A HYBRID APPROACH TO SEGMENTATION OF SPEECH USING GROUP DELAY PROCESSING AND HMM BASED EMBEDDED REESTIMATION

S. Aswin Shanmugam, Hema A. Murthy

Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

Introduction

- Text-to-speech (TTS) systems require accurate and consistent segmentation.
- Group delay segmentation [1] for detecting boundaries and HMMs for acoustic modelling, used in tandem to obtain accurate segmentation.
- Quality of TTS improves with the proposed segmentation algorithm.

Automatic Segmentation using HMM

- Three stage process : (1) flat start initialization of mono-phones [2], (2) embedded training, (3) forced alignment.
- Drawback - boundaries are not represented by the model.

Importance of Syllables in Segmentation

- Phoneme – basic unit in written form, syllable – basic speech production unit.
- Phone transitions not necessarily distinguishable.
- Inter-syllable co-articulation comparatively less.
- Syllable boundary characterized by low energy.

Group Delay Based Segmentation of Syllables [1]

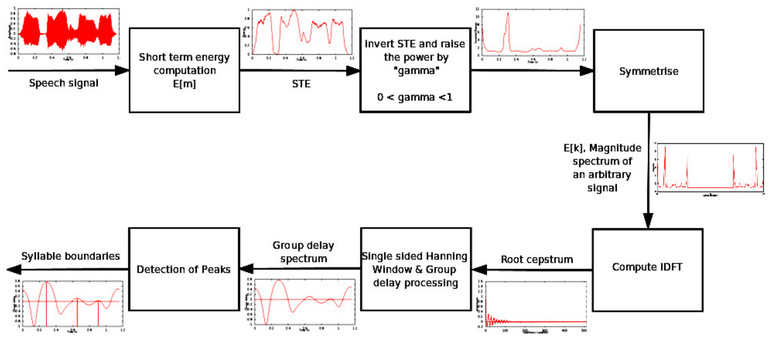

Figure 1: *Group delay segmentation algorithm*

- Short-term energy (STE) used as acoustic cue.
- Causal portion of root cepstrum is minimum-phase [3].
- Size of the window applied is

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

- WSF determines resolution of the group delay function.

Importance of Phase

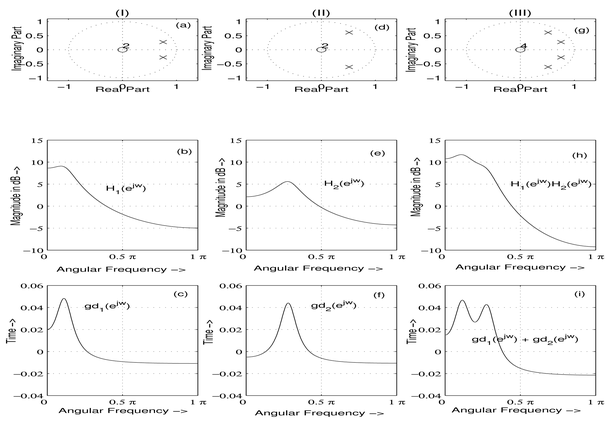

Figure 2: *High resolution property of minimum-phase group delay [1]*

Hybrid Approach

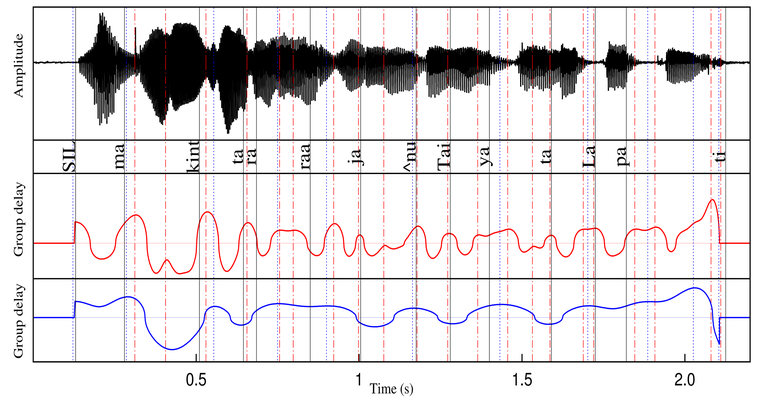

Figure 3: *Syllable boundaries given by HMM based segmentation and GD based segmentation with WSF=10 & WSF=30*

