

Overview of the IBM Mandarin Text-to-Speech System

**Dan-ning Jiang, Qin Shi, Fan-ping Meng, Zhi-wei Shuang,
Xi-jun Ma, Yi Liu, Yong Qin**

IBM China Research Laboratory
Building 19 Zhongguancun Software Park, 8 Dongbeiwang West Road, Haidian District,
Beijing, P.R.C. 100094

{jiangdn, shiqin, mengfp, shuangzw, maxj, liuyyi, yongqin }@cn.ibm.com

Abstract

This paper presents overview of the IBM Mandarin Text-to-Speech system. It is a concatenative speech synthesis system with a data-driven text processing module and probability-based prosody model. The TC-STAR evaluation results showed that the system is state-of-the-art both in intelligibility and naturalness, and the synthesis speech are close to natural speech in overall quality. The system also has the capability to fast-develop new voices, languages, and Chinese dialects by using data-driven methods.

1. Introduction

The IBM Mandarin Text-to-Speech (TTS) system is a concatenative synthesis system. The input text is first processed by a data-driven text processing module, where grapheme-to-phoneme conversion and prosody structure prediction are done. Then, a probability-based prosody model is utilized to estimate the prosody feature distribution under the certain prosody context. Appropriate synthesis units are selected from a large corpus based on the prosody model and other context features. As most Mandarin speech synthesis system, syllable is used as the basic synthesis unit. Finally, the selected segments are concatenated together and the waveform is generated. Pitch and duration features of the segments are usually not modified to preserve the acoustic quality, except some special requirements are encountered. For example, to provide a web-page reader to blinds, the duration is modified shorter to make the speech faster and more information can be given during the same period of time.

The system is start-of-the-art. As the TC-STAR Mandarin speech synthesis evaluation results showed, the system can synthesize speech with high overall quality. By using data-driven methods, the system also has the capability to fast-develop new voices and other Chinese dialects [Li. 2005]. If we can get the text processing module and corpus of a foreign language, it can even develop TTS of that foreign language even though we can not understand it.

The paper is organized as the follows. Section 2 briefly introduces the special features of Mandarin and the issues which should be considered in the Mandarin synthesis system. Section 3 illustrates the Mandarin speech corpus building procedure. Section 4, section 5, and section 6 present the main TTS modules of text processing, prosody model, and unit selection respectively. The TC-STAR evaluation results are discussed in section 7, and the work is summarized in section 8.

2. Issues in Mandarin Speech Synthesis

Distinguished with western languages such as English, Mandarin has its special features. First, each syllable in Mandarin has a relatively fixed structure, composed of an initial part, final part, and tone. The initial part is a consonant; the final part usually is a vowel, but may also contain some consonant, such as /n/. Besides the initial and final, the tone is also used to discriminate different meanings. There are four basic tones in Mandarin, which are high-level, high-rising, falling-rising, and high falling. In addition to these four basic tones, in continuous utterances, some syllable may be read very short and lose its original tone, and it becomes the fifth tone, the neutral tone. As a consequence, the prosody model in Mandarin synthesis system is more complex. The syllable's pitch contour is jointly influenced by the tone and intonation. But meanwhile, the smoothness of the pitch contour between two syllables is not as crucial as that in English.

Another important feature of Mandarin is that a word is composed of several characters and there is no space between two words in text to indicate the word boundary explicitly. Thus, the text processing module has to do lexical word segmentation as the basis of other further processing (see figure 1).

3. Speech Corpus Building

3.1. Script Design

The IBM Mandarin speech synthesis corpus [Zhu. 2002] is designed to cover sufficient phonemic and prosodic events. The basic unit is syllable, which is the same as most other Mandarin TTS systems. In phonetics, three aspects of the segmental variants are considered: a. phonemic context, b. tone context, and c. intonation context. The phonemic context reflects the co-articulation between syllables in continuous utterances. The tone context and intonation context jointly determine the prosodic features of a syllable. The former indicates the local prosodic structure. The latter indicates the universal

intonation. For example, an accented syllable often has an extended pitch range, and the duration of segment followed by a prosodic boundary is usually lengthened. However, currently the Chinese text analyzer still can't predict the accent position reliably. Thus, only the syllable's position in phrase is used as the intonation context.

