interspeech 2008 incorporating SST 08

Non-Segmental Duration Feature Extraction for Prosodic Classification

Amy Dashiell, Brian Hutchinson, Anna Margolis, Mari Ostendorf

Department of Electrical Engineering, University of Washington, Seattle, Washington

{amydash,brianhutchinson,amargoli,mo}@ee.washington.edu

Abstract

This paper presents a set of novel duration features for detecting pitch accent and phrase boundaries, which depend on articulatory timing rather than segmental duration information. The features are computed from the detected syllable nuclei and boundaries, using peaks and valleys in an energy contour but also leveraging information from a simple HMM phone manner class recognizer to increase recall. In experiments on the hand-segmented TIMIT corpus, we obtain greater than 90% F-measure for vowel detection. In prosody detection experiments on the BU Radio News corpus, comparing to a segmental feature baseline, we obtain similar performance for pitch accent detection and slightly worse boundary detection from the new features without the need for phonetic alignments.

Index Terms: prosody, prominence, pitch accent, boundary detection, duration features

10.21437/Interspeech.2008-336

1. Introduction

For speech transcription aimed at spoken document processing and dialog systems, it can be useful to recognize segmentation and emphasis patterns along with the words. One alternative is automatic recognition of prosodic events in speech. Recognition of a variety of prosodic events has been explored in many prior studies, typically combining acoustic and lexical/syntactic features; e.g. [1, 2, 3, 4]. While a range of acoustic cues are useful in prosody modeling, it is well known that duration cues are particularly important [5, 6].

Duration features that have been used include syllable or syllable nucleus durations in [4] and allophone duration and average duration (over a allophone window) in [1]. The features in [2], upon which our baseline segmental feature set is based, include various types of normalized word and phonetic segment duration, where start and end times of the units are found as a by product of the word decoding process or via forced alignment of the speech signal to a specified word sequence. At a minimum, normalization accounts for phone identity, but some systems account for speaker identity or speaking rate as well.

One practical problem with using segmental durations in the feature set, for developers using third-party (off-the-shelf) speech recognition software, is that commercial systems typically do not provide phonetic time alignments with the word recognition output, though word times may be available. In [7], a substitute set of features is proposed that does not require phone time alignments but instead uses word durations normalized by summed phoneme average duration statistics. While they report good results for accent and boundary detection, the approach obscures cues known to occur at the syllable level. (Accented words are longer primarily just in the accented syllable, and phrase final words are longer primarily in the rhyme of the word-final syllable.) In addition, there is evidence that the timing of articulatory gestures may be more relevant to detection of prosodic events than segmental duration. In [8], using magnetometer recordings of the speech articulators, it was observed that phrase-final lengthening is fundamentally an articulatory gesture phenomenon. In other words, the relative timing of the transition between the opening and closing gestures within the syllable helps to distinguish between syllables that are lengthened due to accent vs. phrase-final positioning. Figure 1 illustrates this, showing the differences in lengthening and peak timing for unmarked, accented and phrase-final syllables.

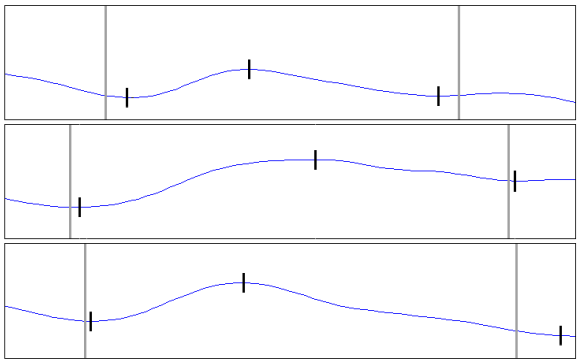

Figure 1: Example energy contours for the word "hall" in normal (top), pitch accented (middle), and phrase-final (bottom) contexts. Full bars mark hand labeled word boundaries, short bars mark the automatic placement of peaks and valleys.

The goal of this work is to develop duration feature extraction algorithms that more closely represent the relative timing of lengthening in different contexts, but also eliminate the need for word-based phonetic segment alignments. The approach approximates articulatory timing by identifying "peaks" and "valleys" in the energy of the speech signal. Peaks mark the location of highest energy within the syllable nucleus; valleys aim to identify syllable boundaries using acoustic cues.

