Phonological Contrast Drives Phonetic Implementation: Evidence from Category Goodness Ratings

Michael C. Stern
Yale University

Phonological Contrast Drives Phonetic Implementation: Evidence from Category Goodness Ratings

Abstract

Previous evidence from category goodness rating tasks has demonstrated that the phonetic reflexes of phonological categories have internal structure, such that some signals are better cues to phonological category membership than others. Puzzlingly, the best-rated exemplars of peripheral vowel categories have often been observed to be more peripheral than even hyperarticulated signals from natural speech, while non-peripheral vowel categories have often failed to demonstrate clear patterns of internal structure. The present study proposes that these puzzles can be explained by a conception of the phonology-phonetics interface whereby phonetic implementation is primarily driven by the maximization of contrast, rather than by particular targets in acoustic or articulatory space. Two experiments were conducted to test this hypothesis. In Experiment 1, native American English speakers rated the category goodness of exemplars of the peripheral [æ] category, and native Turkish speakers rated exemplars of the non-peripheral [y] category. Results confirmed the generalization that peripheral vowels demonstrate clearer patterns of internal structure than non-peripheral vowels. In Experiment 2, native American English speakers rated exemplars of the [æ] category that were extremely peripheral in the F1-F2 space—far more peripheral than even the most hyperarticulated productions from natural speech. Results generally supported an indefinite linear relationship between formants and goodness, consistent with a view of phonetic implementation based primarily on contrast rather than particular targets. I argue that this view is broadly compatible with existing approaches to the phonology-phonetics interface, but would involve certain modifications. I suggest paths for future research to investigate the implications of these modifications.

Phonological Contrast Drives Phonetic Implementation: Evidence from Category Goodness Ratings

Michael C. Stern*

1 Introduction

A division is often drawn between *phonological information*[1] and *phonetic information* (see Cohn and Huffman 2014 for an overview and bibliography). Phonological information is usually conceptualized as lexically contrastive, discrete, and abstract or low-dimensional, whereas phonetic information is non-contrastive, continuous, and concrete or high-dimensional. Given the apparent ontological incommensurability of these two domains (c.f. Poeppel and Embick 2005), understanding the nature of their relationship has been a defining problem in the study of linguistic sound structure.

Attempts to address this problem can be broadly divided into two main approaches: the 'symbolic' approach and the 'dynamic' approach (terminology from Gafos and Benus 2006). Under the symbolic approach of generative phonology, phonological information is primarily constituted by *phonemes* (e.g. Sapir 1933) or *distinctive features* (e.g. Chomsky and Halle 1968)—discrete mental categories that serve to differentiate lexical items—as well as the abstract processes undergone by these units. Phonetic information, on the other hand, is the information involved in the *implementation* of phonological information physically through articulation, as well as the retrieval of phonological information from physical data through perception (e.g. Lenneberg 1967). Since phonological information is viewed as ontologically distinct from phonetic information, a transduction process must mediate an interface between them (Volenec and Reiss 2017). The dynamic approach, on the other hand, denies an ontological division between phonology and phonetics, instead aiming to integrate both in a single model with no interface (e.g. Ohala 1990). Articulatory Phonology (AP: Browman and Goldstein 1989) models human speech using differential equations that capture both its discrete (phonological) and continuous (phonetic) aspects as different levels of description of the same system. In this way, "there is, strictly speaking, no 'mapping' between phonological and phonetic information" (Mücke, Hermes, and Tilsen 2020:140).

Both of these approaches must contend with a defining aspect of the relationship between phonological representation and phonetic implementation: *variance.* The implementation of phonological information varies greatly depending on, for example, phonetic context (e.g. Liberman et al. 1967), the individual talker (e.g. Klatt and Klatt 1990), and many other factors. This, of course, poses a problem for speakers/listeners, who must relate a variable phonetic signal to (more or less) constant phonological information, and vice versa. Early evidence suggested that listeners solve the 'lack of invariance problem' by collapsing subcategorical acoustic variance during perception, only attending to differences between sounds that cross category 'boundaries' (Liberman et al. 1957). However, later evidence from 'category goodness' rating tasks suggested that listeners remain attentive to subcategorical variance in the phonetic signal (e.g. Miller 1994). In a typical category goodness rating task, participants listen to instances of speech sounds that vary slightly in some relevant phonetic dimension(s) and, using a numerical scale, judge how well each sound exemplifies its category. Category goodness rating tasks have consistently demonstrated that the perceptual reflexes of phonological categories have *internal structure,* such that even among sounds which are all identified as cueing the same category, some sounds are perceived as *better* instances of that category than others (e.g. Drouin, Theodore, and Myers 2016). This body of evidence has motivated models that view speech perception as a statistical inference problem, whereby phonetic information cues phonological category membership with varying likelihoods, rather than in a binary fashion (Kronrod, Coppess, and Feldman 2016).

