Overview of SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013

Yansuo Yu, Fengyun Zhu, Xiangang Li, Yi Liu, Jun Zou,
Yuning Yang, Guilin Yang, Ziye Fan, Xihong Wu

Speech and Hearing Research Center
Key Laboratory of Machine Perception (Ministry of Education)
Peking University, Beijing, 100871, China
{yuys}@cis.pku.edu.cn

Abstract

This paper introduces the SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013. A unit selection based approach is adopted to develop our speech synthesis system using audiobook speech corpus. Aiming at roughly labeled corpora with several hundred hours of speech, our system adopts lightly-supervised acoustic model training of speech recognition to select clean speech data with accurate text. Moreover, rich syntactic contexts instead of prosodic structure are utilized to refine traditional acoustic models. Through automatic syntactic parsing, this way can also help to label the corpora of several tens or even hundreds of hours automatically, thus avoiding manually prosodic annotation with time-consuming and expensive effort. In order to solve the problems of memory space expansion and running time burden for acoustic model training of large-scale corpora, a fast training method, which can ensure the accuracy of acoustic model, is realized. Subjective evaluation results show that our system performs well in almost all evaluation tests, especially in the case of large-scale corpora.

Index Terms: speech synthesis, speech data selection, syntactic parsing, unit selection

1. Introduction

We have been investigating many aspects of speech synthesis technology for years, especially in Mandarin. We once attended the Mandarin tasks of Blizzard Challenge at 2009. And this is our second entry to Blizzard Challenge. This year's challenge involves lots of under-researched topics, such as suboptimal recordings, several hundred hours of same speaker's speech without fine labeling, novels with different styles in both dialogue and aside. Aiming at this situation, many novel technologies, including lightly-supervised acoustic model training for speech data selection, speech labeling based on automatic syntactic analysis and a fast model training approach with low resources, are developed to construct our unit selection based speech synthesis system.

The paper is organized as follows. Section 2 introduces the basic situation of the English tasks in Blizzard 2013. An overview of the system will be discussed thoroughly in Section 3. The results of the evaluation are further described in Section 4. Finally, the conclusion is drawn in Section 5.

2. The English Tasks in Blizzard 2013

In Blizzard Challenge 2013, the English evaluation consists of two tasks as follows:

- EH1 - build a voice from the provided unsegmented audio; text is not provided, so must be obtained by participants from the web and aligned with the audio.
- EH2 - build a voice from the provided segmented audio; the accompanying aligned text may be used, or text may be obtained from the web.

For both EH1 and EH2 tasks, the audiobook data is kindly provided by The Voice Factory, from a single female speaker. In EH1 task, this year's challenge provides approximately 300 hours of chapter-sized mp3 files. In EH2 task, approximately 19 hours of non-compressed wav files are prepared and further labeled by Lessac Technologies, Inc. This task remains the same way as previous challenges. In the following sections we will introduce the whole process of constructing the speech synthesis system for both EH1 and EH2.

