Group Delay Based Phone Segmentation for HTS

S. Aswin Shanmugam, Hema A. Murthy
Department of Computer Science and Engineering
IIT Madras
Email: {aswin, hema}@cse.iitm.ac.in

***Abstract*—HMM based speech synthesis (HTS) is a state-of-the art approach to text-to-speech synthesis. Segmentation of the training data is essential for building any text-to-speech system. Most conventional text-to-speech systems use phones as the basic unit of synthesis and use a speech recogniser to automatically segment the data at the phone level. As Indian languages are low resource languages, accurate transcriptions are difficult to obtain owing to paucity of data. Manual labeling at the phone level is not only laborious but also inaccurate. HMM based flat start segmentation doesn't work well at the sentence level. In this paper we propose an event driven approach to obtain better phone boundaries. Syllable-like events are detected in the speech signal and matched with syllabified transcription of the text. The syllables are converted to phoneme sequences and Baum-Welch embedded re-estimation is restricted to the syllable-level. Subjective evaluations indicate that the proposed system has a lower word error rate compared to that of a conventional system that uses flat start for obtaining phone boundaries.**

I. Introduction

HMM based speech synthesis systems (HTS) [1] correspond to the state-of-art today. They are attractive owing to the fact that they are not only small footprint but also provide much better intelligibility compared to the unit selection based systems. The footprint of HTS systems are small because it doesn't store any pre-recorded speech waveforms directly. HTS extracts speech parameters from the waveforms and builds statistical models and uses them for synthesis. For building a good quality HTS system for any language, a well labeled database is required [2].

For high resource languages like English, HMM based phone segmentation is performed using high quality speech recognizers like [3] and [4]. There are no high quality speech recognizers available for Indian Languages. HMM based flat start segmentation [5] is not accurate, even when context dependent HMM's are used. In flat start segmentation method, all models are initialised such that state means and variances are equal to the global mean and variance. Then, embedded training [3] is performed to build models. Using these models, forced Viterbi alignment is performed to obtain segmentation at phone level. Another conventional method for segmentation is bootstrapping, as performed in [6]. Bootstrap labeling is performed in four steps: (1) about five minutes of data is labeled manually at the phone level, (2) phone models are built using the data that is manually labeled, (3) forced Viterbi alignment is performed on the rest of the data using the models built and (4) iteratively models are built and then forced Viterbi alignment is performed until segmentation is satisfactory. The problem with this approach is labeling the bootstrap data manually at phone level. Manual labeling at the phone level cannot be performed accurately because not all phones can be perceived properly in isolation, as only syllables are the basic units of sound production and cognition [7]. The co-articulation between phones within a syllable is very high and hence marking the boundaries for such phones is very difficult.

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

Although the importance of syllables in segmentation of speech cannot be questioned [8] , appropriate acoustic cues are required to detect syllable boundaries. It is shown in [9] that short-term energy (STE) function, when smoothed by performing group delay processing, can be used to detect syllable boundaries. Group delay based segmentation of syllables [9] has been successfully used for building syllable based text-to-speech systems in [10] and [11]. A semi-automatic labeling tool [12] was used to obtain very accurate segmentation at the syllable level. The main motivation of this paper is to use this existing information of syllable boundaries which are accurate, to obtain the phone boundaries.

This paper is organized as follows. Section II, describes in brief, the HMM based speech synthesis system (HTS). In Section III, the group delay based segmentation of syllables is described. In Section IV, the proposed method for obtaining segmentation at phone level for the speech data, that is already segmented at the syllable level, is presented, highlighting the differences between the proposed segmentation method and flat start HMM based segmentation. In Section V, the experiments and results are discussed. In Section VI, the results of subjective evaluation tests are presented. Section VII concludes the paper.