*Many thanks to the study participants, as well as Kyle Gorman and Gita Martohardjono for their contributions to the larger study from which this data came, and CUNY Second Language Acquisition Lab research assistants Ilaria Porru and Erjon Xholi for their help with data collection and processing. Portions of the data reported here were previously reported in my MA thesis (Stern 2020).

[1]I use the term *information* in a broad sense, not intending to invoke any particular definition from information theory, physics, etc.

One puzzling finding from category goodness rating tasks is that, at least for vowels, the best-rated exemplars tend to be *more peripheral* in the formant space than average productions. For example, among sounds that are identified as American English [i], those rated highest tend to have a lower first formant (F1) and higher second formant (F2) than sounds generated during actual speech (e.g. Iverson and Kuhl 2000). Johnson, Flemming, and Wright (1993) explained this finding through what they termed the "hyperspace hypothesis." According to this hypothesis, phonetic targets of phonological categories are *hyperarticulated*, and it is these hyperarticulated targets that are measured by goodness rating tasks. During actual speech production, on the other hand, phonetic targets undergo reduction processes that lead to the less peripheral outputs we usually observe.

However, two puzzles remain. First, best-rated category exemplars actually tend to be more peripheral than even *hyperarticulated* productions (Johnson et al. 1993), casting doubt on a one-to-one relationship between hyperarticulations and best category exemplars. Second, although the hyperspace hypothesis does quite well in explaining the subcategorical structure of peripheral vowel categories like [i], it is not clear how to apply the hypothesis to non-peripheral categories like [ɪ], for which there is nowhere more peripheral in the F1-F2 space to go without encroaching on the space of another category. Indeed, attempts to measure the subcategorical structure of non-peripheral categories have often failed to produce clear patterns (e.g. Sussman and Gekas 1997). The present study is an attempt at solving these two puzzles.

2 This Study

2.1 The Proposal

In response to the puzzles described above, I propose that the phonetic implementation of phonological categories is not governed by particular target articulatory configurations or acoustic patterns, hyperarticulated or otherwise; rather, the goal in phonetic implementation is the *distancing* of phonetic information from the phonetic signatures of the other categories in the speaker's/listener's phonological system. In other words, phonetic implementation is governed primarily by the *maximization of contrast*. This would explain why participants tend to prefer the exemplars of peripheral vowel categories that are most peripheral in the F1-F2 space, even when those exemplars are more peripheral than hyperarticulations: the most peripheral exemplars are the *most different* from the exemplars of the other vowels in the participants' phonological system. This also provides an explanation of the less coherent patterns of goodness ratings for non-peripheral vowels: the phonetic implementation of non-peripheral vowel categories must balance contrasts on multiple sides in the F1-F2 space, giving rise to greater complexity than that involved in the implementation of peripheral vowels. In particular, individuals may vary in the relative weighting of each contrast, similar to the way that individuals vary in the weighting of specific phonetic cues to phonological contrasts (Shultz, Francis, and Llanos 2012), which is known to impact phonetic implementation (Hauser 2019). In this way, the phonetic implementation of categories maintaining a greater number of contrasts (non-peripheral vowels) would be subject to more sources of variation than that of categories maintaining less contrasts (peripheral vowels).

This proposal is reminiscent of previous work on the relationship between phonological contrast and phonetic implementation, most notably Dispersion Theory (Flemming 2004, Liljencrants and Lindblom 1972). However, Dispersion Theory (in its various forms) is primarily a theory of the long-term development of phonological inventories, rather than the real-time phonetic implementation of phonological information. For this reason, the present proposal is closer to the contrast and enhancement approach (Hall 2011) based on the Modified Contrastivist Hypothesis (Dresher 2009), which posits that phonological knowledge comprises only contrastive information represented in a hierarchical fashion, whereas during phonetic implementation, additional redundant information is produced so as to *enhance* the contrasts specified by the phonology (c.f. Keyser and Stevens 2006). However, at least in accounting for the puzzles described above, a division between the primary implementation of contrastive features and the secondary enhancement of those features does not appear to be necessary. For example, F1 and F2 are primary cues to vowel height and backness, respectively, which are uncontroversially contrastive features in American English, the language of the participants in Iverson and Kuhl (2000). Nonetheless, the results of that study (along with many

others) demonstrated that these cues are clearly quite malleable for the sake of contrast enhancement. In this way, enhancement does not seem to be limited to redundant features, but seems to play an important role in the implementation of phonologically active features as well.