We defined Context Variation Unit Vector (CVUV) to quantify the above variations. Each CVUV has four components: left tone context, right tone context, left phonemic class, and right phonemic class. The left tone context and right tone context indicate the tone of the preceding syllable and the following syllable respectively, whose values include five tones plus the phrase boundary. The left phonemic class is the category of the preceding syllable's final, 13 categories in total; the right phonemic class is the category of the following syllable's initial, 20 categories in total.

A tool was implemented to automatically select the scripts from a huge text corpus [Li. 2002]. Using the tool, about 20k sentences of script were selected from 900k sentences. Each sentence in the script contains about 20 syllables. After that, several hundreds of common phrases, as well as some language-specific segments were added manually as the complementary recording materials.

3.2. Voice Recording

Two problems are encountered in voice recordings. First, the speaking style should be consistent with the unlimited-domain speech synthesis task; second, the speaking style in different recording sections should be kept consistent, because the amount of sentences to be recorded is so huge that the recordings have to last several months.

For the first problem, the speaking style should be a normal one with neutral intonation, without intentional emotion, accents, or semantically specific meanings. For the second problem, both a professional speaker and an experienced human monitor are required. To make the supervision of the recording easily and effectively, we also defined detailed specifications for the speaking rate, intonation, articulate uttering and volume.

3.3. Corpus Processing

To build a TTS system, annotations including pronunciation marking, word segmentation, syllable segment alignments, and prosodic event labels are required besides speech data. The pronunciation marking and word segmentation are first automatically done based on the pronunciation dictionary and syntactic rules, and then checked manually. With the help of HMM model, the syllable segment alignments can be given automatically. But it still needs to be manually modified. Problems often occur on the joint points of two adjacent voiced segments and the boundaries between speech and silence.

The prosodic events are represented with a three-level structure: prosodic word, prosodic phrase, and intonation phrase. The prosodic word is the lowest constituent in prosodic structure, generally is a two-syllable group, and it sometimes may also contain one syllable or three syllables. The prosodic phrase is between prosodic word

and intonation phrase, with no obvious break inside it. The intonation phrase, with a complete pitch contour, can be perceived as an isolate phrase.

4. Data-driven Text Processing

The text processing module in Mandarin speech synthesis system should output two kinds of important information: the phoneme string (pinyin) for each syllable and the prosodic structure. These two works are based on linguistic analysis such as lexical word segmentation and part-of-speech tagging [Shi. 2002]. Figure 1 shows paradigm of the text processing module.

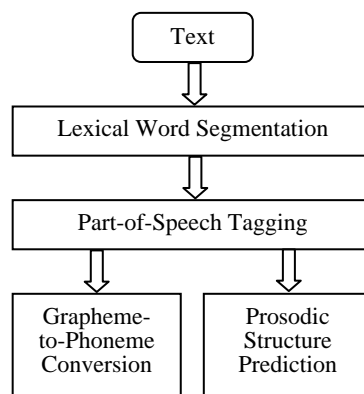


Figure 1. Paradigm of the text processing module.

4.1. Grapheme-to-phoneme Conversion

The difficulty in Mandarin Grapheme-to-Phoneme (G2P) conversion is to pick out one correct pronunciation from several candidates according to the context information. The usual method to solve this problem is to list polyphonic words and characters with correct pronunciations into a dictionary. However, such dictionary can not completely solve G2P problem, so extra pronunciation rules are needed to handle more complicated problems. Due to the above disadvantage, recently, various data-driven methods have been proposed to solve G2P problem.

