interspeech 2008 incorporating SST 08

Examining pitch-accent variability from an exemplar-theoretic perspective

Michael Walsh, Katrin Schweitzer, Bernd Möbius, Hinrich Schütze

[1]Institute for Natural Language Processing
[2]University of Stuttgart, Germany
{michael.walsh,kati.schweitzer,bernd.moebius,hs0711}@ims.uni-stuttgart.de

10.21437/Interspeech.2008-280

Abstract

This paper presents an exemplar-theoretic account of pitch-accent variability. Results from two experiments are reported. The first experiment examines variability / similarity across a range of pitch-accent contour types which vary in frequency of occurrence. The results report similar behaviour across the frequency bins and are indicative of pitch-accent immunity from frequency effects. The second experiment, in order to establish the impact of lexical frequency on pitch-accent production, examines the similarity of pitch-accent contours linked to high and low frequency syllables. The results indicate further autonomy on the part of pitch-accents and offer possible evidence for post-lexical pitch-accent generation.

Index Terms: speech production, exemplar theory, pitch-accents

1. Introduction

Exemplar Theory has successfully accounted for a number of language phenomena, including diachronic language change [1], the emergence of grammatical knowledge [2], syllable duration variability [3, 4], entrenchment and lenition [5], among others. Unlike more traditional, often generative, rule-oriented approaches, at the core of Exemplar Theory is the idea that the acquisition of language is significantly facilitated by repeated exposure to concrete language input. Key elements of Exemplar Theory are the notions of storage, frequency, recency, and similarity. There is an increasing body of evidence which indicates that significant storage of language input exemplars, rich in detail, takes place [7, 8, 9]. These stored exemplars are then employed in the categorisation of new input percepts. Similarly, production is facilitated by accessing these stored exemplars. 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 [5].

To date, however, little if any exemplar-theoretic research has examined pitch-accent prosody (but see [6] for memory-based prediction of pitch-accents and prosodic boundaries). The experiments presented here aim to address this by examining how pitch-accents vary with respect to frequency of occurrence. Two hypotheses motivate these experiments. Firstly, given the considerable body of evidence for the role of frequency in a variety of other linguistic domains one might expect significant frequency effects upon pitch-accent production. Secondly, if pitch-accent production is autonomous and independent of the lexicon, then an examination of pitch-accent on the basis of lexical frequency should yield limited, if any, frequency effects. However, an exemplar-theoretic account would suggest that perceived lexical items could be stored with the accompanying perceived pitch-accent, in which case pitch-accent might not always operate in an autosegmental fashion.

The pitch-accents employed in the experiments below are represented in terms of a number of parameters. The Parametric Representation of Intonation Events (PaIntE) model is employed to approximate the F_0 contours of GToBI labelled pitch-accents [10]. This model facilitates dimensionality reduction and analysis of particular properties (e.g. rising amplitude) of the F_0 contour, which a raw F_0 analysis would not permit so readily. The next section gives a brief introduction to the PaIntE model. This is followed by section 3 which examines the first hypothesis and presents an investigation into pitch-accent variability with respect to pitch-accent frequency of occurrence. Section 4 then examines the second hypothesis and presents further analysis, this time examining pitch-accent variability when associated with tokens of frequent and infrequent syllable types. Overall conclusions and opportunities for future research are presented in sections 5 and 6 respectively.

2. The Parametric Representation of INTonation Events - PaIntE

The PaIntE model approximates stretches of F_0 by employing a phonetically motivated model function [10]. This function consists of the sum 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 functions range. The resulting function has six parameters which describe the contour and were employed in the analysis: parameters a1 and a2 express the gradient of the accent's rise and fall, parameter b describes the accents temporal alignment (which has been shown to be crucial in the description of an accent't shape [11, 12]), c1 and c2 model the ranges of the rising and falling amplitude of the accent's contour, respectively , and parameter d expresses the peak height of the accent. These six parameters are thus appropriate to describe different pitch accent shapes. The model's parameters are illustrated in Fig. 1.

3. Examining pitch-accent variability.

In order to establish whether or not high numbers of pitch-accent exemplars affect subsequent productions, a number of experiments were carried out on pitch-accents found in a German radio news corpus of approximately 1 hour of speech [13]. This corpus has been automatically segmented and manually labelled according to the GToBI(S) annotation schema [14]. The first experiment, detailed below, was performed on tokens of each of the following pitch-accent types (in descending order of frequency): L*H: 1233 tokens, H*L: 704 tokens, H*: 162 tokens, and L*HL: 70 tokens.

