Analysis of Prosody Increment Induced by Pitch Accents for Automatic Emphasis Correction

Yang Zhang*, Gautham Mysore†, Floraine Berthouzoz†, Mark Hasegawa-Johnson*

* University of Illinois, Urbana-Champaign † Adobe Research, San Francisco, USA

{yzhan143, jhasegaw}@illinois.edu, {gmysore, floraine}@adobe.com

Abstract

We are interested in developing an automatic emphasis correction system, which converts any unemphasized word in an utterance into emphasized. Analyzing how prosody changes from unaccented to accented is crucial for the task. While previous works on prosody reconstruction only model the prosody contour itself instead of the increment, we propose a framework to study the prosody increment induced by pitch accents from real speech in a statistically rigorous manner. This framework also infers the degree of emphasis of each word to account for the additional prosody variations due to metalinguistic factors. According to the analysis results, this framework provides a lot of useful insights into the prosody increment, which are consistent with many existing studies on pitch accent and emphasis.

Index Terms: Pitch accent, emphasis, prosody increment, regression analysis, automatic emphasis correction.

1. Introduction

Large amounts of voiceover and narration content are regularly recorded for applications such as podcasts, demo videos, lecture videos, and audio stories. Most people do not have professional voice acting skills, so such content typically does not sound as good as professional voiceovers when recorded by amateurs. Emphasis is one of the important aspects of good voiceover performance. Our problem, therefore, is to develop an algorithm that automatically emphasizes a word that was not emphasized by manipulating the prosody with signal processing techniques. To do this, we need a model that describes how prosody changes from an unemphasized word to an emphasized word.

There have been a lot of studies on prosody behaviors in emphasized words, which conclude that three prosody components are important cues for emphasis [1]: F0, duration, and spectral balance.

F0 has been considered as one of the most important cues of emphasis. The concept of pitch accent [2], which is defined primarily by the shape pitch contour, is the acoustic phenomenon through which an emphasis is expressed. According to [3], a pitch accent is characterized by either a high or a low pitch excursion in pitch contour. The ToBI system [2], further classifies it into different types. Many pitch contour models [4, 5, 3] specifically model this behavior..

Duration in an emphasized word usually gets elongated. According to [6, 7], the elongation is typically 10%-20%. Subjective analyses [8] reveal that duration is a significant cue to perceive a phrasal prominence. As for the internal structure, evidence [9] show that the stress syllable tends to vary most in duration under different prosodic contexts.

This paper is dedicated to Floraine Berthouzoz, who has unfortunately passed away during the development of the work.

Spectral balanced is perceived roughly as loudness, which is proven to be another important cue for emphasis [10]. To produce an emphasized word, a speaker would generally increase the glottal effort. Acoustically, this corresponds to 1) a higher amplitude overall; and 2) a flattening of the speech spectrum, i.e. a higher magnitude in higher frequency [11, 12]. Subjective evaluation shows that the latter phenomenon is a more important cue for perceiving vocal effort, and therefore emphasis.

Existing speech synthesis systems, e.g. [13, 14, 15], directly models prosody behaviors at emphasis. However, they only model the prosody contours themselves instead of the increment from unemphasized to emphasized. Some other works compare prosody with and without emphasis, e.g. [16], but they are based primarily on lab-recorded speech of very constrained utterances. In this paper, we propose a framework to study prosody increments induced by pitch accents, the acoustic correlate of emphasis, in a statistically rigorous manner. In addition, the proposed framework infers the degree of emphasis of each word to account for the additional prosody variations due to metalinguistic factors.

2. The Analysis Methodology

2.1. The Difference Model

The basic idea of our approach is to have two regression models, one for the words/phones with emphasis and one without, and then compute the difference between the two models.

