Automatic, model-based detection of pause-less phrase boundaries from fundamental frequency and duration features

Mahsa Sadat Elyasi Langarani, and Jan van Santen

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

{elyasila, vansantj}@ohsu.edu

Abstract

Prosodic phrase boundaries (PBs) are a key aspect of spoken communication. In automatic PB detection, it is common to use local acoustic features, textual features, or a combination of both. Most approaches – regardless of features used – succeed in detecting major PBs (break score “4” in ToBI annotation, typically involving a pause) while detection of intermediate PBs (break score “3” in ToBI annotation) is still challenging. In this study we investigate the detection of intermediate, “pause-less” PBs using prosodic models, using a new corpus characterized by strong prosodic dynamics and an existing (CMU) corpus. We show how using duration and fundamental frequency modeling can improve detection of these PBs, as measured by the F1 score, compared to Festival, which only uses textual features to detect PBs. We believe that this study contributes to our understanding of the prosody of phrase breaks.

Index Terms: Prosody event detection, Phrase break detection, intermediate phrase boundaries

1. Introduction

Phrase boundaries (PBs) are important in human-human, machine-human (i.e., text-to-speech synthesis, or TTS), and human-machine (i.e., automatic speech recognition, or ASR) communication. In human-human communication, PBs are used to chunk speech into semantic or syntactic units, not only as a natural by-product of how speech is “computed” by the brain or as a result of limitations of the speech production apparatus (e.g., running out of breath) but also as a device to make it easier for the listener to understand the message. In TTS, again intelligibility is key. Finally, in ASR, PBs contain useful information that helps recognition at the word or phoneme level. Automatic detection of PBs is important both for TTS, for training systems that predict PBs from text; and for ASR, as an integral part of the recognition process.

The acoustic-prosodic correlates of PBs involve both fundamental frequency (F0) and temporal features. PBs can be conveyed by, for example, final lowering for PBs at the end of utterances that are statements, final rises for PBs at the end of utterances that are yes/no questions, and continuation rises for non-utterance-final PBs. In the temporal domain, PBs can be indicated by, for example, the presence of pauses or phrase-final lengthening. There has been a large amount of work on the relationship between prosodic information and PBs [1, 2, 3, 4]. An extensive survey on prosody boundaries and prominence in language processing can be found in a review article by Wagner [5].

In ToBI annotation, there are two levels of PBs: intonational PBs (break score “4” in ToBI standard) and intermediate PBs (break score “3” in ToBI standard). Intonational PB’s are frequently indicated by followed pause while intermediate PBs are indicated by phrase-final F0 changes and phrase-final lengthening. While PBs involving pauses (PB^+, intonational PBs) are relatively easy to automatically detect, pause-less PBs (PB^−, intermediate PBs) are much harder to detect [6], for two reasons: 1) F0 contours may pass entirely smoothly through the PB. 2) Lengthening is difficult to assess because phoneme durations depend on many other factors besides the presence of a PB. For example, a 120 ms Schwa is relatively long while a 120 ms /aI/ is relatively short, and a 120 ms stressed OH is short while a 120 ms unstressed OH is long — there is no fixed ms boundary that defines whether or not a vowel is lengthened [7].

The core aim of this study is to detect PB^−’s by applying concise (i.e., having few parameters) quantitative models for F0 and for speech timing. We apply these models to speech recordings that have been orthographically (but not phonemically) transcribed, and combining the information provided by these models. We will use an F0 model that uses the concept of a (left-headed) foot [8] to determine which PB assignment (i.e., specification of between which words PB^−’s are present) provides the best fit of the model, taking advantage of the assumption that feet are necessarily terminated by a PB. We will use a duration model that measures pre-boundary lengthening by predicting the duration of a vowel based on all factors known to affect vowel duration, but excluding boundary-related factors. Simply by comparing observed and predicted durations for vowels in word-final syllables, we can obtain a measure of phrase-final lengthening. We confine ourselves to vowels, because their durational behavior is particularly well-understood (e.g., [9]). Finally, we combine the F0 and duration information to optimally predict PB^−’s. As our reference, we use Amazon Mechanical Turk to obtain consensus PB assignments.

