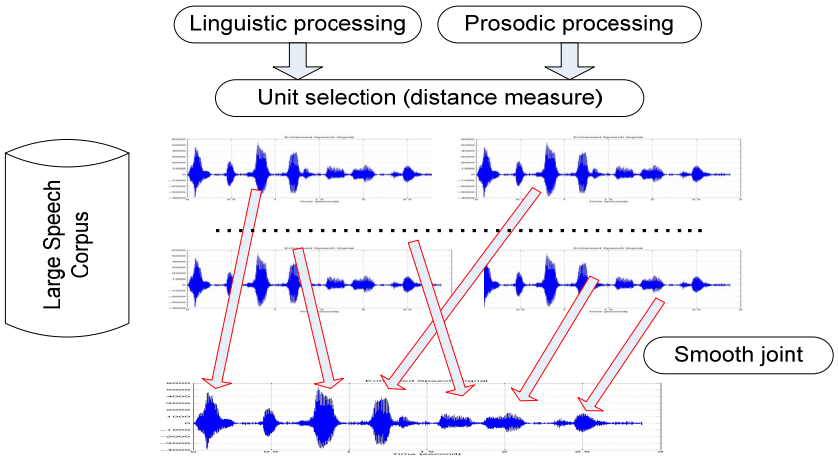

Figure 1. Diagram of acoustic unit selection procedure.

2 Mandarin TTS system

Our Mandarin TTS system is composed of three major common modules, text processing, prosodic processing and acoustic processing modules, as shown in Figure 2. The text processing module performs the text normalization aiming to convert digits, phone numbers, control symbols, punctuations, and non-Chinese-characters to their corresponding pronounceable Mandarin words. In addition, this module deals with the word segmentation, POS tagging, word/phrase prediction, the disambiguities in pronunciation, and the generation of the phonetic transcription of the input text. The pinyin (initial + final) representation is used for each monosyllable character. A backward (right-to-left) longest matching algorithm is used for word segmentation.

Given a word sequence W and a POS sequence T, we have

$$\begin{aligned} P(T \mid W) &= P(T) \cdot P(W \mid T) / P(W) \\ &\cong P(T) \cdot P(W \mid T) \\ &\cong \{P(t_2 \mid t_1) \cdot P(t_3 \mid t_2) \cdot \ldots\}\{P(w_1 \mid t_1) \cdot P(w_2 \mid t_2) \cdot \ldots\} \end{aligned} \quad \textbf{(1)}$$

The POS tagging is based on an N-gram model using dynamic programming to find the best POS sequence.

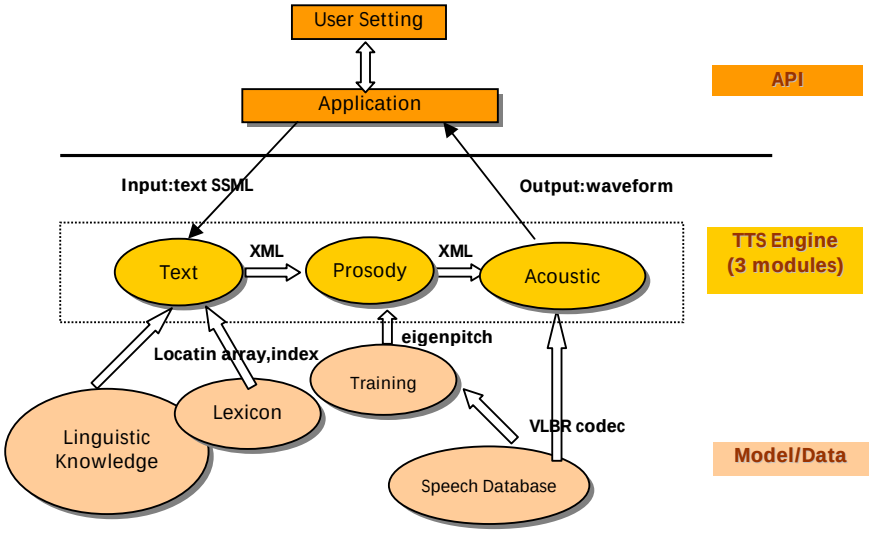

Figure 2. Block diagram of Mandarin TTS system.