Formally, define the two regression models as

$$\begin{aligned} \log \boldsymbol{Y}_e^{(p)} &= \boldsymbol{X}_e^{(p)} \boldsymbol{B}_e^{(p)} + \boldsymbol{U}_e^{(p)} \\ \log \boldsymbol{Y}_u^{(p)} &= \boldsymbol{X}_u^{(p)} \boldsymbol{B}_u^{(p)} + \boldsymbol{U}_u^{(p)} \end{aligned} \tag{1}$$

where the superscript (p) denotes the prosody components (f for F0, d for duration and s for spectral balance). Subscript e denotes the emphasis model and u denotes the unemphasis model. $\boldsymbol{Y}$ is a matrix of independent variables of prosody, which will be defined soon. Each row is an observation, and each column is a prosody parameter. Each observation can be a phone or a word, depending on prosody component, which will be specified later. $\boldsymbol{X}$ is a matrix of explanatory variables, which depict the phonetic, lexical and contextual information of the target word/phone. Each row is an observation, and each column is an explanatory variable. $\boldsymbol{B}$ is the matrix of regression coefficients, whose element (i, j) denotes the coefficient of the j-th prosody parameter on the i-th explanatory variable. Finally, $\boldsymbol{U}$ is a matrix of error terms.

In addition to the common assumptions made by ordinary least square approach [20], we also assume that the corresponding columns in $\boldsymbol{U}_e^{(p)}$ and $\boldsymbol{U}_u^{(p)}$ are independent.

For each prosody component p, the choice of prosody parameters and explanatory variables are identical for the emphasis and unemphasis model, i.e. the columns in $\boldsymbol{X}_e^{(p)}$ and $\boldsymbol{X}_u^{(p)}$ correspond to the same variables, and similarly for $\boldsymbol{Y}_e^{(p)}$ and $\boldsymbol{Y}_u^{(p)}$. The only difference between them is that $\boldsymbol{X}_e^{(p)}$ and $\boldsymbol{Y}_e^{(p)}$ are extracted from words/phones that are labeled as accented, whereas $\boldsymbol{X}_u^{(p)}$ and $\boldsymbol{Y}_u^{(p)}$ are from unaccented words/syllables.

With the two models with identical structure, we can define the difference model:

$$\boldsymbol{B}_d^{(p)} = \boldsymbol{B}_e^{(p)} - \boldsymbol{B}_u^{(p)} \tag{2}$$

Combining equations (1) and (2), for any row vectors $\boldsymbol{x}^{(p)}$, we have

$$\nabla \log \boldsymbol{y}^{(p)} = \boldsymbol{x}^{(p)} \boldsymbol{B}_d^{(p)} + \boldsymbol{u}_d^{(p)} \tag{3}$$

where $\nabla \log \boldsymbol{y}^{(p)}$ can be interpreted as under the phonetic, lexical and contextual setting specified by $\boldsymbol{x}^{(p)}$, the **percentage increase** in the prosodic parameters $\boldsymbol{y}^{(p)}$ if an emphasis is to be placed. This is because when $\nabla \log \boldsymbol{y}^{(p)} \ll 1$, the logarithm difference approximates the percentage increment. Therefore, each element in $\boldsymbol{B}_d^{(p)}$ can be interpreted as the expected **additional** prosody increment if the corresponding explanatory variable increases by 1 unit.

2.2. The F0 Model

In the F0 model, the observation unit is a word. The F0 parameters are motivated by the TILT model [3], which proposes that the accented pitch contours can be divided into two classes - the peak accents and the through accents. In this paper, we are primarily interested in the peak accent. Therefore, we introduce three F0 parameters to depict the emphasis F0 contour, i.e. $\boldsymbol{Y}^{(f)}$:

- ***Peak level*** - the max F0 (in hertz) in or within 20ms around the stressed vowel;
- ***Start level*** - the first F0 value of the accented word;
- ***End level*** - the last F0 value of the accented word;

The explanatory variables of the F0 model, $\boldsymbol{X}^{(f)}$, include:

- **Break Context:**
  Preceding break level - the break level right before the word;
  Proceeding break level - the break level right after the word;
- **Lexical Structure:**
  Length before stress - # phones before the stressed vowel;
  Length after stress - # phones after the stressed vowel;
- **Major Break Context:**
  Preceding major break - distance to the previous level 4 break (in number of phones);
  Proceeding major break - distance to the next level 4 break (in number of phones);
- **Vowel Identities in the Stressed Vowel:**
  Is stress back - 1 if and only if (iff) the stress vowel is back and 0 otherwise;
  Is stress low - 1 iff the stress vowel is low;
  Is stress tense - 1 iff the stress vowel is tense;
  Is stress reduced - 1 iff the stress vowel is back;
- **Vowel Identities in the Final Vowel:**
  Is final back, ***Is final low***, ***Is final tense***, ***Is final reduced*** - defined similarly as above on the final vowel;
- **F0 Context:**
  Preceding F0 - average F0 within 500ms preceding the word;
  Proceeding F0 - average F0 within 500ms proceeding the word;
- **Pitch Accent:**
  Is H* - 1 iff the pitch accent is H*;
  Is L+H* - 1 iff the pitch accent is L+H*;
  Is L*+H - 1 iff the pitch accent is L*+H;
- **Accent Context:**
  Is preceding accent - 1 iff the preceding word is accented;
  Is proceeding accent - 1 iff the proceeding word is accented.

Note that in the unemphasis model, where there is no pitch accent type, all the pitch accent variables are set to zero.

2.3. The Duration Model

Unlike in the F0 model, the observation unit in the duration model is each vowel in a word.

The duration parameter, $\boldsymbol{Y}^{(d)}$ is simply the duration of each vowel in seconds. The explanatory variables, $\boldsymbol{X}^{(d)}$, include:

- **Break Context:** same as in the F0 model;
- **Lexical Structure:**
  Is stress - 1 iff the vowel is the stress vowel of the word;
  Is first - 1 iff the vowel is the first vowel of the word;
  Is last - 1 iff the vowel is the last vowel of the word;
- **Major Break Context:** same as in the F0 model;
- **Current Vowel Identity:**
  Is back, ***Is low***, ***Is tense***, ***Is reduced*** - defined similarly as the vowel identities in the F0 model on the current vowel;
- **Preceding Phone Identity:**
  Is preceding back, ***Is preceding low***, ***Is preceding tense***, ***Is preceding reduced*** - defined similarly as above on the preceding phone;
  Is preceding vowel - 1 iff the preceding phone is a vowel;
  Is preceding stop - 1 iff the preceding phone is stop;
  Is preceding nasal - 1 iff the preceding phone is nasal;
  Is preceding glide - 1 iff the preceding phone is glide;
  Is preceding fricative - 1 iff the preceding phone is fricative;
  Is preceding affricate - 1 iff the preceding phone is affricate;
  Is preceding liquid - 1 iff the preceding phone is liquid;
- **Proceeding Phone Identity:** defined similarly as above on the proceeding phone;
- **Duration Context:**
  Preceding duration - average phone duration within 1s preceding the word;
  Proceeding duration - average phone duration within 1s proceeding the word;
- **Accent Context:** same as in the F0 model.

2.4. The Spectral Balance Model

Similar to the duration model, the observation unit in the spectral balance model is also a vowel in a word.

The parameters in spectral balance model are motivated by the spectral flattening effect observed in existing studies on emphasis and vocal efforts [10], as mentioned in section 1. To capture this effect, our proposed model divides the 16kHz speech into 5 bands: 0-500Hz, 500-1kHz, 1k-2kHz, 2k-4kHz, 4kHz-8kHz, similar to the division in [10]. The 5 parameters of the spectral balance model, $\boldsymbol{Y}^{(s)}$, are then defined as the averaged energy in these 5 bands (in linear scale).

The explanatory variables, $\boldsymbol{X}^{(s)}$, are almost the same as in the duration model, except that the duration context is replaced with the spectral balance context, which consists of 10 variables representing the average energy in the 5 frequency bands within 1s preceding and proceeding the word.

2.5. The Degree of Emphasis

Previous studies [18, 19] agrees that emphasis varies in level or degree. We would like to estimate how the degree of emphasis would affect the prosody increment. However, ToBI avoids annotating the degree of emphasis, considering this to be a metalinguistic variable [2]. Therefore we should simultaneously infer the degree of emphasis for each observation, and estimate the regression coefficients of the prosody parameters.

