The Toshiba Mandarin TTS System for the Blizzard Challenge 2008

Jian Li[1], Dawei Xu[2], Lifu Yi[1], Xiaoyan Lou[1], Jian Luan[1], Xi Wang[1], Liqiang He[1], Jie Hao[1]

[1] Research and Development Center, Toshiba (China) Co., Ltd., Beijing, China
[2] Multimedia Lab, Toshiba Corporate Research and Development Center, Kawasaki, Japan

{lijan,yilifu,louxiaoyan,luanjian,wangxi,heliqiang,haojie}@rdc.Toshiba.com.cn,
dawei.xu@toshiba.co.jp

Abstract

This paper describes the Toshiba Mandarin Text-to-Speech (TTS) system that was submitted to the Blizzard Challenge 2008. The front-end of the system uses machine-learning approaches such as generalized linear models (GLM) and Quantification Method Type 1 (QMT1) to predict pause, duration and F_0 contour. According to the predicted prosody information, the back-end of the system uses Toshiba's own "plural unit selection and fusion" method to create fused speech units which contain pitch-cycle waveforms. The pitch-cycle waveforms are then aligned along the predicted pitch marks and are overlapped with each other to generate the final speech waveforms. This paper also addresses the methods used to prepare the speech corpus and tune the performance of the back-end. The evaluation results showed that our Mandarin TTS was in the leading position among the 12 participating TTS systems.

1. Introduction

Three labs of Toshiba in Japan, the UK and China jointly work on research and development of TTS for Japanese, European languages and Chinese languages, respectively. Recently, we developed our next-generation Mandarin TTS system using Toshiba's own "plural unit selection and fusion" back-end [1]. We decided to take advantage of the Blizzard Challenge to benchmark our Mandarin TTS against those of other research organizations.

The "plural unit selection and fusion" method combines the ideas of both conventional unit selection approach and Toshiba's older closed-loop training method [2][3] so that the synthetic speech sounds not only clear at the segment level, but also smooth throughout the whole sentence.

Prosodic information, including pause, phoneme duration and F_0 contour is predicted explicitly. In the back-end, the predicted prosodic information is first used in the step of multiple unit selection where pitch, duration and pause are combined into the cost function. After the fused speech units were generated from the selected multiple speech units in the unit fusion step, the predicted prosodic information is finally used as the target prosody to modify the fused speech units.

This paper is organized as follows. In section 2, we give an overview of Toshiba's Mandarin TTS system, including prosody prediction and plural unit selection and fusion. Section 3 briefly describes our speech corpus preparation and section 4 briefly describes parameter tuning for the back-end. In section 5, we analyze the evaluation results. In section 6, we discuss the performance of the Toshiba Mandarin TTS system. Then, we address conclusions and future work in section 7. Finally, we express our acknowledgements in section 8.

2. System Overview

Toshiba's TTS system consists of two major parts: front-end and back-end. The major function of the front-end is to analyze the text and then predict prosodic information based on the results of text analysis. In our system, the prosodic information includes pause, duration and F_0 contour.

In back-end, multiple units are selected for a target segment according to target cost and concatenation cost. A new speech unit is then fused from these selected multiple units. The series of new fused speech units for a sentence are modified according to the predicted prosody and concatenated with each other to form the speech waveform.

The details of these two parts will be introduced in the following two subsections.

2.1. Front-end

2.1.1. Text analysis

Firstly, the input text is normalized so that the dates, time, numbers, etc., contained in the input text are converted into proper Chinese character strings. In this year's Blizzard Challenge, all text sentences were normalized when they were released, so this normalization step is not made use of. Then, the normalized text is syntactically analyzed. From text analysis, we can get the following basic linguistic and phonetic information:

- The word segmentation of the sentence.
- The part of speech (POS) of each word in the sentence.
- The pronunciation (pinyin with tone) of each Chinese character in the sentence.

From the basic information mentioned above, we can further get some other linguistic information, such as the position of a Chinese character in a word, the position of a word in the sentence, the length of the word, and so on. The phonetic and linguistic information is the basis for prosodic prediction. They are referred to as *attributes* hereinafter.

2.1.2. Pause

The purpose of the pause model is to predict pauses from a sequence of contextual linguistic attributes for each segmented word. We use a generalized linear model (GLM) to predict pause for our TTS system [4].

