Vowel perception under prominence: Dual roles for F0 and duration

Jeremy Steffman[1] Wei Zhang[2]

[1]The University of Edinburgh [2]McGill University
[1]jeremy.steffman@ed.ac.uk [2]wei.zhang16@mail.mcgill.ca

ABSTRACT

Prosodic prominence modulates vowel production and acoustics. In the present study we test if the same effects play out in perception. We test perception of four American English vowel contrasts, varying formants and duration along a continuum. We test how F0-based prominence modulates vowel categorization, and explore the role of vowel duration as an intrinsic and prominence-lending cue. We find that F0-based prominence modulates vowel perception, dependent on vowel height. High vowels undergo perceptual recalibration in line with hyperarticulation, non-high vowels instead reflect sonority expansion. Lengthening of the vowel is interpreted as an intrinsic vowel quality cue, not as prominence. However, both F0-based prominence and duration interact with formants: more prominent F0 and longer duration enhance categorization along the formant continuum, showing a stronger influence of formant cues under prominence. Results thus show a dual influence for F0 and duration. Both independently impact categorization, and additionally mediate formant cue use.

Keywords: speech perception, vowels, prominence, sonority expansion, hyperarticulation.

1. INTRODUCTION

From a phonological view, a phenomenon in spoken language can be either segmental (e.g., vowels and consonants) or suprasegmental (e.g., presence or absence of a pitch accent). However, as [1] pointed out, it's hard to find a domain of speech dealing with only "suprasegmentals" (F0, duration and intensity). These intricacies have only more recently begun to be explored in terms of speech perception and speech processing [2,3]. Segments differ intrinsically in features such as F0 and intensity [4-6]. In addition, prosodic prominence modifies "segmental features" including VOT and formant structure in vowels [e.g., 7,8]. In other words, many phonetic dimensions play dual roles (both suprasegmental and segmental). For vowels in particular, F0, duration and formants have each been shown to vary systematically both as function of contrastive segmental features, and as a function prosodic prominence. In this study we ask: How does the listener deal with this multiple-mapping in speech perception?

Vowel perception under prominence is an interesting test case because there are two existing frameworks predicting acoustic variation under prosodic prominence in vowel production: *hyper-articulation* and *sonority expansion*. The hyper-articulation model [9] predicts that the vowel's distinctive features are hyper-articulated when the vowel is produced as prominent (e.g., pitch-accented). The sonority expansion model [10] predicts that for a prominent vowel, the vocal tract is more open (i.e., the jaw is lower) so that more energy can radiate from it, resulting in expanded sonority. The speech production literature suggests that the hyper-articulation hypothesis better predicts acoustic variation in high vowels and the sonority expansion hypothesis better predicts the patterns in non-high vowels, in American English vowels at least [7, 11, cf. 12]. This begs the question if perceptual analogs of these effects occur, and if they too vary based on vowel height, as in the production literature.

Previous work has shown that listeners consider prosodic features in segmental perception [e.g., 13, 14]. For example [13] showed that listeners expected a longer VOT to identify a sound as voiceless English /p/ after an intonational phrase boundary than a word boundary, reflecting phrase-initial lengthening (or, strengthening) of VOT in speech production. [14] recently explored how prosodic prominence influenced vowel categorization, finding that listeners used phrasal prosodic prominence to recalibrate perception of the /ɛ/-/æ/ contrast. A prominent vowel was perceived as having undergone sonority expansion with acoustically lower and backer (in the vowel space) formants perceived as /ɛ/ when prominent. Extending [14,15], we test how pitch accent type/shape influences vowel perception, and how vowel duration (along a continuum) impacts vowel perception for four vowel contrasts varying in height and front-ness: /i-ɪ/, /u-ʊ/, /ɛ-æ/, /ʌ-ɑ/. The vowels in each pair differ in their intrinsic duration [5] and F0 [6] making duration and F0 (potential) intrinsic cues for their identification. F0 is predicted to cue prominence, based on the finding that F0 is

more critical for prominence perception than duration [16], and based on the nature of our F0 manipulation.

