Exploiting phone-class specific landmarks for refinement of segment boundaries in TTS databases

Vijayaditya Peddinti, Kishore Prahallad

International Institute of Information Technology, Hyderabad

vijayaditya.p@research.iiit.ac.in, kishore@iiit.ac.in

Abstract

High accuracy speech segmentation methods invariably depend on manually labelled data. However under-resourced languages do not have annotated speech corpora required for training these segmentors. In this paper we propose a boundary refinement technique which uses knowledge of phone-class specific sub-band energy events, in place of manual labels, to guide the refinement process. The use of this knowledge enables proper placement of boundaries in regions with multiple spectral discontinuities in close proximity. It also helps in the correction of large alignment errors. The proposed refinement technique provides boundaries with an accuracy of 82% within 20ms of actual boundary. Combining the proposed technique with iterative isolated HMM training technique boosts the accuracy to 89%, without the use of any manually labelled data.

Index Terms: speech segmentation, under-resourced language, landmarks

1. Introduction

The use of accurately annotated speech data enables development of high quality data driven synthesis techniques [1]. Speech segmentation is a major task in the annotation process. The use of manually segmented data for these accurate annotations has a prohibitive cost. Automatic segmentors are used to tackle this problem to some extent. In TTS databases, transcripts are available during the segmentation procedure. Hence explicit segmentation (text-dependent) procedures, which are better performing than the implicit segmentation (text-independent) procedures are used for TTS segmentation [2].

Hidden-Markov Model (HMM) based segmentation using force alignment is prominently used for explicit-segmentation. The flat start initialization [3] accuracy of HMM based segmentation is comparatively low. Hence manually labelled data or speaker-independent (SI) acoustic models are used as bootstrap for training the acoustic models used for segmentation [4]. In addition to this, methods such as [5], [6], [7], [8] have been proposed to further refine the boundaries predicted by the HMM segmentor. These refinement techniques can be broadly classified into two types:

1) *error correcting methods*, which reduce the systematic bias in the HMM boundaries

2) *alternate boundary detection methods*, which move the HMM boundary to alternate boundary candidates in it's neighbourhood

These refinement methods depend on manually labelled data. It is difficult to obtain manually labelled data, more so in the case of under-resourced languages. Hence alternate techniques are required to guide the label movement during the refinement process. This guidance is mostly obtained from spectral discontinuities in the locality of the HMM boundary. However several spectral discontinuities are observed in close proximity in several cases. Identification of the spectral discontinuity corresponding to the current HMM boundary is a non-trivial task, especially in cases where the initial HMM boundaries are highly erroneous. In this paper a set of boundary type specific cues and cost functions are proposed to identify the spectral discontinuity pertaining to the current HMM boundary. This method exploits the knowledge of phone-class specific sub-band energy events to identify the relevant spectral discontinuity. Kim *et al.* [4] also suggested the use of a phone-class specific spectral discontinuity measure for detection of boundary candidates and phone-class specific time windows to identify the spectral discontinuity corresponding to the current HMM boundary. However these time windows are empirically derived, which requires manual intervention. The proposed method on the other hand eliminates the necessity for manual intervention by using large time windows, to tackle even severe HMM segmentation errors. The process of selecting the relevant boundary is instead guided by using phone-class specific spectral cues.

The paper is organized as follows : Section 2 details the existing techniques for refinement. Section 3.2 describes the landmark based boundary refinement algorithm. Section 3.3 describes the proposed spectral cue based boundary selection procedure. Section 4 analyses the results and discusses improvements due to the proposed method.

2. Existing techniques

To tackle the lack of hand-labelled data in under-resourced languages, a variety of techniques have been suggested. Niekerk *et al.* [1] proposed the use of broad phonetic class label data from other language speech corpora for bootstrapping of HMM models. Kominek *et al.* [9] proposed the use of mel-cepstral distortion (MCD) measure, rather than manual labels for guiding the label movement. Hoffman *et al.* [10] proposed a technique which uses frames in the middle of the automatically labelled segments to guide the label refinement. However none of these methods use the knowledge of boundary specific sub-band events, which are already well researched in the literature ([11],[12]), to bring the performance of the fully automatic methods closer to the supervised techniques. The use of this knowledge base helps tackle errors, which are undetectable using conventional spectral discontinuity based techniques. It helps distinguish between closely occurring spectral discontinuities and helps tackle large misalignment errors by the HMM segmentor.

