A HIERARCHY OF PROMINENCE: THE PRODUCTION AND PERCEPTION OF FOCUS IN AMERICAN ENGLISH

Argyro Katsika[1], Jiyoung Jang[1], Jelena Kriivokapić[2], Louis Goldstein[3] and Elliot Saltzman[4]

[1]University of California, Santa Barbara
[2]University of Michigan
[3]University of Southern California
[4]Boston University

argyro@ucsb.edu, jiyoung@ucsb.edu, jelenak@umich.edu, louisgol@usc.edu, esaltz@bu.edu

ABSTRACT

Pitch accents on stressed syllables mark prominent words. Moreover, articulatory gestures are longer, larger and faster when accented. It is also known that prominence marking interacts with focus structure. However, it is still unclear how many degrees of prominence are encoded phonetically and whether these are perceived. Here, we assess these issues via an electromagnetic articulography study of American English and a corresponding crowd-sourced online perception survey.

Results show that phonetic dimensions (acoustic F0; and kinematic duration, displacement and velocity) differ in the number of degrees of prominence they distinguish. Nonetheless, a hierarchical ordering of the reflected degrees remains consistent across dimensions. Listeners' judgements reflect this: Degrees of prominence are more successfully distinguished by listeners the further away they are from each other hierarchically; neighbouring degrees are distinguished more poorly the lower they are in the hierarchy. An account for this hierarchy is discussed with the framework of Articulatory Phonology.

Keywords: Prominence, focus, prosody, production-perception relationship, Articulatory Phonology.

1. INTRODUCTION

Phrasal pitch accents mark words within phrases as rhythmically or conceptually important, and they are associated to the stressed syllables of the accented words (e.g., [1, 2]). The articulatory movements, called *gestures*, forming the consonant (C) and vowel (V) constrictions in stressed syllables have in turn been found to become longer, larger and faster when accented [3, 4, 5, 6, 7, 8, 9, 10, 11, 12; but see 13]. This spatiotemporal expansion that gestures undergo is often referred to as *prosodic strengthening* (cf. [6]). Later work has presented evidence that it is focus structure and not simply accentuation that causes strengthening ([14, 15, 16] for German; see also [17] for English). This line of research considered different types of focus (i.e., broad focus, narrow focus, contrastive focus, unfocused) and found that phonetic dimensions, such as acoustic F0 and kinematic duration and displacement, increased across focus types. Although there might be typological differences and the set of dimensions used might differ on a language-dependent basis, phrase-level prominence emerges from these findings not only as a multi-dimensional (i.e., employing multiple phonetic dimensions) but also as a multi-level system, which goes beyond the distinction between accented and unaccented. Of course, the presence of a pitch accent automatically denotes an accented vs. unaccented split.

Taken together, the conclusions of previous research suggest three hypotheses: 1) Prominence is organized hierarchically, with levels of the hierarchy possibly representing focus types. However, it is unclear whether all focus types are encoded, and if yes, in what order, since, to date, neither the complete range of degrees of prominence nor the phonetic correlates of these degrees have been established. 2) Whichever set of phonetic parameters a language uses for marking phrase-level prominence, this multi-dimensional system functions as a whole, i.e., with all the phonetic parameters being modified in tandem. Interestingly, [15] presents a dynamical model that captures the observed focus-induced modifications on multiple phonetic dimensions in German by controlling the same, single parameter. 3) Assuming that prominence encodes aspects of information structure, playing thus a significant role in speech comprehension, and given the abundance of phonetic cues used in this encoding, we expect that any degrees of prominence represented phonetically should also be perceived by listeners in their produced order.

Here, we take a first step towards addressing these hypotheses by the means of an electromagnetic

articulography study of American English and a corresponding crowd-sourced online perception survey. By examining both production and perception, a wide range of focus types (i.e., contrastive focus, narrow focus, broad focus, deaccented and unfocused) and multiple phonetic parameters known to be used in prominence marking in American English (acoustic F0; and kinematic duration, displacement and velocity), the current study ultimately aims at examining the hierarchical structure for phrase-level prominence.

2. METHODS

2.1. Production study

The production data from [17] were retrieved and analyzed for this study, and the reader is referred to that study for a full description of the experimental design and procedure. Here, a brief summary of these aspects is offered.

