SIGNIFICANCE OF PSEUDO-SYLLABLES IN BUILDING BETTER ACOUSTIC MODELS FOR INDIAN ENGLISH TTS

Rupak Vignesh S, S. Aswin Shanmugam,and Hema A. Murthy

Indian Institute of Technology Madras

ABSTRACT

Signal processing based landmark detection is precise compared to HMM based alignment, primarily because the location of the landmark is not factored in the estimation of parameters. Acoustic cues for syllable boundaries are usually obtained by exploiting the inherent sonority characteristics of a syllable. As syllabification of the text is based on generalized rules or lexicon definitions, there is a mismatch between the acoustical and the lexical segments for non-native syllabification. In this paper, an attempt is made to modify the syllabification rules for Indian English using acoustic cues obtained from syllable boundaries. The modified syllabifier is used to syllabify the text. Embedded re-estimation is performed using forced alignment at the modified syllable level to obtain refined phoneme boundaries. Indian English Text-to-Speech (TTS) systems are built using labels obtained after (i) embedded re-estimation at the sentence level and (ii) the aforementioned procedure. Reduction in the word error rates for both native Aryan and Dravidian speakers (relatively by 54.1% and 52.4% respectively), suggests that there is a significant synthesis quality improvement in the proposed system.

Index Terms— Indian English, Pseudo syllables, TTS.

1. INTRODUCTION

Building an Indian English TTS system is a challenging task as the pronunciation, syllabification and stress assignment rules differ from that of native English. The lack of standard pronunciation dictionary and letter-to-sound (LTS) rules result in poor acoustic models and hence forced alignment leads to inaccurate boundaries.

Our motivation for this work is based on cross-linguistic observations. Indian languages being phonetic, predominantly have straightforward syllabification rules. On the one hand, the group delay based syllable segmentation [1] has no knowledge of the underlying transcription. It smooths the short-term energy (STE) function by making use of the additive property of Fourier transform phase and deconvolution property of the cepstrum, thereby deriving syllable boundaries from the smoothed STE function. On the other hand, the HMM based segmentation assumes the sequence of underlying phonemes and yields imprecise boundaries when forced alignment is done at the sentence level [2]. The consistent coordination between the lexical and acoustic syllable segments in Indian languages, makes it possible to combine these two different approaches under a common framework by restricting the re-estimation and alignment procedures within syllable boundaries and thus refining the phoneme boundaries [3].

In an earlier attempt [4] at building better Indian English TTS, syllabification and stress assignment rules were modified in standard American English lexicon [5] to suit Indian English. The effect of this modification was found to be insignificant. When words are uttered in isolation, the syllabification corresponds to that present in lexicon definitions. However, in continuous speech, significant coarticulation is present between words that leads to insertion of syllables at word junctures. In Syllable based generalizations in English phonology [6], the author agrees to this notion and formulates a special syllabification rule (Rule V) for inter-word syllable linkages. There were attempts to model word junctures [7, 8] and correct transcriptions based on phonological rules, which resulted in better acoustic models.

The method of constructing psuedo-syllables from temporal envelope of the speech signal was introduced in [9]. The hill-shaped profile on the temporal envelope specifies the syllable onset and offset which is used to cluster phonemes. It has been reported that consonant clusters at syllable boundaries become ambisyllabic. For this work in particular, syllabification must be unambiguous, as wrong syllabification leads to incorrect re-estimation of the monophone models within the syllables.

2. ACOUSTIC CUES FOR SYLLABLE SEGMENTATION

To obtain the initial syllable boundaries, group delay processing of short-term energy (STE) function is used. Root cepstral smoothing [10] is performed on the symmetrized STE function to smooth local energy fluctuations. The minimum phase group delay of the liftered root cepstrum is computed instead of the magnitude spectrum owing to its better resolving power

This work is funded by the Department of Information Technology, Ministry of Communication and Technology, Government of India

[1]. Since this resolution is affected by the size of the lifter on the root cepstrum, the resultant function suffers from insertion and deletion errors. Previously, a semi automatic labelling tool [11] was used to manually correct syllable boundaries. Later, the process was automated by choosing the lifter size such that it always allowed insertions and the syllable boundaries given by HMMs were approximated to only boundaries of high confidence [3].

