# NICT Blizzard Challenge 2010 Entry

*Yoshinori Shiga*[1], *Tomoki Toda*[1,2], *Shinsuke Sakai*[1], *Jinfu Ni*[1], *Hisashi Kawai*[1]
*Keiichi Tokuda*[1,3], *Minoru Tsuzaki*[4], *Satoshi Nakamura*[1]

[1]National Institute of Information and Communications Technology (NICT), Japan
[2]Nara Institute of Science and Technology, Japan
[3]Nagoya Institute of Technology, Japan
[4]Kyoto City University of Arts, Japan
{yoshi.shiga, shinsuke.sakai, jinfu.ni, hisashi.kawai, satoshi.nakamura}@nict.go.jp
tomoki@is.naist.jp, tokuda@nitech.ac.jp, tsuzaki@kcua.ac.jp

## Abstract

This paper details a speech synthesis system developed at NICT for the Blizzard Challenge 2010. The system depends on an HMM-based speech synthesis technique that possesses two distinctive features: HMM training under global-variance constraint on the parameter trajectory and trainable mixed excitation for source-filter vocoding. For this year's entry, we added some modifications to the system we developed for last year's Challenge. The major improvement is on the scheme for the training of the unvoiced filter that is a component of our mixed excitation model. Despite the fact that our excitation modelling has room for further improvement, the official results show that the system achieves reasonable performance for all assessment categories.

**Index Terms**: Blizzard Challenge, statistical parametric speech synthesis, HMM-based speech synthesis, trajectory HMM, mixed excitation, residual modelling

## 1. Introduction

HMM-based speech synthesis [1] has a great advantage over other leading speech synthesis techniques in terms of flexibility and adaptability in synthesising speech with various voice characteristics and speaking styles. However, unnatural speech produced from statistically estimated acoustic features through the parametric source-filter model still poses a challenge. Attempting to achieve speech quality satisfactory for practical application, the NICT Speech Synthesis Team has developed HMM-based synthesis systems, and since 2008 submitted them to the Blizzard Challenge, though we have a long tradition and successful history of corpus-based concatenative speech synthesis [2, 3, 4].

This paper describes the speech synthesis system we submitted to the Blizzard Challenge 2010 (BC2010). The system is basically the same as the one we developed for Blizzard Challenge 2009 [5] and has two original features: HMM modelling via the technique of trajectory training with embedded global variance (GV) [6], and state-dependent excitation modelling based on tree-based state-clustering [7]. The modification on our BC2010 entry relates to the latter, and we detail it in this paper.

The remainder of this paper is organised as follows: Section 2 briefly mentions BC2010; Section 3 outlines our submitted system and details how the excitation model training has been improved, showing results from our preliminary experiment; Section 4 describes in detail how we built voices for BC2010; Section 5 discusses the results of the official listening test; and conclusions are given in Section 6.

## 2. Blizzard Challenge 2010

The Blizzard Challenge is an event that promotes improved understanding and comparison of different techniques for building corpus-based speech synthesizers on the same data. The Challenge consists of building voices from released sets of speech data and synthesizing a prescribed set of test sentences. The synthesised sentences are then evaluated through extensive listening tests by volunteers, speech experts and paid native speakers.

This time, the following corpora were released [8]:

- British English 'rjs' corpus: 4 014 utterances (approx. five hours) of a male speaker released by Phonetic Arts Ltd., Cambridge, UK
- British English 'Roger' corpus: 1 132 utterances (approx. one hour) of a male speaker released by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK
- Mandarin Chinese corpus: 5 884 utterances of a female speaker released by the National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences

This year, speech data were available at 16-kHz and 48-kHz sampling rates for both the English corpora, while the submitted wav-format files must be at 16-kHz sampling rate. Also, two sets of labels are released: standard Festival utterances, created at the University of Edinburgh using the multisyn voice-building tools, and hand-corrected phone labels (based on the original Festival utterances) given with hand-annotated prosodic labels. The latter set was kindly provided by iFlyTek Co., Ltd., China.

