A MULTI-PASS LINEAR FOLD ALGORITHM FOR SENTENCE BOUNDARY DETECTION USING PROSODIC CUES

Dagen Wang, Shrikanth S. Narayanan
{dagenwan@,shri@sipi}.usc.edu
Signal and Image Processing Institute
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089

ABSTRACT

We propose a multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech. It uses only prosodic cues and does not rely on segmentation information from a speech recognition decoder. We focus on features based on pitch breaks and pitch durations, study their local and global structural properties and find their relationship with sentence boundaries. In the first step, the algorithm, which requires no training, automatically finds a set of candidate pitch breaks by simple curve fitting. In the next step, by exploiting statistical properties of sentence boundaries and disfluency, the algorithm finds the sentence boundaries within these candidate pitch breaks. With this simple method without any explicit segmentation information from an ASR, a 25% error rate was achieved on a randomly selected portion of the switchboard corpus. The result from this method is comparable with those that include word segmentation information and can be used in conjunction to improve the overall performance and confidence.

1. INTRODUCTION

A crucial requirement for robust information extraction from speech is automatic determination of sentence boundary [1]. Simple methods relying on signal energy features, such as those used in speech end point detection, are not adequate to address this problem. Recently, significant efforts have been directed toward utilizing higher level linguistic information in the speech such as in creating an hidden event statistical language model wherein sentence boundaries and disfluencies are modeled as hidden events[2][3]. Many of these techniques rely on information such as phone/word segmentation made available by an automatic speech recognizer (ASR). While these approaches have signaled progress, they often suffer from potential difficulties related to dealing with (ASR) errors and the inherent ineffectiveness in modeling hidden events with just segmental information. To address this issue, a promising and frequently sought after solution is to utilize one of the key pieces of information ignored during ASR, viz., prosody related cues.

1.1. Prosody cues

In recent years, there has been increasing attention paid to the use of prosody in automatic speech recognition and understanding. Prosodic cues are known to be relevant in characterizing discourse structure across languages and therefore are expected to play an important role in various information extraction tasks. For example, a long pause in conjunction with a preceding phrase-final low boundary tone, and a subsequent pitch range reset might imply a sentence boundary.

A companion problem with sentence boundary detection is disfluency detection [6]. Disfluencies (e.g., fillers such as "um", repeats, self-repairs) are prevalent in spontaneous speech. In studies of spontaneous speech, it has been found that the probability of disfluency is exponentially proportional to the sentence length [4]. In the context of sentence boundary detection that deals with lengthy speech utterances, disfluency is almost unavoidable. These two problems together with end of utterance detection (EOU) are being studied increasingly through the use of prosodic features.

A popular method is to use statistical learning such as CART style decision trees, as for example, the study by Shriberg et al[1]. More recent results by these researchers indicate a 22.9% sentence boundary detection error rate on the switchboard corpus [5]. In their approach, a variety of spectral and temporal features are first extracted in a local region around a word boundary obtained from an ASR (200 ms across both directions from the trailing and leading edges of the words defining the boundary). A CART style decision tree is then trained using these features.

In this method, since the training data employed come from a large number of speakers, the parameters of the tree provide an "averaged" representation of various speakers' speaking styles. As such, the decision trees could not reflect the specific speaking styles of an individual speaker. Also, features extracted within a local window do not reflect the global (utterance level) speaking styles: Features such as speaking rhythm, global pause distribution are, however, important. A final limitation arises from the ASR word alignment error, which is almost unavoidable with current ASR techniques. Incorrect word boundary hypotheses often lead to higher error rates in decision tree methods [5].

This paper attempts to address some of the challenges and limitations listed above, in proposing an algorithm that is speaker dependent, ASR independent and that uses some global features for sentence boundary detection.

