Detecting Pitch Accent Using Pitch-corrected Energy-based Predictors

Andrew Rosenberg, Julia Hirschberg

Computer Science Department, Columbia University, USA
{amaxwell,julia}@cs.columbia.edu

Abstract

Previous work has shown that the energy components of frequency subbands with a variety of frequencies and bandwidths predict pitch accent with various degrees of accuracy, and produce correct predictions for distinct subsets of data points. In this paper, we describe a series of experiments exploring techniques to leverage the predictive power of these energy components by including pitch and duration features – other known correlates to pitch accent. We perform these experiments on Standard American English read, spontaneous and broadcast news speech, each corpus containing at least four speakers. Using an approach by which we correct energy-based predictions using pitch and duration information prior to using a majority voting classifier, we were able to detect pitch accent in read, spontaneous and broadcast news speech at 84.0%, 88.3% and 88.5% accuracy, respectively. Human performance at pitch accent detection is generally taken to be between 85% and 90%.

Index Terms: prosodic analysis, spectral emphasis

1. Introduction

Automatic detection of pitch accent is at least useful and at most critically important to a number of spoken language processing tasks. In English, accenting and deaccenting of a word provides information concerning its discourse status [1] and surrounding discourse structure [2]. The importance of a given word can be highlighted by either types of pitch accent or the relative height and placement of pitch peaks or intensity excursions. Additionally, pitch accent can provide information to listeners to perform syntactic and semantic disambiguation [3, 4]. Of interest to text-to-speech system developer is the potential of annotating a unit-selection corpus with prosodic information. This allows prosody to be included within the unit selection process to produce more natural, and less ambiguous synthesized speech, as well as offering users greater control of prosodic parameters. Currently, to include this functionality, unit selection corpora need to be manually annotated with prosodic information – a very time-consuming process.

The three major acoustic correlates to pitch accent are pitch excursions, increased intensity and prolonged vowel duration [5, 6]. In [7], we explored the discriminative properties of energy features extracted from a range of frequency subbands. We found that energy features extracted from different frequency subbands, even adjacent and overlapping ones, predict pitch accent with varying degrees of accuracy, and moreover produce correct predictions on different subsets of data points. It was determined that the frequency region between 2 and 20 bark was the most accurate, and robust predictor to pitch accent. Additionally, we found that at least one of the energy-based predictions was correct for upwards of 99% of all words. In this paper, we build upon these results, investigating techniques to leverage these predictions along with pitch and duration information to the ends of constructing a robust, high-accuracy pitch accent detector.

In section 4, we present a number of approaches to using filtered energy-based predictions for pitch accent detection. In particular, we present a technique to improve the accuracy of a majority voting classifier by ’correcting’ those contributions from energy-based classifiers that are believed to be erroneous. We use pitch-based features to classify an energy prediction as ‘correct’ or ‘incorrect’, inverting those predictions that are determined to be ‘incorrect’. This method is described in greater detail in section 4.2. We apply these techniques to three manually annotated corpora, containing read speech (BDC-R), spontaneous speech (BDC-S) and broadcast news (TDT).

Some particularly relevant previous contributions to the task of automatically detecting pitch accent are described in section 2. In section 3 we describe the material we use to evaluate our approach. We present results from our experiments in section 5, and conclude in section 6.

2. Previous Work

The task of automatically identifying pitch accent has received a significant amount of attention (e. g. [8, 9, 10, 11, 12, 13, 14, 15, 16, 17]). Wightman and Ostendorf [18] used decision trees with acoustic and lexical information to classify pitch accent, obtaining accuracy of approximately 84%. Ananthakrishnan and Narayanan [19] approached this problem using a sequential modelling approach. The application of Coupled HMMs was able to correctly classify approximately 80% of words correctly for the presence or absence of pitch accent when using syntactic and acoustic features. Sun [20] found that Bagging and Boosting ensemble learning approaches to significantly improve pitch accent prediction accuracy over a standard CART classifier. Using acoustic and lexical information, detection accuracy of approximately 87% was achieved on a corpus of broadcast news speech. Sluijter and van Heuven showed that accent in Dutch strongly correlates with the energy within a a particular frequency subband, specifically that greater than 500Hz, in both production [21] and perception experiments [22]. Heldner [23, 24] and Fant [25] extended the study of this “spectral emphasis” observation, by examining read Swedish speech. They found the relationship between the energy in one spectral region and the overall energy in the speech signal to be an excellent predictor of pitch accent.