Formally, to incorporate the degree of emphasis, the error term $\boldsymbol{U}_e^{(p)}$ in equation (1) is further decomposed as

$$\boldsymbol{U}_e^{(p)} = \boldsymbol{d}^{(p)}\boldsymbol{\beta}_e^{(p)} + \boldsymbol{E}_e^{(p)} \tag{4}$$

where $\boldsymbol{d}^{(p)}$ is a column vector, and each element represents the degree of emphasis of the corresponding observation. $\boldsymbol{\beta}_e^{(p)}$ is a row vector of the regression coefficients of the corresponding prosody parameters. $\boldsymbol{E}_e^{(p)}$ is the remaining error terms. We assume that the degree of emphasis of (different vowels of) the same word is the same across all prosody components.

2.6. Least Square Estimation

This section briefly introduces the estimation of $\boldsymbol{B}_d^{(p)}$ and $\boldsymbol{\beta}_e^{(p)}$.

First, by the assumptions given in section 2.1, an unbiased estimate of $\boldsymbol{B}_d^{(p)}$ is given by

$$\hat{\boldsymbol{B}}_d^{(p)} = \hat{\boldsymbol{B}}_e^{(p)} - \hat{\boldsymbol{B}}_u^{(p)} \tag{5}$$

where $\hat{\boldsymbol{B}}_e^{(p)}$ and $\hat{\boldsymbol{B}}_u^{(p)}$ are OLS estimates of $\hat{\boldsymbol{B}}_e^{(p)}$ and $\hat{\boldsymbol{B}}_u^{(p)}$ respectively. Since the emphasis and unemphasis models are independent, the variance of $\hat{\boldsymbol{B}}_d^{(p)}$ is simply the sum of the variances of $\hat{\boldsymbol{B}}_e^{(p)}$ and $\hat{\boldsymbol{B}}_u^{(p)}$.

Then, denote $\hat{\boldsymbol{U}}_e^{(p)}$ as the OLS regression residuals of the emphasis model, $\boldsymbol{d}^{(p)}$ and $\boldsymbol{\beta}_e^{(p)}$ are estimated such that

$$\hat{\boldsymbol{d}}^{(p)}, \hat{\boldsymbol{\beta}}_e^{(p)} = \underset{\boldsymbol{d}^{(p)}, \boldsymbol{\beta}_e^{(p)}}{\operatorname{argmin}} \sum_{p=f,d,s} \left\| \hat{\boldsymbol{U}}_e^{(p)} - \boldsymbol{d}^{(p)}\beta_e^{(p)} \right\|_F \tag{6}$$

where $\|\cdot\|_F$ is the Frobenius norm.

The solution to this problem is given by setting $\hat{\boldsymbol{d}}^{(p)}$ and $\hat{\boldsymbol{\beta}}_e^{(p)}$ to the left and right singular vectors corresponding to the largest singular value of the residual matrix.

3. Results and Discussions

3.1. Configurations

The analyses are performed on the Boston University Radio Speech Corpus [17]. The model is trained on the lab data, professional radio portion of the corpus across all the 7 speakers.

We apply the normal distribution to test the statistical significance of each regression coefficient. Coefficients with significance level above 90% are shown in bold; those above 95% are shown in bold and underlined. No significance test is performed for the coefficients on the degree of emphasis.

3.2. The F0 Results

The regression results of the F0 model is shown in table 1. Recall that each coefficient can be interpreted as the additional emphasis increment in F0 if the corresponding variable increases by 1 unit. Particularly, if the variable is an indicator variable (e.g. *is H**), the coefficient represents the additional increment if the corresponding statement is true.

