The NICT/ATR speech synthesis system for the Blizzard Challenge 2008

Ranniery Maia[1,2], *Jinfu Ni*[1,2], *Shinsuke Sakai*[1,2], *Tomoki Toda*[1,3], *Keiichi Tokuda*[1,4]
Tohru Shimizu[1,2], *Satoshi Nakamura*[1,2]

[1]National Institute of Information and Communications Technology (NICT), Japan
[2]ATR Spoken Language Communication Labs, Japan
[3]Nara Institute of Science and Technology, Japan
[4]Nagoya Institute of Technology, Japan

{ranniery.maia,jinfu.ni,shinsuke.sakai,tohru.shimizu,satoshi.nakamura}@atr.jp
tomoki@is.naist.jp,tokuda@nitech.ac.jp

Abstract

This paper describes the development of the NICT/ATR speech synthesizer for the Blizzard Challenge 2008 and discuss the official results. The submitted system is based on the hidden Markov model speech synthesis technology and utilizes an improved excitation approach based on residual modeling, in order to remove artifacts related to the parametric way in which speech is synthesized. Although development time was limited, the results show that the system in question achieves good performance in terms of naturalness and intelligibility.

Index Terms: speech synthesis, statistical parametric speech synthesis, Blizzard Challenge.

1. Introduction

Recent advances in corpus-based speech synthesis have been responsible for many enhancements of known state-of-the-art techniques such as unit concatenation-based [1] and hidden Markov model (HMM)-based [2] approaches. In order to verify the strengths and weakness of several voice development methods, the Blizzard Challenge has been conducted since 2005 [3].

This paper describes the NICT/ATR entry for the Blizzard Challenge 2008. The submitted system is based on synthesis from HMMs and utilizes the improved excitation modeling described in [4, 5, 6] to eliminate the inherent *buzziness* and increase naturalness of the synthesized speech. The referred system represents the second participation of ATR in the Blizzard Challenge as a competing system. In 2006 the XIMERA concatenative speech synthesizer [7] was submitted [8].

The organization of this paper is as follows: Section 2 shows the characteristics of the 2008 version of the Blizzard Challenge; in Section 3 the NICT/ATR speech speech synthesis technology based on HMMs is introduced; Section 4 describes the building process for the submitted voices; and Section 5 shows and discusses the official results. The conclusions are in Section 6.

2. The Blizzard Challenge 2008

The Blizzard Challenge is an event promoted by volunteer researchers around the world in order to better understand and compare different techniques for building corpus-based speech synthesizers on the same data. The challenge itself consists of building the requested voices from the released data and synthesizing a prescribed set of test sentences. The sentences for each synthesizer are then evaluated through extensive listening tests. Volunteers, speech experts, and paid native speakers are the usual subjects who take part in the evaluation.

For the 2008 version of the Blizzard Challenge, the following databases were released:

- **UK English:** 15 hours of a male speaker released by The Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK, under a research-only-purpose license;
- **Mandarin Chinese:** 6.5 hours of a female speaker released by The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China.

By using the databases above, one, two or three of the following voices could be built:

- **Voice A:** using the full UK English database (15 hours);
- **Voice B:** using the ARCTIC subset of the UK English database (approximately 1 hour);
- **Voice C:** using the full Mandarin database (6.5 hours).

The main rules enforced in this year corresponded to the the non-utilization of external data for database alignment[1] and system homogeneity.

3. HMM parametric speech synthesis technology at NICT/ATR

Although ATR has a long tradition on corpus-based speech synthesizers using the unit concatenation approach [9, 10, 7], the entry for the Blizzard Challenge 2008 corresponded to an HMM-based speech synthesis system.

3.1. The basic system

The basic HMM-based synthesizer of ATR is very similar to the one described in [11] except for the parametric mixed excitation employed. Main differences from the baseline technique (as described in [2]) are:

- hidden semi-Markov model as the statistical machine [12];
- parameter generation considering global variance [13].

[1] This also concerned the use of data from Voice A to align Voice B.