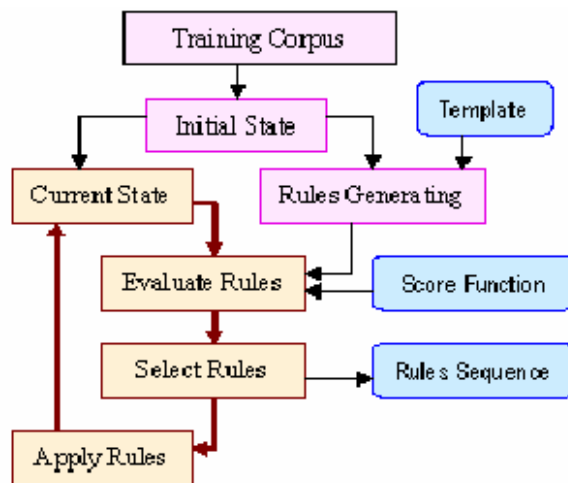


Figure 2. TBL algorithm process.

We used Transformation Based Learning Algorithm (TBL) to solve G2P problem in Mandarin [Zheng, 2005]. TBL is one of the most successful rule-based machine learning algorithms. Figure 2 shows the algorithms' process. In the training corpus, there are two kinds of states: one is initial state which is annotated automatically by the current system, and the other is target state which is corrected manually. The task of the algorithm is to select a set or ordered rules from candidates to transfer the wrong initial states into the target states with minim errors. In order to generate a set of candidate rules automatically, templates are required. A template is composed of several features and the relationship among them. The rule generation space is limited by the templates. Each difference between the initial state and target state will derive a set of rules according to the templates.

In experiments, the TBL training corpus contains 12903 sentences randomly selected from newspapers, novels and oral talks. According to discrepancy among the occurrence frequency of polyphones, G2P conversion accurate rate, and dominating pronunciation rate, 78 key polyphones out of 682 polyphones are selected to be analyzed. These polyphones can't be handled well with the usual method. The results show that TBL algorithm is very effective to generate rules for polyphones. It improves the performance both for the polyphones which have low original accuracy.

4.2. Prosodic Structure Prediction

In our system, there are three levels of prosodic structure: prosodic word, prosodic phrase, and intonation phrase. A prosodic word may contain one or several lexical words and it should be uttered without any break. Except the intonation phrase boundaries are predicted by rules based on punctuations, prosodic word and prosodic phrase boundaries are both predicted by Decision Tree. The prediction process is hierarchy: the lower-level prosodic boundaries are detected first, and then the higher-level prosodic boundaries are detected based on the lower-level prosodic unit.

The problem of prosodic structure prediction can be described as follows. Given a sequence of lower-level units $U = \{u_1, u_2, \dots, u_n\}$ (for prosodic word prediction, the lower-level unit is lexical word; for prosodic phrase prediction, the lower-level unit is prosodic word), suppose the context feature vector of unit u_i is x_i , and a parameter $a_i \in \{0,1\}$ represents whether there is a prosodic boundary after u_i or not. In binary decision tree training stage, each tree-node is assigned with one question from the candidate questions which are pre-defined. In run-time, the decision tree is first traversed by the context feature vector to find a leaf node, and then $P(a_i | u_i)$ can be given by the feature distribution of that leaf node. Finally, a possibility threshold is set to decide whether there is a prosodic boundary after u_i .

In implementation, the training corpus contains about 20,000 sentences, and the testing corpus contains 2,000 sentences out of the training corpus. Both corpora have prosodic labels according to the annotator's listening

perception. The recall rate and precise rate of prosodic word/prosodic phrase prediction are listed in table 1.

	Prosodic word	Prosodic phrase
Recall rate	93.8%	76.5%
Precise rate	95.1%	89.6%

Table 1. Prosodic word and prosodic phrase prediction results.

5. Probability-based Prosody Model

In our previous version of Mandarin TTS, prosody prediction is given by the centroid of a context equivalent cluster. However, in concatenative speech synthesis system, it is difficult to keep the prosody consistent even under the same phonetic representation in the same corpus. Thus, a fuzzy prosody model based on probability is highly expected.

Suppose $S_i (1 \leq i \leq N)$ is the synthesis unit sequence, $CT_i (1 \leq i \leq N)$ is the context feature vector of unit S_i , and $C(S_i, j) (1 \leq j \leq M_i)$ is the j -th candidate of unit S_i in corpus (M_i is the total candidate number of unit S_i). Then, the appropriateness of $C(S_i, j)$ for the unit S_i can be given by the probability $P(C(S_i, j) | S_i, CT_i)$.

