PKU Mandarin Speech Synthesis System for Blizzard 2009

Zhiping Zhang Xingchi Xian Lidong Luo Xihong Wu

Speech and Hearing Research Center, Key Laboratory of Machine Perception (Ministry of Education), Peking University, China

zhangzp@cis.pku.edu.cn, wxh@cis.pku.edu.cn

Abstract

This paper describes the development of PKU mandarin speech synthesis system for Blizzard Challenge 2009, which is built in the framework of corpus-based unit concatenation synthesis. The system employs a trainable VTR model named HTM to label the VTR trajectories in corpus and predict the target VTR features. In addition, a CART based prosody model is built to predict the prosody parameters of the target units. In corpus building, the speech waveform in the corpus is converted to parametric representation by STRAIGHT algorithm. And in voice building, the speech waveform is constructed from the connected STRAIGHT parameters of selected units.

Index Terms: speech synthesis, unit selection, VTR model

1. Introduction

The speech synthesis system of Peking University (PKU) was built based on unit concatenation method which is one of the most popular synthesis technologies nowadays. The system is mainly designed to synthesize mandarin speech, using the provided database from a Chinese female speaker.

In a concatenation system, the key problem is how to select the speech units from a large corpus. The process should meet some criterion to ensure the selected units can be connected as a natural utterance. Generally, the system predicts the features of target units according to the text input, and then selects the proper units from candidate sets to match the predicted features. The matching degree will be measured quantitatively by a cost function. In some previous methods linguistic factors are utilized to define the cost function to reflect the contextual constraint [1]. However, the designing of the cost function strongly relies on expert knowledge and manual tuning. Recently, some works have focused on the physical features of speech, such as f0 and spectrum, and employs statistical models for unit selection [2][3].

We exploit a novel method to select units according to the vocal tract resonance (VTR) and prosody features of the units. The VTR referred is represented as center frequency and bandwidth which are articulation related parameters. Therefore, it is more effective to measure the difference between two units in the VTR space. In our system, a trainable VTR model named hidden trajectory model (HTM) proposed for speech recognition [4] is employed to label the VTR trajectories in corpus and predict the target VTR features. In addition, the f0, energy and duration related features are extracted from the recorded speech database A decision tree model is built to predict these prosodic features.

The rest of this paper is organized as follow: section 2 will give an overview of the system. Section 3 and Section 4 will discuss the VTR model and prosody model respectively. In section 5, the method of unit selection and waveform generation will be introduced. Section 6 will introduce and analysis the evaluation results of the system in Blizzard Challenge 2009. Finally, we will give a conclusion and discussion on the future work.

2. Overview of System Development

The PKU mandarin speech synthesis system is built in the framework of corpus-based unit concatenation. The provided corpus contains 10 hours of speech data recorded by a female speaker. In the corpus, pinyin and boundaries of every syllable instance and prosodic structure are labeled, which are utilized in synthesis.

In our system, a syllable is treated as a concatenation unit, which is represented as a set of features in segment scale, and each segment corresponds to a phoneme in the syllable. For example, a mandarin syllable *hao* can be divided to at least three segments: the initial *h*, *a* and *o* in the final. If the syllable is in the continuous speech, there is a fourth segment acting as the transition from the final to the following initial. The unit features are VTR and prosody related parameters. A VTR model and a prosody model are designed to process these two classes of features respectively.

The development of the system consists of two stages. The first one is corpus building and model training. And the second one is voice building. Figure 1 and figure 2 show the diagrams of the two stages respectively. In the system, the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) parameters are extracted from the waveform of speech, which include f0, spectrum envelop and aperiodicity [5]. These parameters are used to train models and constructed synthesized speech.

In the two stages, VTR model and prosody model play core roles. The next two sections will give a detail presentation of the two models respectively.