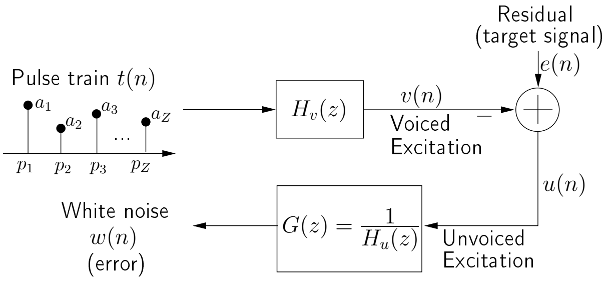

Figure 1: *Excitation model training: filters are calculated assuming an analysis-by-synthesis optimization procedure.*

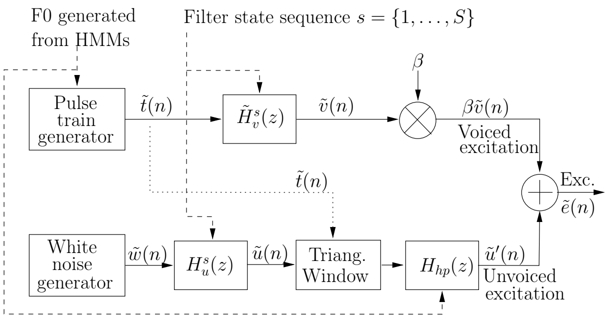

Figure 2: *During the synthesis: excitation signal is constructed from sequence of filter coefficients and* F_0.

3.2. The enhancement: better excitation modeling

In order to improve naturalness and eliminate the inherent *buzziness* the improved excitation model described in [4, 5] is utilized. This scheme is divided into two parts: the first one in which residual signal modeling is performed through an iterative optimization of some state-dependent digital filters; and the second part in which the excitation signal is constructed through the calculated filters and F_0. Each of these stages is outlined in the following (see [5] for more details).

3.2.1. *Training part*

Training for the excitation model starts with the extraction of residual signals and pitch marks from the speech database. Pulse trains are constructed from the latter ones. After that, according to a specific set of clusters of residual and pulse train segments [6], voiced and unvoiced filters for each of these clusters are calculated assuming the analysis-by-synthesis system of Figure 1, where the input is the pulse train $t(n)$ derived from pitch marks, the target is the residual signal $e(n)$, and the error is the sequence $w(n)$, assumed to be white noise. Therefore, the voiced and unvoiced filters, $H_v(z)$ and $H_u(z)$ respectively, are determined for each given cluster of segments by *whitening* the signal $w(n)$.

3.2.2. *Synthesis part*

In the synthesis part, first a *filter state sequence* is defined according to context-dependent labels derived from the input text. After that, the excitation signal is constructed by using the filter coefficients and F_0. The latter is generated from HMMs and utilized to construct the pulse train $\tilde{t}(n)$, as illustrated in the diagram of Figure 2.

Although one would expect that the synthesis diagram of Figure 2 should be exactly the one obtained from Figure 1 by simply reversing the filter $G(z)$, adjustments and empirical approximations are necessary to be taken into account. First, since $w(n)$ is not actually *whitened* during the training stage, the noise component must be attenuated during the synthesis part. This is done here by applying triangular windowing and voicing-dependent high-pass filtering. The final unvoiced excitation component is thus given by

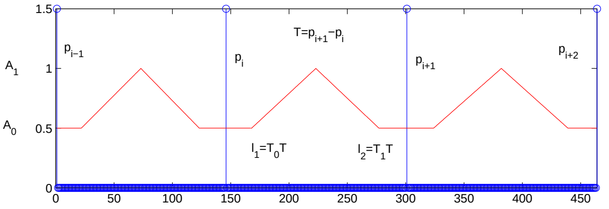

Figure 3: *Pitch-synchronous triangular window* $j(n)$ *and corresponding pulse positions* p_i.

$$\tilde{u}'(n) = \begin{cases} j(n)\left[h_u(n) * \tilde{w}(n)\right], & \text{if } F_0 = 0, \\ h_{hp}(n) * \left[j(n)\left[h_u(n) * \tilde{w}(n)\right]\right], & \text{if } F_0 > 0, \end{cases} \tag{1}$$

where $j(n)$ is a pulse-synchronous (pitch-synchronous) triangular window and $H_{hp}(z)$ is a high-pass filter with cut-off frequency f_c =2 kHz. In fact, the way in which $\tilde{u}'(n)$ is constructed is similar to the form in which the noise part of the harmonic plus noise scheme of [14] is modeled, that is: white noise filtered through an AR system followed by triangular windowing. Figure 3 shows how the window $j(n)$ is determined for each pitch interval T and pulse positions p_i. Here the parameters T_0 and T_1 are considered constant with values $T_0 = 0.15$ and $T_1 = 0.85$, the same approximation employed in [14]. The voiced component gain β is calculated from $\tilde{u}'(n)$ so as the excitation signal $\tilde{e}(n)$ has power one for every 5-ms frame, i.e.,

$$\beta = \sqrt{1 - \frac{1}{N}\sum_{n=1}^{N} \tilde{u}'^2(n)}. \tag{2}$$

In this case $\tilde{v}(n)$ is assumed to have power one in each frame, which is a good approximation since $\tilde{t}(n)$ has power one at pitch interval and $\tilde{H}_v^s(z)$ is normalized in energy. The factor N is the number of samples in each frame.

