On the Correlation between Energy and Pitch Accent in Read English Speech

Andrew Rosenberg, Julia Hirschberg

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

Abstract

In this paper, we describe a set of experiments that examine the correlation between energy and pitch accent. We tested the discriminative power of the energy component of frequency subbands with a variety of frequencies and bandwidths on read speech spoken by four native speakers of Standard American English, using an analysis by classification approach. We found that the frequency region most robust to speaker differences is between 2 and 20 bark. Across all speakers, using only energy features we were able to predict pitch accent in read speech with accuracy of 81.9%.
Index Terms: prosodic analysis, spectral emphasis

10.21437/Interspeech.2006-107

1. Introduction

In English speech, accenting particular words within an utterance serves a number of purposes. Accenting or deaccenting a word can provide information concerning its discourse status [12] and discourse structure [16]. An accented word's "importance" in a discourse may indeed be related to the type of accent with which it is produced or the peak height or amplitude of the accent. Pitch accent may also be employed by listeners to perform syntactic and semantic disambiguation [22, 5]. Automatic detection of pitch accent is therefore critically important to many natural language understanding tasks. Of particular interest to builders of text-to-speech systems is the possibility of automatically annotating large unit-selection corpora for prosodic information, so that prosody can be included in the unit search process to produce more natural synthetic speech and to permit users to specify prosodic variation. Currently such corpora must be manually labeled, a very time-consuming process.

Considerable attention has been given to the automatic identification of pitch accents within an utterance. It has been shown in a number of studies that features based on the pitch (f0), intensity, and duration of a word or its component syllables can be used in concert to achieve accent prediction accuracy between 80% and 85%. While this automatic prediction task has been studied extensively, there is no consensus concerning the best way to leverage highly predictive features from the three information streams.

In this paper, we examine the role of energy in the speech signal as an indicator of pitch accent in Standard American English (SAE). It has long been believed [4] that amplitude is a significant indicator of prosody in general and pitch accent in particular for accent in SAE [2]. Additionally, Sluijter and van Heuven [26] showed that the energy component of a high frequency subband – greater than 500Hz – highly correlates with stress in Dutch speech. To further investigate this correlation for SAE, we construct simple (binary) decision-tree pitch accent classifiers using only energy features to identify those energy features that correlate most strongly with pitch accent. The experiments presented in this paper elaborate upon [26]'s findings by closely examining the correlation between pitch accent and the energy components of a large number of frequency subbands.

In Section 2, we present previous research on the automatic detection of prosodic events, focusing on studies examining the correlation between energy and pitch accent. We describe the data used in our experiments in Section 3. Our experimental method is presented in Section 4. In Section 5, we report on our results. Finally, in Section 6, we conclude and describe some future directions of this research.

2. Previous Work

Much research attention has been paid to the task of identifying intonationally prominent or **accented** words within an utterance (e. g. [1, 6, 7, 8, 9, 17, 23, 25, 29, 30, 32, 33]). While there is consensus that the energy of a word or syllable correlates with pitch accent, how to use the energy information in the speech signal to predict accent has not yet been determined. Sluijter and van Heuven [26] have shown that accent strongly correlates with the energy within a particular frequency subband, namely, that greater than 500Hz in Dutch, using both production [26] and perception [27] experiments. Heldner [13, 14] and Fant [11] examined the role of this "spectral emphasis" in read Swedish speech, finding that the relationship between the energy in a particular spectral region and the overall energy of the signal was an excellent predictor of pitch accent. For SAE, Tamburini [28, 29], reported that the energy components of the 500Hz to 2kHz frequency band were more predictive of prominence than those from either 0 to 500Hz or above 2kHz. Also, Tepperman [30] used the RMS energy extracted from between 60 and 400Hz as a feature in his syllabic stress detection system on non-native British English speech. This research suggests that energy extracted from specific frequency regions rather than the entire spectrum is helpful in the automatic prediction of pitch accent in English. The work presented in this paper examines the energy component of a large set of frequency bands to determine which are most predictive of pitch accent in read SAE.

