A Probabilistic Approach to Selecting Units for Speech Synthesis Based on Acoustic Similarity

Anjana Babu*, Raghava Krishnan K†, Anil K Sao* and Hema A Murthy‡
*School of Computing and Electrical Engineering Indian Institute of Technology Mandi,
†Department of Electrical Engineering Indian Institute of Technology Madras,
‡Department of Computer Science and Engineering Indian Institute of Technology Madras,
Email: anjana.babu@gmail.com, raghav1105@gmail.com, anil@iitmandi.ac.in, hema@cse.iitm.ac.in

***Abstract*—Most unit selection synthesisers sound quite natural when the database consists of a number of realisations of the same sound unit from a large number of contexts. A common problem observed with these synthesisers is unexpected prosody when a new context is presented in the text.**

The objective of this paper is to address this issue and select appropriate units that are relevant to a specific context. Text-to-speech synthesisers propose a number of different features based on the linguistic context to select units. The key contribution in this paper is that the acoustic context rather than the linguistic context is crucial for improving naturalness. A probabilistic framework is proposed for selecting units based on an acoustic framework. Reducing the variability in acoustic context improves both naturalness and intelligibility. Since the context is only specified by acoustics, it can be applied to any language and perhaps even multilingual synthesis. The proposed approach has been tested on 2 Indian languages. An improvement of up to 21.9% in DMOS and 73.93% in WER relative to the conventional system that uses linguistic criteria is observed.

I. Introduction

State-of-the-art high quality speech synthesisers using concatenative waveform synthesis [1] perform very well as this method of synthesis is based on a huge repository of multiple realisations of a unit in different contexts. Since the choice of unit is mostly based on linguistic context, the resulting speech sounds occasionally discontinuous for unseen linguistic contexts. Although phone based approaches to Text to Speech Synthesis (TTS) are more common, a major consortium effort on Unit Selection Speech Synthesis (USS) for Indian languages is based on syllable-like units [2]. Although the syllable captures most of the intra-syllable coarticulation effectively, and only positional context is required for synthesis, variations in inter-syllable duration play an important role in the quality of the output produced. In particular, when the inter-syllable pause and duration of consecutive syllables is not systematic[1], the synthesised sentences appear to have unnatural prosody. The primary objective of this paper is to study the effect of prosodic parameters namely, f_0, duration and energy of adjacent units during synthesis. In particular we include features of adjacent units to select units that are acoustically similar. To set the context, a brief recap of syllable-based units for speech synthesis for Indian languages is now given.

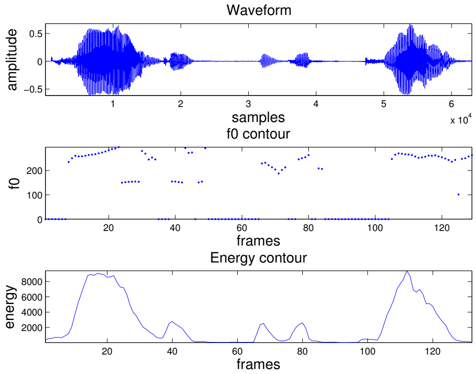


Fig. 1: An example of a speech with artifacts for the text *bhaag bhii khel* which has artifact near *bhii*

It has been shown in [2] and [3] that syllables are the units best suited for synthesis of Indian languages. Syllables being longer in duration, the number of concatenation points is considerably reduced. When units are unavailable for a specific context in the repository, the nearest unit is chosen based on linguistic criteria. This leads to significant prosodic discontinuities in the synthesised output. This is indicated in Figure 1, where discontinuities are observed in f_0 and energy contours, for an utterance synthesised using a conventional synthesiser. Figure 1 shows the discontinuities in pitch, energy and duration of a sentence synthesised using the conventional Indian language unit selection synthesiser.

In previous efforts on building syllable-based synthesisers, the primary effort was on developing appropriate clustering algorithms for syllables. The standard algorithms were not adequate owing to the fact that the syllable, unlike the

[1] More details will follow in the subsequent sections

978-1-4799-2361-8/14/$31.00 © 2014 IEEE

phoneme, contains a significant context. It was observed, that primarily three different contexts were required for clustering, namely - begin, middle and end[2] [4]. There are also approaches that deal with using larger amounts of linguistic information to select units from contexts similar to the target context [5]. [6] proposes a continuity metric where the pitch tracks for successive units in a word are compared, and the closest matching units are selected. [4], [6] and [7] deal with predicting pauses using textual information to make the speaking pattern of the synthesised speech resemble that of the speaker whose speech data is being used. Notwithstanding the research efforts on large context phone based synthesis, it is observed that syllable-based synthesisers are on par with large context phone based approaches. It is observed that syllable-based synthesisers work well for a majority of the cases, but occasionally, a single syllable that is wrongly chosen affects the quality of synthesis significantly.

