IIT Madras's Submission to the Blizzard Challenge 2014

Raghava Krishnan K[1], *S Aswin Shanmugam*[2], *Anusha Prakash*[3], *Kasthuri G R*[2], *Hema A Murthy*[2]

[1]Dept of Electrical Engineering, Indian Institute of Technology Madras, India
[2]Dept of Computer Science and Engineering, Indian Institute of Technology Madras, India
[3]Dept of Applied Mechanics, Indian Institute of Technology Madras, India

hema@cse.iitm.ac.in

Abstract

This paper details the work done by IIT Madras team for the Blizzard Challenge 2014. Two sets of tasks were given - the Hub tasks and the Spoke tasks. The details of the efforts put in to complete the Hub tasks have been presented in this paper. 2-3 hours of speech data along with text in utf-8 format were given. Two text to speech synthesis (TTS) systems - Unit selection synthesis (USS) and Hidden Markov Model based speech synthesis systems (HTS), were submitted for every language, except Assamese. Systems "I" are the primary systems and system "H" are the secondary systems. Syllables and phones were used as the basic units in USS and HTS systems, respectively. Techniques to develop letter to sound rules, predict prosodic phrase breaks, and perform syllable-level and phoneme-level segmentation have been described. To build speech synthesis systems for six Indian languages, a unified framework based on the common phone set and the common question set was used.

Modifications made to the system, post-submission of synthetic speech to the Blizzard Challenge, have also been briefly described.

Index Terms: USS, HTS, hybrid segmentation, common phone set, common question set

1. Introduction

The Blizzard Challenge is an evaluation to compare research techniques across the world for building corpus-based text to speech systems (TTS). It involves two sets of tasks - the Hub tasks, to build native language speech synthesisers and the Spoke tasks, to build bilingual (native language and English) systems. The focus of the 2014 Challenge was on Indian languages. The languages are Assamese, Gujarati, Hindi, Rajasthani, Tamil and Telugu. Given only 2-3 hours of recorded speech along with the corresponding text in utf-8, TTS systems have to be developed.

The IIT Madras team participated only in the Hub tasks. Syllable based unit selection synthesisers (USS) and phone based Hidden Markov Model based speech synthesisers (HTS) were developed. Two systems were submitted per language, except Assamese, where only the HTS system was submitted. Informal listening tests were conducted to determine the best system for each language. The primary systems (systems "I") submitted were Assamese HTS, Gujarati HTS, Hindi USS, Rajasthani USS, Tamil HTS and Telugu USS. The secondary systems (systems "H") submitted were Gujarati USS, Hindi HTS, Rajasthani HTS, Tamil USS and Telugu HTS.

Our main contribution to system building was in terms of:

1. Building syllable based USS systems for 5 Indian Languages.

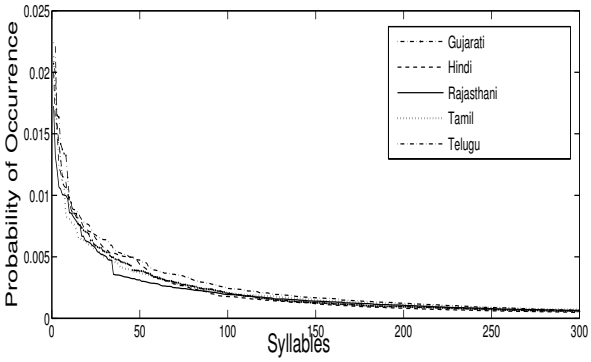

Figure 1: *Probability of occurrence of syllables*

2. Developing letter to sound rules (LTS), where a set of hand written rules specific to a language were designed.
3. A semi-automatic tool [1] that uses the group delay algorithm to obtain syllable labels which were later corrected manually, was used for the languages Hindi and Tamil.
4. A hybrid segmentation approach [2] which uses HMMs in tandem with the group delay algorithm to obtain accurate syllable and monophone labels was used to segment the speech waveforms for the languages Gujarati, Rajasthani and Telugu.
5. Also, a unified framework consisting of a common phone set and a common question set was used to build HTS systems for all languages [3].
6. It was also observed that adding positional context [4], geminate context and phrase break prediction [5] had a major impact on the quality of synthesis.