2. Corpora

2.1. Prosodically Rich Database (PRD)

We selected 100 sentences from the AP Newswire (years 1988-1990), automatically annotated in terms of factors relevant for prediction of duration [7], and used greedy methods to select text with maximal coverage of the resulting feature space [10]. These sentences contained on average 19 words. Two female American English speakers, both experienced actresses / voice talents, were given carte blanche as to how to read...
these sentences as long as their utterances were affectively and
prosodically meaningful, natural, and exciting-sounding. All
sentence-internal punctuation was removed, but the speakers
were instructed to insert PBs as judged appropriate; no instruc-
tions were provided in terms of whether PBs should contain
pauses or involve specific intonational cues. The record-
ings from Speaker 1 were phonetically transcribed and segmented
manually, the recordings from Speaker 2 were graphemically
transcribed manually (i.e., slight deviations from the read text
were corrected) but were segmented automatically using the
HTK toolkit [11]; no manual corrections were made in the latter
case.

2.2. CMU Arctic speech database

We also used the CMU Arctic speech database [12]. We used
speaker SLT, a US English female. The database was
automatically labelled via CMU Sphinx using FestVox labeling
scripts. No hand corrections were made. This corpus con-
tains 1132 utterances; we extracted 100 utterances that were
most similar to those in the Prosodically Rich Database (PRD),
in the following sense. For each sentence in the PRD, we
found the best concordance between all characters of the sen-
tence with all sentences of CMU using a global alignment algo-
rithm (Bio.pairwise2.align function) from BioPython [13]. For
that sentence, we stored the 10 best-aligned sentences from the
CMU corpus. Finally from this set (100 × 10), we extracted 100
sentences with the highest matching scores.

3. Methods

3.1. Group-wise reference boundary assignment

Agreement among human labelers or between the latter and
automated labelers is extremely high for PBs with pauses, or
PB+. This is less the case for PB−’s [6]. To generate a ref-
erece, we used Amazon Mechanical Turk [14], with native
speakers (master participants who have approval ratings of at
least 95%). Their task was to determine, for all 100 sentences
in a database, the location of PBs, regardless of PB type. We
used two contexts: a text-only context and a text-plus-speech
context. In the first context, at any given trial, the labelers were
presented with the text displayed in normal, horizontal format,
accompanied by a vertical list of the words, displayed in the
same order, and each word followed by a button. The task was
to click on any words that were felt should be followed by a
comma or period. The second context was identical, except that
the labeler also listened to the sentence. We hired 15 unique
labelers for each database and context, for a total of 90 unique
labelers.

Disagreement between the labelers (native speakers) was
handled as follows. For example, consider the sentence, S. “I
like cooking dogs and kids.”, which received two different PB
assignments (1), from labelers l1 and l2 (generating boundary
assignments l1(S) and l2(S)) and (2) from labelers l3, ..., l15,
generating corresponding boundary assignments.

1. “I like cooking dogs[PB−] and kids[PB+]”

We will evaluate the performance of an automatically gen-
ated boundary assignment by comparing it with these labeler-
generated assignments, that, naturally, will not be in full agree-
ment. But preliminary to this, we need to assess the agree-
ment among the labelers, which we did as follows. For each
sentence, we split the group of 15 labelers 100x into two sub-
groups, computed the respective unions of the boundary assign-
ments for each group, and then computed the group-wise agree-
ment (Algorithm 1) for these unions, measured via Occurrence
agreement (Equation 1) and Total agreement (Equation 2). Re-
results of group-wise agreement are presented in Table 2.

\[
\text{Occurrence agreement} = \frac{TP}{TP + FP + FN} \times 100 \quad (1)
\]

\[
\text{Total agreement} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (2)
\]

Algorithm 1 group-wise agreement

1. for S in Sentences do
2. L ← {l1,l2,...,l15}
3. A ← all subsets if size 7 of L \( \{\frac{L}{2}\} \)
4. for subset in A do
5. \text{subset' ← L − subset}
6. \text{C(subset) ← } \bigcup_{i=1}^{\text{subset}} l_i(S)
7. \text{C(subset') ← } \bigcup_{i=\text{subset'}} l_i(S)
8. O ← Occurrent agreement(C(subset),C(subset'))
9. T ← Total agreement(C(subset),C(subset'))
10. end for
11: end for
12: report average of O and T

3.2. Constraining the PB search space

The number of (internal) boundary assignments for a sentence of
19 words is 3^{18} (each word can get two types of PBs or noth-
ing and last word has to get \( PB^+ \)), which exceeded the com-
pute power available for this project. We instead constrained
the set of PB assignments considered for a given sentence by
the methods discussed next.

