The CSTR/EMIME HTS System for Blizzard Challenge 2010

Junichi Yamagishi, Oliver Watts

The Centre for Speech Technology Research,
University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom
jyamagis@inf.ed.ac.uk

Abstract

In the 2010 Blizzard Challenge, we focused on improving steps relating to feature extraction and labeling in the procedures for training HMM-based speech synthesis systems. New auditory scales were used for spectral features and F0 representation. We have also adopted finer frequency bands motivated by an auditory-scale for aperiodicity measures, which determine the level of noise in each band for mixed excitation. Further for tighter coupling of the HMM training and automatic labeling processes, we have studied methods for stepwise bootstrap training. The listeners' evaluation scores were much better than those of HTS-benchmark systems. More importantly, we can see some improvements even in speaker similarity, which was known to be the acknowledged weakness of this method. In fact, speaker similarity is not a weak point of this method on the tasks using smaller databases. In terms of naturalness, the new systems outperformed or competed with unit selection systems regardless of the size of speech databases used and moreover competed with hybrid systems on smaller databases.

Index Terms: speech synthesis, HMM, average voice, speaker adaptation

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has become a mainstream method for speech synthesis because of its natural-sounding synthetic speech and its flexibility. It has the potential to go far beyond conventional unit-selection type methods because the speech is generated from a parametric model, which can be modified in various ways. Since HMM-based speech synthesis now has a history of more than 10 years, it is worth briefly summarising progress to date. Research on HMM-based speech synthesis started with the development of algorithms for generating smooth and natural parameter trajectories from HMMs [2]. Next, to simultaneously model the excitation parameters of speech as well as spectral parameters, the multi-space probability distribution (MSD) HMM [3] was developed. To simultaneously model the duration for the spectral and excitation components of the model, the MSD hidden semi-Markov model (MSD-HSMM) [4] was developed. These basic systems employed a mel-cepstral vocoder with simple pulse or noise excitation, resulting in synthetic speech with a "buzzy" quality. To reduce buzziness, a more sophisticated excitation technique, called *mixed excitation* was integrated into the basic system to replace the simple pulse or noise excitation [5]. A high-quality speech vocoding method called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [6] was also used, in conjunction with mixed excitation [7]. STRAIGHT explicitly uses F_0 information to remove periodic components from the estimated spectrum, i.e., it interpolates missing frequency components considering neighboring harmonic components based on an F_0 adaptive smoothing process over a time-frequency region. This enables the generation of better spectral parameters and consequently more natural synthetic speech. Still, all these basic systems had a serious shortcoming: the trajectories generated from the HMMs were excessively smooth due to statistical processing; over-smooth spectral parameters result in synthetic speech with a "muffled" quality which lacks the "sharpness" or "transparency" so easily achieved by concatenative methods. To alleviate this problem, a parameter generation algorithm that considers the global variance (GV) of the trajectory being generated was proposed [8]. In order to reflect within-frame correlations and optimize all the acoustic feature dimensions together, semi-tied covariance (STC) modeling [9] was employed to enable the use of full-covariance Gaussians in the HSMMs [10]. Taken together, these modest incremental improvements have had a cumulative effect. Compared with early buzzy and muffled HMM-based speech synthesis, the latest systems have a dramatically improved quality. They have exhibited good performance in the Blizzard Challenges [11, 12, 13, 14].

The systems mentioned above are *speaker-dependent*. In parallel, we have also been developing a *speaker-adaptive* approach in which "average voice models" are created using data from several speakers. The average voice models may then be adapted using speech from a target speaker (e.g. [15]). To adapt spectral, excitation and duration parameters within the same framework, an extended MLLR adaptation algorithm for the MSD-HSMM has recently been proposed [16]. A more robust and advanced adaptation algorithm called constrained structural maximum a posteriori linear regression (CSMAPLR) has been proposed [15]. We have also developed several techniques for training the average voice model, such as a speaker-adaptive training (SAT) algorithm [17]. To further explore the potential of HMM-based speech synthesis, for the 2007 Blizzard Challenge we combined these advances in the speaker-adaptive approach with our current speaker-dependent system that employs STRAIGHT, mixed excitation, HSMMs, GV, and full-covariance modeling [18]. In the 2008 Blizzard Challenge the same speaker-adaptive approach was used, but the model was trained on more data using a more efficient algorithm and a higher order cepstral analysis was employed [19]. In the 2009 Blizzard Challenge unsupervised and noise-robust versions of the 2008 systems were investigated [20].