2.2 The Experiments

In order to test this proposal, two experiments were conducted (the second of which is still ongoing). Experiment 1 compared category goodness ratings of sounds from a peripheral vowel category (American English [æ]) to those of sounds from a non-peripheral category (Turkish [y])[2] in order to confirm the generalization that peripheral vowels show clearer patterns of internal structure than non-peripheral vowels. Experiment 2 was another category goodness rating task of American English [æ]. However, the stimuli in this experiment were extremely peripheral in the F1-F2 space—far more peripheral than even the most hyperarticulated productions from human speech. Data from Experiment 2 (although preliminary, at this point) will allow us to test whether goodness ratings continue to increase indefinitely as peripherality increases, which would be expected if category goodness is primarily determined by contrastiveness, rather than by particular targets in articulatory or acoustic space.

3 Experiment 1

Experiment 1 consisted of two category goodness rating tasks: native American English speakers judged instances of [æ], and native Turkish speakers judged instances of [y]. The proposal described in Section 2.1 predicts that the peripheral vowel category [æ] will demonstrate a clearer pattern of internal structure than the non-peripheral category [y]. On the other hand, if category goodness is primarily determined by proximity to a particular target signal (or range of signals) rather than contrastiveness per se, then we would expect both categories to display relatively clear patterns of internal structure.

3.1 Method

3.1.1 Participants

Twenty-three native American English speakers participated in the [æ] goodness rating task, and 19 native Turkish speakers participated in the [y] goodness rating task. Participants ranged in age from 18 to 55 and reported no history of speech, language or hearing impairment. The English speakers all reported no higher than beginner proficiency in any language other than English. However, the Turkish speakers resided in the United States (ages of arrival were all 17 or later) and were advanced second language (L2) speakers of English. Potential effects of L2 exposure on the results will be briefly discussed in Section 5.

3.1.2 Stimuli

A matrix of 25 stimuli was synthesized for each vowel space using the KlattGrid synthesizer (Klatt 1980) implemented in Praat. All stimuli were 500 ms in duration, had a pitch contour rising from 112 Hz to 132 Hz over the first 100 ms and falling to 92 Hz over the remaining 400 ms, and a root-mean-square (RMS) intensity value of -20 dB.

Stimuli varied in 30-mel steps of F1 and F2. The F1-F2 ranges were determined based on the results of norming identification experiments (see Stern 2020 for details) and are displayed in Table 1 and Table 2. Note that the norming experiment for [æ] already reflected a preference for very peripheral sounds (c.f. average production values of 660 Hz and 1720 Hz for F1 and F2, respectively, from Peterson and Barney 1952). F3 for all [æ] stimuli was set to 2410 Hz based on Peterson and Barney (1952), and F3 for all [y] stimuli was set to 2320 Hz based on Radisic (2014). Remaining formants were set to the synthesizer's default values.

[2]I classify Turkish [y] as non-peripheral since it is consistently realized with a higher F1 and lower F2 than [i] (Radisic 2014). See Section 5 for further discussion of this point.

		1		2		3		4		5	
		F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
A	mels	809	1488	809	1458	809	1428	809	1398	809	1368
	Hz	735	1921	735	1852	735	1785	735	1720	735	1656
B	mels	839	1488	839	1458	839	1428	839	1398	839	1368
	Hz	774	1921	774	1852	774	1785	774	1720	774	1656
C	mels	869	1488	869	1458	869	1428	869	1398	869	1368
	Hz	813	1921	813	1852	813	1785	813	1720	813	1656
D	mels	899	1488	899	1458	899	1428	899	1398	899	1368
	Hz	854	1921	854	1852	854	1785	854	1720	854	1656
E	mels	929	1488	929	1458	929	1428	929	1398	929	1368
	Hz	896	1921	896	1852	896	1785	896	1720	896	1656

Table 1: F1 and F2 of each [æ] stimulus in Hz and mels. The y-axis represents F1 and the x-axis represents F2. Rows are labeled A-E and columns are labeled 1-5 for ease of reference.

		1		2		3		4		5	
		F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
A	mels	307	1478	307	1448	307	1418	307	1388	307	1358
	Hz	219	1898	219	1830	219	1763	219	1699	219	1636
B	mels	337	1478	337	1448	337	1418	337	1388	337	1358
	Hz	244	1898	244	1830	244	1763	244	1699	244	1636
C	mels	367	1478	367	1448	367	1418	367	1388	367	1358
	Hz	269	1898	269	1830	269	1763	269	1699	269	1636
D	mels	397	1478	397	1448	397	1418	397	1388	397	1358
	Hz	296	1898	296	1830	296	1763	296	1699	296	1636
E	mels	427	1478	427	1448	427	1418	427	1388	427	1358
	Hz	322	1898	322	1830	322	1763	322	1699	322	1636

Table 2: F1 and F2 of each [y] stimulus in Hz and mels.