3. Overview of the System

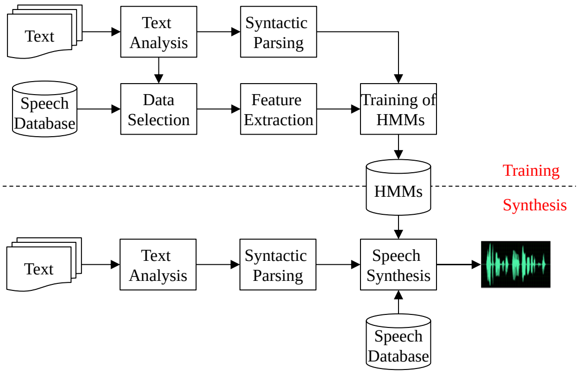

Figure 1: *Flowchart of SHRC-Ginkgo speech synthesis system.*

The overview of the text-to-speech (TTS) system, which consists of both training and synthesis parts, is shown in Figure 1. At training stage, the clean speech with accurate text is firstly chosen from roughly labeled corpora with several hundred hours of speech by means of speech recognition and text alignment. Afterwards the acoustic features including spectral envelope and F_0 are extracted from this chosen speech and the corresponding text are labeled with both phone-related and syntax-related tags through text analysis and syntactic parsing respectively. Based on these acoustic features and the context-dependent labels, the corresponding HMMs are estimated in the maximum likelihood (ML) sense [1]. At synthesis stage, the context-dependent label sequence of synthesized text is first predicted by the front-end text analysis and syntactic parsing. Then this label sequence is adopted to choose optimal waveform segment sequence from speech corpora under the statistical criterions, such as maximum likelihood [2], minimum Kullback-Leibler divergence (KLD) [3] or a combination of both criterions [4]. Finally, all consecutive waveform segments in the optimal sequence are concatenated to produce the synthesized speech. The following subsections will introduce the whole system in detail.

3.1. Data Preparation

3.1.1. Speech Data Selection

Through detailed analysis about the speech data of EH1 task, it can be found that this corpus has the following characteristics: (1) all data come from a number of novels or stories read by the same person; (2) there are not accurate transcriptions along with speech and these texts downloaded from the Internet can't be guaranteed to be consistent with the speech content; (3) The average gain varies from one speech segment to another due to different recording environment; (4) Because the reader will employ various timbre and rhythm to show the characteristics of the fiction roles, such as age, gender and mood, speech segments from the dialogue or the aside differ widely in both acoustic and prosodic aspects. Hence it's necessary to first choose "clean" speech data from raw corpus in order to construct the following speech synthesis system (Here "clean" speech data mainly means that the speech has both relative high quality and quite accurate transcription).

Referring to [5, 6], the basic process of speech data selection is designed as follows. The whole process mainly adopts the method of speech recognition based on transcription-related language model (LM). This LM leads to the effect similar to training set in the process of speech recognition. Based on this, if the recognition result is not identical with raw transcription, it's likely that the transcription has the errors, such as insertion error, deletion error or substitution error. Afterwards, the word error rate (WER) for every sentence is calculated through text alignment and the corresponding speech durations for each-level WER are also obtained, as listed in Table 1. At last, all the aside sentences, whose WERs are zero, are chosen as the final training set for EH1.

Table 1: The WER result of text alignment as well as its corresponding speech durations

Word error rate(%)	Speech durations(hour)
0.0	214.02
≤ 0.1	262.39
≤ 5.0	291.95
≤ 10.0	292.26
≤ 20.0	292.38

3.1.2. Syntactic Parsing

In general, prosodic context referring to prosodic structure is often selected as labels for the context-dependent HMMs. To some extent, this way can capture the suprasegmental characteristics of prosodic parameters. But more rich linguistic information can not been recovered fully from this simple four-layer prosodic structure. Therefore we attempted to introduce more linguistic context from syntactic tree to represent the context-dependent HMMs. In this paper, both internal grammar structure of the sentence and internal collocation relations among the words [7] for syntactic tree are fully adopted to refine traditional acoustic models. Here two categories of syntactic features including grammatical types and position relations for phrases of different levels are considered. Grammatical types mainly involve the types of father phrase, grandfather phrase and others for the previous, current and next words. For position relations, the relative and absolute positions among father phrase, grandfather phrase and others of current word are included. At last, conventional prosodic context are replaced by rich syntactic context in the process of modeling acoustic parameters for both EH1 and EH2. It is noted that our syntactic parsers are trained using the Berkeley parser [8], which achieves high performance across many languages.

3.2. Model Training

In training stage, spectrum (e.g., 39-ordered mel-cepstral coefficients and their dynamic features) and excitation (e.g., F_0, and its dynamic features) parameters are first extracted from the speech database using STRAIGHT [9] and modeled by the corresponding context-dependent HMMs. These parameters are further separated into different streams, in which mel-cepstral coefficients are modeled by continuous HMMs while F_0 observations are modeled by the MSD-HMMs [10]. Specially, a single Gaussian distribution is adopted to model the distribution of state duration. Finally, the context-dependent HMMs for each stream are constructed using the decision-tree-based context-clustering method with the minimum description length (MDL) criterion.