1.2. Pause features

Pause features were found to play a critical role in boundary and disfluency detection problems. In measurements of decision-tree feature usage, the pause-related features were found to be queried the most. [1] A summary of results related to the pause

behavior in spontaneous speech from a study on spontaneous speech by Nakatani and Hirschberg is provided in Table 1 [8]:

Pause Juncture	Mean	Std Dev	N
Fluent Pause	513ms	676ms	1186
Disfluency Pauses	334ms	421ms	346
. Fragment	289ms	377ms	264
. Non-Fragment	481ms	517ms	82

Table1: Characteristics of pause durations in spontaneous speech (from [8])

From this table, we could observe that fluent pause is statistically longer than the disfluent pause. Non-fragment disfluency pause is similar in its length to fluent pause but its frequency of occurrence in the corpus was rare. We could also see the deviations in these statistics are also large. This is attributed to differences in individual speaking styles (i.e., individual differences in pause allocation) and or differences in speaking rates. Our algorithm is specifically useful in dealing with the second situation.

Based on these data, people have tried to employ alternate values for static threshold, such as 400ms, to classify disfluent and fluent pause, but these results were found to be in general "unreliable" [8]. In this paper, instead of setting such static thresholds to detect boundaries, we attempt to model the behavior of the internal breaks and use it to "predict" the correct classification boundary on a per speaker basis.

Another contribution of this work is in the pause representation. In Shriberg's study[1], pause durations are retrieved by automatic speech recognition. In this paper, smoothed pitch duration and break statistics are utilized to obtain sentence boundaries without needing ASR segmentation information. These features are found to be faster to retrieve and enable real time processing possible. Also the inherent word boundary alignment error in ASR is avoided.

The major questions that are addressed next are:

1. How to derive a threshold to robustly narrow down the range of pitch breaks from which we can find the sentence boundaries?

2. How to deal with breaks that are brought about by some type of disfluency instead of a true sentence boundary?

1.3. Disfluency and their phonetic consequences

In Shriberg's phonetic study of disfluency [7], disfluency is divided into 3 phases: reparandum, editing phase and repair. Among these, reparandum and editing phase may have the most significant influence on pitch break duration distributions. There are many special phonetic features characterizing these two phases, including lengthened pause, lengthened reparandum syllable, intonation repetition, word cutoffs and laryngealization. Some of these features may help us to do post processing in distinguishing sentence boundaries from disfluencies. Some of these features, such as lengthened pause deemed to influence pitch breaks' distribution are considered in this study. It should be noted that some of the other properties such as laryngealization are difficult to measure without additional information and are hence not included in this study.

2. MULTI-PASS LINEAR FOLD ALGORITHM

2.1. Pitch break behavior

Figure 1 shows the pitch (F0) contour for an example speech signal (Switchboard corpus). We can see that the sentence boundary (<bd>) shows a larger break duration, and the pauses around the filled pause "Uh" are large as well. These are quite apparent relative to the other breaks seen in the speech stream.

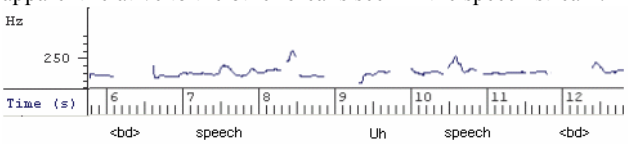

Figure 1: F0 (pitch) contour showing pauses for sentence boundaries and fillers in an example utterance from the Switchboard corpus.

Indeed, in normal spontaneous speech, such phenomena are statistically dominant in the sense that people tend to pause in sentence boundaries and disfluencies [8]. Other types of breaks that occur in speech include intra- and inter-word breaks and breaks corresponding to other manifestations of disfluency (such as repeat, repair and delete). In general, the values corresponding to these breaks tend to be smaller than that for a sentence boundary or filler [8].

2.2. The sorted pitch break map

In this algorithm, we mainly focus on pitch breaks, pitch properties in the break neighborhoods and global statistics such as distribution of the pauses. The procedure used to obtain the pitch values was similar to that in [1]. (ESPS [10] and post processing using Speech Filing System [9])

The key step in this algorithm is to perform the ascending sort operation with respect to the pitch break durations (calculated for each dialog turn). In Figure 2, each point in the figure represents a pitch break. The order of these breaks is no longer the same as the order in which they occur in the utterance, but sorted by the length of their duration.