Predictions for F0 are that prominence-lending F0 patterns should lead listeners to expect a prominent vowel variant: that is, hyperarticulation in high vowels and sonority expansion in non-high vowels. Empirically, this predicts that prominent F0 leads to *fewer* higher vowel responses in high vowel pairs /i-ɪ/, /u-ʊ/ (acoustically higher-vowel F1/F2 required to perceive /i/ and /u/ under an expectation of hyper-articulation). Conversely, prominent F0 is predicted to lead to *more* higher vowel responses in non-high vowel pairs /ɛ-æ/, /ʌ-ɑ/ (aligning with sonority expansion, whereby vowels become lower and backer in the vowel space). F0 as an intrinsic cue predicts the opposite effect for high vowels (more /i/ and /u/ responses with high F0). Duration also cues prominence [17], in which case longer duration may generate the same effects as prominent F0. Duration as an *intrinsic* cue predicts the opposite (more /i/, /u/, /æ/ and /ɑ/ responses with longer duration, as each of these is longer than their counterpart in the pairs tested [5]). We also examine how duration and F0 influence vowel perception in a carrier phrase versus isolation. This allows us to test the potential up-weighting or down-weighting of any of these cues as a function of the presence/absence of phrasal context.

2. METHODS

Stimuli were created by re-synthesizing the speech from a female speaker of American English, recorded in a sound-attenuated booth, using a Shure SM27 Large-diaphragm Condenser Microphone and pop filter. The audio was recorded at 44.1 kHz and 32 bit depth. The speaker was instructed to produce the phrase "I will say *x* now" (where *x* is one of eight target words) with 4 levels of emphasis (no-emphasis, emphasized, more emphasized, very emphasized) on *x*. The emphasis instructions were intended to elicit different levels of F0-based prominence which were used as the basis for resynthesis, described below.

All target words used a /k_d/ frame. The following words were used: "keyed", "kid", "cooed", "could", "ked" (a type of insect), "cad", "cud", "cod". The starting point was the speaker's production of the third level of emphasis for each word. We resynthesized F1 and F2 [18] to create a 7-step formant continuum between each pair of vowels: "kid-keyed", "ked-cad", "could-cooed", "cud-cod". Formant resynthesis was followed by pitch accent (F0) resynthesis. The goal of F0 resynthesis was to create two F0 conditions, one with a relatively prominent F0 contour over the target vowel, and one with a relatively non-prominent contour. To keep the F0 resynthesis relatively controlled while also being

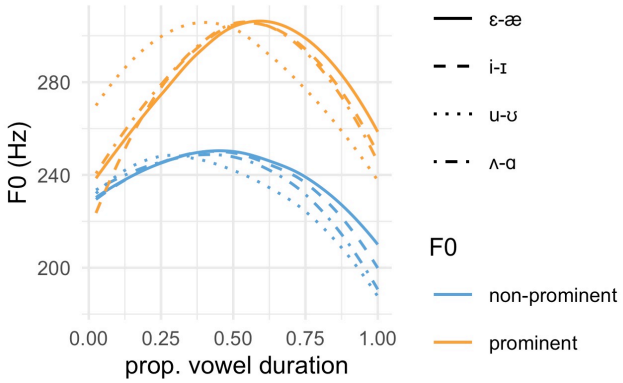

Figure 1: The F0 values for the F0 manipulation for each contrast (indicated by line type).

naturalistic, we resynthesized F0 to match naturally produced contours for each vowel contrast. For a given contrast, a representative less-prominent "no-emphasis" (roughly H*) contour was selected, and a representative more-prominent "more emphasized" (roughly L+H*) contour was selected. These were each overlaid on the continuum using the PSOLA method in Praat [19], with both conditions thus created by resynthesis. The F0 contours for each contrast and prominence condition are shown in Figure 1. Within a contrast the same distinction is apparent: the prominent condition shows a later F0 peak, higher scaling of the peak, and overall higher F0. We are not testing the effect of a specific F0 contour or shape (i.e. if we had controlled the contours across contrasts) but rather testing the effect of these relative differences which are *of the same nature* across contrasts. Finally, we created a duration continuum with five equidistant steps [20], whose endpoints varied by contrast and were set according to the speaker's productions:110-270ms for i/-/ɪ/, 80-250ms for /ɛ-æ/, 150-320ms for /u/-/ʊ/ and 60-250 for /ʌ/-/ɑ/. Altogether, there were 280 stimuli (4 contrasts * 7 formant steps * 2 pitch accents * 5 duration steps). Target words were spliced into the same sentence frame "I'll say __ now". In one version of the experiment they were presented in this carrier phrase, and in another were presented in isolation (a between-subjects manipulation).

Participants were recruited online from Prolific, 30 for each carrier phrase condition for a total of 60. All were self-reported native speakers of American English with normal hearing. Participation was remote, with participants seated in a quiet space using headphones. **The Experiment** was a two-alternative forced choice task. Stimuli were presented auditorily with orthographic representations onscreen and categorized by key press. Each stimulus was presented once for a total of 280 trials (fully randomized). All trials were analyzed.

