Building Sleek Synthesizers for Multi-Lingual Screen Reader

E.Veera Raghavendra[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, {awb,skishore}@cs.cmu.edu

Abstract

In this paper, we are investigating the unit size: syllable, half-phone and quarter-phone to be used for speech synthesis in multi-lingual screen reader in phonetic languages such as Telugu and non-phonetic language English. Perceptual studies show that syllable-level unit performs better for Telugu and half-phone units perform better for English. While syllable based synthesizers produce better sounding speech, the coverage of all syllables is a non-trivial issue. We address the issue of coverage of syllables through approximate matching of syllable and show that such approximation produces intelligible and better quality speech than diphone units. In this paper, we also propose a hybrid synthesizer within the framework of unit selection and also show that the hybrid synthesizer built from pruned database performs as well as hybrid synthesizer built from unpruned database.

Index Terms: speech synthesis, unit selection, unit size, database pruning, and hybrid speech synthesis.

1. Introduction

Our goal is to develop multi-lingual screen reader which can read contents in all official languages of India such as Hindi, Telugu, Tamil etc., including Indian English, and provide support for different computer applications (Email, Internet, Office software) using intelligible, human-sounding synthetic speech.

In concatenative Text-To-Speech (TTS) synthesis, the speech waveform is generated concatenating the pre-recorded 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 pre-recorded sentences uttered by a native and professional speaker of the language. With unit selection, speech synthesis becomes a problem of gathering, annotating, indexing and retrieving from a large database [1]. The size of a unit selection database could vary from 100 MB to 1 GB. Large unit selection databases [2] cause too much hindrance to download and install, and moreover often people in third-world countries (where we plan to deploy our multi-lingual screen reader) use machines with limited storage and CPU power. Thus the difficulty is to come-up a method of reducing the speech database with minimal loss of naturalness and intelligibility.

In recent years, HMM-based parametric speech synthesis method has widely been proposed and made significant progress [3][4]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [5] and the parameters are generated from HMMs under maximum likelihood criterion by using dynamic features [6]. MLSA [7] is used to synthesize the signal from the generated parameters. While the quality of the speech produced by HMM-based technique is intelligible and consistent but it is not as human-sounding as unit selection voices.

Several approaches for reducing the size of unit selection voices have also been proposed. The approach described in [8] leverages the LSM decomposition of information gathered across a given speech segment in the case of unit boundaries. Black and Taylor [9] clustered phonetic and prosodic context using a decision tree. They pruned synthesis units by discarding $1 \sim 4$ instances locating furthest from each cluster center. As a rule of thumb, pruning 20% of units usually makes no significant difference, while up to 50% may be removed without seriously degrading quality [10]. The method proposed in [11] is based on a unified HMM framework. Only instances, single or multiple, with the highest HMM scores are kept to represent a cluster of similar ones. The approach proposed in [12] uses a weighted vector quantization method that prunes the least importance instances.

In this paper, we experiment on unit size: syllable, half-phone and quarter-phone, from the perspective of reducing the unit selection database for phonetic languages (where orthography of the language is phonetic) such as Indian languages Telugu, Hindi, Kannada and Tamil etc., and non-phonetic language such as English. Perceptual studies show that syllable-like unit seems to be more suitable for phonetic languages and half-phone units for English. As a part of experimentation with half-phone and quarter-phone units, we show that a hybrid synthesizer could be built within the unit selection framework, and a pruned database performs as good as unpruned database.

The rest of the paper organized as follows. Description of speech databases used in the experiments are given in section 2. Experiments on unit size are discussed in section 3. Database pruning is described in section 4.

2. Speech Database Used

The quality of unit selection voices depends to a large extent on the variability and availability of representative units. It is crucial to design a corpus that covers all speech units and most of their variations in a feasible size. The speech database used for Telugu is recorded by a female speaker and the duration of this speech data is approximately 2 hours. Each recording utterance contains approximately 15 words. That has led to 2 hours of speech recording. All sentences are recorded in a professional studio and the sentences are read in relax reading style, which is between "formal reading style" and "free talk style", in moderate speed. Recordings are performed in a sound proof room with close-talking microphone. For English, the *Roger Arctic* voice provided in Blizzard 2008 synthesis challenge [18] has been used.

3. Experiments on Unit Size

The following sub sections describes the experiments on different unit size.

3.1. Syllable-like as Unit

In [13], on a Hindi unit selection voice it was observed that the syllable unit performs better than diphone, phone and half phone, and seems to be a better representation for languages such as Hindi. It was also observed that the half phone synthesizer performed better than diphone and phone synthesizers, though not as well as syllable. It should be noted that Hindi is a phonetic language and further experiments on unit size are needed in the case of non-phonetic languages such as English. Moreover, units smaller than half-phones also need to be investigated.