3.2.1. Expert

We hired two linguistically informed experts to manually indi-
cate PBs for each database. They used Praat [15] for annotating
pitch accent labels and PB labels. They also had access to pho-
netic transcriptions and segmentation [16].

3.2.2. Festival

We employed Festival to predict pitch accents and PBs for each
database. Festival predicts PBs at the word level, based on an
algorithm presented in [17]. It also predicts pitch accents at
the syllable level. We moved the pitch accent labels to the word
level, such that if one syllable of a word is accented then the
whole word is accented. Only textual information is used for
this prediction without any acoustic or prosodic information.

3.2.3. Combination of Festival and Expert (Comb)

We combined PB and pitch accent labels from Expert and Festival
by considering their unions. Pitch accent labels in this
method are obtained via the union of Expert-pitch-accent labels
and Festival-pitch-accent labels. $PB^-$ are also obtained via the union of the Expert’s $PB^-$ labels and Festival’s $PB^-$ labels. These methods are different in terms of pitch accent labels and the location of $PB^-$ labels but they all have the same $PB^+$ labels.

3.3. Usage of fundamental frequency model

Naturally, the assignments resulting from the Expert, Festival, and Combo methods, because they are based on unisons, generally contain too many $PB^-$’s. In this section, we describe how using $F_0$ information can be used to select a specific boundary assignment for each sentence, which we will then compare with the group-wise boundary assignments as reference.

Recently [18, 19], we proposed a superpositional model to estimate $F_0$ contour using syllable stress, pitch accent, and PB labels. It decomposes a continuous $F_0$ contour into component curves in accordance with the General Superpositional Model. According to this model [8], the $F_0$ contour for a single-phrase utterance can be written as the sum of a phrase curve and any number of accent curves, one for each foot. In this method [18], the phrase curve consists of two log-linear curves, between the phrase start and the start of the phrase-final foot, and between the latter and the end point of the last voiced segment of the phrase, respectively. We use a combination of the skewed normal distribution and a sigmoid function to model three different types of accent curves. First, the skewed normal distribution is employed to model rise-fall accents that occur in non-phrase-final positions as well as, in statements, in utterance-final positions. Second, a sigmoid function is used to model the rise at the end of a yes/no question utterance. And, third, the sum of the skewed normal distribution and the sigmoid function is used to model continuation accents at the end of a non-utterance-final phrase (for details, see [18]).

We used this model for the present purposes as follows. The syllable stress labels were dictionary-based. We also have PB labels and pitch accent labels resulting from each method (as described in section 3.2). For each sentence, we consider all combinations of occurrence/non-occurrence of the $PB^-$ labels of that sentence (we call these combinations for a given sentence “phrase boundary assignments”). For each assignment, we fit the $F_0$ model, which results in a root weighted mean square error (RWMSE) for that sentence. Then we determine the PB assignment resulting in the lowest RWMSE. In other words, by applying the $F_0$ model to the Expert, Festival, and Combo assignments we can in principle detect the $PB^-$’s. We call these methods: Expert+$F_0$, Festival+$F_0$, and Combo+$F_0$. Table 1, the average best RWMSE between estimated $F_0$ contour and raw $F_0$ is shown for each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Spk1</th>
<th>Spk2</th>
<th>CMU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>5.9048</td>
<td>6.1812</td>
<td>–</td>
</tr>
<tr>
<td>Expert+$F_0$</td>
<td>5.5212</td>
<td>6.1490</td>
<td>–</td>
</tr>
<tr>
<td>Combo</td>
<td>5.5661</td>
<td>6.4945</td>
<td>–</td>
</tr>
<tr>
<td>Combo+$F_0$</td>
<td>4.9943</td>
<td>5.8964</td>
<td>–</td>
</tr>
<tr>
<td>Festival</td>
<td>6.3230</td>
<td>7.1157</td>
<td>4.6707</td>
</tr>
<tr>
<td>Festival+$F_0$</td>
<td>5.7905</td>
<td>6.5460</td>
<td>4.4433</td>
</tr>
</tbody>
</table>