2.1. Additional cues using spectral flux

STE boundaries are not very accurate whenever syllables begin or end with fricative or affricate consonants. In a few cases, sibilants are shown as separate syllables. These boundaries are corrected based on the change in the spectral content of the conjectural phones. A modified version of the spectral flux is used in order to correct only boundaries having abrupt spectral changes as in the case of fricative and affricate consonants. [12] details the exact specifications of the spectral flux function used for this purpose.

2.2. Combining evidences

While group delay cues accurately determine the boundaries of unvoiced stop consonants, spectral flux cues give precise boundaries if one of the conjectural consonants is a fricative or an affricate. The forced aligned boundaries of these consonants are alone moved to the appropriate cues in their vicinity. After this boundary correction, embedded model re-estimation is done on syllable splices to build better models. These improved models are used during forced Viterbi alignment at the sentence level.

Boundary correction is repeated in the second pass, to further refine syllable segmentation. After embedded model re-estimation and forced alignment on syllable splices in the second pass, these splices are concatenated to get the final phone-level alignment. As shown in Figure 1, spectral flux corrects the boundary of /s ah/ and STE corrects the boundaries of /p ao/ and /t ay m/. The problem arises in syllabifying the text such that it corresponds to these cues. It may be possible that the consonant is present in one syllable, but its acoustic realization is in the preceding or subsequent syllable. Such errors lead to incorrect re-estimation of the models.

3. DERIVING SYLLABLES CORRESPONDING TO THE ACOUSTIC LANDMARKS

The conventional method of pronunciation lexicon based syllabification does not allow word juncture coarticulations. To obtain preliminary syllabification, we used the NIST syllabifier [13], written based on the rules given by [6]. The algorithm implementation allows only permissible consonant clusters to form a syllable. Additionally, only [C-V] type word junctures are permitted (Rule V [6]). This preliminary

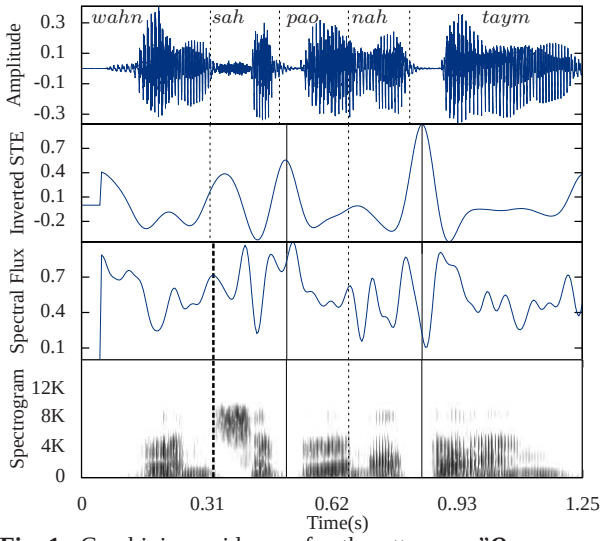

Fig. 1. Combining evidences for the utterance "***Once upon a time***". HMM boundaries (dotted lines) are approximated to the nearest STE (solid lines) and spectral flux (dashed line) peaks according to correction rules.

syllabification is modified based on inferences from acoustic cues that dictate syllable boundaries. The rules for forming these pseudo-syllables are discussed in section 3.2.

3.1. Finding inter-word syllable linkages

Choosing segments for syllabification is critical as word junctures determine syllabification even in the case of monosyllabic words (as shown in Fig.1). A 3-state silence is said to be inadequate to model pauses between words [14]. To find the actual pauses, a special single-state short pause model (sp) is inserted after every word during embedded training. Since forced alignment exploits the knowledge of the phoneme sequence, the 'sp' model is forced to occupy the least possible duration as self transition becomes less likely whenever coarticulation between words occur. These units are deleted from the transcription and connected word segments between the remaining short pauses are chosen for syllabification.

3.2. Inferences from STE and spectral flux functions

The following rules are framed for conjectural consonants based on their energy and spectral properties. Let the syllable-initial consonant be represented as begin phone B_p and syllable-final consonant as end phone E_p. When E_p is an unvoiced stop consonant which is followed by a semivowel or a whisper B_p, there is a significant dip in energy between E_P and its previous phone. So we associate the stop consonant with the semivowel or whisper, making the stop as B_p (Figure 2a: E_p /k/ is associated with the semivowel /l/).

