Comparing accounts of formant normalization against US English listeners' vowel perception

Anna Persson,[1] Santiago Barreda,[2] and T. Florian Jaeger[3]

[1] *Swedish Language and Multilingualism, Stockholm University, Stockholm, SE-106 91, Sweden*[a]

[2] *Linguistics, University of California, Davis*

[3] *Brain and Cognitive Sciences, Data Science, University of Rochester*

(Dated: 13 November 2024)

Human speech recognition tends to be robust, despite substantial cross-talker variability. Believed to be critical to this ability are auditory normalization mechanisms whereby listeners adapt to individual differences in vocal tract physiology. This study investigates the computations involved in such normalization. Two 8-way alternative forced-choice experiments assessed L1 listeners' categorizations across the entire US English vowel space—both for unaltered and for synthesized stimuli. Listeners' responses in these experiments were compared against the predictions of twenty influential normalization accounts that differ starkly in the inference and memory capacities they imply for speech perception. This includes variants of *estimation-free* transformations into psycho-acoustic spaces, *intrinsic* normalizations relative to concurrent acoustic properties, and *extrinsic* normalizations relative to talker-specific statistics. Listeners' responses were best explained by extrinsic normalization, suggesting that listeners learn and store distributional properties of talkers' speech. Specifically, *computationally simple* (single-parameter) extrinsic normalization best fit listeners' responses. This simple extrinsic normalization also clearly outperformed Lobanov normalization—a computationally more complex account that remains popular in research on phonetics and phonology, sociolinguistics, typology, and language acquisition.

[a]anna.persson@su.se

I. INTRODUCTION

One of the central challenges for speech perception originates in cross-talker variability: depending on the talker, the same acoustic signal can encode different sound categories (Allen *et al.*, 2003; Liberman *et al.*, 1967; Newman *et al.*, 2001). This results in ambiguity in the mapping from acoustics to words and meanings. Research has identified several mechanisms through which listeners resolve this ambiguity, ranging from early perceptual processes, to adaptation of phonetic categories, all the way to adjustments in post-linguistic decision processes (for review, see Xie *et al.*, 2023). The present study focuses on the first type of mechanism, early auditory processes that transform and normalize the acoustic input into the perceptual cues that constitute the input to linguistic processing (for reviews, Barreda, 2020; Johnson and Sjerps, 2021; McMurray and Jongman, 2011; Stilp, 2020; Weatherholtz and Jaeger, 2016). We seek to respond, in particular, to recent calls to put theories of adaptive speech perception to stronger tests (Baese-Berk *et al.*, 2018; Schertz and Clare, 2020; Xie *et al.*, 2023).

Evidence for the presence of early normalization mechanisms comes from neuroimaging and neurophysiological studies (e.g., Oganian *et al.*, 2023; Skoe *et al.*, 2021), as well as research on the peripheral auditory system suggesting automatic transformations of the acoustic signal into scale-invariant spectral patterns (e.g., Patterson and Irino, 2014; Smith *et al.*, 2005). Neurophysiological studies have further decoded effects of talker identity from subcortical brain areas like the brain stem, and thus prior to the cortical regions believed to encode linguistic categories (e.g., Sjerps *et al.*, 2019; Tang *et al.*, 2017). This includes brain responses that lag the acoustic signal by as little as 20-50 msecs (Lee, 2009), suggesting very fast and highly automatic processes. While this does not mean that *only* talker-normalized auditory percepts are available to subsequent processing—there is now convincing evidence that subcategorical information can enter listeners' phonetic representations (e.g., Hay *et al.*, 2017, 2019; Johnson *et al.*, 1999; McGowan, 2015; Walker and Hay, 2011)—it does suggest that normalized auditory percepts are available to subsequent processing. By removing (some) cross-talker variability early during auditory processing, normalization offers an elegant and effective solution that can reduce the need for more complex adaptive processes further upstream (Apfelbaum and McMurray, 2015; Xie *et al.*, 2023).

While it is relatively uncontroversial *that* normalization contributes to robust speech perception, it is still unclear what types of computations this implicates. We address this question for the perception of vowels, which cross-linguistically relies on peaks in the distribution of spectral energy over acoustic frequencies (formants).[1] Vowel perception has long been a focus in research on normalization (e.g., Bladon *et al.*, 1984; Fant, 1975; Gerstman, 1968; Johnson, 2020; Joos, 1948; Lobanov, 1971; Miller, 1989; Nearey, 1978; Nordström and Lindblom, 1975; Syrdal and Gopal, 1986; Traunmüller, 1981; Watt and Fabricius, 2002; Zahorian and Jagharghi, 1991; for review, see Barreda, 2020), with some reviews citing over 100 competing proposals (Carpenter and Govindarajan, 1993). Importantly, these accounts differ in the types and complexity of computations they assume to take place during normalization.

On the lower end of computational complexity, *estimation-free* psycho-acoustic transformations involve zero degrees of freedom that listeners would need to estimate from the acoustic input. For example, there is evidence that a transformation of acoustic frequencies (measured in Hz) into the psycho-acoustic Bark-space better describes how listeners perceive differences along the frequency spectrum (in terms of critical bands, e.g., Traunmüller, 1990; Zwicker, 1961; Zwicker *et al.*, 1957; Zwicker and Terhardt, 1980). It is thus possible that cross-talker variability in vowel pronunciations is reduced when formants are represented in Bark, rather than Hz. Similar arguments have been made about other psycho-acoustic transformations (e.g., ERB, Glasberg and Moore, 1990; Mel, Stevens and Volkmann, 1940; or semitones, Fant *et al.*, 2002) most of which share that they log-transform acoustic frequencies—in line with neurophysiological evidence that the auditory representations in the brain seem to follow a roughly logarithmic organization, so that auditory perception is (up to a point) more sensitive to differences between lower frequencies than to the same difference between higher frequencies (e.g., Merzenich *et al.*, 1975; for review, see Saenz and Langers, 2014). While each of these transformations was developed with different applications in mind (e.g., ERB and Bark to explain frequency selectivity, Glasberg and Moore, 1990; or semitones for the perception of musical pitch, Balzano, 1982), psycho-acoustic transformations might suffice for effective formant normalization. If so, this would offer a particularly parsimonious account of vowel perception as listeners would not have to infer talker-specific properties.

The parsimony of psycho-acoustic transformations contrasts with the majority of accounts for vowel normalization, which introduce additional computations. This includes accounts that normalize formants relative to other information that is available at the same point in the acoustic signal (intrinsic normalization, e.g., Miller, 1989; Peterson, 1961; Syrdal and

Gopal, 1986). For example, according to one proposal, listeners normalize vowel formants by the vowel's fundamental frequency or other formants estimated at the same point in time (Syrdal and Gopal, 1986). To the extent that the fundamental frequency is correlated with the talkers' vocal tract size (for review, see Vorperian and Kent, 2007), this allows the removal of physiologically-conditioned cross-talker variability in formant realizations. While such intrinsic accounts arguably entail more computational complexity than estimation-free transformations, they do not require that listeners *maintain* talker-specific estimates over time. This distinguishes intrinsic from extrinsic accounts, which introduce additional computational complexity.

According to extrinsic accounts, normalization mechanisms infer and store estimates of talker-specific properties that then are used to normalize subsequent speech from that talker (Gerstman, 1968; Lobanov, 1971; Nearey, 1978; Nordström and Lindblom, 1975; Watt and Fabricius, 2002; for review, see Weatherholtz and Jaeger, 2016). At the upper end of computational complexity, some accounts hold that listeners continuously infer and maintain both talker-specific means for each formant and talker-specific estimates of each formant's variability (Gerstman, 1968; Lobanov, 1971). These estimates are then used to normalize formants, e.g., by centering and standardizing them (essentially z-scoring formants, Lobanov, 1971), removing cross-talker variability in the distribution of formant values. There are, however, more parsimonious extrinsic accounts that require inference and maintenance of fewer talker-specific properties. The most parsimonious of these is Nearey's *uniform scaling* account, which assumes that listeners infer and maintain a single talker-specific parameter. This parameter (Ψ) can be thought of as capturing the effects of the talker's vocal tract length on the spectral scaling applied to the formant pattern produced by a talker (Nearey, 1978).[2] Uniform scaling deserves particular mention here as it is arguably one of the most developed normalization accounts, and rooted in principled considerations about the physics of sound and the evolution of auditory systems (for review, see Barreda, 2020).

In summary, hypotheses about the computations implied by formant normalization differ in the flexibility they afford as well as the inference and memory complexity they entail. Considerations about the complexity of inferences—essentially the number of parameters that listeners are assumed to estimate at any given moment in time—arguably gain in importance in light of the speed at which normalization seems to unfold. In the present study, we thus ask whether computationally simple accounts are sufficient to explain human vowel perception.

While previous research has compared normalization accounts across languages, most of this work has evaluated proposals in terms of how well the normalized phonetic space supports the separability of vowel categories (Adank *et al.*, 2004; Carpenter and Govindarajan, 1993; Cole *et al.*, 2010; Escudero and Bion, 2007; Johnson and Sjerps, 2021; Syrdal, 1985). This approach is illustrated in Figure 1. These studies have found that computationally more complex accounts—which also afford more flexibility—tend to achieve higher category separability and higher categorization accuracy (for review, see Persson and Jaeger, 2023). This includes Lobanov normalization, which continues to be highly influential in, for example, variationist and sociolinguistic research because of its effectiveness in removing cross-talker variability (for a critique, see Barreda, 2021). It is, however, by no means clear that human speech perception employs the same computations that achieve the best cate-

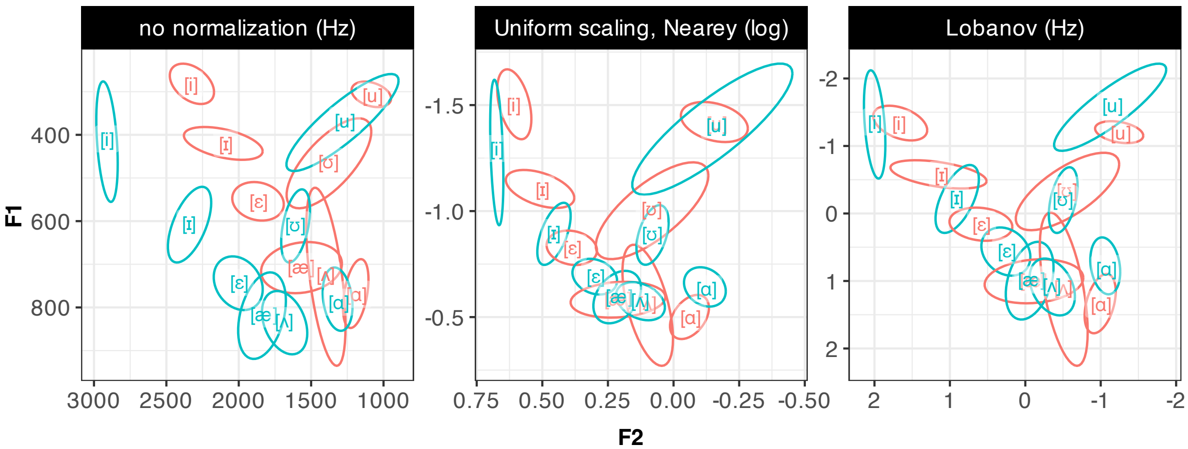


FIG. 1. Illustration of how height, which is positively correlated with vocal tract size, affects vowels' F1 and F2, and how normalization can partially remove this effect. Shown here are realizations of eight monophthong vowels of US English by a short (cyan) and a tall native talker (red). **Panel A:** In the acoustic space, prior to any normalization (Hz). **Panel B:** After uniform scaling (Nearey, 1978). **Panel C:** After Lobanov normalization (Lobanov, 1971). The present study compares these three accounts, along with 17 other normalization accounts. Here and throughout the paper, panel captions indicate the phonetic space in which normalization takes place in parenthesis. Note that this is not necessarily identical to the units of F1 and F2 *after* normalization (e.g., Lovabov normalization results in scale-free z-scores along the formant axes).

gory separability or accuracy (see also discussion in Barreda, 2021; Nearey and Assmann, 2007).

A substantially smaller body of research has addressed this question by comparing normalization accounts against *listeners' perception* (Barreda and Nearey, 2012; Barreda, 2021; Nearey, 1989; Richter *et al.*, 2017; for a review, see Whalen, 2016). Interestingly, these works seem to suggest that computationally simpler accounts might provide a better fit against human speech perception than the influential Lobanov model (Barreda, 2021; Richter *et al.*, 2017). For example, Barreda (2021) compared the predictions of uniform scaling and Lobanov normalization against listeners' categorization responses in a forced-choice categorization task over parts of the US English vowel space. In his experiment, listeners' categorization responses were better predicted by uniform scaling than by Lobanov normalization. Findings like these suggest that comparatively simple corrections for vocal tract size—such as uniform scaling—might provide a better explanation of human perception than more computationally complex accounts (see also Johnson, 2020; Richter *et al.*, 2017).

This motivates the present work. We take a broad-coverage approach by comparing the 20 normalization accounts in Table 1 against the perception of eight monophthongs of US English ([i] as in *heed*, [ɪ] in *hid*, [ɛ] in *head*, [æ] in *had*, [ʌ] in *hut*, [ʊ] in *hood*, [u] in *who'd*, [ɑ] in *odd*).[3] We do so for the perception of both natural and synthesized speech. Our broad-coverage approach complements previous studies, which have typically compared a small number of accounts (up to 3) and focused on parts of the vowel inventory, and thus parts of the formant space (typically 2-4 vowels, Barreda, 2021; Barreda and Nearey, 2012; Nearey, 1989; Richter *et al.*, 2017). The accounts we consider include the most influential examples of psycho-acoustic transformations (Fant *et al.*, 2002; Glasberg and Moore, 1990; Stevens and Volkmann, 1940; Traunmüller, 1981), intrinsic (Syrdal and Gopal, 1986), extrinsic (Gerstman, 1968; Johnson, 2020; Lobanov, 1971; McMurray and Jongman, 2011; Nearey, 1978; Nordström and Lindblom, 1975), and hybrid accounts that contain intrinsic and extrinsic components (Miller, 1989). This broad-coverage approach allows us to assess, for example, whether the preference for computationally simple accounts observed in Barreda (2021) replicates on new data that span the entire vowel space. It also allows us to ask whether accounts even simpler than uniform scaling—such as psycho-acoustic transformations—provide an even better fit to human perception.