Tasks were divided into *hub tasks* and *spoke tasks*. The hub tasks for English speech synthesis were:

- EH1: build a voice from the English 'rjs' database
- EH2: build a voice from the English 'Roger' database, optionally using the provided hand-corrected labels

Explanation about the spoke tasks for English and all tasks for Chinese are omitted since we did not take part in any of these.

# 3. The NICT system

## 3.1. Outline

We used the same system for tasks EH1 and EH2. As already noted, the system is basically the same as the one we developed for BC2009 [5]. Its unique features are (1) trajectory training of HMMs under GV constraint [6] and (2) a trainable excitation model for parametric waveform generation [9][7][10].

The trajectory training [6] provides a unified framework for training and synthesis using a common criterion in consideration of global variance [11]. We apply this method to refine the state output probability densities obtained by the conventional HMM training. The GV-constrained trajectory training optimises the HMM parameters so that the GV of the generated trajectory is close to a natural one. The generated trajectory is therefore calculated analytically, even if we consider the GV in the parameter-generation process.

The excitation model is based on the principle of analysis-by-synthesis speech coders and consists of optimising state-dependent filter coefficients. The optimisation is achieved through iterative minimisation of the difference between synthesised excitation and the residual directly obtained from the speech corpus through inverse filtering. At its synthesis stage, the trained filters serve to generate mixed excitation by inputting a pulse train and white noise into the respective filters. The states are represented by leaves of a decision tree, which is also produced within the excitation training.

The modification we made for our BC2010 entry is related to the latter, i.e., the excitation modelling. In the following sections, we briefly explain our excitation model and then detail the improvements added to the conventional model.

## 3.2. Trainable excitation model

### 3.2.1. Generation of excitation signals

Figure 1 shows the synthesis stage of our excitation model [9], where pulse train $t(n)$ and white noise $w(n)$ are passed through voiced and unvoiced filters, $H_{\mathrm{v}}(z)$ and $H_{\mathrm{u}}(z)$. Outputs from the filters are added together to result in the excitation signal $\tilde{e}(n)$. Associated with each HMM state position $s$, each of the filters has the following transfer function:

$$H_{\mathrm{v}}^{s}(z) = \sum_{l=-M/2}^{M/2} h_s(l) z^{-l}, \tag{1}$$

$$H_{\mathrm{u}}^{s}(z) = \frac{K_s}{1 - \sum_{l=1}^{L} g_s(l) z^{-l}}, \tag{2}$$

where $M$ and $L$ are the respective filter orders. The excitation signal thereby generated will be input into the vocal-tract filter of the source-filter vocoder.

### 3.2.2. Training filters

The model components, filters $H_{\mathrm{v}}(z)$ and $H_{\mathrm{u}}(z)$, and impulse train $t(n)$, are iteratively calculated so as to minimise the error between residual and synthetic excitation. Figure 2 illustrates the procedure diagrammatically.

In terms of vectors and matrices, with $N$ being the total number of samples of the entire database, the filters are determined in a way that minimises the mean squared error $\varepsilon$, given by

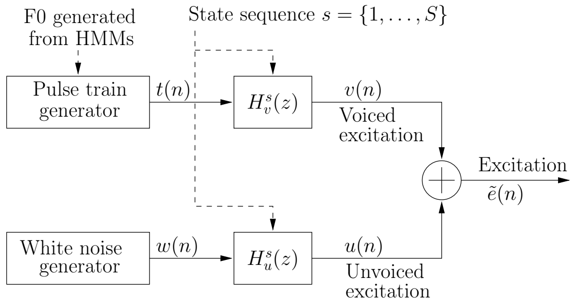

Figure 1: *Excitation signal generation: filters $H_{\mathrm{v}}(z)$ and $H_{\mathrm{u}}(z)$ are associated with each HMM state $s$*