A two-step process is used to model $P(C(S_i, j) | S_i, CT_i)$ [Ma, 2004]. In training stage, the context features are first clustered by the decision tree based on the distribution of the concerned prosodic features; those with similar prosodic features are clustered into the same leaf node. Then for each leaf, Gaussian Mixture Model (GMM) is trained to model the probabilistic distribution of the prosodic features. In unit selection, the decision tree is first traversed according to the context features, and the associated GMM is obtained. Thus, the probability $P(C(S_i, j) | S_i, CT_i)$ is estimated as:

$$P(C(S_i, j) | S_i, CT_i) = \sum_{k=1}^K w_k^{ij} N(\mu_k^{ij}, \sigma_k^{ij})$$

where w_k^{ij} , μ_k^{ij} , and σ_k^{ij} is the weight, mean and variation of the corresponding Gaussian component respectively.

In practice, we build a target model and a transition model for pitch and duration respectively. In the target model, pitch features are represented as a vector of four representative points in the syllable's pitch contour, converted into log values; duration feature is the syllable's duration. In the transition model, pitch features are represented as $(p_{i,1} - p_{i-1,4}, p_{i,2} - p_{i-1,3})$, where $(p_{i,1}, p_{i,2}, p_{i,3}, p_{i,4})$ is the pitch target feature vector of the candidate of unit S_i , and $(p_{i-1,1}, p_{i-1,2}, p_{i-1,3}, p_{i-1,4})$ is the pitch target feature vector of the candidate of the previous unit S_{i-1} ; duration feature is represented as $(d_i^j - d_{i-1}^k)$, where d_i^j , d_{i-1}^k is the duration of the j -th candidate of unit S_i and the duration of the k -th candidate of S_{i-1} respectively. The context feature vector CT_i of S_i is composed of three parts: tone of the syllable and its neighbor, category of the preceding syllable's final and

the following syllable's initial, and position of the syllable in the prosodic word, prosodic phrase, and intonation phrase.

6. Unit Selection

In this procedure, appropriate candidates of the synthesis unit sequence are selected from the speech corpus. It follows the framework proposed in [Hunt, 1996], where dynamic programming is used to search the corpus after the target cost and transition cost are defined. Since dynamic programming is a common algorithm, here we only focus on the definition of the target cost and transition cost.

In general, the cost function should contain terms that reflect the prosodic variations and phonemic variations of the segments. In target cost, the prosodic term is mainly given by the pitch target cost and duration target cost based on GMM probability (as illustrated in the previous section). The phonemic term is estimated through the mismatch of adjacent phone contexts (the preceding syllable's final and the following syllable's initial), because it is difficult to model the phonemic variations of each syllable in different contexts directly. We named the phonemic term as phone similarity cost. To estimate it, a phone similarity table is first computed, which contains the average spectral distance between each phone pairs. Then, suppose the previous phone and the next phone of the unit S_i is P_{prev}^t and P_{next}^t , and those of the candidate segment in its original sentence is P_{prev}^c and P_{next}^c respectively, thus the phone similarity cost is estimated as $d(P_{prev}^t, P_{prev}^c) + d(P_{next}^t, P_{next}^c)$, where d denotes the distance stored in the phone similarity table.

In transition cost, the prosodic term is also given by the related GMM probabilities, while the phonemic term is measured by a time-domain method. It measures the discontinuity of concatenated segments in multiple frequency bands, named as seam cost. To compute seam cost, each sentence in the corpus is first filtered by four pre-designed second-order linear filters. Parameters of each filter are determined through listening experiments by maximizing the correlation between the perceived discontinuity and the spectral discontinuity measured with the filter. Thus, the output of each filter at time t composes a state vector $S(t) = (s_1(t), s_2(t), \dots, s_K(t))$, where K is the total number of the filters. Then, each segment in the corpus is represented by the state vector at its beginning point and ending point. Seam cost between $C(u_{i-1}, k)$ and $C(u_i, j)$ is computed as the distance between the state vector at the ending point of $C(u_{i-1}, k)$ and the state vector at the beginning point of $C(u_i, j)$.