4. Voice building for the Blizzard Challenge

Voice building process for the Blizzard Challenge 2008 can be divided into four steps: (1) database segmentation; (2) feature extraction and labeling; (3) speech parameter extraction; (4) synthesizer and excitation model training. In the next sections each of these parts is covered with details.

4.1. Database segmentation

Database segmentation was performed differently for English and Chinese. To fulfill one of the rules enforced by the Blizzard Challenge this year, no external data was utilized to segment the voices. Thus, Voice B was segmented using solely the ARCTIC subset of the English database.

4.1.1. *Database segmentation for voices A and B*

Pause detection and database alignment for voices A and B were conducted as follows. The entire procedure had as input the Festival [15] *utterances* released to all interested participants. Firstly, phonetic labels and word transcriptions were derived from these files. By using the phonetic labels tied-state triphone HMMs were trained. After that, pause detection was performed by decoding the entire database using a constrained word recognition network, in which pause models could be present between any adjacent pair of words, and a word pronunciation dictionary derived from the provided Unilex lexicon [16]. The word recognition network was derived from the word transcriptions. Once pauses were placed at the appropriate places, HMMs were trained again using the newly constructed phonetic labels. These final acoustic models were utilized to segment the database through forced Viterbi alignment, using the word pronunciation dictionary and new word transcriptions, with pauses located at the appropriate places. Table 1 shows the characteristics of the acoustic models of the aligners used to segment voices A and B.

Table 1: *Characteristics of the aligners for voices A and B.*

Acoustic models	Tied-state triphones
HMM topology	Left-to-right no-skip 5 states
Acoustic features	12-th order MFCC with c_0 plus Δ and $\Delta\Delta$
Output distribution	5-mixture Gaussian

4.1.2. Database segmentation for Voice C

For Voice C, segmentation and pause detection were performed at once through Viterbi alignment. First, phonetic labels were derived from the released sentence prompts through the Chinese XIMERA text processing front-end [7]. The labels were then used to train monophone HMMs with different number of states, all of them left-to-right topology except for a *tee-model* utilized as *pauses between words*. The trained HMMs were eventually utilized to force-align the entire database, considering that pauses existed between any sequence of two words. Eventually, pauses which had duration shorter than a predetermined threshold were eliminated. Table 2 shows the characteristics of the acoustic models of the aligner for Voice C.

Table 2: *Characteristics of the aligner for Voice C.*

Acoustic models	Monophones
HMM topology	Left-to-right no-skip with different number of states and one *tee-model*
Acoustic features	12-th order MFCC with energy plus Δ and $\Delta\Delta$
Output distribution	Single Gaussian

4.2. Feature extraction and labeling

4.2.1. Voices A and B

Contextual features for voices A and B were derived from the provided Festival utterances. However, before feature extraction, the referred files were modified in order to insert pause models at appropriate places, according to the phonetic labels derived by the pause detection procedure described in Section 4.1.1. The modification did not affect the utterances in terms of features or phonetic content. This procedure resulted in better full context labels.

In addition to all the features listed in [17], the following ones included in the released utterances were also employed: *emph* and *b-tone*.

4.2.2. Voice C

Contextual factors for Voice C were extracted through the Chinese XIMERA text processing part [7]. In fact, the only non-speech released information which was actually utilized for building Voice C corresponded to the sentence prompts.

4.3. Speech parameter extraction

The speech parameters extracted to train the synthesizers and excitation models corresponded to: (1) spectral parameters; (2) F_0; (3) pitch marks; (4) residual sequences.

