A Hybrid Approach to Segmentation of Speech Using Group Delay Processing and HMM Based Embedded Reestimation

S Aswin Shanmugam, Hema Murthy

Department of Computer Science and Engineering
Indian Institute of Technology Madras

aswin@cse.iitm.ac.in, hema@cse.iitm.ac.in

Abstract

The most popular method for automatic segmentation is embedded reestimation of monophone HMMs after flat start initialization, followed by forced alignment. This method may not yield accurate boundaries. To address this issue, group delay based processing of short-time energy (STE) is performed on the speech signal to obtain syllable boundaries. The syllable boundaries are accurate, but there are a number of spurious insertions as the text transcription is not used during segmentation. The boundaries obtained using group delay segmentation in the vicinity of the HMM syllable boundaries are used as correct boundaries to reestimate the monophone HMM models, where the monophone HMMs are restricted to the syllable boundaries rather than the whole utterance. The reestimated boundaries are again compared with the group delay boundaries and corrected again. Essentially signal processing for detecting boundaries and statistical segmentation for acoustic modelling work in tandem to obtain accurate segmentation at both phoneme and syllable levels. Considering phones and syllables as basic units, HMM based speech synthesis systems (HTS) are built with the proposed segmentation method. Listening tests indicate that there is an improvement in the quality of synthesis.

Index Terms: HMM, flat start, embedded reestimation, forced alignment, group delay, STE, syllable, HTS

1. Introduction

Segmentation of the speech corpus at syllable/phoneme level is an important phase in many speech processing applications, including training of text-to-speech systems and automatic speech recognition systems. Especially, text-to-speech systems require very accurate and consistent segmentation [1, 2].

The commonly used approach for automatic segmentation is a three stage process : (1) flat start initialization of phoneme HMMs, (2) embedded training, (3) forced Viterbi alignment [3]. This approach requires word, syllable or phoneme transcriptions. These transcription are mapped into phoneme sequences using grapheme to phoneme rules or a pronunciation dictionary. A fundamental drawback of this approach is that boundaries are not represented by this model [4] as HMMs do not use proximity to boundary positions as a criterion for optimality during training [2]. Many methods have been proposed to model phoneme boundaries. For example, support vector machine (SVM) classifiers are used to locate boundaries in [5], a special one-state HMM is used for detecting phoneme boundaries in [4]. In [6], a multi layer perceptron is used to refine phone boundaries.

Phone transitions are not necessarily distinguishable owing to coarticulation in continuous speech. On the other hand, syllable boundaries are more or less distinct, owing to syllable being the fundamental unit of speech production and cognition [7, 8]. The acoustic energy between syllables is significantly lower [9] than at the middle of a syllable.

Although the role of syllables in segmentation cannot be contested [10], appropriate acoustic cues are required to detect syllable boundaries. Syllable boundaries are characterised by low energy[1]. So, short-time energy (STE) can be used as a cue to determine syllable boundaries but it cannot be applied directly owing to local fluctuations [11]. It is shown in [12] that STE function, when smoothed by performing group delay processing, can be used to detect syllable boundaries. As group delay segmentation is agnostic to the transcription, the boundaries can not be relied upon completely.

In earlier work [13], the authors used a semi-automatic labelling tool [14] based on group delay segmentation to obtain segmentation at the syllable level. Then, with the manually corrected syllable boundaries, they restricted embedded reestimation [15] of monophones to the syllable level and obtained accurate segmentation at the phone level. In this study, a novel algorithm is proposed where the purely signal processing based approach of detecting syllable boundaries using group delay processing and the machine learning approach of forced alignment are made to work in tandem to obtain accurate segmentation at both phone and syllable level automatically.

This paper is organized as follows. Section 2, describes in brief, the group delay based segmentation of syllables. Section 3 presents the proposed segmentation algorithm. In Section 4, the experiments and results are discussed. Section 5 concludes the paper.