A major consortium effort on Unit Selection Speech Synthesis (USS) for Indian languages based on syllable-like units [6] is underway. Syllable based USS systems have been built and they have been shown to perform better than diphone based USS systems [7]. The reasons for using syllable as the unit for synthesis are:

- Syllables are the basic units of speech production [8]
- Indian languages are inherently syllable-timed
- Syllables capture co-articulation between phonemes adequately
- Syllables being relatively large units, there is a decrease in the number of concatenation points resulting in the synthesised speech sounding more continuous

- Also, the number of frequently occurring syllables in a language is not more than 300, and the distribution of the frequency of occurrence of these syllables follows a Zipf distribution as can be seen from Figure 1.

The HTS systems submitted to the blizzard challenge are based on the phoneme as the basic unit. A unified approach to building HTS systems has been used. A common framework has been developed, using which systems for new languages can be easily built. A Common phone set having a common representation for similar sounds across 13 Indian languages, and a common question set have been designed.

The paper is organised as follows: Section 2 details the design of LTS rules. Section 3 discusses the techniques used to perform prosodic phrasing for the languages Hindi and Tamil. The method used to segment the speech waveforms into syllables and monophones is described in Section 4. Section 5 and Section 6 describe the methods used to build the USS and HTS systems respectively. Section 7 discusses the results of the Blizzard challenge and Section 8 describes the efforts post submission of synthetic speech to the Blizzard challenge, that have gone into improving the quality of the systems.

2. Letter to Sound Rules

Letter to sound (LTS) rules play a very important role in building a high quality TTS system. The letter to sound rules for the systems built are a set of handwritten rules. The rules have been written to first break each word down into syllables. These syllables are then broken down into aksharas or monophones.

The rules for the Dravidian language systems (Telugu and Tamil) are fairly straightforward, as there is a one-to-one correspondence between the grapheme and phoneme representations. In Tamil though, there are certain symbols which represent more than one sound. These symbols therefore have to be tagged accordingly, depending on whether their pronunciation is voiced or unvoiced. These rules which are specific to Tamil have been listed below

- Symbols such as க, ச, ட, த, ப are tagged as unvoiced when they are at the beginning of a word or when they are appear as geminates.
- Symbols such as க, ச, ட, த, ப are tagged as voiced when they occur in between vowels or when they follow nasals

For the Aryan languages (Hindi, Gujarati, Rajasthani) and Assamese (Tibeto-Burman), the grapheme to phoneme relation is not one-to-one. This is mainly due to the effect of schwa deletion. Schwa refers to the mid-central vowel sound (rounded or unrounded) denoted by the IPA symbol ə. In Aryan languages, this is usually the short vowel sound /a/ which needs to be replaced by the halant symbol in certain cases. For example, in Hindi, the syllable /mil/ is written as /mila/ but is pronounced as /mil/. The rules for schwa deletion are also hand written. The pronunciation rules for Hindi have been derived from [9] and [10]. The rules for other Aryan languages and Assamese are also written similar to the rules of Hindi.

Once the letters forming the word have been appropriately tagged and after applying the rules of schwa deletion (for Aryan languages and Assamese), the rules of syllabification are used to break the words into C*VC* units. Once the words have been broken down into C*VC* syllables, they can be broken down into monophones and aksharas using one-to-one mapping with the phoneme representations of the graphemes forming the

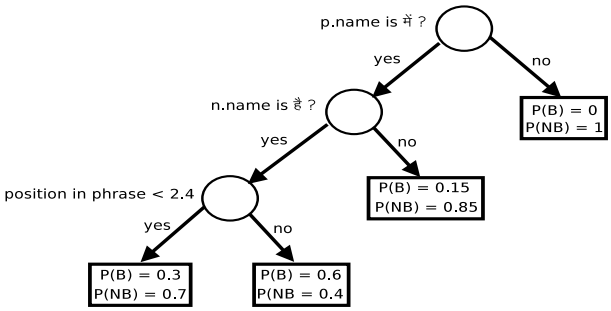

Figure 2: *Structure of CART*

syllables. Aksharas are (C*V) $\bigcup$ C units which are used as substitutes when a syllable is not found in the database. The CV units that are not present in the database are mapped to a similar sounding CV unit depending on their manners of articulation.

The rules to tag the syllables based on positional context and geminate context are also written in the LTS module. The contexts are used for pre-clustering and are described in more detail in Section 5.1.