3. Corpora

3.1. Boston Directions Corpus

The Boston Directions Corpus (BDC) was collected by Nakatani, Hirschberg and Grosz in order to study the relationship between intonation and discourse structure [26]. The corpus consists of spontaneous and read speech from four native speakers of Standard American English, three males and one female, all students at Harvard University. Each speaker was given written instructions and asked to perform a series of nine increasingly complicated direction giving tasks. This elicited spontaneous speech was subsequently transcribed manually, and speech errors were removed. At least two weeks later, the speakers returned to the lab and read the transcripts of their initial spontaneous monologues. The corpus was then ToBI [27] labeled and annotated for discourse structure. For the purposes of the experiments described in this paper we treat the spontaneous and read subcorpora as distinct data sets. The read subcorpus contains approximately 50 minutes of speech and 10818 words. The spontanous subcorpus contains approximately 60 minutes of speech over 11627 words. We use the hand-segmented word boundaries from the ToBI orthographic tier during the extraction of acoustic features, and assume these to be available for both the training and testing sets. We use the ToBI tones tier to provide ground-truth pitch accent labels for training and evaluation. We make only a binary distinction between accented and non-accented words; in this work, we do not attempt to distinguish pitch accent type.

3.2. TDT4

The TDT-4 corpus [28] was constructed by the LDC for the Topic Detection and Tracking shared task, and was provided for use in the DARPA GALE project. As part of the SRI NIGHTENGALE team, Columbia University was provided with automatic speech recognition (ASR) transcriptions of the corpus by SRI [29] and hypothesized speaker diarization results by ICSI Berkeley [30]. The TDT-4 corpus as a whole comprises material from English, Mandarin and Arabic broadcast news (BN) sources aired between October 1, 2000 and January 2, 2001. However, for the experiments presented in this paper, we had one 30-minute broadcast, 20010131_1830_1900_ABC_WNT, annotated for pitch accent. The annotation was performed by a single experienced ToBI labeler and reviewed by one of the authors. The annotator was asked to annotate the ASR transcript with pitch accent labels – since ASR hypothesized word boundaries may not align with those perceived by a human listener, the annotator was asked to mark an ASR hypothesized word as containing a pitch accent if he believed any syllable within the ASR word to contain the realization of a pitch accent. After omitting regions of ASR error, silence and music, the TDT4 material for use contained approximately 20 minutes of annotated speech and 3326 hypothesized words. Note, we use the ASR hypotheses only for word boundaries not for lexical content. The output of an automatic speaker diarization system identified 25 speakers within this show. These hypothesized identities are used to normalize acoustic information to account for speaker differences.

4. Methods

We explored a number of techniques of combining results from the filtered energy experiments with pitch and duration features in order to create a robust pitch accent detection module. In order to eliminate any influence of learning algorithm, every experiment was performed using weka's [31] J48 algorithm, a java implementation of Quinlan's C4.5 algorithm [32]. In order to isolate the learning architecture from the features used, we extract the same acoustic features for each classification experiment.

Pitch and Duration Features

We compute, for each word, the minimum, maximum, mean, root mean squared and standard deviation of pitch (f0) values extracted using Praat's [33] Get Pitch (ac)... function. We also computed each of these features based on speaker normalized pitch values. This normalization was performed using z-score normalization. For the BDC corpus, the true speaker identifies (four male, one female) are known. However, the speaker normalization for the TDT corpus does not use any manual annotation. Instead, we use the hypotheses of a automatic speaker diarization module to determine speaker identity. We included in the feature set, the above features calculated over the first order differences (Δ f0) of both the raw and speaker normalized pitch tracks.

Additionally, we used nine contextual windows to account for local context. These contextual windows were constructed using each combination of two, one or zero previous words and two, one or zero words following the given data point. Based on the pitch content of these regions we performed z-score and range normalization on the maximum and mean raw and speaker normalized f0 of the current word.

We extracted three duration features: the duration of the current word in seconds, the duration of the pause between the current and following word, and the duration of the pause between the current and previous word.

Energy Features

We extracted energy information from 210 distinct frequency bands. These frequency bands were constructed by varying the minimum frequency from 0 bark to 19 bark, and the maximum frequency from 1 bark to 20 bark. 20 bark is the maximum frequency in all of our corpora (see section 3) due to Nyquist rates of 8kHz.