TABLE. I. Normalization accounts considered in the present study. Unless otherwise marked, formant variables (Fs) in the right-handside of normalization formulas are in Hz.

	Normalization procedure	Perceptual scale	Source	Formula
	No normalization	**Hz**	**n/a**	**n/a**
trans-formation	—	log		$F_n^{log} = ln(F_n)$
	—	Bark	Traunmüller (1990)	$F_n^{Bark} = \frac{26.81 \times F_n}{1960+F_n} - 0.53$
	—	ERB	Glasberg & Moore (1990)	$F_n^{ERB} = 21.4 \times \log_{10}(1 + F_n \times 0.00437)$
	—	Mel	Stevens & Volkmann (1940)	$F_n^{Mel} = 2595 \times \log_{10}(1 + \frac{F_n}{700})$
	—	Semitones conversion	Fant et al. (2002)	$F_n^{ST} = 12 \times \frac{ln(\frac{F_n}{100})}{ln}$
intrinsic	Syrdal & Gopal 1 (Bark-distance model)	Bark	Syrdal & Gopal (1986)	$F1^{SyrdalGopal1} = F1^{Bark} - F0^{Bark}$ $F2^{SyrdalGopal1} = F2^{Bark} - F1^{Bark}$
	Syrdal & Gopal 2 (Bark-distance model)			$F1^{SyrdalGopal2} = F1^{Bark} - F0^{Bark}$ $F2^{SyrdalGopal2} = F3^{Bark} - F2^{Bark}$
	Miller (formant-ratio)	log	Miller (1989)	$SR = k(\frac{GMf0}{k})^{1/3}$ $F1^{Miller} = log(\frac{F1}{SR})$ $F2^{Miller} = log(\frac{F2}{F1})$ $F3^{Miller} = log(\frac{F3}{F2})$
extrinsic centering	Nearey's uniform scaling	log	Nearey (1978)	$F_n^{Nearey} = \ln(F_n) - mean(ln(F))$
	Nordström & Lindblom (vocal tract scaling)	Hz	Nordström & Lindblom (1975)	$F_n^{NordströmLindblom} = \frac{F_n}{mean(\frac{F3[F1>600]}{2.5})}$
	Johnson (average formant spacing)	Hz	Johnson(2020)	$F_n^{Johnson} = \frac{F_n}{mean(\frac{F1}{0.5}, \frac{F2}{1.5}, \frac{F3}{2.5})}$
	Nearey's formantwise log-mean	log	Nearey (1978)	$F_n^{Nearey} = \ln(F_n) - mean(ln(F_n))$
	C-CuRE	Hz	McMurray & Jongman (2011)	$F_n^{C-CuRE} = F_n - mean(F_n)$
	—	Bark		
	—	ERB		
	—	Mel		
	—	Semitones conversion		
extrinsic standardizing	Gerstman (range normalization)	Hz	Gerstman (1968)	$F_n^{Gerstman} = 999 \times \frac{F_n - F_n^{min}}{F_n^{max} - F_n^{min}}$
	Lobanov (z-score)	Hz	Lobanov (1971)	$F_n^{Lobanov} = \frac{F_n - mean(F_n)}{sd(F_n)}$

Next, we motivate and describe the two experiments we conducted. Then we compare the normalization accounts in Table 1 against listeners' responses from these experiments.

A. Open Science Statement

All stimulus recordings, results, and the code for the experiment, data analysis, and computational modeling for this article can be downloaded from the Open Science Framework (OSF) at `https://osf.io/zemwn/`. The OSF repository also include extensive supplementary information (SI). Both the article and SI are written in R markdown, allowing readers to replicate our analyses with the click of a button, using freely available software (R Core Team, 2023; RStudio Team, 2020). Readers can revisit the assumptions we committed to for the present project—for example, by substituting alternative normalization accounts or categorization models. Researchers can also substitute their own experiments on vowel normalization for our Experiments 1a and 1b, to see whether our findings generalize to novel data. We see this as an important contribution of the present work, as it should make it substantially easier to consider additional normalization accounts—including variants to the accounts we considered—and to assess the generalizability of the conclusions we reach based on the present data.

II. EXPERIMENTS 1A AND 1B

To compare the performance of different normalization accounts against listeners' perception, we conducted two small web-based experiments on US English listeners' perception of US English vowels. Both experiments investigate listeners' perception of a single talker.

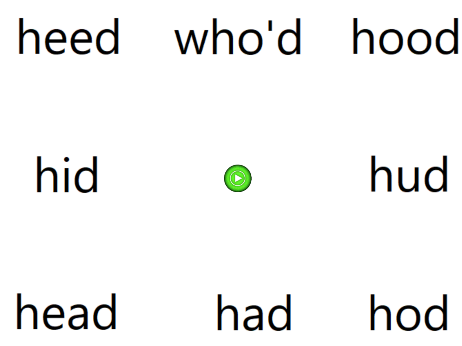


FIG. 2. Screen shot of the eight-alternative forced-choice (8-AFC) task used in both Experiment 1a and 1b.

This choice was made so as to not confound questions about formant normalization with questions about talker recognition, and inferences about talker switches (Magnuson and Nusbaum, 2007). The two experiments employ the same eight-alternative forced-choice vowel categorization task (Figure 2), and differ only in the whether they employed 'natural' (Experiment 1a) or synthesized stimuli (Experiment 1b). To the best of our knowledge, these two experiments are the first designed to compare normalization accounts against listeners' perception over a larger portion of the monophthong inventory of a language.

Experiment 1a employs recordings of *hVd* word productions from a female talker of US English, these recordings are 'natural' in the sense that they were not synthesized or otherwise phonetically manipulated. One consequence of this is that the formant values of these recordings are clustered around the talker's category means, and thus span only a comparatively small part of the phonetic space. This can limit the statistical power to distinguish between competing accounts. Natural recordings furthermore vary not only along the primary cues to vowel quality in US English (F1, F2) but also along secondary cues (e.g., F0, F3, vowel duration, and vowel inherent spectral change—VISC) as well as other unknown properties, which can make it difficult to discern whether the performance of a normalization model is due to the normalization itself or other reasons, e.g., because a normalized cue happens to correlate with another cue that listeners are sensitive to but that is not included in the model.

Experiment 1b thus adopts an alternative approach and uses synthesized vowels. Unlike most previous work, which has used isolated vowels as stimuli (Barreda, 2021; Barreda and Nearey, 2012; Nearey, 1989; Richter *et al.*, 2017), Experiment 1b uses synthesized *hVd* words to facilitate comparison to Experiment 1a. This allowed us to sample larger parts of the F1-F2 space, which has two advantages. First, it allowed us to collect responses over parts of the formant space for which we expect listeners to have more uncertainty, and thus exhibit more variable responses. This can increase the statistical power to distinguish between competing accounts. Second, differences in the predictions of competing normalization accounts will tend to become more pronounced with increasing distance from the category centers. By collecting responses at those locations, we can thus increase the contrast between competing accounts. Critically, an adequate model of formant normalization needs to capture human perception not only for prototypical vowel instances, but also instances of vowels that fall between category means.

The use of synthesized stimuli does, however, also come with potential disadvantages. Synthesized stimuli can suffer in ecological validity, lacking correlations between cues, and across the speech signal (e.g., due to co-articulation) that are characteristic of human speech. This raises questions about the extent to which processing of such stimuli engages the same mechanisms as everyday speech perception. Additionally, it is possible that the use of robotic sounding synthesized speech affects listener engagement. This can lead to an increased rate of attentional lapses, and thus a decrease in the proportion of trials on which listeners' responses are based on the acoustics of the speech stimulus rather than random guessing (compare, e.g., Kleinschmidt, 2020; Tan and Jaeger, 2024). By comparing normalization accounts against both natural and synthesized stimuli, we investigate the extent to which the accounts that best describe human perception depend on the type of stimuli used in the experiment.

A. Methods

1. Participants

We recruited 33 (Experiment 1a) and 33 (Experiment 1b) participants. The majority of these (24 for each experiment) were recruited from Amazon's Mechanical Turk. However, after exclusions we were left with a relatively low number of participants (for Experiment 1a, 19, and for Experiment 1b, 22). We therefore decided to recruit an additional 18 participants from Prolific (9 for each experiment; October 2024). Exclusions described below left 28 and 31 participants for analysis in Experiments 1a and 1b, respectively. Results did not change after inclusion of the new participants from Prolific.

Participants were paid \$6/hour (\$12/hour on Prolific) prorated by the duration of the experiments (15 minutes). Participants only saw the experiment advertised, and could only participate in it, if (i) they were located within the US, (ii) had an approval rating of

99% or higher, (iii) met the software requirements (a recent version of the Chrome browser engine), and (iv) had not previously completed any other experiments on vowel perception in our lab. Before the experiment could be accepted, participants had to confirm that they were (i) native speakers of US English (defined as having spent their childhood until the age of 10 speaking English and living in the United States), (ii) in a quiet room without distractions, (iii) wearing over-the-ear headphones. Participants' responses were collected via Javascript developed by the Human Language Processing Lab at the University of Rochester (Kleinschmidt *et al.*, 2021).

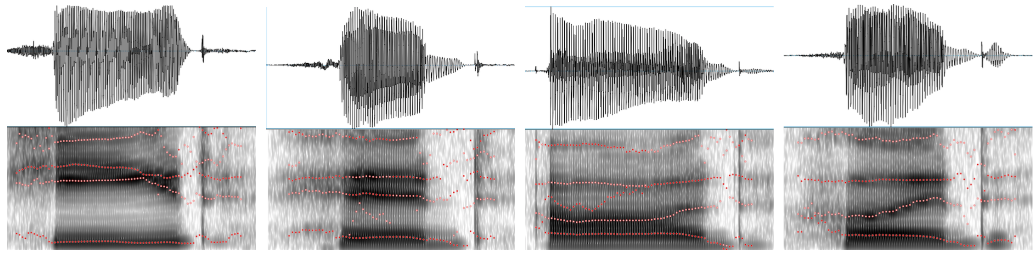
An optional post-experiment survey recorded participant demographics using NIH prescribed categories, including participant sex (Male: 36, Female: 29), age (mean = 36.9 years; SD = 12.2; 95% quantiles = 22.6-66 years), race (White: 48, multiple: 3, Black: 10, Asian: 3, declined to report: 1), and ethnicity (Non-Hispanic: 60, Hispanic: 4, declined to report: 1). All but 1 participant completed the survey.

2. Materials

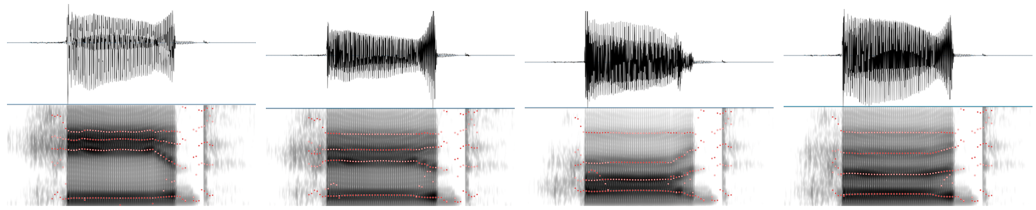
Experiment 1a employed *hVd* word recordings by one adult female talker of a Northeastern dialect (spoken in central Connecticut) from a phonetically annotated database of L1-US English vowel productions (Xie and Jaeger, 2020). Specifically, we used all nine recordings of each of the eight *hVd*-words—*heed*, *hid*, *head*, *had*, *hut*, *odd*, *hood*, *who'd* (the use of "hut" and "odd" rather than "hud" and "hod" follows Assmann *et al.*, 2008; but see Hillenbrand *et al.*, 1995).

The stimuli for Experiment 1b were synthesized from a single *had* recording used in Experiment 1a (see Figure 3 for example spectrograms). Specifically, we used a script (based on descriptions in Wade *et al.*, 2007) in Praat (Boersma and Weenink, 2022) to concatenate the original /h/ with a synthesized vowel and the original /d/ recording. Unlike in Experiment 1a, all eight words thus had an *hVd* context (including "hud" and "hod", rather than "hut" and "odd"). The Praat script first segmented the original *had* token into the three segments /h/, /ae/ and /d/, with the /d/ segment consisting of the voiced closure and burst. The script then estimated the spectral envelope of the /h/ sound by linear predictive coding (LPC; autocorrelation method), and used the resulting coefficients to inversely filter the /h/. This resulted in an /h/ sound with effects of vocal tract removed, leaving the source signal. Next, a glottal waveform was generated at each point in the pitch contour from the original /ae/ sound using the point process to phonation functionality in Praat. This waveform was multiplied with the intensity pattern from the same original /ae/ sound. The resulting sound was concatenated with the neutral fricative /h/ sound, to create a neutral hV-section that did not reflect any vocal tract resonances. The script then created a formant grid that filtered the hV-section to create the intended vowel, and finally concatenated this segment to the final /d/ to create an *hVd* word. For each *hVd* word, the formant grid was populated with the F1, F2 and F3 values that we handed to the script at five time-points transitioning from the /h/ to the steady-state vowel, to the first portion of the voiced closure of the final /d/ segment through linear interpolation, thus holding formants steady until transitioning into the final consonant. Formant bandwidths were 500 Hz at the initial two time-points (the /h/ and beginning of transition to vowel), and then decreased linearly during vowel onset and throughout the final three time-points to 50 Hz (F1), 100 Hz (F2), 200 Hz (F3), 300 Hz (F4), and 400 Hz (F5-F8, following Wade *et al.*, 2007). The bandwidth manipulation implied that the spectral peaks of the formants became more defined and more separated as the vowel unfolded. We used this approach to create synthesized vowels for arbitrary F1-F2 combinations. F3 was set based on those F1-F2 values. Specifically, we ran a linear regression over the natural productions of the talker from Experiment 1a, predicting F3 from F1, F2 and their interaction. We then used that regression to predict F3 values for any F1-F2 combination in Experiment 1b. F4 to F8, as well as vowel duration, were held identical across all tokens (using the automatically extracted vowel duration and mean formant values across the vowel segment from the *had* token used for resynthesis).