A number of perceptual experiments were performed where other units that produced better quality were manually chosen from the repository to replace the offending unit. This was followed by a careful study as to why these issues were observed. It was observed that continuity is based on the quality of adjacent units. Speech being produced by an inertial system, sudden prosodic changes are not perceptually pleasing.

Therefore, the following methodology was adopted: Differences in f_0, energy and duration of consecutive pairs of units were observed. The differences were converted to probability density functions. The units were then chosen based on the values of duration difference, average energy difference and/or average f_0 difference that best fit the distribution.

The rest of the paper is organised as follows. Section II gives an introduction to the conventional USS based systems. Section III describes the approach adopted as an improvement to the USS approach. Section IV gives details of the recorded data used to conduct the experiments. Section V describes the experiments performed. Section VI and VII are results and discussion and Section VIII concludes the work.

II. Unit Selection TTS Systems

Unit selection based speech synthesisers are based on waveform concatenation. To synthesise a sentence from a USS database, the best sequence of units is chosen by minimising the *target cost* and the *concatenation cost*. Similar to the automatic speech recognition framework, the state occupation cost is given by the target cost which is the distance between a candidate unit and the target unit. The concatenation cost is the cost which estimates the quality of concatenation of two consecutive units [8].

Indian Language TTS systems are based on the syllable as a basic unit [2], [4]. Syllable level segmentation is obtained using the *group delay algorithm* which exploits the low energy regions at syllable boundaries to obtain accurate segmentation [9]. The syllables are then clustered according to linguistic context and *classification and regression trees* are built. During synthesis, the sequence of units with minimum target and concatenation cost are selected by performing a Viterbi search through the *classification and regression tree* [4] [8] . To find the optimal path through the candidate clusters, the expression

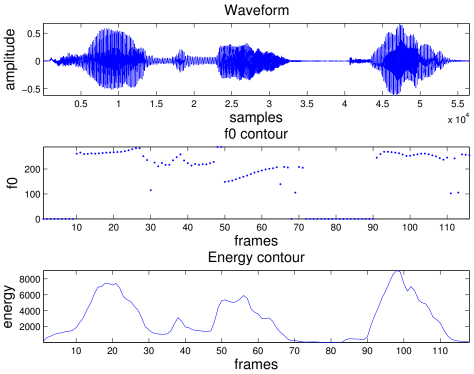


Fig. 2: An example of natural speech for the text *bhaag bhii khel*

$$\sum_{i=1}^{N}[Cdist(S_i) + W \times Jcost(S_i, S_{i-1})]$$

is minimised. $Cdist(S_i)$ is the target cost which is the distance of the unit from the center of the cluster and $Jcost(S_i, S_{i-1})$ is the concatenation cost between every syllable and its previous syllable. W can be used to weight the concatenation cost over the target cost. N is the number of syllables in the utterance [10].

III. Probabilistic Approach to Unit Selection

Most of the concatenative speech synthesisers produce speech which sounds close to natural speech but contains artifacts. Artifacts include short distortion in the audio, click, short high-pitched whine, overlap, simultaneous speech, etc [11]. There has been previous work on removing artifacts, especially using signal processing techniques [12]. But signal processing approaches are not preferred primarily because these signal processing corrections in certain contexts often lead to distortions in other contexts.

An alternative approach would be to avoid processing the speech units, and selecting units which do not introduce artifacts that are perceptible. This requires that such artifacts be studied, and appropriate criteria be introduced while selecting units from the database. An example of speech with artifacts is shown in Figure 1. The figure shows the energy and f_0 plots for the text *bhaag bhii khel* which has three syllables - *bhaag*, *bhii* and *khel*.

As mentioned earlier, a number of experiments were performed to determine the level of f_0 and energy variation that

[2] These correspond to the location of the syllable within a word

can be tolerated by the human ear at the point of concatenation. It was observed that if the variations in average f_0 and average energy of adjacent units are reduced, the discontinuities are minimised, due to the effects of persistence of hearing [13].