3.1.3 Procedure

The experiments were conducted in a quiet room. Instructions and stimuli were presented on a tablet running E-Prime 2.0. Stimuli were presented binaurally through headphones. Participants were instructed to rate how well each sound exemplified its category (the vowel in *hat* for English speakers; the vowel in *kül* for Turkish speakers) on a six-point Likert scale with endpoints labeled "very good" and "very bad". All instructions and labels were presented in participants' native language. The experiment began with six practice trials, after which the participant was prompted to ask any questions. Then, each stimulus was presented twice in random order for a total of 50 trials. The session lasted approximately ten minutes.

3.2 Results and Discussion

Goodness ratings were first *z*-scored by participant to account for the different mean and variance in each participant's distribution of responses, and then averaged to create an overall goodness score for each stimulus. As seen in Figure 1, ratings of [æ] sounds demonstrated a clear pattern whereby ratings increased as F1 and F2 increased. The only apparent exception was at the highest level of F2, where the increase appeared to taper off. If this was due to the detection of a target value of F2, then we would expect ratings to begin to decrease as F2 continues to increase. On the other hand, this apparent tapering of the effect of F2 might be explained by the fact that higher formants tend to be relatively less important than lower formants in vowel perception (Schwartz et al. 1997:21-22 and references cited there). If this were the case, we might expect ratings to simply plateau as F2 continues to increase. A third possibility is that the apparent tapering was simply a spurious effect caused by sampling error. These possibilities will be teased apart in Experiment 2.

	1	2	3	4	5
A	-0.39	-0.40	-0.64	-0.94	-0.92
B	-0.02	0.09	-0.21	-0.40	-0.60
C	0.22	0.24	0.19	-0.04	-0.40
D	0.37	0.58	0.60	0.25	-0.18
E	0.91	0.87	0.73	0.48	-0.38

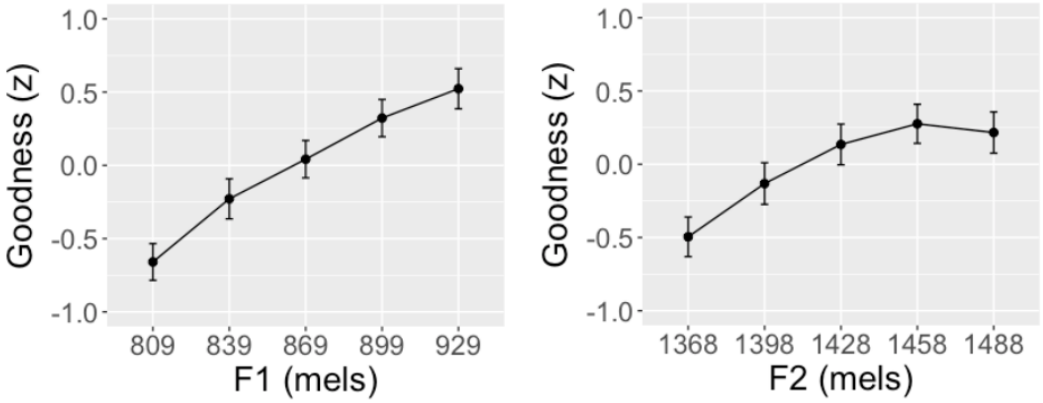

Figure 1: Mean *z*-scored goodness ratings of [æ] stimuli by native American English speakers (top: darker green indicates higher ratings); mean *z*-scored goodness ratings of [æ] stimuli at each level of F1, averaged across F2 (bottom left); and vice versa (bottom right). Error bars indicate 95% confidence intervals.

As seen in Figure 2, ratings of [y] sounds simply hovered around the mean across levels of F1 and F2. The 95% confidence intervals of *z*-scored goodness ratings at every level of F1 and F2 all overlapped with 0 (the mean). No clear pattern was observed.

	1	2	3	4	5
A	0.08	-0.05	-0.08	-0.18	-0.24
B	-0.06	0.17	-0.04	0.00	-0.33
C	0.10	0.30	0.05	0.31	-0.20
D	-0.20	0.30	0.07	0.17	0.02
E	-0.18	0.05	-0.09	-0.01	0.06

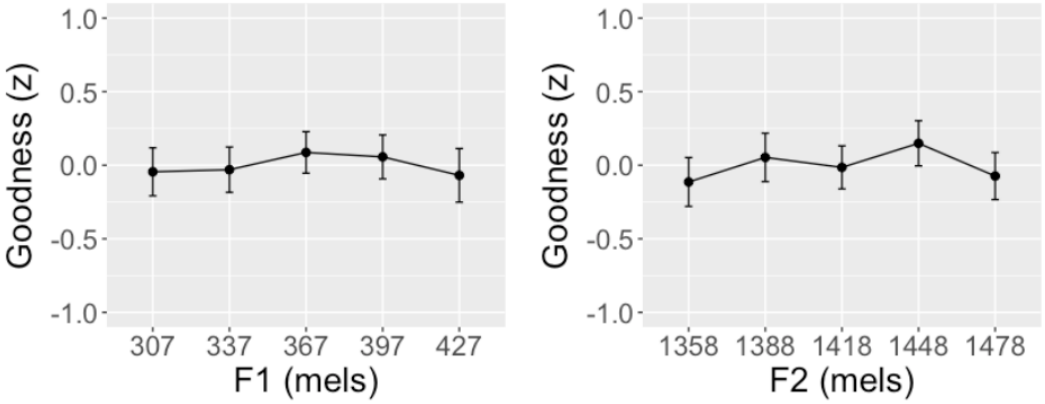

Figure 2: Mean *z*-scored goodness ratings of [y] stimuli by native Turkish speakers.

