An HMM Trajectory Tiling (HTT) Approach to High Quality TTS
– *Microsoft Entry to Blizzard Challenge 2010*

Yao Qian[1] *, Zhi-Jie Yan*[1] *,Yi-Jian Wu*[2] *, Frank K. Soong*[1] *, Guoliang Zhang*[2]*, Lijuan Wang*[1]

[1] Microsoft Research Asia, [2]Microsoft China, Beijing, China
(yaoqian, zhijiey, yijiwu, frankkps, leonzh,lijuanw)@microsoft.com

Abstract

We propose an HMM Trajectory Tiling (HTT) approach to high quality TTS, which is our entry to Blizzard Challenge 2010. In HTT, first refined HMM is trained with the Minimum Generation Error (MGE) criterion; then trajectory generated by the refined HMM is to guide the search for finding the closest waveform segment "tiles" in synthesis. Normalized distances between HMM trajectory and those of the waveform unit candidates are used for selecting final candidates in a unit sausage (lattice). Normalized cross-correlation, a good concatenation measure for its high relevance to spectral similarity, phase continuity and concatenation time instants, is used for finding the best unit sequence in the sausage. The sequence serves as the best segment tiles to closely follow the HMM trajectory guide. Tested in four tasks, {EH1, EH2, MH1 and MH2}, of Blizzard Challenge 2010, the new HTT approach delivers high quality, natural sounding TTS speech without sacrificing high intelligibility. Subjectively, they are confirmed by naturalness and intelligibility listening test scores.

Index Terms: speech synthesis, unit selection, trajectory tiling, Blizzard Challenge

1. Introduction

Corpus-based TTS has significantly improved the voice quality of synthesized speech in past decades. The state-of-the-art, corpus-based speech synthesis systems can roughly be put into two categories: unit selection based waveform concatenation and hidden Markov models (HMMs) based parametric synthesis. Unit-selection based approach [1] can produce natural-sounding speech with occasional glitches or artifacts, particularly with a smaller database. The waveform concatenation synthesis has a large footprint and usually is difficult to modify its voice characteristics. Compared with the unit selection, waveform concatenation based TTS, HMM-based synthesis [2, 3] is parameterized in a source-filter model and statistically trained. The speech generated by the HMMs is fairly smooth and exhibits no concatenation glitches and the segmental or supra-segmental trajectories can be modified rather flexibly. However, limited by its source-filter assumption, the HMM-based TTS still carries an intrinsic, hiss-buzz vocoding flavor which makes it difficult to compete with the waveform concatenation-based TTS in terms of naturalness.

Over the last ten years, a hybrid speech synthesis approach, which combines parametric model based HMM and waveform concatenation-based unit selection approaches, has become more popular. On the one hand, HMMs can insure the smoothness and stability of generated trajectories which can guide unit selection to match their spectrum, pitch and duration information [4-9]. A probabilistic criterion of likelihood is used in selecting units for concatenation [4]. Additionally, Kullback-Leibler divergence (KLD) between target and candidate phone-based HMMs [4,8] and the HMM generated parameter trajectories from HMMs are used to select the potential candidates [5,6]. The units for concatenation can be 5ms frame, HMM state, half-phone, phone, diphone and non-uniform units. An in-depth review is given by Zen et al [10]. On the other hand, unit selection based approach can also improve the quality of HMM-based synthesis by employing stable regions of natural units [11] and using the optimal rich context model sequences [12] to alleviate or eliminate the sound muffling caused by overly smoothed HMM parameters.

Recently, we improve our TTS in both two fronts: HMM and unit selection based approaches. In HMM, the criterion of minimum generation error (MGE) [13] is used to improve HMMs trained by the conventional maximum likelihood (ML) criterion. The generation error in synthesis are first tried in Euclidean distance between generated line spectral pairs (LSPs) and those from original training data, and later extended to log spectral distortion (LSD). The state alignments of HMM generated trajectories are refined simultaneously with the spectral HMM parameters. We also use a minimum error for improving v/u error in F0 generation, in which the posterior probabilities of voiced and unvoiced states are accumulated for finding the optimal v/u switching points in a state [14]. In the unit-selection, we proposed a rich-context unit selection (RUS) approach to high quality TTS [8]. It adopts a prune-and-search procedure, where KLD is used to select potential unit candidates and normalized cross-correlation is used as the final objective measure to search for the optimal unit path [8]. Experimental results show that the voice quality of synthesized speech is significantly improved in comparing with the conventional speech synthesis based on either one of two major approaches.