Table 1: Linear regression coefficients: differences in F0 between emphasized and unemphasized words

	Peak F0	**Start F0**	**End F0**
Intercept	-0.188	-0.089	-0.094
Prec. break level	**0.008**	**0.015**	0.002
Proc. break level	0.005	**0.012**	**-0.045**
Len. before stress	0.000	**-0.018**	0.006
Len. after stress	-0.004	-0.004	**-0.014**
Prec. major break	**-0.001**	**-0.001**	**-0.001**
Proc. major break	**-0.001**	**-0.001**	0.000
Is stress back	0.011	-0.001	**-0.034**
Is stress low	**0.060**	**0.062**	0.020
Is stress tense	-0.020	0.008	-0.009
Is stress reduced	**-0.105**	-0.055	-0.019
Is final back	-0.003	0.015	**0.059**
Is final low	**-0.036**	**-0.047**	-0.025
Is final tense	0.023	0.005	**0.045**
Is final reduced	0.023	0.006	**-0.077**
Prec. F0	0.036	0.006	0.016
Proc. F0	0.018	0.027	0.037
Is H*	**0.151**	**0.112**	**0.117**
Is L+H*	**0.159**	**0.070**	**0.103**
Is !H*	0.011	0.017	-0.014
Is prec. accent	**-0.044**	**-0.029**	**-0.056**
Is proc. accent	**-0.029**	**-0.048**	**-0.028**
R-squared	0.483	0.389	0.495
Degree of emphasis	0.009	0.008	0.003

There are several important observations. First, the most significant factors of the pitch increment is the pitch accent category. *Is L*+H* is removed and set as the baseline category. H* and L+H* have higher pitch increase, and H* has a slightly lower peak increase and a much higher start increase, which agrees with the canonical definition of these accent types.

Second, the vowel identities have a significant impact on the increment, especially on the ending pitch increase. In particular, a low vowel in the stress syllable corresponds to a higher pitch increase, whereas that in the final syllable suppresses the increase. This may be because a low vowel has a wider opening, and thus is more influential with its target pitch level, which is high if in the stressed syllable, and low in the final.

Third, the preceding and proceeding break level context have different effects. The preceding corresponds to a higher increase overall, while the proceeding corresponds to a higher increase in the lexical stress, but a lower increase in the final syllable. This is because a major break is usually followed by a pitch reset, and preceded by a pitch drop to the base level.

Fourth, the F0 context does NOT have a significant effect. This is not a trivial result: F0 context is significant in both the emphasis and the unemphasis models, but not significant in their difference. This indicates that speakers always produce a similar percentage pitch increase for an emphasis, regardless of how high the surrounding pitch is.

Finally, the degree of emphasis result is shown in the last line of table 1. We can observed that the peak pitch gets the most additional increase. Since the Peak F0 is around the stressed syllable, this observation agrees with the conclusion that pitch accents are realized in lexical stress [2].

3.3. The Duration Results

Table 2 presents the regression results of the duration model. *Is prec. tense*, *Is prec. reduced* and *Is proc. affricate* are removed and set as the baseline categories. Here are our major findings.

First, the vowel identities and phonetic context are the most important factors of the duration elongation. The back, high,

Table 2: Linear regression coefficients: differences in duration between emphasized and unemphasized words

Intercept	0.086	Is prec. fricative	0.073
Prec. break level	**-0.011**	Is prec. affricate	-0.070
Proc. break level	**-0.020**	Is prec. liquid	0.012
Is stress	**0.135**	Is proc. back	**0.599**
Is first	0.023	Is proc. low	-0.133
Is last	**0.103**	Is proc. tense	0.089
Len. before stress	0.001	Is proc. reduced	**-0.714**
Len. after stress	0.000	Is proc. vowel	0.080
Is back	**0.033**	Is proc. stop	0.002
Is low	**-0.092**	Is proc. nasal	**0.071**
Is tense	**-0.172**	Is proc. glide	0.053
Is reduced	**0.116**	Is proc. fricative	**0.047**
Is prec. back	0.041	Is proc. liquid	**0.097**
Is prec. low	0.143	Prec. duration	**0.685**
Is prec. vowel	-0.029	Proc. duration	-0.096
Is prec. stop	0.036	Is prec. accent	-0.019
Is prec. nasal	-0.073	Is proc. accent	0.022
Is prec. glide	-0.028		
Degree of emphasis	0.046	R-squared	0.198

lax, and reduced vowels correspond to a greater elongation. The high vowels get greater elongation because they are shorter to start with (when there's no pitch accent) and have more room for elongation. The lax vowels get more lengthening probably because they are more elastic than the tense vowels. The proceeding phone, especially the proceeding consonant, has a much greater impact than does the preceding phone. Of all preceding consonants, liquids lead to the greatest lengthening, followed by nasals, probably because they are voiced consonants.