In the Blizzard 2010 challenge we adopted a speaker-dependent approach for task EH1 where 4 hours of speech data were to be used and a speaker-adaptive approach for tasks EH2 and ES1 where 1 hour and 100 utterances of speech data respectively were to be used. Systems entered in the 2006 and 2008 Challenges were adopted as the basis for speaker dependent and speaker adaptive systems respectively, and the following new improvements to feature extraction and labelling were incorporated into each system:

- New auditory scales for spectral features and F0
- Fine frequency bands motivated by an auditory scale for aperiodicity measures
- Tighter coupling of the HMM training and automatic labeling processes

In the following sections, we explain the details of these techniques and analyse the results in the 2010 challenge.

2. New acoustic features for CSTR HMM-based Speech Synthesis Systems

Our previous HMM-based speech synthesis system models three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [6]) mel-cepstral vocoder with mixed excitation, that is, the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures, as static feature vectors for the HMMs.

2.1. From mel-cepstrum to Bark cepstrum

In mel-cepstral analysis using all-pass filter [21], the vocal tract transfer function $H(z)$ is modelled by M-th order mel-cepstral coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\top$ as follows:

$$H(z) = \exp \boldsymbol{c}^\top \tilde{\boldsymbol{z}} = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, \qquad (1)$$

where $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^\top$. $\tilde{z}^{-1}$ is defined by a first-order all-pass (bilinear) function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \qquad (2)$$

and the warped frequency scale $\beta(\omega)$ is given as its phase response:

$$\beta(\omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}. \qquad (3)$$

The phase response $\beta(\omega)$ gives a good approximation to an auditory frequency scale with an appropriate choice of α.

In general, as the sampling frequency increases, the differences between different auditory frequency scales such as the Mel and Bark scales [22] implemented using a first-order all-pass function become greater. Therefore we tested Bark scale in this challenge. In [23], Smith and Abel define the optimal α (in a least-squares sense) for Bark scale as follows:

$$\alpha_{\mathrm{Bark}} = 0.8517\sqrt{\arctan(0.06583\, f_s)} - 0.1916 \qquad (4)$$

where f_s is the waveform sampling frequency.

2.2. From $\log F_0$ to pitch in mel

Recently we proposed a generalised logarithmic transform of fundamental frequency [24]. In this challenge, we test a different psycho-acoustic transform of fundamental frequency. Rabiner and Schafer define pitch and fundamental frequency as follows [25]:

> Pitch is a subjective attribute of sound that is related to the fundamental frequency of the sound, which is a physical attribute of the acoustic waveform.

Stevens and Volkmann [26] show that the relation pitch measures on the mel-scale and frequency (of a pure tone) is approximated by

$$\mathrm{Pitch}\ [\mathrm{mel}] = 1127 \log\left(1 + \frac{f}{700}\right). \qquad (5)$$

We use the pitch in mel scale and its delta and delta-delta as observation vectors for F0 modeling.

2.3. Auditory-scale motivated frequency-bands for aperiodicity measures

In the conventional systems, five frequency sub-bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz) [7] were used for aperiodicity measures in a similar way to MELP coding [27]. In this challenge, we tested frequency-bands for aperiodicity measures motivated by an auditory scale instead of the five frequency-bands. The Bark critical band ratio can be converted from frequency approximately as follows [28]:

$$\mathrm{Critical\,band\,rate}\ [\mathrm{bark}] = \frac{26.81 f}{1960 + f} - 0.53. \qquad (6)$$

As in the original paper on the Bark scale [22], we directly used the critical bands as frequency-bands for aperiodicity measures (i.e. regard the integers of critical band ratio as edge of frequency bands). This results in 25 frequency bands with non-linear varying bandwidth for speech sampled at 48 kHz sampling frequency. Alternatively it is also possible to warp frequency non-linearly for aperiodicity measures using the equation above and to use equalized frequency bands.

3. Stepwise bootstrap training including regeneration of rich context-dependent labels

Performing automatic labelling of contexts requires well-trained HMMs, whereas training HMMs completely automatically from scratch also requires automatically annotated contextual labels including not only the phoneme sequence but also some additional information such as vowel reduction. Therefore we should couple the labeling process and HMM training process more tightly and should optimise both processes at the same time.