Unit-selection and HMM-based approaches have their own pros and cons. The hybrid approach can combine the strength of these two approaches by: 1) generating a better trajectory by refining HMM parameters; 2) rendering more natural sounding speech by selecting the most appropriate waveform segments to tile (approximate) the generated trajectory. In this paper, we describe our Blizzard Challenge 2010 entry: an HMM trajectory tiling (HTT) based approach to high quality speech synthesis. The feature parameter trajectories generated by the improved HMMs trained in MGE are used to guide waveform unit selection. The maximum cross-correlation, a good concatenation criterion for preserving spectral similarity, phase continuity and finding the best connecting time instants, is used to search optimal waveform units for concatenation.

The rest of paper is organized as follows. In Section 2, we introduce our HTT based approach to high quality speech synthesis. The Blizzard Challenge 2010 tasks, evaluation results and analysis are presented in Section 3. We draw our conclusions in Section 4.

2. HMM Trajectory Tiling

The schematic diagram of HMM trajectory tiling (HTT) based speech synthesis is shown in Fig. 1. In the training stage, HMM parameters are first trained and then refined for their capability of synthesizing training sentence trajectories in minimum generation error. In the synthesis stage, parameter trajectories are firstly generated for constructing a unit sausage; then a best unit path is searched in the sausage; finally the optimal waveform units are concatenated to output speech. The detailed description is given in the following subsections.

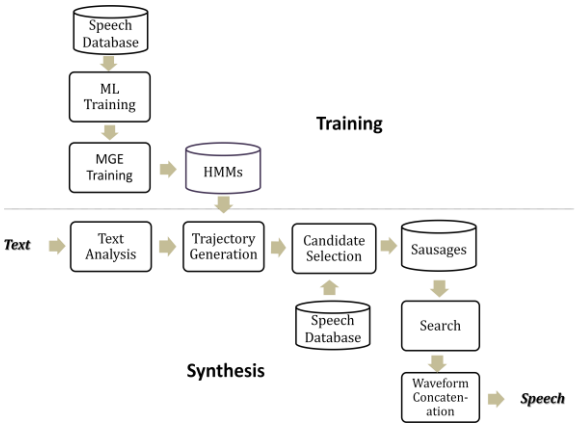

Figure 1: *Schematic diagram of HMM trajectory tiling based speech synthesis.*

2.1. HMM Parameter Trajectory Generation

In HMM-based TTS training, spectral envelope, fundamental frequency, and duration are modeled simultaneously by the corresponding HMMs [2]. In synthesis, for a given text, speech parameter trajectories are generated by the trained HMMs in the maximum likelihood (ML) sense with the dynamic ("delta" and "delta-delta") feature constraints [3]. Speech waveform is finally synthesized from the generated spectral and excitation parameters via the source-filter based production model.

We use line spectrum pair (LSP) [15] as spectral feature in training HMM. LSP parameters have good interpolation property and correlate well with "formants" or spectral peaks. It is also beneficial that perturbation of an LSP parameter is only localized in a nearby frequency region, i.e., perturbing an LSP only affects the LPC spectrum in the neighborhood.

After the ML-based HMM training, the minimum generation error (MGE) training is adopted to optimize HMMs parameters. It adjusts HMM parameters trained by the conventional EM algorithm to minimize the generation error between synthesized and original parameter trajectories of the training data [13]. The state alignments of HMMs are also refined simultaneously together with the LSP parameters.

For a given text sequence, speech parameter trajectories are firstly generated from well trained HMMs by the conventional approach, then formant sharpening is used to reduce the over-smoothing problem of HMMs and to improve the synthesized speech quality. To avoid voiced/unvoiced (v/u) generation errors caused by pitch tracking errors and corresponding flawed v/u decisions, F0 generation is improved by a minimum v/u error approach [14].

