The WISTON Text to Speech System for Blizzard Challenge 2010

Jianhua Tao, Shifeng Pan, Ya Li, Zhengqi Wen, Yang Wang

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{jhtao, sfpan, yli, zqwen, yangwang}@nlpr.ia.ac.cn

Abstract

The paper introduces the speech synthesis system developed by Institute of Automation, Chinese Academy of Sciences(CASIA) for Blizzard Challenge 2010. The large corpus based speech synthesis system, WISTON, was built to synthesize Mandarin speech. In this year, a new prosodic structure prediction model was used, which is more precise and compact than before. Furthermore, two kinds of syllable segmentation methods, i.e. rough segmentation and precise segmentation, were performed on Mandarin speech corpus. The rough segmentation labels were used in prosody models training and unit selection stage. During concatenation stage, these two kinds of segmentation labels are both used to determine the start position and end position of waveform fragment of each unit. Experiment results show that this approach is effective. The evaluation results show that except the similarity is very high, mean opinion score (MOS) and word error rate (WER) of WISOTN system are of average level.

Index Terms: Speech synthesis, WISTON, unit selection

1. Introduction

Large corpus based unit selection approach is always a popular and wide used approach to speech synthesis for its high naturalness and voice quality of synthetic voices, despite the occasional occurrence of inappropriate units [1][2]. The WISTON system is such a unit selection system [3][4]. CASIA has joined Blizzard Challenge with WISTON system since 2008.

WISTON system consists of two main modules: text processing module and unit selection module, i.e. front-end and back-end. The text processing module conducts text pre-processing, word segmentation, part of speech (POS) tagging, phonetic transcription and prosodic structure prediction. The unit selection module conducts the selection of unit, where context dependent pre-selection is performed and a set of Classification and Regression Tree (CART) based models are used to guide the calculation of target cost and concatenation cost.

There are mainly two differences between WISTON system for Blizzard Change 2010 and 2009. One is that a new prosodic structure prediction model is used in front-end, which is more precise and compact than before. The other is that the Mandarin corpus is segmented by two kinds of segmentation method, i.e. rough segmentation and precise segmentation. The prosody models are trained with rough segmentation labels. During the waveform concatenation stage, these two kinds of segmentation labels are both used to determine the start position and end position of waveform fragment of each unit. Experiment results show that this approach is effective.

The rest of this paper is organized as follows. In section 2, a brief system overview is given. In section 3, the text analysis module is introduced. Section 4 introduces the unit selection module, including pre-selection of units, calculation of target cost and concatenation cost, and study on syllable segmentation methods. In section 5, building of WISTON system for Blizzard challenge 2010 is introduced, and the evaluation results are analyzed. The conclusion is presented in section 6.

2. System Overview

Fig. 1 shows the overview of WISTON system.

In the training stage, speech corpus is annotated firstly, including syllable segmentation, pitch contour annotation and