2. Group delay based segmentation of speech

Group delay based processing of short-time energy (STE) can yield syllable boundaries [12, 16]. The algorithm for group delay segmentation [11] is given below :

- Compute STE function $E[m]$ for the given speech utterance.
- Compute $\frac{1}{E[m]^\gamma}$ where $\gamma = 0.01$[2]
- Construct the symmetric part of the sequence by lateral

[1] There are exceptions to this rule but this is not crucial for the current work as will be seen later

[2] A small value of γ is chosen to reduce the dynamic range of STE

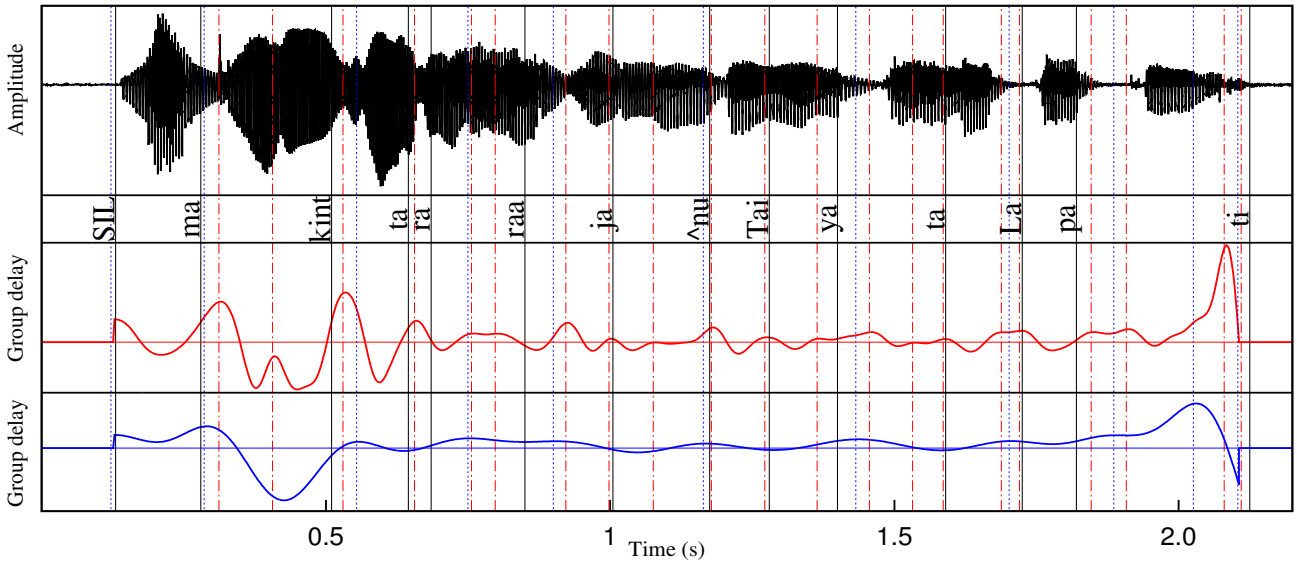

Figure 1: *Syllable boundaries given by HMM based segmentation (black solid lines) and GD based segmentation with WSF=10 (red dashed lines) & WSF=30 (blue dotted lines)*

inversion. The resulting sequence is positive and symmetric. So, it resembles the magnitude spectrum of an arbitrary real signal. Let's call this sequence $E[K]$.

- The IDFT of $E[K]$ is computed. The resultant signal is the root cepsrum [17] and the causal portion of the same is a minimum phase signal [18]. Let's call this causal sequence $e_c[n]$.
- A single sided Hanning window is applied on $e_c[n]$ and its minimum-phase group delay function $E_{gd}[K]$ is computed. The size of the window applied is

$$N_c = \frac{Length\ of\ STE}{Window\ scale\ factor\ (WSF)} \qquad (1)$$

The resolution of the group delay function is controlled by WSF.

- Peaks in the minimum phase group delay function ($E_{gd}[K]$) are detected and they are regions of low energy.

Extensive experimentation with group delay segmentation shows that except for syllables starting or ending with a fricative, nasal; or starting with a semi-vowel, affricate, the boundaries are very accurate [11].

