Blizzard 2008: Experiments on Unit Size for Unit Selection Speech Synthesis

E.Veera Raghavendra[1], *Srinivas Desai*[1], *B. Yegnanarayana*[1], *Alan W Black*[2], *Kishore Prahallad*[12]

[1]International Institute of Information Technology, Hyderabad, India
[2]Language Technologies Institute, Carnegie Mellon University, Pittsburg, USA
{raghavendra,yegna}@iiit.net, srinivasdesai@research.iiit.net, {awb,skishore}@cs.cmu.edu

Abstract

This paper describes the techniques and approaches developed at IIIT Hyderabad for building synthetic voices in Blizzard 2008 speech synthesis challenge. We have submitted three different voices: English full voice, English ARCTIC voice and Mandarin voice. Our system is identified as D. In building the three voices, our approach has been to experiment and exploit syllable-like large units for concatenative synthesis. Inspite of large database supplied in Blizzard 2008, we find that a back-off strategy is essential in using syllable-like units. In this paper, we propose a novel technique of approximate matching of the syllables as back-off technique for building voices.

Index Terms: speech synthesis, unit size, tonal unit, prominence

1. Introduction

Blizzard is a speech synthesis challenge conducted every year since 2005, where teams from academia and industry participate in this challenge by building voices on a benchmark database. Such participation allows to exchange and compare the approaches and techniques for building synthetic voices. International Institute of Information Technology (IIIT) Hyderabad, has participated in Blizzard 2008 challenge for the first time. Our goal was to experiment with syllable-like large units for concatenative synthesis in the context of Blizzard 2008 challenge.

Concatenative synthesis is based on the concatenation of segments of recorded speech. Generally, concatenative synthesis produce the most natural-sounding synthesized speech [1]. This synthesis method uses basic speech units that produce the sounds of the particular language, along with the co articulation, prosody, and transitions of the language [2]. In concatenative text-to-speech (TTS) synthesis, the speech waveform is generated concatenating the prerecorded segments corresponding to a given unit sequence, where the unit may be a phone, diphone, syllable, word or phrase. These segments, referred to as acoustic units, are normally extracted from a prerecorded sentences uttered by a native and professional speaker of the language. The quality of the synthetic speech is thus a direct function of the available units, making unit selection very important. For good quality synthesis, all the units of the language should be present. Moreover, the units should also be generic so that they can be used for unrestricted synthesis, which means that they should have minimum prosodic variations [3]. Syllables on the other hand, are inherently of longer duration and it has been observed that the relative duration of syllables is less dependent on speaking rate variations than that of phonemes [4]. The human auditory system integrates time spans of 200 msecs of speech, which roughly corresponds to the duration of syllables [5]. Syllables also capture the co-articulation between sounds better than the phonemes.

There has been many attempts on syllable based synthesizers [6] [7]. There are few difficulties associated with syllable based speech synthesizers. First, how to handle co-articulation effect of adjacent syllables. Second, how to build up the database of syllable segments. Finally, syllable-based approach has to face the problem with a relatively large inventory of syllables and we can not cover all the syllables of the language in the lexicon.

This paper is organized as follows: Section 2 briefly describes the speech database used in Blizzard 2008 challenge. Section 3 describes the approximate matching for syllables. Section 4 describes the framework for building English full voice. Section 5 explains the approach adapted for building English ARCTIC voice. Section 6 describes the approach used for building Chinese voice, and Section 7 discusses the results of 3 systems.

2. Speech Database Used

Blizzard challenge 2008 has released two databases: UK English and Mandarin. UK English set contains 9508 sentences from 5 domains - novel, newspaper, emphasize, conversation and Semantically Unpredictable Sentences (SUS) and Mandarin set contains 4500 sentences from 2 domains - newspaper and SUS. We have submitted three different types of systems, one is built with full English voice second one is with ARCTIC subset which contains novel style sentences and the third system is Mandarin voice. We selected only novel, newspaper, conversation and SUS sentences for building English full voice and it contains 19907 syllables. To build ARCTIC and Mandarin voice, all the utterances of the corresponding database were used. We have used Festival [8] framework for building ARCTIC and Mandarin voice. To build the English full voice, Festival synthesizer was adapted to our needs.