The prosodic word shown in Figure 3 is predicted by

$$P(J_i = status \mid POS_i, POS_{i+1}) \cdot P(J_i = status \mid nLen_i, nLen_{i+1}) \quad \textbf{(2)}$$

where POS and nLen are the part of speech and the length of the phrase on both side of the boundary to be predicted. The prosodic breaks are predicted by decisions made based on the following N-gram probability (maximum) and a length constraint (search within a limited length).

$$P(J = break_. \mid T) \quad \textbf{(3)}$$

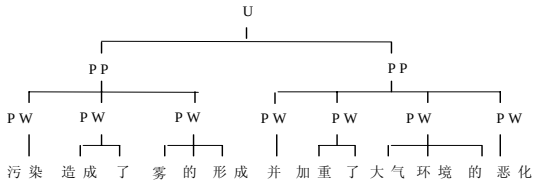

Figure 3. Utterance (U) prosodic tree structure consisting of prosodic word (PW) and prosodic phrase (PP).

The overall prosody is an inherent supra-segmental feature of human's speech to express attitude, emotion, intent and attention, etc [1][4]. In our prosody prediction, we utilize a syllabic prosody template selection approach that is described in Figure 4. 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. 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.

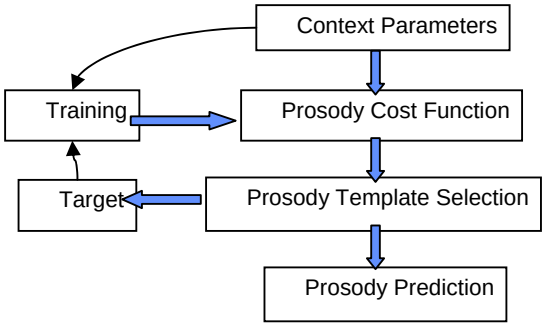

Figure 4. Syllabic prosodic template selection.

In our system, a novel data-driven syllable-based eigenpitch approach is proposed for the modeling of pitch contours [3][8]. The eigenpitch method can reduce the pitch vector dimension in the eigen space while minimizing the energy loss and preserving tonal features and high classification capability. Moreover, duration modeling [9] is also applied using a probability model for approximating the durations.

In the concatenative acoustic processing module, unit selection plays a critical role in reaching high-quality synthetic speech. In our system, the unit selection is based on the output of the prosodic model. The template based prosodic model for syllables includes context information, pitch contours and duration information of the syllable instances. In the synthesis phase, for a given text, the context features of the syllable are extracted from the text through text processing. Using the distance between the context features taken from the text and the context features pre-trained and stored in the prosodic model, a target pitch contour and the duration of the syllable instance are selected so that the defined distance is minimized. Then, the selected pitch contour and duration information are used for selecting the best acoustic unit instance of the syllable from the database inventory as shown in Figure 1.

3 Modular design and interfaces

The text input to the TTS system is formatted into a speech synthesis markup language (SSML). SSML defines tags to control the text structure and to give information about the desired features. Since all the tags defined in SSML are optional, plain text can easily be transformed into an SSML document by embracing it between <speak> and </speak> SSML definition tags.

Extensible Markup Language (XML) based interfaces are defined between three 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. Figure 5 has shown an example of modular XML based interface in Mandarin TTS system.

XML modular interface:

```
<?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 ">
    <frequency>
     <pair time="240" value="301" />
   </frequency>
   <energy>
            <pair time="267" value="74" />
          </energy>
          <break strength="none" />
          </syllable>
        .....
      </speak>
```

Figure 5. XML interface, text processing (blue) and prosodic info (red).

Chinese language specific issues must be taken into account when designing the XML interface [5]. For Chinese, it is natural to use tonal syllables as the basic unit of a TTS system. The word segmentation is a very crucial issue with Chinese text that does not have word boundaries. Thus, <word> element is defined to enhance the automatic word segmentation, and even in the cases where the automatic word segmentation does not work or the user wants to force the system to take a certain word segment. Its attribute is the segmented word. Inside the <word> element, <pos> can be used for facilitating the determination of the pronunciation of a given word, even in the cases where the POS tagging does not work or the user forces the system

to use a certain POS tag. <break> can be used for defining the break strength at different boundaries, such as character boundaries, word boundaries, prosodic phrase boundaries, sentence boundaries, etc.