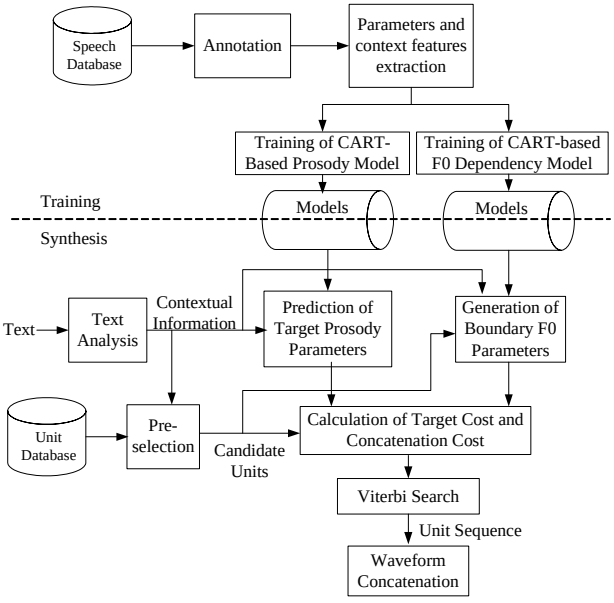


Figure 1: *An overview of WISTON system.*

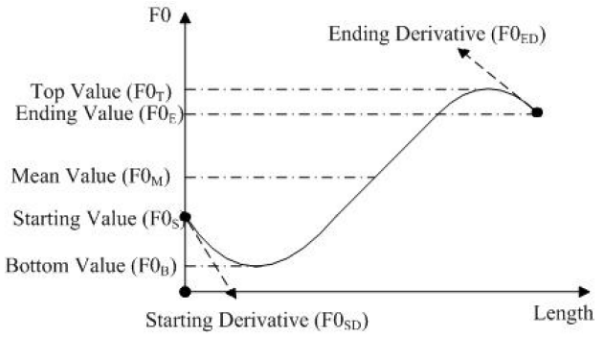


Figure 2: *Seven F0-related parameters used in prosody models training*

prosodic boundary labeling [5]. Secondly, the prosody parameters and contextual information of each unit in corpus are extracted. The prosody parameters include duration of unit (D_{UNIT}), silence duration between two adjacent unit (D_{SIL}) and seven F0-related parameters. These F0-realted parameters are F0 Mean ($F0_M$), F0 top value ($F0_T$), F0 bottom value ($F0_B$), F0 starting value ($F0_S$), F0 starting derivative ($F0_{SD}$), F0 ending value ($F0_E$) and F0 ending derivative ($F0_{ED}$), as Fig. 2 illustrates. Among these prosody parameters D_{UNIT}, D_{SIL}, $F0_M$, $F0_T$ and $F0_B$ are used to train context-dependent CART-based prosody prediction models, and $F0_S$, $F0_{SD}$, $F0_E$ and $F0_{ED}$ are used to train F0 dependency model. The prosody prediction models are used to calculate target cost, and the F0 dependency model is used to calculated concatenation cost.

In the synthesis stage, firstly, the contextual information of the text to be synthesized is analyzed and extracted by text analyzer. Secondly, the pre-selection procedure is conducted according to the contextual information. Then the prosody parameters are predicted by prosody prediction models and F0 dependency models. Then the target cost of each candidate unit and the concatenation costs between each pair of adjacent candidate units can be calculated. The optimal candidate units are selected by Viterbi search. Finally, the waveform fragments of optimal units are concatenated, and silence sections are inserted between some adjacent syllables based on the value predicted by silence model.

3. Text Analysis Module

Firstly, the front-end of WISTON translates the raw text into normalized utterance structure through the following processes, text normalization, word segmentation, Part-Of-Speech (POS) tagging, prosodic structure prediction.

The most important part of the text analysis is prosodic structure prediction. All the other procedures are carried out to improve the performance of prosodic structure prediction from textual features. In our work, we categorize the prosody structures into four levels: syllable, word (prosody word for Mandarin), minor prosody phrase and major prosody phrase. Three Maximum Entropy (ME) models are adopted here to predict boundaries for prosodic word (PW), prosodic phrase (PP) and intonation phrase (IP), respectively. The following sample shows the hierarchal prosodic structure of a sentence.

In text analysis module, the main difference between WISTON 2010 and WISTON 2009 are the prosodic structure prediction model is more precise and compact by automatic feature template selection.

As explained in [6], efficient feature template set and the attribute sets of each feature can greatly improve the prediction performance in machine learning. The procedure in text analysis is almost the same in all the Mandarin speech synthesis systems. However, we can work harder on feature selection to enhance the capability of the ME models and further the text analysis module. Before the prosodic structure prediction, the following textual information is obtained.

- The pronunciation (PINYIN) of each Chinese character in the utterance.
- The word segmentation information.
- The Part-Of-Speech of each word.

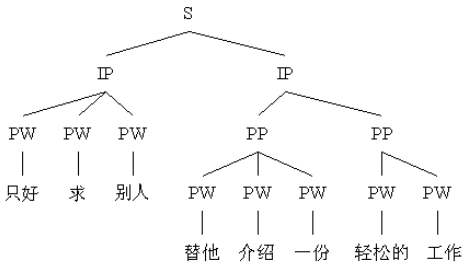


Figure 3: *Hierarchal prosodic structure (zhi2 hao3 qiu2 bie2 ren5 ti4 ta1 jie4 shao4 yi2 fen4 qing1 song1 de5 gong1 zuo4. English meaning: he has to ask others to introduce him an easy job).*

These features are used in ME model as well as their position information. The sliding window method was adopted in feature extraction. The window length is 7, which indicate that all the words from the 3rd of the previous to the 3rd of the next are taken into consideration. The first two kinds of information are based on the input text and can not be changed, whereas the POS information is just intermediate result which is used in prosodic structure prediction. Therefore, we could elaborately define a POS set which is suitable for the subsequent prosodic structure prediction. The current POS set is established by Peking University, which consists of 44 POS tags. This POS set may not be the most appropriate set for prosodic structure prediction for two reasons. Firstly, this POS set is designed for grammatical or syntactic analysis. As shown in Fig. 3, although prosodic structure is related with syntactic structure, they are not the exactly the same. Secondly, feature selection is an important factor which may improve the overall performance in most machine learning algorithms.