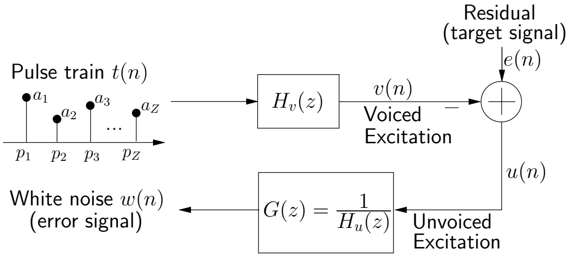

Figure 2: *Excitation model training: both filters are computed based on analysis-by-synthesis optimisation*

$$\varepsilon = \frac{1}{N}\left[\boldsymbol{e} - \sum_{s=1}^{\mathcal{S}} \mathbf{A}_s \boldsymbol{h}_s\right]^{\mathsf{T}} \mathbf{G}^{\mathsf{T}}\mathbf{G} \left[\boldsymbol{e} - \sum_{s=1}^{\mathcal{S}} \mathbf{A}_s \boldsymbol{h}_s\right], \tag{3}$$

where $\mathbf{G}$ is an $N \times N$ matrix containing the impulse response of the inverse unvoiced filter $G(z)$, $\boldsymbol{h}_s = [h_s(-M/2) \cdots h_s(M/2)]^{\mathsf{T}}$ is the impulse response vector of the voiced filter for state $s$, and the term $\mathbf{A}_s$ is the overall pulse train matrix where only pulse positions belonging to state $s$ are non-zero. In this case, each state $s = \{1, \ldots, \mathcal{S}\}$ corresponds to a different HMM state-position covering the entire database, after the Viterbi alignment.

Voiced filter coefficients for a given state $s$ are obtained by making $\partial\varepsilon/\partial\boldsymbol{h}_s = 0$, which results in a linear system for the solution of $\boldsymbol{h}_s$ [9]. On the other hand, the unvoiced filter coefficients for state $s$, $\{g_s(1), \ldots, g_s(L)\}$, and related gain, $K_s$, are determined by performing linear prediction (LP) analysis on the unvoiced excitation signal $u(n) = e(n) - v(n)$ over segments tagged as state $s$.

Apart from the determination of the filters, the positions and amplitudes of $t(n)$, $\{p_1, \ldots, p_Z\}$ and $\{a_1, \ldots, a_Z\}$, with $Z$ being the number of pulses of the entire training database, are modified in the sense of minimising the mean squared error of (3). The procedure to determine the positions and amplitudes resembles multipulse excitation linear prediction coding algorithms [12]. The overall procedure for the design of the filters and optimisation of $t(n)$ is performed in an interchanging way. Either the filter coefficient variation or the mean squared error reduction is applied for the convergence criterion.

The filters vary state by state and the states are defined according to tree-based clustering using a residual signal ML criterion [9]. The excitation training process can hence be enumerated through the following steps [5]: (1) state definition; (2) residual segment classification according to the defined states and (3) iterative filter calculation for each cluster of residual segments, using the procedure described in the previous section.

### 3.3. Improvement on the NICT BC2009 system

#### 3.3.1. *Problem of our conventional system*

Synthetic speech from this model, however, contains an excessive amount of noise. During the training process, as in Fig. 2, the differential signal of the residual $e(n)$ and voiced excitation estimate $v(n)$ is dealt with as a target signal for determining the unvoiced filter coefficients. The differential signal, however, includes an error caused in the statistical optimisation of the voiced filter response. Since the error contaminates the unvoiced excitation target $u(n)$, $w(n)$ is not actually *whitened* and the unvoiced filter tends to be overestimated, which causes the final speech output to be noisy.

Conventionally, we avoided the noisiness by attenuating the unvoiced component of excitation during the synthesis stage. This was done mainly by passing the synthesised unvoiced excitation through a high-pass filter (HPF) with a cut-off frequency of 2 kHz, before it is mixed with the voiced counterpart [9]. This remedy, however, does not remove the root cause of the noisy speech problem. The total volume of perceptible noise can certainly be reduced, but while unvoiced information is almost entirely eliminated for the range below the HPF cut-off frequency, the influence of the voiced component estimation error remains in the range above the frequency.