Besides that, when the quantities of speech increase up to one hundred or even several hundred hours, the conventional training process of acoustic model [11] is not appropriate due to the problems of both memory space expansion and running time burden. First, a amount of full-context HMMs increased dramatically lead to memory space expansion. Second, running time burden mainly comes from both Baum-Welch reestimation and model clustering based on decision tree for every stream. In this paper, a fast training method, which can ensure the accuracy of acoustic model, is realized through the optimization of conventional process of model training.

3.3. Speech Synthesis

A unit selection based approach, similar to [3, 4], is employed to construct our speech synthesis system for EH1 and EH2. For a whole sentence containing N phones, the selection criterion combining the unit likelihood with the distance criterion is adopted as in Equation (1). The unit likelihood mainly involves the probability of acoustic observation $\mathbf{o}_n$ (spectrum and F_0) including static and dynamic features and phone duration d_n for the n^{th} phone. Thus the optimal instance sequence $\mathbf{u}^*$ can be determined using Equation (1).

$$\mathbf{u}^* = \arg\max_{\mathbf{u}} \sum_{n=1}^{N} [LL(\mathbf{u}_n, \lambda_n) - D(\tilde{\lambda}_n, \lambda_n)] \qquad (1)$$

$$LL(\mathbf{u}_n, \lambda_n) = w_o \log P(\mathbf{o}_n|\lambda_n, Q_n) + w_d \log P(d_n|\lambda_n^{dur}) \qquad (2)$$

$$D(\tilde{\lambda}_n, \lambda_n) = \sum_{c\in\{s,p,d\}} \sum_{i=1}^{S} D_{KL}^{c}(\tilde{\lambda}_n^i, \lambda_n^i) \cdot t_i \qquad (3)$$

$$
\begin{aligned}
D^{c}_{KL}(\tilde{\lambda}^{i}_{n}, \lambda^{i}_{n}) \leq (\omega_0 - \tilde{\omega}_0) \log \frac{\omega_0}{\tilde{\omega}_0} + (\omega_1 - \tilde{\omega}_1) \log \frac{\omega_1}{\tilde{\omega}_1} \\
+ \frac{1}{2} tr \left\{ \omega_1 (\mathbf{\Sigma}_i \tilde{\mathbf{\Sigma}}_i^{-1} - \mathbf{I}) + \tilde{\omega}_1 (\tilde{\mathbf{\Sigma}}_i \mathbf{\Sigma}_i^{-1} - \mathbf{I}) \right. \\
\left. + (\omega_1 \mathbf{\Sigma}_i^{-1} + \tilde{\omega}_1 \tilde{\mathbf{\Sigma}}_i^{-1})(\mathbf{m}_i - \tilde{\mathbf{m}}_i)(\mathbf{m}_i - \tilde{\mathbf{m}}_i)^T \right\} \\
+ \frac{1}{2} (\tilde{\omega}_1 - \omega_1) \left| \mathbf{\Sigma}_i \tilde{\mathbf{\Sigma}}_i^{-1} \right|
\end{aligned} \tag{4}
$$

where $\mathbf{u}_n$ is one of the candidate units for the n^{th} phone and $LL(\mathbf{u}_n, \lambda_n)$ is the log likelihood of candidate unit $\mathbf{u}_n$; w_o and w_d are the likelihood weights of acoustic observation and phone duration; Q_n and λ^{dur}_{n} denote the state allocation and the duration model for the n^{th} phone respectively; S is the number of states; $c \in \{s, p, d\}$ represents the index of spectrum, F_0 and duration streams respectively and t_i is the duration of the state i; $\omega_0, \tilde{\omega}_0$ and $\omega_1, \tilde{\omega}_1$ are prior probabilities of the discrete and continuous sub-space (for spectrum and duration, $\omega_0, \tilde{\omega}_0 \equiv 0$ and $\omega_1, \tilde{\omega}_1 \equiv 1$); $N(\mathbf{m}_i, \mathbf{\Sigma}_i)$ and $N(\tilde{\mathbf{m}}_i, \tilde{\mathbf{\Sigma}}_i)$ denote the probability density function of state i for model λ_n and $\tilde{\lambda}_n$ respectively.