We generated 146 synthesized *hVd* recordings that spanned the F1 and F2 space. The specific F1-F2 locations chosen were determined by a mix of modeling (using ideal observers described in the next section to predict listeners' categorization responses) and intuition. Specifically, we selected 64 recordings that we expected to fall within the bivariate 95% confidence intervals (CIs) of the eight US English monophthongs, and 82 recordings that we expected to fall between those CIs. Figure 4 under *Results* shows the distribution of stimuli for both experiments. Of note, our procedure also generated formant combinations that are physiologically unlikely to have all been produced by the same talker during 'normal' vowel production (also known as "off-template" instances, Nearey, 1978).



(a) natural *heed* (b) natural *hid* (c) natural *odd* (d) natural *hood*



(e) resynthesized *heed* (f) resynthesized *hid* (g) resynthesized *hod* (h) resynthesized *hood*

FIG. 3. **Top:** Spectrograms of four natural recordings from Experiment 1a. **Bottom:** Same for four synthesized tokens with similar formant values from Experiment 1b. Additional spectrograms are provided in the SI §2 C.

3. Procedure

The procedure for both experiments was identical. Live instances of each experiment can be found at `https://www.hlp.rochester.edu/experiments/DLPL2S/experiment-A/experiments.html`. At the start of the experiment, participants acknowledged that they met all requirements and provided consent, as per the Research Subjects Review Board of the University of Rochester. Before starting the experiment, participants performed a sound check. Participants were then instructed to listen to a female talker saying words, and click on the word on screen to report what word they heard. On each trial, all eight *hVd*-words were displayed on screen. Half of the participants in each experiment saw the response options organized as in Figure 2 (resembling the IPA representation of a vowel space), half saw the response options in the opposite order (flipping top and bottom and left and right in Figure 2). Each trial started with the response grid on screen, together with a light green dot centered on screen. After 1000 ms, an *hVd* recording played, and participants indicated their response by a mouse-click. After a 1000 ms intertrial interval, the screen reset, and the next trial started.

In both experiments, participants heard two blocks of the materials described in the previous sections, for a total of 144 trials in Experiment 1a and 292 trials in Experiment 1b. Presentation within each block was randomized for each participant in order to reduce confounds due to stimulus order (known to affect vowel perception, Repp and Crowder, 1990, and references therein). Participants were not informed about the block structure of the experiment.

After completing the experiment, participants filled out a language background questionnaire and the optional demographic survey. On average, participants took 9.3 minutes to complete Experiment 1a (SD = 5.5) and 17.9 minutes for Experiment 1b (SD = 6.5).

4. Exclusions

We excluded participants who failed to follow instructions and did not wear over-the-ear headphones (as indicated in the post-experiment survey). We also excluded participants with mean (log-transformed) reaction times that were unusually slow or fast (absolute z-score over by-participant means > 3), or if they clearly did not do the task (e.g., by answering randomly). This excluded 5 participants from Experiment 1a and 2 from Experiment 1b (for details, see SI §2 A).

We further excluded all trials that were unusually fast or slow. Specifically, we first z-scored the log-transformed response times *within each participant* and then z-scored these z-scores *within each trial* across participants. Trials with absolute z-scores > 3 were removed from analysis. This double-scaling approach was necessary as participants' response times decreased substantially over the first few trials and then continued to decrease less rapidly throughout the remainder of the experiment. The approach removes response times that are unusually fast or slow *for that participant at that trial*, while avoiding specific assumptions about the shape of the speed up in response times across trials. This excluded 1.3% of the trials in Experiment 1a and 0.9% in Experiment 1b. This left for analysis 3983 observations from 28 participants in Experiment 1a, and 8970 observations from 31 participants in Experiment 1b.

B. Results

Participants' categorization responses in Experiments 1a and 1b are shown in Figure 4, with larger labels indicating recordings that participants agreed on more.[4] We make two observations. The first pertains to the degree of (dis)agreement between the two experiments. The second observation pertains to the degree of (dis)agreement across participants within each experiment.

1. Similarities and differences between Experiments 1a and 1b

Unsurprisingly, participants in both experiments divided the F1-F2 space into the eight vowel categories in ways that qualitatively resembled each other (after taking into account

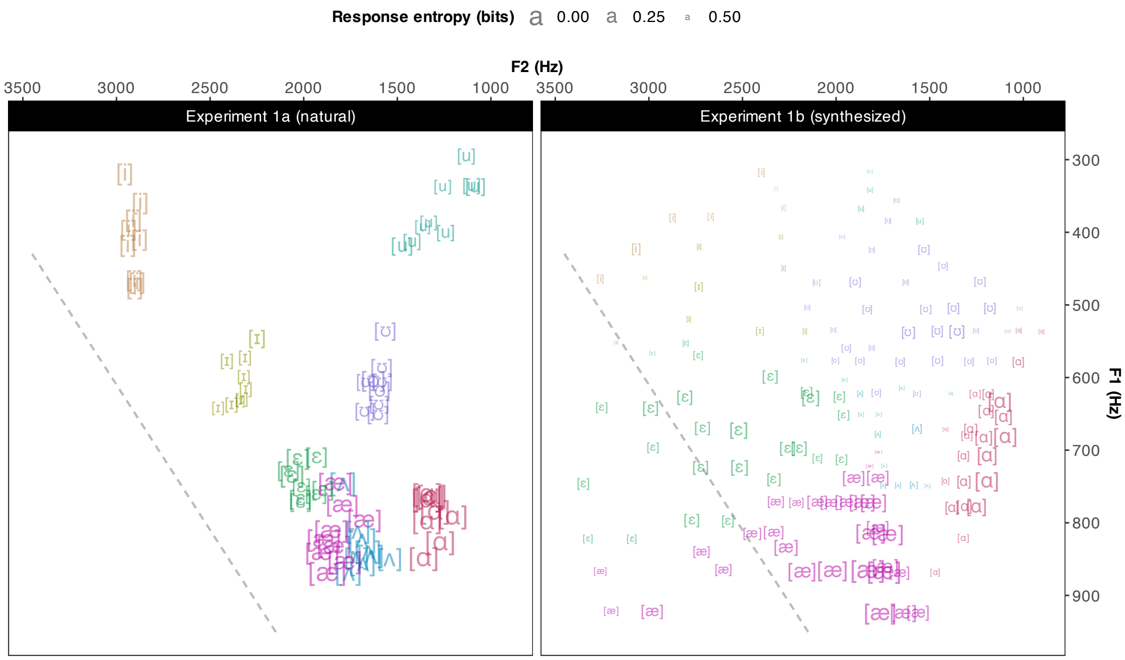


FIG. 4. Summary of listeners' categorization responses in Experiments 1a and 1b in F1-F2 space. The vowel label indicates the most frequent response provided across participants on each test location. Size indicates how consistent responses were across participants, which larger symbols indicating more consistent responses (lower entropy). F1-F2 combinations below the gray dashed line are unlikely to be articulated by the same talker.

that Experiment 1b covers a larger range of F1-F2 values). Also unsurprisingly, there were some differences between participants' responses across the two experiments, at least when plotted in Hz. For example, [u] rarely was the most frequent response in Experiment 1b, even for stimuli with similar F1-F2 values that were predominantly categorized as [u] in Experiment 1a. There are at least two reasons to expect such differences. First, stimuli with similar F1-F2 values across the two experiments still differed in other acoustic properties (e.g. vowel duration or F3). These acoustic differences might have affected participants' responses. Second, it is possible that *formant normalization* affected participants' responses—i.e., the very mechanism we seek to investigate in the remainder of the paper. The two experiments differ in the means, variances, and other statistical properties that some normalization accounts predict to affect perception. As a consequence, Hz might not be the space in which we should expect identical responses across experiments.

Similarly, the two experiments differed in the extent to which participants agreed with each other. Participants in Experiment 1b exhibited overall less agreement in their responses (mean by-item response entropy = 0.45 bits, SE = 0.01) than participants in Experiment 1a (mean by-item response entropy = 0.19 bits, SE = 0.02). This was also confirmed by participants' responses during the post-experiment survey. Compared to participants in Experiment 1a, participants in Experiment 1b reported increased uncertainty about their responses, and that the stimuli were less distinguishable and more robotic-sounding (see SI §2 B).

This increased uncertainty in Experiment 1b was expected—and, indeed, intended by the design: Experiment 1b explored the entire F1-F2 space, including formant combinations located *between* the centers of the natural vowel categories. Experiment 1b therefore achieved its goal of eliciting less categorical response distributions, which is expected to facilitate comparison of competing normalization accounts.[5]

Auxiliary analyses presented in the SI (§2 E) suggest that *some but not all* of the differences in response entropy between the two experiments were caused by the placement of the stimuli in formant space: when comparing categorization responses for tokens from the two experiments with similar acoustic properties (differences of $\leq$ 30 Hz along F1 and F2), response entropies still differed substantially (for N = 40 acoustically similar tokens, mean by-item response entropy for Experiment 1a = 0.14 bits, SE = 0.02; Experiment 1b = 0.4

bits, SE = 0.03). The same section of the SI (§2 E) presents additional analyses grouping acoustically similar tokens in the phonetic space defined by the normalization account we find to best fit listeners' responses. These analyses support the same conclusion.

We see two mutually compatible explanations to this difference in listener agreement between experiments. First, similar to the differences between experiments in the dominant response pattern discussed above, differences in the degree of agreement between participants might originate in *normalization*. Second, it is possible that the relation between formants in the synthesized stimuli or some other unknown acoustic-phonetic differences between the experiments explain the difference in response. For example, the absence of VISC or differences in spectral tilt in the synthesized stimuli might have deprived listeners of information that is actually crucial for establishing phonemic identity (Hillenbrand and Nearey, 1999). This would result in increased uncertainty on each trial, leading to increased entropy of listeners' responses. The computational study we present below shed some light on these two mutually compatible possibilities.

2. Similarities and differences between participants

Since the intended category was known for Experiment 1a, it was possible to calculate participants' recognition accuracy. As also evident in the left panel of Figure 4, participants' most frequent response *always* matched the intended vowel in Experiment 1a. Overall, participants' responses matched the intended vowel on 84.7% (SE = 3.5%) of all trials (Experiment 1b had no such ground truth). This is much higher than chance (12.5%). It is, however, also quite a bit lower than 100%. To better understand the reasons for this, Figure

5A plots the confusion matrix. This suggests that participants' performance was largely affected by confusions between [ɪ]-to-[ɛ] (*hid*-to-*head*), [ɛ]-to-[æ] (*head*-to-*had*), and [u]-to-[ʊ] (*who'd*-to-*hood*).

One plausible explanation for this pattern of vowel confusions lies in the substantial variation that exists across US English dialects (Labov *et al.*, 2006). Differences in the realization of vowel categories, and associated representations, across dialects will directly affect the expected classification for any given token. In addition, listeners might differ in terms of experience with different dialects, or in the dialect they attribute to the talker who produced the stimuli. To test this hypothesis, we calculated the [ɪ]-to-[ɛ], [ɛ]-to-[æ], and [u]-to-[ʊ] confusion rates for each participant in Experiment 1a. These data are summarized in the left panel of Figure 5B. The data in the left panel suggest that most participants in Experiment 1a either heard [ɪ] tokens consistently as the intended [ɪ] (clustering on the left side of the panel) or as [ɛ] (clustering on the right side of the panel). Only a few participants exhibited mixed responses for items intended to be [ɪ]. Tellingly, many of the participants who exhibited increased [ɪ]-to-[ɛ] confusion *also* exhibited increased [ɛ]-to-[æ] confusion. This is precisely what would be expected by listeners who assume a dialect in which these vowels are articulated lower (with higher F1) than in the dialect of the talker in Experiment 1a. A similar, but less pronounced, pattern was also found with regard to [u]-to-[ʊ] confusions.[6] Finally, a qualitatively similar relation between [ɪ]-to-[ɛ], [ɛ]-to-[æ], and [u]-to-[ʊ] confusions was also observed in Experiment 1b (right panel of Figure 5B), though the pattern was unsurprisingly less pronounced given that the stimuli in Experiment 1b by design often fell into the ambiguous region *between* vowels. Taken together, vowel-to-vowel confusion rates