#### 3.3.2. *Periodic/non-periodic decomposition of residual signals*

To resolve this problem more effectively, we employ contamination-free unvoiced-excitation targets for training. The 'clean' targets are extracted directly from the residual signals $e(n)$. The overall training scheme for the unvoiced filter is shown schematically in Fig. 3. The extraction is achieved with a periodic component estimator, which is used to separate the non-periodic component of the residual from the periodic component. Periodic component estimation does also introduce a certain level of error, which directly smears the resulting non-periodic component. However, the contamination should be minor because the periodic component is estimated *locally for each speech segment* whereas in the conventional approach the voiced filter is optimised *over the entire database*.

Since the decomposition is part of the offline training process, one may employ a somewhat computationally-expensive approach. We adopt the following sinusoidal model to represent the periodic (i.e., harmonic) component of the residual:

$$\widetilde{s}(t) = \sum_{k=-J}^{J} A_k(t) \exp\Big\{\jmath \left[\Theta_k(t) + \phi_k\right]\Big\}, \tag{4}$$

where $A_k(t) = \alpha_k t + \beta_k$ and $\Theta_k(t) = \omega_k \left(\gamma t^2 + t\right)$ with $\omega_k = \omega_0 k = 2\pi f_0 k$ and the fundamental frequency $f_0$. Represented by $J$ is the number of harmonics. Obviously, in this model both the frequency and amplitude of each harmonic are approximated in a piecewise linear sense. The problem is finding $\alpha_k$, $\beta_k$, $\gamma$ and $\phi_k$ that minimise

$$\delta = \sum_{t=t_0-N_\mathrm{w}}^{t_0+N_\mathrm{w}} w^2(t) \Big[s(t) - \widetilde{s}(t)\Big]^2, \tag{5}$$

Table 1: *Number of terminal nodes of trees for each task*

| task | number of terminal nodes | |
|---|---|---|
| | voiced filter | unvoiced filter |
| EH1 | 83 | 254 |
| EH2 | 57 | 173 |

where $s(t)$ is the original signal and $w(t)$ is a window function whose length is $2N_\mathrm{w} + 1$. The solution described in [13] is applied to the problem above.

#### 3.3.3. *Modified unvoiced-filter training*

The state-dependent unvoiced filter coefficients $\{g_s(1), \ldots, g_s(L)\}$ and gain $K_s$ can be determined using LP analysis on the non-periodic component signal $u'(n)$ over segments tagged as state $s$. The states are defined using the same decision-tree-based technique of [9]. Thus, a different tree is constructed for the unvoiced filter additionally to the one for the voiced filter. During the synthesis stage, a state sequence is first produced using the 'unvoiced-filter' tree and LP coefficients that correspond to each state are used to generate the unvoiced excitation signal.

## 4. Voice building for BC2010

### 4.1. Speech parameter extraction

For both corpora, spectral envelopes and $F_0$s were estimated from the 16-kHz versions of data with 5-ms frame shifts, using the STRAIGHT analysis [14] and the Snack Sound Toolkit [15], respectively. Each of the spectral envelopes was then converted into the $39^{\mathrm{th}}$-order mel-cepstrum using the Speech Signal Processing Toolkit (SPTK) [16].

### 4.2. Synthesiser training

The released sets of files for BC2010 contain full-context labels for HMM-based speech synthesis. For task EH1, we employed the labels automatically created at the University of Edinburgh, while for task EH2 we used the hand-corrected labels with hand-annotated prosodic information that are provided by iFlyTek Co., Ltd.

Five-state left-to-right no-skip HSMMs for duration, $F_0$ and mel-cepstral coefficients were trained on the basis of the GV-constrained trajectory training [6]. The procedure of training HMMs is completely the same as we followed last year. Details are available in [5].