3. Approximate Matching for Syllables

Syllable is considered as one of the largest unit used in speech synthesis. A syllable and can be typically of the following form: V, CV, VC, CCV, CCCV, and CCVC, where C is consonant and V is Vowel. A syllable can be represented as C*VC*, syllable should contain at least one vowel.

UK English phoneset consists of about 29 consonants and about 25 vowels. Theoretically possible syllable combinations with V, CV, CCV, CVC, CCVC representation are 652525. Syllable based synthesizers can produce very natural synthesis as number of joins are less at concatenation time. But, it is very difficult to cover all possible syllables of language in lexicon. To address this issue, we propose approximate matching of a

10.21437/Blizzard.2008-10

syllable, when it is not found in the database. The hypothesis of using approximate matching is that the end-users of synthetic voices are human beings and hence by replacing a syllable with its approximate match (even if a few phones of the syllable are missing), the perceptual mechanism of human beings will still be able to understand the utterance based on the context. As a result of approximate matching, an utterance could be synthesized using syllables and approximated syllables thus avoiding to back-off to lower level units such as diphones and half-phones. The following algorithm explains the approximate matching of syllable-like [9] units used in this work.

Approximate matching of syllable-like units:

1. break the syllable into 3 parts as /C*$_l$/ /V/ /C*$_r$/
2. if (/C*$_l$/ and /C*$_r$/) is null find /V/ in lexicon and return /V/, otherwise goto step 3
3. if /C*$_l$/ is null goto step 4, otherwise
   - break the /C*$_l$/ into individual consonants like /C_1,C_2,../.
   - Find the unit(/C*$_l$'/) in the lexicon with maximum number of possible consonants in /C*$_l$/ succeeded by vowel /V/ in right to left direction
   - if /C*$_r$/ is null return /C*$_l$'V/, otherwise goto step 4
4. break the /C*$_r$/ into individual consonants like /C_1,C_2,../
   - Find the unit(/C*$_r$'/) in the lexicon with maximum number of possible consonants in /C*$_r$/ preceded by /C*$_l$'V/ from left to right
   - return /C*$_l$'VC*$_r$'/

4. Framework for Building English Full Voice

The synthesis framework has been changed to use global syllable set and approximate matching. The current framework consists two phases in designing the system, building and synthesis.

4.1. Building Phase

During the building phase, the text transcriptions and recorded utterances are passed through lexical analysis and speech analysis respectively. In turn they produce phone sequences and signal features, fundamental frequency, mel-cepstral coefficients (MCEP) and energy. Phone sequence and MCEPs are passed to EHHM [10] for labeling the speech signal with respect to phone sequence of the utterance. EHMM would produce the labels with phones and it's time stamps in the speech signal. Rest of the procedure is broken into

- Creating High Frequency Words.
- Building multi syllable database.
- Duration and F0 modeling for unit selection.

4.1.1. Creating High Frequency Words

The idea of creating high frequency word database is that the quality of synthetic voice could be improved when the high frequency words are used directly for concatenation. The high frequency words of the text transcription are identified by calculating the frequency of each word. The words which have the frequency $>= 25$ are considered as the high frequency word. A separate catalogue is created for each high frequency word. A catalogue is a list contains the speech signal id, starting and ending time stamp of each example of the high frequency word. For example for the word *want*, the catalogue file appears as follows.
eg: *roger_6313 0.625 0.885*
roger_6370 1.725 1.995
roger_5759 1.220 1.475
roger_6544 1.165 1.415
roger_6284 1.260 1.545

4.1.2. Building Multi Syllable Database

If we use multi syllable sequence the quality of synthesis could be improved. naturalness more. Multi syllables are created word level. Each word in the text transcription is broken into syllables and n-gram syllables are generated. For example the word *possession* have */p,@/, /z,e/, /sh,en/* syllables. Corresponding n-grams are shown below, where each gram refers to one syllable.
/p,@/, /z,e/, /sh,en/
/p,@/, /z,e/
/p,@/

A separate catalogue file is created for each multi-syllable. When none-of the syllable are found, approximating matching is applied which is explained in Section 3.

