Relative Frequency Affects Pitch Accent Realisation: Evidence for Exemplar Storage of Prosody

Katrin Schweitzer[1], *Sasha Calhoun*[2,3],
Hinrich Schütze[1], *Antje Schweitzer*[1], *Michael Walsh*[1]

[1] University of Stuttgart, Germany
[2] University of Edinburgh, UK, [3] Victoria University of Wellington, NZ
Katrin.Schweitzer@ims.uni-stuttgart.de, Sasha.Calhoun@vuw.ac.nz

Abstract

This paper looks at the variability of pitch accent realisations on different word types in relation to the relative frequency with which each word type occurs with a particular pitch accent type (among all pitch-accented occurrences of the word). Results indicate that pitch accent realisation variability decreases with increasing relative frequency. This is consistent with Exemplar Theory: relative frequency can be regarded as an encoding of the diversity of prosodic contexts in which a word occurs. If there is less prosodic variability among stored tokens of a word, there will be less variability in production. It seems that pitch contours are stored with words, contrary to the standard assumption that accenting is "post-lexical" in English.
Index Terms: exemplar theory, pitch accents, intonation

1. Introduction

It is well established that frequency of usage affects linguistic realisation across a variety of domains [1, 2]. These findings can be well explained within the framework of *Exemplar Theory* [3, 4, 1] (c.f. section 2), where it is assumed that language input is stored in memory in rich detail as exemplars and that targets for subsequent productions are derived from them. Frequency effects are assumed to result from different numbers of stored exemplars. In the domain of prosody, the variability of syllable duration [5, 6] and pitch accents [7] has been found to be subject to frequency effects: The duration of frequent syllables is less easy to predict from the length of the underlying phones than it is for the case for infrequent syllables [5, 6]; and the shape of frequent pitch accents is more variable than the shape of infrequent ones [7]. Moreover, it has been argued that intonation contours can be stored with lexical items [8].

If pitch accent shapes are stored, this should influence production and it would be expected that pitch accent shape on a specific word type is affected by the diversity of (prosodic) contexts in which that type occurs (e.g. pitch accent type, information status as well as phonological context). It is assumed here that the diversity of contexts in which a specific word occurs is reflected by the diversity of pitch accent types associated with it and that this can be measured by the relative frequencies of the pitch accent types on the word, i.e. the proportion of times a word occurs with a given pitch accent type. The idea behind this is that for words that appear in diverse contexts (i.e.word-accent pairs of low relative frequency) exemplars from a greater number of competing contexts are activated in production. Therefore, greater variability of the resulting F_0-contours is expected. That is, while high *absolute* frequency increases the variability of the F_0-contours [7], we expect high *relative* frequency to decrease the variability since it reflects low diversity of contexts.

The research presented in this paper specifically targets this claim. In particular, we address the following questions:

1. Does the relative frequency of a word-accent pair affect the realisation of the pitch accent?
2. If so, what is the impact of relative frequency on a set of word-accent pairs?
3. What conclusions can be drawn for theories of lexical storage and post-lexical accenting?

In examining the variability of pitch accent tokens in a speech corpus (section 3), a parametric intonation model, known as PaIntE ([9], cf. section 4), is employed to extract meaningful parameters from pitch-accented words of varying relative frequency in a speech corpus. Variability of pitch-accented word types is measured using Euclidean distance (section 5). Linear regression models assess the relationship between relative frequency and variability (section 6). Results are interpreted from a usage-based perspective (section 7).

2. Exemplar Theory

Exemplar Theory is concerned with the idea that language is acquired by repeated exposure to concrete language input, and it has successfully accounted for a number of language phenomena, including diachronic language change and frequency of occurrence effects and grammaticalisation [2], syllable duration variability [5, 6], entrenchment and lenition [1], among others. Central to Exemplar Theory are the notions of exemplar storage, frequency of occurrence, recency of occurrence, and similarity. There is an increasing body of evidence which indicates that significant storage of language input exemplars, rich in detail, takes place in memory [10, 11]. These stored exemplars are then employed in the categorisation of new input percepts. Similarly, production is facilitated by accessing these stored exemplars as production targets. Computational models of the exemplar memory also argue that it is in a constant state of flux with new inputs updating it and old unused exemplars gradually fading away [1].