We utilize a wrapper method in feature selection [7]. The initial feature set is the above mentioned single feature template, including word, word segmentation, POS, the position and distance to the beginning of the sentence and to the end of the sentence. Then, we try to combine two of these templates, and add it into the feature template set, and then train a ME model. The criterion is the F-score of the ME model prediction results, which balance precision and recall at same time. At each step, only the combined template with highest F-score will be added in the final feature template set. The procedure stops only when no improvement is achieved in F-score. Although a better performance can be obtained by using wrapper based feature selection, the whole procedure is time-consuming. To speed up the wrapper based feature selection, we calculate the similarity between every two POS tags and only a certain amount of the combined POS tag with a higher similarity score are calculated in each step. After feature selection, only 29 POS tags are selected and the F-score is improved by 2%. Although 2% is not an impressive improvement, the size of the new ME model become smaller.

4. Unit Selection Module

4.1. Pre-selection

In a corpus based speech synthesis system, there are too many candidate units for each target unit. Conducting unit selection procedure on such a large database is very time-consuming. To decrease the number of candidate units and thus improve

the running speed, a contextual information difference (CID) based pre-selection is conducted. The CID is defined in Eq. (1).

$$CID = \sum_{i=1}^{N} W_i * D_i \quad (1)$$

, where N is the number of contextual information category, D_i is the difference of the *i* th contextual information between current candidate unit and the target unit and W_i is the weight of the *i* th contextual information.

The CID depicts the difference of contextual information between the candidate unit and the target unit to be synthesized. The contextual information used here includes the location of the current speech unit in word, phrase and sentence, the name of syllable, the length of word, phrase and sentence, the boundary types before and after the current speech unit, etc.

After the pre-selection, a small number of candidate units which have the smallest CID will be kept for the later processing.

4.2. Target Cost

Target cost is defined as the difference between the prosody parameters predicted by prosody models and the prosody parameters of candidate unit. In our work, the prosody parameters used for target cost include $F0_M$, $F0_T$, $F0_B$ and D_{UNIT}. The three F0-related parameters denote the pitch register and the pitch range. Eq. (2) shows the definition of target cost,

$$C_T = w_1 * DF0_M + w_2 * DF0_T + w_3 * DF0_B + w_4 * DD_U \quad (2)$$

,where $DF0_M$, $DF0_T$, $DF0_B$ and DD_U denote the difference between the predicted prosody parameters and those of candidate units respectively, and $w_1 \sim w_4$ are weights.

4.3. Concatenation Cost

Prosody parameters involved in concatenation cost include $F0_S$, $F0_{SD}$, $F0_E$ and $F0_{ED}$, as Fig. 2 illustrates. These four parameters can be considered as boundary features of a unit's f0 contour. For two adjacent units, their f0 contours have impacts on each other [8]. Fig. 4 shows some examples of pitch contour of two adjacent Mandarin syllables. As can be see that these four parameters, $F0_E$ and $F0_{ED}$ of the former syllable, $F0_S$ and $F0_{SD}$ of the latter syllable, have some strong and complicated relationship. Therefore, a CART based F0 dependency model is adopted to learn this complicated relationship, which can be used later to predict these boundary parameters of F0 contour [9]. Parameters involved in the concatenation cost include the above four boundary parameters. When we model and predict these four parameters, features listed in Table 1 are used.

The predicted value by this dependency model can be considered as the expected boundary F0 value by adjacent syllables. Therefore the difference between these predicted values and actual values of candidate unit can be used to measure the concatenation cost of F0. For spectrum, the continuity of spectrum across the concatenation point can be used to measure the concatenation cost of spectrum. Therefore, the total concatenation cost can be calculated as Eq. (3) demonstrates.

$$C_C = w_5 * DF0_S + w_6 * DF0_E + w_7 * DF0_{SD} + w_8 * DF0_{ED} + w_9 * D_{spec} \quad (3)$$