2.1.1. Participants and recording apparatus

There were eight native speakers (2 male, 6 female; mean age: 23) of American English to this study. Kinematic data were collected using the AG501 3D electromagnetic articulograph (EMA; Carstens Medizinelektronik). Receiver coils were attached to the tongue dorsum, tongue body's center, tongue tip, upper and lower lips, upper and lower incisors, left and right ears, and nose. Audio data were collected simultaneously at a sampling rate of 16 kHz.

2.1.2. Stimuli and experimental design

The following nine English words were used as test words: *bee*, *baby*, *melody*, *military*, *design*, *banana*, *humanity*, *matinee*, and *salmonella*. Test words were placed in frame sentences appropriate for their meaning, which were in turn paired with prompt sentences designed to elicit the test words in five focus types: 1) unfocused, following a narrowly focused item (UF), 2) de-accented by virtue of following a contrastively focused item (DA), 3) accented under broad focus (BF), 4) accented under narrow focus (NF), or 5) accented under contrastive focus (CF). To illustrate, a pair of prompt-target sentences for the test word *bee* in the de-accented condition is given in (1).

(1) **Prompt:** *Is it the zoologist who will be testing the bee with the stripes?*

Target: *No, the botanist will be testing the bee with the stripes*

The stimuli were presented on a monitor. The participant read the prompt sentence silently, and the target sentence aloud. Eight blocks of the stimuli were recorded; stimuli in each block were presented in a different random order.

2.1.3. Data analysis

Custom software (Tiede, Haskins Laboratories) was used to semi-automatically detect the C gesture of the stressed syllables of the test words and to label important timepoints in their kinematic progression on the basis of velocity criteria (see Figure 1). Labial and coronal C gestures were labeled on the lip aperture and tongue tip vertical displacement trajectory respectively.

The following timepoints are relevant for the current analyses: gestural onset, peak velocity of constriction's forming movement, and constriction's target (i.e., achievement). Based on these timepoints, the measures of (a) maximum displacement (as the absolute difference between constriction's target and gestural onset) and (b) peak velocity of the formation were calculated. Formation duration (i.e., the interval from onset to release) is not directly reported here, since it is analyzed in [17].

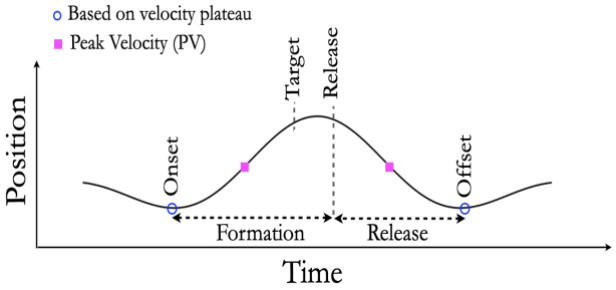

Figure 1: Schematic representation of a constriction gesture and labelled timepoints.

Analysis of acoustic F0 is underway and currently includes the subset of the data that served as the stimuli for the perception study. Stressed vowels were segmented in Praat [18], with their acoustic boundaries being determined at the onset and offset of F2. Maximum F0 at the midpoint of the vowel was automatically extracted using a Praat script.

Statistical analyses were performed by linear mixed effects models using the lme4 package [19] in R [20]. For the dependent variables of displacement, peak velocity, and F0, the fixed effect of Focus Type (UF, DA, BF, NF, CF) and the random effects of Speaker and Word were considered (i.e., *bee* and *melody* for F0; all test words for the kinematic measures). Post-hoc pairwise comparisons with Holm adjustment were assessed using emmeans [21].

2.2. Perception study

In the crowd-sourced online perception survey the fourth repetition of the stimuli for the test words *bee* and *melody* from the production study were assessed by 185 native speakers of American English (age groups: 18-25: 179 speakers; 26-30: 5 speakers; 51-65: 1 speaker). Productions from the middle of the experiment (4th block) were selected as they were considered the least susceptible to possible order effects (e.g., newness of task at the beginning or tiredness at the end). The survey was prepared and presented using Qualtrics [22]. There were eight versions of the survey, each using the data from one speaker from the production study. In each version, there were 23 participants, except for two versions that had 24 participants and one version that had 22.

