ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Word Prominence Detection using Robust yet Simple Prosodic Features

Taniya Mishra, Vivek Kumar Rangarajan Sridhar, Alistair Conkie

AT&T Labs Research, Florham Park, NJ, USA
{taniya,vkumar,adc}@research.att.com

10.21437/Interspeech.2012-408

Abstract

Automatic detection of word prominence can provide valuable information for downstream applications such as spoken language understanding. Prior work on automatic word prominence detection exploit a variety of lexical, syntactic, and prosodic features and model the task as a sequence labeling problem (independently or using context). While lexical and syntactic features are highly correlated with the notion of word prominence, the output of speech recognition is typically noisy and hence these features are less reliable than the acoustic-prosodic feature stream. In this work, we address the automatic detection of word prominence through novel prosodic features that capture the changes in F0 curve shape and magnitude in conjunction with duration and energy. We contrast the utility of these features with aggregate statistics of F0, duration and energy used in prior work. Our features are simple to compute yet robust to the inherent difficulties associated with identifying salient points (such as F0 peaks) in the F0 contour. Feature analysis demonstrates that these novel features are significantly more predictive than the standard aggregation-based prosodic features. Experimental results on a corpus of spontaneous speech indicate that prominence detection accuracy using only the new prosodic features is better than using both lexical and syntactic features.

Index Terms: Word prominence detection, prosodic features.

1. Introduction

Intonational prominence is an important aspect of spoken communication. Speakers use prominence to indicate the focus of an utterance, the introduction of new topics, the information status of a word (new or given), their emotion or attitude about the topic being discussed, or simply to draw the listener's attention. Identifying these discourse-salient elements is important for automatic spoken language understanding, but before that can be done, the prominence bearing units (syllables or words) need to be correctly identified. This makes automatic prominence detection a significant subtask of spoken language understanding.

Automatic prominence detection is a challenging problem. There is a fairly significant body of existing work on this topic. Researchers have investigated the use of several lexical, syntactic, and prosodic features for this task. In this work, we focus primarily on prosodic features as they are robust to speech recognition errors and hence offer more reliability in the prediction of surface level events for downstream applications.

The main prosodic cues correlated with prominence are fundamental frequency (F0), duration and energy. Prominence bearing words are generally marked by the presence of an F0 peak (or valley), longer duration, and increased energy. Most studies that have used prosodic features for prominence detection, however, parameterized F0, energy and duration using aggregate statistics such as mean, slope, variance, max and min [1, 2]. A fair amount of salient information is lost in the aggregation process, especially about the shape and amplitude of the F0 peak. The information loss is likely to negatively impact the accuracy of the prominence detection task, given the strong correlation between the F0 peak and perceived prominence [3].

However, automatic identification of a pitch peak/valley is non-trivial. The peak may occur in the voiceless regions of the F0 contour, or the peak may be flat, the valley may be shallow, the rises and falls may be ill-defined, or spurious peaks created by segmental perturbations may hide the true peak [4]. In this work, we have developed *easily computable* prosodic features that capture the changes in F0 curve magnitude and shape, along with changes in duration and energy, without requiring explicit identification of particularly salient points (such as F0 peaks, valleys, onsets, offsets, etc.) within the F0 contour. Hence, they are *robust* to the aforementioned problems often accompanying identification of salient points in the F0 contour.

Prior work on prominence detection has been performed at word-level, syllable-level, and syllable-nucleus-level. We chose words as the basic unit for prominence detection based on the results of a comparative study by Rosenberg and Hirschberg [5] that showed that prominence detection is most successful at the word level. Using words as units also allows us to compare the results of prominence detection using prosodic features with that obtained by using lexico-syntactic features.

2. Data and Features

2.1. Data

We used a subset of the Switchboard corpus that had manually-corrected word segmentations and hand-marked prominence labels [6]. It consists of about 67k word instances (excluding silences and noise). Roughly one-third of these words were marked as prominent. In a departure from the usual ToBI-based prominence marking, prominent syllables were marked with "*" indicating that the labeler perceived it as prominent, or with "*?" indicating uncertainty. Since we used words as the unit of analysis, we transduced the prominence markers to the word-level, and collapsed the prominence labels to simply 1 (prominent word) or 0 (non-prominent word). Word boundaries are also hand-labeled in this corpus.