Up to now, little exemplar-theoretic research has examined pitch accent prosody (but see [12] for memory-based prediction of pitch accents and prosodic boundaries, [8] for evidence of word storage with intonation contours, and [7] for evidence of frequency effects on within-type pitch accent variability) and to the authors' knowledge this paper represents the first attempt, from a usage-based perspective, to examine the relationship between the diversity of prosodic contexts in which a word occurs

and the variability of realisations of one specific pitch accent type.

3. Data

The corpus used in this study is the Boston Radio News Corpus, a collection of radio news broadcasts [13]. A subset of the corpus has been labelled prosodically using ToBI [14]. The analysed data set contains approximately 1 h of speech from five professional speakers (3 female, 2 male). One speaker produced nearly half the tokens in our data set; however, as tokens from her were fairly evenly spread across the frequency bins (see below), we believe this did not unduly influence results.

For each nuclear H* and L+H* accent (as well as the downstepped versions of these accents), four parameters reflecting its shape were extracted using a parametric intonation model (PaIntE, [9], see section 4). Outlying tokens, i.e. tokens where one or more of the dimension values fell within the upper or lower 2.5 percentile of the dimension's range, were removed. As the purpose of the study was to examine the variability between tokens of a word-accent pair, types with only one occurrence were not analysed so that the nuclear H* dataset (referred to as *HN*) finally comprises 1425 tokens and 465 types, the nuclear L+H* set *LHN* 306 tokens (123 types).

For each word type, the frequency of the combination of this word type with either H* or L+H* was calculated as was the frequency of how often the word occurred with any pitch accent. The ratio of these two values measures the relative frequency of the word-accent-pair:

$$h(w_a) = \frac{n(w_a)}{n(w_x | x \in acc)} \tag{1}$$

where $h(w_a)$ is the relative frequency of word w and accent a. $n(w_a)$ is the absolute frequency of the combination of these two, and acc is the set of accent types, hence $n(w_x)$ corresponds to the absolute frequency with which the w occurs with any pitch accent. It is assumed here that this value reflects the diversity of the prosodic context in which the word occurs. The higher the proportion of contexts with a different pitch accent, the lower the relative frequency of the particular accent on this word. The word "school" for instance, occurred 19 times with a pitch accent, in eight of these occurrences the accent has been labelled as H*. Thus, the relative frequency of the word-accent pair "school–H*" is $h(school_{H*}) = \frac{8}{19} \approx 0.421$.

The frequency distribution of the word-accent pairs is a typical LNRE-distribution (large number of rare events), where few types occur often and many types occur rarely. Due to this fact, the datasets are highly unbalanced: Firstly, the number of tokens that are analysed per type varies. Secondly, splitting the data into frequency bins, i.e. grouping types with the same number of tokens together in one bin, reveals that the number of types per frequency bin varies, as well. Therefore several data reductions have been carried out with to balance the data.

In the first reduction, those types were excluded that were the only ones occurring within a certain frequency bin. The datasets that are modified in such a way are referred to as *HN-mod* and *LHNmod* in the following.

In a second step, a balanced dataset for nuclear H* (*HN-bal*) was created, where some of the low frequency types (two or three tokens) were randomly excluded so that the number of low frequency types and higher frequency types (4 or more tokens) was the same. This was done to prevent the many low-frequency tokens from outweighing effects of the fewer tokens with higher frequency. The statistical analysis for the random set was validated in a 100-fold cross-validation. For L+H* no such analysis could be carried out, since this set would have consisted of only 32 types, and would thus have lacked sufficient statistical explanatory power.

The last data reduction equals the number of tokens that are analysed per type as well as the number of types that are analysed per frequency bin. Two tokens per type were randomly selected, as were 10 types per frequency bin. This reduced the data to only 196 types (and 392 tokens). Only types from the frequency bin 2-6 went into the analysis, since there were not enough types in the higher frequency bins. The analyses carried out on this drastically reduced random dataset (referred to as *HNequ*) were repeated 1000 times.

4. Determination of pitch accent shape

A parametric intonation model, known as PaIntE [9], was employed to represent pitch accent shape using a small number of linguistically meaningful parameters. The model approximates stretches of F_0 by employing a phonetically motivated model function [9]. This function operates on a three-syllable-window, i.e. the span of the accented syllable and the syllables adjacent to it, if they are in the same intonation phrase. The function is composed by addition of two sigmoids (rising and falling) with a fixed time delay which is selected so that the peak does not fall below 96% of the function's range.