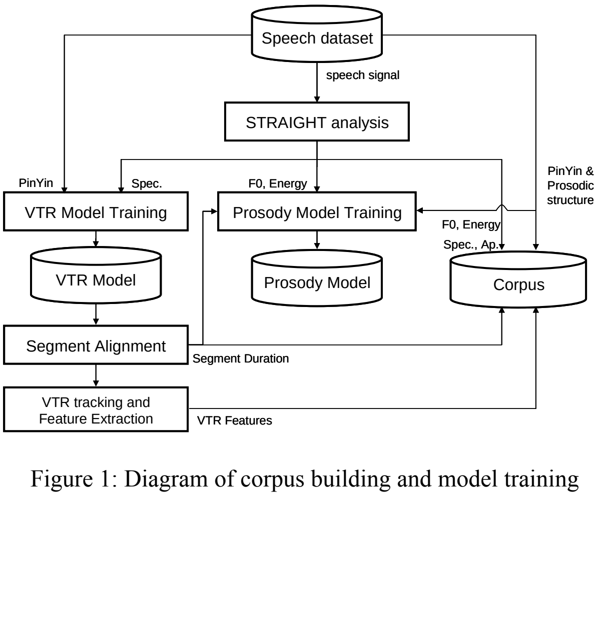

Figure 1: Diagram of corpus building and model training

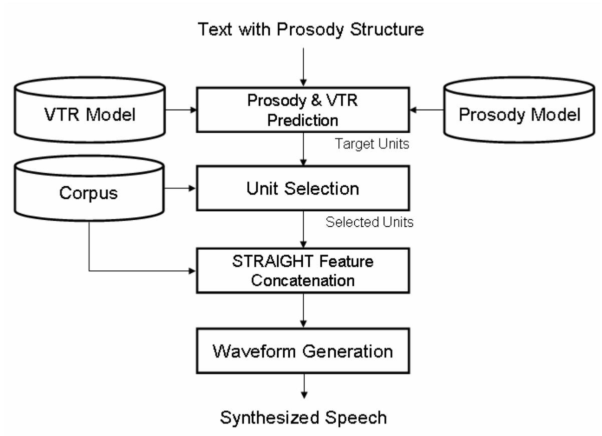

Figure 2: Diagram of voice building

3. VTR model

The VTR features of speech units are the static and dynamic VTR frequency *VF* and Δ*VF*, static and dynamic VTR bandwidth *VB* and Δ*VB*. They are the mean values of the VTR parameters and their differences in a segment domain respectively. They should be extracted from the VTR trajectories. However, it is not easy to exact the trajectories from the speech waveform directly. In the system, we employ HTM to solve the problem.

3.1. Hidden trajectory model

Hidden trajectory model (HTM) was proposed by Deng for speech recognition [4], which describes the dynamic structure of speech in the hidden vocal tract resonance (VTR) space instead of the observed feature space employed by conventional HMM based speech model.

The VTR is related to but is also different from the formant, which is defined as the energy prominence in spectrum. The former can be viewed as the underlying physical mechanism of the latter and keeps continuous during a speech, even if in some consonant segments when the formant can not be measured [6].

In the VTR space, the fluent speech can be described as smooth trajectories driven by a sequence of phoneme specific targets as shown in Figure 3. Though, the VTR trajectories can not be extracted directly from waveform, they can be mapped to acoustic features, such as LPCC [4], by a mapping function. Based on the principle, the HTM builds a framework to learn the phonetic targets as VTR variables and the smoothing function which is related to the coarticulation from observed features.

In HTM, the phonetic context can be captured utilizing a highly compact set of parameters. The advantage of the structure model is not only the requirement of less training data but also the better generalization to other speech styles, speech rate and speakers.

3.2. Modified HTM for synthesis

In our system, the observed features are the discrete cosine transform (DCT) cepstrum which is converted from STRAIGHT spectrum envelop (in log scale) by DCT. The spectrum envelop can be viewed as the frequency response of a speech synthesis filter *H*(*z*) which is integrated by three sub-filters connected in series as shown in (1).