For Chinese, the pitch contours play a very important role in rendering TTS speech. The same baseform syllable with a different tone 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. Element <syllable> is introduced to define the given character. The elements <frequency> and <energy> are brought in to describe prosodic features, pitch and volume, in the (time, value) format in order to have a better representation capability for prosodic features.

In this section, we have mainly discussed the major issues that are specifically relevant for Chinese TTS. For more detailed information on the other parts, please refer to [10].

4 Experimental results

4.1 Setup

The text, prosodic and acoustic processing modules of our Mandarin Chinese TTS system were evaluated in the TC-STAR project. The evaluation included tests for word segmentation, POS tagging, character-to-syllable conversion (so-called grapheme to phoneme conversion (G2P) for Western languages such as English), prosodic features (pitch and duration) and synthetic speech. An independent evaluation agency generated an evaluation text corpus for evaluation, and our TTS engine was evaluated by using this text corpus. 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.

For evaluating the character-to-syllable conversion, there are 103 files in total, and each file has about 15 sentences. For evaluating the word segmentation and POS tagging, there are 603 files in total and each file has about 20 sentences, while for evaluating prosodic processing module of Mandarin Chinese TTS, there are 6 files in total, each containing about 10 sentences. A delexicalised approach is used for the evaluation of the prosodic processing module [6]. The predicted prosodic features in our Mandarin TTS system are used for unit selection in the acoustic processing module. However, the predicted prosody is not directly forced in the synthesized waveform but instead the synthetic waveform is composed of concatenated acoustic units. As a result, the pitch contours and the durations in the synthetic waveform are in fact taken from the acoustic units.

The evaluation agency produced 6 paragraphs as the input of the acoustic module including all features from the text analysis and the prosody features. The judgment test measures intelligibility and naturalness using Mean Opinion Score (MOS) listening test.

4.2 Text Processing Module

The symbolic pre-processing module performs the tokenization, the POS tagging and the phonetic transcription of the input text. The phonetic transcription information is derived for each word in the way that the word is spoken in isolation. The POS coding is partly based on the formal definition specified in the LC-STAR project, among others: NOM (name), ADJ (adjective), ADV (adverb), PRE (preposition), DET (determinant). Chinese is a tonal, monosyllabic language with 5 tones. The tone of each syllable is described by its pitch contour, and each 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. It consists of five parts, the alphabet, the initials, the finals, the tone marks and the syllable-dividing mark. Each character corresponds to one syllable. In the TC-STAR evaluation, the different tests and the results obtained in them are discussed in more detail below.

4.2.1 Word segmentation

The system processed the text and produced a tokenized version of the text including word boundaries. An independent evaluation agency checked the word segmentation against the reference transcriptions. The correct reference segmentations were created using the LC-STAR lexicon that is used in the TTS system. The evaluation corpus contained 2352 words taken from Mandarin Chinese under the 863 project.

Correctly segmented word rate = 98.21%

4.2.2 POS tagging

The evaluation of POS tagging was carried out on correctly segmented words only. The LC-STAR Mandarin POS tag set was modified to have 18 tags in the set and used in the evaluation. The evaluation was performed as a comparison of the POS tagging at the output of the text processing module with the manual reference POS tagged corpus.

Average error rate of POS tagging = 1.786%

4.2.3 Character-to-syllable conversion

The grapheme-to-phoneme conversion for Mandarin is rather straightforward, since one can always fall back to the mono-syllable level for pronunciations, called Pinyin. At that level many of the syllables have only one pronunciation, but there are a few hundred syllables in total that may have multiple pronunciations that can be disambiguated based on the context. Multiple pronunciations usually only differ in tone. In the vast majority of cases, the phonetic representation can be obtained using the lexicon and the results of word segmentation and POS tagging. The evaluation is carried out by performing a comparison between the phonetisation at the output of the text processing module of the speech synthesis system and the manual phonetisation. The error rate is applied as an evaluation metric with and without tonal markers. As the consequence, the error rates in the character-to-syllable conversion are proposed and defined as follows.