Six parameters, illustrated in figure 1, are used to describe the contour: parameter *b* locates the peak of the accent within the three-syllable window, parameters *c1* and *c2* model the ranges of the rising and falling slope of the accent's contour, *d* corresponds to the actual height of the peak and parameters *a1* and *a2* (not displayed in the figure) denote the "amplitude-normalised" steepness of the rising and falling slope [15].

Four of the PaIntE parameters were employed in the analyses: parameters *c1* and *c2* to determine the ranges of an accent's falling or rising slope, respectively, *b* to analyse the accents' temporal alignment and *d* to measure the height of the accent's peak. The *a* parameters were excluded from the analysis because in cases where only one sigmoid is used, the *a* value of the other sigmoid is meaningless, since there is no slope, and would interfere with the analysis. For *c* parameters, this is not the case, because the value for the sigmoid that is not used, is set to 0 which reflects the actual properties of the accent. To normalise for speaker differences the PaIntE parameters were z-scored for each speaker separately.

5. Calculation of pitch accent variability

To calculate a measure of variability, each of the accent tokens was represented as a 4-dimensional vector (one dimension for each z-scored PaIntE parameter). For a given word type and a given accent, the average distance between the tokens of a word-accent pair in the 4-dimensional space was calculated. This was done by calculating a centroid (i.e. a vector composed by the mean value of each dimension) and then calculating the Euclidean distance d between the centroid $\bar{x}$ and x_i:

$$d(x_i, \bar{x}) = \sqrt{\sum_{j=1}^{4} (x_{i_j} - \bar{x}_j)^2} \tag{2}$$

The average distance of all tokens of a type to their centroid gives a measure of the variability of the type. For example, in the case of the word "school" in the *HN* dataset, the centroid

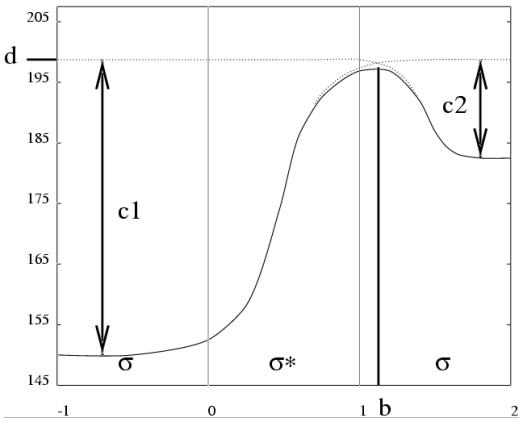

Figure 1: The PaIntE model function is the sum of a rising and a falling sigmoid with a fixed time delay. The time axis is normalised to the syllables' lengths. The parameters are calculated over the span of the accented syllable (starred) and its immediate neighbours (if the accented syllable is not followed/preceded by an intonation phrase break).

Dataset	*HN*	*LHN*	*HNmod*	*LHNmod*	*HNbal*	*HNequ*
p-value	< 0.01	< 0.05	< 0.01	< 0.05		
coefficient	-0.2967	-0.3804	-0.2901	-0.3637		
std. error	0.0918	0.1591	0.0922	0.1591		
repetitions					100	1000
significance					79%	7%
tendency					7%	5%

Table 1: Overview of the significance of the linear regression models for the tested datasets. For datasets that were created by random selection of subsets of the data, the number of repetitions is given as well as the proportion of significant cases ($\alpha = 0.05$) and of cases that showed a tendency ($p < 0.08$).

represents eight instances of "school-H*", and the average distance of the eight instances to this centroid is interpreted as the variability value for the type "school-H*".

6. Results

Linear regression models were fitted to assess the relationship between the relative frequency of a word-accent pair and the variability of the tokens of this type(there were no other variables in the models). This was done for all the above mentioned datasets (section 3). Table 1 gives an overview of the significance of the models. For both analysed accent types, the regression models for the complete dataset (*HN* and *LHN*) yielded a significant p-value indicating a correlation between the relative frequency of a word type occurring with the respective accent and the variability among the tokens of this type. This effect held also over the modified datasets *HNmod* and *LHNmod* where extremely frequent types, that were the only ones in their frequency bin, were removed. Figure 2 illustrates the effect for *HN*, i.e. it depicts the variability of tokens of a word-H* pair plotted against the relative frequency of this pair. Hence, each point in the graph represents a type (e.g, "school – H*"). The regression line illustrates a decrease of variability with increasing relative frequency. The same effect can be observed for L+H*: the greater the relative frequency of a word type and