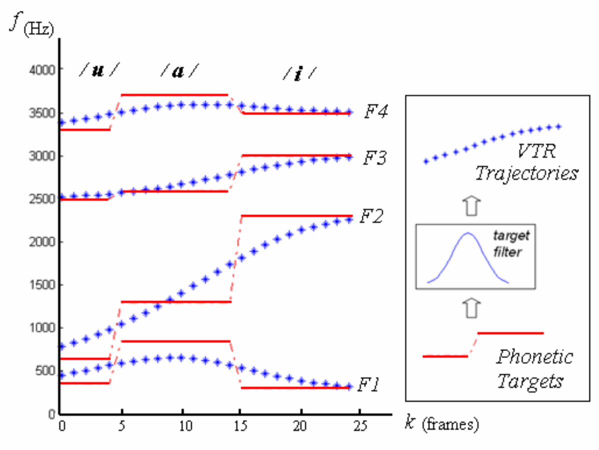

Figure 3: *Phonemic targets convert to trajectories by a target filter.*

$$H(z) = S(z)\cdot V(z)\cdot L(z) \qquad (1)$$

The *S*(*z*) named source filter will shape the spectra of excitation as that of the glottal source. The filter can be designed as a mono pole filter as (2).

$$S(z) = \frac{1}{1-\mu z^{-1}} \qquad (2)$$

The parameter μ is related to the bandwidth of the filter, which is different in the voiced speech and the unvoiced speech.

The *T*(*z*) is the vocal tract filter which models the effect of vocal tract to the source. As shown as (3), *T*(*z*) is designed as a filter of five conjunctive pole pairs to simulate 4 VTRs.

$$T(z) = \prod_{i=1}^{4} \frac{1}{(1-z_i z^{-1})(1-z_i^* z^{-1})} \qquad (3)$$

The all-pole filter simulates 4 VTRs, The complex roots can be represented by frequency and bandwidth pairs as given in (4)

$$z_i = e^{-\pi b_k/F_s + j2\pi f_i/F_s} \qquad z_i^* = e^{-\pi b_k/F_s - j2\pi f_i/F_s} \qquad (4)$$

where F_s is the sampling frequency.

The *L*(*z*) reflects the role of lip radiation, which can be simply modeled as a form of derivation.

$$L(z) = 1 - z^{-1} \qquad (5)$$

As analysis above, the integrated filter *H*(*z*) can be represented as a zero-pole filter in which the poles reflect the roles of vocal source and vocal tract, and the zero reflect the role of lip radiation. Figure 4 shows the framework of the modified model. As shown in the figure, the hidden space of the modified HTM is constructed by vocal source, vocal tract and lip radiation related parameters.

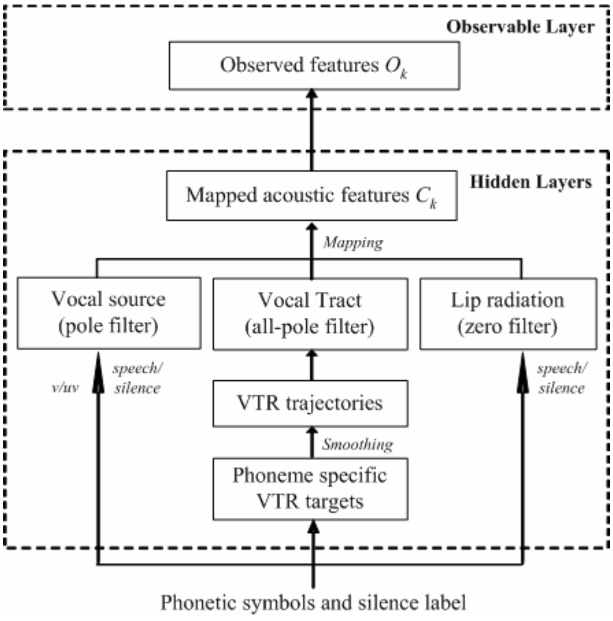
Observable Layer
Observed features O_k
Hidden Layers
Mapped acoustic features C_k
Mapping
Vocal source (pole filter)
Vocal Tract (all-pole filter)
Lip radiation (zero filter)
v/uv
speech/ silence
speech/ silence
VTR trajectories
Smoothing
Phoneme specific VTR targets
Phonetic symbols and silence label