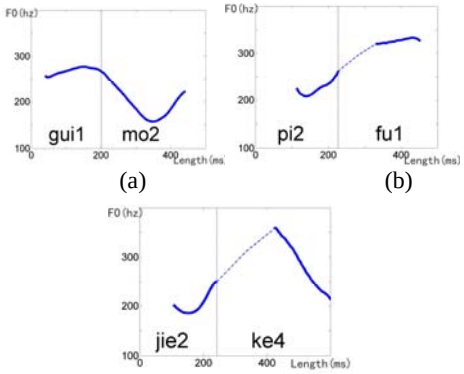


Figure 4: *Examples showing the impact of adjacent syllables' F0 contours on the current one [9].*

Table 1. *Features used in predicting* $F0_S$, $F0_{SD}$, $F0_E$ and $F0_{ED}$

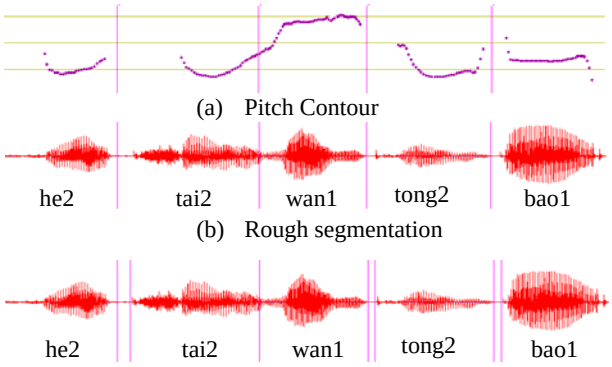
Features in predicting $F0_S$ and $F0_{SD}$	Features in predicting $F0_E$ and $F0_{ED}$
frequently used text information (tone, initial/final identity, prosody structure, etc)	frequently used text information (tone, initial/final identity, prosody structure, etc)
previous syllable's $F0_E$ and $F0_{ED}$	following syllable's $F0_S$ and $F0_{SD}$
pause length before current syllable	pause length after current syllable
current syllable's final part length	following syllable's initial part length

, where $DF0_S$, $DF0_E$, $DF0_{SD}$ and $DF0_{ED}$ denote the difference between the predicted $F0_S$, $F0_E$, $F0_{SD}$, $F0_{ED}$ and those of candidate units respectively, D_{SPEC} denote the discontinuity of spectrum across the concatenation, and $w_5 \sim w_9$ are weights.

4.4. Study on Segmentation of Corpus

For Mandarin speech synthesis system using unit selection approach, syllable is usually chosen as the basic unit. Though the prosodic and acoustic features inside a syllable are well preserved, the continuity of prosodic and acoustic parameters on syllable boundary is still an important and difficult target to achieve.

The segmentation of corpus is very important for both prosody models training and unit waveform concatenation. We have tried two kinds of syllable segmentation methods before. One is precise segmentation, which means a precise syllable boundary is preserved and the short silence or the changeover section with low energy between adjacent syllables is segmented. The other is a rough segmentation, which means the short silence or changeover section between two adjacent syllables is divided into two parts, and assigned to the corresponding syllable respectively. Some segmentation examples are shown in Fig. 5.



(c) Precise segmentation

Figure 5: *An example of two kinds of segmentation on Mandarin speech*

Fig. 5 (b) shows that sometimes there is a short silence (or short pause) between two adjacent syllables. The duration of this short silence may be 30 ms or less. In Fig. 5 (b), the rough segmentation is performed, which means these short silence sections will be divided and merged into the corresponding syllables. The silence sections that will be segmented out finally are only those with long duration, which usually exist on phrase level pause or even higher level pause. After the whole corpus is segmented by this rough segmentation method, the training data can be gathered, including syllable duration, silence duration and context features. Then the context-dependent CART-based models for syllable duration and silence duration can be trained respectively. As can be inferred, the syllable duration predicted by syllable duration model is usually a little longer than that of natural speech. However, the added portion contributed by short silence is much smaller than syllable duration itself. Therefore this difference can be ignored. As to the silence duration predicted by silence model, it is usually zero at syllable level boundary and word level boundary. At phrase level boundary or sentence level boundary it is usually an effective value. This is determined by Mandarin speech itself and the rough segmentation method. During the unit selection stage, the optimal unit is selected by minimize the sum of target cost and concatenation cost. However the prosody parameters that are used in unit selection stage have weak ability to control the waveform at boundary region where there's no excitation, i.e., no F0, as Fig. 5 (a) and Fig. 5 (b) illustrate. Therefore when concatenating two adjacent units, sometimes the changeover section between two synthesized units seems to be too long or mismatch, which degrades the naturalness of synthetic voice.