II. HMM BASED SPEECH SYNTHESIS SYSTEM

In the training phase of an HMM based speech synthesis system (HTS) [1], from the speech data, spectral parameters (Mel generalized cepstral coefficients) and it's dynamic features, excitation parameters (log of the fundamental frequency) and it's dynamic features are extracted. Initially, context independent (monophone) HMM models are built. HTS uses three types of context: prosodic, linguistic and phonetic [13]. The phonetic context used in HTS by default is a pentaphone context; for every phone two preceding and two succeeding phones are used. With these contexts, context dependent HMM models are built after being initialized with corresponding context independent HMM models. Tree based context clustering [14] based on the question set is performed to tie states. Tree based context clustering is performed to address two problems associated with the modeling and use of context dependent phones: (1) insufficient data to build all context dependent HMMs separately and (2) unseen models.

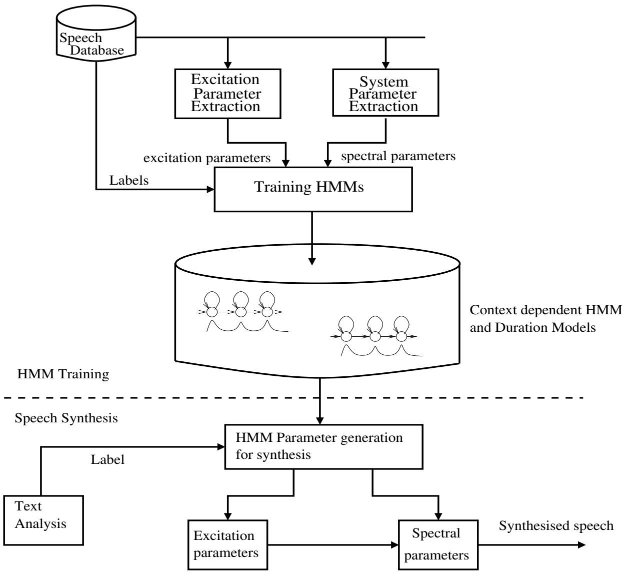

Fig. 1. Overview of HMM-Based Speech Synthesis

During synthesis, the required context-dependent HMMs are concatenated to obtain the sentence HMM. Appropriate models are chosen by traversing the decision tree [15] built during tree based context clustering. The parameter generation algorithm explained in [16] is applied to get the speech observation vector that maximizes the output probability. The speech waveform is synthesized from the generated spectral and excitation parameters using a source-filter model. An overview of the HTS system is illustrated in Figure 1[1].

III. GROUP DELAY BASED SEGMENTATION OF SYLLABLES

Syllable boundaries can be determined by performing group delay based processing of the short-term energy (STE) function [17]. The group delay segmentation algorithm [18] for obtaining syllable boundaries is shown in Figure 7 (last page). In the context of speech synthesis, boundaries must be consistent and accurate. To enable this a semi-automatic tool named ”DONLabel” was developed [12]. The tool requires a set of parameters to be tuned based on the characteristics of the speaker, to obtain accurate syllable boundaries.

DONLabel shows the boundaries given by group delay segmentation and allows the user to insert, modify or move an existing boundary. When the parameters are optimised such that more boundaries are obtained than what is suggested by the transcription, labeling corrections primarily reduce to deletion of boundaries. Movement of boundaries is required occasionally for fricatives and semivowels.