Syllables are sound units of the form C*VC*, where C stands for consonant and V for vowel. The nucleus of the syllable which is a vowel, constitutes the voiced part. Syllable being the basic unit of sound production, a slight variation in the pitch from one syllable to the next would not affect the quality of synthesised speech to a large extent.

The boundary of a syllable is usually characterised by a region of low energy irrespective of whether the syllable ends with a voiced or unvoiced sound. If there is a large difference in the energies between two adjacent syllable nuclei, it leads to unnaturalness. Figure 2 shows a waveform of natural speech. On comparing the energy contours of the two waveforms (Figure 2 and Figure 1), we can see that there are large fluctuations in energy from one syllable to the next in the synthesised waveform. Similar results were observed in the case of other waveforms as well. We can therefore conclude that correcting the energy pattern can perhaps lead to a major improvement in the naturalness of synthesised speech.

Another factor that introduces artifacts into synthesised speech is the large variation in the duration of units selected for synthesis. Syllable timed languages are characterised by minimal variation in the duration of syllables across an utterance. The earlier experiments suggested that the utterance sounded more or less natural if the syllable timing is preserved across the utterance.

An approach has been developed to select units in such a way that the difference in average f_0, average energy and duration of adjacent units are optimal. For each candidate unit i of each syllable n, a score is computed which is given by:

$$\begin{aligned} S_{(i,n)} = {} & P\big((d_{(i,n)} - d_{(j,n-1)})|N_d(\mu_{n-1,n}, \sigma_{n-1,n})\big) \times \\ & P\big((e_{(i,n)} - e_{(j,n-1)})|N_e(\mu_{n-1,n}, \sigma_{n-1,n})\big) \times \\ & P\big((f_{0_{(i,n)}} - f_{0_{(j,n-1)}})|N_{f_0}(\mu_{n-1,n}, \sigma_{n-1,n})\big) \end{aligned} \quad (1)$$

where $(d_{(i,n)} - d_{(j,n-1)})$ is the difference in duration between the i^{th} candidate of the n^{th} syllable and j^{th} candidate of the $(n-1)^{th}$ syllable, $(e_{(i,n)} - e_{(j,n-1)})$ is the difference in average energy, $(f_{0_{(i,n)}} - f_{0_{(j,n-1)}})$ is the difference in average f_0 and $\mathcal{N}_{f_0}(\mu_{n-1,n}, \sigma_{n-1,n})$, $\mathcal{N}_d(\mu_{n-1,n}, \sigma_{n-1,n})$, $\mathcal{N}_e(\mu_{n-1,n}, \sigma_{n-1,n})$ are the distributions of difference in average f_0, duration and average energy respectively. The sequence of units is selected such that the score is maximised over the entire utterance.

The motivation behind conducting a study on duration difference was, that in concatenative synthesis, if a long unit is followed by a short unit, it is perceived as an overlap, or at times, as though two different speakers are speaking simultaneously (echo). But ideally, most of the Indian languages are syllable timed and overlap between units should not occur. In practice, it is not possible to speak with exactly the same duration throughout database. So that calls for some amount of duration difference that can be considered as natural. Various experiments were performed to analyse these effects. The details of the experiments performed are given in Section V.

IV. Data Used

Data from 2 languages (1 Dravidian and 1 Aryan) are used to build systems to test the new approach. The languages are Hindi and Tamil. Details about the amount of data used is given in the table below. The data was recorded in a completely noise free studio environment from a native speaker of the language at 48KHz sampling rate at 16 bit PCM resolution.

TABLE I: Language Databases Used

Language	Hours of Data	Speaker	Language Family
Hindi	6.45	Male	Aryan
Tamil	10	Female	Dravidian

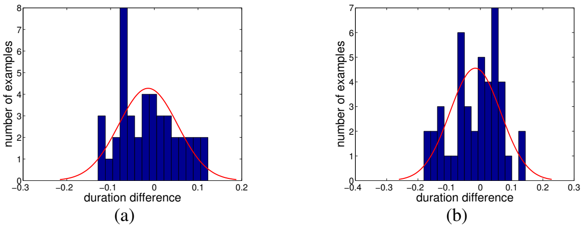
V. Experiments

Various experiments were performed with different combinations of the features - average energy difference, average f_0 difference and duration difference.

