INTERSPEECH 2009 BRIGHTON

Prosodic effects on vowel production: evidence from formant structure

Yoonsook Mo[1], *Jennifer Cole*[1], *Mark Hasegawa-Johnson*[2]

[1] Department of Linguistics, University of Illinois, Urbana, Illinois
[2] Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois
ymo@illinois.edu, jscole@illinois.edu, jhasegaw@illinois.edu

10.21437/Interspeech.2009-668

Abstract

Speakers communicate pragmatic and discourse meaning through the prosodic form assigned to an utterance, and listeners must attend to the acoustic cues to prosodic form to fully recover the speaker's intended meaning. While much of the research on prosody examines supra-segmental cues such as F0 and temporal patterns, prosody is also known to affect the phonetic properties of segments as well. This paper reports on the effect of prosodic prominence on the formant patterns of vowels using speech data from the Buckeye corpus of spontaneous American English. A prosody annotation was obtained for a subset of this corpus based on the auditory perception of 97 ordinary, untrained listeners. To understand the relationship between prominence perception and formant structure, as a measure of the 'strength' of the vowel articulation, we measure the steady-state first and second formants of stressed vowels at vowel mid-points for monophthongs and at both 10% (nucleus) and 90% (glide) positions for diphthongs.

Two hypotheses about the articulatory mechanism that implements prominence (Hyperarticulation vs. Sonority Expansion Hypothesis) were evaluated using Pearson's bivariate correlation analyses with formant values and prominence 'scores'— a novel perceptual measure of prominence. The findings demonstrate that higher F1 values correlate with higher prominence scores regardless of vowel height, confirming that vowels perceived as prominent tend to have enhanced sonority. In the frontness dimension, on the other hand, the results show that vowels perceived as prominent tend to be hyperarticulated. These results support the model of the supra-laryngeal implementation of prominence proposed in [5, 6] based on controlled "laboratory" speech, and demonstrate that the model can be extended to cover prosody in spontaneous speech using a continuous-valued measure of prosodic prominence. The evidence reported here from spontaneous speech shows that prominent vowels have expanded sonority regardless of vowel height, and are hyperarticulated only when hyperarticulation does not interfere with sonority expansion.

1. Introduction

Spoken utterances are composed of hierarchically structured phonological prosodic units with prominence relationships among them. Prosodic structures, in particular the edges of prosodic units and prominent elements, are signaled through the modulation of segmental and supra-segmental phonetic patterns. In everyday conversation, listeners must be sensitive to this phonetic detail in order to reconstruct prosodic structures and to understand the pragmatic and discourse meaning the speaker is conveying through the prosodic form encoded in an utterance.

Prior studies demonstrate that both supra-segmental features (pitch, loudness) and segmental features (like formants) are modulated by prosodic prominence, but most previous studies examine prominence that marks only narrow (contrastive) focus in controlled "laboratory" speech [1, 2, 3, 4, 5]. The present study seeks to extend our understanding of the influence of prosodic prominence on vowel formants in American English by analyzing the everyday conversational speech of ordinary, not professional speakers. The prosodic features are identified by untrained listeners on the basis of auditory impression only.

Prosodically prominent words have distinct formant patterns. To model the production mechanisms that underlie phonetic variation arising from prominence, researchers have proposed two distinct and partly contradictory hypotheses: the Hyperarticulation Hypothesis and the Sonority Expansion Hypothesis. Beckman and her colleagues [3] proposed that pitch accent (prominence) enhances intrinsic sonority (Sonority Expansion Hypothesis) based on findings from an optoelectronic tracking study of jaw height. That is, accented vowels have a more open vocal tract with less impedance of the airway. On the other hand, in his X-ray microbeam study [4], de Jong refuted the sonority expansion hypothesis. He found that features such as lip roundness and protrusion, which are not directly related to sonority expansion, are enhanced and thus proposed that stress (prominence) induces hyperarticulation, that is, expanded or enhanced articulation of the distinctive features of segments.

Findings from later acoustic and articulatory studies [5, 6], however, do not provide a unified account for the effect of prominence on vowel formant structures. Erickson [5] demonstrated that the jaw does not always move along with the tongue in the same direction. When high vowels are emphasized, she found that the jaw moves downward but the tongue dorsum move forward and upward, resulting in lower F1 and higher F2, supporting the Hyperarticulation Hypothesis. Conversely, Cho [6] in his EMA study found that the high vowel /i/ has a higher F1 and higher F2 when prominent, suggesting that the Sonority Expansion Hypothesis has precedence, and that prominence induces hyperarticulation only when hyperarticulation does not interfere with enhanced sonority. In a later acoustic study by Lee and colleagues [10], employing the prosodically labeled Boston University Radio News corpus with news stories read by 4 professional news announcers, inconsistent effects of pitch accents on formant structures are reported. The lax vowels /ɪ, ɛ, ʌ/ are lowered under prominence, and tense high vowels /i, u/ are raised by some speakers, while there is no effect for other speakers.