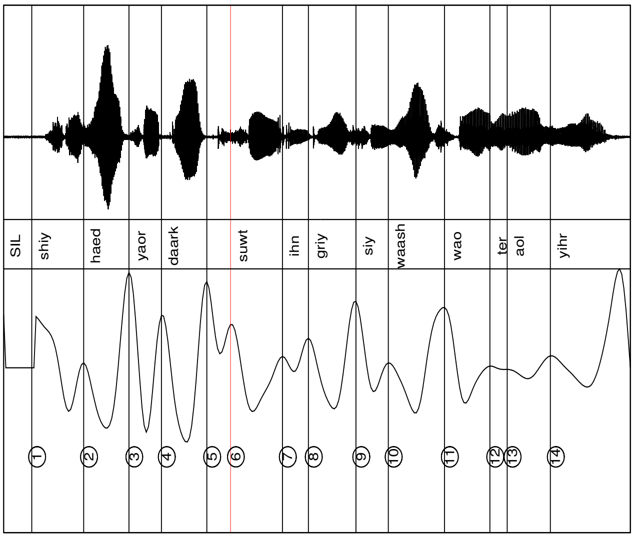

Fig. 2. Group Delay based Labeling

Figure 2 [10] illustrates the process of segmenting a waveform at the syllable level using an example TIMIT sentence (sa1): “She had your dark suit in greasy wash water all year”. The top panel shows the speech waveform, the middle panel shows the syllable transcription, and the bottom panel shows the group delay function [9]. The peaks in the group delay function correspond to syllable boundaries. Most of the syllable boundaries are detected by the group delay function, except for the extra boundary within the syllable /suwt/, which is marked as number 6 in Figure 2. The extra boundary is manually deleted and the label file is saved.

[1] This figure has been redrawn from [13]

IV. Enforcing syllable boundaries during Embedded Re-estimation

The process of obtaining phone level segmentation from syllable label files is explained here. As the syllable boundaries have been obtained already, each utterance in the speech data is split at the syllable level and stored separately. For example, if the utterance “text_001” has “n” segments (syllables), the waveform is divided into “n” files and stored separately as “text_001-1” to “text_001-n”. Spectral parameters (Mel frequency cepstral coefficients) are extracted from these syllable waveforms. Baum-Welch embedded re-estimation [5] is performed on each of these syllables iteratively to build phone HMM models. Using these models, phone level alignment is performed on the syllable waveforms. By combining the phone level alignment of the syllables constituting an utterance, phone segmentation corresponding to that utterance is obtained.

Figure 4 summarises the above procedure with an example. When phone alignment is performed within syllable waveforms, the timestamps of the boundaries obtained are relative to the end time of the previous syllable in the utterance. To obtain the phone label file for the entire utterance, the end time of the previous syllable is added to all phone boundaries in the current syllable. For example, when phone alignment is performed on the syllable /baar/ (बार्) 0.728 second, i.e. the end time of the previous syllable in the utterance /ii/ (ई), is added to the boundaries of the phones /b/, /aa/, and /r/. After the phone alignment is performed within the syllable /baar/ (बार्), the end time of /b/ is 0.079 second. When the phone label file is formed, the end time of /ii/ (ई) is added and the end time of /b/ (ब्) becomes 0.807 second.

Syllable Label			Phone Label		
Beg	**End**	**Unit**	**Beg**	**End**	**Unit**
0.000	**0.234**	SIL	**0.000**	**0.234**	SIL
0.234	**0.363**	इ /i/	**0.234**	**0.363**	i
0.363	**0.516**	से /see/	**0.363** 0.398	0.398 **0.516**	s ee
0.516	**0.647**	क /ka/	**0.516** 0.576	0.576 **0.647**	k a
0.647	**0.728**	ई /ii/	**0.647**	**0.728**	ii
0.728	**1.113**	बार् /baar/	**0.728** 0.807 0.982	0.807 0.982 **1.113**	b aa r
1.113	**1.228**	SIL	**1.113**	**1.228**	SIL

TABLE I. Phone labels from syllable labels

Table 1 shows a sequence of known syllable-boundaries for the phrase “इसे कई बार” in an utterance and the boundaries of the phones that make up the syllable, which are obtained using the procedure explained in this section. The start time and the end time of a syllable (given in bold) shown in Table 1, remain unaltered in the phone label, as Baum-Welch embedded re-estimation is restricted to the syllable-level.