Second, in terms of lexical structure, *Is stress* and *Is last* are both significant with positive values, but *Is first* is not. This shows that, unlike F0, duration elongation for pitch accent is realized in both the stress syllable and the final syllable.

Third, the duration increment is positively correlated with the preceding averaged duration. This shows that when the phone rate is low, the percent elongation gets greater. This is different from the F0 case, where the increment is not significantly correlated with the surrounding pitch.

3.4. The Spectral Balance Results

The regression results of the spectral balance model are given in table 3. Similar to the duration model, *Is prec. tense*, *Is prec. reduced* and *Is proc. affricate* are removed and set as the baseline categories. To save space and reduce distractions, only the results of the first four frequency bands are displayed.

Our first observation is that there is a significant spectral flattening effect, as can be shown in the intercept terms and the degree of emphasis. In both rows, the coefficients increase as the frequency increases, and the gap between the first band and the second is greater than the other gaps. These results reveal that higher frequency bands are more amplified than lower bands, when there is a pitch accent.

In terms of lexical structure, the stress and final vowels get most amplification, and the first vowels get significantly less. This observation also implies that the lexical stress and final syllable are primary locations to realize a pitch accent.

The phonetic context has significant and consistent effect on the energy increment across frequency bands, especially the preceding phone identity. The baseline of the preceding phone category is tense vowel and reduced vowel. Therefore, preceding consonants lead to a much smaller amplification than preceding vowels. One explanation is that the concatenation of vowels enables glottal effort to gradually rise to a higher level.

Table 3: Linear regression coefficients: differences in spectral balance between emphasized and unemphasized words

	0-500	**500-1k**	**1k-2k**	**2k-4k**
Intercept	-0.155	0.732	**1.000**	**1.381**
Prec. break level	**0.106**	**0.121**	**0.149**	**0.176**
Proc. break level	0.019	**-0.057**	**-0.038**	**-0.078**
Is stress	**0.308**	**0.508**	**0.528**	**0.360**
Is first	-0.264	**-1.209**	**-1.393**	**-1.203**
Is last	**0.223**	**0.424**	**0.346**	**0.298**
Len. before stress	0.000	0.002	0.000	-0.001
Len. after stress	-0.001	-0.001	-0.002	**-0.004**
Is back	**0.192**	**0.399**	**0.331**	-0.037
Is low	**0.170**	0.061	**-0.119**	**0.171**
Is tense	-0.007	**0.184**	**0.120**	-0.091
Is reduced	0.062	-0.081	0.113	**0.517**
Is prec. back	-0.167	-0.448	-0.605	-0.032
Is prec. low	-0.590	-0.321	-0.002	-0.366
Is prec. vowel	0.216	-0.659	-0.944	**-1.234**
Is prec. stop	-0.303	**-1.417**	**-1.613**	**-1.531**
Is prec. nasal	-0.363	**-1.479**	**-1.651**	**-1.684**
Is prec. glide	-0.500	**-1.726**	**-1.733**	**-1.405**
Is prec. fricative	-0.205	**-1.120**	**-1.316**	**-1.307**
Is prec. affricate	-0.362	**-1.703**	**-1.745**	**-1.462**
Is prec. liquid	-0.049	**-1.138**	**-1.326**	**-1.264**
Is proc. back	0.073	-0.520	**-1.391**	-0.511
Is proc. low	0.159	0.013	0.352	0.531
Is proc. tense	-0.150	0.029	-0.671	0.021
Is proc. reduced	-0.069	0.277	**1.497**	0.070
Is proc. vowel	-0.010	0.372	0.358	-0.331
Is proc. stop	-0.088	-0.099	0.018	-0.076
Is proc. nasal	**0.148**	**0.398**	**0.301**	**0.150**
Is proc. glide	0.262	0.053	0.353	0.088
Is proc. fricative	**0.161**	**0.385**	**0.331**	**0.324**
Is proc. liquid	**-0.163**	-0.009	-0.061	-0.046
Prec. 0-500	-0.138	0.319	0.130	0.105
Prec. 500-1k	**-0.717**	**-1.581**	**-1.240**	-0.686
Prec. 1k-2k	**0.901**	**1.504**	**1.358**	0.791
Prec. 2k-4k	-0.374	-0.285	-0.272	-0.280
Prec. 4k-8k	-0.035	-0.184	-0.160	-0.110
Proc. 0-500	0.022	-0.377	-0.111	-0.106
Proc. 500-1k	0.548	**1.424**	**1.138**	0.637
Proc. 1k-2k	**-0.751**	**-1.344**	**-1.273**	-0.690
Proc. 2k-4k	**0.384**	0.281	0.272	0.343
Proc. 4k-8k	0.034	0.153	0.129	0.014
Is prec. accent	**-0.121**	-0.032	0.059	-0.005
Is proc. accent	**-0.085**	0.029	**0.129**	-0.005
R-squared	0.258	0.363	0.285	0.268
Degree of emphasis	0.481	0.862	0.912	0.987