2.2. Unit Sausage (Lattice) Construction

The distance between the parameter trajectories generated from HMMs and original units is used to select potential units for constructing a unit sausage (lattice). The features we used for HMMs training are LSP, gain and F0. The distances of these three features per frame are defined by

$$d_{F0} = |\log(\mathrm{F0_t}) - \log(\mathrm{F0_c})| \quad (1)$$

$$d_G = |\log(\mathrm{G_t}) - \log(\mathrm{G_c})| \quad (2)$$

$$d_\omega = \sqrt{\frac{1}{I}\sum_{i=1}^{\mathrm{I}} w_i(\omega_{t,i} - \omega_{c,i})^2} \quad (3)$$

$$w_i = \frac{1}{\omega_{t,i} - \omega_{t,i-1}} + \frac{1}{\omega_{t,i+1} - \omega_{t,i}} \quad (4)$$

where the absolute value of F0 and gain difference in log domain between target frame $F0_t$, G_t and candidate frame $F0_c$, G_c are computed, respectively. It is an intrinsic property of LSP that clustering of two or more LSPs creates a local spectral peak and the proximity of clustered LSPs determines its bandwidth. Therefore, the distance between adjacent LSPs is more critical than the absolute value of individual LSP. The inverse harmonic mean weighting (IHMW) function [16] used for vector quantization in speech coding or directly applied to spectral parameter modeling and generation [17]. We compute the distortion of LSP by a weighted root mean square (RMS) between I-th order LSP vectors of the target frame $\omega_t = [\omega_{t,1}, \dots, \omega_{t,I}]$ and a candidate frame $\omega_c = [\omega_{c,1}, \dots, \omega_{c,I}]$, defined in Eq. 3, where w_i is the weight for i-th order LSP and defined in Eq. 4.

The distance between target unit u_t and candidate unit u_c in the corpus is defined in Eq. 5, where $\bar{d}$ is the mean distance of constituting frames. Generally, different weights need to be assigned to different feature distances due to their dynamic range difference. To avoid the weight tuning, we normalize the distances of all features to a standard normal distribution with zero mean and a variance of one and the resultant normalized distance is

$$d(u_t, u_c) = N(\bar{d}_{F0}) + N(\bar{d}_G) + N(\bar{d}_\omega) \quad (5)$$

In order to generate a compact sausage, we employ three pruning techniques: 1) Context pruning allows only unit hypotheses with same label as target; 2) Beam pruning retains only unit hypotheses within a preset distance to the best unit hypothesis; 3) Histogram pruning limits the number of surviving unit hypotheses to a maximum number.

2.3. Normalized Cross-Correlation (NCC) based Search in Sausage and Concatenation

Normalized Cross-Correlation (NCC) is used as the objective measure of concatenation for searching the optimal unit path that generates the best unit sequence [8]. Fig 2 illustrates maximum NCC based waveform search and concatenation in HTT based speech synthesis. Given two waveform units: the front unit W_f and the back unit W_b, the window length L used for computing NCC is placed at the end of W_f and the beginning of W_b. We set the offset to be within the range of $[-L/2, L/2]$, so that W_b is shifted in this range to find the maximum NCC. The best connecting time instant which matches spectral similarity and preserve phase continuity is then found by finding the offset which maximizes NCC. As a result, the smoothest waveform concatenation is achieved at this best concatenation instant of offset.

At each potential concatenation point of each waveform unit pair in the sausage, we first calculate the maximum NCC and

the corresponding time (sample) offset. Then, the unit sequence that yields the maximal accumulated cross-correlation is chosen as the optimal path. This is obtained by using dynamic programming-based Viterbi algorithm in the sausage. Finally, adjacent waveform units along the optimal path are shifted by the best offset and concatenated with triangular cross-fading.

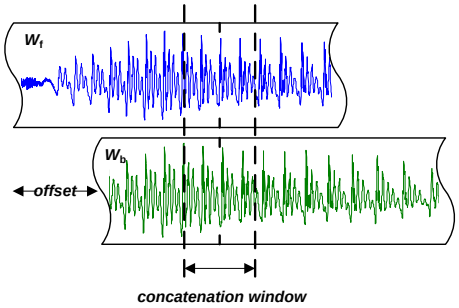

Figure 2: *An illustration of maximum NCC based waveform search and concatenation in HTT based speech synthesis.*

2.4. Flexible Units for Concatenation

In general, the longer the units are used, the less concatenation points are needed and the higher the voice quality. However, the corpus is usually not large enough to provide adequate longer units. To alleviate this problem, our HMM trajectory tiling (HTT) approach provides flexible units for concatenation. For a given target unit, HTT can select unit hypotheses with the same length for concatenation. If some target units are too long to find the better potential candidates or have no matched hypotheses after pruning in the corpus, it can back off to shorter unit. According to our informal perception test, 5ms or state based unit for small (e.g. one hour) corpus, state or half-phone for medium (e.g. 4~6 hours) corpus, and phone or diphone for large (e.g. 8~10 hours) corpus are appropriate.