Figure 4: The framework of the modified HTM

In the hidden space, the hidden VTRs are described as target driven trajectories. In the model, each phoneme is characterized by a VTR target which is represented as a 4-dimension vector of frequency. The phonetic symbols of speech can be converted to a sequence of targets $T(t)$ in VTR space. Then a target filter $h(\tau)$ will smooth the target sequence as a continuous trajectory $f(t)$.

$$f(t) = h(\tau) * T(t) \qquad (6)$$

In addition, the vocal source filter, working as a pole filter, will shape the voiced and unvoiced excitation by corresponding pole parameters. The lip radiation will work as differentiator. From the observed features, these parameters of the filter can be estimated in the training procedure.

3.3. Mapping from hidden parameters to observed features

The mapping from the trajectories of zero-pole to DCT cepstrums can be realized by two steps. The first step is to get the zero-pole filter's spectra response in log scale. As the sub-filters are connected in series, the log spectra S_n (n is the index of the spectrum coefficient) of the integrated filter can be calculated as summation of that of the three sub-filters.

$$S_n = S_{S,n} + S_{T,n} + S_{L,n} \qquad 0 \le n < N \quad (7)$$

In (7), $S_{S,n}$, $S_{T,n}$ and $S_{L,n}$ are the *n-th* log spectrum coefficients of $S(z)$, $V(z)$ and $L(z)$ respectively. According to (2)-(5), they can be derived from the parameters in the corresponding sub-filter as (8)-(10)

$$S_{S,n} = -\log_{10}(\left|e^{j2\pi n\omega_0/F_s} - \mu\right|^2) \qquad 0 \le n < N \quad (8)$$

$$S_{T,n} = -\sum_{i=1}^{I}(\log_{10}(\left|e^{j2\pi n\omega_0/F_s} - e^{(-\pi b_T^i + j2\pi f_T^i)/F_s}\right|^2) + \log_{10}(\left|e^{j2\pi n\omega_0/F_s} - e^{(-\pi b_T^i - j2\pi f_T^i)/F_s}\right|^2)) \qquad 0 \le n < N \quad (9)$$

$$S_{L,n} = \log_{10}(\left|e^{j2\pi n\omega_0/F_s} - 1\right|^2) \qquad 0 \le \mathrm{n} < \mathrm{N} \quad (10)$$

where ω_0 is the frequency resolution of the coefficients.

The second step is the discrete cosine transform to S_n as (11), which gets K orders DCT coefficients C.

$$C_k = \sum_{n=0}^{N-1} S_n \cos(\pi k(n + 1/2)/N) \qquad 0 \le k < K \quad (11)$$

The derived cepstrum C from the hidden space can be viewed as a kind of prediction to the observed feature extracted from the speech. The prediction error is modeled as a Gaussian random variable v with a mean C_{bias} and a variance σ as shown in (7).

$$O_k = C_k + v_k$$
$$v_k \sim N(v_k; C_{bias,k}, \sigma_k) \qquad 0 \le k < K \quad (12)$$

3.4. Model Training

In our system, one hour of speech data in the corpus is utilized to train the HTM. The model parameters include two classes. The first class contains VTR target frequencies of phonemes, VTR bandwidths of segments and the pole parameter μ in the source sub-filter. The second class is about the cepstrum prediction error and contains the mean and the variances of segments.

The model parameters are updated iteratively to maximize the observation likelihood P which is calculated as (13) for the t-th frame.

$$P[t] = \sum_{k=0}^{K-1}(\log(\frac{1}{\sqrt{2\pi\sigma_{m[t],k}{}^2}}) - \frac{(O_k[t] - C_k[t] - C_{bias,k})^2}{2\sigma_{m[t],k}{}^2})) \qquad (13)$$

The variables in (13) have been stated previously, and the subscript $m[t]$ is the segment index in the t-th frame.