GLM is a generalization of the multivariate linear regression model [5]. It can handle attribute interactions and allows the use of different distributions. We assume that the error distribution of pause obeys a Bernoulli distribution, which our experiments show outperforms a normal distribution. Accordingly, the Logistic GLM [5] is applied to handle the Bernoulli distribution.

For each segmented word, we use the attributes of the left 3 words and the right 2 words as the attribute set, which include POS and word length. The attribute set is automatically selected by stepwise regression, which is a totally data-driven method. Open tests show the proposed method outperforms CART.

After pause prediction, we can get the distance of a Chinese character to the next and previous pauses, which are important for F_0 prediction and duration prediction.

2.1.3. Duration

Quantification Method Type 1 (QMT1) is used for duration modeling. For duration prediction, the linguistic and phonetic attributes, such as the part of speech, the tone of the phoneme, distance to the previous and next pauses, and so on, are discrete variables of QMT1. The duration of phoneme i is predicted by the formula:

$$\hat{d}_i = \sum_k \sum_m a_{ikm} \delta(k, m) \qquad (1)$$

where $\delta()$ is the characteristic function:

$$\delta(k, m) = \begin{cases} 1 \cdots \text{if } k\text{th attribute of the phoneme} \\ \qquad \text{falls into the category } m \\ 0 \cdots \text{otherwise} \end{cases}$$

The coefficients a_{ikm} of the QMT1 model are trained to minimize

$$\sum_{j=1}^{N_i} (d_{ij} - \hat{d}_{ij})^2$$

where d_{ij} is the duration of jth sample of phoneme i in the training corpus and N_i is the total number of samples of phoneme i in the training corpus.

2.1.4. F_0 contour

Mandarin is a syllabic tonal language. Consequently, the F_0 contour is based on the sequence of syllables in an utterance. The shape of an F_0 contour is highly related to the tones of the corresponding syllables.

In our system, we use a codebook-based F_0 contour model [6][7], which contains two parts for representing an F_0 contour: one is the shape of an F_0 contour for a single syllable, and the other is the offset level of an F_0 contour on the frequency axis. For simplicity, in this paper we also refer to the shape of an F_0 contour as a F_0 *pattern*.

In the training phase, for each tone a codebook of representative F_0 patterns is firstly obtained from the speech corpus by the vector quantization clustering method. Then, for a F_0 contour r_i in the corpus, we can calculate the approximation error e_{ij} if it is generated by the representative F_0 pattern c_j in the codebook of the corresponding tone. The approximation errors and the attributes t_j of the F_0 contour r_i are used to train the QMT1 models M_j, which are used in the prediction phase to select an optimal representative F_0 pattern from the codebook. In the prediction phase, we firstly get the phonetic and linguistic attributes from the text analysis results for each syllable S_i. Then we predict the approximation error $\hat{e}_{ij}$ to each F_0 pattern j in the codebook using the QMT1 models M_j by formula (1) mentioned in section 2.1.3. The F_0 pattern in the codebook with the minimal error is selected. In the training phase, the QMT1 coefficients are trained to minimize

$$\sum_{j=1}^{N_i} (e_{ij} - \hat{e}_{ij})^2$$

Similarly, we can train and predict the offset of a F_0 contour using the QMT1 method.

For all the F_0 contour r_i in the training data, new clusters of G'_j are made so that the predicted approximation error $\hat{e}_{ij}$ according to QMT1 models M_j is the minimal. With the training data in G'_j, we renew the representative F_0 pattern c'_j in the codebook. The process of building F_0 codebook, training the approximation error prediction QMT1 models, training the offset prediction QMT1 models and renew the cluster is repeated until the sum of the total approximation error in all clusters converges.

During prediction, after the optimal F_0 pattern is selected and the F_0 offset level is predicted for a syllable, we generate the F_0 contour for the syllable by combining the predicted pattern and predicted offset. Then, the F_0 contour is expanded or contracted by the predicted duration. And finally, we concatenate the F_0 contours of all syllables to generate the F_0 contour of the whole sentence.

2.2. Back-end

In the back-end, the "plural unit selection and fusion" method is used to generate speech units from speech corpus. This method is different from the conventional unit selection approach in that it uses two steps to create a speech unit: unit selection and unit fusion. In the first step, multiple speech units, rather than a single unit as in conventional unit selection, are selected for each target segment according to their target costs and concatenation costs to the neighboring units. In the second step, the selected multiple units are averaged to generate a new fused speech unit. In the Toshiba Mandarin TTS system, initials and finals of PinYin syllable are treated as phonemes. An initial is always a consonant. A final can be a vowel or a vowel with a nasal coda. Each fused speech unit contains pitch-cycle waveforms for every halfphone. Then to generate the speech waveforms, the fused speech units are modified according to the predicted prosody and concatenated with each other.