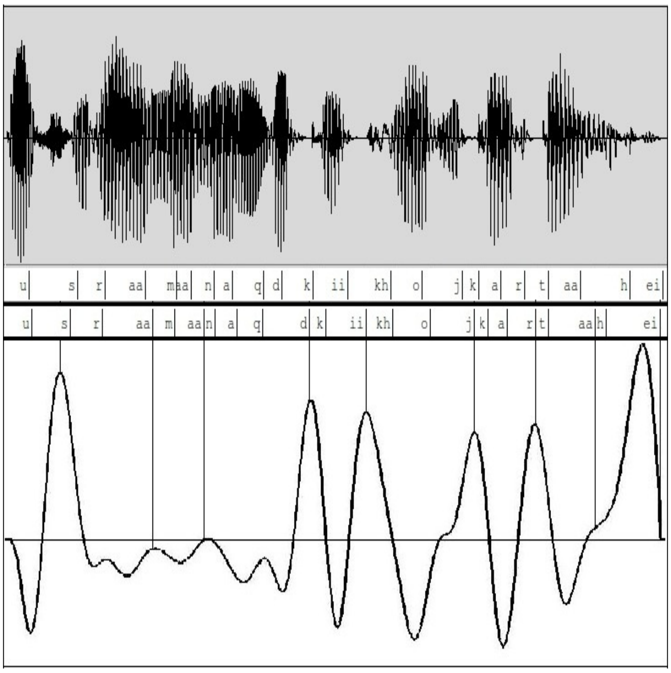

Fig. 3. Flat Start Vs Syllable Enforced Phone Boundaries

The segmentation of an example phrase “उस रामानंद की खोज करता है” is shown in Figure 3. The top panel shows the waveform, panel 2 from top shows the segmentation at phone level given by flat start segmentation, panel 3 is the segmentation given by the proposed method and the bottom panel shows the group delay function. The syllable boundaries are highlighted by the vertical lines. Except for the unvoiced fricative /s/ in the boundary of the syllable /us/, no other syllable boundary was moved during manual correction of labels using DONLabel. Three boundaries (peaks in the group delay function without vertical lines over it) were given within a syllable, that was also deleted during manual correction.

The phone boundaries obtained using flat start segmentation in Figure 3, were observed to be inaccurate. For example, the syllable /kii/ is segmented only as /ii/ by flat start segmentation. On observing the waveform, it can be seen that the segment marked as /ii/ by flatstart segmentation, also has a stop consonant in it. The boundary of the phone /j/ in syllable /khoj/ being incorrect, is also clearly visible in the waveform.