A *syllable* can be typically of the following form: V, CV, CCV, CVC, CCVC. It can be represented as C*VC*, where C is consonant and V is vowel. All Indian language scripts have a common phonetic base, and a universal phoneset consists of about 35 consonants and about 15 vowels. Theoretically possible syllable combinations in Indian language with V, CV, CCV, CVC, CCVC representation are 680415. 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 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 units used in this work.

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 go to step 3
3. if /C*$_l$/ is null go to 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 go to 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*$_r$'/

To evaluate the syllable based synthesizer which employs approximate matching, we have conducted subjective and objective evaluations in comparison with a diphone based synthesizer. The subjects participated for Telugu are native speakers and are also fluent in English as we don't have any native UK accent speakers in our group. We selected a set of 10 sentences from Telugu and English news bulletin. Two or three syllables of each utterance are approximated using nearest syllable-like unit approach. The five persons who participated in these perceptual tests do not have any experience in speech synthesis. Each listener is subjected to MOS i.e score between 1 (worst) to 5 (best) and AB-Test i.e the same sentence synthesized by two different synthesizers is played in random order and the listener is asked to decide which one sounded better. They also had the choice of giving the decision of equality. As a part of objective evaluations Mel-Cepstral Distortion (MCD) [16] are calculated between original and synthesized wave files. Lower the MCD value the better it is. Informally we have observed an absolute difference of 0.2 in MCD values of two synthesizers in comparison indicate that the synthesizers produce perceptually different voices.

The results shown in Table 1 indicate the syllable based synthesizer employing approximate matching performs better than diphone based synthesizer for Telugu. The MOS scores in Table 1 show that approximate matching does not degrade the intelligibility of synthesis in comparison with diphone synthesis. The similar technique can also be applied to rest of the Indian languages as they have common phonetic base. This indicates that approximate matching is a useful technique for developing syllable based synthesizers in Indian languages with out worrying about back-off synthesizers using lower level units. However, for English, diphone based synthesizer seems to be better than syllable. The poorer performance of syllable level unit for English could be due to the fact that syllabification in English is not as simple as in phonetic languages. In this work we have used syllabification as specified in Unilex lexicon that comes with *Roger* voice. Further experiments need to be conducted to study how the errors in syllabification process affect the building of syllable based synthesizer in English and also with large number of listening subjects.

Table 1: *Syllable (Syl) Vs Diphone.*

Test	Telugu			English		
	Syl	Diphone	Similar	Syl	Diphone	Similar
AB-Test	20/50	15/50	15/50	14/50	32/50	4/50
MOS	2.63	2.592	-	3.19	4.07	-
MCD	5.563	5.812	-	4.875	3.325	-

3.2. Half-Phone Size Unit

In [13], half-phones were considered only for vowels. But in this paper, we are investigating the synthesizers where half-phones are created for all the phones. In implementing half-phone synthesizer, each phone is represented by two half phones. Two phone symbols are defined for each phone in the phoneset, for example phone /m/ is represented by /m_1/ and /m_2/. Where /m_1/ represents first half-phone and /m_2/ represents second half-phone. Labels at half phone level are derived by equally dividing the phone segment into two half phones. The lexicon parser is also modified accordingly, to generate appropriate half-phone strings. To imitate the diphone synthesizer, individual trees are built for each half-phone by tagging previous half-phone. In later stage duration models were also built for each half-phone.

Table 2 shows the subjective and objective evaluation of syllable and half-phone based synthesizers. Please note that the five subjects participated in this perceptual study are different

from the subjects participated in Table 1. Different subjects are participated for different experiments to avoid any bias the subjects might hold. From Table 2, we observe that syllable based synthesizer performs better than half-phones for Telugu while half-phone based synthesizer performs better for English.

Table 2: *Syllable Vs Half-Phone.*

Test	Telugu			English		
	Syl	Half Phone	Similar	Syl	Half Phone	Similar
AB-Test	25/50	16/50	9/50	16/50	27/50	7/50
MOS	2.93	2.7	-	3.37	3.7	-
MCD	5.563	5.707	-	4.875	4.426	-

3.3. Quarter-Phone Size Unit

In implementing quarter-phone synthesizer, each phone is represented by four quarter phones. Labels at quarter-phone level were derived by equally dividing the phone segment into four parts. As explained in the Section 3.2, phoneset, lexicon and labels are modified accordingly.

Table 3 shows the subjective and objective evaluation of half-phone and quarter-phone synthesizers. It could be observed that half-phones perform better than quarter-phone for Telugu and English.