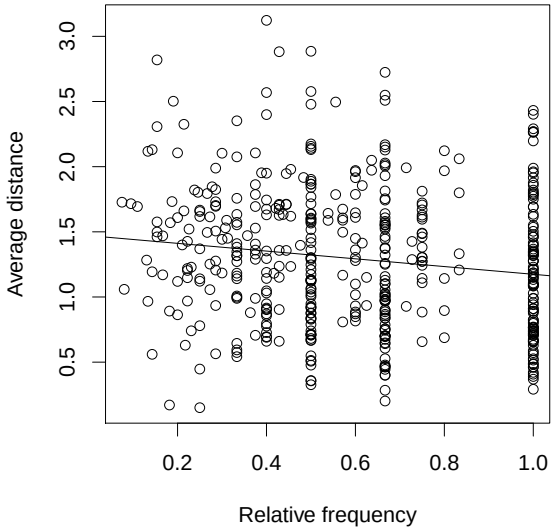

Figure 2: Relationship between the relative frequency of word-H* pairs and the variability among the tokens of each of the types.

L+H*, the lower the variability of L+H* tokens of this type. This is shown for *LHN* in figure 3.

For H*, where it was possible to balance the dataset and still keep a reasonable number of tokens, the random selection of a smaller number of low-frequency types and the subsequent calculation of the model was done 100 times. It yielded significance in 79 of 100 repetitions. In all the significant cases, variability of the types decreased with increasing relative frequency.

For the fully equalized set, however, the effect did not seem to hold. Since the selection of a subset of the original dataset involved two randomisation processes, the number of repetitions was increased to 1000 repetitions. The regression model was significant in only 66 cases; though in those cases the effect was the same (decreasing variability for increasing relative frequency). However, this dataset is reduced drastically, compared to the original set (cf. section 3). While the original *HN* dataset comprises 1425 tokens of 465 types with tokens ranging in their frequency between two and 58 (though only one type is as frequent as this), the reduced dataset *HNequ* consists only of 196 types and only two instances of each type went into the analysis. Moreover, the types ranged in their frequency only between two and six, i.e. the higher frequency types were not analysed. It is therefore debatable, whether such a drastic reduction of the data still represents the original dataset in an appropriate way.

7. Discussion and Outlook

The analyses described above indicate a relationship between the relative frequency with which a word occurs with a certain pitch accent and the variability of the realisations of this accent. The greater the relative frequency of a word-accent pair, the less variable are the realisations of the accent.

Such an effect presents a challenge for traditional, autosegmental theories of intonation. These theories are in general silent on the effect of frequency on pitch accent type and re-

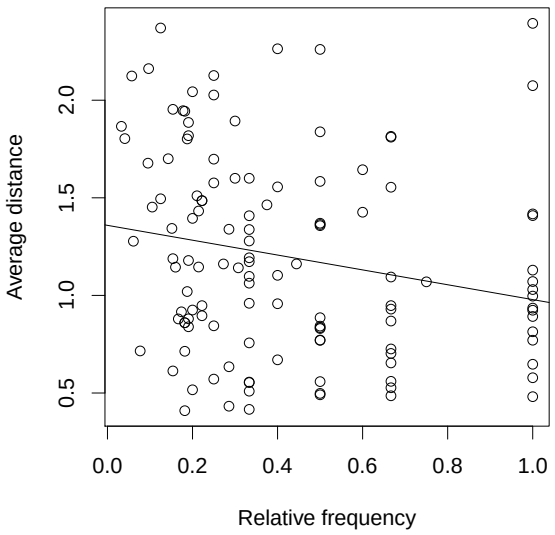

Figure 3: Relationship between the relative frequency of word-L+H* pairs and the variability among the tokens of each of the types.

alisation. However, once a particular accent type has been assigned to a given word, it is assumed that the realisation of the pitch contour on that word is related purely to the phonological context, e.g. the position in the phrase and how near other accents are [16]. While these factors are undoubtedly still relevant, our results seem to show that accent-word frequency also plays a part in explaining pitch contour variation. Along with our previous results (see section 1), these findings suggest that pitch contour realisation cannot be considered to be purely a "post-lexical" process in English.