4.3.1. Spectral parameters and F_0

Spectral parameters and F_0 were calculated from speech at every 5 ms. Spectral parameters corresponded to mel-cepstral coefficients that can directly synthesize speech through the utilization of the mel log approximation (MLSA) filter [18]. Based on analysis-synthesis experiments, where properties related to the residual extraction for the excitation model [4] were also taken into account, the number of mel-cepstral coefficients in each 5-ms frame for Voice C was 19 whereas for voices A and B was 25. Mel-cepstral analysis was performed through 25-ms Hamming windows with 20-ms overlaps. F_0 was extracted by using the Snack Sound Toolkit [19].

After training the synthesizers, mel-cepstral analysis using a smoothed periodogram, as described in [11], was performed and the models of the synthesizer mapped onto the obtained coefficients to create new HMMs. The reason for this is that mel-cepstral coefficients derived from smoothed periodogram seem to produce better synthesized speech through the application of the global variance-based parameter generation approach [13]. However, one might wonder why this sort of coefficients were not utilized in the first place to train the synthesizer. The explanation is that these coefficients do not present good properties in terms of residual extraction by inverse filtering. Thus, it was a matter of consistency. The spectral parameters used to train the synthesizers should be the ones utilized to extract residual signals and define filter states for the excitation models.

4.3.2. Pitch marks and residual signals

Pitch marks and residual sequences were necessary for training the excitation models. The former ones were obtained by the Snack Sound Toolkit [19] while residual signals were derived from the original speech waveforms through inverse filtering using the MLSA structure.

4.4. Synthesizer and excitation model training

Training of the synthesizer and corresponding excitation model for Voice A took approximately 3 weeks on a DualCoreXeon 2.33GHz 32GB machine. For the other voices computational time was considerably smaller.

States for the excitation models were defined according to the phonetic decision tree approach [6]. In this method, states are derived from the corresponding synthesizers by performing context clustering on the distributions of mel-cepstral coefficients using solely phonetic questions and large stopping criterion. Thus, in this way, gross phonetic information are conveyed by the filter states. Although the bottom-up clustering procedure described in [6] presents better performance this approach was utilized because development time was crucial. In total, 232, 130 and 257 filter states were produced for voices A, B and C, respectively.

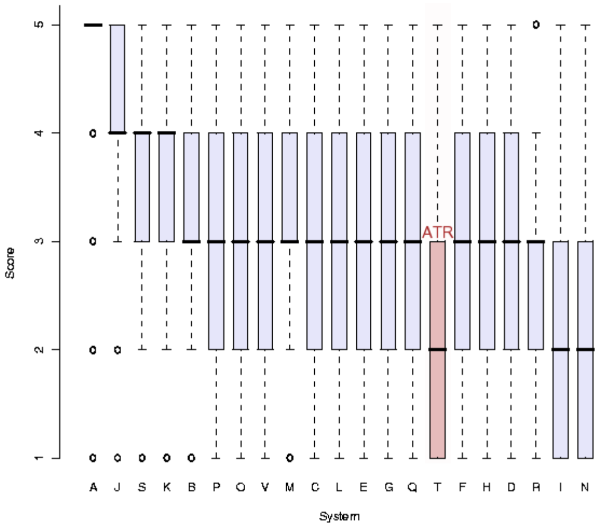

Figure 4: *Similarity to the original speaker for Voice A considering all the listeners.*

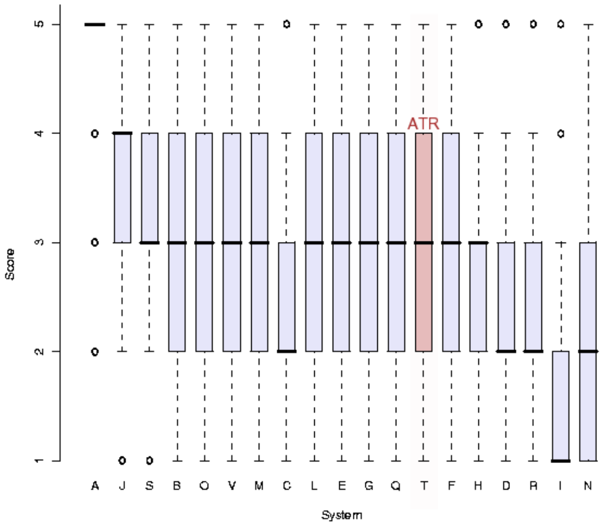

Figure 5: *Similarity to the original speaker for Voice B considering all the listeners.*

5. Results of the official listening tests

The experimental tests conducted during the Blizzard Challenge 2008 evaluated the submitted systems according to three different criteria:

1. similarity to the original speaker, on a scale from 1 - "Sounds like a totally different person" to 5 - "Sounds exactly the same person";
2. naturalness, on a scale from 1 - "Completely unnatural" to 5 - "Completely natural";
3. word error rate (WER).