4.1.3. Duration and F0 modeling

Prosody is a complicated phenomenon of spoken language. Simply speaking, it controls the flow of an utterance. The major components of the prosody are duration and fundamental frequency (F0).

One traditional method of determining these duration and F0 is to use a rule-based system, using rules based on the context in which the segment is set. These rules increase or decrease segment durations and F0 along a scale determined by the identity of the phone uttered during the segment. But with the large amount of data the process of manually deriving the rules becomes tedious and time consuming. Hence, rule based methods are limited to small amount of data. Statistical methods are good when dealing with large amount of database. There are various tools available for statistical modeling. In our experiments we are using WAGON, tool which comes with Edinburgh speech tools [11].

WAGON is a classification and regression tree. A tree is a binary tree, constructed based on questions concerning prosodic and phonetic context. Duration and F0 depends on the phones which are succeeded and preceded by current phone. In our context we conceived previous and next 2 phones as context for duration modeling. Where as 10 left and right phones are considered for predicting F0 of the phone.

4.2. Synthesis Phase

The synthesis phase consists of three steps, namely

- Target Prediction
- Target Cost
- Waveform Generation

4.2.1. Target Prediction

The input sentence to be synthesized is converted to phonemes, if required, and broken into words and syllables. Initially, the high frequency words are searched in the database, if the word is not found then the word is divided into syllable and n-gram syllables. Available n-gram syllable is taken from the syllable database. If the n-gram is not found, the mono syllable database is searched using syllable approximate matching technique.

4.2.2. Target Cost

There are multiple examples available in the database for each desired unit. To select the best unit from *'n'* examples, we consider duration and pitch are the selection factor components. We derive the desired duration and pitch information from WAGON model, which were generated during the duration and f0 modeling for the given sentence in phoneme level. The best unit is selected using Euclidean Distance between duration and f0 of the desired and *'n'* examples of the desired unit.

4.2.3. Waveform Generation

After units are selected for synthesizing, they can not be concatenated directly. It affects the natural sounding and articulation of the sentence. The ending and starting of the two joining units causes undesired discontinues between the subsequent units. To reduce the discontinues the two joining units must be smoothed, cross-fading [12] algorithm is applied for smoothing.

5. Building ARCTIC Voice

To build ARCTIC voice we have made use of acoustic driven modeling technique of prominence as explained below. Syllables in English can be stressed and unstressed. A stressed syllable could be accented if it appears in accented phrase. Thus syllables could be categorized into unstressed, stressed-unaccented and stressed-accented. Such categorization can be obtained from the lexical information and syntactic parsing. However, there exists acoustic variations due to style and complexities involved in uttering a sentence.

To model such variations, acoustic driven modeling of prominence is required. A simple methodology that could be adapted is to use Hidden Markov Models (HMM) for each of the three categorization of a vowel: i.e., unstressed vowel, stressed-unaccented vowel, and stressed-accented vowel. During the process of Viterbi alignment, these HMMs are connected in parallel and let the acoustics dictate the suitable HMMs, thus indicating the vowel to be unstressed, stressed-unaccented and stressed-accented based on the acoustic evidence.

For initialization, HMMs were trained for unstressed, stressed-unaccented and stressed-accented vowels as indicted by lexical analysis. Given these initialized models, the HMMs were used in forced-alignment. The labels obtained in this first-pass were treated as first order approximation. HMMs were retrained based on the labels obtained in the first-pass for a few more iterations on the ARCTIC set. The process of labeling and retraining was repeated for three times, and the final HMM models were used in forced-alignment to obtain the final set of labels for three different categories of vowels.

In informal perceptual studies, it was found that the use of unstressed, stressed-unaccented and stressed-accented type of vowels produced more natural and consistent speech than our default unit selection voice where no such information was used.