7. Evaluation Results

We participated in the TC-STAR Mandarin TTS component evaluation in Mar. 2006. In system delivering, we used the IBM female voice. To make our corpus size comparable with other participants', 5000 sentences are

included in the corpus both for prosody model training and unit selection.

The evaluation texts are twelve news paragraphs. We generated synthesis samples and then sent them back to the evaluation agent. The evaluation is subjective. Table 2 lists all the subjective measures and their illustrations.

Measures	Illustrations
Overall Quality (OQ)	The quality of the sound.
Listening Effort (LE)	The effort you were required to make in order to understand the message.
Pronunciation (Pr)	Any anomalies in pronunciation?
Comprehension (C)	Certain word hard to understand?
Articulation (A)	The sound distinguishable?
Speaking Rate (SR)	Average speed of the sound.
Naturalness (N)	The naturalness of the sound.
Ease Listening (EL)	Easy or difficult to listen to the voice for long time?
Pleasantness (Pl)	The pleasantness of the voice.
Audio Flow (AF)	The continuity or flow of the audio.

Table 2. The subjective measures and their illustrations.

All the subjective scores are in the scale of 0~5. Higher score means the synthesis sample is closer to natural speech. The evaluation results of natural speech are also given for reference.

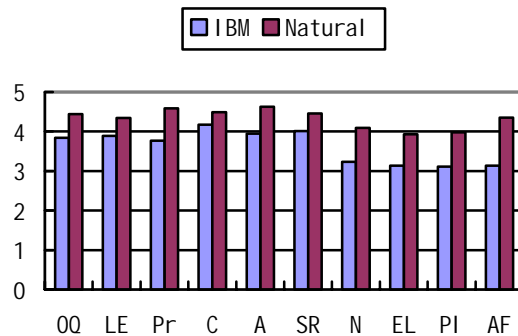


Figure 3. Evaluation results of the IBM mandarin TTS samples and natural speech.

Figure 3 shows the evaluation results. It can be seen that samples synthesized by our system are close to natural speech in overall quality. Subjects rarely feel hard to understand the words. However, the scores of measures which reflect the naturalness are only slightly more than 3, which indicate that the synthesis speech is still distinguished from natural speech.

8. Summary

This paper presents the overview of IBM Mandarin Text-to-Speech system. The system is a concatenative speech synthesis system. In text processing module, based on lexical word segmentation and POS tagging provided

with NLP processing, data-driven methods of TBL and decision tree are used to handle Grapheme-to-Phoneme conversion and prosodic structure prediction. Then, the prosodic features of each synthesis unit are modeled in a probabilistic way, by using decision tree and GMM. Finally, dynamic programming algorithm is performed in unit selection after the target costs and transition cost are defined.

The system can synthesize speech with high overall quality. Moreover, with the use of data-driven methods, less prior knowledge is required. Thus, the system has the capability to fast-develop new voices, languages and dialects.

9. References

- Zhu, W.B., Zhang, W., Shi, Q., et al. (2002) Corpus Building for Data-driven TTS Systems. In: *Proc. of IEEE TTS Workshop*.
- Li, H.P., Chen, F.X., Shen, L.Q. (2002). The Context Variation Unit Vector. In: *Proc. of ICSLP*.
- Ma, X.J., Zhang, W., Zhu, W.B., et al. (2004) Probability Based Prosody Model for Unit Selection. In: *Proc. of ICASSP*.
- Hunt, A., Black, A. (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: *Proc. of ICASSP*.
- Shi, Q., Ma, X.J., Zhu W.B., et al. (2002) Statistic Prosody Structure Prediction. In: *Proc. of IEEE TTS Workshop*.
- Zheng, M., Shi, Q., Zhang W., et al. (2005) Grapheme-to-Phoneme Conversion Based on TBL Algorithm in Mandarin TTS System. In: *Proc. of Eurospeech*.
- Li, H.P., Zhang, W. (2005) Adapt Mandarin TTS System to Chinese Dialect TTS Systems. In: *Proc. of EuroSpeech*.