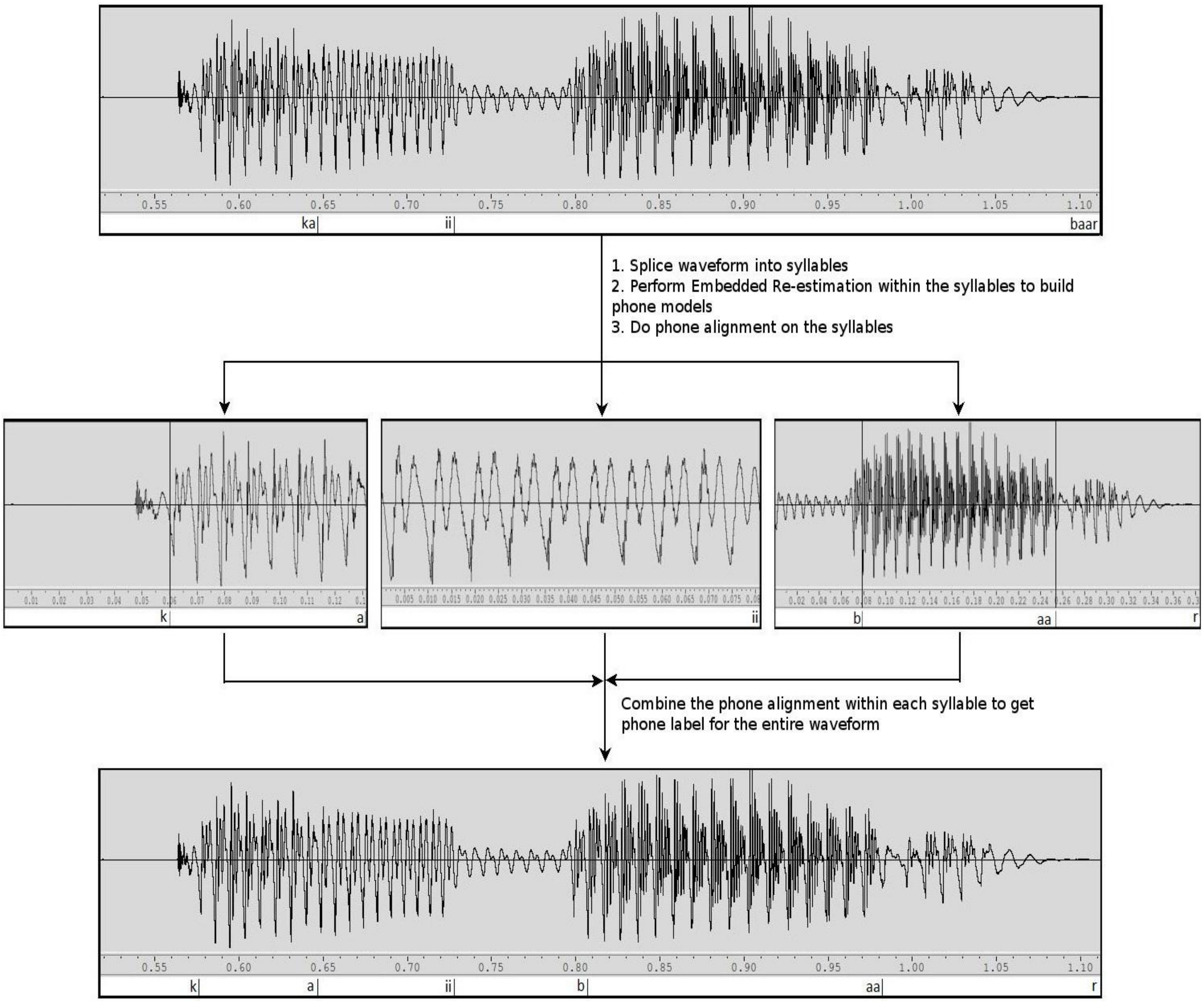
0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05 1.10
ka
ii
baar
1. Splice waveform into syllables
2. Perform Embedded Re-estimation within the syllables to build phone models
3. Do phone alignment on the syllables
k
a
ii
b
aa
r
Combine the phone alignment within each syllable to get phone label for the entire waveform
0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05 1.10
k
a
ii
b
aa
r

Fig. 4. Syllable Enforced Embedded Re-estimation

V. Experiments and Results

1.35 hours of Hindi data (595 utterances) has been used for training both systems. All utterances have been recorded in a studio environment at 48000Hz, 16bits per sample, spoken by a single native Hindi male speaker. For the experiments in this paper HTK-3.4.1 and HTS-2.2 were used. All 595 utterances were labeled at the syllable level using DONLabel [12]. HMM models with five states and two mixture components per state were used to model each phone for the proposed group delay based phone segmentation approach. The phone set[2] and question set described in [6] have been used. The utterance structures obtained using Festival [19] were used during the training phase. A 108 dimensional feature vector consisting of Mel-generalised cepstral (mgc) coefficients (35), their delta (35), and acceleration coefficients (35), 3 dimensional excitation features, that is, log F0 and its dynamic features were extracted from the speech files. Four-stream HMM models (composite models which models both spectral and excitation parameters) with five states and a single mixture component per state were used to model each phone. Five-stream duration models, with a single state and a single mixture component per state, were also generated for each phone.