For each word, we extracted the maximum, minimum, mean, root-mean-squared and standard deviation of energy. Additionally, we used the same nine contextual windows to account for local pitch content to normalize out local context from the energy information. Based on the content of these nine regions we performed z-score and range normalization on the maximum and mean energy of the current word.

4.1. Simple decision trees

In order, to have a point of comparison for our experiments with filtered energy features, we first performed pitch accent classification using feature vectors containing the pitch, duration and unfiltered energy features.

In [7], based on experiments with the BDC-read corpus, it was hypothesized that the frequency region between 2 and 20 bark contains energy information that would be the most robustly discriminative of pitch accent. To evaluate this claim, we ran classification experiments on all three corpora with feature vectors containing energy features drawn from the 2-20 bark frequency subband along with pitch and duration features.

4.2. Voting classifiers

Using an ensemble of classifiers, each trained using only energy features extracted from a single frequency subband, we constructed a simple majority voting classifier. For each data point, 210 predictions were obtained – one from each filtered energy-based classifier. The ultimate prediction for each data point was the class ('accented' or 'non-accented') predicted by at least 106 energy-based classifiers. In the case of a tie, the data point was assigned to the 'non-accented' class – the majority class in all corpora.

We also evaluated the performance of a number of variants of a weighted majority voting classifier. First, we weighted the predictions by the J48 confidence scores. Second, we weighted each prediction by the cross-validation accuracy of the classifier which generated it. Third, we weighted the predictions by the product of the J48 confidence scores and this estimated expected accuracy.

We observed that on all corpora, the oracular coverage of the 210 predictors was over 99%. That is, at least one energy-based classifier produced a correct prediction for nearly every word in every corpora. We performed two experiments examining ways of using pitch and duration information to determine which predictors will be correct for a given word.

In the first experiment, we constructed our feature vector using the pitch and duration features along with the 210 raw predictions from the filtered energy-based classifiers. When evaluating this type of classifier in a cross-validation setting, particular attention was paid to guarantee that none of the elements of the testing set were used in constructing the predictions included in the training set feature vector. To that end, for each training and testing set, an additional ten-fold cross validation scenario was run over the training set in order to produce predictions for use in the training feature vector. The testing set predictions were based on energy-based classifiers trained on the full training set.

The expectation in constructing this type of classifier is that rules would automatically be learned that would either associate predictions from frequency bands or associate pitch features that might distinguish when one frequency band might be more predictive than another. In figure 1 we can observe and instance of the former relationship. The behavior represented by this clipping of the decision tree says that for a given word, following some number of previous decision, if the speaker normalized mean pitch is below 0.6, then predict deaccented. If this pitch value is greater than or equal to 0.6, then trust the prediction made by the energy classifier trained on energy information within the frequency band between 8 and 16 bark. One possible explanation behind this type of decision is that this particular energy-based classifier is fairly accurate in a specific pitch environment, but fairly inaccurate in others. This type of branch inspired the next type of classification scheme, in which we make explicit the use of pitch-based features to correct energy-based predictions.

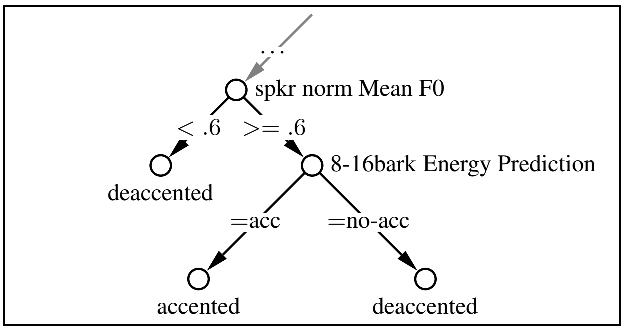

Figure 1: Detail view of single pitch-based classifier

In our final classifier design, we make the relationship between pitch and duration information and filtered energy based predictions explicit. For each frequency band, we build a pitch and duration-based classifier that predicts when the energy-based prediction from the given frequency band will be correct, and when it will be incorrect.

Again, when performing the ten-fold cross-validation on this two stage classifier, we pay particular attention to making sure that no data point in the test set is ever used in producing a training set prediction.