Compared with Toshiba's older closed-loop training method [2][3], in which we created an optimal diphone speech unit for all phonetic contexts, the current method selects speech units for the target segment for the specific phonetic context in the target sentence so that speech units are more locally suitable for each target sentence.

Compared with conventional unit selection, the final speech unit is fused from multiple optimal speech units using the new method, rather than using just a single selected speech unit. Synthetic speech using conventional unit selection often suffers from discontinuities at unit boundaries and instability in voice quality throughout a sentence. As the speech corpus becomes larger and larger, the risk of discontinuities at boundaries becomes smaller, but the voice quality of the synthetic speech may often sound more unstable because in a larger speech corpus, the voice quality itself can vary more over the length of the corpus. The bottleneck in the conventional unit selection approach can be abstracted by saying that for a particular target segment in the sentence, we usually cannot find a speech unit that exactly matches the phonetically ideal one. The phonetic distance between the ideal and the found unit is the fundamental reason for

distortion in the synthetic speech. However, with the "plural unit selection and fusion" method, we search for multiple speech units around the ideal one and then average them to create a new unit which is closer to the ideal unit in the phonetic space. Consequently, synthetic speech using the new method results in less distortion in quality.

The details of each step of the "plural unit selection and fusion" method are described in the following subsections.

2.2.1. Selection of multiple speech units

Instead of a single speech unit, multiple speech units are selected for a target segment in a sentence. The cost evaluation for the speech units is composed of the commonly used target cost and concatenation cost:

- Target cost:
  - Phonetic context to the target segment;
  - Duration cost;
  - F_0 cost at the beginning and ending point;
- Concatenation cost:
  - Mel-cepstrum at the beginning and end-point.

First, dynamic programming is used to search for an optimal path of primary speech units from the beginning to the end for a sentence. Then, secondary speech units are selected based on the target cost and their concatenation cost with the neighboring primary speech units. The number of secondary speech units can vary from speaker to speaker as well as according to the system configuration, such as memory footprint. Basically, more secondary speech units can help to improve the robustness of the synthetic speech. The primary unit and the secondary units for the same target segment are fused together to form a new speech unit. The number of the unit to fuse, therefore the number of secondary units plus one, is an important parameter for the performance of the back-end. We further address with this issue in section 4.

2.2.2. Unit fusion

First, the speech units are lengthened or shortened to fit the predicted number of pitch cycle in the target segment. The pitch-cycle speech waveforms of the primary and secondary speech units are then decomposed into four sub-bands and averaged in every sub-band. Given the sampling frequency noted as f_s, the division boundaries of the sub-bands are $f_s/16$, $f_s/8$, and $f_s/4$. Finally, a formant-emphasis filter is used to make the fused speech unit sound clearer.

2.2.3. Unit concatenation

The pitch-cycle speech waveforms of the fused speech unit are aligned according to the predicted pitch marks and overlapped with each other to form the final speech waveform of the synthetic speech.

3. Speech corpus preparation

For our system, the speech corpus preparation mainly consists of segmenting the training sentence into words, tagging syntactic information about the words, and labeling the phoneme boundaries. Word segmentation and word syntactic information can be obtained by the text analysis procedure described in section 2.1.1, so it will not be described here. In the following subsections, the automatic and manual steps to build the speech corpus will be briefly introduced.

3.1. Manual check of text and pinyin

The recorded speech does not always correspond to the prompt text due to mispronunciation by the speaker or other reasons. It is necessary to correct the text and pinyin according to the recorded speech; otherwise, the wrong units may be selected.

The workload of this manual check is relatively small and we found the number of errors to be quite small.

3.2. Automatic phoneme segmentation

The forced alignment tool of HTK [8] was used to automatically segment the speech. MFCC features were used. The frame length was 20ms and the frame overlap rate was 10ms. The database is large enough to train speaker-dependent phone models.