3. Proposed method

The proposed method uses boundary specific sub-band event information for both detection and selection of the alternate boundary candidates. The boundary candidates are detected using spectral discontinuities in sub-bands specific to the current class of boundary. From these detected boundaries the ideal candidate is selected using cost functions whose elements are spectral cues extracted from the neighbourhood of the boundary and from the center of phones on either side of the boundary. The proposed method distinguishes closely occurring discontinuities which pertain to different HMM boundaries by using different cues for each class of HMM boundary. Further care is taken to accurately place boundaries around burst segments which are of really short duration, as even minor alignment errors could lead to a mislabelling of the burst segment.

All the refinement techniques described in Section 2 are combined with the isolated HMM training procedure (described in [6]) to further increase the segmentation accuracy. The refined labels are used for isolated training, where each HMM model is initialized and iteratively re-trained, exclusively using the segments of the corresponding phone. The boundaries obtained using this combined training procedure are found to be superior [4]. Hence the proposed boundary refinement technique was also combined with isolated training.

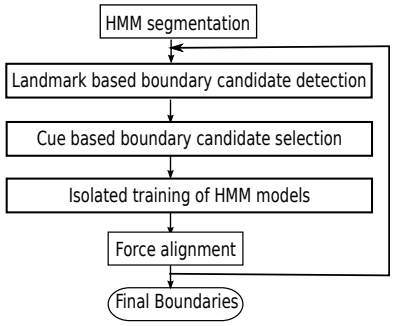

Figure 1: *Flow chart of proposed segmentation process*

3.1. HMM segmentation

Acoustic models for HMM segmentation were trained using embedded training technique with flat start initialization as the availability of manual labels or speaker-independent (SI) acoustic models is not assumed for bootstrap. Based on the analysis of segmentation performance with various HMM configurations described in [6], the HMM configuration was selected as three state left to right context-independent (CI) models, without a skip state (except for pauses), with one gaussian per state. The feature set used is 12 mel-frequency cepstral coefficients (mfcc) and normalised energy calculated with a frame length of 20ms and hop size of 5ms, along with delta and acceleration coefficients, resulting in a vector of length 39. Sub-phoneme labels were used for stop closures and bursts, as they have different acoustic properties. The separation of stops into closures and bursts must be accompanied a mechanism for detection of stop stop interactions (SSI), which lead to incomplete stops (missing bursts or closures). Hence intra and inter word stop-stop interactions and other incomplete stop possibilities were considered as pronunciation variations during the alignment. Geminates, which are contiguous occurrences of the same consonant, were mapped to corresponding single consonant [6]. Forced alignment was done using these acoustic models to derive the initial segment boundaries. These boundaries are used as an input to the refinement procedure.

3.2. Landmark based boundary candidate detection

Landmarks are time points of lexically significant acoustic events. Consonantal landmarks represent instances of sudden signal change eg., consonant release or consonant closure. Consonantal landmarks correspond to segment boundaries, hence the accurate identification of these landmarks implies increased accuracy in the detection of the phonetic segment boundaries. Hence these landmarks can be used for guiding the boundary refinement process. Chitturi *et. al,* [13] proposed a landmark based boundary refinement procedure for HMM labels using multi-class support vector machines (SVM). However this method requires manual labels for training the landmark detection SVMs. On the other hand, the accuracies of signal processing based landmark detection methods are considerably low for use in high accuracy segmentation. However in the case of explicit segmentation, the expected landmark can be predicted from the phonetic transcript and the HMM boundary provides an approximate localization of the landmark. The proposed technique uses this information for boosting the accuracy of the signal processing based landmark detection methods. In this paper we do not discuss the refinement of inter-vowel boundaries and boundaries between vowels and glides.

The consonantal landmark expected at each HMM boundary is obtained using a mapping table described in [12]. The use of this table requires the knowledge of the phoneclasses and their constituent phones in a language. The consonantal landmarks are classified into sonorant(s), burst(b) and glottal(g) landmarks. Each landmark is further divided into **'+'** and **'-'** based on the expected energy change across the landmark. An increase in energy across the landmark corresponds to a **'+'** landmark and a decrease in energy corresponds to a **'-'** landmark.