3. The Corpus

For this work we used data from the Boston Directions Corpus (BDC), collected by Nakatani, Grosz, and Hirschberg for a study of the relationship between intonation and discourse structure [15]. This 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 increas-

ingly complicated direction-giving tasks. Their audience was a confederate, who was to trace the routes given on a map, as the directions were given. This elicited spontaneous speech was subsequently transcribed, and speech errors removed. At least two weeks later, the speakers returned to the lab and read the transcripts. The corpus was then ToBI [24] labeled and also labeled for discourse structure.

The material used for our current study consists only of the read speech from this corpus. This subcorpus contains 50 minutes of speech and 10825 words. We employ the hand-segmented word boundaries from the ToBI orthographic tier during the extraction of energy features, and we assume that word boundaries are available in both training and test sets. We use the ToBI tonal tier to provide ground truth pitch accent labels for the training and testing of our classifiers. However, we make only a binary distinction between accented and non-accented words; we do not attempt to classify pitch accent type.

4. Method

To examine the correlation between energy and pitch accent, we have taken an analysis by classification approach. We constructed a feature vector for each manually-segmented word whose elements contained only features derived from the energy of the speech signal. Using the pitch accent annotation from the manual ToBI labeling, we assigned a binary class to each feature vector indicating whether the word is uttered with a pitch accent or not. Using this labeled data and ten-fold cross validation, we ran classification experiments to determine how predictive of pitch accent the energy components of various frequency subbands are. We used the *weka* machine learning environment's [34] C4.5 implementation, J48, a decision-tree algorithm, for classification.

The features we examined were computed from the energy component of a variety of frequency subbands. These subbands were derived from the Bark scale, using a Bark-to-Hertz transformation function of $hertz = 600 * \sinh(bark/6)$ [10]. We varied the lowest frequency of the subbands from the bark edges 0 to 19 and varied the bandwidth from 1 to 20 bark. The maximum frequency of any subband was 20 bark due to the 8kHz Nyquist rate of the BDC speech material. These combinations yielded 210 frequency subbands from which we extracted energy features for analysis by classification. We performed the filtering and energy extraction using the Praat speech analysis tools [3].

Our energy features included the minimum, maximum, mean, standard deviation, and root mean squared (RMS) of the energy, as well as features designed to capture the dynamics of the energy within the word. These features included the z-score of the maximum energy in the context of the current word, the mean slope, and a four-way classification describing the shape of the energy contour over the word (rising, falling, peak or valley).

Whether a word is perceived as accented or not is determined by its acoustic properties relative to its surrounding intonational context [18]. Therefore we included in the feature vector five normalized energy features based on the surrounding region. We varied the size of this contextual window in six different ways: 1) two previous and two following words, 2) one previous and one following word, 3) one previous word, 4) two previous words, 5) one following word, 6) two previous words and one following word. The energy features calculated over these regions included:

- The difference between maximum energy in the current word and the mean energy in the region, normalized by the standard deviation of the energy in the contextual window.
- The difference between mean energy in the current word and the mean energy in the region, normalized by the standard deviation of the energy in the contextual window.
- The difference between maximum energy in the current word and the maximum energy in the region, normalized by the standard deviation of the energy in the contextual window.
- The maximum energy in the current word normalized by the energy range realized in the contextual window.
- The mean energy in the current word normalized by the energy range realized in the contextual window.

We follow the American School of intonational description for SAE (e.g. [21]) in assuming that pitch accents, while interpreted as a property of the word, are aligned with the lexically stressed syllable of that word. Therefore, the detection of pitch accents in SAE may profit from information found at the syllable or syllable nucleus level. To that end, we automatically determined syllable boundaries, as well as start and end times of syllable nuclei using algorithms based on [19] and [20], respectively. In order to identify the most predictive region of analysis within a word we ran the classification experiments under four different configurations: using energy information extracted from 1) the entire word, 2) only the component syllable nuclei, 3) the longest syllable in the word and 4) the longest syllable nucleus in the word. We chose to include the longest syllable and syllable nucleus with the two larger regions due to earlier experimental results that indicate that pitch accent correlates with a lengthening of the accented vowel (e. g. [33]) and that the canonically stressed syllable of a polysyllabic word tends to be the longest (e. g. [31]).