Table 1: Average of Root weighted mean square error (in Hz) between fitted $F_0$ contour and raw $F_0$ (only voiced parts are considered)

3.4. Usage of duration model

As mentioned in the Introduction, phrase-final lengthening is a well-established prosodic cue for PBs, with some of the earlier work reporting lengthening at many types of boundary (e.g., [20]), not just at the boundaries considered by ToBI. We will use a simple model that expressed vowel duration as a sum of product terms, with each component of a product depending on a specific factor (e.g., stress, post-vocalic consonant) [7, 21, 22]. Special cases of the sum-of-products model include the additive model (each product term has just one factor) and the multiplicative model (a single product term containing all factors). Using this model, it was shown that phrase-final lengthening is largely confined to phrase-final syllables, with much weaker lengthening for earlier syllables [7]. We therefore confine our attention to vowels in phrase-final syllables.

The duration of a vowel depends on many features in addition to position in the phrase. The sum-of-products model was used to take into account these factors in order to evaluate the presence of lengthening. We fitted the additive version of the model using the following features: the phoneme whose duration is of interest, next phoneme, previous phoneme’s stress label (binary), current syllable’s stress label (binary), and current word’s accent label (binary). Key is that we did not include position in the phrase as a feature in this prediction. Also note that we excluded both sentence-initial and sentence-final vowels, since this would confound the parameter estimates for the features included in the analysis.

By letting $D_{Obs}$ be the observed duration of the $i^{th}$ vowel in a sentence and $D_{Pred}$ the predicted duration using the duration model, we define the ratio of the observed to the predicted duration of the vowel as $R_i = D_{Obs}/D_{Pred}$. Then, we extract a sequence of ratios, normalized per sentences (Equation 3).

$$\text{Sig} = \frac{R_i}{\text{Median}\{R_i | i \notin PB\}}$$ (3)

Thus, the sequence $\text{Sig}$ is a vector that, by construction, provides hints about which vowels may be lengthened, and thus about possible PBs. After extracting the $\text{Sig}$ vectors for all sentences for each of the six methods (three labeling methods, and whether or not $F_0$ information was used), we trained a logistic regression model [23] to predict the PB assignments. In each case, we split the data into 10 partitions, and applied 10-fold cross validation. The suffix “-Dur” is used to represent the usage of the duration model in a given method. We note, however, that the estimation of the duration parameters and hence of $D_{Pred}$ was not part of the cross-validation procedure. However, given the extremely small number of parameters compared to vowel tokens (30 compared to over 2,500), the risk of over-training is minimal.
Figure 1: This figure summarizes the F1 score of all methods in two contexts (text and text-plus-speech) for the three speakers. Three different colors DarkGray, Silver, and WhiteSmoke are used for representing results of the Expert, Comb and Festival methods, respectively. Also we use three patterns to show which prosodic information is used. Medians and means are represented by solid horizontal black line and black star in each box-plot, respectively. The p-values are based on the Exact Wilcoxon test.

4. Experiments

We used the two contexts (text and text-plus-speech) to study the effect of prosodic information (duration and \(F_0\)) on PB detection. For each speaker, we extracted PB assignments of each sentence via each method \(X\) (\(X = \text{Expert}, \text{Festival}, \text{or Comb}\)), their combination with \(F_0\) information \((X + F_0)\), with duration information \((X + \text{Dur})\), and with \(F_0\) and duration information \((X + F_0 + \text{Dur})\). These assignments are compared with the group-wise reference assignments. Because most word boundaries are not PBs (roughly 80% of word boundaries are not PBs), percentage of correct predictions is a biased measure. We use the F1 score (Equation 4) as performance measure. Since the location of PB+ is the same for all methods, they are not considered in results of this study.

\[
F1 = \frac{2TP}{2TP + FP + FN} \quad (4)
\]

The F1 scores for all methods are summarized in Figure 1. In comparison between the three methods (boxes without pattern), Expert performs better than Festival since the experts have access to all the acoustic/prosodic/textual information (Figure 1 DarkGray Vs. WhiteSmoke). Also, we expect that Comb performs worse than Expert (more PB assignments to choose from causes higher False Positive) and better than Festival (more PB assignments cause higher True Positive). The ordering showed in Equation 5 suggests that adding more acoustic/prosodic informations, results in more accurate assignments.

\[
\text{Expert} > \text{Comb} > \text{Festival} \quad (5)
\]

The order of the above equation does not change when \(F_0\) information (Figure 1, boxes with backslash pattern) is incorporated (Equation 6).