To further speed up the subsequent search process, three pruning techniques [12] including context pruning, beam pruning and histogram pruning, are also employed in the process of pre-selection. Then dynamic programming search can be applied to find the optimal unit sequence in the above maximum likelihood sense. Finally, the cross-fade technique [13] is adopted to smooth the phase discontinuity at the concatenation points and the waveforms of every two consecutive units in the optimal sequence are concatenated to generate the synthesized speech.

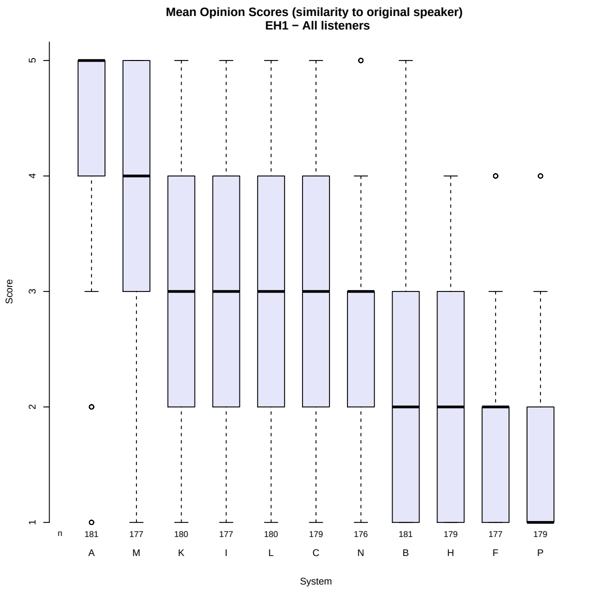

Figure 2: *Results of MOS on speaker similarity for EH1.*

4. Results and Discussion

This section will discuss the evaluation results of our system in Blizzard Challenge 2013 in detail. Our system is identified as M, whereas system A, B and C are benchmark systems. System A is the natural speech, system B is the Festival unit selection benchmark system and system C is the HTS statistical parametric benchmark system.

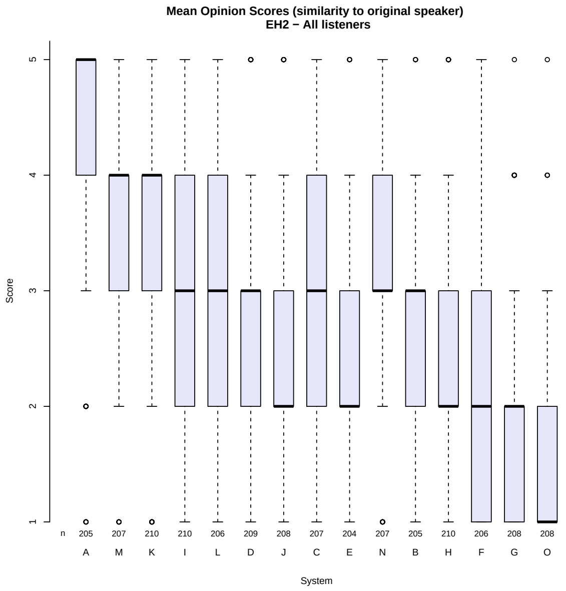

Figure 3: *Results of MOS on speaker similarity for EH2.*

4.1. Similarity test

Figure 2 and Figure 3 shows the results of similarity scores of all systems for EH1 and EH2. It can be seen that our system achieves the best similarity to original speaker for EH1 and EH2. Moreover the results of Wilcoxons signed rank tests further show that the difference between system M and any other systems on similarity is significant at 1% level for EH1. The high similarity score of our system can be attributed to use the original segment of a large corpus, even though there are no modifications to adapt concatenated units to new context.