For each training set, we use ten-fold cross-validation to generate filtered energy-based pitch accent predictions for each frequency region. We, then, for each energy-based classifier, train a second classifier using pitch and duration features that classifies each training-set energy prediction as either 'correct' or 'incorrect'. Predictions that are classified as 'incorrect' are inverted. Thus, a 'accent' prediction classified as 'incorrect' becomes 'non-accented' and vice versa. Since, this correction is performed independently for each filtered energy-based classifier, we are left with 210 'corrected' pitch accent predictions. We then combine these into a final prediction using a majority voting scheme.

5. Results and Discussion

	BDC-R	BDC-S	TDT
Pitch/Dur Corrected Voting	84.0%	88.3%	88.5%
Pitch/Dur + Predictions	78.8%	77.5%	80.3%
Majority Voting	81.8%	81.8%	83.7%
'Best' Band Energy	80.0%	79.0%	81.1%
No Filtering	79.8%	79.1%	81.1%

Table 1: Pitch Accent Classification Accuracy

Our baseline experiment ('No Filtering'), which uses pitch, duration and unfiltered energy features to train a standard decision tree, yields the lowest accuracy on all corpora. Replacing the unfiltered energy features with corresponding energy features extracted from the frequency band between 2 and 20 bark ("Best' Bark Energy') does not yield significantly different results on any corpus. The hypothesis that the band between 2 and 20 bark would yield the most robust and discriminative energy features was based on experiments on the BDC-read corpus. On this corpus, we observe a statistically insignificant gain in accuracy of 0.02%. This band does not improve the accuracy on either other corpora – even insignificantly reducing it on BDC-spon. While the energy features extracted from the frequency region between 2 and 20 bark are able to predict pitch accent significantly better than unfiltered energy features, when combined with pitch and duration information, the impact of this improvement is severely diminished.

Based on the 210 predictions per data point using exclusively those energy features extracted from each frequency subband ('Majority Voting'), a simple majority voting classifier achieves classification accuracy that is significantly better than the baseline experiment on the TDT and BDC-spon corpora. Weighted voting classifiers, where each prediction is weighted by either J48 confidence score, cross validation accuracy, or the product of the two, do not yield significantly different results from the majority voting classifier.

When we included the 210 energy-based predictions into a feature vector ('Pitch/Dur + Predictions')along with the pitch and duration features, the classification accuracy was reduced below that of the majority voting classifier. We expected the decision tree to learn associations between pitch features and energy predictions, or to identify mutually reinforcing sets of predictions. However, even the baseline classifier outperforms this approach.

The two-stage classification technique ('Pitch/Dur Corrected Voting'), where pitch information is used to correct energy-based predictions before voting, demonstrated the best classification results on all corpora. On the BDC-spontaneous and TDT corpora the accuracy was 88.3% and 88.5% respectively. The human agreement on pitch accent identification is generally taken to be somewhere between 85% and 90%, depending on genre, recording conditions and particular labelers [18, 27]. These results represent a significant improvement over the baseline classifier, and approach human levels of competence. The fact that the accuracy on the TDT corpus is not significantly different from that obtained on the BDC material indicates that the technique is relatively indifferent to the fine grained accuracy of word boundary placement. Recall, the BDC corpus word boundaries were manually defined, the TDT word boundaries are a result of ASR output. While this technique produces the highest accuracy predictions on BDC-read (84.0%), the improvement over the baseline classifier is much more modest than that achieved on the other two corpora. It is possible that non-professional speakers produce read speech without pitch and duration information that can be successfully used by this classification technique.

6. Conclusion

We have presented a number of experiments on the use of filtered energy based predictors to accurately detect pitch accent. In particular, we described a two-stage classification technique which predicts pitch accent at rates close to human performance. This technique proceeds as follows. First, energy-based features extracted from 210 frequency subbands are used to generate a set of predictions for each data point. Pitch and duration features are then used to classify each prediction for each data point as correct or incorrect. Predictions labeled as incorrect are inverted; predictions of 'accent' were changed to 'no accent' and vice versa. Finally, a majority voting classifier was used to combine these 210 corrected predictions. On a corpus of read speech (BDC-read), this technique yielded accuracy of 84.0%. On spontaneous speech (BDC-spontaneous), the accuracy was 88.3%, and on a corpus of broadcast news from multiple speakers with ASR-generated word boundaries, the technique achieved accuracy of 88.5%, approaching human performance on a similar task. This high accuracy performance on disparate corpora demonstrates that this technique is robust to genre, speaker and recording condition differences, as well as noise in word boundary locations. We plan, however, to investigate why this technique yielded less improvement over baseline on non-professional read speech, than BN or spontaneous speech. This work has shown the success of applying ensemble-based techniques to the task of detecting pitch accent – we intend to study these applications more thoroughly. One drawback of the technique presented in this paper.is that it is very resource consuming to train and test. While there are many opportunities for parallelization, each data point requires 420 classifications in order for pitch accent to be detected. While previous work has determined that energy information drawn from individual frequency regions is largely non-redundant, we plan on running a combinatorial analysis to identify redundant sets of frequency regions.

7. Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

8. References

[1] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational Lunguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[2] J. Hirschberg and J. Pierrehumbert, "The intonational structure of discourse," in *Proc. of 24th Annual Meetinc og the Assoc. for Computational Linguistics*, 1986, pp. 136–144.

[3] P. J. Prince, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *JASA*, vol. 90, no. 6, pp. 2956–2970, 1991.

[4] J. Bos, A. Batliner, and R. Kompe, "On the use of prosody for semantic disambiguation in verbmobil," in *VERBMOBIL memo*, 1995, pp. 82–95.

[5] D. L. Bollinger, "A theory of pitch acent in english," *Word*, vol. 14, pp. 109–149, 1958.

[6] M. Beckman, *Stress and non-Stress*. Foris Publications, Dordrect, Holland, 1986.

[7] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read english speech," in *Proc. INTERSPEECH*, 2006.

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

[9] K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *ICSA International Conference on Speech Prosody*, 2004, pp. 509–512.

[10] A. Conkie, G. Riccardi, and R. C. Rose, "Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events," in *EUROSPEECH'99*, 1999, pp. 523–526.

[11] R. Delmonte, "Slim prosodic automatic tools for self-learning instruction," *Speech Communication*, vol. 30, pp. 145–166, 2000.

[12] A. Eriksson, G. C. Thunberg, and H. Traunmüller, "Syllable prominence: A matter of vocal effort, phonenetic distinctness and top-down processing," in *EUROSPEECH'01*, 2001, pp. 399–402.

[13] R. Kompe, "Prosody in speech understanding systems," *Lecture Notes in Artificial Intelligence*, vol. 1307, pp. 1–357, 1997.

[14] Y. Ren, S.-S. Kim, M. Hasegawa-Johnson, and J. Cole, "Speaker-independent automatic detection of pitch accent," in *ICSA International Conference on Speech Prosody*, 2004, pp. 521–524.

[15] A. M. C. Sluijter and V. J. van Heuven, "Acousic correlates of linguistic stress and accent in dutch and american english," in *Proc. ISCLP96*, 1996, pp. 630–633.

[16] F. Tamburini, "Automatic prominence identification and prosodic typology," in *Proc. InterSpeech 2005*, 2005, pp. 1813–1816.

[17] A. Waibel, *Prosody and Speech Recognition*. London: Pitman, 1988.

[18] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[19] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *Proc. ICASSP*, 2005.

[20] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proc. ICSLP*, 2002.

[21] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *JASA*, vol. 100, no. 4, pp. 2471–2485, 1996.

[22] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *JASA*, vol. 101, no. 1, pp. 503–513, 1997.

[23] M. Heldner, E. Strangert, and T. Deschamps, "A focus detector using overall intensity and high frequency emphasis," in *Proc. of ICPhS-99*, 1999, pp. 1491–1494.

[24] M. Heldner, "Spectral emphasis as an additional source of information in accent detection," in *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 57–60.

[25] G. Fant, A. Kruckenberg, and J. Liljencrants, "Acoustic-phonetic analysis of prominence in swedish," in *Intonation, Analysis, Modelling and Technology*, A. Botinis, Ed. Kluwer, 2000, pp. 55–86.

[26] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *Working Notes of AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation*, 1995.

[27] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *Proc. of the 1992 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[28] S. Strassel and M. Glenn, "Creating the annotated tdt-4 y2003 evaluation corpus," http://www.nist.gov/speech/tests/tdt/tdt2003/papers/ldc.ppt, 2003.

[29] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent innovations in speech-to-text transcription at sri-icsi-uw." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1729–1744, 2006.

[30] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *RT-04F Workshop*, November 2004.

[31] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," in *ICONIP/ANZIIS/ANNES*, 1999, pp. 192–196.

[32] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[33] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5(9-10), pp. 341–345, 2001.