We also label the boundaries between the closure and the release of plosives and affricates, because we model the closures and release bursts separately. Since it is difficult to model the closures using HMMs, we only train a single HMM for each plosive and affricate. Then, after forced alignment, we set the closure/release (CR) boundary at one third of the way through the plosive/affricate. Of course, these CR boundaries are not good enough and need refinement.

3.3. Automatic refining the segmentation

The performance of forced alignment is not good enough for the purposes of TTS. One reason is that the HTK was originally developed for ASR usage, and the purpose of forced alignment is to maximize the likelihood of the whole sentence rather than to find the best segmentation position of each phoneme. Another reason is that due to the large frame overlap rate, phone boundaries can only located at positions that are multiples of 10ms.

We developed an automatic refinement tool as a post-process to refine the phoneme boundaries. The tool focuses on refining boundaries between voiced and unvoiced phonemes (VU boundaries) and CR boundaries (as discussed in the previous section).

There are large differences between the features of voiced unvoiced phonemes in the time domain, such as energy and zero-crossing rate. And the boundary between the closure and release burst of a plosive or affricate can be easily found using energy. It is easy to develop a simple tool utilizing these features to find VU and CR boundaries near the forced alignment boundaries that are better than those given by forced alignment. Informal listening tests showed that synthetic speech based on refined segmentation is better than that based on forced alignment segmentation.

3.4. Manual check of the segmentation

Some automatically refined boundaries are still not good enough. For example, some canonically unvoiced phonemes are realized as voiced, in which case time domain features cannot help to improve them. Another example is that some boundaries given by forced alignment are far from the real positions and the real boundaries are not located in the searching region when refining. So we carried out a manual check based on the refined boundaries.

Besides bad VU and CR boundaries, we also manually checked the boundaries between voiced initial /l/, /m/, /n/, /r/ and their following phonemes, since the forced alignment boundaries for these phonemes are often bad, and they are not refined by the automatic tool.

In Mandarin there are some syllables that don't have an initial part (syllable onset), such as "yi" and "an", which results in many boundaries between two successive finals. For example, "xi1 an1" is the name of a city in west China. The boundary between final "i1" and final "an1" is of this type. However, in our system, neither automatic refining nor manual check has been applied to improve these boundaries, even though these boundaries as determined by forced alignment are often not ideal. The most important reason for this is that these boundaries are not easy to identify, even manually.

4. Parameter tuning for the back-end

In back-end, some parameters need to be optimized according to the language or the speaker. Two important parameters were tuned for the Blizzard Challenge. One was the number of units to be fused, N, and the other was the strength of the formant emphasis filter that is applied to the fused pitch-cycle waveforms [9].

When we set N to 1, our method is equivalent to conventional unit selection approach. When N is increased, the stability or smoothness of synthetic speech is increased while the clearness of the segments may be degraded. A formant emphasis filter is used to improve the clearness of the fused segments. Figure 1 shows the concept of tuning the number of units in fusion to get the best overall quality and tuning the parameter of the formant emphasis filter to improve the clearness further.
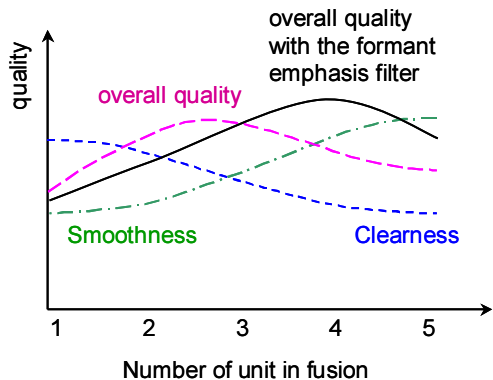

Figure 1: *Tuning the number of unit in fusion*

Because the parameter of the emphasis filter depends on N, the task of system tuning is to search for a good combination of these two parameters, as judged by subjective evaluation. An informal subjective evaluation was carried out and the combination of fusing 4 units and using a medium setting of the formant emphasis filter was selected as the best combination. This combination for Mandarin is different from that used for English in last year's Blizzard Challenge [9].

5. Evaluation results

Three aspects of the synthetic speech were evaluated in the challenge: the naturalness of the synthetic speech, the intelligibility of the synthetic speech, and the similarity to the original voice. Among the 12 systems participating, our system got the highest MOS for naturalness, the second highest MOS in similarity to the original speech, and had the smallest error rates in the SUS test for intelligibility. The identifier letter of our system was *U*.