The BDC has not been annotated for syllable or phone identities and boundaries. We therefore cannot provide precise error rates for these automatic segmentation approaches. However, we were able to compare the automatically derived syllable counts to the canonical pronunciation forms of the ToBI orthographic tier. The syllable boundary detector had an insertion rate of 20% and a deletion rate of 36%. The nuclei detector had an insertion rate of 14% and a deletion rate of 39%. As the actual pronunciations may differ significantly from the canonical forms, and pairs of deletion/insertion errors may be attributable to alignment problems, we make no claims as to the veracity of these error rates; they should be taken merely as estimates of the true accuracy of the automatic syllable segmentation systems.

It has been proposed that the extraction of energy from frequency subbands is helpful because it isolates the formants of the vocal portion of the speech signal from the fundamental frequency [26, 13]. In order to evaluate this claim on our data, we used Praat's automatic formant tracking algorithm to extract the frequency and bandwidth of the first and second formants of the vocal portions of the speech signal. We chose to examine only the first two formants because performance of the formant tracker degraded significantly at greater numbers. We then extracted energy components of the formants for each frame of the speech. We constructed the feature vector described above based on the energy extracted from the automatically determined formant bandwidths as opposed to a static frequency range.

In total, we explored eight experimental configurations: We constructed 210 classifiers, one for each frequency subband, extracting energy features from either 1) the whole word, 2) only syllable nuclei, 3) only the longest syllable, and 4) only the longest

syllable nucleus. For each of these 4 configurations we constructed speaker-dependent classifiers (one for each speaker) and a speaker-independent classifier (using data from all 4 speakers).

5. Results and Discussion

The classification accuracy produced by our machine learning experiments indicates significant differences in the discriminative power of energy information extracted from distinct frequency subbands.[1] Across all experimental scenarios – extracting energy from four different regions within a word, and looking at speakers individually or all together – the mean relative improvement of the most predictive subband over the least predictive was 14.8%. As an example, Figure 1 shows the classification accuracies on all speaker data using energy information extracted from the whole word with a frequency bandwidth of 1 bark.