In order to further examine the reliability of the observed patterns (or lack thereof), a linear

mixed effects model of goodness ratings of stimuli in each vowel space was run in R using the `lme4` package (Bates et al. 2015). The models included fixed effects of F1, F2, trial number, and all interactions. Trial number was included as a fixed effect since goodness ratings have been argued to reflect adaptation to the particular range of stimuli presented (Lotto, Kleunder, and Holt 1998); we would expect adaptation of this sort to be indicated by a significant main effect of trial number, or one or more significant interactions with trial number. Fixed effects were scaled and centered. Both models included random intercepts by participant and by stimulus, and random slopes for F1, F2, and trial number by participant. The model of [y] ratings included a random slope for trial number by stimulus, but this random slope was excluded from the model of [æ] ratings because its inclusion led to a singular fit. Significant main effects and interactions are reported below.

The model of [æ] ratings revealed significant effects of F1 ($B = 0.536$, $SE(B) = 0.081$, $t(29.598) = 6.640$, $p < .001$) and F2 ($B = 0.347$, $SE(B) = 0.087$, $t(28.890) = 4.005$, $p < .001$), confirming the patterns observed in Figure 1. The model also revealed a significant effect of trial number ($B = -0.106$, $SE(B) = 0.049$, $t(21.981) = -2.143$, $p = .043$) and a significant interaction between F2 and trial number ($B = 0.078$, $SE(B) = 0.028$, $t(1047.436) = 2.787$, $p = .005$). The negative main effect of trial number suggests that participants generally decreased their ratings throughout the course of the experiment, while the positive interaction between F2 and trial number suggests that this effect was reversed for stimuli with higher F2. In other words, the positive effect of F2 on goodness ratings became more pronounced throughout the course of the experiment.

No main effects were significant in the model of [y] ratings; however, there was a small significant interaction between F1 and F2 ($B = -0.087$, $SE(B) = 0.039$, $t(21.016) = -2.198$, $p = .039$). In order to interpret this interaction, I calculated Spearman's coefficient (ρ) for the correlation between F1 and goodness at each level of F2, and between F2 and goodness at each level of F1. Consistent with the statistical interaction, the correlation between F1 and goodness tended to slightly decrease as F2 increased, and likewise for the correlation between F2 and goodness as F1 increased. However, the correlation (ρ) between F1 and goodness ranged only from .096 to -.092 and the correlation between F2 and goodness ranged only from .050 to -.129. Therefore, the statistical interaction does not seem to indicate any meaningful relationship between formants and goodness.

4 Experiment 2

It was observed in Experiment 1 (see Figure 1) that as F1 and F2 increased, [æ] category goodness increased, even as the formants became higher than those heard in normal productions (Peterson & Barney 1952) and even hyperarticulated productions (Johnson et al. 1993). If the category goodness of phonetic signals is primarily determined by the *distance* of those signals from the signals of other categories, as proposed above, then we would expect this linear relationship to continue indefinitely. On the other hand, if goodness is determined by proximity to a (range of) target values for F1 and F2, we would expect ratings to begin to decline at some point as F1 and F2 continue to increase.

In Experiment 2, another matrix of 25 [æ] stimuli was synthesized that was even more peripheral in the F1-F2 space than the [æ] matrix from Experiment 1. In particular, ratings of these stimuli will help determine whether the apparent plateau of goodness ratings at the highest levels of F2 in Experiment 1 were due to (1) detection of a target range of F2, in which case we would expect a decline in goodness ratings at higher F2 values; (2) the de-emphasis of higher formants as cues to vowel category membership, in which case we would expect a continual plateau of goodness ratings as F2 increases; or (3) sampling error, in which case we would expect the linear trend to continue as F2 increases. Data collection for Experiment 2 was interrupted by COVID-19 (see Section 4.1.1), so the results reported here should be regarded as preliminary.

4.1 Method

4.1.1 Participants

Inclusion criteria for native American English speakers were identical to those from Experiment 1. At this point, only seven participants have been tested on Experiment 2, due to interruption by the COVID-19 pandemic.