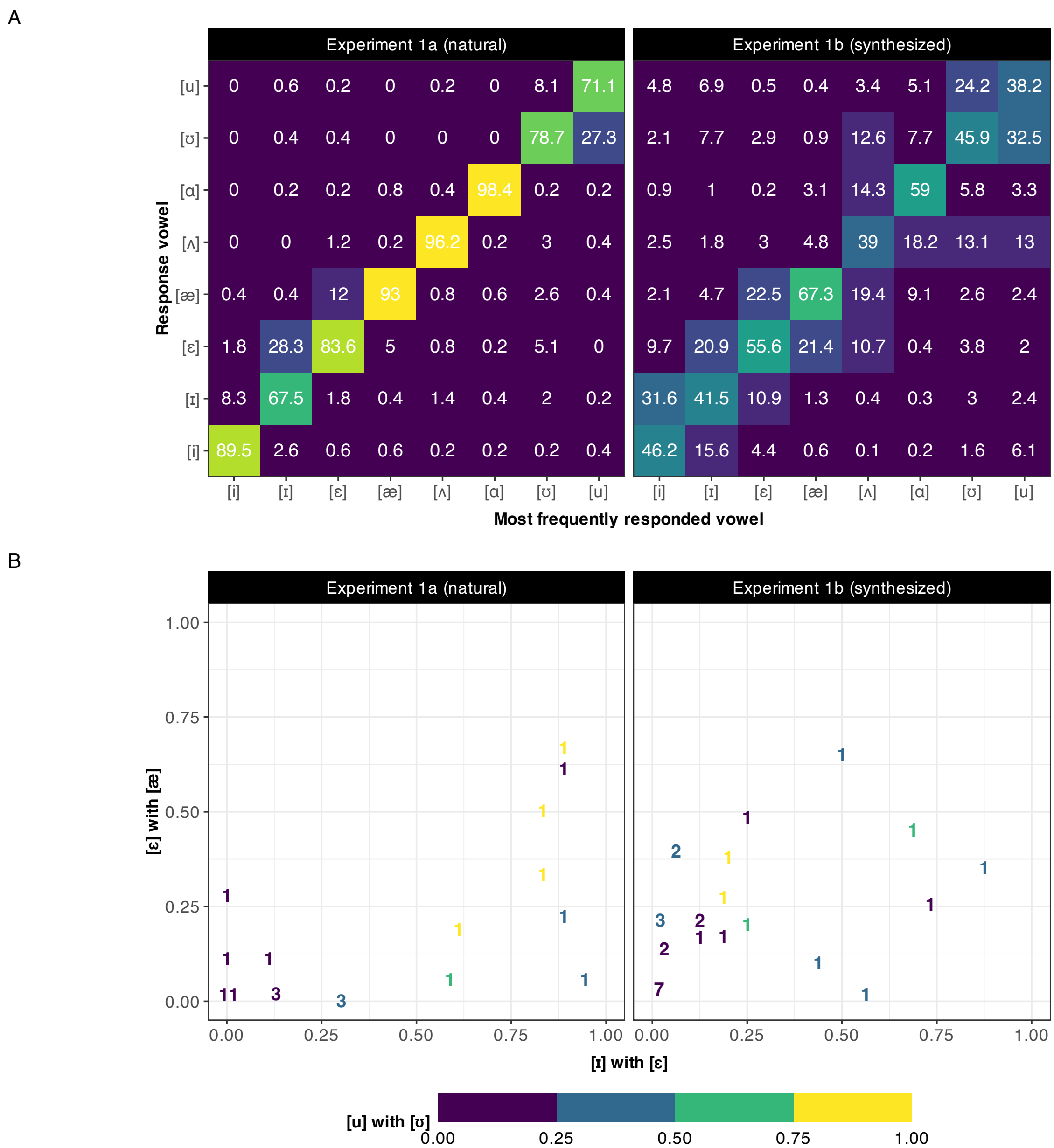


FIG. 5. Category confusability in Experiments 1a and 1b. **Panel A** summarizes the category confusability. Since correct responses were not defined for Experiment 1b, we grouped items along the x-axis based on most frequent response that listeners provided (for Experiment 1a, this was always identical to the intended response). Response percentages sum to 100 in each column, showing the response distribution depending on the most frequent response. **Panel B** summarizes individual differences across listeners, in terms of the listener-specific confusability of [ɪ] with [ɛ] (x-axis), [ɛ] with [æ] (y-axis), and [u] with [ʊ] (color fill).

in Experiments 1a and 1b suggest that systematic dialectal differences contributed to the relatively low categorization accuracy.

This highlights two important points. First, the data from Experiment 1a demonstrate the perceptual challenges associated with an unfamiliar talker: in the absence of lexical or other context to distinguish between the eight available response options, listeners can only rely on the acoustic information in the input. In such a scenario, even listeners who are in principle familiar with the dialect spoken by the talker have comparatively little information to determine the talker's dialect, making apparent what Winn (2018) aptly summarizes as "speech [perception] is not as acoustic as [we] think". Second, when dialect variability is taken into account, listeners' recognition accuracy improved substantially. After removing 8 listeners who heard more than 50% of the [ɪ] items as [ɛ], *all* vowels were correctly recognized at least 87.1% of the time (overall accuracy = 94.8%). This suggests that dialect differences affected the recognition of all vowels. This aspect of our results serves as an important reminder that formant normalization is only expected to erase inter-talker variability associated with *physiological* differences: variation in dialect, sociolect, or other non-physiologically-conditioned variation pose separate challenges to human perception, and require additional mechanisms (see discussion in Barreda, 2021; Weatherholtz and Jaeger, 2016). This introduces noise—variability in listeners' responses that cannot be accounted for by normalization—to any comparison of normalization accounts, potentially reducing the power to detect differences between accounts.

III. COMPARISON OF NORMALIZATION ACCOUNTS

In order to evaluate normalization accounts against speech perception, it is necessary to map the phonetic properties of stimuli—under different hypotheses about normalization—onto listeners' responses in Experiments 1a and 1b. Previous work has done so by directly predicting listeners' responses from the raw or normalized phonetic properties of stimuli (Apfelbaum and McMurray, 2015; Barreda, 2021; Crinnion *et al.*, 2020; McMurray and Jongman, 2011; Nearey, 1989). For example, McMurray and Jongman used multinomial logistic regression to predict eight-way fricative categorization responses in US English (see also Barreda, 2021).

Here we pursued an alternative approach by committing to a core assumption common to contemporary theories of speech perception: that listeners acquire implicit knowledge about the probabilistic mapping from acoustic inputs to linguistic categories, and draw on this knowledge during speech recognition (e.g., TRACE, McClelland and Elman, 1986; exemplar theory, Johnson, 1997; Bayesian accounts, Luce and Pisoni, 1998; Nearey, 1990; Norris and McQueen, 2008; ASR-inspired models like DIANA or EARSHOT, ten Bosch *et al.*, 2015; Magnuson *et al.*, 2020). Using a general computational framework for adaptive speech perception (ASP, Xie *et al.*, 2023) we trained Bayesian ideal observers to capture the expectations that a 'typical' L1 adult listener might have about the formant-to-vowel mappings of US English. We approximated these expectations using a database of L1-US English vowel productions (Xie and Jaeger, 2020)—transformed to reflect the different normalization accounts. We then ask which of the different ideal observer models—corresponding to different hypotheses about formant normalization—best predicts listeners' responses in Experiments 1a and 1b.

Training ideal observers on a database of vowel productions has the advantage that it reduces the degrees of freedom (DFs) used to predict listeners' responses. For example, using ordinary multinomial logistic regression trained on our perceptual data to predict eight-way vowel categorization as a function of F1, F2 and their interaction would require up to 28 DFs. This problem increases with the number of cues considered. By instead training ideal observers on phonetic data that are independent of listeners' responses, the ASP-based approach we employ uses only two DFs to mediate the mapping from stimuli properties to listeners' responses, regardless of the number of cues considered. Over the next few sections, we describe how this parsimony is made possible through a commitment to strong linking hypotheses motivated by theories of speech perception.

A. Methods

1. A general-purpose categorization model for J-AFC categorization tasks

Figure 6 summarises ASP's categorization model for a J-alternative forced-choice task (for an in-depth description, we refer to Xie *et al.*, 2023). The model combines Bayesian ideal observers (as used in e.g., Clayards *et al.*, 2008; Feldman *et al.*, 2009; Norris and McQueen, 2008; Xie *et al.*, 2021; for a closely related approach, see also Nearey and Hogan, 1986) with psychometric lapsing models (Wichmann and Hill, 2001). To reduce researchers' degrees of

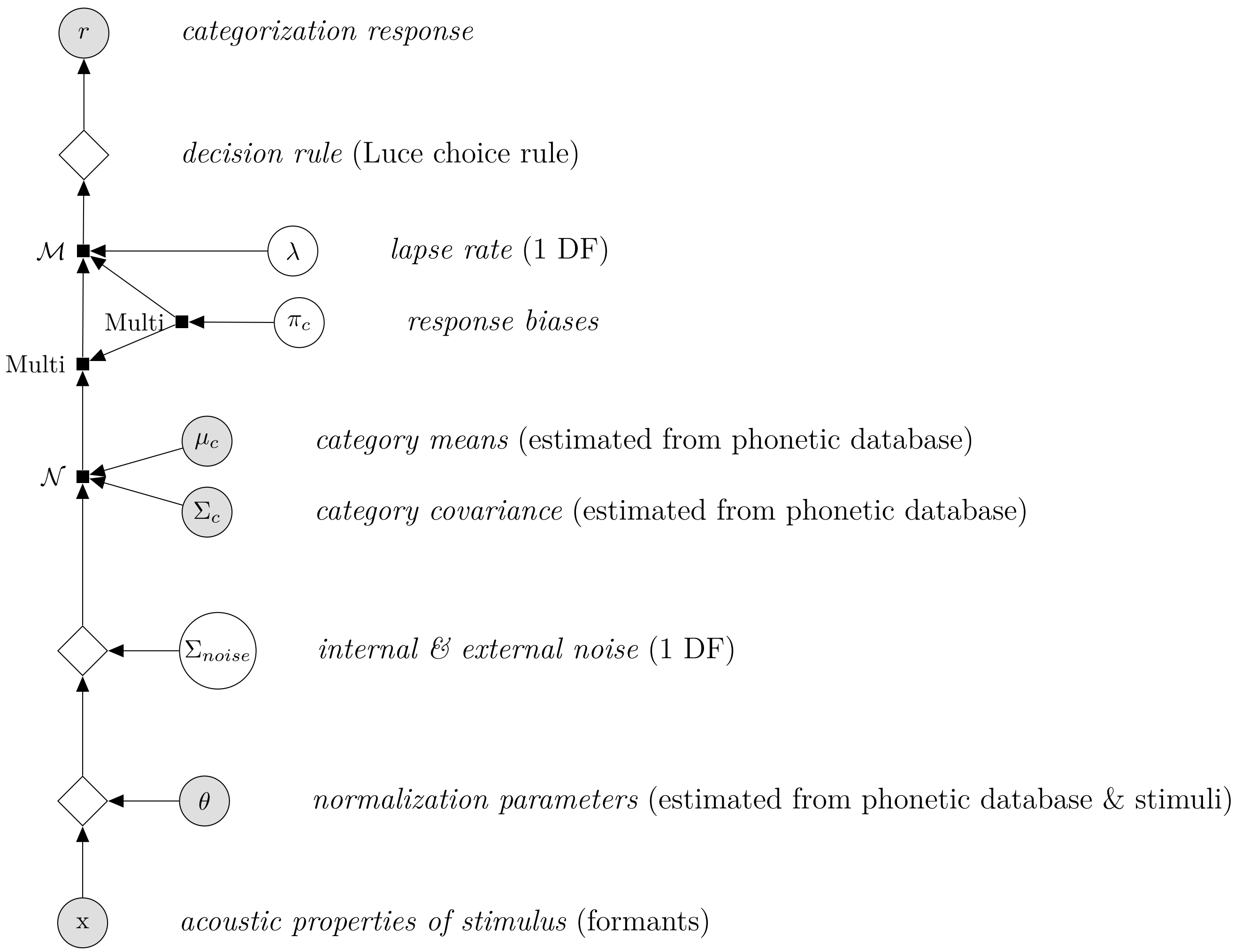


FIG. 6. Graphical model of ASP's general categorization framework (adapted for the current purpose from Xie et al., 2023, Figure 4). Here $J = 8$ (the eight vowel response options in Experiments 1a and 1b). We use this framework to compare normalization accounts against listeners' categorization responses from Experiments 1a and 1b. Filled gray circles represent variables that are known to the researcher. Empty circles represent latent variables that are not observable. Diamonds represent variable-free processes, annotated with the distributions resulting at that level of the model: $\mathcal{N}$(ormal), Multi(nomial), and $\mathcal{M}$(ixture) distributions.

freedom, we adopt all assumptions made in Xie *et al.* (2023), and do not introduce additional assumptions.

Starting at the bottom of the figure, the acoustic input x is normalized. Here, we follow most previous evaluations of normalization accounts, and focus on the point estimates of formants at the center of the vowel as the inputs to normalization. This leaves open the question of how considerations of additional cues to vowel identity (e.g., VISC) or formant dynamics might affect the findings we report below (a point to which we return in the general discussion). Specifically, the main analysis we present here focus on x = F1 and F2. As one anonymous reviewer pointed out, this focus on F1-F2 might underestimate the potential of *intrinsic* normalization accounts, which might perform better when more acoustic-phonetic features are considered. The SI, §3 E, thus reports additional analyses that instead employ F1-F3. These analyses indeed find that the fit of intrinsic normalization accounts improves more than that of extrinsic accounts when F3 is included in the analysis. However, the best-fitting accounts were still the same extrinsic accounts we find to best fit listeners' responses when only F1 and F2 is considered.

The specific computations applied to the input x depend on the normalization accounts (see Table 1). We use θ to refer to the parameters required by the normalization account. For example, for Nearey's uniform scaling account (Nearey, 1978), θ is the overall mean of all log-transformed formants. For Lobanov normalization (Lobanov, 1971), θ is a vector of means and standard deviations for each formant (in Hz). The normalized input is then perturbed by perceptual and environmental noise. Following Feldman *et al.* (2009), this noise is assumed to be Gaussian distributed centered around the transformed stimulus with noise variances that are independent and identical for all formants (i.e., Σ_{noise} is a diagonal matrix, and all diagonal entries have the same value).

Next, the likelihood of the normalized percept under each of the eight vowel categories is calculated, $p(F1, F2|vowel)$. This requires specifying listeners' expectations about the cue-to-category mapping (listeners' likelihood function). We followed Xie *et al.* (2023) and previous work and assume that each vowel maps onto a multivariate Gaussian distribution over the phonetic cues, here bivariate Gaussians over F1 and F2 (cf. Clayards *et al.*, 2008; Feldman *et al.*, 2009; Kleinschmidt and Jaeger, 2015; Norris and McQueen, 2008; Xie *et al.*, 2021). We also followed previous models in assuming a single dialect template—i.e., a single set of bivariate Gaussian vowel categories (Nearey and Assmann, 2007). The analyses of participants' responses we provided above in the description of Experiments 1a and 1b suggest that this assumption is wrong. However, more appropriate alternatives—such as hierarchical or mixture models with multiple dialect templates—will require substantial additional research as well as larger databases of vowel recordings that have high resolution both within and across dialects. We return to this issue in the general discussion.