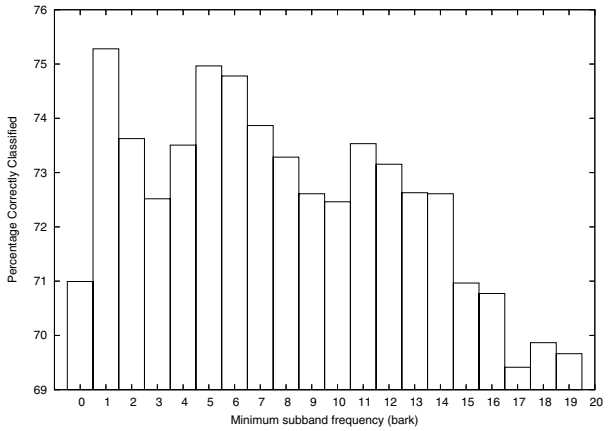

Figure 1: Pitch accent prediction accuracies with bandwidth of 1 bark

While our experiments confirm the claim that accent is realized through increased energy in a particular frequency subband, the classification results do not support previous findings as to the regions in which such useful information is to be found. Neither in speaker-independent nor in speaker-dependent experiments, did we find the most predictive band to be 500-2000Hz, as found in [28, 29], or >500Hz as suggested by Sluijter and van Heuven [26] or 60-400Hz as used in [30]. However, there **are** experimental configurations in which the classifications based on energy contributions from these regions are not significantly worse than the most predictive band. For our data, the frequency range that yields the most predictive features is between 3 and 18 bark (312Hz to 6000Hz) with energy information drawn from the entire word. This band correctly classifies 76% of all words on average, over 10-fold cross-validation, compared to a majority class baseline of 42.4%. The precision and recall for detecting accented words is 71.6% and 73.4%, respectively. The most predictive features used in this classification were the normalized maximum energy of the word relative to the maximum and mean energy in three contextual windows: 1) 1 previous and 1 following word, 2) 2 previous and 1 following word, and 3) 2 previous and 2 following words. However, this subband generates results significantly worse than the best performing subband for one of the four speakers – regardless of the region of analysis within each word. Interestingly enough this speaker is one of the three male speakers, not the female speaker. The band from 2 to 20 bark (203Hz to 8kHz)[2], while not being the most predictive region in any experimental configuration, is only significantly worse than the best performing band in one[3][4]. With energy drawn from the entire word, the subband between 2 and 20 bark correctly classifies 75.5% of all words. The precision and recall for detecting accented words is 70.5% and 72.5%. The most predictive features in this classification are identical to those used for the 3 to 18 bark band. Due to both the predictive power of this energy component within this frequency subband and its robustness to a variety of speakers and types of analysis, we believe this to be the best region from which to extract energy information for the prediction of pitch accent, based on our data. The Nyquist rate of our corpus is between 19 and 20 bark, so it is impossible to tell whether the band is more accurately described as "all frequencies above 2 bark" or strictly between 2 and 20 bark. Classifiers trained on the first and second formant information, analyzed together or separately, perform significantly worse than those based on the energy component of the subband between 2 and 20 bark; on average they yield 6.4% relative accuracy reduction. It is possible that formants higher than the second contain discriminative energy information as well. Additionally, since the formant tracking algorithm is errorful particularly for vowels in which the first and second formants tend to overlap, it is possible that these errors limit the usefulness of these features in our study.

After observing that distinct subbands predicted pitch accent with varying accuracies, we analyzed the classification results to determine the degree to which the correct classifications overlapped. We would expect a high degree of overlap if our data are such that there are distinct sets of words that are harder or easier to classify using energy features. We found however there is a relatively small intersection of correct predictions, even between overlapping or adjacent subbands. Moreover, 10823 out of 10825 data points were correctly classified by at least one of the 210 classifiers. To exploit the predictive power of the individual classifiers, we set up a voting scheme, where each classifier classified a given data point and the majority classification was used as the final hypothesis. Using this voting classifier, the accuracy improves to 81.9% with precision of 76.7% and recall of 82.5%. This is a very high accuracy, given that we are ignoring $f0$ and duration information in these experiments and relying entirely upon energy features.

The results described above are based on energy features of the whole word. Experiments using energy features extracted from portions of the word (only the longest syllable, only the longest syllable nucleus, all syllable nuclei) each produced significantly poorer classification accuracy. However, the energy features derived from all of the syllable nuclei in the word predicted pitch accent significantly better than those extracted from only the longest syllable nucleus. These findings indicate that there is, in fact, energy information both in relatively short syllable nuclei and outside the syllable nuclei entirely which is useful for predicting pitch accent. Additionally, we found that the classification accuracies obtained by using the longest syllable in the word and the longest syllable nucleus in the word as the region of analysis did not sig-

[1] Statistical significance was determined by χ^2 with $p \leq 0.001$.

[2] NB:8kHz was the Nyquist rate of the corpus.

[3] Classification of speaker h1, a male, with features extracted from the longest nuclei of the word.

[4] Statistical significance determined by χ^2 indicating $p <= 0.05$

nificantly differ. In the context of detecting whether an individual syllable is accented or not, this is an interesting result; there is a clear parallel between this and the observation that the duration of a syllable is equivalently predictive of syllabic prominence as the duration of a syllable nucleus [28].

6. Conclusions and Future Work

In this paper, we have described an analysis by classification approach to determining how the energy contributions from different frequency bands correlate with pitch accent in read SAE. Our experiments confirm that the energy component from different frequency subbands predict pitch accent with differing degrees of success. Specifically, we have found that the band between 3 and 18 bark to be the most predictive on our whole dataset, with data taken from all four speakers. However, the band between 2 and 20 bark predicts pitch accent significantly better than the band from 3 to 18 for one speaker, while not predicting significantly worse for the other three. As the Nyquist rate of our corpus is between 19 and 20 bark, it is moot whether this band is more accurately reported as "all frequencies above 2 bark" or strictly between 2 and 20 bark.