In the model training, the first class of parameters should be updated by corresponding gradients as (14)

$$\theta^{w+1} = \theta^w + \lambda \sum_t \frac{\partial P[t]}{\partial \theta} / T \qquad (14)$$

where λ is a fixed step. The calculation of the derivatives to these parameters can be derived from formulas discussed above. Here will not give the detail.

To the second class of parameters, cepstrum bias C_{bias} and the variances of the M-th segment σ_M can be updated by averaging the prediction bias and residuals in relevant frames.

$$C_{bias,k} = \sum_{t=1}^{T}(O_k[t] - C_k[t])/T$$
$$\sigma_{M,k} = \sqrt{(\sum_{m[t]=M}(O_k[t] - C_k[t] - C_{bias,k})^2)/T_M} \qquad (15)$$

In model training, the boundaries of segments are initialized by conventional HMM. From the 900-th update, the boundaries of phonetic segments will be refined by Viterbi algorithm every 300 times of update. In the framework of HTM, one frame of acoustic feature corresponds to L frames (the frame shift is 10ms) of hidden targets. In principle, the number of all the possible targets sequences will be for one frame. When L=17 in our method, the lattice of Viterbi alignment is too large. To reduce the computing cost, two-stage alignment is adopted. The first stage is the first 4 times of alignment, in which the L is set to 9 and the boundary is limited to be changed less than 30 frames. After that, the second stage is fine alignment. L is set to 17, and the boundary change is limited to 2 frames. In this way the number of nodes can be reduced significantly.

3.5. VTR trajectory tracking and feature prediction

To the training data in the corpus, the smoothed VTR trajectories and segment alignment get in the last step of training are used to label the training data. To the other data, the VTR trajectories will be tracked along with the segment alignment by iterative Viterbi algorithm used in training procedure. Figure5 shows a demo of the labeled VTR trajectories for a period of voices.

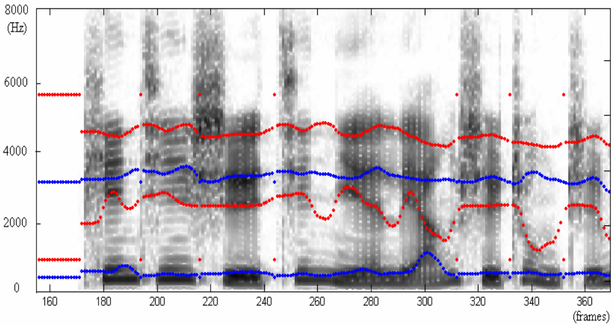

Figure 5: VTR trajectories tracked by HTM

In voice building, the VTR model will predict the VTR features of target units. According to the segment duration predicted by prosody model and training results of HTM, such as VTR targets of phonemes and target filter parameters, the system will synthesis VTR trajectories, from which the VTR features can be generated.

4. Prosody model

The prosody features include f0, delta f0, energy, delta energy and duration. Thus, all the f0 and energy related values used for training are actually mean value for each segment. The segment boundary given by VTR is employed for duration model training. f0 is extracted by fixed point analysis by STRAIGHT and the energy is collected from STRAIGHT spectrum envelop.

4.1. Model Training

We use CART (Classification and Regression Trees) to train the prosody model. The model is built on segment scale in accordance with the VTR model. Features used in CART training are based on segment-, syllable-, word-, sub-phrase- and phrase-level, shown in Table 1. The whole data in mandarin corpus is used for training the prosody model.

4.2. Prediction

CART is employed to predict f0, delta f0, energy, delta energy and duration based on the prediction features used in training. All the prediction features can be obtained from the given test utterances in the Challenge. The predicted duration is also used to generate the formant trajectory from the VTR model. The generated prosodic features and VTR features from VTR models for each segment are grouped into syllable level that we utilize as target units.