1) Duration difference between units
2) Average pitch difference
3) Average energy difference
4) Duration difference and average pitch difference
5) Duration difference and average energy difference
6) Average pitch difference and average energy difference
7) Duration difference, average pitch difference and average energy difference

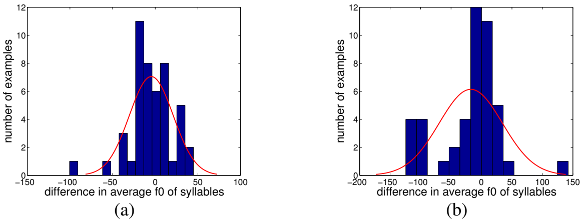
To conduct the experiments, the duration of syllables were computed. The duration difference between every pair of successive units in the sentence was computed. The probability of the duration difference of units at syllable position n and $n+1$ was estimated using the statistics obtained from natural data. The best possible sequence of units was obtained using Viterbi algorithm. A similar procedure was adopted for the other features listed above. Note that the traditional target and concatenation cost functions are not used for selecting units.

An important issue that has to be taken into account is the effect of phrase on f_0, energy and duration. Various works on prosody models suggest that there is a direct influence of phrase on the f_0 contour [14]. Intuitively, this suggests that the differences in the values of f_0 and energy of syllables at phrase boundaries could be larger compared to syllables located at



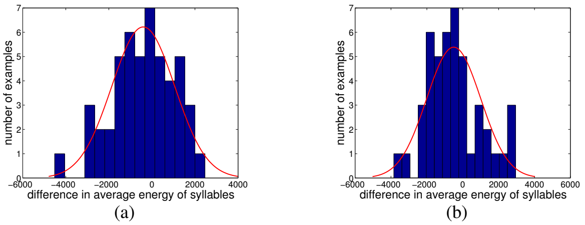
(a) (b)

Fig. 3: Distribution of duration difference of syllables at (a) 2^{nd} and 3^{rd} position (χ^2 = 11.2578, degrees of freedom = 12, confidence interval = 0.95) and (b) 13^{th} and 14^{th} position (χ^2 = 13.1642, degrees of freedom = 12, confidence interval = 0.95)



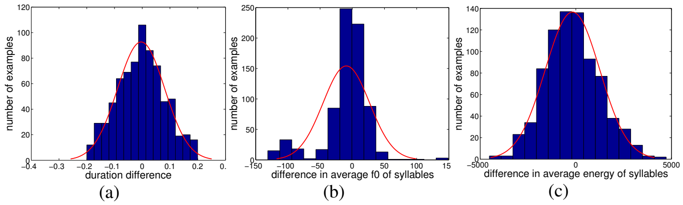
(a) (b)

Fig. 4: Distribution of average pitch difference of syllables at (a) 2^{nd} and 3^{rd} position (χ^2 = 14.3335, degrees of freedom = 8, confidence interval = 0.95) and (b) 9^{th} and 10^{th} position (χ^2 = 14.2072, degrees of freedom = 8, confidence interval = 0.95)



(a) (b)

Fig. 5: Distribution of average energy difference of syllables at (a) 2^{nd} and 3^{rd} position (χ^2 = 5.2933, degrees of freedom = 10, confidence interval = 0.95) and (b) 13^{th} and 14^{th} position (χ^2 = 10.6529, degrees of freedom = 12, confidence interval = 0.95)



(a) (b) (c)

Fig. 6: (a) Duration difference (χ^2 = 9.7929e+03, degrees of freedom = 14, confidence interval = 0.95), (b) Average pitch difference (χ^2 = 6.0652e+04, degrees of freedom = 17, confidence interval = 0.95), and (c) Average energy difference (χ^2 = 740.3880, degrees of freedom = 25, confidence interval = 0.95)

other positions in the sentence. The logical conclusion arrived at is that the location of syllables in the sentence is important. But the extent to which this is relevant in syllable timed languages is to be investigated. Therefore, separate experiments were conducted for the case where statistics are computed considering the location of syllables in an utterance and the case where statistics are computed irrespective of the location of the syllable in the utterance.

The plots of difference in duration, average f_0 and average energy are shown for a set of 10 word sentences in Hindi. Also, the duration difference between a pair of syllables were found irrespective of the location of the syllable pair in the sentence. Observations show that the distribution of differences in duration between syllable pairs follows a normal distribution. These studies were performed to determine whether the location of the syllable has an effect on the duration of the unit.