4.1.2 Stimuli

A matrix of 25 [æ] stimuli (see Table 3) was synthesized using the same parameters and inter-stimulus distances as those from Experiment 1. These stimuli differed from those in Experiment 1 only in the values of the first two formants. Stimulus A5 (the least peripheral) in this matrix had the same formant values as stimulus E1 (the most peripheral) from Experiment 1.

		1		2		3		4		5	
		F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
A	mels	929	1608	929	1578	929	1548	929	1518	929	1488
	Hz	896	2216	896	2139	896	2065	896	1992	896	1921
B	mels	959	1608	959	1578	959	1548	959	1518	959	1488
	Hz	939	2216	939	2139	939	2065	939	1992	939	1921
C	mels	989	1608	989	1578	989	1548	989	1518	989	1488
	Hz	983	2216	983	2139	983	2065	983	1992	983	1921
D	mels	1019	1608	1019	1578	1019	1548	1019	1518	1019	1488
	Hz	1029	2216	1029	2139	1029	2065	1029	1992	1029	1921
E	mels	1049	1608	1049	1578	1049	1548	1049	1518	1049	1488
	Hz	1076	2216	1076	2139	1076	2065	1076	1992	1076	1921

Table 3: F1 and F2 of each extremely peripheral [æ] stimulus in Hz and mels.

4.1.3 Procedure

The procedure of Experiment 2 was identical to that of Experiment 1.

4.2 Results and Discussion

Data was analyzed using the same methods described in Section 3.2. As seen in Figure 3, although the pattern was less clear (possibly due to the much smaller sample), the linear relationship between F1 and goodness continued; the only apparent exception was at the highest level of F1. Figure 3 also suggests a positive relationship between F2 and goodness; however, this relationship was even less robust than it was in Experiment 1. Nonetheless, the lack of a downward trend seems quite clear, contra the hypothesis that the plateau observed in Experiment 1 was due to detection of a target F2 value.

The linear mixed effects model of goodness ratings included the same fixed effects and random effects structure as the model of [æ] ratings from Experiment 1. The model revealed a significant main effect of F1 (B = 0.550, $SE(B)$ = 0.166, t(6.033) = 3.313, p = .016), confirming the pattern observed in Figure 1. There were no main effects of F2 or trial number. However, the model did reveal significant interactions between F1 and trial number (B = 0.114, $SE(B)$ = 0.053, t(306.172) = 2.172, p = .031) and F2 and trial number (B = 0.161, $SE(B)$ = 0.053, t(306.244) = 3.047, p = .003), suggesting that the positive effects of F1 and F2 became more pronounced throughout the experiment.

5 General Discussion and Conclusion

Taken together, the results reported above support a view of phonetic implementation based primarily on the maximization of contrast, rather than the achievement of particular targets in phonetic space. For peripheral vowels, this leads to a preference for phonetic signals that are as peripheral as possible; for non-peripheral vowels, competing contrastive forces on multiple sides give rise to a complexity that is more difficult for goodness rating tasks to measure.

Previous experimental evidence has demonstrated that degrees of phonological contrast (Hall 2013) are reflected by different degrees of articulatory movement (Hall et al. 2017), and that the presence of a minimal pair in the lexicon differing in a single phonetic dimension increases hyper-articulation of that dimension (Wedel, Nelson, and Sharp 2018). These findings suggest that the

	1	2	3	4	5
A	-0.55	-0.58	-0.52	-0.38	-1.01
B	0.07	-0.35	-0.02	-0.58	-0.29
C	0.25	0.10	0.31	-0.28	-0.06
D	0.82	0.24	0.36	0.46	0.02
E	0.80	0.49	0.50	0.30	-0.10

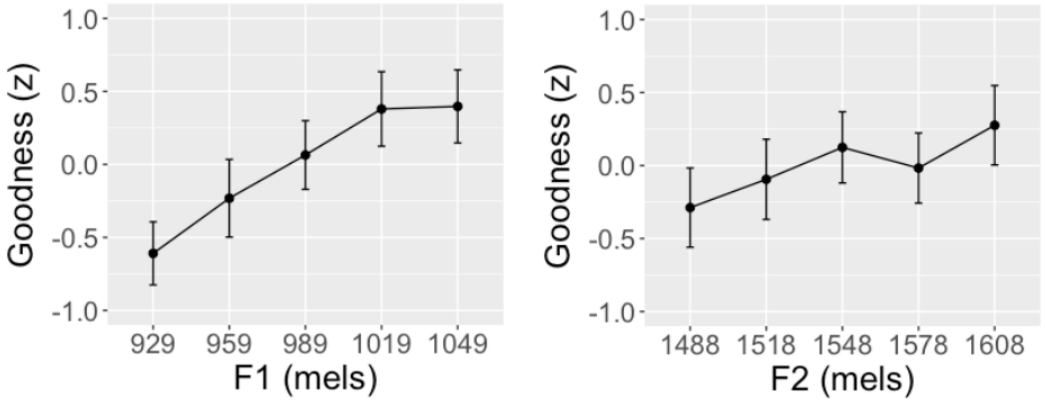

Figure 3: Mean *z*-scored goodness ratings of extremely peripheral [æ] stimuli.

influence of phonological contrast on phonetic implementation is not binary, but rather varies continuously depending on information-theoretic factors like functional load and entropy. The present proposal, that phonetic implementation is governed primarily by phonological contrast, in combination with a gradient view of phonological contrast (Hall 2013), offers an intuitive explanation of these findings. More research is needed to understand the full range of effects of gradient contrastive factors on phonetic implementation.