### 4.3. Excitation training

The orders of excitation filters were $M = 512$ and $L = 64$. Residual signals as excitation-training targets were extracted by passing speech through the *inverse* mel-log-spectrum approximation (MLSA) filter [17], for which the mel-cepstral coefficients extracted above were used. The periodic component of the residual was estimated on the technique described in Section 3.3, using 20-ms-width window and 5-ms frame shift. Once states were defined as terminal nodes of the trained decision trees, voiced and unvoiced filters of the excitation model were calculated using the procedure described in [9]. The number of terminal nodes (i.e., number of clusters) for each tree is shown in table 1.

Figure 4 shows typical unvoiced filter responses obtained from the improved training and the conventional training, for

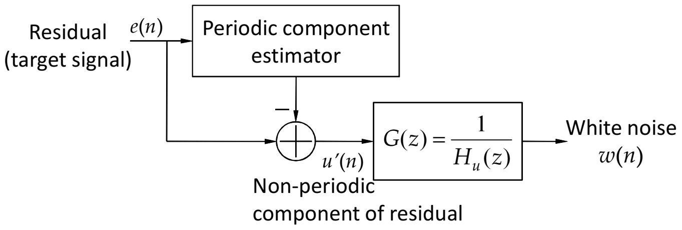

Figure 3: *Proposed training scheme for unvoiced filter coefficients*

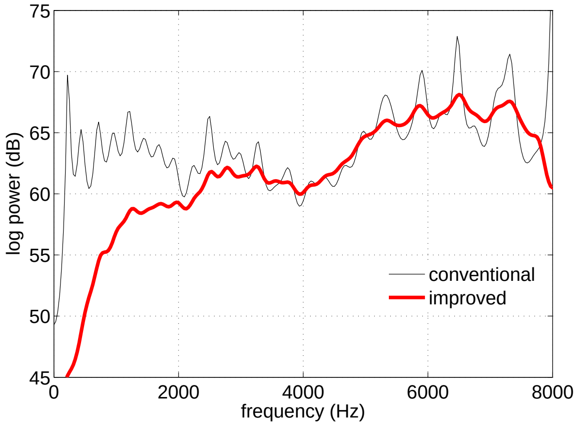

Figure 4: *Comparing unvoiced filter responses from the conventional (thin black line) and improved (thick red line) training schemes*

both of which the ‘rjs’ corpus was used as training data. These responses correspond to the second state of the 5-state HMM of the English sound /i:/ in a certain context. It can be observed from this figure that spectral energy in the low frequency range is sufficiently low for the improved training, but not for the conventional training.

### 4.4. Experiments

We conducted a subjective evaluation to confirm the effectiveness of the modified framework for unvoiced excitation training.

#### *4.4.1. Conditions and procedure*

A listening test was performed with five subjects consisting of four speech synthesis experts and one with no experience in speech research. The full 4 014 utterances of the British English ‘rjs’ corpus were used for training HMMs and the excitation model. The test took the form of an AB forced preference, with the utterances of 20 sentences taken from the Blizzard Challenge 2009 test set (the first ten sentences from each of the ‘news’ and ‘novel’ categories), with the aim of comparing the quality of speech from the conventional and improved training both with and without the application of the HPF mentioned in Section 3.3.1. The test was carried out in a quiet room with the listeners using headphones.

#### *4.4.2. Results and Discussion*

Figure 5 shows the listeners’ preference for each type of test pair. Since the unvoiced filter is estimated separately from the voiced filter estimation, clear speech with little noise is synthesised from the excitation model trained with the improved method. For this reason, as shown in Figure 5(a), speech from the improved training is preferred by listeners in 97% of all cases, if no HPF is applied to the synthesised unvoiced excitation signals during synthesis.