Once the likelihood function for each vowel is specified, the posterior probability of each vowel is obtained by combining its likelihood with its prior probability or response bias π_c, according to Bayes theorem:[7]

$$p(vowel = c|F1, F2) = \frac{\mathcal{N}(F1, F2|\mu_c, \Sigma_c + \Sigma_{noise}) \times \pi_c}{\sum_{c_i} \mathcal{N}(F1, F2|\mu_{c_i}, \Sigma_{c_i} + \Sigma_{noise}) \times \pi_{c_i}} \tag{1}$$

Up to this point, the model is identical to a standard Bayesian ideal observer over noisy input (Feldman *et al.*, 2009; Kronrod *et al.*, 2016) for which the input has been transformed based on the normalization account. ASP's categorization model adds to this the potential that participants experience attentional lapses—or for other reasons do not respond based on the input—on some proportion of all trials (λ, as in standard psychometric lapsing models, Wichmann and Hill, 2001). On those trials, the posterior probability of a category is determined solely by participants' response bias, which we assume to be identical to the response bias on non-lapsing trials (following Xie *et al.*, 2023). This results in a posterior that is described by weighted mixture of two components, describing participants' posterior on non-lapsing and lapsing trials, respectively:

$$p(vowel = v|F1, F2) = (1-\lambda)\frac{\mathcal{N}(F1, F2|\mu_c, \Sigma_c + \Sigma_{noise}) \times \pi_c}{\sum_{c_i} \mathcal{N}(F1, F2|\mu_{c_i}, \Sigma_{c_i} + \Sigma_{noise}) \times \pi_{c_i}} + \lambda\frac{\pi_c}{\pi_{c_i}} \tag{2}$$

Finally, a decision rule is applied to the posterior to determine the response of the model, conditional on the input (one of the eight vowels in Experiments 1a and 1b). We followed the gross of research on speech perception and assume Luce's choice rule (Luce, 1959; for discussion, see Massaro and Friedman, 1990). Under this choice rule, the model can be seen as sampling from the posterior, responding with each category proportional to that category's posterior probability.

Next, we describe how we estimated the θs, μ_cs and Σ_cs for each normalization account from a phonetic database. We use this database as a—very coarse-grained—approximation of a the speech input a 'typical' listener might have experienced previously. By fixing θ, μ_c and Σ_c based on the distribution of phonetic cues in the database, we substantially reduce the DFs that are allowed to mediate the mapping from stimulus properties to listeners' responses (following Xie *et al.*, 2023). In addition, this approach naturally penalizes overly complex models by validating these against out-of-sample data. Finally, we describe how we fit the remaining parameters as DFs to participants' responses from Experiments 1a and 1b.

2. Modeling listeners' prior experience (and guarding against overfitting): $\boldsymbol{\theta}$, $\boldsymbol{\mu_c}$, and $\boldsymbol{\Sigma_c}$

By fixing θ, μ_c, and Σ_c based on a database of vowel *productions*, we impose strong constraints on the functional flexibility of the model in predicting listeners' responses. This benefit is made possible by committing to a strong linking hypothesis—that listeners' categories are learned from, and reflect, the distributional mapping from formants to vowels in previously experienced speech input (e.g., Abramson and Lisker, 1973; Massaro and Friedman, 1990; Nearey and Hogan, 1986). The database we use to approximate listeners' prior experience was originally developed to compare the production of L1 and L2 speakers (Xie and Jaeger, 2020). It contains 9-10 recordings of the eight *hVd* words from each of 17 (five female) L1 talkers of a Northeastern dialect of US English (ages 18 to 35 years old). Since Experiments 1a and 1b used recordings of one of these talkers, we excluded that talker prior to fitting training ideal observers on the data. In total, this yields 5842 recordings that are annotated for F0, F1-F3, and vowel duration. The SI (§3 A 1) summarizes the distribution of these cues, and how the different normalization accounts affect those distributions.

To avoid over-fitting the ASP model to the database, we used 5-fold cross-validation: we randomly split the Xie and Jaeger (2020) database into five approximately evenly-sized folds (following Persson and Jaeger, 2023). This split was performed within each vowel to guarantee that all five folds had the same relative amount of data for each vowel category.

These splits were combined into five training sets, each containing one of the folds (20% of the data). This way, each training set was different from the others, increasing the variability between sets.[8]

For each training set and for each normalization account, we then estimated the required normalization parameters θ for all talkers, and normalized all formants based on those talker-specific parameters. This yielded 5 (training sets) * 20 (accounts) = 100 normalized training sets. For each of these normalized training sets, we fit the category means, μ_c, and covariance matrices, Σ_c, of all eight vowels, using the `R` package `MVBeliefUpdatr` (Jaeger, 2024).[9]

This yielded 100 ideal observer models, five for each of the 20 normalization accounts in Table 1. Of note, the 20 ideal observers fit on each fold differ *only* in the assumptions they make about the normalization that is applied to cues before they are mapped onto the eight vowel categories. Figure 7 visualizes the resulting bivariate Gaussian categories for four of the 20 normalization accounts. This illustrates one advantage of the cross-validation approach: it takes a modest step towards simulating differences across listeners' prior experience (represented by the five different folds).

3. Transforming the stimuli from Experiments 1a and 1b into the normalized phonetic spaces

Next, we transformed the stimuli of Experiments 1a and 1b into the formant space defined by the 20 normalization accounts in Table 1. This requires estimating the required normalization parameters θ for each experiment and normalization account. We calculated

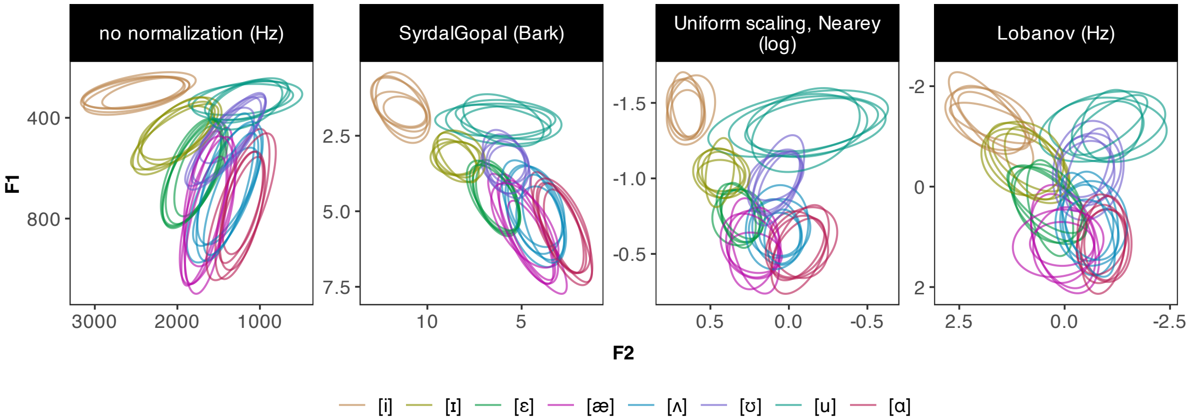


FIG. 7. Visualizing the bivariate Gaussian categories (prior to adding Σ_{noise}) of four example normalization accounts in F1-F2 space. Separate ellipses are shown for each of the five training sets (each set corresponds to one set of eight ellipses). The relative stability of the category ellipses across training sets indicates that the database is sufficiently large for the present purpose.

these θs over all stimuli (of each experiment and normalization account). For example, for Nearey's uniform scaling account (Nearey, 1978), we calculated the overall mean of all log-transformed formants over all stimuli. For Lobanov normalization (Lobanov, 1971), we calculated the mean and standard deviation of each formant (in Hz) over all stimuli. For each combination of experiment and normalization account, we then normalized the stimuli using those parameter estimates. The SI (§3 A 2) summarizes the θ parameters of all normalization accounts for each experiment and how they relate to the values obtained from the training sets. For reasons outlined in that same section, we did not expect a clear relation between an account's ability to predict listeners' responses for an experiment, and the degree to which the account's normalization parameters differed between the experiment and the training database (and, indeed, no such relation was found).

Combining the 100 normalized training sets described in the previous section with the matching normalized stimuli from each of the two experiments yielded 200 data sets.

4. *Noise (Σ_{noise}) and attentional lapses (λ)*

Finally, we describe the two parameters of the ASP model that we fit against listeners' responses in Experiments 1a and 1b. These two parameters constitute the only DFs that mediate the link from ideal observers' predictions to listeners' responses, and which are fit to listeners' responses. The first DF (Σ_{noise}) models the effects of internal (perceptual) and external (environmental) noise on listeners' perception. While previous work provides estimates of the internal noise in formant perception, these estimates were obtained under *assumptions* about the relevant formant space. For example, Feldman *et al.* (2009) estimated the internal noise variance to be about 15% of the average category variance along F1 and F2. This estimate was based on the assumption that human speech perception transforms vowel formants into Mel, without further normalization. Since we aim to *test* which normalization account best explains speech perception, we cannot rely on this or other internal noise estimates obtained under a single specific assumption. Additionally, internal noise can vary across individuals and external noise can vary across environments (a point particularly noteworthy, given that we conducted Experiments 1a and 1b over the web). We thus allowed the noise variance Σ_{noise} to vary in fitting participants' responses. Following Feldman *et al.* (2009), we assumed that perceptual noise had identical effects on all formants in the phonetic space defined by the normalization account (see also Kronrod *et al.*, 2016). This reduces Σ_{noise} to a single DF, regardless of the normalization account (for details, see SI §3 A 3).

The magnitude of Σ_{noise} affects the slope of the categorization functions that predict listeners' responses from stimulus properties (here, F1 and F2): higher Σ_{noise} imply more shallow categorization slopes. To facilitate comparison of Σ_{noise} values across normalization accounts, we report results in terms of the best-fitting *noise ratios* (τ^{-1}), rather than Σ_{noise}s. Specifically, Σ_{noise} is best understood *relative* to the inherent variability of the vowel categories (Σ_c). This variability in turn depends on the phonetic space defined by the normalization account. We thus divide Σ_{noise} by the mean of the diagonals of all Σ_cs to obtain the *noise ratio* τ^{-1}. For example, noise ratio of 0 corresponds to the absence of any noise, and a noise ratio of 1 corresponds to noise variance of the same magnitude as the average category variance along F1 and F2 in the phonetic space defined by the normalization account.[10] Figure 8B illustrates the effects of this noise ratio for Nearey's uniform scaling account.

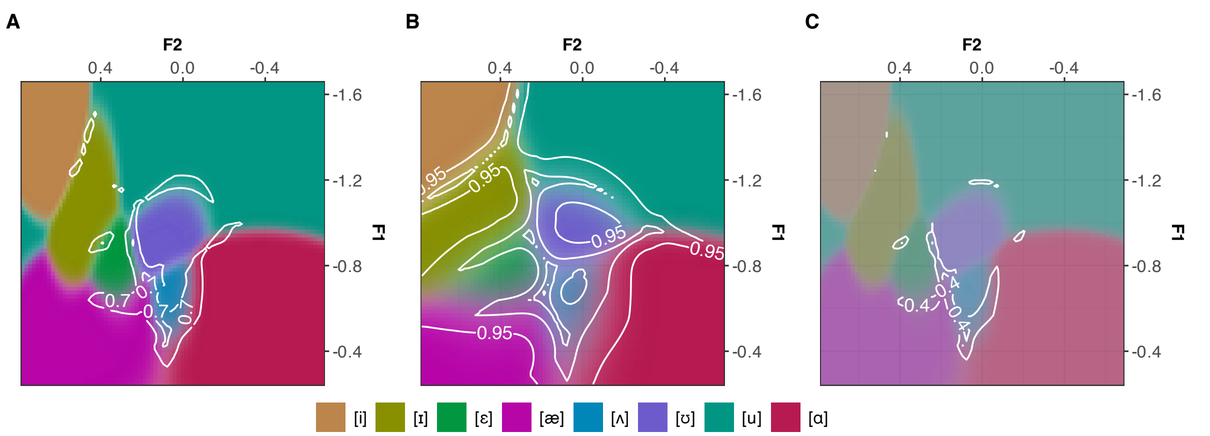


FIG. 8. Illustrating the consequences of perceptual and external noise (Σ_{noise}) and attentional lapse rates (λ) on the predicted posterior distribution of vowel categorizations. Shown are the average predicted posteriors across all five folds for Nearey's uniform scaling account. **Panel A**: Predicted posterior distribution for noise ratio $\tau^{-1} = \lambda = 0$. **Panel B**: Same for $\tau^{-1} = 1$ and $\lambda = 0$. **Panel C**: Same for $\tau^{-1} = 0$ and $\lambda = 0.5$. Transparency of a color is determined by that vowel's posterior probability. Contours indicate the highest posterior probability of any vowel (at .4, .5, .7, .95 probability level).

Second, participants can attentionally lapse or for other reasons reply without considering the speech input. We thus allowed lapse rates (λ) to vary while fitting human responses. This introduces a second DF, which we fit against listeners' responses. Together, the inclusion of freely varying lapse rates and a uniform response bias allows the ASP models to capture that some unknown proportion of listeners' responses might be more or less random, rather than reflecting properties of the vowel stimuli. This is illustrated in Figure 8C.

Finally, participants can have response biases that reflect their beliefs about the prior probability of each category. However, to reduce the DFs fit to participants' responses, we did *not* fit this response bias against listeners' responses (thus avoiding $J - 1 = 7$ additional DFs). Instead, we assumed uniform response biases—i.e., that listeners believed all eight response options in the experiments to be equally likely ($\forall c\ \pi_c = .125$). This decision implies that our models would not be able to capture any potential non-uniformity in listeners' response biases—including potential effects of additional acoustic differences (the absence of [h] in *odd* or the coda [t], rather than [d] in *hut*) and orthographically particular response options in Experiment 1a ("who'd", "odd", and "hut"). We do, however, see no reasons to expect this decision to bias the comparison of normalization accounts.