Oura et. al. proposed a training method to use N-best contextual labels for a single utterance in HMM-based speech synthesis [29]. In this challenge, we adopted a simpler and more practical solution, that is, stepwise bootstrap training. First we perform the initial labelling and train HMMs. Then using the trained HMMs, we perform automatic labelling of time alignment information, vowel reduction, and pause detection. The labelling results in a new set of HTS full-context labels having refined time alignment information which enables us to train new HMMs from scratch in a bootstrap manner.

From the speech database and labels that include an initial phoneme segmentation, we first train a set of speaker-dependent context-dependent multi-stream left-to-right MSD-HSMMs [7]. To begin with, monophone MSD-HSMMs are trained from the initial segmentation, converted to context-dependent MSD-HSMMs and re-estimated. Then, decision-tree-based context clustering is applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again. The clustering processes are repeated until convergence of likelihood improvements. Then we perform automatic labelling using weighted finite-state transducers (WFST). The whole process is further repeated using regenerated labels refined with the trained models in a bootstrap fashion. The inner loop for iterative clustering was followed 10 times and the outer loop for refinement of the automatic labels using WFST was followed 10 times. The same procedures can be used for speaker-adaptive systems.

Note that the final labels differ from the initial labels in terms of not only time alignment but also contexts. For instance, the final contextual labels had only half the number of pauses of the initial labels.

4. The Blizzard Challenge 2010

The Blizzard Challenge is an annual evaluation of corpus-based speech synthesis systems, in which each participating team builds a synthetic voice from common training data, then synthesizes a set of test sentences. Listening tests are adopted to evaluate the systems in term of naturalness, similarity to original speaker and intelligibility. In the Blizzard Challenge 2010, two English speech databases consisting of 4 hours of speech uttered by a British male speaker RJS and 1 hour of speech data uttered by a different British male speaker ROGER, and a Mandarin speech database consisting of about 9.5 hours of speech uttered by a Beijing female speaker were released. We entered only the English evaluation this year.

The initial contextual labels for the data were automatically generated using Unilex [30] and Festival's Multisyn module, with no further modification. The English phonetic, linguistic and prosodic context factors used were similar to those in [31]. To investigate the effect of corpus size, three systems were built: one built using 4 hours of speech data from the RJS database (EH1 task), the second one built using 1 hour of speech data from the ROGER database (EH2 task), and the third one built using the first 100 sentences (corresponding to 6 minutes) of the ROGER database (ES1 task). Note that the ROGER voices were adapted from the RJS model instead of an average voice model since there was not enough speech data sampled at 48kHz for training an average voice model. All the feature analysis steps were carried out using the 48 kHz speech data and downsampled only after vocoding speech. The use of higher sampling frequency and details of its use in training have been reported in [24, 32].

4.1. Listening Tests

English synthetic speech was generated for a set of 468 test sentences, including 368 sentences from broadcast, news, and novel genres (used to evaluate naturalness and similarity) and 100 semantically unpredictable sentences (used to evaluate intelligibility). To evaluate naturalness and similarity, 5-point mean opinion score (MOS) and comparison category rating (CCR) tests were conducted. The scale for the MOS test was 5 for "completely natural" and 1 for "completely unnatural". The scale for the CCR tests was 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person" compared to a few natural example sentences from the reference speaker. To evaluate intelligibility, the subjects were asked to transcribe semantically unpredictable sentences and the average word error rates (WER) were calculated from these transcripts. The evaluations were conducted over a six week period via the internet.

4.2. Experimental Results

Figures 1–3 show the evaluation results on naturalness in the EH1 (4 hours), EH2 (1 hour) and ES1 tasks (6 min), respectively. Figures 4–6 show the evaluation results on speaker similarity in EH1 task (4 hours), EH2 task (1 hour) and ES1 task (6 min), respectively. In these figures, systems "V" corresponds to the 2010 CSTR/EMIME HTS system. "A", "B" and "C" correspond to real speech, the Festival "Multisyn" benchmark speech synthesis system [33] and the HTS benchmark system [7], respectively. The Festival system uses a conventional unit-selection method. The HTS Benchmark system is a standard statistical parametric system using speaker-dependent HMMs, which can be trained from scratch by using HTS toolkit version 2.1 and STRAIGHT. This system was highly rated in terms of naturalness and intelligibility in the 2005 Blizzard Challenge. Further "M", "J", and "T" are hybrid systems of unit selection and HTS methods.