The basic approach is to first identify peaks and valleys, as described in section 2 and assessed in section 3, then extract duration features from these time points as discussed in section 4. We compare the new features to segmental durations in pitch accent and boundary detection experiments in section 5. Key findings are highlighted in section 6.

2. Identification of Peaks and Valleys

2.1. Prior Work

Our first task of finding the "peaks," or highest energy points for each syllable, is effectively the task of vowel detection. Much work has been done on the problem of finding vowel locations. In many cases, the vowels are used as a proxy for speaking rate [9, 10, 11], but other uses include language identification [12].

Accepted after peer review of full paper

In [9, 12], the energy in the spectrogram is summed over selected subbands, and post-processed to reduce noise. The resulting peaks in these energy trajectories are considered to be vowels. In [10] subband cross-correlation is a central component to their vowel finding algorithm. It is important to note that these algorithms were developed for applications for which the exact placement of the vowel is not critical.

Finding "valleys" is closely related to syllable segmentation. The task of syllable segmentation has been studied for purposes of improving speech recognition accuracy and concatenative speech synthesis. In [13], four different algorithms are compared on the TIMIT corpus. In three of the algorithms, the syllable boundaries are minima of a filtered, rectified, and down-sampled time-domain signal. In the fourth algorithm, the short-term Fourier Transform is filtered, rectified, and mapped to 9 critical bands. The syllable boundaries are the union of the signal maxima from each band. A data-driven learning approach is used in [14], where a neural network is trained with spectral discontinuity features.

2.2. Peaks

We compare three different techniques for finding the peaks of syllable nuclei, including two previously described methods based on energy contours and a simple Hidden Markov model (HMM) approach. The results are combined using a decision tree classifier to improve over the individual methods.

The first peak-finding algorithm, referred to here as the **envelope** approach, is based closely on work by Pfau and Ruske [9]. A modified loudness curve is created by summing over specific energy bands and then smoothing with a Gaussian filter. The local maxima in the loudness curve are found and tested against two criteria. First, the zero-crossing rate must be below a certain threshold, and second, the peak magnitude must be greater than some percentage of the moving average. Maxima that meet these two criteria are considered peaks. The second method, developed by researchers at USC [10], uses a temporal and spectral correlation algorithm, and is referred to as the **USC** approach. The final **manner** approach uses manner-of-articulation HMMs, placing peaks in the center of vocalic segments in a Viterbi alignment. The manners used for this research include 3 vocalic classes (full vowel, reduced vowel, and syllabic consonant) and 5 non-vocalic classes (stop, nasal, fricative, approximant, and silence). We trained tri-manner left-to-right HMMs using the HTK toolkit [15]. Each HMM has three emitting states, with diagonal covariance Gaussian mixture distributions. The number of component Gaussians is tuned to a development set, never exceeding the number of phones in the given manner class. The 39-dimensional acoustic features consist of cepstral mean normalized mel-frequency cepstral coefficients plus delta and acceleration coefficients. Decoding uses a manner class bigram, with the grammar scale factor and phone insertion penalty tuned on a development set.

To exploit the differences in the errors made by the various peak methods, we merge the three sets of hypothesized peaks and keep or discard each merged peak using a decision tree classifier. The merging algorithm starts with all peaks detected by the envelope approach, adds peaks from the USC set for any segments that are not already associated with an envelope peak, and finally adds a peak at the center of any vocalic segment detected by the manner HMM that was not associated with a peak from one of the other algorithms. This approach to merging favors the energy-based peak algorithms, introducing manner peaks only when a false negative is suspected, which most often occurs on reduced syllables. For each peak in the merged set, fifteen features are extracted to be used as input to a decision tree that predicts which of the merged peaks should be discarded. The features include the type of peak, the manner and duration of the current, previous and following segments, and the number of hypothesized peaks according to each method. The peaks are labeled as "keep" for the one closest to the reference within a vocalic manner segment and "discard" for all others. Using the IND toolkit [16], a decision tree is learned using 10-fold cross-validation on the training data.