The present study provides a further test of these two hypotheses and their partly contradictory prediction about F1 of high vowels. This study extends our understanding of the acoustic effects of prosodic prominence by way of three methodological innovations. First, the current study employs speech excerpts from spontaneous conversational speech produced by multiple ordinary speakers of American English, and therefore, unlike most previous studies, the types of prosodic prominence examined include both broad focus prominence (prominence which marks new information to the discourse) as well as narrow focus prominence (prominence

which emphasizes a word to express negation, correction, or contrastive focus). Second, the status of a word as prominent is based on the judgments of untrained, ordinary listeners performing a real-time transcription task, using only auditory impression. Third, instead of looking at a few vowels (/ɑ/ in Beckman et al., /ɑ, i, ɛ/ in Erickson, and /ɑ, i/ in Cho), the current study examines all the vowels of American English except the diphthong /ɔɪ/. Hence this study examines the contribution of formants as generalized cues to prominence for untrained ordinary listeners.

2. Methodology

2.1. Transcription task

97 ordinary listeners from undergraduate linguistics courses at the University of Illinois at Urbana-Champaign participated in transcription tasks. A total of 54 speech excerpts from 38 speakers were extracted from the Buckeye corpus of American English spontaneous speech [7]. In the transcription task, listeners were provided with a minimal definition of prominence. Then, they were seated at a computer, equipped with individual headphones and provided with a printed transcript with all punctuation and capitalization removed. They marked words heard as prominent in real time, while listening to speech excerpts presented in randomized order, without the aid of any visual display of the speech. Each excerpt was transcribed by 10-22 ordinary listeners. Transcription data from 6 transcribers are excluded: some because they did not follow the instructions and some because they identified themselves as non-native speakers of American English on the language background questionnaire. Transcriptions are pooled across transcribers, and each word is thereby assigned a probabilistic prominence score (P-score) that specifies the fraction of listeners who mark the word as prominent. The P-score ranges from 0 to 1, as shown in Fig. 1. For example, in Fig. 1, no listener hears the first word 'I' as prominent (P-score = 0.0) but the word 'today's' is marked as prominent by all the listeners (P-score = 1.0).

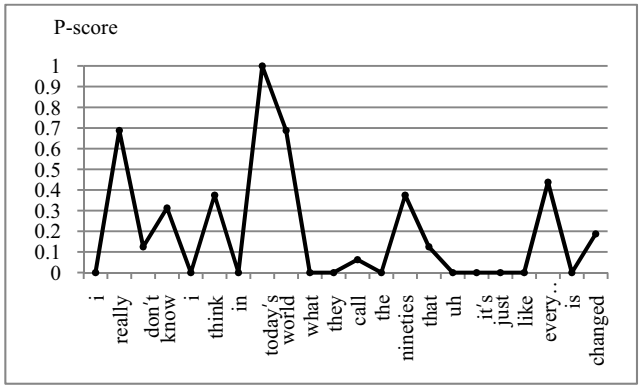

Figure 1: *Graph of P(rosody)-scores for words in a small portion of one excerpt from speaker 2.*

2.1.1. Reliability test

Multi-rater agreement scores using Fleiss' kappa statistic are calculated in order to evaluate whether untrained listeners reliably and systematically perceive prosodic prominence. The kappa scores are tested for significance using a z-test, with results as summarized in Table 1. All the Fleiss' kappa scores are well above chance ($p < .001$, $z=2.33$ with 99% confidence interval), confirming that untrained ordinary listeners' perception of prosodic prominence is not random, but systematic and reliable beyond chance levels.

2.1.2. Acoustic measurements

The stressed vowels of each word are identified based on a reference dictionary [9] so that all vowels analyzed in this study are lexically stressed, and only the effects of sentence level prominence are under examination. The distribution of stressed vowels in the database is summarized in Table 2. The waveforms for each excerpt are aligned with word and phone transcriptions and formant values are measured for all stressed vowels. The steady-state first and second formant values are automatically measured at vowel mid-points for monophthongs and at 10% (nucleus) and 90% (glide) positions for diphthongs. The measured formant values are normalized within phone and speaker using equation (1).

$$z_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_j} \quad (1)$$

where $x_{i,j}$ is the measured formant value of the ith token of type j, $\bar{x}_j$ is the average formant of type j, s_j is the standard deviation of type j, and the number of different types equals the number of speakers times the number of phoneme labels.