Finally, as for the break context, we can see that a higher preceeding break level corresponds to a more significant amplification and spectral flattening, whereas a higher proceeding break level corresponds to a smaller amplification. This can be explained by the decaying vocal effort through time. Right after a major prosody break, the speaker usually has more breath to produce greater vocal effort compared to the end of a phrase.

4. Conclusion and the Future Direction

In this paper, we propose a framework to analyze prosody increment induced by pitch accents in natural speech. The results on the Boston University radio speech corpus reveal that the break context, lexical structure and phonetic information are among the most significant factors on the prosody increments. The framework also infers the degree of emphasis as an important factor. The analysis provides theoretical support of our automatic emphasis system, which will be our future direction.

5. References

[1] A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth, "Can we tell apart intonation from prosody (if we look at accents and boundaries)?" in *Intonation: Theory, Models and Applications*, 1997.

[2] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "Tobi: a standard for labeling english prosody." in *The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16, 1992*, 1992.

[3] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1, pp. 169–186, 1994.

[4] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, pp. 347–355, 1998.

[5] C.-y. Tseng, S.-h. Pin, Y. Lee, H.-m. Wang, and Y.-c. Chen, "Fluent speech prosody: Framework and modeling," *Speech Communication*, vol. 46, no. 3, pp. 284–309, 2005.

[6] C. Coker, N. Umeda, and C. Browman, "Automatic synthesis from ordinary english test," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 3, pp. 293–298, 1973.

[7] D. H. Klatt, "Vowel lengthening is syntactically determined in a connected discourse." *Journal of phonetics*, vol. 3, no. 3, pp. 129–140, 1975.

[8] L. Hitchcock and S. Greenberg, "Vowel height is intimately associated with stress accent in spontaneous american english discourse." in *INTERSPEECH*, 2001, pp. 79–82.

[9] H. Kim and J. Cole, "The stress foot as a unit of planned timing: evidence from shortening in the prosodic phrase." in *INTERSPEECH*, 2005, pp. 2365–2368.

[10] A. M. Sluijter, V. J. Van Heuven, and J. J. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 503–513, 1997.

[11] J. F. Brandt, K. F. Ruder, and T. Shipp Jr, "Vocal loudness and effort in continuous speech," *The Journal of the Acoustical Society of America*, vol. 46, no. 6B, pp. 1543–1548, 1969.

[12] R. Glave and A. Rietveld, "Is the effort dependence of speech loudness explicable on the basis of acoustical cues?" *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 875–879, 1975.

[13] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[14] C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis," *Group*, vol. 1, no. L2, p. L3, 2000.

[15] A. Raux and A. W. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 700–705.

[16] Y. Xu and C. X. Xu, "Phonetic realization of focus in english declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[17] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "Boston university radio speech corpus," *Philadelphia: Linguistic Data Consortium*, 1996.

[18] D. R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 25, no. 3, pp. 313–342, 1997.

[19] C. W. Wightman, "Perception of multiple levels of prominence in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1881–1881, 1993.

[20] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003.