Because the scales utilized to evaluate criteria 1 and 2 are ordinal, similarity and naturalness scores are expressed in terms of medians, and comparison among the systems is conducted through inspection of box-plots. On the other hand, the internal scale utilized to evaluate criterion 3 allows comparison of means [20].

In all the box-plots of this section the NICT/ATR system corresponds to letter "T" and original speech to letter "A".

5.1. Voices A and B

5.1.1. *Similarity to the original speaker*

Figures 4 and 5 show box-plots of similarity scores for voices A and B, respectively, considering all the speakers. It can be noticed that the submitted system obtains better performance for Voice B. One possible explanation for these results is that for small databases unit concatenation systems tend to synthesize speech with more artifacts. Consequently, in this case, HMM synthesizers such as the ATR submission tend to stand out among the other systems, giving the impression that they produce speech which sounds closer to the original speaker.

Table 3 shows similarity scores according to each group of listeners. The results show that the only difference occurs for the *Speech Experts* group, in which Voice B was considered better than Voice A. Therefore, it becomes more evident that the increase in database might have resulted in improvement of the unit concatenation-based entries, giving the sensation that they sound more similar to the original speaker when compared with the submitted system. This effect was thus more easily noticed by speech synthesis experts.

Table 3: *Similarity scores for voices A and B according to each listener group.*

Voice	All	UK students	Volunteers	Speech experts	Indian students
A	2	2	3	2	2
B	3	2	3	3	2

5.1.2. *Naturalness degree*

Figures 7 and 6 show the naturalness scores considering all the listeners for voice A and B, respectively. In this case the results for Voice A are apparently better compared to the ones obtained in the similarity to original speaker case. This emphasizes perhaps a weak point of the submitted system: it produces good synthesized speech that does not sound very close to the original waveforms.

Table 4 shows naturalness scores for voices A and B according to each listener group. The results are exactly the same.

Table 4: *Naturalness scores for voices A and B according to each listener group.*

Voice	All	UK students	Volunteers	Speech experts	Indian students
A	3	2	3	3	3
B	3	2	3	3	3

5.1.3. *Word error rate*

Figure 8 shows the WER for voices A and B considering UK students (actual natives speakers of voices A and B) paid to participate in test. The ATR entry achieves great performance for this criterion. One interesting aspect is that Voice A achieves an intelligibility degree which is higher than that for natural speech (entry "A"). Considering all the listeners together, WER for voices A and B were 14% and 29%, respectively, for the submitted system and 14% for natural speech.

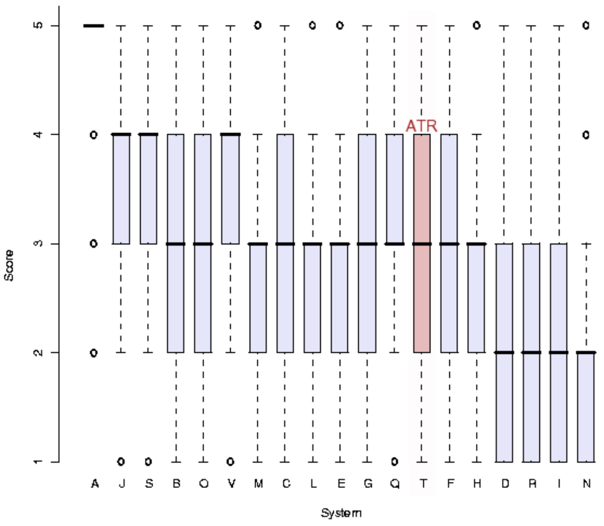

Figure 6: *Naturalness scores for Voice B considering all the listeners.*

5.2. Voice C

In a general the results achieved by the Mandarin entry were better than the ones obtained by voices A and B. Voice C got very good numbers concerning naturalness and WER.

Figures 9 and 10 show respectively box-plots of similarity to the original speaker and naturalness for Voice C considering all the speakers. Naturalness score was 4.0 whereas 3.0 was obtained in the similarity criterion. Therefore, for the Mandarin voice one can again notice for the submitted system that in spite of producing close-to-natural speech the synthesized waveform does not sound very similar to the original speaker.