Table 3: *Half-Phone Vs Quarter-Phone.*

Test	Telugu			English		
	Half Phone	Quarter Phone	Similar	Half Phone	Quarter Phone	Similar
AB-Test	23/50	10/50	17/50	16/50	16/50	18/50
MOS	2.7	2.5	-	3.57	3.41	-
MCD	5.707	5.963	-	4.426	4.649	-

3.4. Hybrid Technique of Synthesis for Half-Phone Size Unit

The listening experiments on half-phone based synthesis indicated that the half-phones had many perceptual discontinuities due to large number of joins. In order to produce a smoother version of half-phone based synthesis we investigated the use of Mel-Log Scale Approximation (MLSA) [7] based synthesis technique used in CLUSTERGEN [17]. The idea is to let the unit selection frame work select the appropriate half-phone units. Once the selection of units has been made, the corresponding MCEP and F0 parameters from the original features are used to synthesize the utterance. Such technique could be viewed as hybrid technique of synthesis, as it captures natural trajectories at half-phone unit level, and also allows to modify F0 during MLSA synthesis. It should be noted that the proposed hybrid method relies primarily on unit selection and thus differs from other implementations of hybrid synthesizers such as in [14][15].

In order to evaluate the hybrid approach, we compared the half-phone based hybrid synthesizer with CLUSTERGEN for Telugu and English. The subjective and objective analysis shown in Table 4 indicates that the proposed hybrid synthesis method performs better than CLUSTERGEN for Telugu and English as it adapts naturalness and smoothness from unit selection and statistical parametric techniques respectively.

Table 4: *CLUSTERGEN (Clu) Vs Half-Phones using MLSA (HP MLSA).*

Test	Telugu			English		
	Clu	HP MLSA	Similar	Clu	HP MLSA	Similar
AB-Test	16/50	30/50	4/50	18/50	22/50	10/50
MOS	2.8	3.4	-	3.1	3.6	-
MCD	5.532	5.008	-	4.325	4.011	-

4. Pruning on Half-Phone Size Unit

For the purposes of multi-lingual screen reader, we need slim and faster synthesizers as it supports multiple languages and to deploy it on low-end machines. So, we tried to reduce the database size to make synthesizer faster and slimmer. First we constructed the half-phone based context dependent decision-tree [9], a binary tree with categorical questions associated with each branching node. The categorical questions could be contextual features such as previous unit, next unit, and acoustic-phonetic features such as stress, onset, coda, vowel, and articulatory positions. The decision trees are generated to obtain minimum within-unit distortion for each split. This criterion would assure minimum spectral variations for the context-dependent phones within each cluster or leaf node. Therefore our hypothesis was that the context-dependent phone cluster could be substituted with one unit which is prosodically balanced. The advantage of a single-instance cluster is its simplicity and compact size.

4.1. Deducting prosodically consistent unit

As described above, every leaf node in the decision trees represent a similar features. The selection of consistent unit from the cluster can be done using prosodical features such as pitch, duration and energy.

Each decision tree has N clusters or leaf nodes and each cluster has M candidates as shown in below equations.

$$L = l_1, l_2, l_3, ..., l_N \tag{1}$$

$$C = c_1, c_2, c_3, .., c_M \tag{2}$$

Prosodic features, pitch, duration and energy are extracted from each candidate and arranged in a matrix format. Each value of the matrix is normalized with maximum value of corresponding column as range of each feature varies. Later, mean is calculated over all the feature vectors and considered as the threshold. Euclidean Distance is estimated between the candidate feature vector and mean vector. A statistically consistent unit is deduced by choosing smallest distance candidate unit as shown in equation 4.

$$d_i = \sqrt{\sum_j (c_{ij} - m_j)^2} \tag{3}$$

where m is the mean vector

$$D = argmin_i(d_i) \tag{4}$$

Figure 1 shows the 3D diagram of the candidates on one cluster (in blue color), centroid (red color) of the cluster and unit which is closest to the centroid (green color). Table 5 shows the size of the databases reduce to 11034 units for Telugu and 3213 units for English after applying the pruning technique. Table 6 shows the subjective and objective analysis of half-phone based

synthesizers built from pruned and unpruned speech databases. The perceptual scores shown in Table 6 indicate that the half-phone synthesizer build from pruned database performs better than the synthesizer built from unpruned database. It also suggests that the pruning technique employed here is good at selecting units which are prosodically balanced. The higher MCD values for synthesizer built from pruned database could be justified from the fact that a single instance is chosen as a representative of 20 or more units, which results in quantization error.