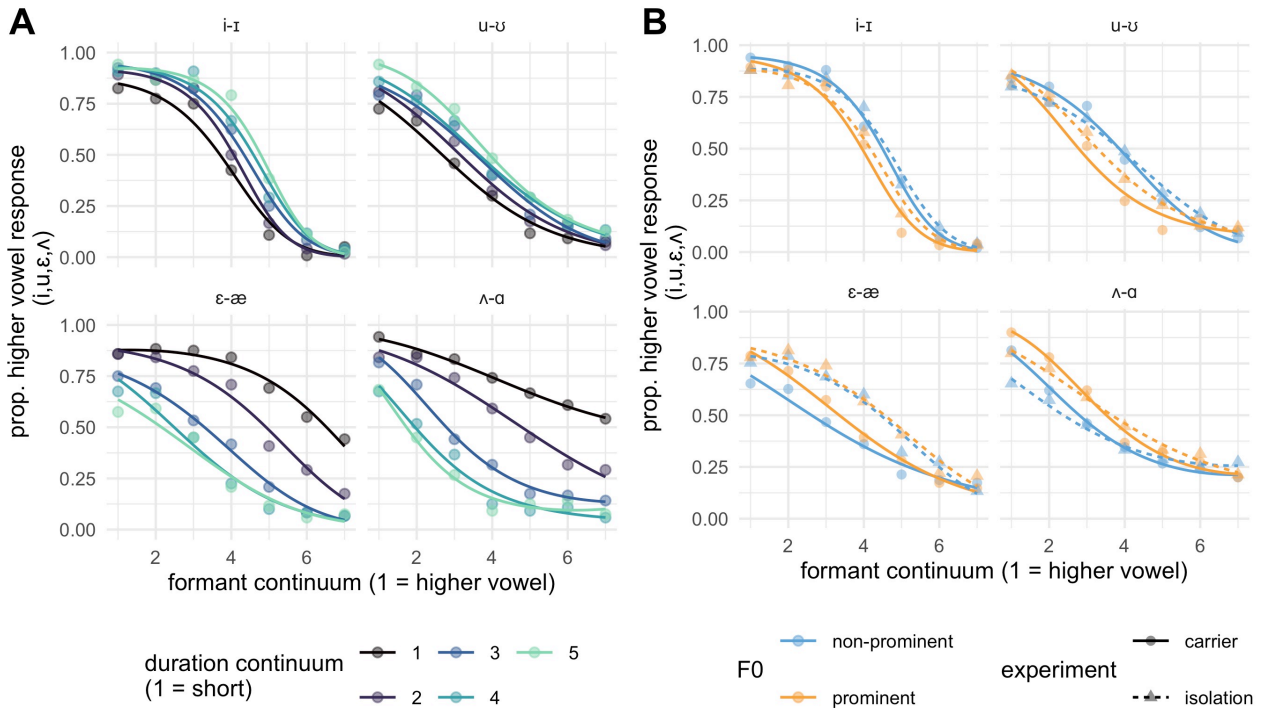

Figure 2: Panel A shows categorization responses as a function of formants (x axis) and duration (line coloration), with the proportion of higher vowel responses for each contrast plotted (/i,u,ɛ,ʌ/). Panel B shows categorization pooled by duration and split by the carrier phrase (line type) and F0 (coloration) manipulations.

Statistical analysis of the data was carried out with Bayesian mixed effects logistic regression, implemented using *brms* [21]. We fit a separate model for each of the four contrasts we tested. The dependent variable was coded with the higher vowel in the pair mapped to 1 (/i/, /ɛ/, /u/, /ʌ/ mapped to 1; /ɪ/, /æ/, /ʊ/, /ɑ/ mapped to 0). The fixed effects in each model were formant continuum step, duration continuum step (both scaled and centered), prominence condition (prominent mapped to 0.5, and non-prominent mapped to -0.5), and experiment (isolation mapped to 0.5 and carrier phrase mapped to -0.5). All interactions were included as well. Random effects were specified as by-participant random intercepts, with all fixed effects and interactions as random slopes, save for experiment (as it was not a within-participant manipulation). We report the median for an estimate's posterior and the probability of direction (pd), computed using the R package *bayestestR* [22] which gives the percentage of a posterior distribution with a given directionality. A distribution precisely centered on 0 (no effect) would have a pd of 50, while one with a consistently estimated positive or negative effect will have a pd approaching 100. pd > 97.5 corresponds to the 95% credible intervals (CrI) excluding zero, which we consider a credible effect. All of the data, code for running the statistical models, full model summaries, and an expanded version of Fig. 2 are available on the OSF at: https://osf.io/cfsru/.