Table 1. *Results of Fleiss' kappa multi-rater agreement scores, and corresponding z-statistics, for P-score annotation. Results reported separately for each of six transcriber groups annotating identical materials*

z=2.33, α=0.01		Exp.1				Exp. 2	
		Run 1		Run 2		Run 1	
		Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2
Prom	Kappa	0.373	0.421	0.394	0.407	0.356	0.400
	z	19.43	20.48	18.15	18.31	15.31	19.56

Table 2. *Distribution of stressed vowels in the database*

vowel	ɑ	æ	ʌ	ɔ	aʊ	aɪ	ɛ
Freq.	173	290	407	121	52	309	463
vowel	ɝ	eɪ	ɪ	i	oʊ	ʊ	u
Freq.	122	214	475	306	211	72	183

3. Results

Pearson's bivariate correlation analyses are performed with the probabilistic P-scores and normalized formant values in order to see the relationship between prosodic prominence and vowel formant structures in American English, when prominence is based on the perception of untrained listeners. The results demonstrate that most vowels show a significant correlation between perceived prominence and the first and second formants values, as summarized in Table 3. Overall, the formant values of monophthongs are correlated with prosodic prominence, but neither the nucleus nor the glide part of a diphthong show consistent patterns of correlation between formant values and prominence. More specifically, the first formant values (F1) are positively correlated with perceived prominence in most monophthongs (9 out of 10 monophthongs). The second formant values (F2) are negatively correlated with prominence in the vowels, /ɑ, ʌ, ɛ, u/ , the glide portion of /aʊ/, and the nucleus portion of /oʊ/, while they are positively correlated in the vowels, /i/, and the nucleus portion of /aʊ/.

Multiple linear regression analyses were performed in order to test the extent to which formants can predict listeners' ratings of prosodic prominence (Fig. 2). The results illustrate that in 10 out of 14 vowels, a certain portion of the variation in listeners' responses to prominence can be explained based on variation in the patterns of formant structures. Statistically meaningful regression models of perceived prominence can be established for all the monophthongs except the vowel /ʊ/ based on the observed variation in formant structures. On the other hand, except the vowel /aʊ/, the perceived prominence of diphthongs is not significantly correlated with formant variation. Notably, even when prominence is correlated with formant variation, only a small portion of the P-score variance (1.4 to 20.0%) can be explained by the variation of formant structures.

Table 3. *Pearson's bivariate correlation coefficients (R) for the correlation between P-scores and F1, and P-scores and F2. ** represents a significant correlation with a 99% confidence interval and * represents a significant correlation with a 95% confidence interval.*

Vowel		ɑ	æ	ʌ	ɔ	ɛ	ɝ
R	F1	.106	.181**	.226**	.201*	.186**	.335**
	F2	-.185**	-.045	-.164**	-.131	-.095*	-.024

Vowel		ɪ	i	ʊ	u	aʊ	
						10%	90%
R	F1	.131**	.142**	.285**	.178**	.097	-.270*
	F2	-.010	.251**	-.021	-.216**	.248*	-.345**

Vowel		aɪ		eɪ		oʊ	
		10%	90%	10%	90%	10%	90%
R	F1	-.011	-.068	.026	.077	.013	-.082
	F2	-.130*	.018	.093	.150*	-.171**	-.026

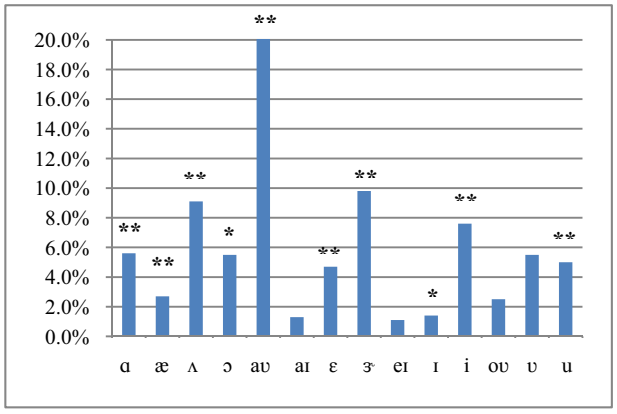

Figure 2: *Bargraph of R^2 values from simple linear regression models of perceived prominence on the basis of the first and second formant values as predictors. ** represents a significant correlation with a 99% confidence interval and * represents a significant correlation with a 95% confidence interval.*