3. Phrase break prediction

Prediction of phrase like breaks for Indian languages has been tackled in various ways in [5], [11] and [12]. For Hindi, there are certain words called case markers that usually signify the end of a phrase. Text corresponding to 5 hours of speech was used to identify such case markers and build the *classification and regression tree* (CART) to predict phrase breaks for Hindi. The identity of these case markers and their neighbouring words along with the position of the word in a phrase were the features used to build the CART. A portion of a CART tree has been shown in Figure 2 as an example. A few examples of these case markers are given in Table 1.

Table 1: *Examples of case markers for Hindi*

Case Marker	**No. of occurrences**	**No. of occurrences succeeded by pause**
है	2686	2491
थी	720	656
था	647	276
पर	1034	458
को	1857	622

For languages like Tamil, which are replete with complex words (formed by the concatenation of more than one word), low level linguistic features such as features at the syllable level have to be used. For such languages, word terminal syllables which acted like case markers were identified as in [11] and these word terminal syllables were used as cues to build the CART to predict phrase breaks for Tamil. Text corresponding to 5 hours of speech was used to identify such word terminal syllables. Identity of present, previous and next words, and the identity of the word terminal syllables of the present, previous and next words were the features used to build the CART using these 5 hours of text. A few examples of the word terminal syllables used for phrase prediction for Tamil have been given in Table 2.

Table 2: *Examples of word terminal syllables for Tamil*

Case Marker	No. of occurrences	No. of occurrences succeeded by pause
ேவ	262	132
னால்	532	332
வும்	317	136
க_	1085	373
யை	262	122

4. Segmentation of data

4.1. Syllable Level Segmentation

4.1.1. Group Delay Based Segmentation

For Tamil and Hindi, group delay based semi-automatic segmentation was performed. Syllable boundaries can be determined by performing group delay based processing of the short-time energy (STE) function [13]. STE can be used as an acoustic cue for finding syllable boundaries. But fluctuations in STE make it ineligible for using it directly. Therefore, group delay processing [14] is performed to smooth the STE. The algorithm for group delay based segmentation is given in [15]. The extent of smoothing is controlled by a parameter called the window scale factor (WSF). WSF depends on the syllable rate. As in TTS, only a single speaker is used for recording the data, WSF can be tuned to get the best approximate labeling output. But TTS requires boundaries to be consistent and accurate. Hence, a semi-automatic labeling tool [1] shown in Fig 3 was developed. Boundaries can be added, deleted and moved using this tool.

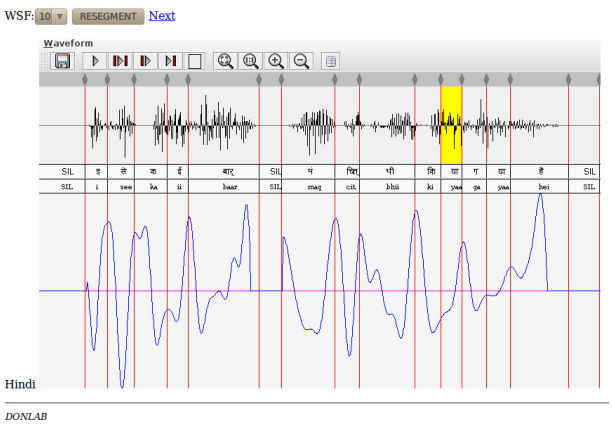

Figure 3: *Labeling a given Hindi sentence using the labeling tool.*

4.1.2. Hybrid Segmentation

For Telugu, Gujarati and Rajasthani, we performed segmentation automatically using a hybrid segmentation approach [2]. In this approach, HMM based segmentation and group delay based segmentation are performed in tandem to obtain accurate segmentation automatically. WSF in group delay based segmentation is tuned to over-estimate the syllable boundaries. This results in many spurious boundaries, but the correct boundaries are not misplaced. The group delay boundaries in the proximity of the boundaries given by HMMs are considered as correct boundaries [1]. Then, embedded reestimation is restricted to the syllable level and monophone HMM models are built. Using these models syllable level alignment is performed and boundary corrections are repeated again.

4.2. Phone Level Segmentation

The procedure explained in [16] was used to obtain phone boundaries from syllable boundaries. The popular approach for automatic segmentation is performing embedded training [17] of flat start initialized monophone HMMs followed by forced alignment [17]. As segmentation at syllable level is already available, embedded reestimation is restricted to the syllable level rather than the entire utterance and monophone HMM models are built. Using these HMMs, forced alignment is performed within the syllable to obtain segmentation at the phone level. These monophone labels are then concatenated to obtain akshara labels which are used for fallback.