6. Building Mandarin Voice

Mandarin is the official spoken language of the People's Republic of China. In spoken Chinese, words are made up of one, two or more syllables. Each of the syllables is written with a separate character. Each character has its own meaning, though many are used only in combination with other characters, every character is given exactly the same amount of space, no matter how complex it is. There are no spaces between characters and the characters which make up multi-syllable words are not grouped together, so when reading or processing Chinese, one has to work out what the characters mean and how to pronounce them, and also which characters belong together. Hanyu Pinyin is the standard Chinese pronunciation system to represent characters using the Latin alphabet. Pinyin means "spell sound", or the spelling of the sound. The pinyin system also uses diacritics for the four tones of Mandarin, usually above a non-medial vowel.

Along with the speech database, the transcription was also provided in terms of sequence of Pinyin characters with tone markers. We made use of these transcription available in Pinyin characters to build the Mandarin voice. We have built two different systems which can synthesize unrestricted text of Mandarin. 1) System-1: The first system is built by hard binding the Pinyin character and associated tone together as a basic unit. The idea is to capture the exact pronunciation of the different variations of Pinyin character. The number of basic units in system-1 are 1460. 2) System-2: The second system is built with Pinyin characters alone as the basic unit of the system and tone is considered as a stress feature of the Pinyin character. Thus tone is loosely binded to each character. The number of basic units in system-2 are 399.

During synthesis, we synthesize an utterance using system-1. If the synthesis fails due to coverage of Pinyin characters then we go through system-2. We found that around 13 utterances had a rare Pinyin character which was not covered by either by system-1 or system-2. In order to comply to Blizzard 2008 challenge requirements, we manually removed that particular Pinyin character (), and synthesized the rest of the utterance.

7. Results and Discussion

The evaluation results of Blizzard challenge 2008 for these three systems are discussed in this section. In following figures, the label of IIIT Hyderabad system is identified as D. System A denotes the natural speech.

7.1. Similarity Test

Figure 1, 2 and 3 shows the Boxplots [13] of similarity scores of all systems for Voice A, B and C (Mandarin). From these figures we can observe that our system "D" is as good as many other systems in similarity scores. This can be attributed to the fact that multi syllable sequence can be used for synthesis.

7.2. Mean Opinion Score Test

The Boxplots of mean opinion scores of all systems for Voice A, B and C(Mandarin) are shown in Figure 4, 5 and 6. We observe that while syllable sequence preserve the naturalness, the intelligibility and consistency is not as good as other systems. We hope to improve this aspect in further challenges.