Figure 5 shows the spectrogram corresponding to a part of a sentence synthesized using Flat Start (FS) segmentation based HTS voice. The spectrogram corresponds to the words “कर होती हैं”. The word “हैं” (highlighted in the figure) is not synthesized properly. The formant structure in Figure 5 is not clear towards the end and it is noisy at the beginning.

[2] The phone set used can be downloaded from the link : http://lantana.tenet.res.in/ilsl12.pdf

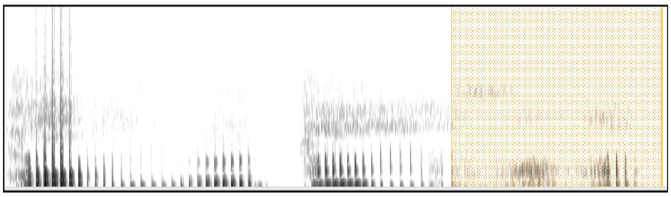

Fig. 5. Wideband Spectrogram of the portion of a sentence synthesized using HTS built with HMM based flat start segmentation

Figure 6 shows the spectrogram corresponding to the same portion of the sentence synthesized using the HTS voice built with our proposed Group Delay (GD) based segmentation. All the words are synthesized clearly. The formant structure is also better.

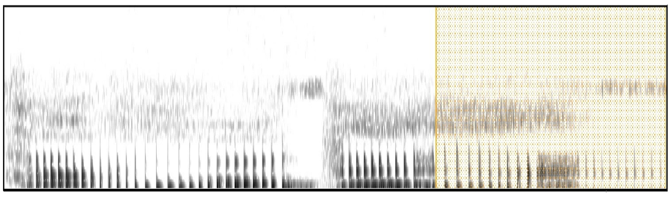

Fig. 6. Wideband Spectrogram of the portion of a sentence synthesized using HTS built with GD based segmentation

VI. Performance Evaluation

Degradation MOS (Naturalness) [6] was calculated and for intelligibility, semantically unpredictable sentences[3] were synthesized. Participants at the test were asked to listen to each sentence only once and transcribe the sentence. From the transcribed sentences Word Error Rate (WER) was calculated.

System	DMOS (Naturalness)	WER
Hindi (GD Segmentation)	2.98	3.19%
Hindi (FS Segmentation)	2.89	5.04%

TABLE II. DMOS and Word Error Rate

Pair comparison (PC) test [20] was also performed. Same sentences were synthesized with both the systems and the participants were asked to choose which system was better for each sentence during PC test. We denote the HTS system built with Group Delay (GD) based segmentation as "A" and the HTS system built with Flat Start (FS) HMM based segmentation as "B". During PC test, the order in which the sentences were 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. Scores in column 1 and 2 of Table 3 is the percentage with which the first term of the pair is preferred. "A-B+B-A" in Table 3 denotes the order independent preference in percentage of system "A". It is calculated as

$$\text{``A-B+B-A''} = \frac{(\text{``A-B''} + (100 - \text{``B-A''}))}{2} \tag{1}$$

A-B	B-A	A-B+B-A
66.67	13.33	76.67

TABLE III. Pair Comparison Tests, where A is the HTS system built with GD based segmentation and B is the HTS system built with Flat Start HMM based segmentation

All the subjective evaluation tests were performed in a quiet environment with good quality headphones. A live synthesizer is available at http://lantana.tenet.res.in/hts/.

VII. Conclusion

In this paper, a novel approach for initialising the monophone HMMs in HTS based systems is proposed. A event-driven approach is used to direct the segmentation process. Syllable-like events are detected in the speech signal and matched with the transcription. The syllables are transcribed into phones and embedded re-estimation is performed at the syllable level. Although DMOS (naturalness) does not improve significantly, the WER has improved significantly. This suggests that use of signal processing cues to improve phone boundaries can aid in better speech synthesis.

Acknowledgment

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 like to thank Akshay Bhandari, Raghava Krishnan, Shrey Dutta, Shreya Khare and Anusha Prakash for helping with MOS tests. The authors would also like to thank all the participants of the MOS tests.