The position of the expected landmark is detected in the speech signal around the HMM boundary. The algorithm described here for landmark candidate detection is adapted from the algorithm described in [11]. In this method the consonantal landmarks are detected using discontinuities in corresponding sub-band energy curves. Frequency ranges of the sub-bands used for detection of each landmark type are given in Table 1.

Table 1: *Frequency ranges of sub-bands for landmark type*

Band	Frequency Range (in Hz)	Landmark
1	0-400	g
2	800-1500	
3	1200-2000	
4	2000-3500	s & b
5	3500-5000	
6	5000-8000	

A search for discontinuities in corresponding sub-band energy curves, in the locality of the HMM boundary, using one-dimensional edge detection provides the required landmark candidates. A window w_{b_i} is selected around the i^{th} HMM boundary b_i for edge detection. This window is calculated as

$$w_{b_i} = [-1 * max(\frac{b_i - b_{i-1}}{2}, 50), max(\frac{b_{i+1} - b_i}{2}, 50)] \quad (1)$$

The minimum spread of the window on either side of the HMM

boundary was taken as 50ms, to account for severe alignment errors of the HMM segmentor. Further, if the segment on the right side of the boundary is a stop burst, the right bound of the window is calculated till the middle of the phone following the burst. The short duration of bursts and the presence of stop closures in the left context of the bursts, results in a portion of the closure being assigned to the burst segment in HMM boundaries generated after embedded training [10]. Hence the actual stop burst does not necessarily fall at the mid of the HMM bounded segment.

The large time windows and low edge detection thresholds, used for boundary candidate detection, enable the inclusion of required landmark in the set of possible candidates. However they also result in a large number of candidates for each boundary.

3.3. Cue based boundary candidate selection

A cue based selection algorithm is used to select the boundary from the landmark candidates. Park [12] and Liu [11] proposed two different acoustic cue based techniques for selection of landmark candidates. Liu [11] uses empirically determined thresholds to eliminate extraneous landmark candidates. Park [12] uses the knowledge of manual boundaries to build probabilistic distributions of cues at segment boundaries. Both these methods necessitate manual intervention. Further these methods are designed for use in the ASR scenario where the exact nature of the landmark is not known prior to the selection. However as explicit segmentation provides prior knowledge of the expected landmark, we proposed a selection procedure which eliminates the necessity for manual intervention.

The proposed method is based on the hypothesis that the behaviour of acoustic cues in the immediate neighbourhood of the ideal boundary candidate, is similar to those at the center of the phones on either side of the boundary. Hence the absolute differences between the acoustic cues in the immediate neighbourhood of the ideal boundary and those at the center of the phones on the respective sides are minimal. Further as the ideal boundary has a spectral discontinuity, there is maximal difference of acoustic cues in the immediate neighbourhood. These differences are used as elements of a cost function. The boundary candidate (bc_j) which maximizes the cost function corresponding to the expected landmark is selected as the boundary.

A typical cost function is composed of three elements *viz.*,

1. $\epsilon_l = |\alpha_{IL} - \alpha_{CL}|$
2. $\epsilon_r = |\alpha_{IR} - \alpha_{CR}|$
3. $\epsilon_i = |\alpha_{IR} - \alpha_{IL}|$

where α is an acoustic cue calculated over a span of 10ms, in the immediate neighbourhood of the boundary candidate bc_j on left (IL) and right (IR) side, or over a span of 10ms at the center of the phone on the left (CL) and right (CR) side of the HMM boundary b_i.

Acoustic cue, α corresponds to one of the following

1. E_H = Average high band (1.2-8KHz) energy
2. E_G = Average glottal band (0-400Hz) energy
3. E_{5-6} = Average energy in bands 5 and 6 (see Table 1)

These cues are selected as they help in distinguishing different types of landmarks. A subset of these cues is used in each cost function, specific to the landmark type.

An ideal landmark candidate minimizes ϵ_l and ϵ_r while maximizing ϵ_i. Combining these elements in a simple additive unweighted cost function we have

$$C_{lm}(bc_j) = -\epsilon_l - \epsilon_r + \epsilon_i \quad (2)$$

where lm is the landmark type and bc_j represents the j^{th} boundary candidate.