The argument put forward here is theoretically compatible with either of the two broad approaches to the phonology-phonetics interface described in Section 1. Within the framework of generative phonology, the notion of phonological categories as abstract mental objects can be maintained; the present proposal would simply require that the output of the transduction function underlying phonetic implementation is not neuromuscular information corresponding to particular phonetic targets (c.f. Volenec and Reiss 2017), but rather *directions* in phonetic space. In this sense, the units of phonetic implementation might be best conceptualized as vectors with particular directions in relevant phonetic dimensions, and magnitudes corresponding to factors like cue weight (Shultz et al. 2012), degree of contrast (Hall 2013, Hall et al. 2017), and presence versus absence of a cue-specific minimal pair (Wedel et al. 2018).

The notion that phonetic implementation is driven by spatiotemporal forces is familiar in dynamic approaches to the phonology-phonetics interface (e.g. Hoole and Mooshammer 2002). However, the current proposal suggests that the forces driving the movements of the articulators might be better modeled with *repellers* rather than the *attractors* of AP. Of course, this is merely a speculative suggestion; working out the mathematical details and empirical implications of such a modification would require a great deal of future research.

Before concluding, I will briefly address a few potential criticisms of the present argument arising from both study limitations and theoretical problems. First, American English [æ] is known to exhibit variation due to dialectal differences and phonetic context, although this variation has been argued to be disappearing among younger speakers (Coggshall and Becker 2009). However, variation would be expected to *obscure* patterns in internal category structure, while the present results demonstrated relatively clear patterns. In addition, it might be argued that, phonologically, [y] differs from [i] only in terms of rounding, calling into question the status of [y] as a non-peripheral vowel. Although Turkish [y] is realized with consistent phonetic differences from [i] in height and backness (Radisic 2014), the present argument would be strengthened with evidence from a

more unambiguously non-peripheral vowel. A further reason to confirm these results with other non-peripheral vowels is that F3 likely plays an important role in [y] category goodness, which may have led to a de-emphasis of F1 and F2 as cues to category goodness, causing (or at least inflating) the greater randomness observed in the internal structure of [y] relative to [æ]. Next, as mentioned in Section 3.1.1, the native Turkish speakers in the present study all had substantial exposure to English, and it is known that exposure to an L2 can impact the L1 sound system (e.g. Chang 2013). Although unclear patterns of internal structure in non-peripheral vowels have already been observed in monolinguals (e.g. Sussman and Gekas 1997), it is possible that the present lack of a pattern was (at least partially) due to effects of L2 exposure. Evidence from monolingual Turkish speakers would therefore strengthen the present conclusions. Finally, it might be contended that the pattern observed here, whereby [æ] goodness appeared to increase indefinitely as F1 and F2 increased, was simply a result of participants adapting to the particular stimulus range with which they were presented (c.f. Lotto et al. 1998). This possibility was supported by a significant main effect of trial number in Experiment 1 and significant interactions between formants and trial number in both experiments, suggesting that participants only came to prefer the most peripheral vowels as they became familiar with the stimulus range. However, if category goodness is determined in part by the relationship of each sound to other relevant sounds (as suggested by stimulus range effects), this in itself can be seen as evidence for the role of contrast in determining category goodness, against a view in which phonetic implementation is based on particular targets.

In conclusion, it was observed that the peripheral vowel category [æ] exhibited clearer patterns of internal phonetic structure than the non-peripheral category [y] (Experiment 1), and the category goodness of [æ] sounds generally increased indefinitely as the first two formants increased, even as the sounds became extremely peripheral in the F1-F2 space (Experiment 2). I argued that these results support a view in which phonetic implementation is primarily driven by the maximization of contrast, rather than the achievement of particular phonetic targets. More research is needed to flesh out the details of the relationship of this proposal to existing models of the phonology-phonetics interface.

References

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed effects models using lme4. *Journal of Statistical Software* 67:1-48.

Browman, Catherine P., and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology* 6:201-251.

Chang, Charles B. 2013. A novelty effect in phonetic drift of the native language. *Journal of Phonetics* 41:520-533.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Coggshall, Elizabeth L., and Kara Becker. 2009. The vowel phonologies of African American and white New York City residents. *Publication of the American Dialect Society* 94:101-128.