3. Blizzard Challenge 2010 Evaluation

3.1. Tasks

The Blizzard Challenge was hold annually to evaluate corpus-based speech synthesis on common databases [18]. Each participant need take the released speech data, build synthesis system, and generate a given set of testing sentences. All synthesized testing sentences from all participants are evaluated by extensive listening test. This year, we participated in two British English tasks and two Mandarin Chinese tasks as following,

- Hub task EH1: 5 hour (4000 utterances) 'rjs' database spoken by a male professional speaker with RP accent and supplied by Phonetic Arts.
- Hub task EH2: 1 hour (1000 utterances) subset of the 'Roger' database spoken by a male speaker with RP accent and provided by the University of Edinburgh.
- Hub task MH1: 9 hour (6000 utterances) database supplied by the Chinese Academy of Sciences.
- Hub task MH2: 1 hour (800 utterances) subset of the MH1 data.

3.2. System Setup

Speech signals of all databases are sampled at 16 kHz, windowed by a 25-ms window with a 5-ms shift, and the 40th order LPC coefficients are transformed into static LSPs and their dynamic counterparts. Five-state, left-to-right HMM phone models, where each state is modeled with a single Gaussian, diagonal covariance output distribution, are adopted. The phonetic and prosodic contexts are used as the question set in growing a decision HMM tree for state clustering. Standard Festival labels (full-context labels) produced by the University of Edinburgh are used for both British English tasks. The labels of database in EH2 were hand-corrected by iFLYTEK. We use our own system [17] to automatically generate labels for both Mandarin Chinese Tasks.

As mentioned in Section 2, decision tree-tied HMMs are trained by conventional ML criterion and optimized by MGE training, then generated parameters are used to select potential candidates for unit sausage construction, finally a best unit path is searched for waveform concatenation. Since the sizes of databases in four tasks are rather different, we use half-phone for EH1, state for EH2 and MH2, and phone for MH1 as units for concatenation, respectively.

3.3. Evaluation Results and Analysis

There are three performance evaluation metrics: naturalness, intelligibility and similarity. Naturalness and similarity are 5-pints mean opinion scores (MOS) in term of the naturalness of synthesized sentences and similarity with original speaker. The score 5 is the best while the 1 is the worst. The intelligibility is measured by dictation results for synthesized semantically unpredictable sentences. English is measured by word error rate (WER) while Chinese is measured by character error rate (CER).

Our entry is represented by letter J in all result Figures. System A is natural speech; System B is a Festival Benchmark system: this is a standard Festival unit-selection voice built using the same method as used in the CSTR entry to Blizzard 2007; System C is an HTS_2005 Benchmark system with: this is a standard speaker-dependent HMM-based voice, built using a similar method to the HTS entry to Blizzard 2005.

The similarity, naturalness and intelligibility scores of EH1 task for all systems by all listeners are shown in Figs 3, 4 and 5. Our system achieved the second best on both similarity and naturalness and the best on intelligibility. The intelligibility score in Fig 5 shows the word error rate (WER) of our system is 15%, which is slightly better than 16% of the second best system R. The significance test at 1% level indicates that both our system and R system have no significant difference compared with natural speech A. In general, HMM-based approach achieves the best intelligibility among all TTS systems. Our new HTT approach can synthesize high quality speech without sacrificing its intelligibility.