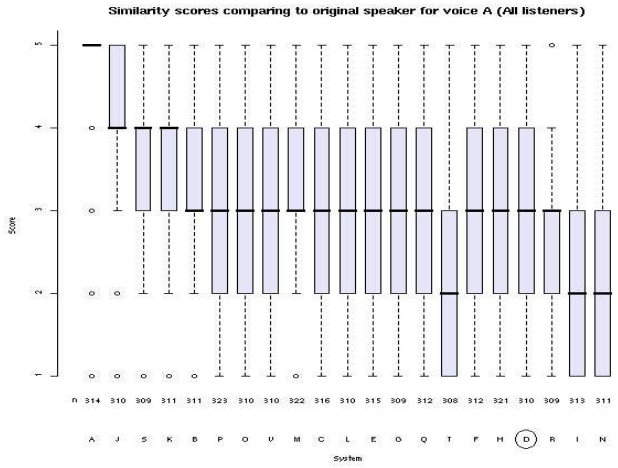

Figure 1: Boxplot of similarity scores for Voice A

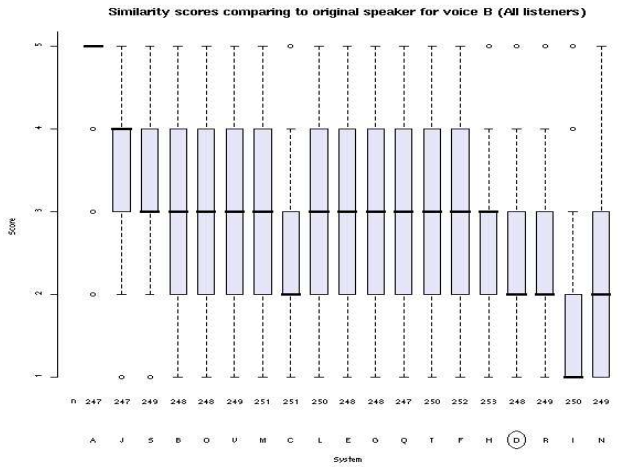

Figure 2: Boxplot of similarity scores for Voice B

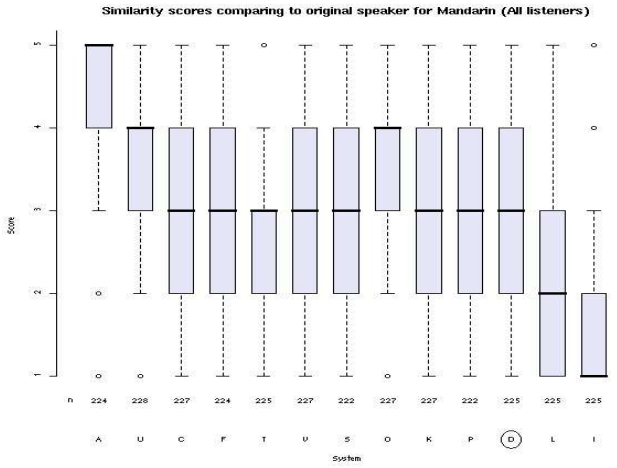

Figure 3: Boxplot of similarity scores for Voice C

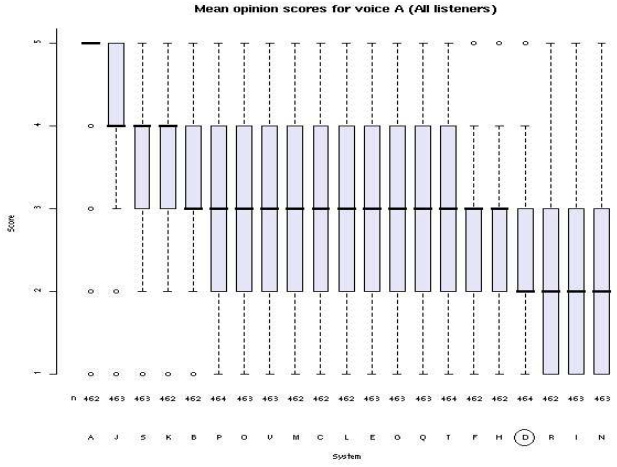

Figure 4: Boxplot of MOS scores for Voice A

7.3. Word Error Rate Test

Figure 7, 8 shows the results of word error rate (WER) tests of all systems for Voice A and B. Figure 9, 10 and 11 shows the result of character error rate (CER), Pinyin (without tone) error rate (PER) and Pinyin tone error rate (PTER) tests of all systems for Voice C (Mandarin). While system "D" can sound highest WER for voice A, it is interesting to note that MOS scores are not that lower and WER is highest penalty because of approximating the syllables.

8. References

[1] Black, A. and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis," in Proceedings of Eurospeech 97, vol2 pp 601-604, Rhodes, Greece.

[2] Dutoit, T., "An introduction to text-to-speech synthesis," Kluwer Academic Publishers, 1997.

[3] Hunt, A. and Black, A., "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", Proceedings of ICASSP1996, vol. 1, pp. 373-376, 1996.

[4] Greenberg, S., "Understanding speech undestanding: Towards a unified theory of speech perception," In Proc. ESCA Workshop on The Auditory Basis of Speech Perception, Keele Univ., UK, 1996.

[5] Hasuenstein, A., "Using syllables in a hybrid HMM-ANN recognition system," in Proceedings of EUROSPEECH, Rhodes, Greece, 1997, vol. 3, pp. 1203-1206.

[6] Kishore, S.P. and Black, A., "Unit Size in Unit Selection Speech Synthesis," in Proceedings of Eurospeech, Geneva, Switzerland, pp.1317-1320, 2003.

[7] Thomas, S., Rao, M. N., Murthy, H. A. and Ramalingam, C. S., "Natural Sounding TTS Based on Syllable-like Units," in Proceedings EUSIPCO, Florence, Italy, Sept 4-8, 2006.