Table 1. *Features used in CART training*

Feature used in CART training (to predict f0, duration, energy)
current segment name
the initial of the current syllable
the final of the current syllable
the tone of the current syllable
position of the current syllable in the current prosodic word (forward, backward)
position of the current syllable in the current sub-phrase (forward, backward)
position of the current syllable in the current prosodic phrase (forward, backward)

5. Unit selection and concatenation

5.1. Unit corpus building

To guarantee the synthesis quality, units are stored on syllable scale. Generally, there are two groups of features in the corpus: one is used for synthesis and the other is used for selection. Each unit is associated with a bunch of features which are used for synthesis including f0, envelop and aperiodicity extracted by STRAIGHT toolkits. Features for calculating the cost are VTR and prosody features which are introduced in previous sections.

5.2. Unit selection

Given a sequence of target units predicted by VTR and prosody model for each utterance, the system uses a Viterbi search to find the minimal cost path. The target cost for each candidate at time t is calculated as follows:

$$C^t(t_i,u_i)=\sum_{i=1}^{n} w_j^t C_j^t(t_i,u_i) \qquad (16)$$

where t_i is the i-th target unit, u_i is the i-th candidate unit, and $C_j^t(t_i,u_i)$ is the j-th sub-cost for certain feature. w_j^t is the cost weight for j-th sub-cost, which is manually tuned to maximize the performance. In selection stage, only those that have same syllable identity, tone identity and segment number with specific target unit will be counted as candidates. Table 2 shows the features used in the cost functions.

Table 2. *Features used in CART training*

Target cost
VTR frequency
Delta VTR frequency
VTR bandwidth
Delta VTR bandwidth
f0
Delta f0
Energy
Delta Energy

For join cost, only the adjacency property is used. If two units are continuous in the corpus, the cost equals zero. Otherwise, a much bigger value is set as the join cost.

5.3. Waveform construction

The selected units form a sequence of syllables. f0, spectrum envelop, aperiodicity for each unit are concatenated respectively and synthesized into waveform by STRIAGHT synthesis algorithm. The parametric analysis-synthesis strategy is adopted to ensure the smooth connection of units without phase discontinuous.

6. Evaluation

The evaluation results of the Blizzard Challenge 2009 are discussed in this section. Our system is identified as N and the natural speech is identified as A.

6.1. Similarity Test

Figure 6 shows the similarity test results for Mandarin hub task (MH). It can be shown that the system achieved a median similarity level in all the systems. This can be attributed to use the original segment of a large corpus, even though there are no modifications to adapt concatenated units to new context.

Figure 7 shows the result for Mandarin Sub task 2 (MS2), which simulates the performance of synthesized voices through the telecom channel. We make no special treatment to entries for MS2. The similarity score has degraded to some extent.

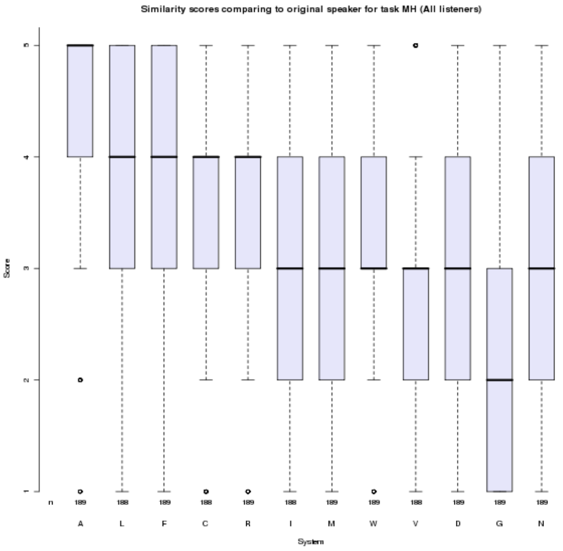

Figure 6: *Similarity scores comparing to original speaker in MH task (All listeners)*

6.2. Naturalness Test

In Blizzard Challenge 2009, naturalness test are carried on Mean Opinion Score measurement. The Boxplot in Figure 8 illustrates that our system could preserve the naturalness as many other systems.

Figure 9 shows Mean Opinion Scores in MS2. Though the degradation of quality is supposed to happen, the result does not show a significant difference compared to the MH of our system.