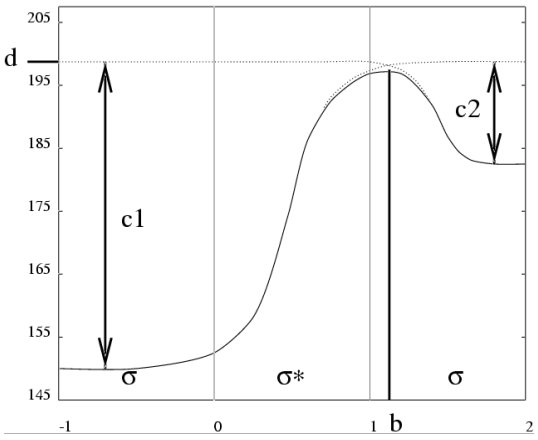

Figure 1: *The PaIntE Model. This is Fig. 1 from [15]. The model function is the sum of a rising and falling sigmoid and the time axis is normalised to the syllable's length.*

3.1. Experiment 1.

The purpose of this experiment is to establish the extent to which frequency of occurrence might affect pitch-accent production. For tokens of the pitch-accent types listed above, each token was modelled using the full set of PaIntE parameters. Thus, each token was therefore represented in terms of a 6-dimensional vector. Vectors which contained outliers (dimensional values which fell outside the whiskers of box-and-whiskers plots) were removed. This led to some reduction in the data set, namely: L*H: 1006 token vectors; H*L: 526 token vectors; H*: 118 token vectors; and L*HL: 61 token vectors. The parameters are computed over the span of the accented syllable and its immediate neighbours (see Fig. 1). For each of the pitch-accent types the following steps were carried out:

- For each 6-dimensional pitch-accent category token calculate the z-score value for each dimension. The z-score value represents the number of standard deviations the value is away from the mean value for that dimension and allows comparison of values from different normal distributions. The z-score is given by:

$$Dim_{z-score} = \frac{Dim_{value} - Dim_{mean}}{Dim_{standard-deviation}} \quad (1)$$

  Hence, at this point each pitch-accent is represented by a 6-dimensional vector where each dimension value is a z-score.

- For each token z-scored vector calculate how similar it is to every other z-scored vector within the same pitch-accent category using the cosine of the angle between the vectors. This is given by:

$$cos(\vec{i}, \vec{j}) = \frac{\vec{i} \bullet \vec{j}}{\| \vec{i} \| \| \vec{j} \|} \quad (2)$$

  where *i* and *j* are vectors of the same pitch-accent category and • represents the dot product. Each comparison between vectors yields a similarity score in the range [-1,1], where -1 represents high dissimilarity and 1 represents high similarity.

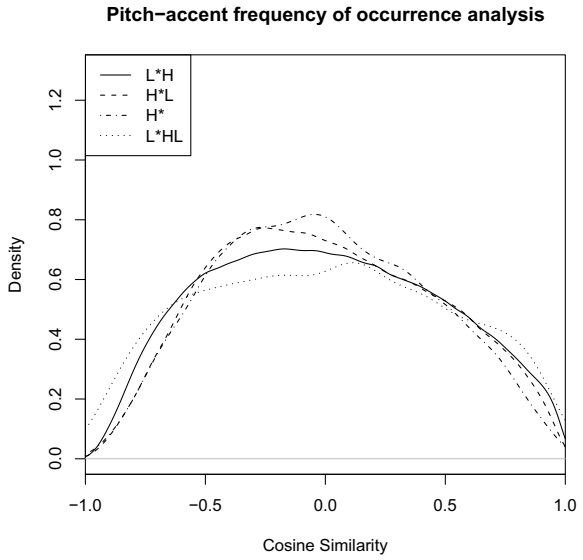

Figure 2: *Density plot of similarity values for 4 pitch-accents.*

3.2. Results 1.

The results of the z-scored vector similarity comparisons for each pitch-accent type are plotted in Fig. 2. Each density plot indicates how similar pitch-accents of the same type are to each other. Before analysing the graph it is important to note that pairwise comparison of each contour using the (two-sample) Kolmogorov-Smirnov test yielded statistically significant differences in each case ($p < 0.05$ in all cases). However, while the differences in distributions appear to rest between 0.0 and -0.5, Fig. 2 would nevertheless seem to indicate that pitch-accent frequency of occurrence has little, if any, effect on pitch-accent production. Similarity comparison scores within high frequency L*H productions are distributed in a similar fashion to those within the low frequency L*HL productions. Nor is there any interesting deviation within those extremes; H*L and H* behave in much the same way. Why is this ?