4.2. Naturalness test

Figure 4 and Figure 5 shows the results of MOS on naturalness of all systems for EH1 and EH2. As we can see, our system achieved the best performance (not including the natural speech system A) on naturalness among all the participant systems. And the Wilcoxons signed rank tests also show that the difference between M and any other participant systems on naturalness is significant.

4.3. Intelligibility test

Figure 6 and Figure 7 shows the results of the overall word error rate (WER) test of all systems for EH1 and EH2. The results show that our system achieves the 3th and 4th lowest WER among all the systems for EH1 and EH2 respectively. And as well as previous Blizzard Challenge evaluations, the intelligibility of HMM-based parametric synthesis method usually can achieve better performance than unit selection methods.

4.4. Paragraph test

In addition to three above tests, 60-point mean opinion scale (MOS) tests was further conducted to evaluate different aspects of novel paragraph, such as overall impression, pleasantness, speech pauses, stress, intonation, emotion, and listening effort. These evaluation results show that our system is the best system for EH1 and top-3 system for EH2 in all the seven aspects.

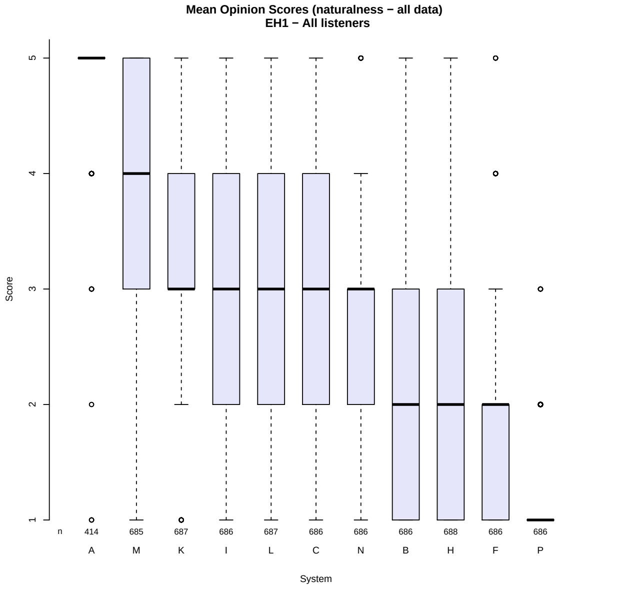

Figure 4: *Results of MOS on naturalness of sentences for EH1.*

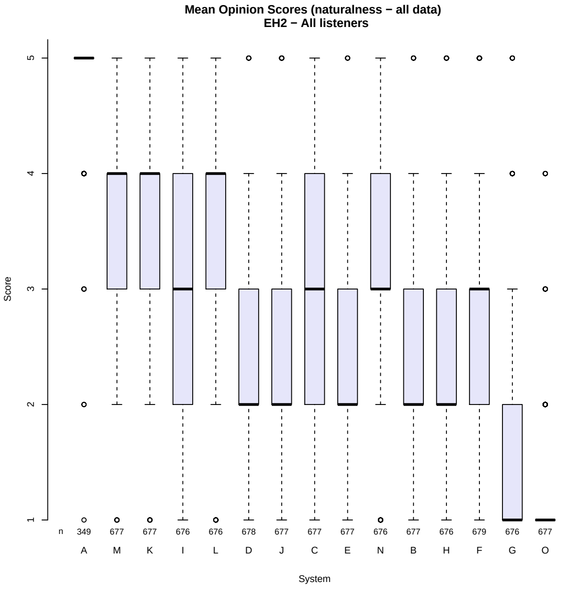

Figure 5: *Results of MOS on naturalness of sentences for EH2.*

Figure 8 and Figure 9 shows the results of MOS on overall impression of all systems for EH1 and EH2. It can be seen that the our system obtains more advantage on performance than the other systems as the speech data increases. This benefits from two aspects: first, our system could choose more clean speech data corresponding to accurate text from roughly labeled corpora with several hundred hours of speech; second, rich syntactic contexts may model complex prosodic variations more accurately than prosodic structure in the case of large corpora.