In Fig. 5 (c), the precise segmentation is performed. The exact boundary of each syllable is segmented, and many short silence sections occur. The models trained by these data have such features that the syllable duration modeling is more exact, and many short silence sections occur even at syllable level boundary. Though most of these short silence predicted by silence model are very short, e.g. 10ms or 20ms, etc. Too many short inserted silence sections will make the synthetic voice sound not coherent, or sound a little discontinuous.

To achieve a better performance, these two segmentation labels are both used in the new WISTON system. That is, the models trained by rough segmentation labels are used to predict all the prosody parameters which are then used to perform the selection of units. However, when picking out the waveform fragments of each optimal unit, the precise segmentation labels are used. With this method there's nearly no short silence at syllable level and word level boundary predicted by silence model, which will help to improve the continuity of synthetic voice. Meanwhile, the 'clean' waveform fragments of units derived by precise segmentation will also help to improve the continuity of synthetic voice.

Considering those plosive consonants which have a short silence (pause) preceded them, such as the consonant of 'tong2' in Fig. 5 (b), it will be better if the waveform fragments with starting position labeled by rough segmentation are used to conduct the concatenation. With this special process, the plosive consonants in synthetic voice sound more natural and clear.

Experiments were carried out to evaluate this new method. Three voices were synthesized. Voice A was synthesized only using the precise segmentation. Voice B was synthesized only using the rough segmentation. Voice C was synthesized using the approach introduced above. 20 sentences were synthesized for each voice. 10 speech experts were asked to perform the preference test. Test 1 was conducted between voice A and Voice B. The result is shown in Table 2. As seen from Table 2, these two voices have almost equal performance. Then test 2 was conducted between voice B and voice C. Table 3 shows the result. As we can see, voice C is better than voice B.

Table 2. *result of test 1*

Prefer A(%)	Prefer B(%)	Equal(%)
33	35	32

Table 3. *result of test 2*

Prefer B(%)	Prefer C(%)	Equal(%)
22	42	36

5. Evaluation

5.1. System Building for Blizzard 2010

WISTON system was built for Mandarin hub task 1(MH1) of Blizzard 2010.

The Mandarin corpus consists of 5884 utterances, uttered by a professional female speaker. The speech signals were sampled at 16 kHz with 16 bit sampling precision.

The whole corpus was firstly annotated, including segmentation, pitch contour annotation and prosodic boundary labeling. During segmentation stage, the rough segmentation and precise segmentation method were both performed. After annotation, the prosody parameters and contextual information of each unit in corpus were extracted to train CART-based prosody models and F0 dependency models. At the same time, the unit database was also constructed according to the method introduced in section 4.3. Weights $w_1 \sim w_9$ of Eq. (2) and Eq. (3) are set by empirical values.

5.2. Evaluation Results

The mean opinion score (MOS), similarity and word error rate (WER) were evaluated for MH1. The results are shown in Fig. 6 – Fig. 8, where system A identifies natural speech, system C identifies HTS-2005 benchmark system and the identifier of WISTON is K. For similarity evaluation, our system has a good performance, which is one of the three highest-score systems (natural voice excluded). For MOS score, our system only ranked average level. As to WER, our system is still of

average level. These results reminder us there's still many works need to be done, especially on improving the MOS score and reducing WER.

6. Conclusion

In this paper, the WISTON system built for Blizzard challenge 2010 by CASIA is introduced. There are two new features on WISTON system of this year. One is a new prosodic structure prediction model is used in front-end, which is more precise and compact. The other is the two kinds of segmentation labels are both used to determine the start position and end position of waveform fragment of each unit during concatenation stage. The evaluation results show that the similarity of our system is very high. However, MOS and WER of our system are of average level. Many works need to be done focusing on these two aspects.

7. Acknowledgment

The work was supported by the National Science Foundation of China (No. 60873160), 863 Programs (No. 2009AA01Z320) and China-Singapore Institute of Digital Media (CSIDM).

8. References

[1] A. Hunt and A. Black, "Unit selection in a concatenate speech synthesis system using a large speech database", in proc. ICASSP. 1996, pp. 373-376.

[2] A. Black, H. Zen, K. Tokuda, "Statistical parametric speech synthesis", in proc. ICASSP 2007, pp. 1229-1232.

[3] J. H. Tao, J. Yu, L.X. Huang, etc, "The WISTON Text to Speech System for Blizzard 2008", in Blizzard Challenge Workshop, 2008.