Table 5 shows similarity and naturalness scores for Voice C according to each listener group. The ATR system achieves great results in terms of naturalness degree among paid native speakers of Chinese. Like in the Voice A case, speech experts gave a 2.0 score for similarity.

Table 5: *Similarity and naturalness scores for Voice C according to each listener group.*

Crit.	All	Natives in China	Natives in UK	Volunteers	Speech experts
Sim.	3	3	3	3	2
Nat.	4	4	4	3	3

Table 6 shows the character error rate (CER), Pinyin (without tone) error rate (PER), and Pinyin (with tone) error rate (PTER) for Voice C according to each listener group. The results were considered very good.

Table 6: *CER, PER and PTER for Voice C (%). Results for natural speech are in parentheses.*

Group	CER	PER	PTER
All	17.0 (13)	9.5 (5.8)	11.7 (8)
Natives in China	22 (18)	11.9 (7.7)	15 (12)
Natives in UK	16.3 (6.8)	9.1 (3.8)	10.1 (4.2)
Volunteers	8.1 (4.1)	1.8 (1.4)	3.2 (2.3)
Speech experts	11.1 (8.9)	6.9 (4.6)	7.8 (5.2)

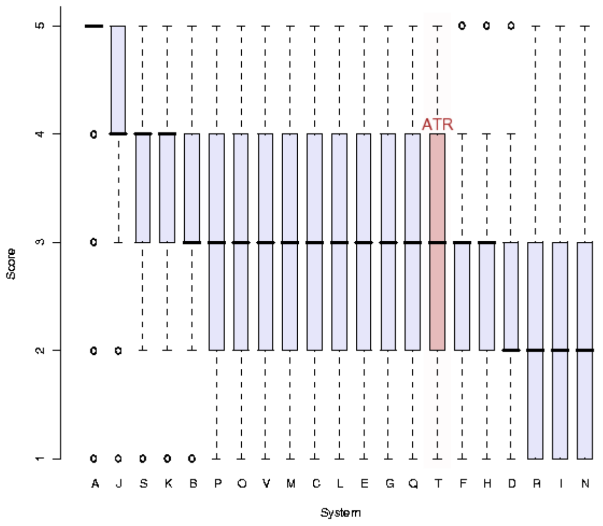

Figure 7: *Naturalness scores for Voice A considering all the listeners.*

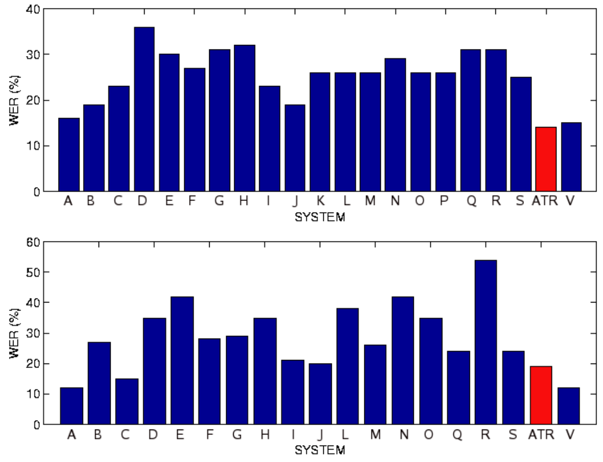

Figure 8: *WER according to paid UK listeners for voices A (top) and B (bottom). Entry "A" corresponds to original speech.*

5.3. Discussion

Despite the fact that the submitted voices obtained good performance in terms of naturalness and intelligibility, a more adequate spectral parameterization could have resulted in better scores for the criterion similarity to the original speaker. The current spectral parameters were chosen in order to keep consistency between the synthesizers and corresponding excitation models. Although high-order mel-cepstral coefficients extracted as shown in [11] represent better choice for HMM synthesizers since they enable a better reproduction of high frequency components, they do not present good characteristics in terms of residual extraction. Owing to this problem the approach described in 4.3.1 was employed. Eventually, it was verified that if the spectral parameters used to train the synthesizers were higher-order mel-cepstral coefficients extracted as described in [11], and the ones employed to extract residual signals were lower-order mel-cepstral coefficients obtained as [18], synthesized speech would sound more *clean* despite the inconsistency. However, since time was limited the voices whose training had already started had to be submitted.