Figure 1 shows the syllable level segmentation of the phrase "மகிந்தராஜனுடைய தளபதி" (makintararaajaˆnuTaiya taLapati[3]). The black solid lines in Figure 1 show the syllable boundaries given by flat start initialized embedded training of monophone HMMs followed by forced Viterbi alignment. The top panel shows the waveform and the other two panels show two different group delay functions. The difference between the two functions is the change in value of the empirical parameter, WSF. The parameter can be chosen such that it always introduces insertions. The function in the middle pane is obtained with WSF as "10" and the dashed red lines are the corresponding boundaries. The other function is obtained with WSF as "30" and the dotted blue lines are the corresponding boundaries. Although with WSF as "10", the number of spurious boundaries is large (resolution is high), nevertheless the correct boundaries are not misplaced. Such a group delay function is first obtained for every utterance.

[3]Tamil text written in ITRANS [19]

3. Hybrid method

Some segments given by HMM based segmentation suffer from gross durational errors, where the duration of one segment is significantly larger than the other. Hidden semi-Markov models (HSMM) [20] have explicit state duration probability distributions. All states of monophone HSMMs were initialized with global mean and global variance, followed by embedded reestimation. When alignment is performed using HSMMs, it was observed that although durational errors reduced significantly, the boundaries obtained by forced alignment using HSMMs were worse than those obtained with HMMs. So, with the segmentation given by HSMMs, HMM models are built and alignment is performed. This corrects durational errors in most of the places and the location of boundaries are also not very inaccurate.

If the syllable does not end with a nasal, fricative or if it's not followed by a nasal, fricative, affricate or a semi-vowel; the boundary of that syllable is moved to the nearby region of low energy given by the group delay function with high resolution explained in Section 2.

During embedded training [15], transcription is used to construct a composite HMM for each utterance by concatenating phone HMMs. Embedded Baum-Welch re-estimation is performed and all monophone HMM models are updated simultaneously.

As segmentation at the syllable level is available, waveforms are spliced at the syllable level and embedded training is performed on these syllables. Now, the composite HMMs obtained will be syllable HMMs and not utterance HMMs. Embedded Baum-Welch re-estimation is restricted to the syllable boundaries and monophone models are built [13]. These models are more robust as reestimation is performed on shorter segments of speech. Using the new monophone models, forced

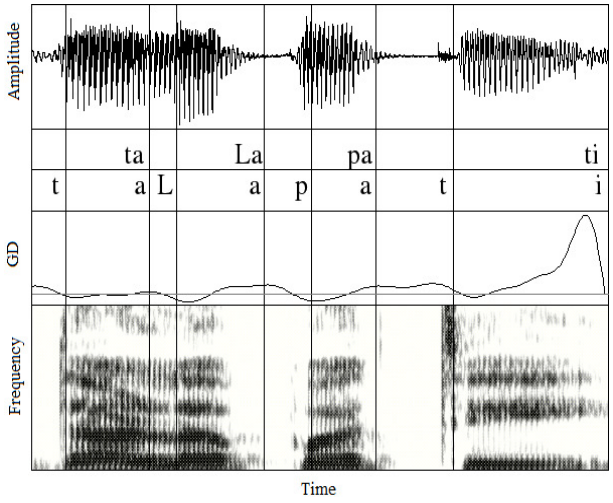

Figure 2: *Phone level segmentation after alignment within syllables*

alignment is performed on the entire utterance. The boundaries obtained using forced alignment are again compared with that of the group delay boundaries and corrected as before. This gives the final syllable level segmentation.

To obtain segmentation at the phone level, embedded Baum-Welch re-estimation is performed at the syllable level again to refine the monophone models. With these models phone level alignment is performed on the syllable waveforms [13]. By combining the phone level alignment of the syllables constituting an utterance, phone segmentation corresponding to the entire utterance is obtained as shown in Figure 2. In Figure 2, the top panel shows the waveform and the bottom panel shows the spectrogram. The panel above the spectrogram shows the group delay (GD) function. Syllable level alignment obtained using the proposed method is given below the waveform and then, phone level segmentation obtained after forced alignment within the syllables is shown. Figure 3 shows the sequence of steps involved in the proposed method.