6.3. Intelligibility Test

For Mandarin, three measures of intelligibility are computed:
Character error rate (CER)
Pinyin + tone error rate (PTER)
Pinyin in error rate (PER)

All the test sentences are Semantically Unpredictable Sentences. Figure 10 and 11 show the test results in MH1 and MS2 respectively. Our system does not perform well in the two tests for the reason that all the selected units are original segments from the corpus without modifications in durations, energy and F0. The discontinuity is more notable in SUS utterances than in normal utterances.

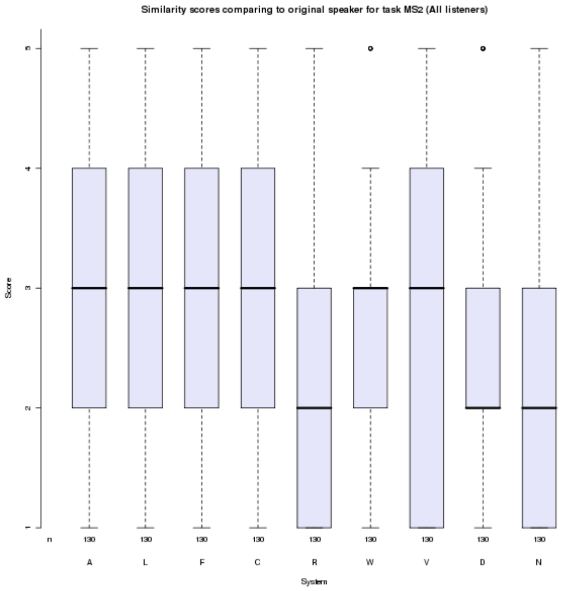

Figure 7: *Similarity scores comparing to original speaker in MS2 task (All listeners)*