2.3. Valleys

Similarly to peaks, three different methods are employed to find valleys, which are designed to mark syllable boundaries. The **envelope** method uses the local minima in the modified loudness curve from the [9] algorithm. In a second approach, **spectral discontinuity** is computed by differentiating each energy band and then summing the derivatives at each time. The idea is that syllable onsets often occur at a sharp transition between sounds, resulting in a large spectral discontinuity value that is less likely within a syllable [14]. Finally, a third set of valleys are hypothesized using the **manner** alignments together with simple rules based on the maximum onset principle [17]: make the onset as large as possible while preserving the property that the segments in the onset increase in sonority. There are known exceptions to this simple rule that cannot be identified based on manner alone, so the rule-based syllabifications are imperfect but useful nonetheless.

We combine the valleys hypothesized by the various methods to have exactly one valley before and after each peak, with the exception that two valleys are used to mark edges of silent regions. Valleys at word boundaries (taken from the recognizer) and silent regions are first placed deterministically, and other hypothesized valleys within the given peak to peak region are discarded. Valleys in other regions are predicted from the set hypothesized by the different algorithms using a regression tree, for which we explored two alternative strategies, as follows.

The **per-valley** regression tree combination produces one feature vector per candidate valley. Training examples are labeled by the (signed) offset to the reference valley within their peak-to-peak region. The regression tree estimates this offset for each hypothesized valley in a region, and the valley with the smallest absolute offset is chosen. If there are no hypothesized valleys, then one is inserted in the middle of the region.

The **per-region** regression tree combination produces one feature vector per candidate peak-to-peak region. Features include the manner of the current, previous and following segments, the number of hypothesized valleys for each region, and the median valley location per valley-finding method. Training examples are labeled by the relative location of the reference valley within a region, in the range $[0, 1]$. The tree predicts this location, which may not coincide with an actual valley.

3. Peak/Valley Detection Experiments

While the ultimate goal for extracting peaks and valleys is to construct novel duration features, it is informative to assess the intermediate results of vowel detection and syllabification, both for choosing between algorithm alternatives and for understanding limiting factors in the approach.

The various peak/valley detection algorithms were evaluated on the TIMIT corpus, which contains read speech from 630 speakers. We divide the training set into a 3440 utter-

Table 1: *Peak Method Results.*

Method	Prec.	Recall	F-Measure
Envelope	92.6	87.4	90.0
USC	96.1	81.8	88.4
Manner	94.4	91.4	92.9
Combined	94.1	90.6	92.3

Table 2: *Valley Method Results.*

Method	Prec.	Recall	F-Measure	Avg Dist (ms)
Envelope	79.9	81.6	80.7	64
Spec. Disc.	63.1	85.4	72.6	78
Manner	67.6	85.6	75.5	69
Per-Valley	80.2	85.8	82.9	36
Per Region	80.5	85.4	82.9	38

ance training set and a 256 utterance development set, and report results on the standard test set. From the hand-marked phone level annotation, we automatically construct reference peaks and valleys as follows. Reference peaks are located at the energy maximum within the interior 50% of each vocalic segment, where the energy is computed using a five-frame triangle-window moving-average of the unsmoothed energy measurements produced by the envelope method. Reference valleys are located by finding syllable boundaries using the highest speaking rate syllabification generated using the tsylb2 program [18] and then placing a valley in the center of ambisyllabic segments or at the segment boundary for regular syllable boundaries.

Since the reference peaks are not associated with hand-marked times, we considered any peak within a labeled vowel region (with a ± 10 ms tolerance) to be a correctly identified peak. The results obtained for each of the three methods are presented in Table 1, together with the combined result. The combined method, relying largely on energy-based peaks, obtains a similar F-measure to that of the manner method while preserving the timing advantage of the energy peaks.

To evaluate the quality of a valley sequence, we not only measure the precision and recall, but we also compute the average distance from the reference valley for each correct detection. Specifically, we aligned the hypothesized valleys to the reference valleys using a modified Levenshtein distance with a fixed penalty (1) for inserting and deleting a valley and a "substitution penalty" for aligning two valleys equal to the distance (in seconds) between them. The results of the various valley methods are presented in Table 2. For computing precision and recall, we considered any valley aligned to a reference boundary to be correctly identified. The combined methods lead to the best performance, with the main impact being a reduction in the distance to the reference valley.