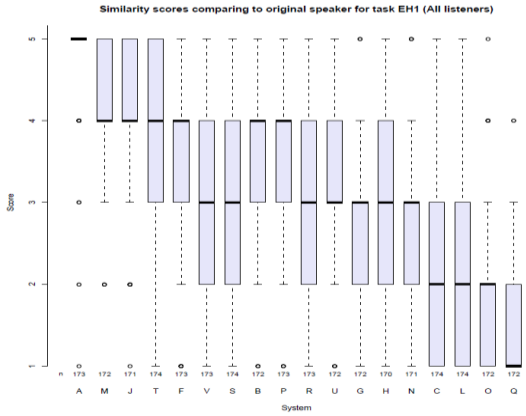

Figure 3: *The similarity score of EH1 task for all systems by all listeners. Our entry is represented by letter J.*

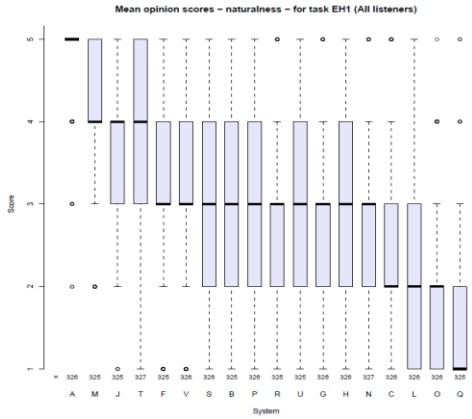

Figure 4: *The naturalness score (MOS) of EH1 task for all systems by all listeners. Our entry is represented by letter J.*

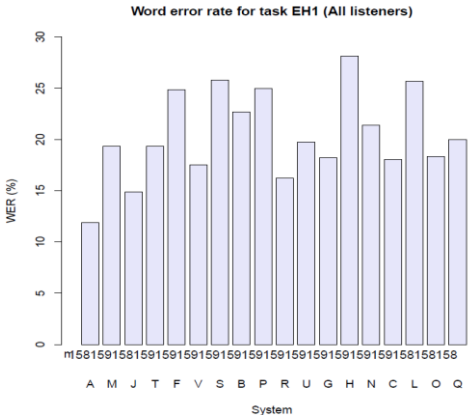

Figure 5: *The intelligibility score (WER) of EH1 task for all systems by all listeners. Our entry is represented by letter J.*

The similarity, naturalness and intelligibility scores of MH1 task for all systems by all listeners are shown in Figs 6, 7 and 8. The mean similarity score of our system is the best one, i.e., our score is 3.9 and significantly higher than the second best 3.4 of system H. Our system, system C and system R all achieved 3.9 of the mean naturalness scores and tied for the first place. Both our system and system R perform character error rate (CER) 22% of intelligibility score, which is slightly worse than 21% of the best system C. It is a little bit intractable to compare our system with system C since we use our own automatic labeling system, which is rather different from system C.

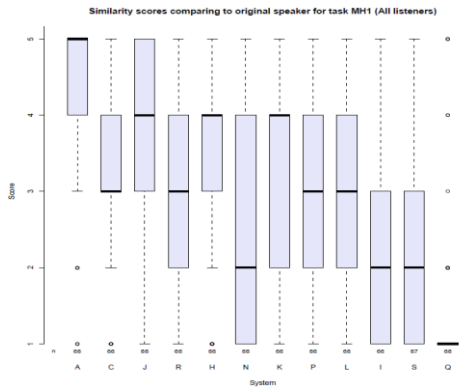

Figure 6: *The similarity score of MH1 task for all systems by all listeners. Our entry is represented by letter J.*

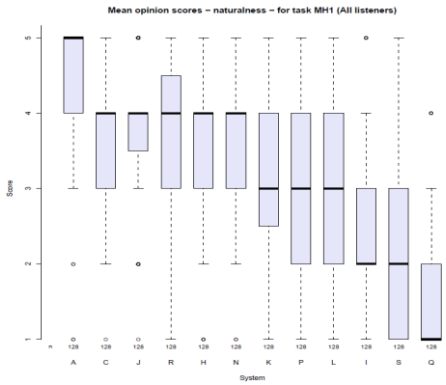

Figure 7: *The naturalness score (MOS) of MH1 task for all systems by all listeners. Our entry is represented by letter J.*

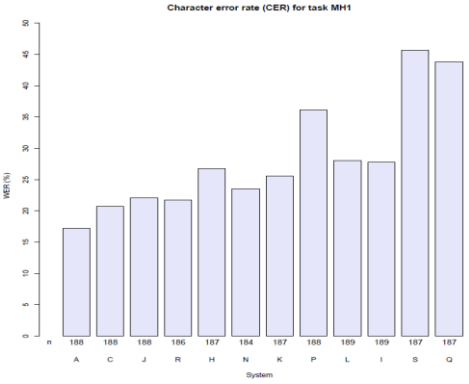

Figure 8: *The intelligibility score (CER) of MH1 task for all systems by all listeners. Our entry is represented by letter J.*

For the small size database tasks: EH2 and MH2, the performance evaluation metrics of our systems are all within top 3. The corresponding figures are shown in Appendix.