The HMM boundaries, though erroneous, are used to identify the center region of the phone segments; as the HMM bounded phone segments have considerable overlap with the actual phone segments. However the average energies calculated from the center of the stop burst segment cannot be assumed to represent the burst landmark energy profile, as bursts do not necessarily fall at the center of the HMM bounded segment (see 3.2). Hence the cost functions corresponding to **b** landmarks are designed without considering the cues from the center regions (CL and CR). Instead, the high band energy of the closure segment before the burst is compared with the average high band energy of all silence segments in the current utterance (E_{sil}).

The **+g** and **+b** landmarks, corresponding to the scenario of a voiced phone following a stop burst, occur within a short interval. Cues in the glottal band (E_G) are insufficient to differentiate these two closely occurring landmarks, leading to boundary selection errors. Energy in the bands 5-6 (E_{5-6}) can be used to distinguish the **+g** and **+b** landmarks, as burst landmarks are accompanied by broadband noise present even in these high bands. Hence in the case where a **+g** landmark follows the **+b** landmark, a separate cost function is used to identify the ideal **+g** candidate.

Table 2 summarizes the details of elements in the cost function, for each landmark (LM). The cost functions are same for both the '+' and '-' type landmarks, unless explicitly specified.

Table 2: *Elements of the cost function for each landmark type*

LM	Elements		
	ϵ_l	ϵ_r	ϵ_i
s	$\lvert E_{H_{CL}} - E_{H_{IL}}\rvert$	$\lvert E_{H_{CR}} - E_{H_{IR}}\rvert$	$\lvert E_{H_{IL}} - E_{H_{IR}}\rvert$
+g	$\lvert E_{H_{CL}} - E_{H_{IL}}\rvert$	$\lvert E_{G_{CR}} - E_{G_{IR}}\rvert$	
+g after +b	$\lvert E_{H_{CL}} - E_{H_{IL}}\rvert$	$\lvert E_{G_{CR}} - E_{G_{IR}}\rvert$	$E_{5-6_{IL}} - E_{5-6_{IR}}$
-g	$\lvert E_{G_{CL}} - E_{G_{IL}}\rvert$	$\lvert E_{H_{CR}} - E_{H_{IR}}\rvert$	
+b	$\lvert E_{H_{IL}} - E_{sil}\rvert$		$\lvert E_{H_{IL}} - E_{H_{IR}}\rvert$
-b		$\lvert E_{H_{IR}} - E_{sil}\rvert$	$\lvert E_{H_{IL}} - E_{H_{IR}}\rvert$

4. Experiment

4.1. Experimental Setup

The TTS database used is a single speaker Telugu language database comprising of 4800 utterances (at 16KHz), with a total duration of 30.6 hours. The performance evaluation was done on five hours of manually labelled utterances in the database. The phone sequences for explicit segmentation were automatically predicted from the orthographic transcripts.The iterations of isolated training combined with landmark refinement were performed until the average shift of boundaries in successive iterations started increasing.

4.2. Results

Table 3 has a comparison of baseline HMM segmentation method (BL), the iterative isolated training method using embedded HMM labels (IT) (described in [2]), proposed landmark

based refinement algorithm without iterative isolated training (LM) and proposed landmark based refinement with iterative isolated training (LM+IT). The accuracies of these segmentation techniques are measured as percentage of boundaries within 5ms, 10ms and 20ms deviation from the manual boundary.

Table 3: Performance of the refinement methods

Boundary Type	Deviation from manual boundary	Percentage of boundaries within specified deviation			
		BL	IT	LM	LM+IT
Total	5ms	28.9	25.9	34.8	37.0
	10ms	50.7	52.1	59.5	65.0
	20ms	73.3	81.9	82.2	88.6
s	5ms	38.5	40.4	37.0	44.2
	10ms	69.4	68.5	64.8	74.0
	20ms	94.0	89.8	89.4	93.4
g	5ms	35.3	19.9	44.5	45.1
	10ms	59.4	45.6	73.4	75.3
	20ms	76.2	78.4	88.7	93.4
b	5ms	9.6	25.6	18.9	22.8
	10ms	20.2	51.7	35.7	47.5
	20ms	48.9	85.3	66.8	82.4

Isolated training using baseline labels, disregarding the knowledge of the spectral cues, results in the HMM models capturing some errors in baseline labels introduced due to embedded training. This can be clearly observed in the performance of the isolated training procedure, in the placement of boundaries corresponding to **g** landmarks. However it can be seen that the LM refinement method increases the accuracy of boundaries to 88.7% ($< 20ms$). On the use of these refined boundaries in the iterative isolated training procedure the accuracy is boosted to 93.4%.