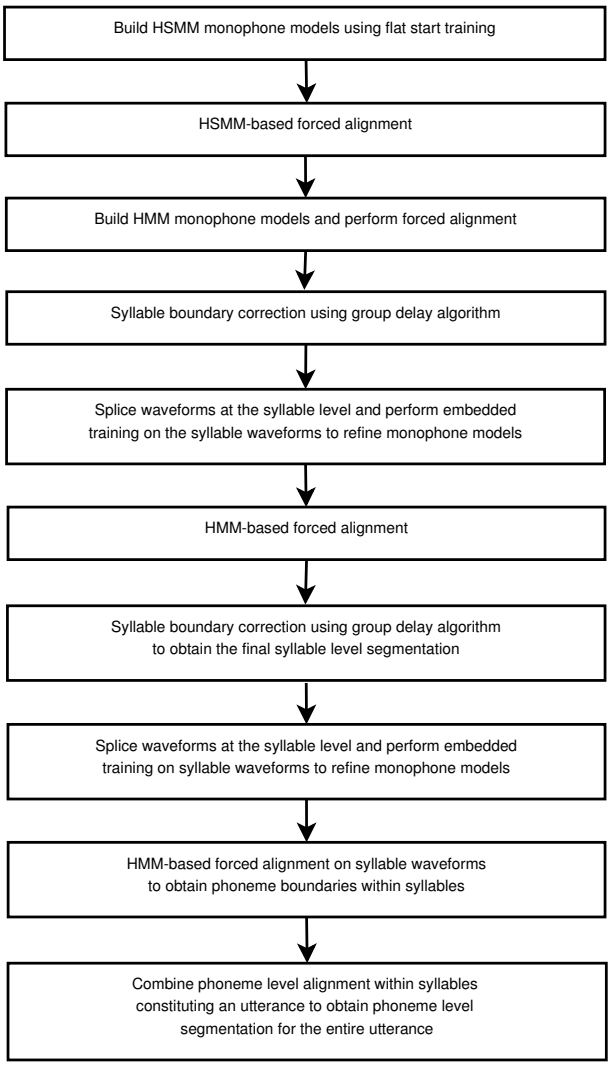

Figure 3: *Steps involved in the proposed hybrid method*

4. Experiments and results

4.1. Experimental setup

730 utterances of Tamil data spoken by a single native female speaker is used. 35 phonemes are used for Tamil. The syllabification rules for Tamil are explained in [21]. The phoneset[4] used is given in [22]. For training HSMMs, mel cepstral coefficients and log of the fundamental frequency are used as features. All monophones are modelled as 5-state, 1-mixture HSMMs. For training HMMs, MFCC's are used and all monophones are modelled as 3-state, 2-mixture models. HTS-2.2 [23] and HTK-3.4.1 [15] are used. WSF is set as "10" in group delay segmentation algorithm.

4.2. Segmentation accuracy

After obtaining phone level segmentation as explained in Section 3, monophone HMM models are built using it. Using these models forced alignment is performed on all 730 utterances in the database. The total acoustic likelihood during forced Viterbi alignment increased for all these utterances when the hybrid method is used compared to flat start embedded training of HMMs. The average log probability per frame increased from "-74.16" to "-73.17".

Table 1 shows the average log probability per frame for different types of phonemes. Although, boundary corrections were not performed for semi-vowels, fricatives and nasals, even their log probability has increased significantly.