- With WSF "10", the number of spurious boundaries is large (resolution is high) but correct boundaries are not misplaced
- 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, boundary corrections are performed

Hybrid Segmentation Algorithm

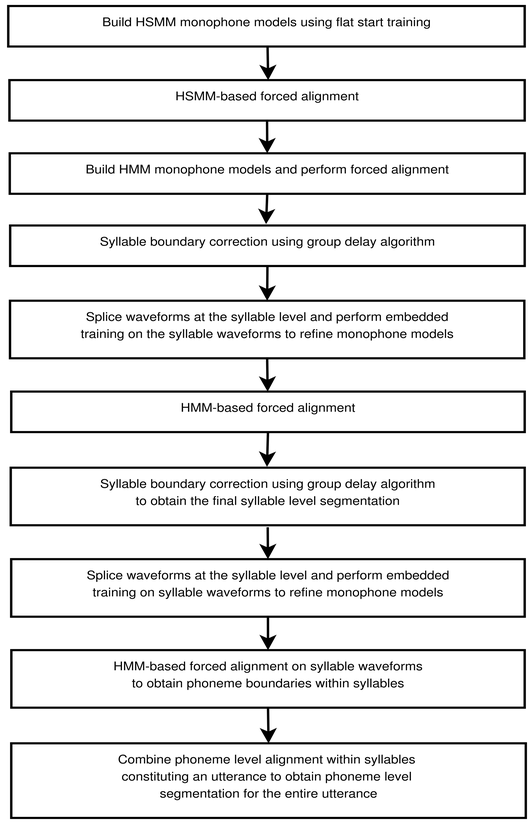

Figure 4: *Steps involved in the proposed hybrid method*

Segmentation Accuracy

- Increase in acoustic likelihood during forced alignment.

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*

- The syllable "ta" marked in the spectrogram is identified correctly only by the proposed method.

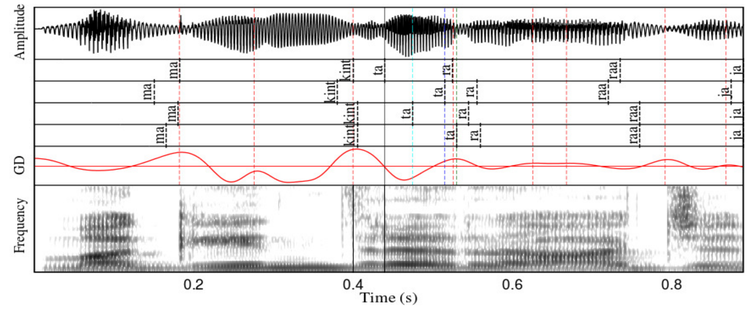

Figure 5: *Syllable level segmentation given by the proposed hybrid method compared to HMM,HSMM and HSMM followed by HMM*

Quality of TTS

- HMM based speech synthesis systems (HTS) [4] built for Tamil.
- Both Phone HTS and Syllable HTS [5] show improved WER

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 (WER)*

- Order independent pair comparison tests also show improvement.

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*

Acknowledgement

The authors would like to thank the Department of Electronics & Information Technology (DeitY), Ministry of Communications & IT, Government of India, for funding this research under the project "CSE/1112/129/DITX/HEMA".

Conclusion

- Machine learning techniques are robust on the average, while signal processing techniques are accurate on the particular.
- An attempt has been made to synergize the benefits of knowledge-based domain specific signal processing and machine learning to obtain accurate segmentation.

References

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

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

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

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

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

The authors would like to thank the Department of Electronics & Information Technology (DeitY), Ministry of Communications & IT, Government of India, for funding this research under the project "CSE/1112/129/DITX/HEMA"