Exemplar Theory provides evidence for storage of words and phrases rich in phonetic detail [7, 8, 9, 16]. It seems plausible that such storage might also include pitch-accent information. For example, the German word "oder" (meaning "or") is very often placed in sentence-final position and produced with a rising F_0 contour to indicate a question rather than a disjunction. It seems likely and intuitive, from the perspective of Exemplar Theory, that this word would be stored, with this interrogative intonation, as an exemplar in memory, rather than have the word be incrementally augmented with the pitch-accent during the production phase. However the expectations of Exemplar Theory, as it currently stands, are not clear. On the one hand Bybee [1] notes that high frequency sequences undergo processes of entrenchment. If this is always the case one might expect to see greater levels of similarity for the high frequency L*H curve in Fig. 2, i.e. a peak towards the right hand side of the graph due to entrenched behaviour. However, on the other hand, Schweitzer and Möbius [3] found that high frequency syllable productions yielded increased duration variability. Interestingly here neither entrenchment nor increased variability occur. One possible explanation of this uniformity of behaviour is that while pitch-

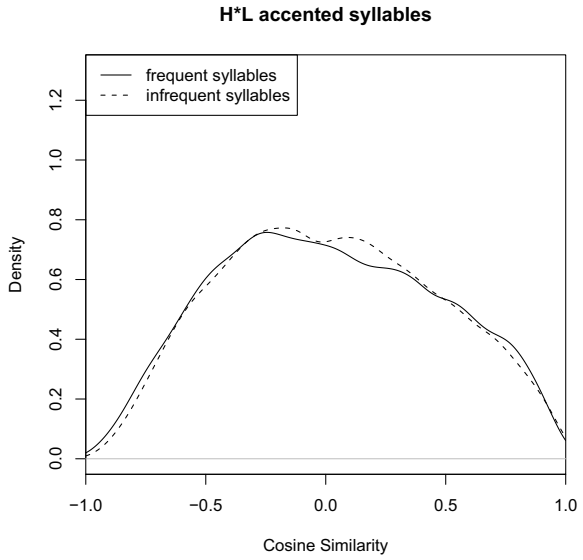

Figure 3: *Similarity of H*L accents occurring on frequent and infrequent syllables.*

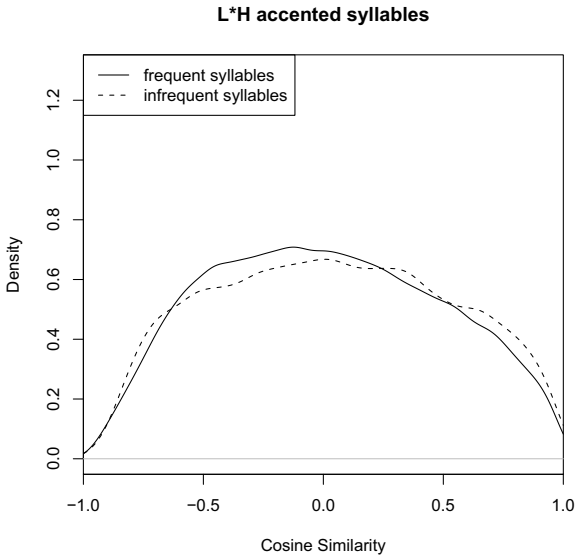

Figure 4: *Similarity of L*H accents occurring on frequent and infrequent syllables.*

accents might well be stored in line with exemplar-theoretic expectations, stored exemplars might only be used for the purposes of perception (e.g. speaker identification, detection of questions, sarcasm, etc.). Pitch-accent information for production however could be added incrementally, i.e. after lexical access (perhaps exemplar access) has taken place. This would be in keeping with Levelt et al.'s theory of lexical access in speech production [17]. If this indeed is the case then *lexical* frequency should have no impact on pitch-accent behaviour either. Section 4 below seeks to establish whether lexical frequency has any effect.