While adding \(F_0\) information to Expert and Festival improves the F1 scores, it does not improve the performance of the Comb method \((\text{Comb} + F_0 \equiv \text{Comb})\). A reason for that is the \(F_0\) model that we used is an optimization-based method. In \(\text{Comb} + F_0\) method, the number of optimization parameters increases by combining PB labeling of two methods (Festival and Expert) which causes the model to be over-fitted to the \(F_0\) contour.

\[
\text{Expert} + F_0 > \text{Expert} > \text{Comb} + F_0 \equiv \text{Comb} > \text{Festival} + F_0 > \text{Festival} \Rightarrow \text{Expert} + F_0 > \text{Comb} + F_0 > \text{Festival} + F_0 \quad (6)
\]

For studying the effect of phrase-final pre-lengthening, we apply duration information \("+Dur\) to the three methods (Figure 1, boxes with slash pattern Vs. boxes without pattern). The \(\text{Expert} + \text{Dur}\) shows lower performance than the Expert in the two contexts (text and text-plus-speech). In the Festival case, \("+Dur\) results in significant improvement for the PRD; however, this improvement can not be seen in the CMU arctic database. A reason for that might be the complexity of the PRD sentences compared to the CMU arctic database. Therefore, adding the duration information to the methods not only changed the ordering on Equation 5, but also shows different behavior for different speakers and methods (Equation 7).
While using $F_0$ information and duration information individually produced minor improvement, their combination resulted in major improvements, especially in the Comb and Festival cases (in Figure 1, Silver boxes with no pattern vs. Silver boxes with ‘×’ pattern, and WhiteSmoke boxes vs. WhiteSmoke boxes with ‘×’ pattern). Equation 8 shows the relationship between $X^{+Dur}_0$ methods for the two speakers (Speaker 1, and Speaker 2).

In the early part of this section we mentioned that the higher number of PB assignments in the Comb method was the reason that Comb performed worse than Expert (Equation 5). However, in $Comb^{+Dur}_0$, $+Dur$ and $+F_0$ appeared to filter out incorrect PB assignments, resulting in better performance by decreasing the False Positives.

$$Comb^{+Dur}_0 \geq Festival^{+Dur}_0 \geq Expert^{+Dur}_0$$ (8)

The equality of Festival’s performance in CMU(text) and CMU(text-plus-speech) implies that the PB’s of the CMU speaker matched the grammatical PBs. In other words, for the CMU arctic database using acoustic-prosodic information had less effect on the decision by the labelers.

Also, we employ the Exact Wilcoxon Test [24] to assess whether the following pairs: $(X, X_+F_0)$, $(X, X^{+Dur})$, and $(X, X^{+Dur}_0)$ are significantly different (the P-values are shown in Figure 1).

5. Conclusion

In this study, we presented how prosodic information can be used for the detection of pause-less phrase breaks. We find that the combination of duration and $F_0$ information results in the best performance.

These models have three advantages. First, they use very few parameters, making the methods usable in cases where few data are available. This is in particular the case in special populations, such as dialect groups or individuals with speech or language challenges. Second, they make use of global as well as local information available in an utterance. Third, they may allow us to “connect” this line of research with linguistics research, because the models are grounded in such research.

Limitations of the study include the following. First, we limited the search space by not considering the very large space of possible boundary assignments for a given sentence. Considering such large spaces may cause further decrease in performance of the $+F_0$-only methods. Obviously, search spaces can be reduced based on text-based methods (e.g., [25]), but even their sizes may still pose challenges. Second, both the Expert and Comb methods incorporated prosodic information, perhaps giving our approach a special advantage. However, this is not the case for the Festival method based results, which are, in fact, the most powerful.
6. References