Method	Average log probability per frame					
	Nasals	Fricatives	Semi-Vowels	Vowels	Stop Consonants	Overall
Hybrid	-72.22	-79.23	-73.36	-70.77	-82.15	-73.17
HMM	-73.74	-81.72	-75.46	-70.99	-85.04	-74.16

Table 1: Average log probability per frame

Figure 4 shows the waveform in the top panel, followed by syllable aligned transcription given by the proposed method, HMMs, HSMMs and flat start training of HSMMs followed by HMMs. The group delay function is shown below the transcrip-

[4]The phone set used can be downloaded from the link : "http://www.iitm.ac.in/donlab/ilsl12.pdf"

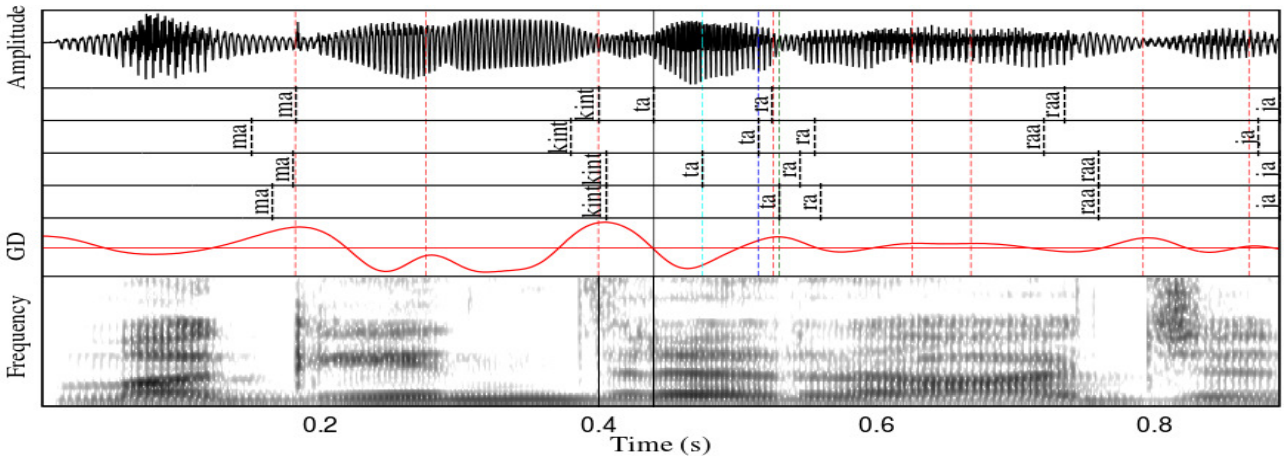

Figure 4: *Syllable level segmentation given by the proposed hybrid method compared to HMM,HSMM and HSMM followed by HMM*

tion and the spectrogram is shown in the bottom panel. The red dashed lines are boundaries given by group delay segmentation. The boundary of syllable "kint" is slightly misplaced when flat start training of HSMMs followed by HMMs are used. When boundary correction is performed in the proposed method, group delay peak in the proximity is taken and the location of the boundary is rectified. The syllable "ta" marked in the spectrogram is identified correctly only by the proposed method (black solid line). Both, the statistical approach and the group delay approach failed to find the boundary correctly. Group delay based approach didn't detect the boundary as the succeeding syllable "ra" starts with a semi-vowel. But in the proposed method, when the two approaches work in tandem, phone models are built in a robust manner, which lead to the correct alignment of the syllable.

4.3. Text-to-speech systems

With the segmentation given by the hybrid method and the conventional HMM based method, HMM based speech synthesis systems (HTS) [23] are built. Both, Phone based HTS and syllable based HTS [21] are built for Tamil. In syllable based HTS, monosyllables, with only begin, middle and end context with respect to the word are used. Subjective evaluations were performed on these TTS systems. First, semantically unpredictable sentences (SUS) were synthesized and participants of the test were made to transcribe them. Then, word error rate (WER) was calculated. The hybrid approach results in a lower WER as indicated in Table 2

System	WER
HTS - Syllable (HMM Segmentation)	11.11%
HTS - Syllable (Hybrid Segmentation)	7.07%
HTS - Phone (HMM Segmentation)	4.04%
HTS - Phone (Hybrid Segmentation)	1.01%