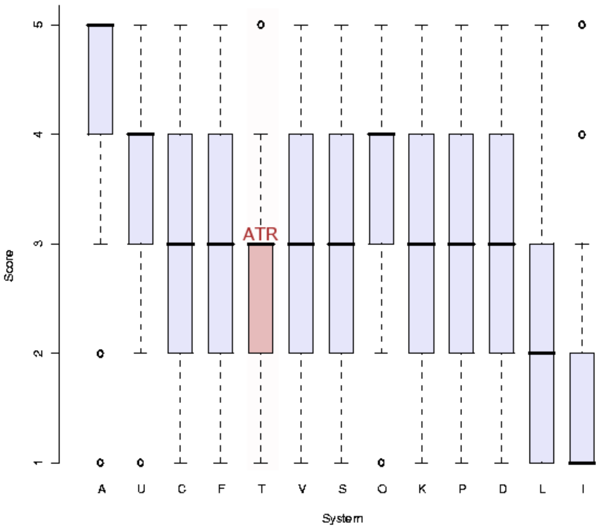

Figure 9: *Similarity to the original speaker for Voice C considering all the listeners.*

As positive aspects from the participation we could mention the development of the approach utilized for database segmentation, the hacking which enabled the inclusion of pause models in the Festival utterances, and the simple mistakes which should not be done when voices are requested to be built in a limited period of time.

6. Conclusions

This paper described the NICT/ATR entry for the Blizzard Challenge 2008. The system is based on the statistical parametric speech synthesis technology and presents as enhancement the utilization of an excitation model based on analysis-by-synthesis training using residual as target signals. In general, good results in terms of naturalness and intelligibility degrees were obtained.

7. Acknowledgements

The authors would like to thank Prof. Minoru Tsuzaki for the fruitful discussions.

8. References

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

[2] 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. of EUROSPEECH*, 1999.

[3] http://festvox.org/blizzard/index.html.

[4] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in *Proc. of INTERSPEECH*, 2007.

[5] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation approach for HMM-based speech synthesis based on residual modeling," in *Proc. of ISCA Speech Synthesis Workshop*, 2007.

[6] R. Maia, T. Toda, K. Tokuda, S. Sakai, and S. Nakamura, "On the state definition for an excitation model in HMM-based speech synthesis," in *Proc. of ICASSP*, 2008.

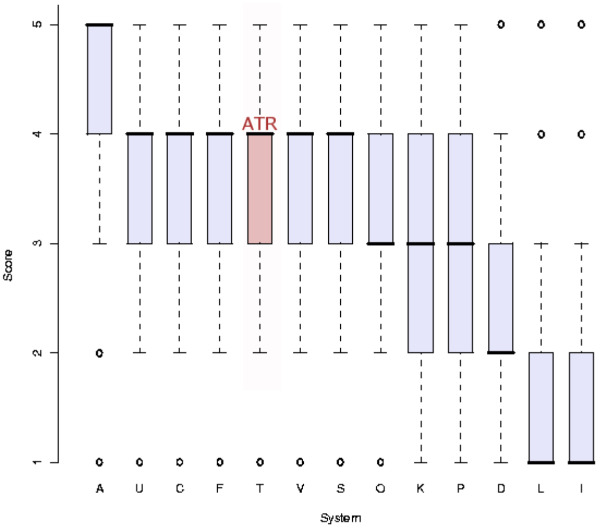

Figure 10: *Naturalness for Voice C considering all the listeners.*

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

[8] T. Toda, H. Kawai, T. Hirai, J. Ni, N. Nishizawa, J. Yamagishi, M. Tsuzaki, K. Tokuda, and S. Nakamura, "Developing a test bed of English text-to-speech system XIMERA for the Blizzard Challenge 2006," in *Proc. of Blizzard Challenge Workshop*, 2006.

[9] Y. Sagisaga, K. Kaiki, and N. Iwahashi, "ATR ν-TALK speech synthesis system," in *Proc. of ICSLP*, 1992.

[10] 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.

[11] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis for Blizzard Challenge 2005," in *Proc. of EUROSPEECH*, 2005.

[12] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf & Syst.*, vol. E90-D, May 2007.

[13] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, pp. 816–824, May 2007.

[14] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. of EUROSPEECH*, 1994.

[15] http://festvox.org/festival.

[16] http://www.cstr.ed.ac.uk/projects/unisyn.

[17] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis applied to English," in *Proc. of IEEE Speech Synthesis Workshop*, 2002.

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

[19] http://www.speech.kth.se/snack.

[20] 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. of the Blizzard Challenge Workshop*, 2007.