In contrast, Figs. 5(b)–(f) show that the HPF is still necessary even for an excitation model from the improved training. When the HPF is applied for models from the improved and conventional training, the model from the improved training can produce slightly better quality speech, although this is not significant. Careful listening by the authors revealed that the types of noise perceptible in the background differ depending on the approach. With respect to the improved training, it is an intermittent type of noise arising segmentally, whereas the noise from the conventional method is rather stationary. The former noise is considered to be generated at state boundaries, where the excitation filter response can change dramatically, because no dynamic features are used in the current model of excitation.

Based on the experimental results obtained, we decided to apply the HPF to the system for our BC2010 entry, although due to the reason above little improvement was expected over our BC2009 system in the quality of speech synthesised.

## 5. System performance

### 5.1. Results and discussion

BC2010 evaluated all the submitted systems under three main categories: (1) naturalness, (2) similarity to the original speaker and (3) word error rate (WER) [18]. Since the scales used to evaluate categories 1 and 2 are ordinal, those scores are given as medians, and comparisons among the systems are made through inspection of boxplots. On the other hand, the internal scale used to evaluate category 3 allows comparison of means [19].

The boxplots that appear in this section for naturalness and similarity scores represent the opinions of native speakers (students) who were paid to perform the listening tests. The plots showing WER reflect the opinions of all the native speakers who participated in the evaluation. It should also be noticed that in all the plots the NICT system corresponds to letter ‘G’ and natural speech to ‘A’.

Figures 6, 7 and 8 show naturalness scores, similarity scores and WERs, respectively, for all the systems submitted to tasks EH1 and EH2. It can be seen that the NICT BC2010 system performed reasonably well for all the evaluation categories in both tasks EH1 and EH2. It might not be appropriate comparing the scores between different Blizzard Challenges, but as far

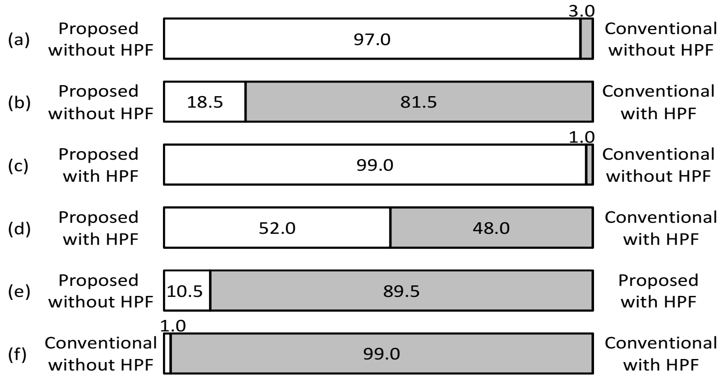

Figure 5: *Subjective evaluation results; figures indicate preference scores (%)*

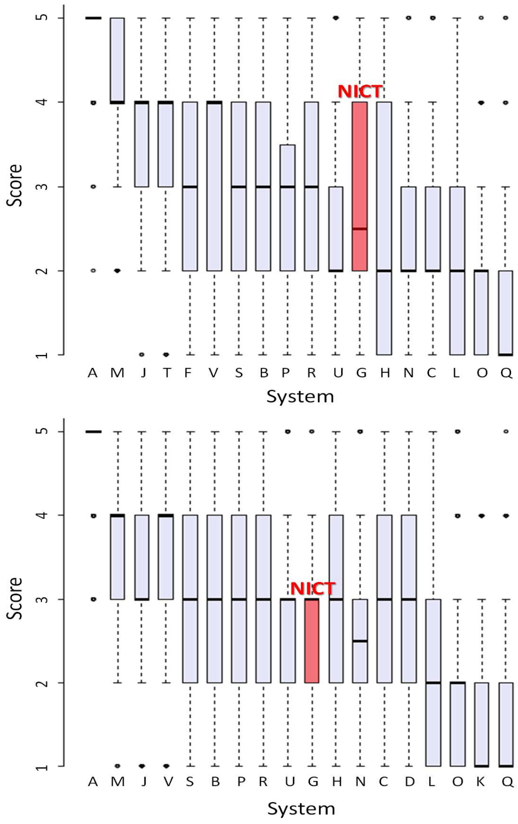

Figure 6: *Naturalness scores for tasks EH1 (top) and EH2 (bottom), assessed by paid native speakers*