[4] J. H. Tao, Y. Li, S. F. Pan, etc, "The WISTON Text to Speech System for Blizzard 2009", in Blizzard Challenge Workshop, 2008.

[5] Jianhua Tal, "Acoustic and Linguistic Information Based Chinese Prosodic Boundary Labeling", TAL 2004.

[6] Fengjian Li, Guoping Hu and Renhua Wang, 2004, Prosody Phrase Break Prediction Based on Maximum Entropy Model, Journal of Chinese information processing, 18(5), pp. 56-63e

[7] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection", *Artificial Intelligence*, Vol. 97, No. 1, pp. 273-324, 1997.

[8] G. Kochanski, C. Shih, "Prosody Modeling with Soft templates", Speech Communication, 2003. 39. pp. 311-352.

[9] J. Yu and J.-H. Tao, "A novel prosody adaptation method for mandarin concatenation-based text-to-speech system," Acoust. Sci. & Tech., 2009, pp. 33–41.

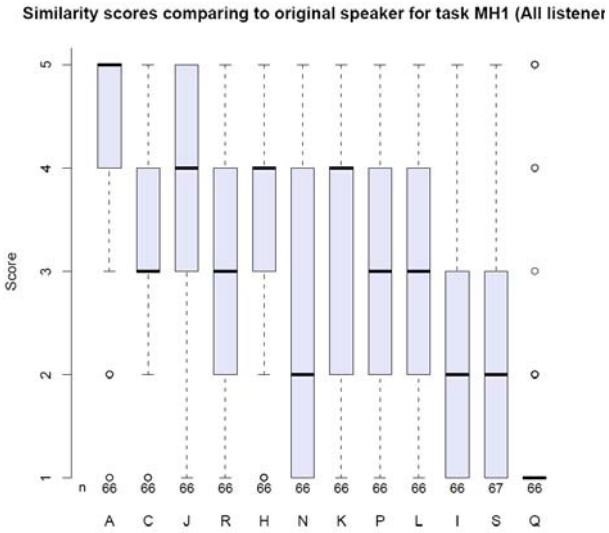


Fig 6: *Similarity scores by all listeners (K: WISTON system)*

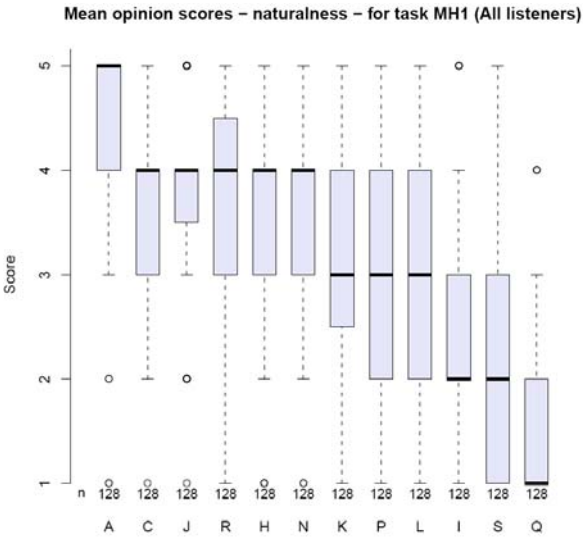


Fig 7: *MOS scores by all listeners (K: WISTON system)*

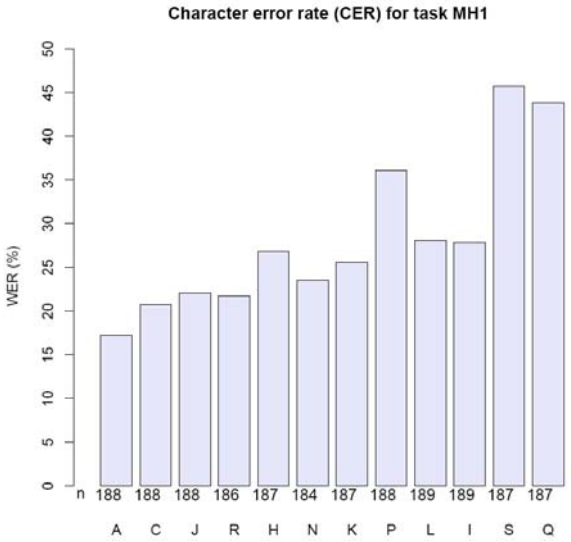


Fig 8: *WER by all listeners (K: WISTON system)*