4. Conclusions

In this paper, we present our Blizzard Challenge 2010 entry: an HMM Trajectory Tiling (HTT) approach to high quality speech synthesis. The parameter trajectories are first generated by refined HMMs which are trained with MGE. The HMM trajectory is then used to guide waveform unit selection to synthesize output speech. When tested in four tasks, {EH1, EH2, MH1, MH2}, of the Blizzard Challenge 2010, the output speech rendered by HTT sounds both natural and highly intelligible.

5. References

[1] Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K., "ATR v-TALK speech synthesis system", In Proc. ICSLP, pp. 483–486, 1992.

[2] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", In Proc. Eurospeech, 1999.

[3] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", In Proc. ICASSP, pp. 1315–1318, 2000.

[4] Ling, Z.-H., Wang, R.-H., "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion", In Proc. ICASSP. pp. 1245–1248, 2007.

[5] Black, Alan W, Bennett, Christina L., Blanchard, Benjamin C., Kominek, John , Langner, Brian , Prahallad, Kishore , Toth, Arthur, "CMU Blizzard 2007: a hybrid acoustic unit selection system from

statistically predicted parameters", In Proc. Blizzard Challenge Workshop, 2007.

[6] Hira, T., Yamagishi, J., Tenpaku, S., "Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis", In Proc. ISCA SSW6, 2007.

[7] Kawai, H., Toda, T., Yamagishi, J., Hirai, T., Ni, J., Nishizawa, N., Tsuzaki, M., Tokuda, K., "XIMERA: a concatenative speech synthesis system with large scale corpora", IEICE Trans. J89-D-II, no.12, pp.2688-2698, 2006.

[8] Yan, Z.-J., Qian, Y. and Soong, F.K., "Rich-context unit selection (RUS) approach to high quality TTS", In Proc. ICASSP, 2010.

[9] Mizutani, N., Tokuda, K., Kitamura, T., "Concatenative speech synthesis based on HMM", In Proc. Autumn Meeting of ASJ. pp. 241–242, 2002.

[10] Zen, H., Tokuda, K. and Black, Alan W., "Statistical parametric speech synthesis", Speech Communication Volume 51, Issue 11, pp. 1039-1064, 2009.

[11] Gonzalvo, X ., Gutkin, A., Socoró, J. C., Iriondo, I., Taylor, P. "Local minimum generation error criterion for hybrid HMM speech synthesis", In Proc. Interspeech, pp. 416-419, 2009.

[12] Yan, Z.-J., Qian, Y. and Soong, F.K., "Rich context modeling for high quality HMM-Based TTS", In Proc. Interspeech, 2009.

[13] Wu, Y.-J. and Wang R.H., "Minimum generation error training for HMM-based speech synthesis", In Proc. ICASSP, 2006.

[14] Qian, Y., Soong, F. K., Wang, M.M., Wu, Z.Z., "A minimum v/u error approach to F0 generation in HMM-based TTS", In Proc. Interspeech, 2009.

[15] Soong, F. K., and Juang, B. H., "Line spectrum pair (LSP) and speech data compression", In Proc. ICASSP, pp.1.10.1-1.10.4.,1984.

[16] Laroia, R., Phamdo, N., and Farvardin, N., "Robust and efficient quantization of speech LSP parameters using structured vector quantizers", In Proc. ICASSP, pp. 641–644, 1991.

[17] Qian, Y., Soong, F.K. , Chen, Y.N and Chu, M., "An HMM-based Mandarin Chinese text-to-speech system", In Proc. ISCSLP 2006 , Springer LNAI Vol. 4274, pp.223-232, 2006.

[18] http://www.synsig.org/index.php/Blizzard_Challenge_2010

6. Appendix