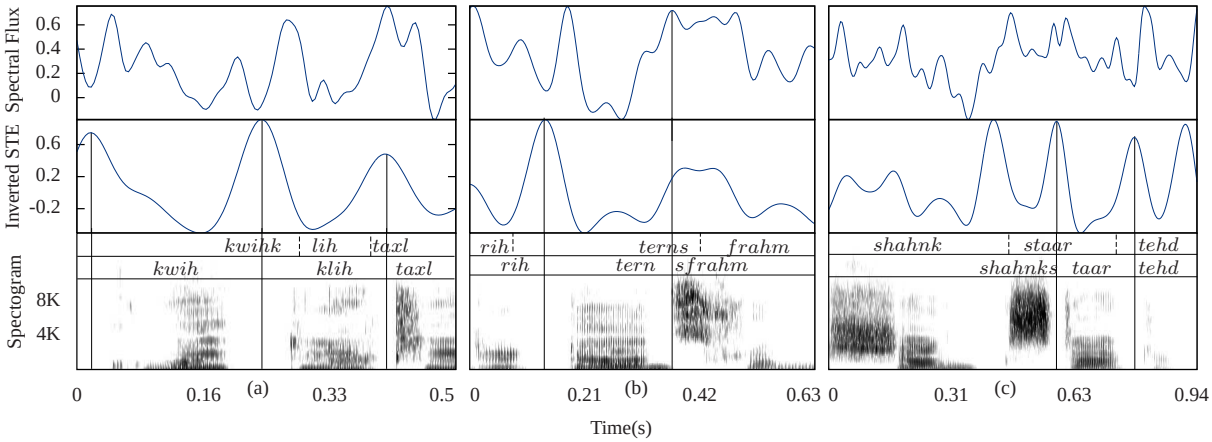

Fig. 2. Forming pseudo-syllables with the help of acoustic cues for (a) **Quick little**, (b) **Returns from** and (c) **Shunk started**. The top transcription panel shows the nist syllabification along with forced aligned boundaries (dashed lines). The bottom transcription panel shows the modified syllabification.

Fricatives and affricates are characterized as having high energy in the high frequency regions of the spectrum and hence contribute to spurious peaks in the smoothed STE. When both B_p and E_p belong to these categories, the spectral change between the two phones is not prominent. Since both the spectral flux and STE boundaries between syllables B_p and E_p are ambiguous, these consonants are merged into the same syllable (Figure 2b: E_p /s/ is associated with /f/).

Consonant associations at word junctures ([C-C]), are not permitted by the syllabifier (Rule V [6]). These are observed when a fricative B_p occurs between two unvoiced stop consonants. Due to the coarticulation between B_p and the stop consonant E_p, the boundaries given by both spectral flux and STE functions are of low confidence. For this reason, the fricative B_p is associated with the word-final stop consonant E_p (Figure 2c: B_p /s/ is associated with word-final /k/). This rule applies to intra-word syllables as well (Explain: /ehks-pleyn/, Extent: /ehks-tehnt/).

3.3. Syllabification accuracy

Indian languages are broadly classified into Aryan and Dravidian languages. The syllabification rules were applied to native Hindi (Aryan) and Tamil (Dravidian) speakers' English. 200 randomly chosen utterances from CMU Arctic database [15] were taken for each speaker. As the syllables do not correspond to the actual syllables for English, we refer to these as pseudo-syllables. Insertions and deletions of phones can still occur in these pseudo syllables. If the actual B_p of the succeeding syllable is present in the current syllable as E_p, it corresponds to an insertion. Deletion occurs when the actual E_p of current syllable is present in the succeeding syllable as B_p. The respective insertion and deletion rates were found to be 0.47% and 1.97% for the Hindi speaker and 0.28% and 1.8% for the Tamil speaker.

4. EVALUATION

4.1. Indian English database

The Indian English database consists of 5 hours of 4186 utterances each for Hindi male and Tamil female speakers. The utterances are taken from the CMU Arctic database [15] and Chandamama stories [16], recorded in an acoustic treated studio setup. The Indian English pronunciation lexicon (IE dict) consists of 6413 words, phonetically transcribed by only two people to maintain maximum consistency. Letter-to-sound rules are built from this pronunciation lexicon using CART [17] for predicting the transcription for out of vocabulary words. A total of 38 phones from the CMU phoneset are present.