This dataset has been previously used for word prominence detection using prosodic and lexico-syntactic features in [7], and those results serve as a benchmark for our evaluations.

2.2. Features

The word-level prosodic and lexico-syntactic features used in our prominence detection experiments are outlined below. The prosodic feature set contains two subsets of features, the new features that we developed for the word prominence detection task, and the set of standard features that are generally used for the same. The new prosodic features are indicated using bolding and an asterisk ("*").

Prosodic Features

We investigated the use of the following prosodic features for word-level prominence detection. These features were computed from pitch and energy contours extracted over 10 msec intervals for each utterance using ESPS Waves. Pitch halving and doubling were automatically cleaned up using an implementation of Bagshaw's 'defiltering' algorithm [8]. The F0 curve was smoothed over the unvoiced frames using weighted linear interpolation, where the weight vector was energy times voicing. We constructed these features from both raw and speaker normalized (z-score) pitch and energy contours.

(P1) * Area under the F0 curve (AFC): This feature is intended to capture the raised F0 and the increased duration that is often associated with prominent words. It is the integral of the smoothed F0 and duration within the interval of the word, as shown in the equation below.

$$AFC = \sum_{i \epsilon \text{interval}} (t_i \times F0_i)$$

(P2) * Energy-F0-Integral (EFI): Since word prominence is also often accompanied by increase in energy, we included it in the integral, as shown here (α is a scaling factor that was set to 0.10).

$$EFI = \sum_{i \epsilon \text{interval}} (t_i \times F0_i \times \alpha \cdot \text{RMS-energy}_i)$$

(P3) * Voiced-to-unvoiced ratio (VUR): This feature was developed to act as *measure of reliability*. AFC and EFI are calculated on the smoothed F0. However, smoothing the F0 contour by interpolation over the unvoiced segments of the word may create spurious peaks or valleys in the F0 contour [3], and the resulting AFC and EFI may not reflect the "true" shape of the contour. VUR informs the model how much the AFC and EFI features should be trusted. If the VUR is less than 0.5, then a majority of the segments in the word are unvoiced and most of F0 contour is obtained by smoothing, hence the AFC and EFI are less reliable than if the VUR was greater than 0.5.
(P4) * Average difference between low and high frequency components (DLH): A two component Gaussian mixture model was used to cluster the F0 values in the smoothed F0 curve into high and low frequency clusters and the difference in the means of the two clusters was computed.
(P5) * F0 curve shape (SHP): For each word, we used isotonic regression to estimate the likelihood of the F0 curve associated with the word being (i) a rising curve, (ii) a falling curve, (iii) a curve containing a peak, and (iv) a curve containing a valley. Isotonic regression measures departure from monotonicity and this algorithm identifies peaks and valleys as points in the curve where the greatest departures from monotonicity occur. A more detailed description of the algorithm can be found in Mishra, 2008 [3]. A significant aspect of this estimation algorithm is that it is robust to the presence of temporally short-lived segmental perturbations that may have higher F0 values than the true peak (valley) in the curve.

Since a word is perceived as prominent in relation to other words around it, for each word, we compute the same set of four likelihood values for the preceding and next word. This set of 12 values makes up the shape vector.
(P6) * F0 peak/valley amplitude and location (FAMP and FLOC): If a peak, as estimated by the aforementioned isotonic regression approach, is encountered in the F0 contour of a word, its *location* is computed as its relative distance from the beginning of the word, and its *amplitude* is computed as the distance from the mean of the GMM-based low frequency component in the word interval. If a valley is most likely, then location is computed similarly and the amplitude is the distance from the mean of the GMM-based high frequency component.
(P7) Duration of the word (STANDARD-DUR): The duration of the word in number of 10 msec frames.
(P8) Aggregate statistics (AGG-STATS): This includes the mean, median, max, min, and variance of F0 and energy computed per word.