Monophone labels for Assamese were obtained using HMM based segmentation.

5. Unit Selection Synthesis (USS)

The Unit Selection TTS systems are built using the Festival speech synthesiser [18] based on the Festvox framework. The systems are built using the syllable as the basic unit. A set of hand written pronunciation rules are written to split the text into syllables and CART is built using linguistic and acoustic context as described in [19]. The appropriate sequence of syllables for synthesising output speech are chosen by performing a Viterbi search through a set of target clusters. The set of units with optimum target and concatenation cost are chosen and concatenated to synthesise speech as explained in [20].

5.1. Pre-clustering

Syllables are tagged separately based on their position in a word [4]. The acoustic/prosodic properties of the same syllable varies across the position in a word. The syllables are therefore tagged as beg, mid or end depending on their position in the word.

Indian languages are replete with geminate consonants. Gemination happens when a spoken consonant is elongated for a longer period than a short consonant. These generally are points where segmentation errors could occur. Also, using a syllable from a geminate context to synthesise speech that is not in a geminate context would not sound appropriate. Hence, syllables that are part of geminate contexts are tagged separately, and lesser preference is given to using geminate syllables in a non-geminate context and vice-versa.

6. A common framework for building phone based HTS systems

When working in a multilingual framework, there is need for standardisation and a more generic approach to system building. Most Indian languages can be broadly divided into two categories- Aryan and Dravidian. And the script of these languages are derived from the Brahmi script. This can be observed in the arrangement of native script glyphs according to the place and manner of articulation. Further, these languages share phonetically similar sounds, around 10-12 vowels and 33-35 consonants. Observing these common traits across the lan-

[1] There are exceptions. Boundary corrections are performed only if the syllable boundary does not contain a fricative, nasal or a semi-vowel

Label	IPA	Hindi	Rajasthani	Gujarati	Assamese		Tamil	Telugu
					P	G		
a	/a/	अ	अ	અ	-	-	அ	అ
ax	/ɔ/	-	-	ઑ	অ	অ	-	-
aa	/a: /	आ	आ	આ	আ	আ	ஆ	ఆ
u	/u/,/ʊ /	उ	उ	ઉ	উ,ঊ	উ	உ	ఉ
eu	/uu /	-	-	-	-	-	ஊ	-
uu	/u: /	ऊ	ऊ	ઊ	-	ঊ	ஊ	ఊ
e	/e/	-	-	-	এ	এ	எ	ఎ
ee	/e: /	ए	ए	એ	-	-	ஏ	ఏ
ei	/ɛ: /	ऐ	ऐ	ઐ	-	-	-	-
ai	/aI/	-	-	-	-	-	ஐ	ఐ
oi	/oj/	-	-	-	ঐ	ঐ	-	-
j	/dʒ/	ज	ज	જ	জ,য	জ,য	ஜ	జ
jh	/dʒʱ /	झ	झ	ઝ	ঝ	ঝ	-	ఝ
n	/n/	न,ऩ	न	ન	ণ,ন	ন	ந,ண	న
nd		-	-	-	-	-	ந	-
l	/l/	ल	ल	લ	ল	ল	ல	ల
lx	/l̪/	-	ळ	ળ	-	-	ள	ళ
w	/ʋ /	व	व	વ	ব	ব	வ	వ
f	/f/	फ़	फ़	-	-	-	ஃப	-
q		◌ॅ	◌ॅ	◌ૅ	-	◌ৎ	-	◌ం
hq		◌ः	◌ः	◌ઃ	◌ঃ	◌ঃ	-	◌ఃా
mq		◌ॅं	◌ॅं	◌ૅં	◌ঁ	◌ঁ	-	-

Figure 4: *Partial common phone set*

guages, a common phone set has been designed, from which a common question set is derived. This is mainly for the purpose of building systems language-independently, in a more efficient and faster way.

6.1. Common Phone Set

A standard phoneme notation is used across the languages. This is the common phone set [3], which is the superset of phonemes across 13 Indian languages. In the common phone set, similar sounds across different languages are denoted by a label. This makes it easier when referring to phonemes in different languages.