3. RESULTS

First we consider the effects of formants and duration for each vowel contrast. We find an expected effect of the formant continuum for each contrast (/i/-/ɪ/: β = -3.70, pd = 100; /u/-/ʊ/: β = -3.14, pd = 100; /ɛ/-/æ/: β = -2.15, pd = 100; /ʌ/-/ɑ/: β = -2.05), shown as a decrease in categorization responses left to right along the x-axis in Fig 2A.

Duration credibly impacted categorization for each contrast as well. For both high vowel contrasts, longer duration led to an increase in /i/ and /u/ responses, respectively (/i/-/ɪ/: β = 0.60; /u/-/ʊ/: β = 0.40, pds = 100), while longer duration in the non-high contrasts increased /æ/ and /ɑ/ responses (/ɛ/-/æ/: β = -1.48, /ʌ/-/ɑ/: β = -1.83, pds = 100). Duration effects clearly support the intrinsic duration predictions [5]: longer duration favors perception of an intrinsically longer vowel, especially for non-high vowels [23].

F0 additionally had a credible effect for each contrast, which showed opposing directionality as a function of vowel height. In high vowel contrasts higher F0 decreased higher vowel responses (/i/-/ɪ/: β = -0.75, /u/-/ʊ/: β = -0.65, pds = 100). The effect showed an opposite directionality credible effect in non-high vowel contrasts whereby higher F0 *increased* higher vowel responses (/ɛ/-/æ/: β = 0.28, pd = 99; /ʌ/-/ɑ/: β = 0.67, pd = 100). Both of these

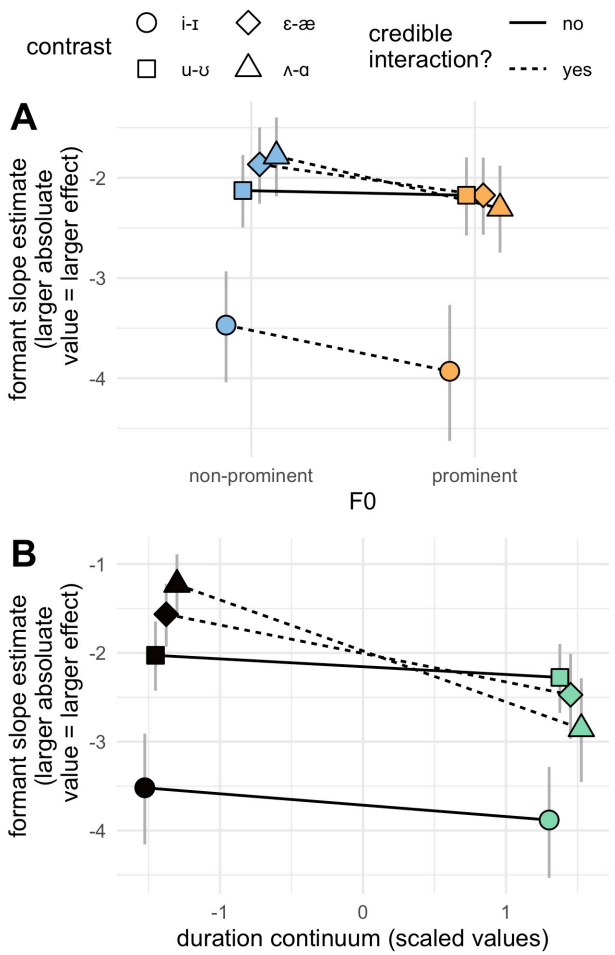

Figure 3: Estimates and 95%CrI from the models for the effect of formant step (y axis), as a function of F0 condition (Panel A) and duration continuum (Panel B).

effects are visible as differences across line coloration in Figure 2B. The carrier phrase manipulation exerted a credible main effect for only /ɛ/-/æ/ ($\beta = 0.97$, pd = 100), and otherwise did not interact with any predictors or show a main effect for other contrasts. The F0 effect confirms the prominence predictions: prominent F0 for high vowels decreases higher vowel responses, consistent with an expectation of acoustic hyperarticulation. Prominent F0 for non-high vowels increases higher vowel responses, consistent with an expectation of acoustic sonority expansion.