The plots for distribution of duration difference (Figure 3), average pitch difference (Figure 4) and average energy difference (Figure 5) are shown for syllables located at various positions in the sentence. It is seen that all the distributions have mode at zero and can be modelled by a Gaussian distribution. χ^2 goodness of fit tests were performed. Interestingly, without considering the location of syllables, the distributions obtained did not exhibit properties of a normal distribution as shown in Figure 6. As a result of this, the location of syllables in an utterance were preserved while performing the experiments.

An interesting observation made during the experiments was that *geminate context* of the syllables are important while selecting units [15]. Geminates refer to two identical consonants occurring without a vowel between them. Geminates are very common in Indian languages. Some examples are *amma* which means ‘mother’ in Tamil and *sattar* which means ‘seventy’ in Hindi. The word *sattar* is syllabified as *sat* and *tar*. Using a syllable *sat* from a geminate context to form the word *satrah* was found to be inappropriate. Syllables have therefore been selected according to geminate context. Context information about the location of a syllable in the word, i.e., whether the syllable is at the beginning, middle or end of the word, is also used. In this study, we have restricted the context information of syllables to geminates and location in a word. The syllables of a word were tagged according to their position in the word as *_beg*, *_mid* or *_end* and *_eg* (ends with geminate) or *_bg* (begins with geminate) and appropriate syllables were selected during synthesis. In order to improve the intelligibility of synthesised speech, inter-syllable pauses were inserted wherever required.

Figure 7 shows an example utterance synthesised using difference in average energy, average f_0 and duration. It can be observed from Figure 1 that the f_0 contour of the utterance synthesised using the traditional USS is very discontinuous and the utterance synthesised using the proposed approach has a more continuous f_0 contour, almost similar to that of natural utterance shown in Figure 2. Also, it can be observed that the energy contour of traditional USS appears to be very non-uniform while the energy contour of the utterance synthesised by the proposed approach is more uniform. Figure 8 shows variations in duration difference, energy difference and f_0 difference of units between a natural utterance, and an utterance synthesised using difference in f_0, energy and duration approach.

TABLE II: Results for DMOS and WER for the different methods used

	DD		**f_0D**		**ED**		**DD, ED**		**DD, f_0D**		**ED, f_0D**		**DD, ED, f_0D**		**USS**	
Language	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER	DMOS	WER
Hindi	3.13	3.03	3.63	5.67	3.24	8.40	3.18	15.95	3.63	17.46	3.05	15.06	3.29	11.94	3.59	7.02
Tamil	3.10	10.78	3.48	8.97	3.94	1.96	3.32	7.01	3.55	5.55	3.47	5.20	3.27	15.47	3.23	7.52

DD-duration difference, f_0D-average f_0 difference, ED-average energy difference, USS-Unit Selection Synthesis

VI. Evaluation

Tests of Degradation Mean Opinion Score (DMOS) and Word Error Rate (WER) [16] were conducted for each of the 7 methods mentioned in Section V. The scores for the tests conducted have been mentioned in Table II. The tests were conducted across 2 different languages namely Hindi and Tamil.

VII. Discussions

As seen in Figure 8 the variations in average f_0 difference, duration difference and average energy difference for a synthesised sentence is minimal as expected. However, in the case of the natural sentence, it can be seen that there is some variation between the features of adjacent units. Therefore, some amount of variation in these features can be considered as natural. This motivated a probability based approach rather than absolute distance measures.

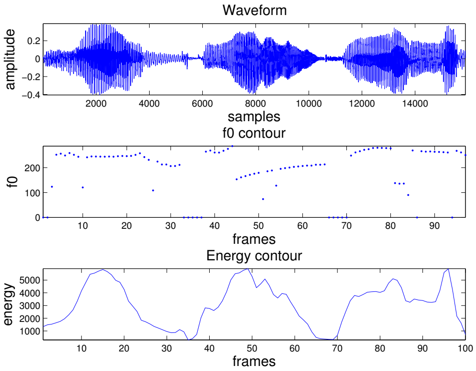


Fig. 7: An example of speech synthesised using the above mentioned approach for the text *bhaag bhii khel*