Table 4 summarizes the measure of the contribution of each formant value to the regression models. With regard to the regression models of perceived prominence for monophthongs, F1 is included in the regression models of perceived prominence for 9 vowels, /ɑ, æ, ʌ, ɔ, ɛ, ɝ, ɪ, i, ʊ/, and F2 is included for 5 vowels, /ɑ, ʌ, ɛ, i, u/. On the other hand, only 3 of the diphthongs, /aʊ, aɪ, oʊ/ show formant values as predictors of perceived prominence, and in all three cases, only the nucleus (not the off-glide) shows contribution, and the formant in question is always F2.

Table 4. *Summary of the results of multiple regression analyses of F1 and F2 as predictors of P-scores. ** represents a significant correlation with a 99% confidence interval and * represents a significant correlation with a 95% confidence interval.*

V	predictors	Beta	V	predictors	Beta
ɑ	F1	.193*	u	F1	.124
	F2	-.252**		F2	-.179*
æ	F1	.178**	aʊ	F1_10	.223
	F2	-.029		F2_10	.339*
ʌ	F1	.266**		F1_90	-.264
	F2	-.214**		F2_90	-.263
ɔ	F1	.234*	aɪ	F1_10	.039
	F2	-.168		F2_10	-.153**
ɛ	F1	.208**		F1_90	-.074
	F2	-.128**		F2_90	.012
ɝ	F1	.335**	eɪ	F1_10	.001
	F2	-.019		F2_10	.043
ɪ	F1	.136**		F1_90	.073
	F2	-.032		F2_90	.136
I	F1	.139*	oʊ	F1_10	.113
	F2	.249**		F2_10	-.205**
ʊ	F1	.285*		F1_90	-.109
	F2	-.018		F2_90	.061

4. Discussion

The current study examines whether variation in the formant patterns of the lexically stressed vowel is predictive of how ordinary listeners perceive prosodic prominence for a word, and if so, what hypothesis concerning the mechanism of prosody production can best account for the relationship between the formant structures and perceived prominence.

The findings from Pearson's bivariate correlation analyses show that variation in formant structures are significantly correlated with listeners' perception of prosodic prominence. Looking closely, prosodic prominence identified by untrained listeners is positively correlated with F1 for all monophthongs except the low vowel, /ɑ/, which requires perhaps the most open vocal tract of any vowel. This suggests that regardless of intrinsic vowel height, except for the vowel /ɑ/, monophthongs have a more open vocal tract when perceived as prominent than when they are not prominent. This supports the Sonority Expansion Hypothesis, but not the Hyperarticulation Hypothesis, which predicts that intrinsically high vowels including /i, ɪ, ʊ, u/ will be articulated as higher (with lower F1) when marked for prosodic prominence. The findings for F2, by contrast, support the Hyperarticulation Hypothesis: F2 values of 3 back monophthongs, 2 back diphthongs, and 1 front monophthong are negatively correlated with perceived prominence, while those of the front high vowel /i/ and the front portion of the diphthong /aʊ/ are positively correlated with perceived prominence. This partly supports the Hyperparticulation Hypothesis in the frontness dimension, showing that the most extreme front vowel is more front while other vowels are more back. However, other phonologically front vowels fail to show enhanced fronting when prominent. To summarize, the present study finds that

Sonority Expansion prevails over Hyperarticulation, and that prominent vowels show Hyperarticulation only in the frontness dimension, at least for some vowels, but only when Hyperarticulation does not conflict with Sonority Expansion.

The results of Pearson's bivariate correlation analyses between formant values and perceived prominence of diphthongs, however, show only a sporadic correlation between formants and prominence, suggesting that prominence influences the formant patterns of diphthongs less effectively than those of monophthongs.