5. *Fitting normalization accounts to listeners' responses*

For each of the 200 combinations of experiment, normalization account, training set, we used constrained quasi-Newton optimization (Byrd *et al.*, 1995, as implemented in `R`'s `optim()` function) to find the λ and τ^{-1} values that best described listener's responses. Specifically, we used the 100 ideal observers described in the previous sections, applied them to the normalized stimuli of the experiment, and determined which λ and τ^{-1} maximized the likelihood of listener's responses (for details, see SI §3 A 3). This procedure yielded five maximum likelihood estimates for both λ and τ^{-1} for each combination of experiment and normalization account—one for each training set. All results presented below were validated and confirmed by grid searches over the parameter spaces (SI, §3 F).

We compare normalization accounts in terms of the likelihood of listeners' responses under these maximum likelihood estimates of λ and τ^{-1}. Comparing accounts in terms of their data likelihood follows more recent work (e.g., Barreda, 2021; McMurray and Jongman, 2011; Richter *et al.*, 2017; Xie *et al.*, 2023). Previous work has instead compared normalization accounts in terms of their accuracy (e.g., Johnson, 2020; Nearey and Assmann, 2007; Persson and Jaeger, 2023), or correlations with human response proportions (e.g., Hillenbrand and Nearey, 1999; Nearey and Assmann, 1986). Both of these approaches are problematic. Correlations between the predictions of a model and human responses can be high even when the model's predictions are systematically 'off'. Imagine three items for which listeners respond [ɪ] 10%, 30%, and 50% of the time. If a model predicts 30%, 50%, and 70% [ɪ] responses, respectively, for the same items, its predictions will perfectly correlate with listeners' response proportions, and yet be systematically wrong. Similarly, a model can achieve the highest possible accuracy in predicting listeners' responses simply because it always predicts the most frequent response (see discussion of criterion choice rule in Massaro and Friedman, 1990). In contrast, the likelihood of listeners' responses under a model is a direct measure of how well the model captures the distribution of listeners' responses conditional on the stimulus properties. In particular, data likelihood will be maximized if, and only if, the model-predicted posterior probabilities of each vowel for each stimulus are identical to the proportion with which those vowels occur in listeners' responses.

B. Results

We begin by comparing the fit of different accounts against listeners' responses in Experiments 1a and 1b. Given the comparatively large number of accounts compared here, we provide initial conclusions based on the best-fitting accounts along with the description of the results (more in-depth discussion is provided in the general discussion). Following this comparison, we visualize how different normalization accounts predict the formant space to be divided into the eight vowel categories.

1. Comparing normalization accounts in terms of fit against human behavior

Figure 9 compares how well the different normalization accounts fit listeners' responses in Experiments 1a and 1b. All accounts performed well above chance guessing (chance per-token log-likelihood in both experiments: $ln(\frac{1}{8})$=-2.08) but also well below the highest possible performance (in Experiment 1a, per-token log-likelihood = -0.46, in Experiment 1b: -1.15).

Normalization significantly improved the fit to listeners' responses relative to no normalization. This was confirmed by paired one-sided t-tests comparing the maximum likelihood values for each normalization account against those in the absence of normalization (all ps $<$ .05 except for Gerstman normalization, log-transformation and semitones-transformation and Experiment 1a; see SI §3 B 1). Not all normalization accounts achieved equally good fits,

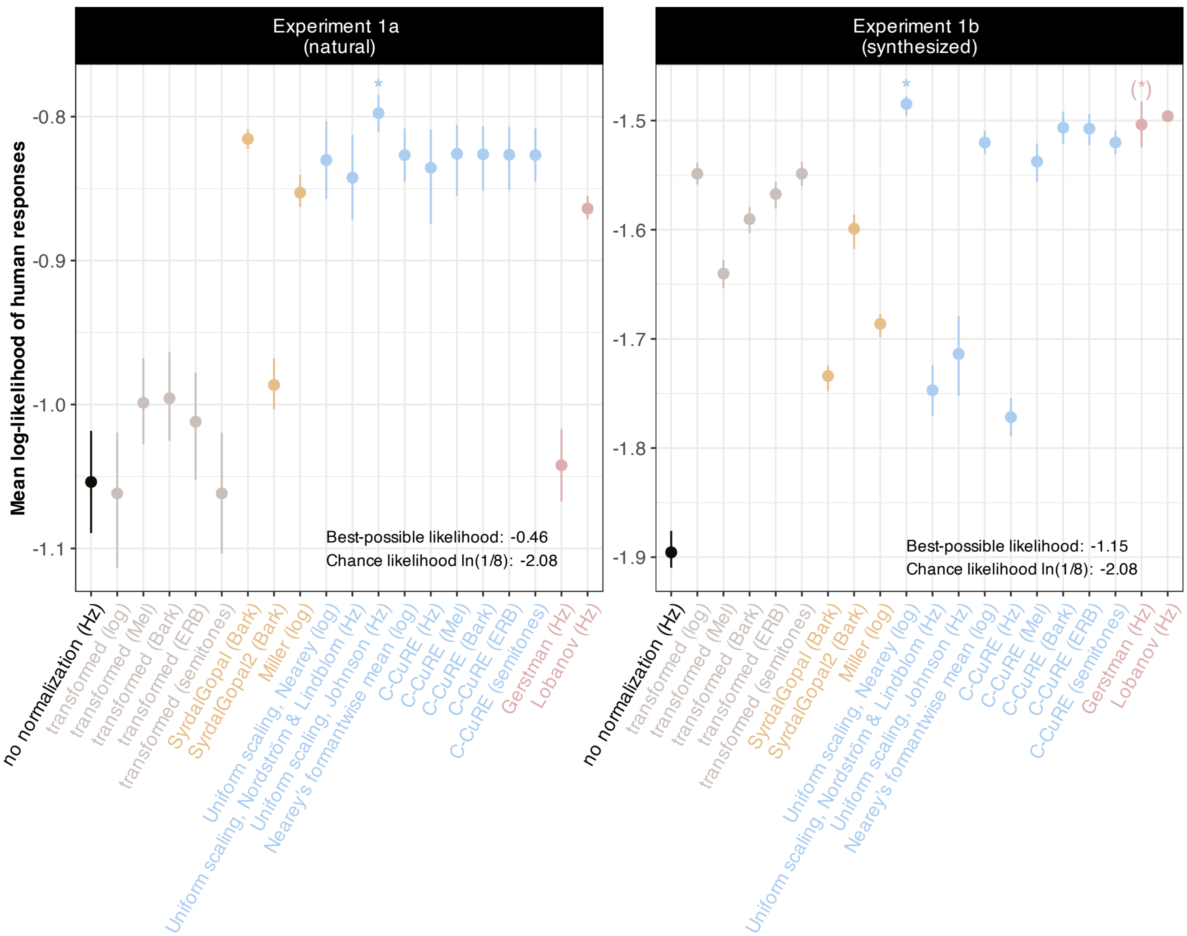


FIG. 9. Comparison of normalization accounts against listeners' responses. Point ranges indicate mean and 95% bootstrapped CIs of the per-token log-likelihoods summarized over the five training sets (higher is better), normalized by the number of listener responses in each experiment. Accounts that fit listeners' responses to an extent that is statistically indistinguishable from the best-fitting account are marked by (*). Note that per-token likelihoods cannot be directly compared across experiments because the best-possible likelihoods differ across experiments (due to differences in stimulus placement and other factors).

however: only some extrinsic accounts fit listeners' behavior well across both experiments. This supports two conclusions. First, it suggests that the normalization mechanisms operating during human speech perception involve computations that go beyond estimation-free transformations into psycho-acoustic spaces. Second, it suggests that the input to these computations is not limited to intrinsic information—i.e., that the computations draw on information beyond what is available in the acoustic signal *at that moment*. In particular, extrinsic normalization requires the estimation and memory maintenance of talker-specific properties from the speech signal.

While the accounts that achieved the best fit against listeners' responses differed between experiments, both were variants of uniform scaling. For Experiment 1a, Johnson normalization account provided the best fit (per-token log-likelihood = -0.8, SD = 0.02 across the five crossvalidation folds), while Nearey's uniform scaling account provided the best fit to Experiment 1b (per-token log-likelihood = -1.48, SD = 0.01). Both accounts essentially slide the representational 'template' of a dialect—here the eight bivariate Gaussian categories of an ideal observer—along a single line in the formant space. They differ only in *which* space this linear relation between formants is assumed. The same two accounts still fit listeners' responses best when F3 was included in the analysis in addition to F1 and F2 (SI, §3 E).[11] This suggests that formant normalization might involve comparatively parsimonious maintenance of talker-specific properties: in its simplest form, uniform scaling employs a single formant statistic to normalize all formants. In contrast, computationally more complex accounts like Lobanov normalization might require the estimation and maintenance of two formant statistics (mean and standard deviation) for each formant that is normalized (e.g., a total of four formant statistics for F1 and F2, or six statistics for F1-F3).

Also of note is that accounts that were particularly stable across experiments operate in log space, whereas accounts that operate in Hz space seemed to display a more volatile performance (e.g., both standardizing accounts but also C-CuRE Hz, Nordström & Lindblom and Johnson normalization). That accounts operating over log-transformed formants fit human behavior better should not be surprising. While questions remain about the exact organization of auditory formant representations, it is uncontroversial that the perceptual sensitivity to acoustic frequency information is better approximated by a logarithmic scale than by a linear scale (see Moore, 2012). As a result, a 30 Hz difference in an F1 of 300 Hz (a 10% change) is expected to be perceptually more salient than a 30 Hz change in an F2 of 2500 Hz (a 1.2% change).[12] In summary, variability in how well different accounts predict human behavior across the two experiments highlights the importance of psycho-acoustic transformations for human speech perception. This also highlights the importance of comparing normalization accounts against multiple types of data.

2. *Visualizing the consequences of different normalization mechanisms*

Before we turn to the general discussion, we briefly visualize how different normalization mechanisms affect vowel categorization. This sheds light on *why* the accounts differ in how well they fit listeners' responses. Figure 10 visualizes the categorization functions predicted by four different normalization accounts, using the best-fitting λ and τ^{-1} values for each account (i.e., the values that lead to the fit shown in Figure 9). Figure 10 highlights three points. First, a comparison across panels A-C shows different normalization accounts can result in very different predictions about how the acoustic space is carved into categories.

Second, the best-fitting parameters (shown at the top of each panel) were relatively comparable across accounts but differed more substantially across experiments. Specifically, the best-fitting estimates of lapse rates λ were generally comparable across the two exper-

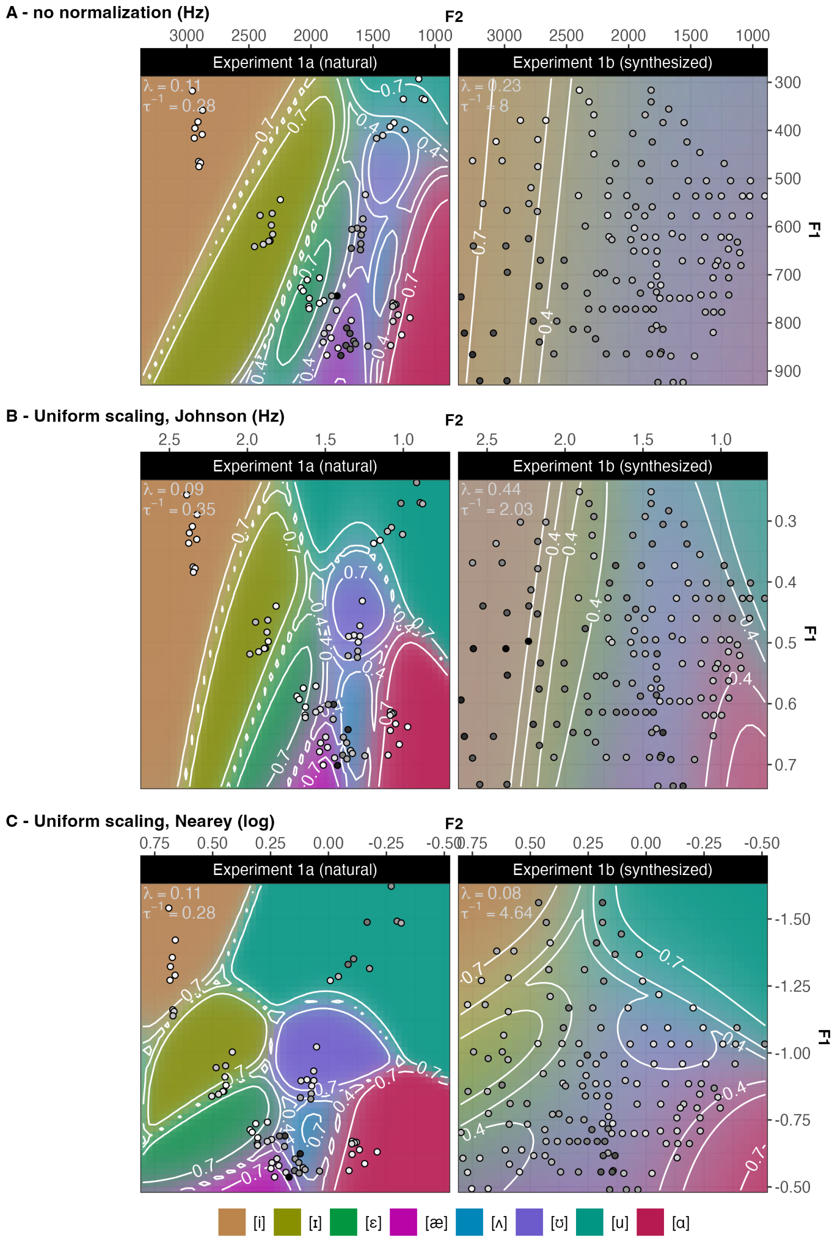


FIG. 10. Predicted categorization functions over the F1-F2 space under three different normalization accounts. For each account, we show the predicted posterior probabilities of all eight vowels obtained by averaging over the maximum likelihood parameterizations (of λ and τ^{-1}) for the five training sets (shown at top of each panel). **Panel A:** absence of normalization shown for reference. **Panel B:** the best-fitting account for Experiment 1a. **Panel C:** the best-fitting account for Experiment 1b. Contours indicate the highest posterior probability of any vowel. Points indicate location of test stimuli. The increasing brightness of points indicates a better match between the account's prediction and listeners' responses (higher log-likelihood; see text for detail).

iments (with the exception of Nordström & Lindblom and Johnson normalization, which exhibited substantially higher lapse rates in Experiment 1b; SI §3 B 2). This suggests that participants in both experiments were about equally likely to pay attention to the stimulus. The best-fitting noise ratios τ^{-1}, however, differed substantially across experiments, and were 10 times larger for Experiment 1b (mean $\tau^{-1} = 4.32$, SD = 2.52 across normalization accounts) than for Experiment 1a (mean $\tau^{-1} = 0.42$, SD = 0.46). This difference most likely reflects the fact that the synthesized stimuli in Experiment 1b left listeners with substantially more uncertainty about the intended category, as discussed during the description of the experiments.