Lexico-Syntactic features

To compare the discriminative power of prosodic features to that of lexical and syntactic features, we investigated the use of the following set of lexical and syntactic features in our prominence detection experiments. For each word, w_i, each of these features was computed over a three word window: w_{i-1}, w_i, and w_{i+1}.
(L1) Word identity (WI): In our final prominence detection model, we do not want to use word identity as a feature because this feature does not generalize to unseen data. We have however used word identity as a feature for experimental comparison.
(L2) Part-of-speech tags (POS): The POS tags were hand-marked using the Penn Treebank tagset.
(L3) Word type: Each word was classified either as a content word or as a function word by using the POS information.
(L4) Number of syllables in word: Three values of this feature were considered; 1-syllable, 2-syllables, and more-than-2-syllables. Syllabification was performed using the AT&T's Natural Voices TTS system.
(L5) Break tags (BT): Three categories were considered. No break, small breaks (corresponding to a comma in punctuated text) and big breaks (corresponding to a terminal punctuation, e.g., period, question mark or exclamation in punctuated text).

3. Prominence Detection Experiments

In our experiments, we model prominence detection as a binary classification task, i.e., we classify a given word as prominent (1) or non-prominent (0) based on different subsets of the input features. We used two ensemble methods, Random Forest and AdaBoost, to train these models. We used Random Forest and Adaboost as implemented in R and an R package, Rattle [9], both of which are freely available open source software.

Random forest is an ensemble of unpruned classification and regression trees in which randomness is injected in tree growing in two ways. Each tree is grown on a different random sampling — with replacement — of the data, and at each node, a random selection of features is used for splitting. AdaBoost is also an ensemble method that combines several "weak" learners (in our experiments, they are trees) to construct a strong classifier. However, in AdaBoost, the weak learners are trained by reweighting rather than resampling.

We used ensemble learners for building these models because they run efficiently on large unbalanced datasets with high dimensionality, require little parameter tuning, and can rank features in terms of their importance to the model. To demonstrate the benefits of the ensemble-learner-based models, we contrast their performance with simple CART models.

We built models using 8 different subsets of the features outlined in Section 2.2 for our experimental evaluation. Prediction results of models developed from each feature subset using three classification methods, namely, CART, AdaBoost and Random Forest (RF), are in Table 1. In this table, LXSYN refers to the lexical and syntactic feature set, PROS to the prosodic feature set, WI refers to the Word Identity feature, and BT to the Break Tag feature. For these experiments, we randomly selected 70% of the data for training, 15% for validation, and the remaining 15% for testing. The training and test partitions were not explicitly constructed to guarantee that no speaker appears in both partitions. However, this is a relatively minor issue due to the diverse speaker composition of Switchboard. The baseline accuracy of our test set, obtained by assigning the majority class (non-prominent) to all test examples is 69%. The same training, validation, and test sets were used for all experimental conditions. In the experiments in which Word Identity was used as a feature, we were unable to build a RF-based model because the R implementation of traditional Random Forest cannot handle variables that have more than 32 levels.

Feature Set	CART	AdaBoost	RF
All LXSYN	75.4%	77.5%	NA
Only WI	73.9%	74.6%	NA
LXSYN w/out WI	75.6%	76.8%	77.9%
LXSYN w/out BT	73.4%	74.2%	75.0%
All PROS	74.9%	75.8%	77.2%
Only AGG-STATS	74%	74.6%	74.1%
Only NEW PROS	74.6%	75.6%	77.2%
New PROS and all LXSYN (no WI)	78.2%	79.5%	81.5%

Table 1: *Results of word prominence detection models.*

Random Forest computes a measure of importance of each of the features that were used to build the model. RF has two main measures of feature importance: The first is the scaled average of the prediction accuracy of each variable, and the second is the total decrease in node impurities splitting on the feature over all trees using the Gini index [10]. The top-10 most important features according to the Gini index are shown in Table 2.

Feature	Gini ↓
Duration of the word (STANDARD-DUR)	1250.29
* Voiced-to-unvoiced ratio (VUR)	895.84
* F0 peak/valley location (FLOC)	806.24
* Energy-F0-Integral (EFI)	787.29
* Area under the F0 curve (AFC)	686.62
Std(energy)	649.79
* F0 peak/valley amplitude (FAMP)	599.94
Mean(F0)	571.14
* Znormed Energy-F0-Integral (EFI)	568.60
Std(duration)	545.84

Table 2: *Gini index based feature importance.*

3.1. Discussion of results

The experiment results shown in Table 1 and Table 2 have several implications. The benchmark for comparison are the results obtained by Sridhar et al. 2008 [7] in performing the same task of word prominence detection using the same dataset, so relevant comparisons are made while discussing the results.