The contribution of each formant measure to listeners' perception of prominence is modeled by multiple linear regression models. The results of multiple regression analyses show that formant patterns influence the perception of prosodic prominence, but not much (Fig. 2): the variation in formant patterns of monophthongs does not account for much variation in listeners' responses to prosodic prominence. Only 1.4 to 20.0% of the variation in perceived prominence can be accounted for on the basis of formant variation. The regression models illustrate that the effect of formant values as cues to prosodic prominence is significant but relatively small compared to the effects of other suprasegmental features including duration, overall intensity, and bandpass filtered intensities reported in the prior studies [11, 12]. Comparing the effects of F1 with those of F2 as a cue to prominence (Table 4), the variation in the patterns of F1 not only is included in the regression models for a greater number of vowels, but also contributes more to explain the variance in listeners' responses to prosodic prominence, suggesting that expanded sonority is a more reliable cue to prosodic prominence for ordinary listeners than enhanced distinctiveness in the front/ back dimension.

Taking the findings from both Pearson's bivariate correlation analyses and regression analyses together, we can confirm that prosodic prominence expands sonority of vowels in spontaneous conversational speech, and untrained ordinary listeners are sensitive to this property, responding to the phonetic variation of raised F1s and peripheral F2s in identifying the locations of prosodic prominence. However, it is also shown that the effect of prominence on F1 of monophthongs, reflecting expanded sonority, contributes more reliably to the perception of prosodic prominence by untrained ordinary listeners than the effect on F2 of monophthongs, reflecting peripheral articulation in the frontness dimension.

5. Conclusions

Nearly one hundred ordinary listeners of American English transcribed the prosody of spontaneous conversational speech produced by multiple speakers, demonstrating that untrained ordinary listeners reliably and systematically perceive prosodic prominence. Consistent with the findings from prior controlled laboratory studies [6, 8], the current study confirms that formant variation correlates with perceived prominence: (1) in particular, positive correlations between perceived prominence and F1 support the Sonority Expansion Hypothesis, (2) in the frontness dimension where enhancing distinctive features does not interfere with sonority expansion, prosodic prominence is associated with hyperarticulation. Lastly, this study shows that the phonetic variation in the patterns of formant structures contributes to untrained ordinary listeners' perception of prosodic prominence, and in particular, increased F1, reflecting expanded sonority, is a more reliable cue to prosodic prominence than peripheral F2.

6. Acknowledgements

This research is supported by NSF grants IIS 07-03624 and IIS 04-14117 to Jennifer Cole and Mark Hasegawa-Johnson. I would like to thank the members of Prosody-ASR group for their comments and Eun-Kyung Lee and Zack Hulstrom for data collection.

7. References

[1] Sluijter, A. M. C. and van Heuven, V. J., "Acoustic correlates of linguistic stress and accent in Dutch and American English", The proceedings of ICSLP 96, 1996.

[2] van Bergem, D. R., "Acoustic vowel reduction as a function of sentence accent, word stress and vowel class", Speech Communication. 12:1-23, 1993.

[3] Beckman, M. E., Edwards, J. and Fletcher, J., "Prosodic structure and tempo in a sonority model of articulatory dynamics", in Docherty, G. J. and Ladd, D. R. [Eds.], Laboratory Phonology II: Gesture, Segment, Prosody, 68-86, Cambridge University Press, Cambridge 1992.

[4] De Jong, K. J., "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", Journal of the acoustical society of America, 97(1):491-504, 1995.

[5] Erickson, D., "Articulation of extreme formant patterns for emphasized vowels", Phonetica, 59:134-149, 2002.

[6] Cho, T., "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a, i/ in English", Journal of the Acoustical Society of America, 117(2):3867-3878, 2005.

[7] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E., "Buckeye Corpus of Conversational Speech" (2nd release), Department of Psychology, Ohio State University (Distributor), Columbus, OH, Online: www.buckeyecorpus.osu.edu, 2007.

[8] Harrington, J., Fletcher, J. and Beckman, M., "Manner and place conflicts in the articulation of accent in Australian English", In Broe, M. B. and Pierrehumbert, J. B. [Eds.], Laboratory Phonology V: Acquisition and the Lexicon, 40-51, 2000.

[9] Hasegawa-Johnson, M. and Fleck, M., "ISLE Dictionary version 0.2.0", Illinois Speech and Language Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, Online: http://www.isle.uiuc.edu/dict/index.html, accessed on Oct. 19, 2007.

[10] Lee, E-K., Cole, J. and Kim, H., "Additive effects of phrase boundary on English accented vowels", Proceedings of the 3rd Speech Prosody Conference, Dresden, Germany, 2006.

[11] Mo, Y., "Acoustic cues of prosodic prominence to naïve listeners of American English", Proceedings of the 34th Annual Meeting of the Berkeley Linguistic Society, 2008.

[12] Mo, Y., "Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception", Proceedings of the 4th Speech Prosody Conference, Campinas, Brazil, 739-742, 2008.