4. Examining pitch-accent autonomy

In order to discern the impact of lexical frequency on pitch-accent realisations, frequent and infrequent syllables which were produced with H*L and L*H accents were extracted from the same corpus used in Experiment 1. Analysis by Müller et al. [18] was employed to define criteria for syllable frequency and infrequency; the criteria being that high frequency syllables have a probability of occurrence in excess of 0.001, and low frequency syllables have a probability less than 0.00005. These syllable frequency categorisation criteria were based on syllable probabilities induced from multivariate clustering and have been successfully employed in previous exemplar-theoretic research into syllable duration variability [3, 4]. This process yielded the following numbers of tokens per pitch-accent and frequency-bin types:

- H*L - 167 frequent vs. 149 infrequent syllable tokens
- L*H - 308 frequent vs. 231 infrequent syllable tokens

Note that for both pitch accent types there are a similar number of syllable tokens in each syllable frequency bin, in particular for H*L bearing syllables. This is due to a large number of infrequent syllable types having a small number of tokens each, and a smaller number of frequent syllable types having a larger number of tokens each. Given these data sets the following experiment was performed firstly on H*L accented syllables (frequent and infrequent), and then on L*H accent syllables. It is also important to note that since the number of tokens across the syllable frequency bins is similar, and the pitch-accent type is the same, any effects observed must have something to do with syllable frequency and not merely pitch-accent frequency.

4.1. Experiment 2.

This second experiment operates in a similar fashion to Experiment 1. For each pitch-accent token of a particular type (H*L or L*H) realised on a given syllable token of a particular frequency bin (frequent or infrequent) do the following:

- As in Experiment 1 calculate the z-score value for each dimension so that the pitch-accent is represented by a z-score vector. It is important to note that the syllable frequency bins for the pitch-accent under investigation are combined to allow calculation of the mean value, from which the standard deviation and z-score are calculated.
- Using the cosine of the vectors determine how similar each possible pair of pitch-accent vectors within a syllable frequency bin are to each other.

4.2. Results 2

Figures 3 and 4 present density plots of cosine similarities found for both pitch-accent types and syllable frequency bins. In Fig. 3 plots similar to those in Fig. 2 can be seen; H*L accents realised on low frequency syllables and high frequency syllables appear to have very similar distributions. The Kolmogorov-Smirnov test here was not significant ($p = 0.0544$). Here, as in the pitch-accent types in Experiment 1, there appears to be little effect of frequency, this time lexical frequency, on the behaviour of the pitch-accent contours. A Kolmogorov-Smirnov test on the results in Fig. 4, examining L*H accents, was significant ($p < 0.001$). However, as with H*L pitch-accents on syllables, the overall behaviour of L*H appears to be indepen-

dent of lexical frequency.

5. Discussion

At the outset, the goal of this research, motivated by a growing body of exemplar-theoretic evidence from a variety of linguistic domains, was to establish what extent, if any, frequency of production affects pitch-accent realisations, and to determine whether or not lexical frequency also has a role to play.

The results from Experiment 1 and 2 are interesting in that they neither support predictions of entrenchment nor variability, both of which are exemplar-theoretic expectations. Indeed, both experiments would indicate that pitch-accent realisations appear to be largely immune to frequency effects and operate autonomously or autosegmentally (as is often noted in the literature [19, 20]). This might well be an argument for post-lexical generation of pitch-accent such as is found in Levelt et al.'s hierarchical, feed-forward model [17]. However, the increasing weight of evidence behind Exemplar Theory might recommend a hybrid approach where pitch-accent exemplars are used for perception only.

6. Future Work

In addition to the results discussed above, this research also opens up a number of avenues for future research. For example, given the parametric nature of the PaIntE model, what effects might analysis on the basis of single dimensions yield? It is possible that while the overall 6-dimensional distributions of pitch-accents might not conform to exemplar-theoretic expectations, single dimensions e.g. temporal alignment, might well behave as predicted. In addition, it might also be worth investigating articulatory dimensions such as those proposed by [21, 22, 23]. A further question which arises concerns the possibility of a hybrid model, whereby pitch-accent exemplars are perceived and stored, but are only used for recognition purposes (given that the results do not support exemplar-based pitch-accent production). If this is the case, then evidence for pitch-accent exemplar storage, which is currently lacking, needs to be established.