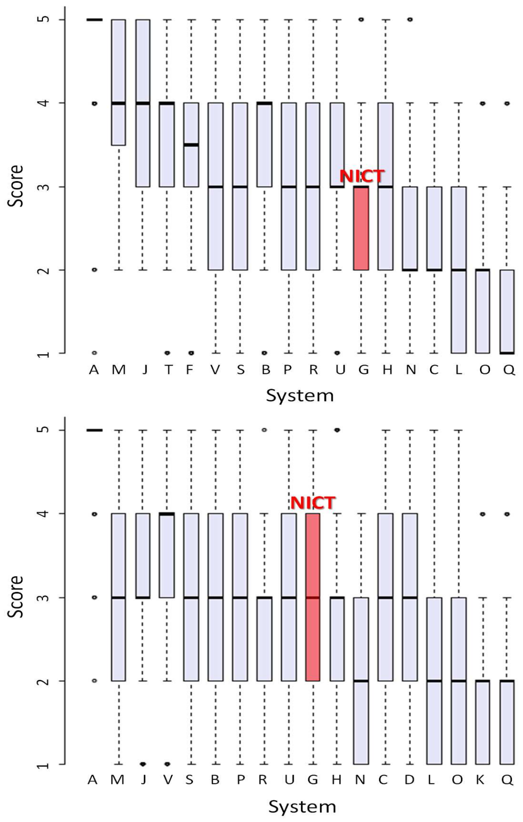

Figure 7: *Similarity scores for tasks EH1 (top) and EH2 (bottom), assessed by paid native speakers*

as task EH2 (ARCTIC subset of 'Roger') is concerned our system achieved better scores in all categories this time than it did at BC2009. However, the differences are slight and, all in all, the evaluation scores obtained from the listening test are similar to those obtained at BC2009. While our new training scheme has the potential to estimate the unvoiced filter response more accurately than the conventional training, the HPF was still necessary to reduce the intermittent noise introduced, as already noted. The use of the filter causes the resulting speech quality to be close to the quality obtained from the conventional training.

## 6. Conclusions

This paper described the NICT entry for the Blizzard Challenge 2010. Preparing for the Challenge this year, we addressed an underlying problem of our conventional system and proposed a new training framework for our mixed excitation modelling. In this framework the unvoiced excitation signal is generated through a filter whose coefficients are trained directly from the non-periodic component of the residual signal.

Although the proposed training is considered to have potential, we discovered that the unvoiced filter response changing state by state during the synthesis stage causes perceptible

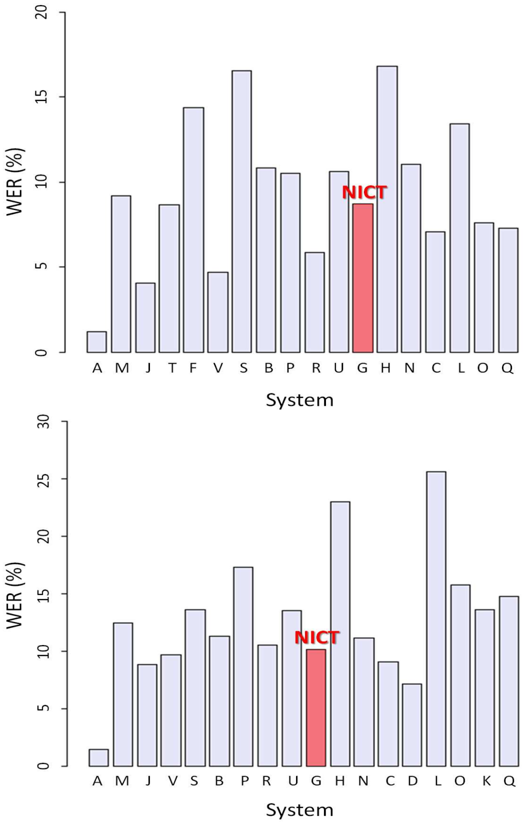

Figure 8: *Word error rates for tasks EH1 (top) and EH2 (bottom), assessed by all the native speakers*