It is obvious from the results in Table 1 that the features under investigation are predictive of prominence. The prediction accuracies range from 73.4% to 81.5%, which is significantly higher than the 69% baseline accuracy of the test set. We also see that using ensemble methods for model training produces significantly better models than simple CART trees.

Comparing rows 1 and 2 of Table 1, we see that word identity (WI in the table) is a fairly strong feature for word prominence detection. The best prominence detection accuracy using just the Word Identity feature is 74.6% compared to the 77.5% accuracy that is achieved when all the lexico-syntactic features are used. This result is in line with the results reported in Sridhar et al. Their best WI-based CRF model had an accuracy of 75.66% compared to the 76.04% accuracy that was obtained when all the lexico-syntactic features were used.

However, comparison of rows 2 and 3 of Table 1 shows that not much accuracy is lost when Word Identity is not in the model; the other features of the LXSYN set appear to have captured much of the same information. The implication of this result is as follows. Given that Word Identity is not a generalizable feature across multiple domains with varying vocabulary, the use of complimentary features that can offer similar discriminative power is extremely attractive. This is one point of departure of our work from [7]. In that work, the two main lexico-syntactic features were Word Identity and Accent Ratio. *Accent Ratio* is an estimate of the number of times a word was seen as prominent in training. Like Word Identity, Accent Ratio too does not generalize well to previously unseen data.

In rows 4 and 5, we compare the prominence detection accuracies obtained by using either lexico-syntactic features (row 4) or prosodic features (row 5) under comparable experimental conditions, i.e., without post-processing of the recognized text (we use reference text that assume 100% accuracy) to include break tag information. In this scenario, the words certainly would not have break tag (BT) markings so we have not used this feature when building the LXSYN model. Under these conditions, the prosodic-feature-based model is more accurate than the lexico-syntactic model. The best performance of the LXSYN model without BT is 75% while that of the PROS model is 77.2%. Given the importance of break tags (evidenced by the lowered prediction accuracy of LXSYN features without BT), it should certainly be used for prediction of word-level prominence in text-to-speech synthesis systems where the input text is punctuated; or in an offline SLU task where the input speech has been punctuated before prominence detection.

The implication of the above results are significant in comparison with previous work in [7]. In that work, even though the detection accuracy of the prosodic feature set is better than that of POS tags, it is significantly worse than using Word Identity or Accent Ratio alone. This is perhaps because the standard aggregate prosodic features (mean, median, variance, etc. of F0 and energy) in that work — like AGG-STATS here (row 6; table 1) — are not as discriminative as the lexico-syntactic features.

This leads us directly into a discussion of the prosodic features. The new prosodic features (row 7) appear to be more predictive than the aggregate prosodic features (row 6). It also appears that when combined together in a model (row 5), the AGG-STATS features do not add to the predictive power of the new prosodic features (see rows 5 and 7). This is further evidenced by the RF-estimated feature importance list presented in Table 2. Of the top-5 most important features according to the Gini Index, four are the new prosodic features that we developed for word-level prominence detection. Duration of the unit (word in this case) is the only standard prosodic feature that occurs in the top-5, although it is the most important feature. Six of the top-10 most important features according to the Gini Index are the new prosodic features. This indicates the new prosodic features developed in this work are more discriminative than the aggregate statistics of F0, duration and energy, which are commonly used for the prominence detection task.

Overall, the feature importance list in Table 2 shows this: (1) As a set, the new prosodic features are more discriminative than the commonly used aggregate statistics of F0, duration and energy. (2) Word duration, a standard feature, is more discriminative than each new feature individually. (3) The Energy-F0-Integral (EFI) is more discriminative than the area under the F0 curve (AFI), which makes sense since EFI robustly captures the change in F0, duration *and* energy accompanying prominence production, while AFI only captures F0 and duration changes. (4) The one new prosodic feature that has little to no discriminative power is DLH (the GMM-based estimate of average difference between low and high frequency components).