Since noise is assumed to be independent of category variability (see also Feldman *et al.*, 2009; Kronrod *et al.*, 2016), differences in noise ratios can substantially change the categorization function. This is particularly evident for the accounts that had more variable performance across the two experiments. For example, Johnson normalization (Panel B) resulted in very different best-fitting categorization functions for Experiments 1a and 1b.

Third and finally, Figure 10 also shows how well accounts fit listeners' responses for each test stimulus (opaqueness of the points). This begins to explain *why* some accounts fit listeners' responses in Experiment 1b less well. For example, the Johnson normalization account (Panel B) predicts the responses to the test stimuli in Experiment 1a well, but fails to predict the responses to the test stimuli in Experiment 1b. This drop in performance seems to be primarily driven by stimuli that are unlikely to be articulated by the same talker (lower left, cf. dashed line in Figure 4). This might suggest that this account was over-engineered to explain naturally occurring productions—the type of data, it was originally tested on (Johnson, 2020). A plausible account of normalization, however, should be able to explain human perception to any type of stimulus, including synthesized stimuli. The SI (§3 B 3) presents more detailed by-item comparisons of normalization accounts that might be of interest to some readers.

IV. GENERAL DISCUSSION

Research on vowel normalization has an influential history. Cognitive scientists have long aimed to understand the organization of frequency information in the human brain (Siegel, 1965; Stevens and Volkmann, 1940), and how it helps listeners overcome cross-talker variability in the formant-to-vowel mapping (e.g., Fant, 1975; Joos, 1948; Nordström and Lindblom, 1975). Auditory processes that normalize speech inputs for differences in vocal tract physiology are now recognized to be an integral part of speech perception (Johnson and Sjerps, 2021; McMurray and Jongman, 2011; Xie *et al.*, 2023). Here, we set out to investigate what types of computations are implicated in the normalization of the frequency information that plays a critical role in the recognition of vowels.

Our results support three theoretical insights. First, human speech perception draws on more than psycho-acoustic transformations or intrinsic information, in line with previous research on normalization (Adank *et al.*, 2004; Ladefoged and Broadbent, 1957; Nearey, 1989). Rather, formant normalization seems to involve the estimation and storing of talker-specific formant properties. Second, computationally simple uniform scaling accounts provide the best fit to listeners' responses, suggesting comparatively parsimonious maintenance of talker-specific properties. This replicates and extends previous findings that uniform scaling or similarly simple corrections for vocal tract size provide a better explanation for human perception than more complex extrinsic accounts (Barreda, 2021; Richter *et al.*, 2017). It is impossible to rule out more complex approaches to perceptual normalization given the large number of possible alternatives. However, given that uniform scaling provides a parsimonious explanation for human formant normalization, and the current absence of empirical evidence for more complex computations, we submit that researchers ought to adapt uniform scaling as the working hypothesis. Third, the psycho-acoustic representation assumed by different normalization accounts matter, as indicated by the comparison of otherwise computationally similar accounts (e.g. Nearey's vs. Johnson's uniform scaling).

The results contribute to a still comparatively small body of work that has evaluated competing normalization accounts against listeners' perception, whereas most previous work evaluates accounts against intended productions. Complementing previous work, we took a broad-coverage approach: the present study compared 20 of the most influential normalization accounts against listeners' perception of *hVd* words with eight US English monophthongs in both natural and synthesized speech. This contrasts with previous work, which has typically focused on subsets of the vowel system, either using natural *or* synthesized speech, and considering a much smaller subset of accounts (typically 2-3 at a time). By considering a wider range of accounts, a wider range of formant values and vowel categories, and multiple types of speech, we aimed to contribute to a more comprehensive evaluation of competing accounts.

Next, we discuss the theoretical consequences of these findings for research beyond formant normalization. Following that, we discuss limitations of the present work, and how future research might overcome them.

A. Consequences for theories of speech perception and beyond

Understanding the perceptual space in which the human brain represents vowel categories—i.e., the normalized formant space—has obvious consequences for research on speech perception. To illustrate how far reaching these consequences can be, we discuss a few examples. For instance, research on *categorical perception* has found that vowels seem to be perceived less categorically than some types of consonants. Recent work has offered an elegant explanation for this finding: the perception of formants—relevant to the recognition of vowels—might be more noisy than the perception of the acoustic cues that are critical to the recognition of more categorically perceived consonants (Kronrod *et al.*, 2016). This is a parsimonious explanation, potentially preempting the need for separate explanations for the perception of different types of phonemic contrasts. Kronrod and colleagues based their argument on estimates they obtained for the relative ratio of meaningful category variability to perceptual noise (τ, the inverse of our noise ratios, $\tau^{-}1$). Critically, this ratio depends both on (i) the perceptual space in which formants are assumed to be represented (Kronrod at al. used Mel-transformed formant frequencies), and on (ii) whether the meaningful category variability is calculated prior to, or following, normalization (Kronrod et al. assumed the former, which increases estimates of category variability). Our point here is not to cast doubt on the results of Kronrod *et al.* (2016) —the fact that the best-fitting noise ratios in our study were relatively similar across accounts (while varying across experiments) suggests that the result of Kronrod and colleagues are likely to hold even under different assumptions about (i) and (ii)—but rather to highlight how research on the perception and recognition of vowels depends on assumptions about formant normalization. For example, similar points could be raised about experiments on statistical learning that manipulate formant or other frequency statistics (e.g., Chládková *et al.*, 2017; Colby *et al.*, 2018; Wade *et al.*, 2007; Xie *et al.*, 2021). Such experiments, too, need to make assumptions about the space in which formants are represented. If these assumptions are incorrect, this can affect whether the experimental manipulations have the intended effects, increasing the chance of null effects or misinterpretation of observed effects.

Understanding the perceptual space in which the human brain represents vowel categories also has consequences for research beyond speech perception, perhaps more so than is sometimes recognized. For instance, in sociolinguistics and related fields, Lobanov remains the norm for representing vowels due to its efficiency in removing cross-talker variability (for review, see Adank *et al.*, 2004; Barreda, 2021). However, as shown in the present study, removing cross-talker variability is not the same as representing vowels in the perceptual space that listeners actually employ. Here, we do *not* find Lobanov to describe human perception particularly well. On the contrary, we find no support for the hypothesis that human speech perception employs these more complex computations that have been found to perform best at reducing category variability. This should worry sociolinguists. In order to understand how listeners infer a talker's background or social identity, it is important to understand the perceptual space in which inferences are actually rooted. Critically, the representations resulting from formant normalization presumably form an important part of the information that listeners use to draw social and linguistic inferences. It should thus be obvious that the use of normalization accounts that do not actually correspond to human perception can both mask real markers of social identity, and 'hallucinate' markers that are not actually present. For example, in order to determine how a talker's social identity influences their vowel realizations, it is important to discount *all and only* effects that listeners will attribute to physiology, rather than social identity (Disner, 1980; Hindle, 1978).

Similar concerns apply to dialectology, research on language change, second language acquisition research, etc. For example, the perceptual space in which vowels are represented is critical to well-formed tests of hypotheses about the factors shaping the organization of vowel inventories across languages of the world (Lindblom, 1986; Stevens, 1972, 1989). It is essential in testing hypotheses about the extent to which the cross-linguistic realization of those systems is affected by perceptual processes (Flemming, 2010; Steriade, 2008), or by preferences for communicatively efficient linguistic systems (e.g., Hall *et al.*, 2018; Lindblom, 1990; Moulin-Frier *et al.*, 2015). Similarly, tests of the hypothesis that vowel *articulation* during natural interactions is shaped by communicative efficiency do in obvious ways depend on assumptions about the perceptual space in which talkers—by hypothesis—aim to reduce perceptual confusion (cf. Buz and Jaeger, 2016; Gahl *et al.*, 2012; Scarborough, 2010; Wedel *et al.*, 2018). The same applies to any other line of research that aims to understand the perceptual consequences of formant variation across talkers, including research on infant- or child-directed speech (Eaves Jr *et al.*, 2016; Kuhl *et al.*, 1997), and research on whether non-native talkers are inherently more variable than native talkers (Smith *et al.*, 2019; Vaughn *et al.*, 2019; Xie and Jaeger, 2020). In short, the perceptual space in which vowels are represented is a critical component of understanding the structure of vowel systems, the factors that shape them, and the ways in which they are used in natural language.

B. Limitations and future directions

As mentioned in the introduction, we take it as relatively uncontroversial *that* normalization is part of human speech perception. Independent of any benefits that such normalization conveys for speech perception, its existence is supported by evidence from cross-species comparisons and neuro-physiological studies (for review, see Barreda, 2020). There are, however, important questions as to how decisions we made in comparing normalization accounts against each other might have affected their fit against listeners' responses.

For instance, we followed previous work in focusing on formants, and specifically estimates of the formants in the *center* of the vowel. There is, however, ample evidence that formant dynamics throughout the vowel can strongly affect perception (Assmann and Katz, 2005, Hillenbrand and Nearey (1999); Nearey and Assmann, 1986). In addition, there are proposals that entirely give up the assumption that formants are the primary cues to vowel identity (e.g., whole-spectrum accounts, Hillenbrand *et al.*, 2006). While these proposals might provide a more informative representation of vowels, we consider it unlikely that they would entirely remove the problem of cross-talker variability. For instance, Richter *et al.* (2017) still found benefits of normalization even when the entire frequency spectrum throughout vowels was considered (in the form of Mel-Frequency Cepstral Coefficients and their derivatives). For the present work, auxiliary analyses in the SI (§3 E) replicated our core findings when

F3 was included in the model. Still, it remains unclear whether the inclusion of additional cues, such as VISC, or additional formant dynamics, would alter the results of the present study.

As is the case of any computational work, the present study committed to a number of assumptions that are not critical, but were necessary in order to deliver clear quantitative predictions. Quantitative tests of theories—as we have done here—require assumptions about *every* aspect of the model. Here, this included all the steps necessary to link properties of the stimuli to listeners' responses. For this purpose, we adopted the ASP framework (Xie *et al.*, 2023), and visualized the graphical model that links stimuli (x) to responses (r) in Figure 6.

Many of the assumptions we made should be relatively uncontroversial—e.g., the decision to include both external (environmental) and internal (perceptual) noise in our model. While these noise sources are often ignored in modeling human behavior, it is uncontroversial that they exist. Other assumptions we made were introduced as simplifying assumptions for the sake of feasibility—e.g., we expressed the effect of both types of noise through a single parameter that related the average within-category variability of formants to noise variability in the transformed and normalized formant space. In reality, however, environment noise can have effects that are independent of internal noise, and internal noise likely affects information processing at multiple (or all) of the steps shown in Figure 6. Such simplifying assumptions are both inevitable, and not necessarily problematic: as long as they do not introduce systematic bias to the evaluation of normalization accounts, they should not limit the generalizability of our results.

Some of our assumptions, however, might be more controversial. For example, we assumed that category representations can be expressed as multivariate Gaussian distributions in the formant space. This assumption, too, is a simplifying assumption—it simplified the computation of likelihoods—rather than a critical feature of the ASP framework we employed. While human category representations are unlikely to be Gaussians, the alternative, e.g., exemplar representations, would come with its own downsides, such as increased sensitivity to the limited size of phonetic databases and substantial increases in computation time (exemplar representations afford researchers with much larger degrees of freedom). For researchers curious how this and other assumptions we made affect our results, our data and code are shared on OSF.

Like previous work, we further assumed that all listeners in our experiments use the same underlying vowel representations—the same dialect template(s). However, as already discussed, it is rather likely that not all of our listeners employed the same dialect template(s). An additional analysis reported in the SI (§3 D) thus compared normalization accounts against only the subset of listeners who employed the dialect template used by the majority of participants (see lower-left of Figure 5B). This left only 20 participants for Experiment 1a (71.4%) and 23 for Experiment 1b (82.1%), substantially reducing statistical power. Replicating the main analysis, uniform scaling accounts again fit listeners' behavior well across both experiments. The best-performing account for Experiment 1a did, however, differ from the one obtained for the superset of data (the intrinsic Syrdal & Gopal achieved the best fit to listeners' responses in Experiment 1a for the shared dialect subset; see SI, §3 D).

A related assumption was introduced by the use of a phonetic database to approximate listeners' vowel representations. This deviates from most previous evaluations of normalization accounts (McMurray and Jongman, 2011; Barreda, 2021; but see Richter *et al.*, 2017), and reflects our commitment to a strong assumption made by most theories of speech perception: that listeners' representations reflect the formant statistics previously experienced speech input. By using a phonetic database to estimate listeners' representations, we *substantially* reduced the degrees of freedom in the evaluation of normalization accounts, reducing the chance of over-fitting to the data from our experiments. Our approach does, however, also introduce two new assumptions.