A partial set is shown in Fig 4. The complete set can be found at http://www.iitm.ac.in/donlab/ilsl12.pdf. The labels are a sequence of Roman alphabets. No special characters are used in the label notation. Since the number of phonemes exceeds the number of alphabets, certain suffixes are used. Aspiration is denoted by suffix h, retroflex place of articulation is denoted by suffix x, etc. IPA symbols are used as reference. Assamese has separate phoneme and grapheme columns, so that the native script can be recovered from the transliterated text.

6.2. Common Question Set

A question set is required to perform tree based clustering in HTS. These questions are based on the acoustic-phonetics of the phonemes in a language. A very detailed analysis of the phoneme characteristics is required to design the question set. Continuing with the idea of a common phone set, a common question set has been designed. The common question set is a superset of questions across 6 Indian languages- Bengali, Hindi, Malayalam, Marathi, Tamil and Telugu [3]. It was found that these six languages covered most of the labels in the common phone set. Any additional label that was not present in the common question set, was included along with a similar sound.

The utf-8 text was first broken down in terms of the common phone set labels and segmentation of speech data was performed as mentioned in Section 4. Phone based HTS systems were then built using the common question set.

7. Results and Discussion

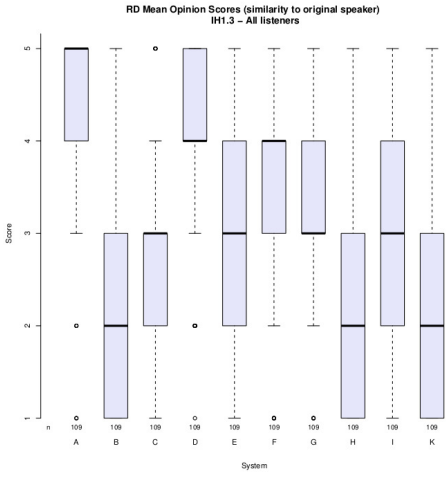

Figure 5: *Results for Hindi (a)*

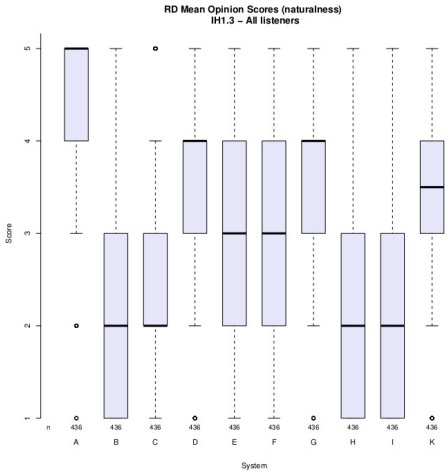

Figure 6: *Results for Hindi (b)*

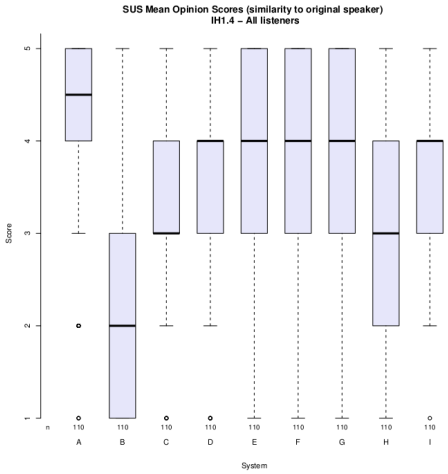

Figure 7: *Results for Rajasthani (a)*