4.3. Naturalness

We note several interesting findings and system improvements in the results:

EH1 task (Figure 1)
Our new HTS system "V" was not as good as the hybrid type systems "M", "J", and "T". However, there was no significant difference between the Festival benchmark unit selection system "B" and our new HTS system. Moreover the new HTS system was found to be significantly better than the HTS benchmark system "C".

EH2 task, (Figure 2)
Our new HTS system "V" was the second best in the EH2 takes where a hybrid system "M" was the best. The new HTS system was found to be significantly better than both the benchmark systems "B" and "C".

ES1 task (Figure 3)
In the ES1 task, our new HTS system "V" and system "M" were the equal best.

4.4. Speaker similarity

We can see several improvements in speaker similarity, which was the acknowledged weakness of the HMM-based speech synthesis method [19]:

EH1 task (Figure 4)
Our new HTS system "V" was rated as the average: The Festival benchmark unit selection system "B" was better than our new HTS system "V". However, the new HTS system was found to be significantly better than the HTS benchmark system "C".

EH2 task, (Figure 5)
Our new HTS system "V", and hybrid systems "M" and "J" were the equal best in the EH2 takes. The new HTS system was found to be significantly better than both the benchmark systems "B" and "C".

ES1 task (Figure 6)
In the ES1 task, our new HTS system "V" was the best.

Since lower speaker similarity of HMM-based speech synthesis was known to be its acknowledged weakness, this is an important achievement for us. In fact, speaker similarity is no longer a weak point of this method in tasks EH2 and ES1.

4.5. Comparison by speech synthesis methods

We can summarise the results above by speech synthesis methods below:

Comparison with the HTS benchmark system
The new HTS system "V" got significant improvements compared to HTS benchmark systems "C" in terms of both naturalness and similarity in all tasks (EH1, EH2, and ES1).

Comparison with the Festival benchmark system
Compared to Festival unit-selection benchmark system "B", the new HTS system "V" are found to be equally good in terms of naturalness and to be worse only in terms of speaker similarity in the EH1 task. On the other hand, in the EH2 task, the new system was rated as significantly better then the Festival benchmark system in terms of both naturalness and similarity.

Comparison with the hybrid systems "M" and "J"
Compared to hybrid speech synthesis systems "M" and "J", the new HTS system "V" are found to be worse in terms of both naturalness and similarity in the EH1 task. In the EH2 task, the new system was rated as good as the hybrid systems in terms of speaker similarity. However, its naturalness was found to be worse than system "M", but as good as system "J".

5. Conclusions

In the 2010 Blizzard Challenge, we experimented with improvements to feature extraction and labeling steps in the training of HMM-based speech synthesisers. New auditory scales were used for spectral features and F0. We also adopted finer frequency bands motivated by an auditory-scale for aperiodicity measures. In addition, to tighter couple the HMM training and automatic labelling processes, we tried a method of stepwise bootstrap training. Listeners' evaluation scores were much better than those of the HTS-benchmark systems. More importantly, we can see some improvements even in speaker similarity, which was this method's acknowledged weakness. In fact, speaker similarity is not a weak point of this method in smaller tasks such as EH2 any more and its score was as good as those of hybrid systems. This is an important achievement for us. In terms of naturalness, the new systems outperformed or competed with unit selection systems regardless of the size of speech databases used and moreover competed with hybrid systems on smaller databases. However, on larger speech databases, ratings for speaker similarity of the voices do not reach high enough levels and thus we need to improve this aspect of their performance.

6. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org).

This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

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, Sep. 1999, pp. 2374–2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorigthms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, Jun. 2000, pp. 1315–1318.

[3] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[4] 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, no. 5, pp. 825–834, May 2007.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. EUROSPEECH 2001*, Sep. 2001, pp. 2263 °!'2266.

[6] H. Kawahara, I. Masuda-Katsuse, and A. 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, pp. 187–207, 1999.

[7] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[8] 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, no. 5, pp. 816–824, May 2007.

[9] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, Mar. 1999.