As a reminder, the production study involved pairs of prompt and target sentences; the prompt was a question and the target was its response. Survey participants listened to the target sentence (produced by the speaker that corresponded to their survey's version) twice at their own pace while viewing on their screen two prompt questions. One question was the original pair in terms of focus type to the target sentence and one that was not. To avoid signaling focus type by other cues, the initial word, which was either 'no' or 'oh', from each target utterance was omitted. The task of the participant was to select the question to which the target sentence they heard was a better response. The pairs of questions the listeners heard represented all possible combinations of the five focus types, yielding 10 combinations. In total, 20 trials were included in each survey (10 focus types combinations x 2 test words x 1 speaker as source). The survey took about 15 minutes to complete. In total, 7400 judgements were acquired.

Survey responses were analyzed in terms of percentage of successful selection between focus types. Additional analyses of kinematic (i.e., duration, maximum displacement, and peak velocity of C gesture formations) and acoustic (i.e., maximum F0 of vowels) dimensions were performed to the subset of data from the production study that was also used in the perception survey. The goal was to examine correlations between selection success rate (in %) and each of these dimensions as well as correlations within each pair of phonetic dimensions. Specifically, for each focus type combination tested in the given version of the survey, we calculated the difference in value (i.e., value in target focus type minus value in competitor focus type) on each phonetic dimension between the corresponding two tokens from the production study. Correlations were analyzed using the GGally [23] package in R [20].

3. RESULTS

3.1. The effect of focus type on phonetic dimensions

The statistical analysis detected main effects of Focus Type on all phonetic dimensions examined (F0: $\chi^2(4)=134.0$, $p<0.001$; Displacement: $\chi^2(4)=78.45$, $p<0.001$; peak velocity: $\chi^2(4)=30.01$, $p<0.001$). The post-hoc pairwise comparisons are listed in Table 1.

		CF	NF	BF	DA
NF	F0	*n.s.*			
	DISP	*n.s.*			
	PV	*n.s.*			
BF	F0	0.008	*n.s.*		
	DISP	0.0001	0.07 (*m.*)		
	PV	0.09	*n.s.*		
DA	F0	<0.0001	<0.0001	<0.0001	
	DISP	<0.0001	0.0001	*n.s.*	
	PV	0.0008	0.006	*n.s*	
UF	F0	<0.0001	<0.0001	<0.0001	*n.s.*
	DISP	<0.0001	<0.0001	0.007	*n.s.*
	PV	0.0001	0.001	*n.s.*	*n.s.*

Table 1: *p* values for pairwise comparisons by Focus Type for F0, displacement (DISP) and peak velocity (PV). Non-significant (*n.s.*) and marginally significant ($0.05 < p < 0.01$) (*m.*) comparisons are indicated.

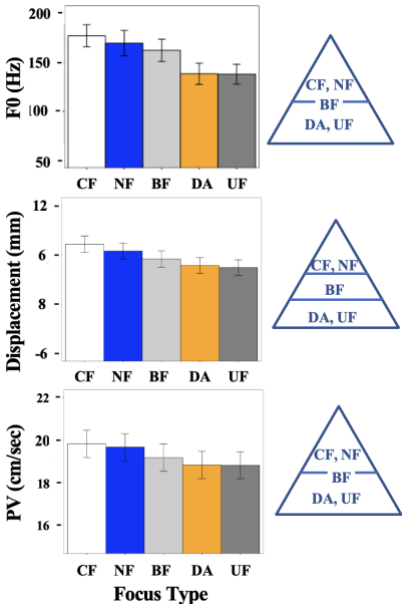

Figure 2: Mean values (with standard error) of stressed vowel's maximum F0 (Hz), and C gesture's formation displacement (mm) and peak velocity (cm/sec) by focus type (DA, UF, BF, NF, CF). A schematic representation of the focus types distinguished per phonetic dimension is juxtaposed.

As per our hypothesis (1), pairwise comparisons indicated that phonetic dimensions differentiated

primarily among focus types and not simply between presence (in all focused conditions) and absence (unaccented and de-accented) of pitch accent. However, phonetic dimensions differed in the number of focus types, and thus degrees of prominence, they distinguished. As illustrated in Figure 2, acoustic F0 and kinematic peak velocity distinguished two degrees: CF, NF > DA, UF with BF not being clearly distinct from either, and kinematic displacement distinguished three degrees: CF, NF > BF > DA, UF. Note that kinematic duration, reported for the same set of data in [17] presented four degrees: CF > NF > BF > DA, UF. Nonetheless, a hierarchical ordering of the reflected degrees remained consistent across dimensions, with CF being on the high extreme and UF on the low.