[3] e.g. "आकाश मे सफ़ेद हाथी नाचता गाता जा रहा है"

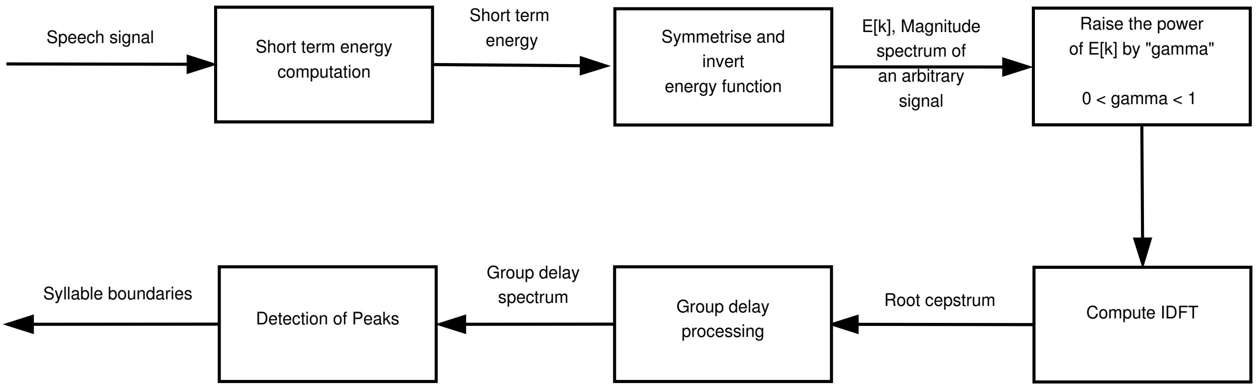
Speech signal
Short term energy computation
Short term energy
Symmetrise and invert energy function
E[k], Magnitude spectrum of an arbitrary signal
Raise the power of E[k] by "gamma"
0 < gamma < 1
Compute IDFT
Root cepstrum
Group delay processing
Group delay spectrum
Detection of Peaks
Syllable boundaries

Fig. 7. Steps involved in group delay segmentation algorithm for obtaining syllable boundaries

References

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

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

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

[4] X. Huang, F. Alleva, H.W. Hon, K.F. Hwang, M.Y. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7(2), pp. 137–148, 1992.

[5] 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.

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

[7] Steven Greenberg, "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*. IEEE, 1996, pp. 1–8.

[8] Jacques Mehler, Jean Yves Dommergues, Uli Frauenfelder, and Juan Segui, "The syllable's role in speech segmentation," *Journal of Verbal Learning and Verbal Behavior*, vol. 20, pp. 298–305, 1981.

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

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

[11] Vinodh M.V., Ashwin Bellur, Badri Narayan K., Deepali Thakare M., Anila Susan, Suthakar N.M., and Hema A. Murthy, "Using polysyllabic units for text to speech synthesis in Indian languages," in *National Conference on Communication(NCC)*, 2010, pp. 1–5.

[12] P.G. Deivapalan, Mukund Jha, Rakesh Guttikonda, and Hema A. Murthy, "DONLabel: An automatic labeling tool for indian languages," in *National Conference on Communication (NCC)*, February 2008, pp. 263–266.

[13] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, pp. 1039–1064, 2009.

[14] J. J. Odell, P. C. Woodland, and S. J. Young, "Tree-based state clustering for large vocabulary speech recognition," in *International Conference on Speech, Image Processing and Neural Networks*, 1994.

[15] J. J. Odell, "The use of context in large vocabulary speech recognition," PhD dissertation, Cambridge University, 1996.

[16] K. Tokuda, T. Yoshimura, T. Masuko, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP*. IEEE, 2000, pp. 1315–1318.

[17] T. Nagarajan and Hema 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.

[18] 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.

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

[20] P.L. 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.