We have found that the differences in predictive power between frequency subbands is not merely one of varied accuracy, but, rather, that different subbands can accurately detect pitch accents on different sets of words. Using a voting scheme, we were able to construct a classifier based on every subband we examined – base frequency from 0 to 19 bark, bandwidth from 1 to 20 bark. This voting scheme predicts pitch accent with 81.9% accuracy. Extending this approach, we will investigate an automatic method to determine which frequency regions contain the most predictive energy information for pitch accent detection based on other features of a given word.

As our results are dependent on manual transcription, our method currently can only be applied to the task of annotating transcribed corpora – TTS inventories, for example – for pitch accent. Future experiments using automatic word segmentation will determine how robust the results are to segmentation error. In our future work, we will also repeat our experiments on spontaneous speech to determine how similarly pitch accent is realized in the two genres with respect to energy. Additionally, we will incorporate our parallel research into the usefulness of additional features based on f0 and duration into a more general pitch accent classifier.

7. Acknowledgments

The authors would like to thank Martin Jansche and Dan Ellis for their helpful comments. Work reported in this paper was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023.

8. References

[1] P. C. Bagshaw. *Automatic prosodic analysis for computer aided pronunciation teaching*. PhD thesis, University of Edinburgh, 1994.

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

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

[4] D. L. Bollinger. A theory of pitch acent in english. *Word*, 14:109–149, 1958.

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

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

[7] 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*, pages 523–526, 1999.

[8] R. Delmonte. Slim prosodic automatic tools for self-learning instruction. *Speech Communication*, 30:145–166, 2000.

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

[10] A. J. F. et al. Speech processing by man and machine. In T. H. Bullock, editor, *Recognition of Complex Acoustic Signals*. Report of Dahlem Workshop, Berlin, 1977.

[11] G. Fant, A. Kruckenberg, and J. Liljencrants. Acoustic-phonetic analysis of prominence in swedish. In A. Botinis, editor, *Intonation, Analysis, Modelling and Technology*, pages 55–86. Kluwer, 2000.

[12] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Lunguistics*, 12(3):175–204, 1986.

[13] 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*, pages 57–60, 2001.

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

[15] J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of the 34th conference on Association for Computational Linguistics*, pages 286–293, 1996.

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

[17] R. Kompe. Prosody in speech understanding systems. *Lecture Notes in Artificial Intelligence*, 1307:1–357, 1997.

[18] D. Ladd. *The Structure of Intonational Meaning*. Indiana University Press, Bloomington, 1980.

[19] P. Mermelstein. Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, 58(4):880–883, 1975.

[20] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 945–948, 1998.

[21] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980.

[22] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90(6):2956–2970, 1991.

[23] Y. Ren, S.-S. Kim, M. Hasegawa-Johnson, and J. Cole. Speaker-intependent automatic detection of pitch accent. In *ICSA International Conference on Speech Prosody*, pages 521–524, 2004.

[24] 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*, volume 2, pages 12–16, 1992.

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

[26] A. M. C. Sluijter and V. J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4):2471–2485, 1996.

[27] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, 1997.

[28] F. Tamburini. Prosodic prominence detection in speech. In *Proc. 7th International Symposium on Signal Processing and its Applications = ISSPA2003*, pages 385–388, 2003.

[29] F. Tamburini. Automatic prominence identification and prosodic typology. In *Proc. InterSpeech 2005*, pages 1813–1816, 2005.

[30] J. Tepperman and S. Narayanan. Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners. In *Proc. ICASSP*, volume 1, pages 937–940, 2005.

[31] J. P. H. van Santen. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546, 1992.

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

[33] C. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481, 1994.

[34] 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 International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pages 192–196, 1999.