Identifier letter *A* was the real speaker which was used as a refer system. According to the information given by the organizer, identifier letter *C* is an HMM synthesizer HTS.

5.1. Naturalness

A 5-point scale mean opinion score (MOS) was used to evaluate the naturalness of the synthetic speech. Figure 2 shows the MOS of all of the systems. Our system U got the highest MOS of 3.7 on naturalness. The MOS for the original voice A was 4.4. Pairwise Wilcoxon signed rank tests showed that all of the TTS systems were significantly lower in MOS than the original voice.
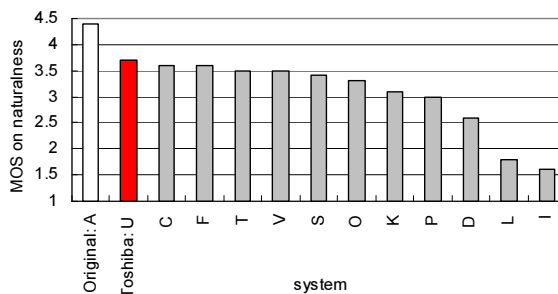

Figure 2: *MOS of the original voice* (A) *and all the systems by all listeners*

5.2. Intelligibility

Thirteen semantically unpredictable sentences (SUS) were used in the intelligibility evaluation. Subjects were required to transcribe the sentences using Chinese characters. Since in Mandarin, a pinyin syllable can correspond to multiple Chinese characters, an error rated counted by character might not necessarily reflect the performance of a TTS system. In the Blizzard Challenge, Pinyin-without-tone Error Rate (PER) and Pinyin-with-Tone Error Rate (PTER) were calculated in addition to Character Error Rate (CER). The error rate of PER is mainly due to the perception of articulation and is related to the back-end of TTS. When we subtract PER from PTER, we can further find the error only due to tone. And when we subtract PTER from CER, the remaining error may be caused by the sense of word. But semantically unpredictable sentences should make it difficult to recognize the sense of the word, and should thus increase this remaining error. Figure 3
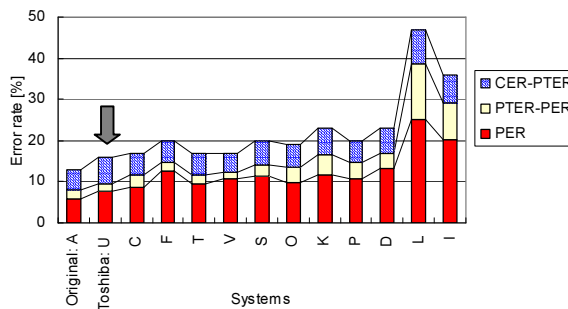

Figure 3: *Mean error rate of the original voice* (A) *and all the systems for articulation only* (PER)*, tone* (PTER-PER) *and the remaining error for character* (CER-PTER)*.*

shows the error rate in percentage for each system with three categories: PER, PTER-PER, and CER-PTER. PER, PTER and CER can be also found in this figure in the accumulated values. The order of systems is the same as figure 2.

Among all the TTS systems, our system U had the lowest PER of 7.8%, the lowest PTER of 9.5%, and the lowest CER of 16%. The intelligibility of our system U achieved the same level as the original voice. Pairwise Wilcoxon signed rank tests showed no significant difference between U and A in PER, PTER or CER. Other systems of C and T achieved the same level. Figure 3 shows that our system was also very low in PTER-PER, which means that intelligibility of tone was also very good.

5.3. Similarity

A 5-point scale was applied to rate the similarity of the synthetic speech to the original speaker. Figure 4 gives the similarity scores of all systems. The order of systems is the same as figure 2. The mean score of our system U is the second highest 3.4, which is only a little lower than system O. Supposing that some other TTS systems in the challenge use conventional unit selection, this result of similarity shows that the "plural unit selection and fusion" method doesn't change the voice quality too much compared with conventional unit selection.
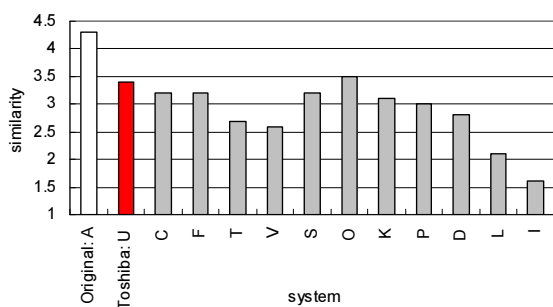

Figure 4: *Similarity scores of the original voice* (A) *and all the systems by all listeners*

6. Discussion

With Toshiba's "plural unit selection and fusion" method, we submitted very fluent Mandarin speech samples for the Blizzard Challenge 2008. The evaluation results showed that our Mandarin TTS system was in the leading position in this challenge given the same speech corpus and without consideration of limits on system resources.