Table 2: Word error rate

Pair comparison (PC) test [24] was also performed. For this comparison, the HTS system built with the hybrid segmentation method is referred to as "A" and the HTS system built with HMM based segmentation is referred to as "B". During PC test, the order in which the sentences are played creates a bias. So, both "A-B" test and "B-A" test were performed with different sets of sentences. In "A-B" test, synthesized sentences of system "A" were played first and during "B-A" test, synthesized sentences of system "B" were played first. "A-B+B-A" in Table 3 gives the order independent preference in percentage of system "A". It is calculated as

$$\text{"A-B+B-A"} = \frac{(\text{"A-B"} + (100 - \text{"B-A"}))}{2} \quad (2)$$

HTS - Syllable			HTS - Phone		
A-B	B-A	A-B+B-A	A-B	B-A	A-B+B-A
75	20	77.5	70	15	77.5

Table 3: Pair comparison tests

A live synthesizer is available at "http://www.iitm.ac.in/donlab/segmentation/hts.php".

5. Conclusions

Speech recognition and synthesis require accurate segmentation of data so that robust models can be built for the fundamental units. Although machine learning has been very successful for segmentation, there is a requirement of huge amount of data. Even then, the segmentation obtained is not accurate. Machine learning techniques are robust on the average, while signal processing techniques are accurate on the particular. In this paper, an attempt has been made to synergize the benefits of knowledge-based domain specific signal processing and machine learning to obtain accurate segmentation.

6. Acknowledgements

The authors would like to thank the Department of Information Technology, Ministry of Communication and Technology, Government of India, for funding the project, "Development of Text-to-Speech synthesis for Indian Languages Phase II", Ref. no. 11(7)/2011HCC(TDIL). The authors would also like to thank Dr. V. Ramasubramanian, PESIT for his valuable suggestions.

7. References

[1] A. W. Black and J. Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *ICASSP*, 2009, pp. 3785 – 3788.

[2] A. Sethy and S. S. Narayanan, "Refined speech segmentation for concatenative speech synthesis." in *ICSLP*. ISCA, 2002, pp. 149–152.

[3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4)," *Cambridge University Engineering Department*, 2002.

[4] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models." in *INTERSPEECH*. ISCA, 2013, pp. 2306–2310.

[5] H.-Y. Lo and H. min Wang, "Phonetic boundary refinement using support vector machine," in *ICASSP*, vol. 4. IEEE, April 2007, pp. 933–936.

[6] K.-S. Lee, "Mlp-based phone boundary refining for a tts database," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 981–989, May 2006.

[7] O. Fujimura, "Syllable as a unit of speech recognition," in *ICASSP*, vol. 23. IEEE, February, 1975, pp. 82–87.

[8] S. Greenberg, "Speaking in shorthand- a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[9] R. W. M. Ng and K. Hirose, "Syllable: A self-contained unit to model pronunciation variation," in *ICASSP*. IEEE, March 2012, pp. 4457–4460.

[10] J. Mehler, J. Y. Dommergues, U. Frauenfelder, and J. Segui, "The syllable's role in speech segmentation," *Journal of Verbal Learning and Verbal Behavior*, vol. 20, pp. 298–305, 1981.

[11] 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.

[12] V. Kamakshi Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3, pp. 429–446, 2004.

[13] 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.

[14] 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.

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

[16] 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.

[17] J. Lim, "Spectral root homomorphic deconvolution system," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 3, pp. 223–233, Jun 1979.

[18] T. Nagarajan, V. Kamakshi Prasad, and H. Murthy, "Minimum phase signal derived from root cepstrum," *Electronics Letters*, vol. 39, no. 12, pp. 941–942, Jun 2003.

[19] A. Chopde, "ITRANS," http://www.aczoom.com/itrans/.

[20] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis," in *Proc. of ICSLP*, 2004.

[21] A. Pradhan, S. Aswin Shanmugam, A. Prakash, V. Kamakoti, and H. A. Murthy, "A syllable based statistical text to speech system," in *EUSIPCO*, 2013.

[22] 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.

[23] "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.

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