1. Pinyin based evaluation: counted as correct if the pinyin is matched, regardless of possible tone differences.
   Error Rate = 0.3807%

2. Pinyin and tone based evaluation: counted as correct if both the pinyin and the tone are matched or correct.
   Error Rate = 1.8225%

3. Tone based evaluation: counted as correct if the tone is matched or correct, regardless of possible pinyin differences.
   Error Rate = 1.6607%

4.3 Prosodic Processing Module

4.3.1 Evaluation metrics

Usually, acoustic objective measures are used in the evaluation of the prosodic module. For instance, to have a first evaluation of the segmental duration model usually the MSE (mean square error) is used. This metric compares, for each phoneme, the prediction given by the model with the duration measured in a human reference speech. However, the correlation between these objective measures and the perceptual judgment is not very high. Therefore, in order to evaluate the prosody we rely on judgment tests of prosody. To assess this module reliably, all the systems under the comparison will share both the input and the backend. The backend performs a re-synthesis of natural sentences. This avoids the distortions that can occur in synthetic speech. Therefore, the assessment of naturalness and quality of intonation is easier. The subjects are instructed not to take into account possible noises or acoustic distortions. The systems are evaluated using an absolute scale going from very unnatural to completely natural in the range from 1 to 5.

For each test paragraph, the evaluation agency produced the input of the prosody module including word segmentation, POS tags, pronunciation in tonal Pinyin, etc. For the evaluation of the prosody module for Mandarin Chinese, two evaluation schemes are proposed and tested. The natural sentences are uttered by the baseline speaker whose speeches are used in the acoustic database of the TTS system.

1: Judgment tests using delexicalised sentences.

The delexicalised utterances are generated purely based on the prosody information defined in the XML interface [6]. Only the melody and temporal information are preserved meanwhile the lexical information is completely omitted to reduce the cross effect for pure prosody evaluation. Voiced sounds are generated using only the first and second harmonic sinusoidal functions and unvoiced sounds are rendered as silence. The f0 and energy (for voiced sounds) and duration (for voiced and unvoiced) are consistent with the prosody information. The amplitude of the second harmonic component is one fourth of the first one. A judgment test is performed by naive subjects. Each subject reads the original text sentences and judges the prosody using a 5-point scale for rating of naturalness from very natural to very unnatural.

2: Functional test using delexicalised sentences.
For each utterance, the subjects have to choose which sentence from a set of 5 is the most appropriate for the particular prosody. The sentences, designed by the linguistic experts, should differ either in phrase modality, boundaries, phrase accent or number of syllables. One of the sentences is the correct one.

4.3.2 Evaluation results

Figure 6 showed the overall performance evaluation of prosodic module in Mandarin TTS system using judgment test and functional test, denoted as evaluation 1 and 2. Sentence level evaluation results are also given in Figure 7 and Figure 8, respectively.

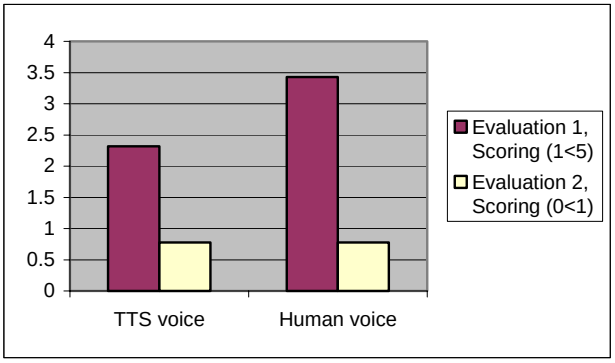

Figure 6. Evaluation performance of prosodic module in Mandarin TTS.

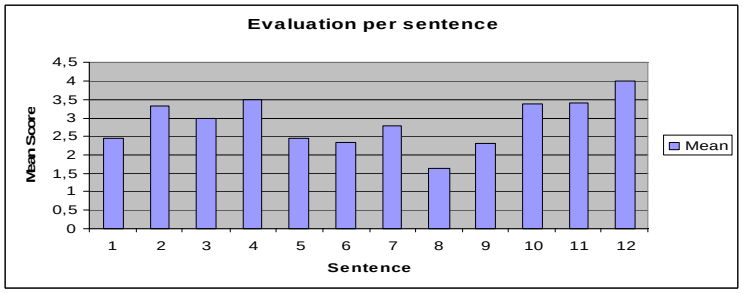

Figure 7. Judgment test using delexicalised utterances.