The approach of "plural unit selection and fusion" has another advantage in that it is very scalable for memory footprint and computational complexity. With this method, we can also easily create a fast TTS system for an embedded application with small memory footprint and calculation power. The computationally intensive unit fusion step can be done offline by experimentally synthesizing many sentences in the target domain, and then only the most commonly fused speech units are selected into the fast TTS system. With this technology, we can tailor TTS systems that vary from very small to very large memory footprints, while maintaining high voice quality [10]. In the Blizzard Challenge, we configured the back-end as for a large memory footprint system with unlimited computational resources.

Our prosody modules are very compact compared to the speech units. The Blizzard Challenge evaluation showed that our prosody prediction is robust enough. For example, the increasing from PTER from PER is small, which means the F_0 prediction is good.

7. Conclusion and future work

This paper describes the Toshiba entry that was in the leading position in the Mandarin portion of the Blizzard Challenge 2008.

On intelligibility, our system achieved almost the same level as the original voice. The naturalness of our TTS system was 3.7, which was quite competitive for real applications. However, there is still much room for improvement in both similarity and naturalness. In the future, we will improve both the front-end and back-end, including in the following ways:

- Utilize more linguistic and phonetic information such as prosodic layers. We already have promising research results for predicting prosodic layers [11], but we still have work to do to incorporate these techniques into applications with very small memory footprints.
- Survey better F_0 contour methods. We will continue to pursue more natural F_0 prediction, especially for systems with large memory footprints. And we will try to find some F_0 contour methods that are able to deal with expressive voice better.
- In the speech corpus preparation phase, try to modify the boundaries between two consecutive finals.
- Continue the effort to improve the naturalness of the speech in the back-end.

8. Acknowledgements

9. References

[1] Mizutani, T. and Kagoshima T., "Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method", *IEICE Transactions on Information and Systems*, Vol. E88-D, No.11, pp.2565 – 2572, 2005.

[2] Kagoshima, T. and Akamine. M., "Automatic Generation of Speech Synthesis Units based on Closed Loop Training," *Proc. ICASSP97*, pp.963-966, 1997.

[3] Akamine, M. and Kagoshima, T., "Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Speech System (TOS Drive TTS)," *Proc. ICSLP'98*, pp.1927-1930, 1998.

[4] Yi, L., Li, J. Lou, X. and Hao, J., "Phrase Break Prediction Using Logistic Generalized Linear Model", *Proc. of Interspeech 2006*, pp1308 –1311, 2006.

[5] McCullagh, P., Nelder, J. A., "*Generalized Linear Models*", Chapman & Hall press, 1989.

[6] Kagoshima, T., Morita, M., Seto, S., Akamine, M., "An F_0 Contour Control Model for Totally Speaker Driven Text to Speech System", *Proc. of ICSLP1998*, pp1975-1978, 1998.

[7] Suh, C. K., Kagoshima, T., Morita, M., Seto, S., Akamine, M., "Toshiba English Text-To-Speech Synthesizer (TESS)", *Proc. of Eurospeech1999*, pp.2111-2114, 1999.

[8]   Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. "*The HTK Book*", (for HTK Version 3.4), Cambridge, United Kingdom, 2006.

[9]   Buchholz, S., Braunschweiler, N., Morita, M., Webster, G., "The Toshiba entry for the 2007 Blizzard Challenge", *Proc. of Blizzard Workshop 2007* (in Proc. SSW6), Bonn, Germany, 2007.

[10] Tamura, M., Mizutani, T. and Kagoshima, T., "Fast Concatenative Speech Synthesis Using Pre-Fused Speech Units Based on the Plural Unit Selection and Fusion Method", *IEICE Transactions on Information and Systems*, Vol. E90-D, No.2, pp.544-553, 2007.

[11] Xu, D., Wang, H., Li, G., Kagoshima, T., "Parsing hierarchical prosodic structure for Mandarin speech synthesis," *Proc. ICASSP 2006*, pp.745-748, 2006.