noise in the synthetic speech. We applied a HPF to attenuate that noise, but this caused the resulting speech quality to be close to that obtained from our conventional approach. In order to properly reduce the noise caused by the discontinuities of the unvoiced filter response at state boundaries, we are now investigating introduction of dynamic features of the filter responses. We expect that the new excitation training introduced in this paper will become effective once the noise is thoroughly reduced and the application of the HPF becomes unnecessary.

Apart from the excitation modelling, from our participation in the Blizzard Challenge this year we learned that it is essential to use ample time for the adjustment of the system after prototype voice building. We had to introduce the modification quite rapidly and had little spare time to tweak the system within the limited development period. Elaborate adjustments for the finishing touches, in particular, are vital for comprehensive evaluation such as the Blizzard, where even a single glitch included in a synthetic sentence, e.g., an impossibly sharp resonance or excessively-sudden $F_0$ jump, can be fatal.

## 7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Proc. EUROSPEECH’99*, (Budapest, Hungary), pp. 2347–2350, Sept. 1999.

[2] Y. Sagisaka, K. Kaiki, and N. Iwahashi, “ATR $\nu$-TALK speech synthesis system,” in *Proc. ICSLP*, 1992.

[3] W. N. Campbell and A. W. Black, “CHATR: a multi-lingual speech re-sequencing synthesis system,” *Tech Rept IEICE*, vol. SP96-7, pp. 45–52, 1996.

[4] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, “XIMERA: a new TTS from ATR based on corpus-based technologies,” in *Proc. ISCA Speech Synthesis Workshop*, 2004.

[5] R. Maia, T. Toda, S. Sakai, Y. Shiga, J. Ni, H. Kawai, K. Tokuda, M. Tsuzaki, and S. Namakura, “The NICT entry for the Blizzard Challenge 2009: an enhanced HMM-based speech synthesis system with trajectory training considering global variance and state-dependent mixed excitation,” in *Proc. Blizzard Challenge Workshop*, 2009.

[6] T. Toda and S. Young, “Trajectory training considering global variance for HMM-based speech synthesis,” in *Proc. ICASSP*, 2009.

[7] R. Maia, T. Toda, K. Tokuda, S. Sakai, and S. Nakamura, “A decision tree-based clustering approach to state definition in an excitation modeling framework for HMM-based speech synthesis,” in *Proc. INTERSPEECH*, 2009.

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

[9] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, “An excitation model for HMM-based speech synthesis based on residual modeling,” in *Proc. SSW6*, 2007.

[10] Y. Shiga, T. Toda, S. Sakai, and H. Kawai, “Improved training of excitation for HMM-based parametric speech synthesis,” in *Proc. Interspeech2010*, (Makuhari, Japan), Sept. 2010.

[11] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” *IEICE Trans.*, vol. E90-D, pp. 816–824, Mar. 2007.

[12] W. Chu, *Speech Coding Algorithms*. Wiley-Interscience, 2003.

[13] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, Ecole Nat. Supèrieure Télécommun., France, Jan. 1996.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “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*, vol. 27, Apr. 1999.

[15] `http://www.speech.kth.se/snack`.

[16] `http://sp-tk.sourceforge.net/`.

[17] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *Proc. ICASSP*, 1992.

[18] V. Karaiskos, S. King, R. Clark, and C. Mayo, “The Blizzard Challenge 2008,” in *Proc. Blizzard Challenge Workshop*, 2008.

[19] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the blizzard challenge 2007 listening test results,” in *Proc. Blizzard Challenge Workshop*, 2007.