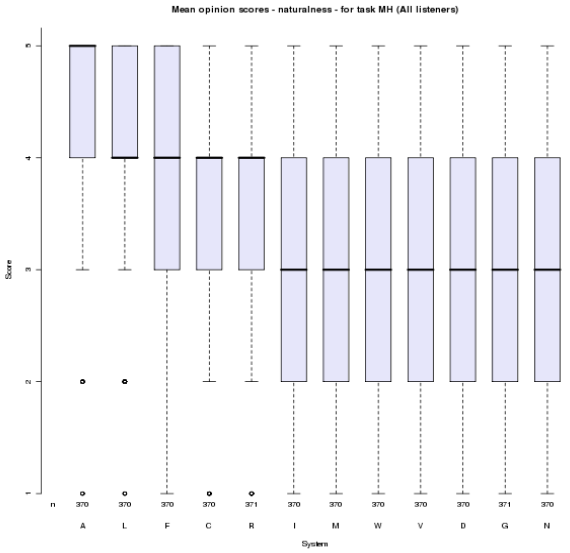

Figure 8: *Naturalness scores in MH task (All listeners)*

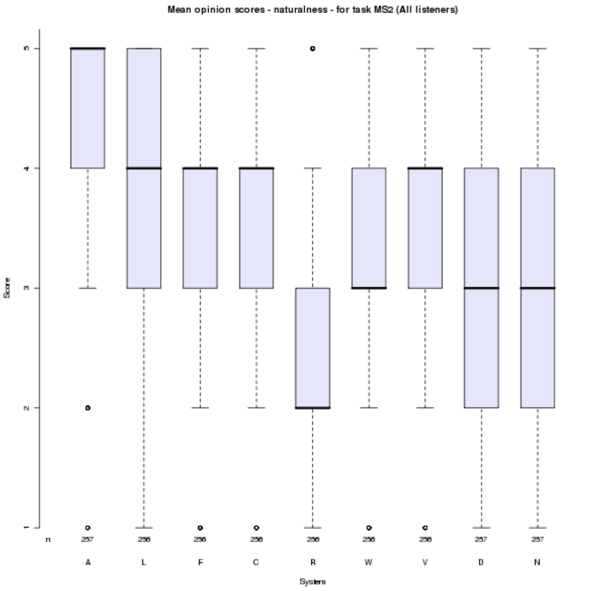

Figure 9: *Naturalness scores in MS2 task (All listeners)*

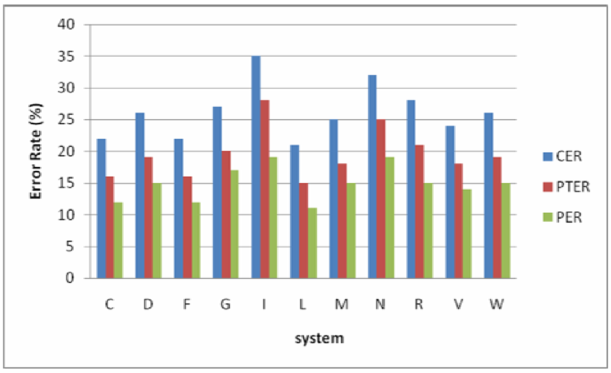

Figure 10: *CER, PTER and PER in MH task (All listeners)*

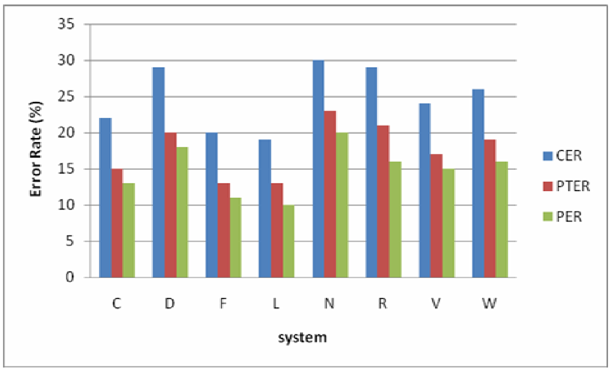

Figure 11: *CER, PTER and PER in MS2 task (All listeners)*

7. Conclusions

This paper introduces the development of the PKU mandarin speech synthesis system for Blizzard Challenge 2009. In this system, we exploit a novel concatenation synthesis method based on VTR and prosody models. The result of evaluation proved the feasibility of the method.

The VTR model based synthesis should perform well in the synthesis with a small corpus or the speaker translation with a little adaptive data. However, the related tasks are not completed for the limited development time. One of our future works will focus on the speaker conversion technology based on VTR model.

The prosody model also plays an important role. One ongoing work of us is the research on a structured pitch model which can be used to predict pitch trajectory instead of segmental pitch parameter prediction by CART. Within the same framework, the duration and energy prediction can achieve more precision.

Further more, the VTR and prosody characters of selected units need be modified by STRAIGHT toolkits to reduce the mismatch to the target units and generate more natural speech voice.

8. Acknowledgements

The work was supported in part by the National Natural Science Foundation of China (60435010; 60535030; 60605016), the National High Technology Research and Development Program of China (2006AA01Z196; 2006AA010103), the National Key Basic Research Program of China (2004CB318105; 2004CB318005), and the New-Century Training Program Foundation for the Talents by the Ministry of Education of China.

9. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in Proc. of ICASSP, pp. 373–376, 1996

[2] Donavan R. D, "Trainable Speech Synthesis", doctor thesis of Cambridge University, 1996.

[3] Ling Z. H, and Wang R. H., "HMM-based unit selection using frame sized speech segments", in Proc. of ICSLP, 2034-2037, 2006.

[4] Deng, L., Yu, D. and Acero A., "A Bidirectional Target-Filtering Model of Speech Coarticulation and Reduction: Two-Stage Implementation for Phonetic Recognition", IEEE Tran. on Audio, Speech and Language Processing, 14(1):256-265. 2006.

[5] Kawahara, H., Masuda-Katsuse, I. and Cheveigne, de A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, 27:187-207, 1999.

[6] Deng L. and O'Shaughnessy D., Speech Processing—A Dynamic and Optimization-Oriented Approach. Chapt. 2, Sec. 2.4. Marcel Dekker, New York, USA. 2003.