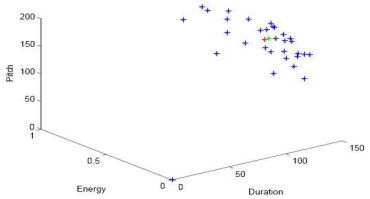

Figure 1: Scatter diagram of instances on one leaf node for half-phone /s_1/

Table 5: *Statistics of the databases before and after pruning.*

Language	No.Of.Units in original database	No.Of.Units in scale down database	% of units in the database
Telugu	342388	11034	3.22
English	79271	3213	4.05

Table 6: *Half-Phones using MLSA (HP MLSA) Vs Half-Phone using MLSA on pruned database (HP Avg MLSA).*

Test	Telugu			English		
	HP MLSA	HP Avg MLSA	Similar	HP MLSA	HP Avg MLSA	Similar
AB-Test	16/50	20/50	14/50	16/50	19/50	15/50
MOS	3.2	3.7	-	3.5	3.8	-
MCD	5.008	5.121	-	4.011	4.225	-

5. Conclusion

In this paper, we addressed the issues of unit size, going for hybrid synthesizer and database pruning for speech synthesis. We built Telugu and English synthesizers using different techniques: syllable, half-phone, quarter-phone, CLUSTERGEN and hybrid techniques. We conducted subjective and objective evaluations to evaluate each of these synthesizers in comparison with other. The evaluation on syllable based synthesizer indicate that the approximate matching of syllables is a useful and viable technique to build syllable based synthesizers for Indian languages without requiring any back off synthesizers. We have also observed that half-phone units seems to perform better than syllable units for English. In comparison with quarter-phones, half-phones performed better for Telugu and English. The proposed hybrid technique hinging on the unit selection framework seem to be perform better than CLUSTERGEN. Finally, we have showed the hybrid synthesizer built from pruning based on prosodic features performs better than the synthesizers built from unpruned thus making the possibly of using the synthesizer in multi-lingual screen readers on low end machines.

6. Acknowledgments

We would like to thank Venkatesh Keri, Gopala Krishna, and Srinivas Desai for useful discussion and suggestions during this work. We also thank all the people of LTRC and graduate students of IIIT Hyderabad for their participation in the perceptual tests.

7. References

[1] Huang, X., Acero, A., and Adcock, J., "WHISTLER: A Trainable Text-to-Speech System", Proceedings of ICSLP1996, Philadelphia, vol. 4, pp. 2387-2390, 1996.

[2] 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.

[3] Zen, H., and Toda, T., "An Overview of Nitech HMMbased Speech Synthesis System for Blizzard Challenge 2005", in Proc. of Eurospeech, pp. 93-96, 2005.

[4] Black, A., Zen, H., and Tokuda, K, "Statistical Parametric Synthesis", ICASSP 2007, pp. IV-1229-IV-1232, Hawaii, 2007.

[5] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. iFlytek USTC of Eurospeech, pp. 2347-2350, 1999.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., "Speech parameter generation algorithms for hmm-based speech synthesis", in Proc. of ICASSP, pp. 1315-1318, 2000.

[7] Imai, S., "Cepstral analysis synthesis on the mel frequency scale", in ICASSP 83, 1983, pp. 93-96.

[8] Bellegarda, J. R., "LSM-based unit pruning for concatenative speech synthesis", in Proc. Int. Conf. Acoust., Speech, Signal Process., Honolulu, HI, Apr. 2007, pp. IV-521-IV-524.

[9] Black, A. W., and Taylor, P. A., "Automatically Clustering Similar Units for Units Selection in Speech Synthesis", Proceedings of Eurospeech1997, vol. 2, pp. 601-604, 1997.

[10] Black A. W., and Lenzo, K., "Optimal Data Selection for Unit Selection Synthesis", in Proc. 4th ISCA Speech Synth. Workshop, Perthshire, Scotland, paper 129, August 2001

[11] Hon, H., Acero, A., Huang, X., Liu, J., and Plumpe, M., "Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems", Proceedings of ICASSP1998, vol. 1, pp. 293-296, 1998.

[12] Kim, S. H., Lee, Y. J., and Hirose, K., "Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization", Proceedings of Eurospeech2001, pp. 2231- 2234, 2001.

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

[14] Zhen-Hua Ling, and Ren-Hua Wang, "HMM-Based Unit Selection Using Frame Sized Speech Segments", Interspeech 2006 . ICSLP, Pittsburgh, PA, pp. 2034-2037, 2006.

[15] Black, A., Bennett, C., Blanchard, B., Kominek, J., Langner, B., Prahallad, K., and Toth, A. (2007). "CMU Blizzard 2007: a hybrid acoustic unit selection system from statistically predicted parameters Blizzard Challenge 2007 Workshop", Bonn, Germany, pp. 1-5, 2007.

[16] Toda, T., Black, A., and Tokuda, K. "Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech", pp 31-36, 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA.

[17] Black, A. (2006), "CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling", Interspeech 2006 - ICSLP, Pittsburgh, PA, pp. 1762-1765.

[18] http://www.synsig.org/index.php/Blizzard_Challenge_2008