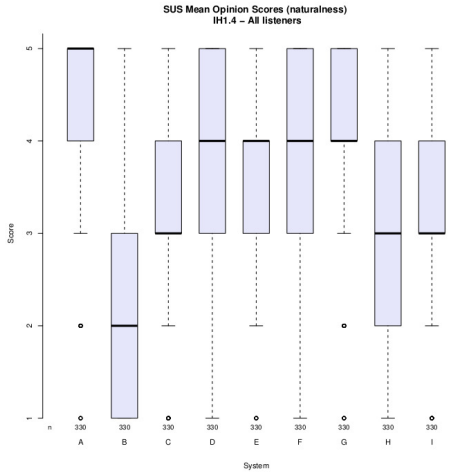

Figure 8: *Results for Rajasthani (b)*

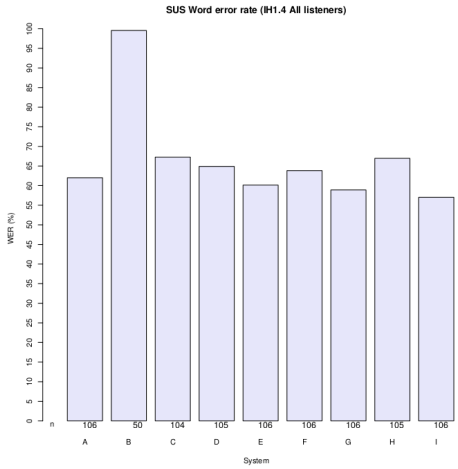

Figure 9: *Results for Rajasthani (c)*

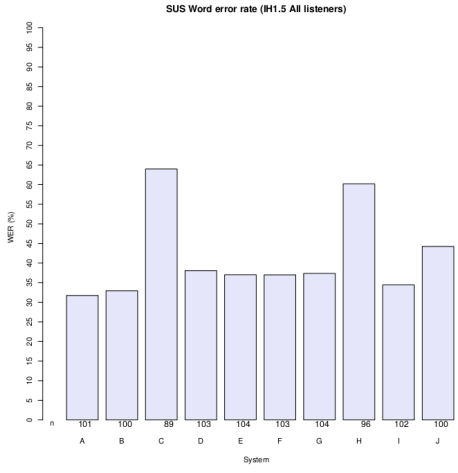

Figure 10: *Results for Tamil*

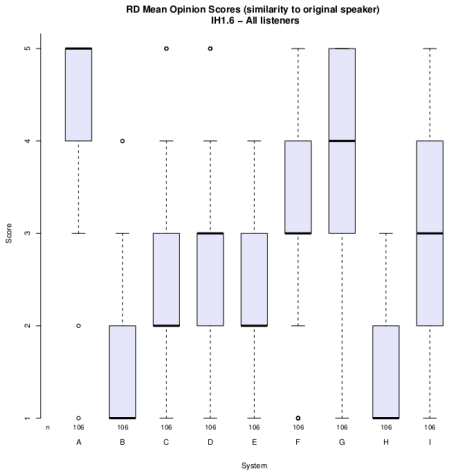

Figure 11: *Results for Telugu*

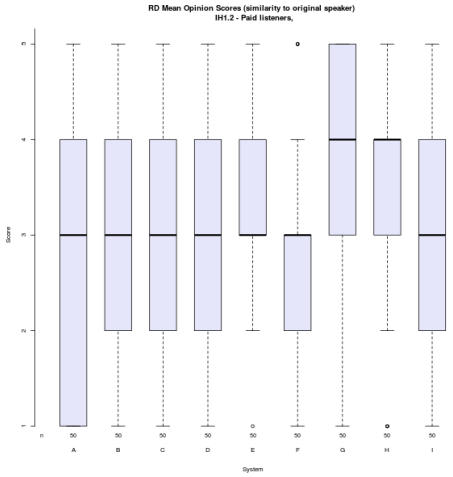

Figure 12: *Results for Gujarati*

Some of the results of the Blizzard Challenge 2014 are presented here. The primary systems submitted by our team are denoted by "I" and the secondary systems by "H". From the Figures 5, 7 and 11, it is apparent that for the systems "I", which are all USS systems, the similarity to the original speaker is preserved in most cases. Clearly, systems "H" for the same figures, which are HTS systems, have lower MOS scores. From Figure 12, similarity to the original speaker in Gujarati is preserved for both USS ("H") and HTS ("I"). But from Figures 6 and 8, it can be observed that naturalness suffers a great deal. This is mostly because of the problem of overlap. This phenomenon of overlap is caused mainly because of the concatenation of acoustically dissimilar units. From Figures 9 and 10, it can be seen that the WER is as good or in some cases even lesser than some of the other systems.

8. Post-Submission of synthetic speech

Some techniques that improved the synthesis quality, after submitting the synthetic speech for Blizzard Challenge, are mentioned here. Firstly, the transcription had to be correct. Indian

text generally do not have any punctuation marks. Commas were introduced only in places where the speaker had paused. This led to more accurate segmentation using the automatic hybrid segmentation method.