[10] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *IEICE Trans. Inf. & Syst.*, vol. E91-D, no. 6, pp. 1764–1773, Jun. 2008.

[11] A. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. EUROSPEECH 2005*, Sep. 2005, pp. 77–80.

[12] C. Bennett and A. Black, "The blizzard challenge 2006," in *Proc. Blizzard Challenge 2006*, Sep. 2006.

[13] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Bonn, Germany, Aug. 2007.

[14] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, September 2008.

[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[16] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[17] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[18] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[19] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, Sep. 2008.

[20] J. Yamagishi, M. Lincoln, S. King, J. Dines, M. Gibson, J. Tian, and Y. Guan, "Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework," in *Proc. Blizzard Challenge Workshop*, Edinburgh, U.K., Sep. 2009.

[21] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito, and S. Imai, "Spectral estimation of speech based on mel-cepstral representation," *IEICE Trans. Fundamentals*, vol. J74-A, no. 8, pp. 1240–1248, Aug. 1991, in Japanese.

[22] E. Zwicker and B. Scharf, "A model of loudness summation," *Psych. Rev.*, vol. 72, pp. 2–26, 1965.

[23] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. on Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Jul. 1999.

[24] J. Yamagishi and S. King, "Simple methods for improving speaker-similarity of hmm-based speech synthesis," in *Proc. ICASSP 2010*, Dallas, TX, Mar. 2010, pp. 4610–4613.

[25] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Found. Trends Signal Process.*, vol. 1, pp. 1–194, January 2007.

[26] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. pp. 329–353, 1940.

[27] A. McCree and T. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. on Speech Audio Process.*, vol. 3, no. 4, pp. 242–250, Jul. 1995.

[28] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 97–100, 1990. [Online]. Available: http://link.aip.org/link/?JAS/88/97/1

[29] K. Oura, Y. Nankaku, T. Toda, K. Tokuda, R. Maia, S. Sakai, and S. Nakamura, "Simultaneous acoustic, prosodic, and phrasing model training for TTS conversion systems," in *Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on*, 2008, pp. 1–4.

[30] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. EUROSPEECH-99*, vol. 2, Budapest, Hungary, Sep. 1999, pp. 823–826.

[31] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[32] A. Stan, J. Yamagishi, S. King, and M. Aylett, "The romanian speech synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate," *Speech Communication*, vol. ??, no. ?, pp. ??–??, 2011, (under review).

[33] K. Richmond, V. Strom, R. Clark, J. Yamagishi, and S. Fitt, "Festival Multisyn voices for the 2007 Blizzard Challenge," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Bonn, Germany, Aug. 2007.

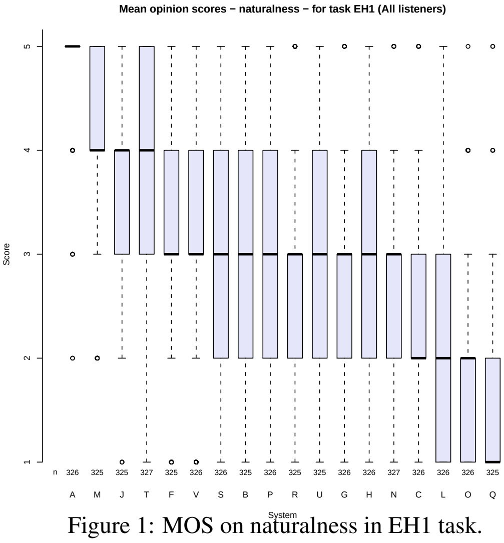

Figure 1: MOS on naturalness in EH1 task.

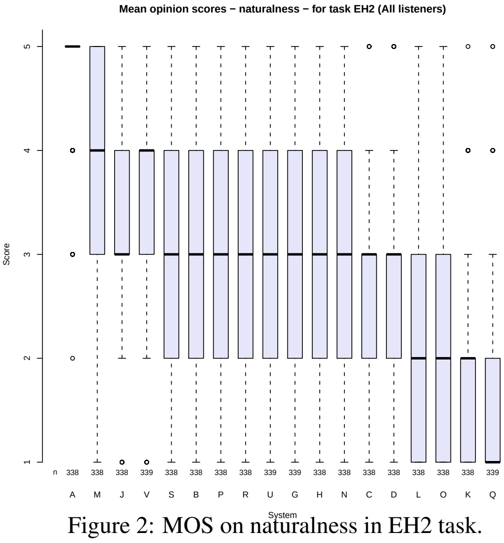

Figure 2: MOS on naturalness in EH2 task.