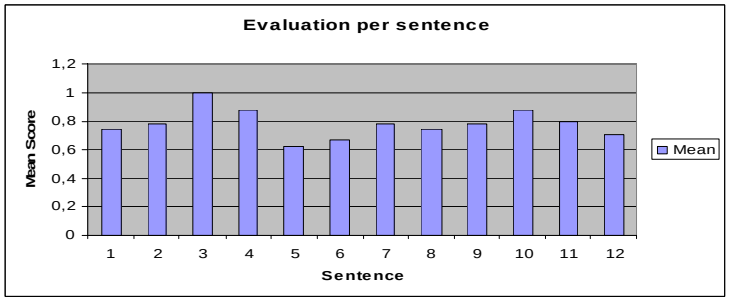

Figure 8. Functional test using delexicalised utterances.

4.4 Acoustic Processing Module

The evaluation agency produced 6 test paragraphs as the input of the acoustic module including all features from the text analysis and the prosody features. The prosody features are based on the reading of the sentences by a professional speaker. However, the acoustic modules are tuned to one particular speaker. In order to provide the modules with a prosody description matched with their voices, the sentences are read by the baseline speakers, i.e., the same speakers that recorded the baseline corpus. The judgment test measures intelligibility and naturalness using a 5-point scale against natural speech read by baseline speaker as shown in Table 1.

Table 1. Evaluation of acoustic module, synthetic TTS and natural speeches.

	Natural speech	Synthetic speech
Intelligibility	4.39	3.24
Naturalness	4.29	2.61

4.5 TTS system

In order to evaluate the system as a whole, a black box test was used. Subjects were asked to indicate their subjective impression of global quality aspects of synthetic output by means of rating scales. The evaluation agency selected 12 paragraphs. A judgment test was performed by naive subjects. Each subject has to rate the outputs in terms of several aspects of the voice using a scale from 1 to 5 as shown in Table 2. It clearly showed the synthetic speech is far below quality of human speech indicating some improvement needs.

Table 2. Evaluation of entire TTS system, synthetic TTS and natural speeches.

System	Natural speech	Synthetic speech
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

4 Discussions and Conclusions

In this paper, we have focused on different aspects related to the development and evaluation of Mandarin Chinese TTS systems. The Mandarin TTS system is described

as three main modules. The algorithms used in our system are briefly outlined. Then the XML based interfaces between each of the module pairs are proposed especially from the viewpoint of the Chinese language. Furthermore, we have introduced meaningful module-wise evaluation schemes and presented real evaluation results we have achieved on the Mandarin Chinese language in the TC-STAR project. The evaluation methods and results are discussed for the text, the prosodic and the acoustic processing modules as well as entire system.

The experiments summarized in this paper clearly show that it is both possible and very flexible to develop the system in the modular way. The experimental results are promising in terms of performance and the results provide information on potential improvements. It can be concluded that this paper provides a very promising framework for system development and evaluation that can be used more widely and extended to fit the particular needs in the future research and development work.

Acknowledgments. This work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

References

1. Bellegarda, J., Silverman, K., Lenzo, K., Anderson, V.: Statistical prosodic modeling: from corpus design to parameter estimation. IEEE Trans. Speech and Audio Processing, Vol. 9, No.1, pp. 52-66 (2001).
2. Dutoit, T.: An introduction to text-to-speech synthesis. Kluwer, Dordrecht (2001).
3. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, Dordrecht (2000).
4. Lai, W., Wang, Y., Chen, S.,: A new pitch modeling approach for Mandarin speech. In Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland (2003).
5. Li, A.: Chinese prosody and prosodic labeling of spontaneous speech. In Proceedings of International Workshop on Speech Prosody, Aix-en-Provence, France (2002).
6. Sonntag, G., Portele, T.: PURR - a method for prosody evaluation and investigation. Journal of Computer Speech and Language, Vol.12, No.4, pp. 437-451 (1998).
7. Sproat, R.: Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer, Dordrecht (1998).
8. Tian, J., Nurminen, J.: On analysis of eigenpitch in Mandarin Chinese. In Proceedings of 4th International Symposium on Chinese Spoken Language Processing, HongKong, China (2004).
9. Tian, J., Nurminen, J., 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).
10. Tian, J., Wang, X., Nurminen, J.: SSML extensions aimed to improve Asian language TTS rendering. In W3C Workshop on Internationalizing the Speech Synthesis Markup Language. Beijing, China (2005).