We next considered the potential *enhancement role* of both F0 and duration in vowel perception in examining the interaction of each with formant cues. Figure 3A shows the estimate for the effect of formants across F0 conditions, extracted using the *estimate_trends* function [24]. Note that a larger absolute value (more negative) effect corresponds to an up-weighting of formant cues. All contrasts but /u/-/ʊ/ showed a credible interaction between F0 (prominence) and formants (pds > 99). As shown in Figure 3A, this interaction reflected a stronger/larger effect of formants in the prominent F0 condition: prominence-based enhancement of formant cue use. Comparable duration-based enhancement was found for non-high vowel contrasts, shown in Figure 3B, both of which showed a credible interaction between vowel duration and formant continuum. In line with the F0 effect, this suggests longer (more prominent) durations lead listeners to upweight formant cues in perception, for non-high vowels in particular.

4. DISCUSSION AND CONCLUSION

This study finds clear support for duration as an intrinsic cue to vowel quality, while, conversely, the pattern of F0 results is consistent with F0 cuing prominence. Critically, the prominence effect varied by vowel height, comporting with the patterns of prominence strengthening in the speech production literature [9,10], while adding new data on how pitch accent type mediates prominence perception in this regard, thus building on [14,15]. We also found that F0-based prominence enhanced formant cue usage for three of the four contrasts tested. The same was observed for the non-high vowel contrasts as a function of duration, in which longer durations led a larger effect of formant cues. We suspect this effect is restricted to non-high vowels due to the higher weight of duration cues for these contrasts (see βs for duration in text and Figure 2A), which may lead to more salient duration-based enhancement.

In summary, results show a dual role for F0-based prominence in vowel perception: one that mediates which vowel category is perceived (as a function of prominence strengthening), and also enhances the perception of formant cues. Duration also serves a dual role in the sense that it constitutes an intrinsic cue to vowel quality, and additionally leads to upweighting of formants (for non-high vowels only). The present study thus advances our understanding of prominence effects on vowel perception across the vowel space in American English and addresses the question of multiple prominence functions in speech perception. More generally, this result, in the vein of other recent work, underlines the importance of considering prosodic features in segmental/lexical processing [cf. 25,26]. In future research, we plan to test the possibility of enhancing effects for other prominence cues and segmental cues, with the general prediction that prominence should enhance the use of segmental cues, e.g. VOT. Future work will also benefit from testing how these effects extend to L2 learners of English, and how they are impacted by speaker-specific properties and generalize across speakers. Finally, we are interested to test how they relate to distributional learning by varying the co-occurrence distributions of duration and F0 in the stimuli (e.g., longer durations tend to occur with a prominent F0).

5. REFERENCES

[1] Lehiste, I. (1970). Suprasegmentals. Cambridge, MA: MIT Press.

[2] Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, *40*(2), 141-201.

[3] McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In *The handbook of phonetic sciences* (pp. 489-520). Blackwell.

[4] Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, *31*(4), 428-435.

[5] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

[6] Chen, W. R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, *87*, 101063.

[7] Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878.

[8] Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, *59*(2–3), 134–149.

[9] De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

[10] Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty & D. R. Ladd (Eds.), *Gesture, Segment, Prosody (Papers in Laboratory Phonology)*. (pp. 68–86). Cambridge: Cambridge University Press.

[11] Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: Evidence from formant structure. *INTERSPEECH*, 2535–2538.

[12] Garellek, M., & White, J. (2015). Phonetics of Tongan stress. *Journal of the International Phonetic Association*, *45*(1), 13-34.

[13] Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19-EL25.

[14] Steffman, J. (2021). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, *36*(5), 586-611.

[15] Steffman, J. A. (2020). *Prosodic prominence in vowel perception and spoken language processing*. University of California, Los Angeles.

[16] Jasmin, K., Tierney, A., Obasih, C., & Holt, L. (2022). Short-term perceptual reweighting in suprasegmental categorization. *Psychonomic Bulletin & Review*, 1-10.

[17] Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126–152.

[18] Winn, M. (2016). Vowel formant continua from modified natural speech (Praat script). http://www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt.

[19] Boersma, P., & Weenink, D. (2022). Praat: Doing phonetics by computer (version 6.1.09). http://www.praat.org

[20] Winn, M. (2014). Make duration continuum (Praat script).http://www.mattwinn.com/praat/Make_Duration_Continuum.txt

[21] Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28.

[22] Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. Journal of Open Source Software, 4(40), 1541.

[23] Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013-3022.

[24] Makowski D, Ben-Shachar M, Patil I, & Lüdecke D (2020). Estimation of Model-Based Predictions, Contrasts and Means. *CRAN*.

[25] McQueen, J. M., & Dilley, L. C. (2020). Prosody and spoken-word recognition. In *The Oxford handbook of language prosody* (pp. 509-521). Oxford University Press.

[26] Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, *108*, 104034.