4.2. Segmentation accuracy

Monophone models built using hybrid segmented labels are compared with flatstart initialized monophone HMMs in terms of the average log probability per frame. Vowels are modelled as 5-state 2-mixture models, consonants as 3-state 2-mixture models and short pause as 1-state 2 mixture model. MFCC features are used for training HMMs. The *WindowScaleFactor* (WSF)[1] for smoothing STE and spectral flux functions are set as 6 and 2 respectively. An overall increase in the average log probability per frame from -68.53 to -67.89 was observed for the Hindi speaker and -73.55 to -73.03 for the Tamil speaker.

Observe from Table.1 that even with flatstart HMMs, the average likelihood increases significantly (from -77.65 to -

[1] *WindowScalefactor* is defined as

$$\text{WSF} = \frac{\text{Size of the STE/SF function}}{\text{Size of the lifter}}$$

Table 1. Normalized likelihood scores for different category of phones using the following methods: **I. Flatstart HMMs (CMU dict), II. Flatstart HMMs (IE dict)** and **III. Hybrid segmentation (IE dict)**

Speaker	Method	Average log probability per frame						
		Semi vowels	Stop Consonants	Fricatives	Affricates	Vowels	Nasals	Total
Native Tamil female	I	-79.97	-83.87	-80.03	-81.57	-72.42	-74.72	-77.65
	II	-79.38	-83.83	-80.40	-81.14	-64.94	-74.30	-73.55
	III	-78.86	-83.17	-80.04	-80.73	-65.18	-72.89	-73.03
Native Hindi male	I	-76.44	-82.23	-77.32	-79.67	-64.24	-70.76	-71.11
	II	-75.54	-82.63	-78.64	-79.37	-61.25	-69.80	-68.53
	III	-75.07	-81.59	-77.45	-78.84	-60.70	-69.50	-67.89

73.55 for the Hindi Speaker and -71.11 to -68.53 for the Tamil Speaker) when IE dict is used as the pronunciation lexicon.

4.3. Comparison of TTS systems

Hidden Markov model based speech synthesis systems (HTS) [18] are built[2] using labels obtained after (i) sentence level embedded re-estimation and (ii) hybrid segmentation. In the first set of listening tests, word error rate is calculated for semantically unpredictable sentences (SUS) synthesized using these systems. In the second set of listening tests, pair comparison test is performed on synthesized sentences with both "A-B" and "B-A" pairs, mainly to avoid cognitive bias. Participants are asked to rate each system in the pair. These scores are normalized using mean opinion scores for natural sentences. The results are listed below. An equal preference of 17.26% and 18.28% for both systems are observed in the case of native Tamil and Hindi speakers respectively.

Table 2. Results of the listening tests

Speaker	Method of Initialization	WER (%)	DMOS	Preference (%)
Native Tamil female	Flatstart	12.4	3.19	23.21
	Hybrid Segmentation	5.9	3.59	59.52
Native Hindi male	Flatstart	17.6	2.93	8.08
	Hybrid Segmentation	7.9	3.84	72.73

4.4. Influence of L1 rhythm

Performing group delay processing of the STE function to capture syllable boundaries, requires lesser degree of vowel compaction in unstressed syllables. This is because unstressed syllables might be considered as local energy fluctuations and get liftered during cepstral smoothing. As a quantification of speech rhythm, interval measures (VarcoV, %V) [19] and rate normalized pairwise variability index for vowels (nPVI-v) [20] were computed for 4 speakers. 100 utterances from CMU Arctic database were taken for the following speakers ($Nativity_{Gender}$): $American_M$, $American_F$, $Hindi_M$ and $Tamil_F$. The phone-level alignments for American speakers were given by CMU Sphinx [21] and the alignments for Indian speakers were given by the hybrid segmentation algorithm. The reduction in both nPVI-v and VarcoV [22], for both native Hindi and Tamil speakers, suggests that there is an influence of syllable-timed rhythm in Indian English.