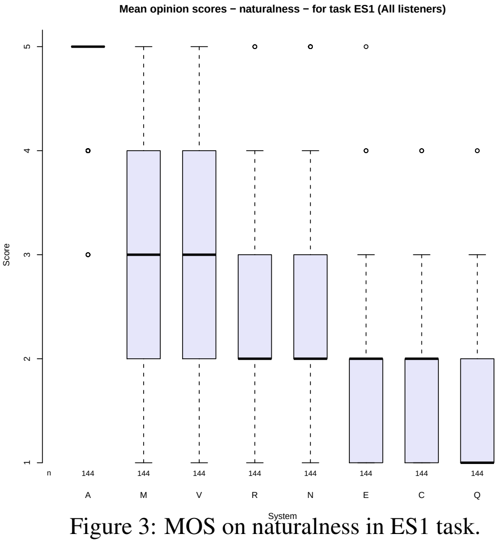

Figure 3: MOS on naturalness in ES1 task.

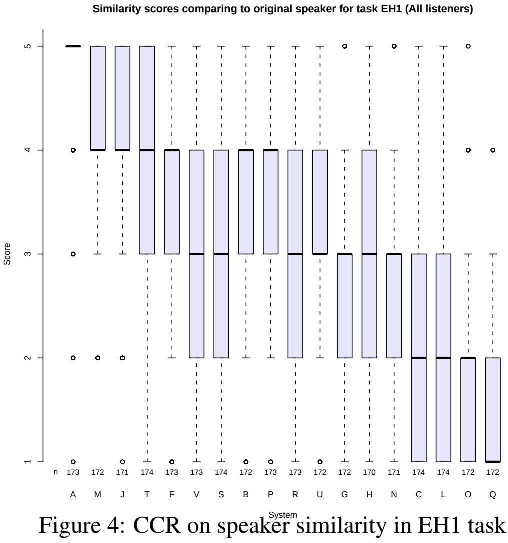

Figure 4: CCR on speaker similarity in EH1 task.

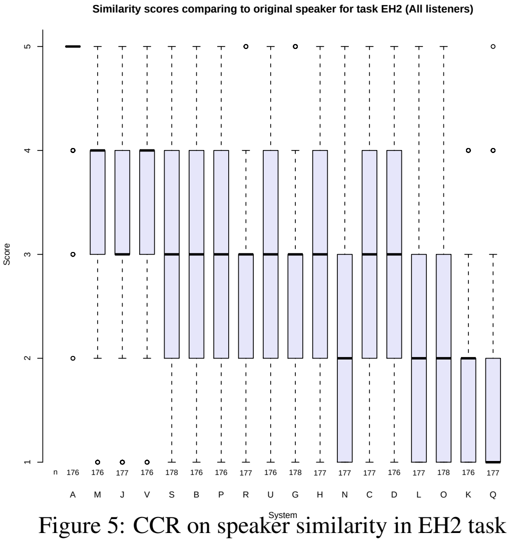

Figure 5: CCR on speaker similarity in EH2 task.

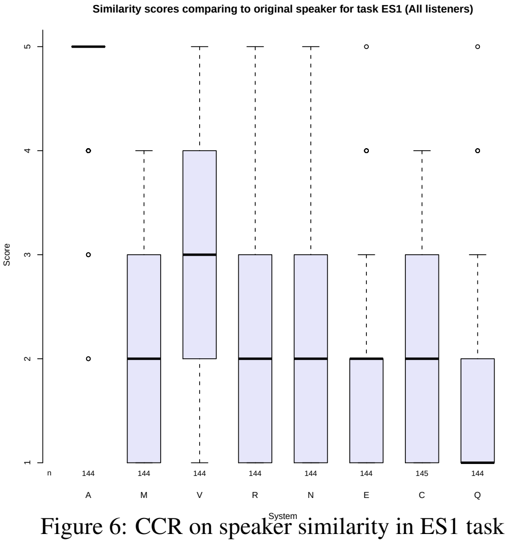

Figure 6: CCR on speaker similarity in ES1 task.