An exemplar-theoretic view of pitch accenting, on the other hand, expects word-based storage of accent contours, and resultant frequency effects. An exemplar-theoretic explanation for this phenomenon is as follows: In production, exemplars of the target word are activated. If more of these exemplars come from different contexts (e.g. are accented with pitch accents other than the intended one), the F_0-shapes from which the production target is derived will be more variable, because they include not only realisations with the intended accent but also realisations with competing pitch accents.

The competing realisations are expected to cause "noise" in the production of the intended accent. Hence, the larger the proportion of competing pitch accents, the greater the variability among the realisations of the target accent. In other words, accent-word pairs with a small relative frequency (more competitors) are expected to be more variable than accent-word pairs with a high relative frequency (less competitors).

For word types that have a high relative frequency with one accent type, this result is consistent with the proposal in [8] that certain words have "fixed" intonation, linked to specific pragmatic meanings of that word, e.g. discourse expressions like "really".

Our analyses demonstrate an effect of relative frequency on the realisation of pitch accents. However it is possible that the changes in variability result from an imbalance in dataset in terms of number of tokens and types. To rule this out, we would need to repeat the analysis on a much larger dataset. Unfortunately, we are not aware of a substantially larger corpus annotated for accent type. However, in the future we are looking to adapt our methodology for larger available corpora which have some prosodic annotation (not including accent type).

8. Acknowledgements

This work was supported by the German Research Foundation (Collaborative Research Center SFB-732) and a British Academy Postdoctoral Fellowship to S. Calhoun.

9. References

[1] J. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition and contrast," in *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper, Eds., 2001, pp. 137–157.

[2] J. Bybee, "From usage to grammar: The mind's response to repetition," *Language*, vol. 84, pp. 529–551, 2006.

[3] R. M. Nosofsky, "Attention, similarity, and the identification-categorization relationship," *Journal of Experimental Psychology: General*, vol. 115, no. 1, pp. 39–57, 1986.

[4] S. D. Goldinger, "Words and voices—perception and production in an episodic lexicon," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 33–66.

[5] A. Schweitzer and B. Möbius, "Exemplar-based production of prosody: Evidence from segment and syllable durations," in *Speech Prosody 2004 (Nara, Japan)*, 2004, pp. 459–462.

[6] M. Walsh, H. Schütze, B. Möbius, and A. Schweitzer, "An exemplar-theoretic account of syllable frequency effects," in *Proceedings of ICPhS (Saarbrücken)*, 2007, pp. 481–484.

[7] K. Schweitzer, M. Walsh, B. Möbius, A. Riester, A. Schweitzer, and H. Schütze, "Frequency Matters: Pitch accents and Information Status," in *Proceedings of EACL-09*, Athens, Greece, 2009.

[8] S. Calhoun and A. Schweitzer, "Can intonation contours be lexicalised? Implications for Discourse Meanings." in *Prosody and Meaning (Trends in Linguistics)*, G. Elordieta and P. Prieto, Eds. De Gruyter Mouton, subm.

[9] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Third Intern. Workshop on Speech Synth (Jenolan Caves)*, 1998, pp. 311–316.

[10] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego: Academic Press, 1997, pp. 145–165.

[11] S. P. Whiteside and R. A. Varley, "Dual-route phonetic encoding: Some acoustic evidence," in *Proceedings of ICSLP (Sydney)*, vol. 7, 1998, pp. 3155–3158.

[12] E. Marsi, M. Reynaert, A. van den Bosch, W. Daelemans, and V. Hoste, "Learning to predict pitch accents and prosodic boundaries in dutch," in *Proceedings of the ACL-2003 Conference (Sapporo, Japan)*, 2003, pp. 489–496.

[13] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Electrical, Computer and Systems Engineering Department, Boston University, Boston, MA, Tech. Rep. ECS-95-001, 1995.

[14] K. Silverman, M. Backman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for Labeling English Prosody," in *Proceedings of the 1992 International Conference on Spoken Language Processing (Banff, Kanada)*, vol. 2, Banff, Canada, 1992, pp. 867–870.

[15] G. Möhler, "Improvements of the PaIntE model for F_0 parametrization," Institute of Natural Language Processing, University of Stuttgart, Tech. Rep., 2001, draft version.

[16] D. Ladd, *Intonational Phonology*, 2nd ed. Cambridge, UK: Cambridge University Press, 2008.