5. Conclusions

This paper introduces the development of the SHRC-Ginkgo speech synthesis system for Blizzard Challenge 2013. Many new technologies are exploited to construct our unit-selection speech synthesis system for the non-standard speech database. This system could realize automatically cleaning and labeling of large-scale corpora by means of speech recognition, text alignment and syntactic parsing. The evaluation results of Blizzard Challenge 2013 further indicate that our system can generate more natural synthesized speech in the novel domain than the other systems, especially in EH1. Some important problems of the audiobook synthesis are still needed to be solved in the future work, such as emotion expression of different roles, fast training of acoustic model based on large-corpora, and so on.

6. Acknowledgements

The work was supported in part by the National Basic Research Program of China (2013CB329304) and National Natural Science Foundation of China (No. 61121002, No.91120001).

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 *EUROSPEECH*, 1999, pp. 2347–2350.

[2] R. E. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge University, 1996.

[3] Z. J. Yan, Y. Qian, and F. K. Soong, "Rich-context unit selection (RUS) approach to high quality TTS," in *ICASSP*, 2010, pp. 4798–4801.

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

[5] X. G. Li, Z. H. Pang, and X. H. Wu, "Lightly supervised acoustic model training for mandarin continuous speech recognition," *Lecture Notes in Computer Science*, vol. 7751, pp. 727–734, 2013.

[6] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[7] Y. S. Yu, D. C. Li, and X. H. Wu, "Prosodic modeling with rich syntactic context in hmm-based mandarin speech synthesis," in *IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

[8] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *Proceedings of Human Language Technology conference - North American chapter of the Association for Computational Linguistics annual meeting(HLT-NAACL)*, Rochester, NY, USA, April 2007, pp. 404–411.

[9] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP*, 1999, pp. 229–232.

[11] K. Tokuda and H. Zen, "Fundamentals and recent advances in hmm-based speech synthesis," in *Tutorial of INTERSPEECH*, Brighton, UK, 2009.

[12] Y. Qian, Z. J. Yan, Y. J. Wu, F. K. Soong, X. Zhuang, and S. Y. Kong, "An HMM trajectory tiling (HTT) approach to high quality TTS," in *INTERSPEECH*, 2010, pp. 422–425.

[13] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," in *Proceedings of Speech Synthesis Workshop*, Pittsburgh, PA, USA, 2004, pp. 37–42.