Cohn, Abigail C., and Marie K. Huffman. 2014. Interface between phonology and phonetics. In *Oxford Bibliographies in Linguistics*, ed. M. Aronoff. New York: Oxford University Press.

Dresher, B. Elan. 2009. *The Contrastive Hierarchy in Phonology.* Cambridge: Cambridge University Press.

Drouin, Julia R., Rachel M. Theodore, and Emily B. Myers. 2016. Lexically guided perceptual tuning of internal category structure. *Journal of the Acoustical Society of America* 140:EL307-EL313.

Flemming, Edward. 2004. Contrast and perceptual distinctiveness. In *Phonetically-Based Phonology*, ed. B. Hayes, R. Kirchner, and D. Steriade. Cambridge: Cambridge University Press.

Gafos, Adamantios I., and Stefan Benus. 2006. Dynamics of phonological cognition. *Cognitive Science* 30:905-943.

Hall, Daniel Currie. 2011. Phonological contrast and its phonetic enhancement: Dispersedness with dispersion. *Phonology* 28:1-54.

Hall, Kathleen Currie. 2013. A typology of intermediate phonological relationships. *The Linguistic Review* 30:215-275.

Hall, Kathleen Currie, Hanna Smith, Kevin McMullin, Blake Allen, and Noriko Yamane. 2017. Using optical flow analysis on ultrasound of the tongue to examine phonological relationships. *Canadian Acoustics* 45:15-24.

Hauser, Ivy. 2019. Effects of phonological contrast on within-category phonetic variation. Doctoral disseration, University of Massachusetts Amherst.

Hoole, Phillip, and Christine Mooshammer. 2002. Articulatory analysis of the German vowel system. In *Silbenschnitt und Tonakzente*, ed. H. P. Auer, P. Gilles, and H. Spiekermann, 129-152. Tübingen: Niemeyer.
Iverson, Paul, and Patricia K. Kuhl. 2000. Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics* 62:874-886.
Johnson, Keith, Edward Flemming, and Richard Wright. 1993. The hyperspace effect: Phonetic targets are hyperarticulated. *Language* 69:505-528.
Keyser, Samuel Jay, and Kenneth Noble Stevens. 2006. Enhancement and overlap in the speech chain. *Language* 82:33-63.
Klatt, Dennis H. 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America* 67:971-995.
Klatt, Dennis H., and Laura C. Klatt. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87:820-857.
Kronrod, Yakov, Emily Coppess, and Naomi H. Feldman. 2016. A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin Review* 23:1681-1712.
Lenneberg, Eric. 1967. *Biological Foundations of Language*. New York: Wiley.
Liberman, Alvin M., Franklin S. Cooper, Donald P. Shankweiler, and Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review* 74:431-461.
Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54:358-368.
Liljencrants, Johan, and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48:839-862.
Lotto, Andrew J., Keith R. Kleunder, and Lori L. Holt. 1998. Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America* 103:3648-3655.
Miller, Joanne L. 1994. On the internal structure of phonetic categories: A progress report. *Cognition* 50:271-285.
Mücke, Doris, Anna Hermes, and Sam Tilsen. 2020. Incongruencies between phonological theory and phonetic measurement. *Phonology* 37:133-170.
Ohala, John J. 1990. There is no interface between phonetics and phonology: A personal view. *Journal of Phonetics* 18:153-171.
Peterson, Gordon E., and Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24:175-184.
Poeppel, David, and David Embick. 2005. Defining the relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics: Four cornerstones*, ed. A. Cutler, 103-120. New Jersey: Laurence Erlbaum Associates.
Radisic, Milica. 2014. An ultrasound and acoustic study of Turkish rounded/unrounded vowel pairs. Doctoral dissertation, University of Toronto.
Sapir, Edward. 1933. La réalité psychologique des phonèmes. *Journal de Psychologie Normale et Pathologique* 30:247-265.
Shultz, Amanda A., Alexander L. Francis, and Fernando Llanos. 2012. Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America* 132:EL95-EL101.
Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée, and Christian Abry. 1997. The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25:255-286.
Stern, Michael C. 2020. Testing the perceptual magnet effect in monolinguals and bilinguals. MA thesis, The Graduate Center, City University of New York.
Sussman, Joan E., and Brian Gekas. 1997. Phonetic category structure of [ɪ]: Extent, best exemplars, and organization. *Journal of Speech, Language, and Hearing Research* 40:1406-1424.
Volenec, Veno, and Charles Reiss. 2017. Cognitive phonetics: The transduction of distinctive features at the phonology-phonetics interface. *Biolinguistics* 11:251-294.
Wedel, Andrew, Noah Nelson, and Rebecca Sharp. 2018. The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language* 100.

Yale Linguistics Department
370 Temple St
New Haven, CT 06511
michael.stern@yale.edu