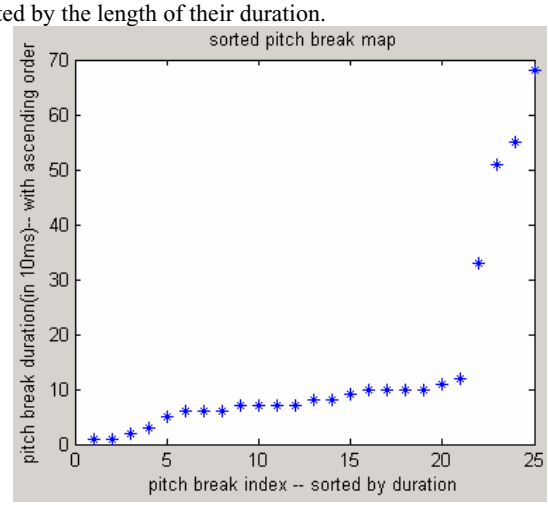

Figure 2: Sorted pitch break map for one dialog turn. Note that a turn may contain more than one sentence boundary.

Analysis of several such sorted pitch break maps led to the following observations:

1. The map tends to be roughly "decoupled" into 2 regions. The lower portion, corresponding to the shorter duration breaks, has most of the pitch break points occurring in a

spontaneous speech utterance. The upper region corresponding to larger duration breaks is more sparsely populated. It was found that the lower region is related to most of the intra-word and inter-word breaks. The upper portion is related to longer breaks corresponding to filled pauses, disfluency, hesitation and other special breaks (with unexplained speaker behavior in some cases).

2. When the incidence of disfluency increases, the second part tends to have more points and the "transition area" between the two parts tends to get increasingly populated.

2.3. Linear fold algorithm

Now the problem is to translate these observations into a model. Classical statistical detection was not appealing primarily due to the aim for performing the detection in an "unsupervised" basis and on a per speaker basis of this non-linear behavior of pause distribution. From the first observation above, it appears that the pitch breaks could be decoupled into two parts. The idea hence is to use two lines to fit the pitch break map in the mean square sense. For any n pitch breaks, there are total (n-1) methods to cut the points into 2 groups. The implementation is done so as to cut the points in the sorted pitch break map into one of these two regions and doing a linear fit to each group independently. Then we choose the cut which has the minimal mean squared deviation. Applying such an operation to the pitch break map of Fig. 2 now yields the fitted map of Figure 3:

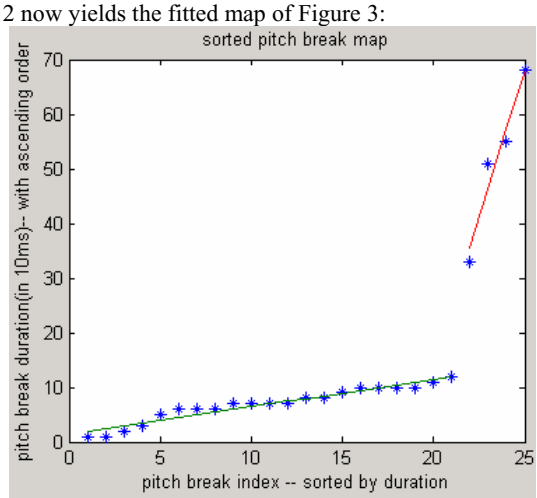

Figure 3: Linear fold algorithm

As could be seen, this one pass linear algorithm could fit the data well most of the time. The lower line slope is decided by the small breaks that safely encompass most of normal intra-word and inter-word breaks. The upper line represents fluent pause, disfluency, hesitation and other special breaks. The breaks in the upper segment are the special focus of our study.

A test with a representative portion of the switchboard corpus (100 dialog turns) showed that 91% of these utterances can be handled well with this simple linear fold algorithm. Yet there are cases where there is a rich transition area between these two linear lines and/or the upper segment encompasses too many points that it is highly overlapped in the transition region. The simplest method do deal with this would be to recursively applying this algorithm. The only difference is removing the lower half segment of the model in each iteration.