The overall accuracy (inclusive of even inter-vowel and boundaries between vowels and glides) of the baseline labels is 79.0% ($< 20ms$). On the inclusion of errors, in the placement of sub-phoneme boundaries before the stop bursts, the baseline accuracy falls to 73.3%. Using the proposed method the accuracy was increased to 88.6% (including the sub-phoneme stop burst boundaries). [10] and [4] report increases in accuracy to 88.4% and 94.8% respectively. However these refinement methods operate on labels whose baseline accuracies are 80.2% and 87.3% respectively. The current method on the other hand is able to operate on labels with initial accuracies as low as 73.3% due to the use of large time windows, during the detection and selection of boundary candidates. The use of these large windows is made possible due to use of boundary specific cost functions, which help select the desired boundary from a number of boundary candidates within a large time window.

The baseline accuracies in the placement of boundaries corresponding to **+g** landmark, following a **+b** landmark (**+g after +b**), are as high as 97.2% ($< 20ms$). This is a misleading statistic as 59.0% of these **+g** boundaries encroach on the preceding burst segment. Stop burst segments are typically of small durations ($\sim 25ms$) and a boundary error of even a few milliseconds could lead to mislabelling of the burst segment. Use of these mislabelled segments in unit selection synthesis, results in missing bursts in the synthesized output. As stop bursts are crucial for perception of stops, these mislabelling errors are detrimental to synthesis quality and have to be avoided during refinement. The proposed landmark based refinement algorithm reduces the error in the placement of the **+g** boundaries, while ensuring that the burst segment is not mislabelled due to encroachment. The proposed landmark based spectral correction technique ensures that these encroachments are reduced to 18.9% while maintaining the accuracy at 97.0% for this class of boundaries. The use of phone-class specific cue (E_{5-6}) ensures the proper assignment of these boundaries.

5. Conclusion

In this paper we explored the use of knowledge base, in the form landmark specific cost functions, to guide the boundary movement during the refinement process. The use of the knowledge base helped to distinguish between closely occurring spectral discontinuities and to correct large alignment errors of the HMM segmentor. The overall segmentation accuracy was increased to 89% (within 20ms from manual boundary) from the baseline accuracy of 73%. Thus the proposed refinement procedure is suitable for segmentation of TTS databases in under-resourced languages where the initial HMM labels have low accuracies, due to lack of properly annotated speech corpora for boot-strapping the acoustic models. The only resource necessary for applying the proposed method to a speech database in a new language is the knowledge of phone-classes and their constituent phones in the language, required to identify the expected landmarks.

6. References

[1] D. R. van Niekerk and E.Barnard, "Phonetic alignment for speech synthesis in under-resourced languages," in *Proceedings of the Interspeech*, Brighton, U.K., Sep. 2009.

[2] I. Mporas, T. Ganchev, and N. Fakotakis, "A hybrid architecture for automatic segmentation of speech waveforms," in *Proceedings of ICASSP*, Las Vegas, USA, Mar. 2008, pp. 4457–4460.

[3] S. J. Young et. al, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[4] Y. Kim and A. Conkie, "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in *In Proceedings of ICSLP*, Denver, SA, 2002, pp. 145–148.

[5] S. Jarifi, D. Pastor, and O. Rosec, "A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis," *Speech Commun.*, vol. 50, no. 1, pp. 67–80, 2008.

[6] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech & Language*, vol. 24, no. 2, pp. 273–288, 2010.

[7] J. Matoušek and J. Romportl, "Automatic Pitch-Synchronous Phonetic Segmentation," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 1626–1629.

[8] J. A. Antonio and A. Bonafonte, "Towards Phone Segmentation For Concatenative Speech Synthesis," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004, pp. 139–144.

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

[10] S. Hoffmann and B. Pfister, "Fully Automatic Segmentation for Prosodic Speech Corpora," in *Proceedings of INTERSPEECH*, 2010.

[11] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," Ph.D. dissertation, MIT, 1995.

[12] C. Park, "Consonant landmark detection for speech recognition," Ph.D. dissertation, MIT, 2008.

[13] R. Chitturi and M. Hasegawa-Johnson, "Novel Entropy based moving average refiners for HMM Landmarks," in *Proceedings of ICSLP*, 2006.