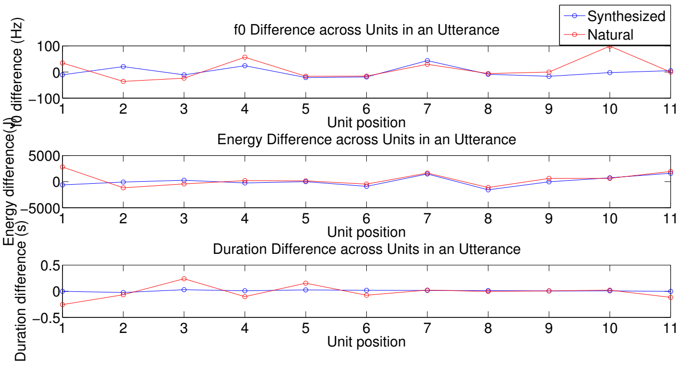


Fig. 8: Comparison of the variations in duration difference, energy difference and f_0 difference of adjacent units between a natural and a synthesised utterance using approach (7) mentioned in Section V

From the DMOS and WER scores in Table II, it is observed that all the approaches mentioned in Section V perform well with respect to naturalness and intelligibility of speech. It is observed that for Hindi, the approaches that use f_0 difference perform the best, while for Tamil, the approaches that use energy difference perform the best. In general, approaches that use f_0 difference perform better. This suggests that if f_0 difference and energy difference are given suitable weights while combining, the resulting system might perform better. Also, it can be observed that duration difference alone is not sufficient to improve the naturalness of synthesised speech while the intelligibility of synthesised speech improves significantly. Accurate segmentation of the speech database is another factor that influences the naturalness of the synthesised utterance. Errors in segmentation deter naturalness significantly.

VIII. Conclusion

From the experiments conducted above, we can conclude that reducing acoustic variation between adjacent units plays a major role in improving the naturalness and intelligibility of speech. Moreover, f_0, energy and duration play an equally important role. It is seen that reasonably natural sounding speech is obtained even without considering linguistic features.

Acknowledgment

The authors would like to thank DeitY for funding the project, *Development of Text to Speech Systems for Indian Languages* (11(7)2011-HCC(TDIL)). The authors would also like to thank Akshay Bhandari, Shreya Khare, Asha T, Abhijit Pradhan, Saad, Kasthuri S, Gandhi A, Musfir M and Anusha Prakash for their invaluable help in conducting the DMOS and WER tests.

References

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis.* Dordrecht: Kluwer, 1997.

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

[3] S. Kishore, R. Kumar, and R. Sangal, "A data driven synthesis approach for Indian languages using syllable as basic unit," in *Proceedings of International Conference on NLP (ICON)*, 2002, pp. 311–316.

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

[5] N. H. Samsudin, E. K. Tang, and K. Chuah, "Adjacency analysis for designing unit selection speech model on micro prosodic level," in *Proceedings of National Computer Sciences Postgraduate Colloquium (NaCSPC)*, 2005.

[6] V. R. Lakkavalli, P. Arulmozhi, and A. G. Ramakrishnan, "Continuity metric for unit selection based text-to-speech synthesis," in *Proceedings of International Conference on Signal Processing and Communications (SPCOM)*, 2010, pp. 1–5.

[7] 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 *Proceedings of 8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, 2013, pp. 209–214.

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

[9] P. V. K., T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communications*, vol. 42, pp. pp.429–446, 2004.

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

[11] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7869–7873.

[12] C. S. Seelamantula and T. V. Sreenivas, "Blocking artifacts in speech/audio: Dynamic auditory model-based characterization and optimal time-frequency smoothing." *Signal Processing*, vol. 89, no. 4, pp. 523–531, 2009.

[13] G. A. Miller and J. Licklider, "The intelligibility of interrupted speech," *The Journal of the Acoustical Society of America*, vol. 22, p. 167, 1950.

[14] Hiroya, "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing." in *The Production of Speech*. New York: Springer-Verlag, 1983, pp. 39–55.

[15] S. Kar, "Gemination in bangla: An optimality theoretic analysis," *Dhaka University Journal of Linguistics*, vol. 1, no. 2, pp. 87–114, 2008.

[16] R. B, S. L. Christina, G. A. Rachel, S. Solomi V, M. K. Nandwana, A. Prakash, A. S. S, R. Krishnan, S. K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *Proceedings of 8th ISCA Workshop on Speech Synthesis*, 2013, pp. 311–316.