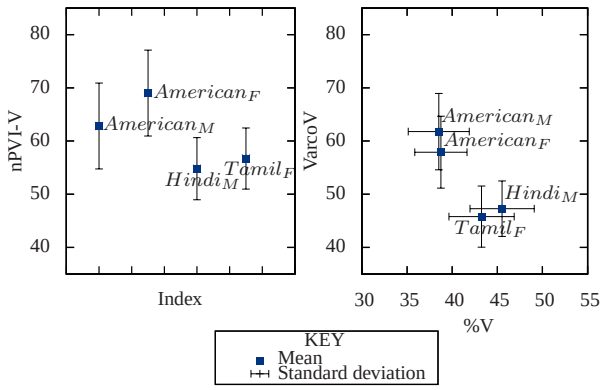

Fig. 3. Plots showing rhythm metric scores for the 4 speakers.

5. CONCLUSION

Although English in India was acquired from the British, it is influenced by syllable-timed rhythm and phonological relationships ingrained in the native language of the speaker. In this work, an attempt has been made to exploit this syllable-timedness with the help of an acoustically driven pronunciation correction at the syllable level. The results presented in this paper do indicate that the native tongue's characteristics must be factored into the pronunciation modelling of the same speaker speaking a different language. Such cross-lingual studies will aid in building not only better TTS systems but also ASR systems, where pronunciations are corrected based on the accent.

[2]samples available online at http://www.iitm.ac.in/donlab/hts/ie.php

6. REFERENCES

[1] V. Kamakshi Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," in *Speech Communication, vol. 42, no. 3*, 2004, p. 429446.

[2] S Aswin Shanmugam, Hema Murthy, et al., "Group delay based phone segmentation for hts," in *Twentieth National Conference on Communications (NCC)*. IEEE, 2014, pp. 1–6.

[3] S Aswin Shanmugam and Hema Murthy, "A hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation," in *INTERSPEECH*. ISCA, 2014.

[4] Rohit Kumar, Rashmi Gangadharaiah, Sharath Rao, Kishore Prahallad, Carolyn Penstein Rosé, and Alan W Black, "Building a better Indian English voice using "more data".," in *SSW*, 2007, pp. 90–94.

[5] Carnegie Mellon University, "The CMU pronunciation dictionary," 2000, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[6] Daniel Kahn, *Syllable-based generalizations in English phonology.*, Ph.D. thesis, Massachusetts Institute of Technology, 1976.

[7] Egidio Giachin, Aaron Rosenberg, and Chin-Hui Lee, "Word juncture modeling using phonological rules for hmm-based continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90.*,. IEEE, 1990, pp. 737–740.

[8] C-H Lee, E Giachin, LR Rabiner, R Pieraccini, and AE Rosenberg, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech & Language*, vol. 6, no. 2, pp. 103–127, 1992.

[9] Raymond WM Ng and Keikichi Hirose, "Syllable: A self-contained unit to model pronunciation variation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4457–4460.

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

[11] 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)*, IIT Bombay, India, February 2008, pp. 263–266.

[12] S Aswin Shanmugam, "A hybrid approach to segmentation of speech using signal processing cues and Hidden Markov Models," `http://lantana.tenet.res.in/thesis.php`, MS Thesis, Department of Computer Science and Engineering, IIT Madras, India, July 2015.

[13] W. Fisher, "The tsylb2 program: Algorithm description. nist. part of the tsylb2-1.1 package," 1996.

[14] Scott S Chen, Ellen M Eide, Mark J.F. Gales, Ramesh A Gopinath, Dimitri Kanevsky, and Peder A Olsen, "Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News," in *ICAŞSP*, 1999, vol. 1, pp. 37–40.

[15] John Kominek and Alan W Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[16] Wikipedia, "Chandamama monthly magazine," `https://en.wikipedia.org/wiki/Chandamama`.

[17] Alan W Black, Kevin Lenzo, and Vincent Pagel, "Issues in building general letter to sound rules," 1998.

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

[19] Emmanuel Ferragne and François Pellegrino, "A comparative account of the suprasegmental and rhythmic features of British English dialects," in *Modelisations pour l'identification des Languages*, 2004.

[20] Esther Grabe and Ee Ling Low, "Durational variability in speech and the rhythm class hypothesis," in *Papers in laboratory phonology*, 2002.

[21] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," 2004.

[22] Laurence White and Sven L Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.