[8] Black, A., Taylor, P. and Caley, R., "The Festival speech synthesis system," http://festvox.org/festival, 1998.

[9] Raghavendra, E. V, Yegnanarayana, B., Black, A. and Prahallad, K., "Building Sleek Synthesizers for Multi-Lingual Screen Reader," in Proceedings of Interspeech, Brisbane, Australia, September 2008.

[10] Prahallad, K., Black, A. and Mosur, R., "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in Proceedings of ICASSP, France 2006.

[11] The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speech_tools/.

[12] Gang-Janp Lin; Sau-Gee Chen; Wu, T., "High quality and low complexity pitch modification of acousticsignals," in Proceedings ICASSP-95, vol 5, pp. 2987-2990, 1995.

[13] http://en.wikipedia.org/wiki/Box_plot

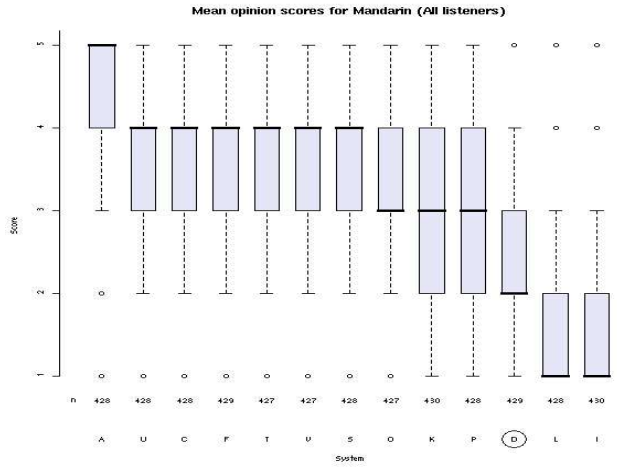

Figure 5: Boxplot of MOS scores for Voice B

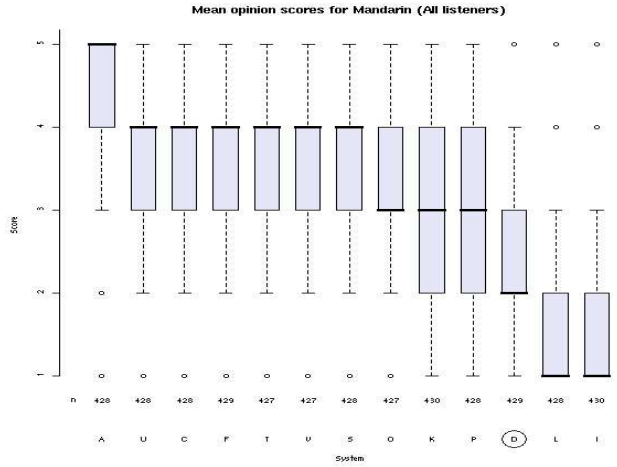

Figure 6: Boxplot of MOS scores for Voice C

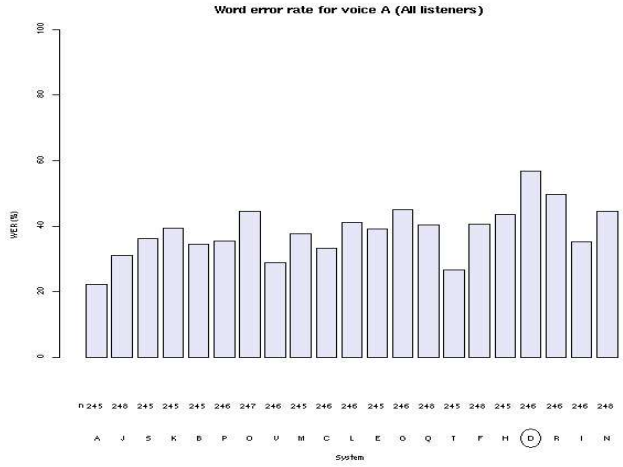

Figure 7: Mean WER for Voice A

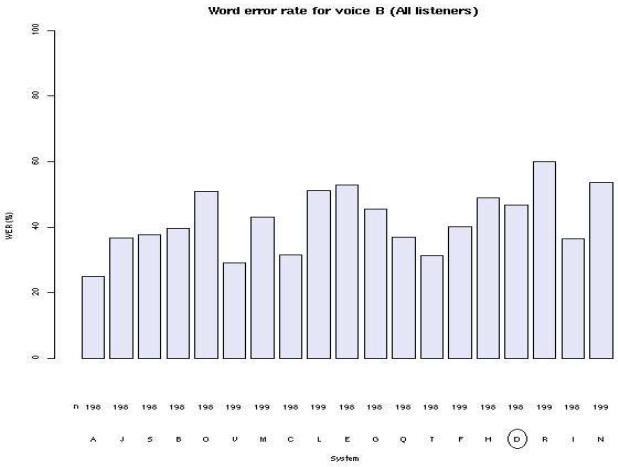

Figure 8: Mean WER for Voice B

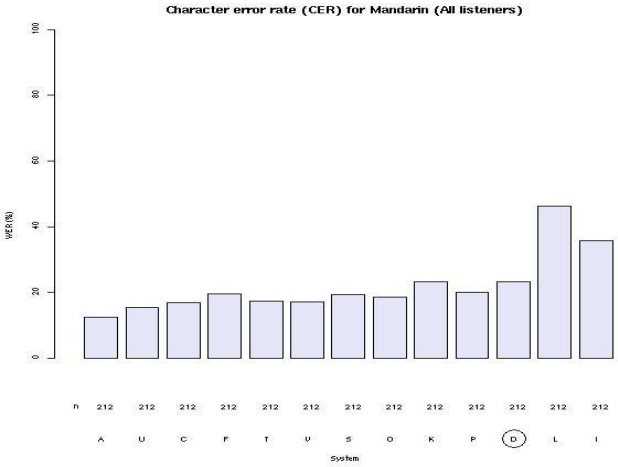

Figure 9: Mean CER for Voice C

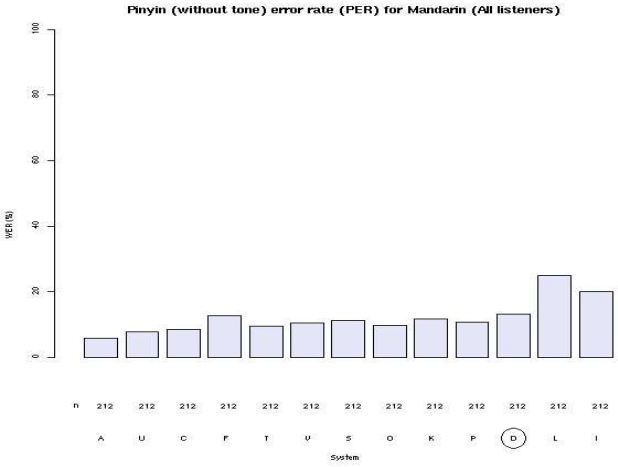

Figure 10: Mean PER for Voice C

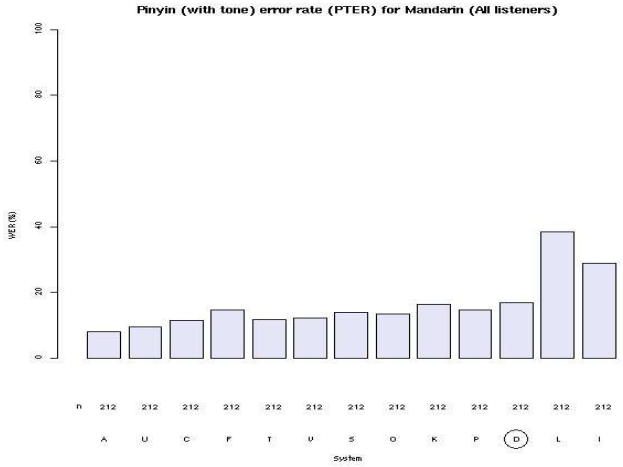

Figure 11: Mean PTER for Voice C