3.2. Success rate for focus type distinction

Overall, 76% of the total number of judgements successfully selected the target focus type. Figure 3 plots the distribution of successful selections across target-competitor pairs. Listeners' judgments systematically reflected the hierarchical ordering encoded phonetically. In particular, degrees of prominence are more successfully distinguished perceptually the further away they are from each other hierarchically (e.g., 85% in CF vs. DA), whereas neighbouring degrees are distinguished more poorly, especially when lower in the hierarchy (e.g., DA vs. UF is at chance).

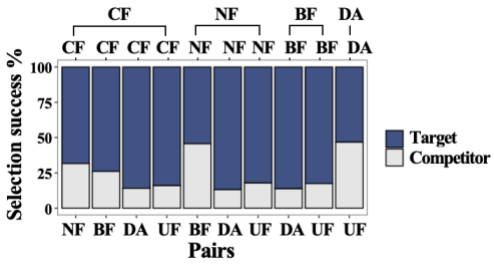

Figure 3: Listeners' selection success rate (%) across target-competitor pairs. Focus types (CF, NF, BF, DA, UF) of each member of the pair are shown on each side of the percentage bar.

3.3. Correlations

Significant correlations were found between selection success rate and all phonetic dimensions as well as within each pair of dimensions (Fig. 4). These patterns, combined with those in 3.1, suggest that phonetic dimensions are modified in tandem across the same hierarchical order of degrees, per hypothesis (2). This seems to facilitate listeners decode focus information, as in hypothesis (3), presumably because of the abundance of phonetic cues. Future work will assess the contribution of individual dimensions to this production-perception relationship.

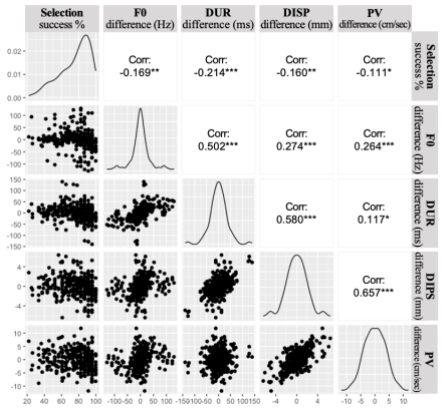

Figure 4: Correlations of selection success rate (in %) and target-competitor difference in F0 (in Hz), duration (DUR in ms), displacement (DISP in mm), and peak velocity (PV in cm/sec).

4. DISCUSSION

Results confirm that prominence encodes focus structure and not solely accentuation [14, 15, 16]. Furthermore, a hierarchical structure of prominence emerges that is reflected in perception. Different dimensions encode different levels of this structure, but this might be an epiphenomenon of different degrees of granularity of the measures used or the perceptual functionality of each phonetic dimension (e.g., velocity does not have a psychoacoustic correspondent). Still, the different dimensions seem to be interconnected. Some of these connections might simply be due to physiological reasons. For instance, it is well established that peak velocity varies with displacement [24, 25]. Others, like the relationship between duration and displacement, are more challenging to capture (see e.g., [6]).

Assuming a single dynamical parameter controlling all of these effects [15] is conceptually compelling, but it is less clear how such an account would capture typological differences. A different account can be offered from within the framework of Articulatory Phonology [e.g., 26], in which prosodic modulations are instantiated by modulation gestures that control the spatial (spatial μ-gestures) and/or temporal (temporal μ-gestures) profile of the C and V constriction gestures [27] and the tone gestures [cf. 28] that overlap with them. The degree of μ'gestures' effect increases at higher hierarchical levels. Based on our findings, English has both a spatial (for effects on displacement and peak velocity) and a temporal μ-gesture (for effects on duration) coordinated with the stressed syllable. There is also a pitch accent gesture coordinated with the same syllable. The higher the prominence level hierarchically, the stronger the strength of the μ-gestures, and thus the stronger their effects on the constriction and pitch accent gestures that overlap with them.

6. ACKNOWLEDGMENTS

The work was supported by collaborative NSF Grants #1551513 / 1551649 / 1551428 / 1551695.

7. REFERENCES

[1] Beckman, M.E., Pierrehumbert, J. 1986. Intonation structure in English and Japanese. *Phonology Yearbook* 3, 255-310.

[2] Silverman, K., Beckman, M.E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard labeling English prosody. *Proc. of the International Conference on Spoken Language Processing 2*, 867–870.