4. Non-Segmental Duration Features

Our original feature set is based on the prosodic features used for sentence segmentation in [2], and contains nearly 100 different features including various normalizations of pitch, energy, word duration, pause duration, time into current story, and segmental durations. Except for the segmental durations, which require phone alignments, the features can be computed from the word alignments. We refer to this non-segmental set as the "basic" set. The segment duration features include various normalizations of average and maximum vowel duration, last rhyme duration, last vowel duration, and vowel and rhyme durations of the primary stressed syllable, where the normalization factors are estimated on a separate speaker-independent broadcast news corpus (TDT4). These features were computed from phone alignments, using a stress-marked dictionary to determine primary stress. Given the peaks and valleys, we aimed to replace the segmental features with new features that do not depend on phone alignments.

Since peak/valley detection is independent of word recognition, the number of syllables defined by the peaks and valleys may not match up with the number of syllables in the dictionary entry for that word. However, because we are only computing features on the last and stressed syllables, we can select those syllables using simple rules. The last syllable is taken to be the region between the two valleys surrounding the last peak in the word. The stressed syllable is chosen via a simple left-to-right alignment of the hypothesized peaks to the syllables in the dictionary entry, allowing skips of reduced syllables in the dictionary when there are fewer hypothesized peaks than dictionary syllables. When there are multiple dictionary pronunciations, the first is used. Explicitly aligning to multiple pronunciations would improve performance, but we conjecture that the differences are minimal.

For each of these syllables three features are created: rhyme duration (the distance between a peak and its following valley), syllable duration (the distance between two adjacent valleys), and ratio (the syllable-to-rhyme duration ratio). The ratio feature is an indicator of the relative location of the peak in the syllable. (The average and maximum vowel durations in the segmental feature set are not replicated in our new feature set.) Two normalized versions of each feature are also created. The first roughly accounts for speaking rate. The syllable duration (or rhyme duration, or ratio) is divided by the average duration over a sliding window of 5 syllables centered on the current syllable. The second normalization accounts for inherent differences in phone durations. The total syllable (or rhyme) duration is divided by the sum of the average durations for the associated phones (from a pronunciation dictionary), using the same table as for segmental normalization.

5. Prosody Recognition Experiments

We perform our prosody recognition experiments using a section of the Boston University Radio News corpus [19] read by speaker "f2b." This professionally read corpus is annotated with phonetic alignments derived from the orthographic transcriptions, part-of-speech (POS) tags, and tone and phrase boundary tags based on the ToBI prosodic labeling conventions for American English [20]. The prosodic labels (done by human labelers) include 7 pitch accent categories and a phrase break index ranging from 1 to 6 for each word boundary. We map all pitch accents to a single "pitch accent" class and all phrase break indices above 4 to the boundary class. (This is similar to previous work on ToBI label recognition with this corpus, such as [1], though [4] defines the boundary class to include break indices 3 and above.) The training set consists of 16 broadcast stories, totaling 5087 annotated words; the test set consists of 4 lab stories, totaling 2112 annotated words. About half the words in the training set have pitch accents, while roughly 20% precede a boundary.

We use BoosTexter [21] as our classifier, following [22]. Experiments were performed with several different feature sets, with and without POS tags; results are summarized in Table 3.

Table 3: *Prosody Recognition Results.*

Results for Pitch Accents (F-Measure)		
Feature Set	w/o POS	w/ POS
Basic	85.6	86.0
Basic + Segmental	85.3	85.9
Basic + Per-Valley	85.8	85.1
Segmental Stress + Word-Dur	82.4	85.1
Per-Valley Stress + Word-Dur	83.1	84.7
Results for Boundaries (F-Measure)		
Feature Set	w/o POS	w/ POS
Basic	78.5	78.1
Basic + Segmental	80.1	81.2
Basic + Per-Valley	77.9	80.3
Segmental Str/Last Rhyme	69.5	80.0
Per-Valley Str/Last Rhyme + Ratios	63.5	73.1