To estimate the best model that we can build for this dataset, we used all the new prosodic features (except DLH) and all the lexico-syntactic features (except word identity) to train three models (see last row of Table 1). The best performance accuracy of this model is 81.5%, which is a significant improvement.

4. Related Work

Automatic detection of word prominence can be abstracted into two main sub tasks. One, the choice of an appropriate feature set that is highly correlated with the notion of prominence or emphasis. Second, a suitable learning framework to either classify or recognize the labels by exploiting the chosen feature set. A majority of previous work [2] has typically used aggregate statistics of F0, energy, and duration such as *mean,max, standard deviation, slope of pitch*, etc. These features are simple to compute and offer reasonable robustness when normalized with respect to the speaker. However, they provide an extremely coarse representation of the macroscopic prosodic contour. Other work has investigated spectral features and aspects such as loudness [11] as markers of prominence. Fine grained representation of the prosodic features has been addressed in [12] through the use of quantized n-grams of the normalized contour. While this approach works well in practice, it is restricted to the window of the n-gram and leads to sparseness for larger windows. Parametric approaches such as TILT intonation model [13], pitch plateau representation [14] and the work presented in [15] also aim to model the shape of the contour using rise, fall and connections. However, these features are computationally more expensive compared to the features presented in this work.

From a modeling standpoint, word prominence detection has been addressed as a sequence of local classifications or through sequential models such as hidden Markov models or conditional random fields [16]. The approach presented in this work uses a local classification framework as it is highly suitable for incremental understanding and generation, i.e., one does not have to wait for the entire utterance to be decoded to find the optimal sequence. The use of ensemble methods for pitch accent (a potential marker of word prominence) detection has been addressed previously in [17]. But the experiments were performed on read speech in contrast to spontaneous speech in this work. Furthermore, the features were based on aggregate statistics of prosodic contour.

5. Conclusion

In this paper, we addressed the automatic detection of word prominence through the use of novel prosodic features that capture the changes in the F0 curve shape and magnitude along with duration and energy. These features are simple to compute yet robust to the inherent difficulties associated with identifying salient points in the F0 contour (such as F0 peaks, valleys, onsets and offsets). We conducted several experiments using a bench-marked dataset to test the predictive power of these new prosodic features. Feature analysis showed that these novel features are substantially more predictive than the standard aggregation-based prosodic features. Under comparable experimental conditions, i.e., realtime spoken language understanding of spontaneous speech input, we found that word prominence detection using our prosodic features was more accurate than that obtained by using lexical and syntactic features.

6. References

[1] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, no. 1-2, 1993.

[2] J. M. Brenier, D. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *In Proceedings of Eurospeech*, 2005.

[3] T. Mishra, "Decomposition of fundamental frequency contours in the general superpositional intonation model," Ph.D. dissertation, Oregon Health and Science University, Portland, Oregon, 2008.

[4] N. M. Veilleux and M. Ostendorf, "Prosody/parse scoring and its application in atis," in *Proceedings of HLT*, 1993, pp. 335–340.

[5] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *Proc. of NAACL-HLT*, 2009.

[6] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 119–121.

[7] V. K. Rangarajan Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proceedings of Speech Prosody*, 2008.

[8] P. Bagshaw, "Automatic prosodic analysis for computer aided pronunciation teaching," Ph.D. dissertation, University of Edinburgh, UK, 1994.

[9] G. Williams, "Rattle: A Data Mining GUI for R," *The R Journal*, vol. 1, no. 2, pp. 45–55, 2009.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, ser. Springer Series in Statistics. Springer, 2009.

[11] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *Journal of the Acoustical Society of America*, pp. 1038–1054, 2005.

[12] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 8, no. 4, pp. 797–811, 2008.

[13] P. Taylor, "The Tilt intonation model," in *Proc. of ICSLP*, 1998.

[14] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 690 –701, 2007.

[15] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in american english continuous speech," *International Journal of Speech Technology*, vol. 8, pp. 33–44, 2005.

[16] M. L. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proceedings of ACL*, 2004.

[17] X. Sun, "Pitch accent prediction using ensemble machine," in *Proceedings of ICSLP*, 2002, pp. 16–20.