[3] Beckman, M. E., Edwards, J., Fletcher, J. 1992. Prosodic structure and tempo in a sonority model of articulatory dynamics. In: Docherty G. J., Ladd D.R. (eds), *Papers in Laboratory Phonology II: Segment, gesture, prosody*, Cambridge University Press, 68–86.

[4] Beckman M.E., Edwards, J. 1994. Articulatory evidence for differentiating stress categories. In Keating, P.A. (ed), *Phonological Structure and Phonetic Evidence*, Cambridge University Press, 7-33.

[5] Cho, T. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English. *JASA* 17, 3867-3878.

[6] Cho, T. 2006. Manifestation of prosodic structure in articulatory variation: Evidence from lip kinematics in English. In Goldstein, L., Whalen, D.H., Best, C.T. (eds), *Papers in Laboratory Phonology VIII: Varieties of Phonological Competence (Phonology and Phonetics*), Mouton de Gruyter, 519-548.

[7] de Jong, K. 1991. An articulatory study of consonant-induced vowel duration changes in English. *Phonetica*, 48, 1-17, 1991.

[8] de Jong, K. 1995. The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation, *JASA* 97, 491–504.

[9] de Jong, K., Beckman, M.E., Edwards, J. 1993. The interplay between prosodic structure and coarticulation, *Language and Speech*, 36, 197–212.

[10] Fowler, C.A. 1995. Acoustic and kinematic correlates of contrastive stress accent in spoken English. In: Bell-Berti, F., Raphael, J.J. (eds), *Producing Speech: Contemporary Issues: For Katherine Safford Harris*, Woodbury: American Institute of Physics, 355-373.

[11] Harrington, J., Fletcher, J., Beckman, M.E. 2000. Manner and place conflicts in the articulation of accent in Australian English. In: Broe, M. (ed.), *Papers in Laboratory Phonology 5*, Cambridge University Press, 40–55.

[12] Harrington, J., Fletcher, J., Roberts, C. 1995. Coarticulation and the accented/unaccented distinction: Evidence from jaw movement data. *Journal of Phonetics* 23, 305-322.

[13] Katsika, A., Tsai, K. 2021. The supralaryngeal articulation of stress and accent in Greek, Journal of Phonetics, 88, 101085.

[14] Hermes, A. Becker, J., Mücke, D., Baumann, S., Grice, M. 2008. Articulatory gestures and focus marking in German. *Proc. of Speech Prosody 2008*, 457-460.

[15] Roessig, S., Mücke, D. 2019. Modeling dimensions of prosodic prominence. *Frontiers in Communication* 4, article 44.

[16] Mücke, D., Grice, M. 2014. The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics* 44, 47-61.

[17] Katsika, A., Jang, J., Krivokapić, J., Goldstein, L., Saltzman, E. 2020. The role of focus in accentual lengthening in American English: Kinematic analyses. *Proceedings of the 10th International Conference on Speech Prosody 2020*, Tokyo, Japan.

[18] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program], Version 6.3.03, retrieved from http://www.praat.org/.

[19] Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[20] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[21] Lenth, R.V. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.5.3.https://CRAN.R-project.org/package=emmeans.

[22] Qualtrics (2005). Provo, Utah, USA. Version 2021. Available at: https://www.qualtrics.com

[23] Schloerke, B. Cook, D., Larmarange, L., Briatte, F., Marbach, M., Thoen, E., Elberg, A. Crowley, J. 2021. GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally

[24] Munhall, K.G., Ostry, D.J., Parush, A. 1985. Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance* 11, 457–474.

[25] Ostry, D.J., Munhall, K.J. 1985. Control of rate and duration of speech movements. *JASA* 77, 640–648.

[26] Browman, C.P., Goldstein, L.M. 1992. Articulatory phonology: An overview. *Phonetica* 45, 155-180.

[27] Saltzman, E., Nam, H., Krivokapić, J., Goldstein, L. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proc. of the Speech Prosody Conference 2008*, 175-184.

[28] Mücke, D., Nam, H., Hermes, A., Goldstein, L. 2012. Coupling of tone and constriction gestures in pitch accents. In: Hoole, P., Bombien, L., Pouplier, M., Mooshammer, C., Kühnert, B. (eds), *Consonant Clusters and Structural Complexity*. Mouton de Gruyter, https://doi.org/10.1515/9781614510772.205.