7. Acknowledgements

This research was funded by the German Research Council (DFG, Grant SFB 732).

8. References

[1] Bybee, J., "From usage to grammar: the minds response to repetition", Language,84:529–551,2006.

[2] Abbot-Smith, K. and Tomasello, M.,"Exemplar-learning and schematization in a usage-based account of syntactic acquisition", The Linguistic Review, 23:275–290,2006.

[3] Schweitzer, A. and Möbius, B., "Exemplar-Based Production of Prosody: Evidence from Segment and Syllable Durations", Proceedings of the Speech Prosody 2004 Conference,459–462, Nara, Japan.

[4] Walsh, M., Schütze, H., Möbius, B. and Schweitzer, A., "An Exemplar-Theoretic Account of Syllable Frequency Effects", Proceedings of the Sixteenth International Congress of Phonetic Sciences (ICPhS 2007),481–483, Saarbrücken, Germany.

[5] Pierrehumbert, J.B., "Exemplar dynamics: Word frequency, lenition and contrast", Frequency effects and emergent grammar, John Benjamins, Amsterdam, 2001.

[6] Maris, E., Reynaert, M., van den Bosch, A., Daelemans, W. and Hoste, V.,"Learning to predict pitch accents and prosodic boundaries in Dutch", Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2993), 489–496, Sapporo, Japan.

[7] Johnson, K., "Speech perception without speaker normalization: An exemplar model", Talker Variability in Speech Processing, Academic Press, San Diego, 1997.

[8] Croot, K. and Rastle, K.,"Is there a syllabary containing stored articulatory plans for speech production in English?", Proceedings of the 10th Australian International Conference on Speech Science and Technology,376–381, Sydney, Australia, 2004.

[9] Whiteside, S.P. and Varley, R.A.,"Dual-route phonetic encoding: Some acoustic evidence", Proceedings of the 5th International Conference on Spoken Language Processing,3155–3158, Syndey, Australia, 1998.

[10] Möhler, G., "Describing intonation with a parametric model", Proceedings of the International Conference on Spoken Language Processing, 7:2851–2854, 1998.

[11] van Santen, J. and Möbius, B., "A quantitative model of F0 generation and alignment", In Antonis Botinis (ed.) Intonation - Analysis, Modelling and Technology,269-288, Kluwer, Dordrecht, The Netherlands, 2000.

[12] Kohler, K.J.,"Macro and micro F0 in the synthesis of intonation", In Kingston, J. and Beckman, M.E., eds, Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech, Cambridge University Press, 115–138,Cambridge,UK,1990.

[13] Rapp, S., "Automatisierte Erstellung von Korpora für die Prosodieforschung", PhD Thesis., Universität Stuttgart, 1998.

[14] Mayer, J., "Transcribing German Intonation – The Stuttgart System", Technical Report, Universität Stuttgart, 1995.

[15] Möhler, G., and Conkie, A., "Parametric modeling of intonation using vector quantization", Proceedings of 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia.

[16] Hay, J. and Bresnan, J.,"Spoken Syntax: The phonetics of giving a hand in New Zealand English", The Linguistic Review,23,2006.

[17] Levelt, W.J.M. and Roelofs, A., and Meyer, S.A.,"A theory of lexical access in speech production", Behavioral and Brain Sciences,22,1–75,1999.

[18] Müller, K., Möbius, B. and Prescher, D., "Inducing Probabilistic Syllable Classes Using Multivariate Clustering", Proceedings of the 38th meeting of the Association of Computational Linguistics,225–232, Hong Kong, 2000.

[19] Ladd, R.D., "Intonational Phonology", Cambridge University Press,Cambridge, UK, 1996.

[20] Pierrehumbert, J., "Tonal elements and their alignments", In M. Horne (ed), Prosody: Theory and Experiment—Studies Presented to Gösta Bruce, Kluwer, Dordrecht, The Netherlands, 2000.

[21] Fujisaki, H., Wang, C., Ohno, S. and Gu, W., " Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model", Speech communication,47: 59-70, 2005.

[22] Xu, Y.,"Speech melody as articulatorily implemented communicative functions", Speech Communication 46: 220-251, 2005.

[23] Kochanski, G. and Shih, C.,"Prosody modeling with soft templates",Speech Communication 39: 311–352, 2003.