Note that the "basic" features are only the non-segmental ones from [2] (pitch, energy, word duration, etc), while the "segmental" are phone-alignment-based duration features. Results on the development set were similar for the per-valley and per-region methods, so we include just the per-valley results. For both pitch accent and boundary classification, we compare the segmental duration set with the new feature set (both paired with the basic features). For the pitch accent case we compare the segmental stressed rhyme and vowel duration features against the new per-valley stressed rhyme, syllable and ratio features; we include word duration also because in preliminary experiments we observed that it is the most useful feature for predicting pitch accent.[1] For boundary detection, we compare the segmental stressed and last rhyme features against the new per-valley stressed and last rhyme features along with ratios. The number of BoosTexter training rounds was tuned to the development set; 400 was selected for the boundary task, 200 for emphasis. Subsequent analysis indicates that our boundary detection classifier may have been over-trained, and that a smaller number of rounds would have been preferable.

When combined with the basic and POS features, the two sets of duration features have similar results but neither is useful for emphasis detection. If either the POS or the basic features are omitted, the results are similar for both feature sets for emphasis, but there is a degradation associated with the non-segmental features for boundaries. Alone, both feature sets are informative, since the F-measure is much higher than prediction based on priors (roughly F=20%). Anecdotal inspection of errors suggests that the new non-segmental approach might be improved by using additional features from the energy contour.

6. Conclusion

This work created non-segmental duration features that achieve similar performance to their segmental counterparts for pitch accent detection, but slightly worse performance in boundary detection. If phone alignments are not available, these new features could be substituted for traditional duration features with little degradation in performance. Furthermore, the features appear to be robust to different data sets, in that the systems were trained on TIMIT but tested on the BU Radio News corpus. As a byproduct of this work, a vowel nucleus detector and a syllable boundary detector were created, both of which are competitive with current systems [9, 13]. An important next step is to assess the features on speaker-independent data from different genres.

[1] This may be since, as reported by others, it helps distinguish content words, which are more likely to be emphasized.

7. References

[1] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proc. ICASSP*, vol. 1, 2004, pp. 509–512.

[2] Y. Liu *et al.*, "Enriching speech recognition with sentence boundaries and disfluencies," *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1526–1540, 2006.

[3] A. Nenkova *et al.*, "To memorize or to predict: Prominence labeling in conversational speech," in *Proc. HLT-NAACL*, 2007, pp. 9–16.

[4] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. ASLP*, vol. 16, no. 1, pp. 216–228, 2008.

[5] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. ASLP*, vol. 2, no. 4, pp. 469–481, 1994.

[6] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, 2004, pp. 105–114.

[7] J. Buckow *et al.*, and robust features for prosodic classification," in *Proc. TSD*, 1999, pp. 193–198.

[8] D. Byrd, "Articulatory vowel lengthening and coordination at phrasal junctures," *Phonetica*, vol. 57, pp. 3–16, 2000.

[9] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. ICASSP*, vol. 2, 1998, pp. 945–948.

[10] S. Narayanan and D. Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *Proc. ICASSP*, 2005, pp. 413–416.

[11] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP*, vol. 2, 1996, pp. 1261–1264.

[12] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[13] R. Villing, T. Ward, and J. Timoney, "Performance limits for envelope based automatic syllable segmentation," in *Proc. ISSC*, 2006, pp. 521–526.

[14] S. Wu, M. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP*, vol. 2, 1997, p. 987.

[15] S. Young *et al.*, "HTK - Hidden Markov Model Toolkit," Cambridge University Engineering Dept. and Entropic Ltd., 2006.

[16] W. Buntine, "IND decision tree software," NASA Ames Research Center, 1995.

[17] D. Kahn, "Syllable-based generalizations in English phonology," Ph.D. dissertation, MIT, 1976.

[18] W. M. Fisher, "tsylb2-1.1 syllabification software," NIST, 1996.

[19] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," Boston University, Tech. Rep. ECS-95-001, 1995.

[20] K. Silverman *et al.*, "ToBI: A standard for labeling English prosody," in *Proc. ICSLP*, 1992, pp. 867–870.

[21] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 2, no. 39, pp. 135–168, 2000.

[22] D. L. Hillard, "Automatic sentence structure annotation for spoken language processing," Ph.D. dissertation, University of Washington Dept. of Electrical Engineering, 2008.