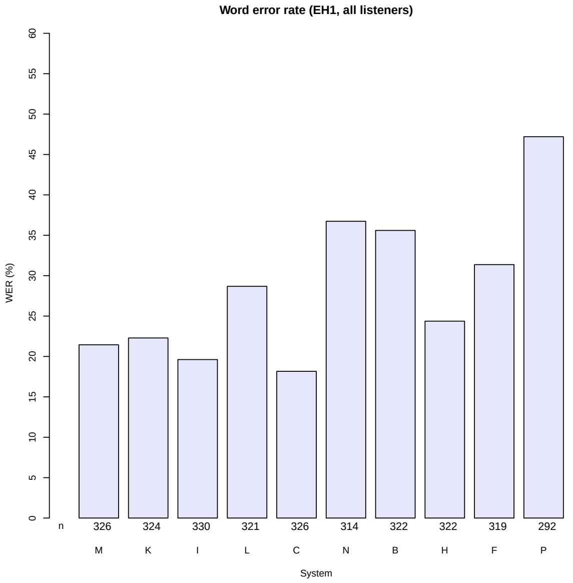

Figure 6: *Results of word error rate (WER) for EH1.*

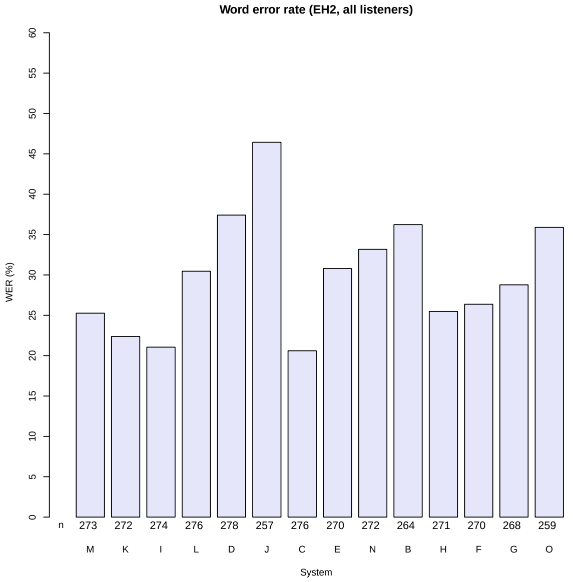

Figure 7: *Results of word error rate (WER) for EH2.*

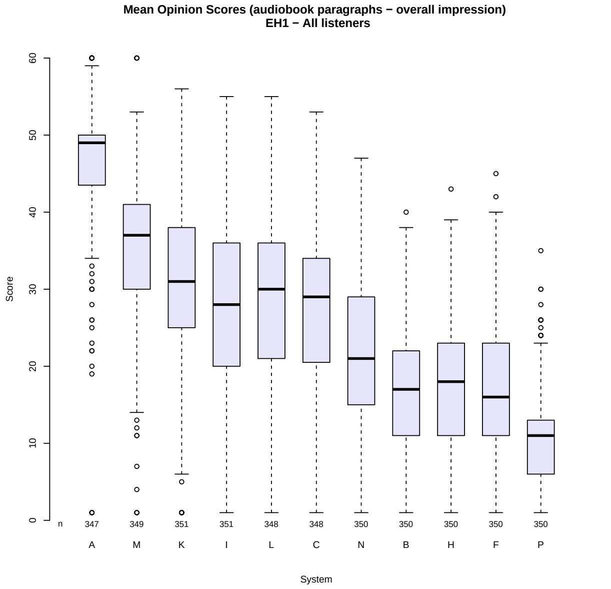

Figure 8: *Results of MOS on overall impression of audiobook paragraphs for EH1.*

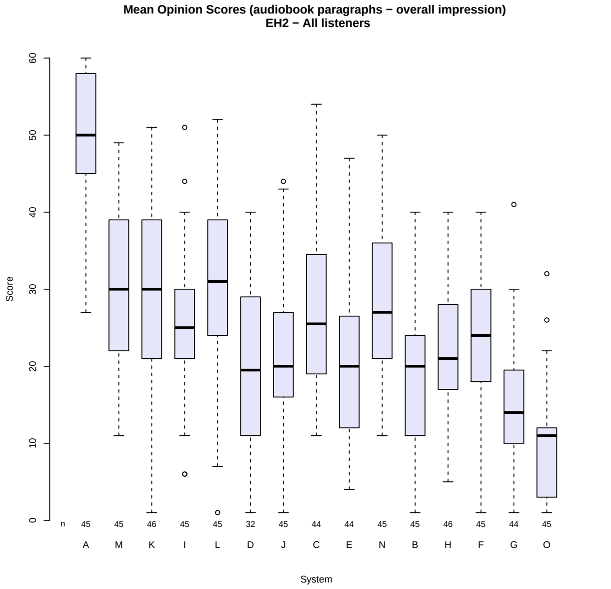

Figure 9: *Results of MOS on overall impression of audiobook paragraphs for EH2.*