Secondly, the database was pruned to remove the "bad" units, so that these units were not selected for synthesis. This largely removed the errors in data recording and segmentation. Certain acoustic cues were used as pruning criteria. This led to an appreciable improvement in the quality of synthesised speech output. The fallback method was also modified. In case a syllable was not present in the database, aksharas were used as fallback units. An additional level of fallback to monophones in the absence of aksharas was also implemented. A pair comparison test was conducted [21], [16] for Hindi. 15 evaluators were asked to evaluate about 12 sentences each. The preference for systems built using a pruned database compared to the default system used for the Blizzard Challenge is quite clear from Table 3.

Table 3: *Results of Pair comparison test*

Score	A-B	B-A	A-B+B-A
Pruned database System	93.33	6.67	93.33

9. References

[1] P. Deivapalan, M. Jha, R. Guttikonda, and H. A. Murthy, "DONLabel: An automatic labeling tool for Indian languages," in *National Conference on Communication (NCC)*, February 2008, pp. 263–266.

[2] S. A. Shanmugam and H. Murthy, "A hybrid approach to segmentation of speech using group delay processing and HMM based embedded reestimation," in *Accepted for presentation in INTERSPEECH*, 2014.

[3] B. Ramani, S. Lilly Christina, G. Anushiya Rachel, V. Sherlin Solomi, M. K. Nandwana, A. Prakash, S. Aswin Shanmugam, R. Krishnan, S. Kishore, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *SSW8*, 2013, pp. 291–296.

[4] Venugopalakrishna.Y.R., Vinodh.M.V., H. A. Murthy, and C. Ramalingam, "Methods for improving the quality of syllable based speech synthesis," in *Proc. of Spoken Language Technology (SLT) 2008 workshop*, Goa, India, December 2008, pp. 29 –32.

[5] A. Bellur, K. B. Narayan, K. Krishnan, and H. A. Murthy, "Prosody modeling for syllable-based concatenative speech synthesis of Hindi and Tamil," in *Communications NCC, 2011 National Conference on*. IEEE, 2011, pp. 1–5.

[6] H. Patil, T. Patel, N. Shah, H. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. Kishore, S. Prasanna, N. Adiga, S. Singh, K. Anand, P. Kumar, B. Singh, S. Binil Kumar, T. Bhadran, T. Sajini, A. Saha, T. Basu, K. Rao, N. Narendra, A. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. Murthy, "A syllable-based framework for unit selection synthesis in 13 Indian languages," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, Nov 2013, pp. 1–8.

[7] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Proceedings of EUROSPEECH*, 2003, pp. 1317–1320.

[8] J. Cholin, N. O. Schiller, and W. J. Levelt, "The preparation of syllables in speech production," *Journal of Memory and Language*, vol. 50, p. 4761, 2004.

[9] M. Ohala, *Aspects of Hindi phonology*. Motilal Banarsidass Publishers Pvt. Ltd, 1983.

[10] M. Choudhury, "Rule-based grapheme to phoneme mapping for Hindi speech synthesis," *90th Indian Science Congress of the International Speech Communication Association (ISCA), Bangalore, India*, 2003.

[11] A. Vadapalli, P. Bhaskararao, and K. Prahallad, "Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for Indian languages," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 209–214.

[12] N. S. Krishna, "Multilingual Text-to Speech Synthesis," M S Thesis, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, May 2004.

[13] T. Nagarajan and H. A. Murthy, "Subband-based group delay segmentation of spontaneous speech into syllable-like units," *EURASIP Journal of Applied Signal Processing*, vol. 17, pp. 2614–2625, 2004.

[14] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its application to speech processing," *Sadhana*, vol. 36, no. 5, pp. 745–782, November 2011.

[15] V. K. Prasad, "Segmentation and Recognition of Continuous Speech," PhD Dissertation, Department of Computer Science and Engg., Indian Institute of Technology Madras, Chennai, India, May 2002.

[16] S. Aswin Shanmugam and H. A. Murthy, "Group delay based phone segmentation for HTS," in *National Conference on Communications 2014 (NCC-2014)*, Kanpur, India, Feb. 2014.

[17] S. Young and P. Woodland, "HTK: Speech recognition toolkit," http://htk.eng.cam.ac.uk/.

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

[19] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. Eurospeech '97*, Rhodes, Greece, September 1997, pp. 601–604.

[20] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (München, Germany)*, vol. 1, 1996, pp. 373–376.

[21] P. Salza, E. Foti, L. Nebbia, and M. Oreglia, "MOS and pair comparison combined methods for quality evaluation of text to speech systems," in *Acta Acustica*, vol. 82, 1996, pp. 650–656.