Figure 4 shows the results of two successive such operations. The left figure is the result of the first fit while the right figure is the result of the second fit.

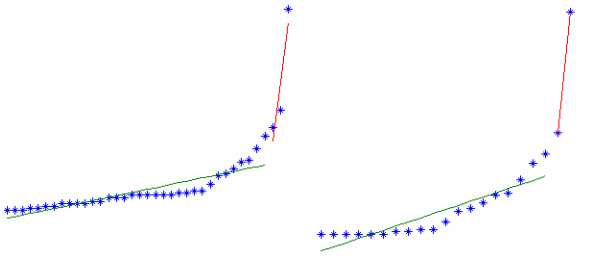

Figure 4: Illustrating the two pass linear fold algorithm: result of the first pass is to the left and the second pass to the right.

2.4. Post processing

The upper segment resulting from a multi linear fold fit includes not only fluent pause, but also disfluent pause. As mentioned earlier, in contrast to the work based on localized information [1], we propose to utilize more global utterance information.

Figure 5 shows some candidate breaks in the upper-most segment after the linear fold operation (in the time axis). Based on published results and our own observations, we summarize the following properties to help find the true sentence boundaries:

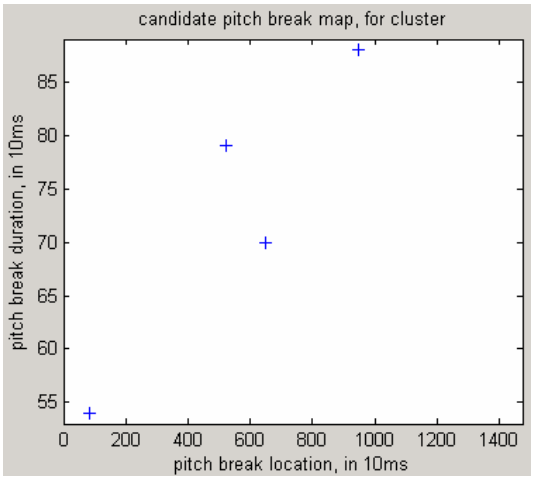

Figure 5: Pitch breaks in time axis

Property 1: Disfluency is more likely to appear earlier in a sentence. Shriberg's study [4] in spontaneous corpus (e.g., SWBD/ATIS/AMEX) supports this fact. The first break point from the left in Fig. 5 reflects this case. Since such a break appears so early in a sentence, it could not be sentence boundary, and we normally remove such breaks in our algorithm.

Property 2: Sentence boundaries should be distributed evenly and could not appear too close to one another . The second and third breaks in Fig. 5 represent such a case. Since it is almost impossible to associate real linguistic events to such a pair of points, this implies the presence of at least one of them related to disfluency.

Property 3: Occurrence of pitch resets associated with sentence boundaries. If we observe a pitch reset, which is a flag for a sentence boundary, we declare a boundary detection.

Property 4: Break duration values. If no other cue is available, when we wish to choose a sentence boundary marker from a pair of candidates, we could simply pick the one with the longest duration.

The implementations of these decisions are illustrated in Fig. 6:

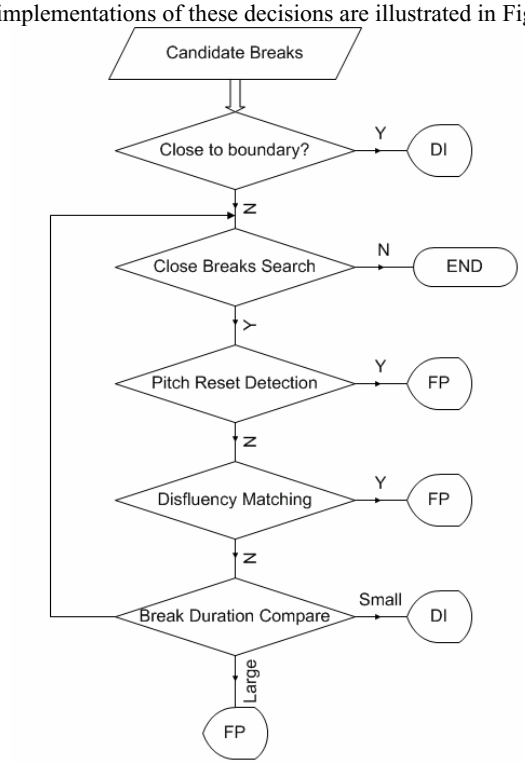

Figure 6: Post processing

3. EXPERIMENTAL RESULTS

Experiments were performed on the switchboard database. It should be noted that the data considered for these experiments did not include utterances with any serious back channeling or other strong background noise and music, etc. These phenomena severely degrade the boundary detection and investigations of robust approaches against such effects are outside the scope of this work.

We randomly chose speech from 100 dialog turns and performed the operation proposed in this paper. The overall error rate in sentence boundary detection was 25%.

	Number
Sentence boundaries	351
Missed boundaries	25
False alerts	63

Table 2: Results of sentence boundary detection.

Here we give an analysis of the errors:

False alerts were the major source of errors. This is similar to the findings of [8]. Phonetic similarities between disfluency and fluent pause were the major reason. Richness of such errors also is related to the corpus we chose. In "clean" speech, such errors are not expected to be so significant.

The existence of "missed boundary" error often related to the variability in the user speaking state, especially while the user tends to speed up the speech, or when he tends to be emotional (e.g., excited). These errors were however less dominant.

4. DISCUSSIONS AND FUTURE WORK

The best results so far yield error rates of about 22.9%[5] using a range of features and a decision tree based analysis on data from the same corpus. In contrast, the proposed method which does not rely on any ASR-based information, and uses just prosodic cues, obtained reasonable good and comparable results. It is also very interesting to combine the result of this method with statistical methods, either as a post processing, or even as an additional set of features to further boost the confidence of the performance.

The post processing method we considered was a simple decision tree. The difference here is these rules are "learned" from human knowledge, but not directly by the machine. Machine learning techniques can facilitate efficient real time processing. Support vector machines, and fuzzy inference techniques appear to be good candidates for the boundary and disfluency detection problems; these are topics for future work.

5. REFERENCES

[1] E. Shriberg, A. Stolcke, D. Hakkani-Tur, & G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", Speech Communication 32 (2000), pp. 127-154. Special Issue on Accessing Information in Spoken Audio.

[2] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tur, & Y. Lu (1998), "Automatic detection of sentence boundaries and disfluencies based on recognized words". In R. H. Mannell and J. Robert Ribes (Eds.), Proceedings of ICSLP (Vol. 5, pp. 2247-2250). Sydney.

[3] Stolcke, A., and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech., Proceedings of ICSLP (Vol. 2, pp. 1005-1008). Philadelphia.

[4] Shriberg, E. 1996. Disfluencies in Switchboard. Proc. ICSLP 1996, Philadelphia, October 1996, vol. addendum, pp. 11. C14.

[5] E. Shriberg and A. Stolcke (2003), Prosody Modeling for Automatic Speech Recognition and Understanding To appear in Mathematical Foundations of Speech and Language Processing, M. Johnson, S. Khudanpur, M. Ostendorf, and R. Rosenfeld (eds.), Volume 138 in IMA Volumes in Mathematics and its Applications, Springer-Verlag.

[6] Y. Liu, E. Shriberg, & A. Stolcke (2003), Automatic disfluency identification in conversational speech using multiple knowledge sources. Proc. Eurospeech, Geneva.

[7] Shriberg E. (1999). Phonetic Consequences of Speech Disfluency. Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organizers), Proc. International Congress of Phonetic Sciences, pp. 619-622, San Francisco.

[8] C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech", JASA, pp. 1603-1616, 1994.

[9] Speech Filing
http://www.phon.ucl.ac.uk/resource/sfs/

[10] ESPS code release,
http://www.speech.kth.se/wavesurfer/links.html