First, our approach assumes that the mixture of dialect template(s) used by talkers in the database sufficiently closely approximates those of the listeners in our experiments. Some validation for this assumption comes from the additional analysis reported in the preceding paragraph: when we subset listeners to only those who used the majority dialect template, this improved the fit of all normalization accounts—as expected, if the category representations we trained on the phonetic database primarily reflect those listeners' representations (see SI, §3 D). Future work could further address this assumption in a number of ways. On the one hand, dialect analyses like the ones we presented for our listeners (in Figure 5B) could compare listeners' templates against the templates used by talkers in the database. Alternatively or additionally, researchers could see whether our results replicate if ideal observers are instead trained on other databases that have been hypothesized to reflect a 'typical' L1 listeners' experience with US English. Finally, it might be possible in future work to use larger databases of vowel recordings to train separate ideal observers for all major dialects of US English, and to try to *estimate* for each listener which mixture of dialects their responses are based on.

Second, we made the simplifying assumption that listeners' category representation—or at least the representations listeners' drew on during the experiment—are talker-*independent* (we trained a single set of multivariate Gaussian categories, rather than, e.g., hierarchically organized set of multiple dialect templates). While this assumption is routinely made in research on normalization and beyond, it might well be wrong (see e.g., Xie *et al.*, 2021).

Finally, the evaluation of normalization accounts in the present study shares with all previous work (e.g., Apfelbaum and McMurray, 2015; Barreda, 2021; Cole *et al.*, 2010; McMurray and Jongman, 2011; Nearey, 1989; Richter *et al.*, 2017) another simplifying assumption that is clearly wrong: the assumption that listeners *know* the talker-specific formant properties required for normalization. Specifically, we normalized the input for each ideal observer using the maximum likelihood estimates of the normalization parameters over all stimuli for the respective experiment. For example, for the evaluation of the ideal observer trained on Lobanov normalized formants against listeners' responses in Experiment 1a, we used the formant means and standard deviations of the stimuli used in Experiment 1a to normalize F1 and F2. While this follows previous work, it constitutes a problematic assumption for the evaluation of extrinsic normalization accounts. For extrinsic accounts, the approach adopted here would seem to entail the ability to predict the future: even on the first trial of the experiment, the input to the ideal observers were formants that were normalized based on the normalization parameters estimated over the acoustic properties of *all* stimuli. Listeners instead need to *incrementally infer* talker-specific properties from the speech input (Barreda and Jaeger, submitted; Nearey and Assmann, 2007; Xie *et al.*, 2023). An important avenue for future research is thus the development and evaluation of incremental normalization accounts.

The present data only allow an initial, rather tentative, look at this question. For example, for Experiment 1a, for which each trial had a known correct answer (the vowel intended by the talker), we can assess whether participants' recognition accuracy improved across trials, as would be expected if listeners need to incrementally infer the talker-specific normalization parameters. Figure 11A suggests that this was indeed the case: the non-parametric listeners' average recognition accuracy improved over the course of the experiment from about 65% to 88%, with most of the improvements occurring during the first ten trials. To address potential confounds due to differences in the distribution of stimuli across trials, we used a generalized additive mixed-effect model to predict listeners' accuracy from log-transformed trial order while accounting for random by-participant and by-item intercepts and slopes for the log-transformed trial order (blue lines). Still, this result should be interpreted with caution, as Experiment 1a was not designed to reliably address questions about incremental changes across the experiment.

Figure 11B shows how the fit of the best-fitting normalization model changes across trials. We used a generalized additive mixed-effect model to predict the log-likelihood of listeners' responses from log-transformed trial order while accounting for random by-participant and by-item intercepts and slopes for the log-transformed trial order (blue lines). Given that our evaluation of normalization accounts assumed that the normalization parameters were already known on the first trial of the experiment, we would expect that the likelihood of listeners' responses under a normalization model would improve the more input listeners have received (i.e., as the simplifying assumptions of our evaluation become increasingly more plausible). For Experiment 1a, this indeed appears to be the case. However, no clear evidence for such incremental improvements in the fit of the normalization model is observed for Experiment 1b. In short, the present data does not support decisive conclusions about the extent to which normalization proceeds incrementally.

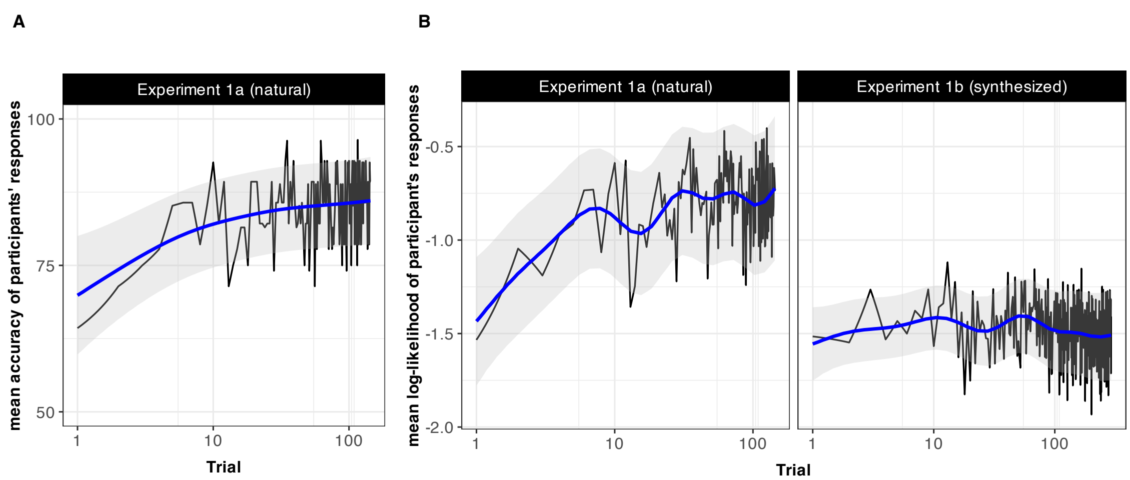


FIG. 11. **Panel A**: Changes across trials in listeners' average accuracy in recognizing the vowel intended by the talker in Experiment 1a, averaged across items and participants (black line). Blue line shows a generalized additive mixed-effects model predicting accuracy from log-transformed trial order, with 95% CIs. **Panel B**: Log-likelihood of listeners' responses under the best-fitting normalization account at each trial, averaged across items and participants (Johnson's uniform scaling for Experiment 1a and Nearey's uniform scaling for Experiment 1b). Blue lines show generalized additive mixed-effects models predicting log-likelihood from log-transformed trial order, with 95% CIs.

C. Concluding remarks

We set out to compare how well competing accounts of formant normalization explain listeners' perception of vowels. We developed a computational framework that makes it possible to compare a large number of different accounts against multiple data sets. The code we share on OSF makes it possible to 'plug in' different accounts of vowel normalization, different phonetic databases, and different perception experiments. This, we hope, will substantially reduce the effort necessary to conduct similar evaluations on other datasets, dialects, and languages.

Comparing 20 of the most influential normalization accounts against L1 listeners' perception of US English monophthongs, we found that the normalization accounts that best describe listeners' perception share that they (1) learn and store talker-specific properties and (2) seem to be computationally very simple—taking advantage of the physics of sound generation to use as few as a single parameter to normalize inter-talker variability in vocal tract size. While the number of studies that have compared normalization accounts against *listeners'* behavior remains surprisingly small, these two results confirm the findings from more targeted comparisons that were focused on 2-3 accounts at a time (Barreda, 2021; Nearey, 1989; Richter *et al.*, 2017). Overall then, we submit that it is time for research in speech perception and beyond to consider simple uniform scaling the most-likely candidate for human formant normalization.

ACKNOWLEDGMENTS

Earlier versions of this work were presented at 2023 ASA meeting, ExLing 2022, at the Department of Computational Linguistics at the University of Zürich and at the Department of Swedish language and multilingualism at Stockholm University. We are grateful to Maryann Tan, Chigusa Kurumada, and Xin Xie for feedback on this work. We thank Travis Wade for clarifications on the synthesis procedure used in his study. We thank Leslie Li and Xin Xie for sharing their database of L1-US English *hVd* productions, and the JASA copy editing staff for help with the Latex formatting. This work was partially funded by grants to AP from Kungliga Vetenskapsakademien, Kungliga Vitterhetsakademien, and the Department of Swedish Language and Multilingualism at Stockholm University, as well as grants to TFJ by the Helge Ax:son Johnson foundation, the Stockholm University Board of Human Science (Funding for Strategic Investments), and the Stockholm University Faculty of Humanities' Research School (Kvalitetssäkrande medel grant).

AUTHOR CONTRIBUTIONS

AP designed the experiments and collected the data, with input from TFJ. TFJ programmed the experiments with input from AP. AP analyzed the experiments, with input from TFJ. AP and TFJ wrote the code to implement and fit the normalization models, with input from SB. AP developed the visualizations within input from SB and TFJ. AP wrote the first draft of the manuscript with edits by SB and TFJ.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics approval

This study was reviewed and approved Research Subjects Review Board (RSRB) of the University of Rochester (STUDY00000417) under the OHSP and UR policies, and in accordance with Federal regulation 45 CFR 46 under the university's Federal-wide Assurance (FWA00009386).

V. REFERENCES

[1]Some hypotheses hold that robust speech perception does not require normalization, and that research on normalization has over-estimated its effectiveness because studies tend to consider only a fraction of the phonetic information available to listeners (for review, see Strange and Jenkins, 2012). For vowel recognition, for example, listeners might use cues other than just formants (Hillenbrand *et al.*, 2006; Nearey and Assmann, 1986), and/or might use information about the dynamic development of formant trajectories over the entire vowel rather than just point estimates of formants at the vowel center (e.g., Shankweiler *et al.*, 1978). We return to this in the general discussion but note that even studies who use much richer inputs have found that normalization provides a better fit to listeners' perception (Richter *et al.*, 2017).

[2]Under uniform scaling accounts, listeners essentially 'slide' the center of their category representations (e.g, the 'template' of vowel categories for a given dialect) along a single line in formant space, with Ψ determining the target of this sliding. Later extensions of this account maintain its memory parsimony but increased its inference complexity by allowing both intrinsic (the current F0) and extrinsic information (the talker's single mean of log-transformed formants) to influence the inference of Ψ (Nearey and Assmann, 2007).

[3]We use Johnson's (2020) implementation of Nordström and Lindblom (1975). We group both Nordström and Lindblom (1975) and Johnson (2020) with the centering accounts, as they are essentially variants of uniform scaling, differing in their estimation of Ψ. We also include both versions of Syrdal & Gopal's Bark-distance model. The two versions differ only in their normalization of F2, and have not previously been compared against human perception.

[4]Shannon (1948) response entropy is defined as $H(x) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$. The maximum possible response entropy for an eight-way response choice is 3 bits, which means that all eight vowels are responded

equally often. The minimum response entropy = 0 bits, which means that the same vowel is responded all the time.

[5]Note that participants in Experiment 1a exhibited high agreement on [ʌ], [æ], and [ɑ], despite the close proximity between, and partial overlap of, these vowels in F1-F2 space. To understand this pattern, it is important to keep in mind that the recordings for [ʌ] and [ɑ] differed from the recordings for other stimuli in their word onset ("odd" for [ɑ]) or offset ("hut" for [ʌ]).

[6][u] has been undergoing changes in many varieties of US English. Whereas the talker in Experiment 1a produces [u] with low F1 and F2 (high and back), other L1 talkers of US English produce this vowel considerably more forward (higher F2).

[7]For Gaussian noise and Gaussian category likelihoods, the resulting noise-convolved likelihood is a Gaussian with variance equal to the sum of the noise and category variances (Kronrod *et al.*, 2016).

[8]We intentionally did *not* split the data within talkers since normalization accounts are meant to make speech perception robust to cross-talker variability. Further, splitting the data by speaker rather than by vowel category avoids the potential for biases in the normalization parameter estimates for different speakers in the case of missing or unbalanced tokens across vowel categories, see (Barreda and Nearey, 2018). Additional analyses not reported here confirmed that the same results are obtained when splits are performed within talkers and within vowels (except that this lead to smaller CIs, and thus *more* significant differences, in Figure 9). These analyses can be replicated by downloading the R markdown document this article is based on from our OSF (see comments in our code).

[9]Alternatively, it would be possible to treat these parameters as DFs in the link to listeners' responses, and infer them from the responses in Experiments 1a and 1b (cf., Kleinschmidt and Jaeger, 2016). This approach would afford the model with a high degree of functional flexibility, regardless of which normalization approach is applied (similar to previous approaches that have employed, e.g., multinomial logistic regression).

[10]This ratio is a generalization of the inverse of the "meaningful-to-noise variance ratio (τ)" used in Kronrod *et al.* (2016). However, whereas Kronrod and colleagues committed to the simplifying assumption that all categories have identical variance (along all formants), we allowed category variances to differ between vowels, and between F1 and F2 (matching the empirically facts). We merely assume that the *noise* variance is identical across all formants (in the phonetic space defined by the normalization account, e.g., log-Hz for uniform scaling and Hz for Lobanov).

[11]Additional analyses reported in the SI (§3 C) overall replicated this result for subsets of Experiments 1a and 1b, with Nearey's uniform scaling achieving the best fit to listeners' responses in both experiments. For Experiment 1a, we excluded responses to the two *hVd* stimuli that differed from the other stimuli in the preceding (*odd*) or following phonological context (*hut*). For Experiment 1b, we excluded responses to any stimuli that were physiologically implausible for the talker (stimuli below the diagonal dashed line in Figure 4). As requested by a reviewer, the SI §3 B 4 also reports the accuracy of predicting listeners' responses for all normalization accounts. The best performing accounts achieved 61.8% for Experiment 1a (Johnson normalization), and 29.2% for Experiment 1b (Nearey's uniform scaling), compared to 52.3% and 16.9%, respectively, without normalization.

[12]In line with this reasoning, additional tests found that Johnson normalization would provide the best fit to Experiment 1b if it was applied to log-transformed formants (instead of Hertz).

Abramson, A. S., and Lisker, L. (**1973**). "Voice-timing perception in Spanish word-initial stops," Journal of Phonetics **1**(1), 01–08, doi: 10.1016/S0095-4470(19)31372-5.

Adank, P., Smits, R., and van Hout, R. (**2004**). "A comparison of vowel normalization procedures for language variation research," The Journal of the Acoustical Society of America **116**(5), 3099–3107, doi: 10.1121/1.1795335.

Allen, J. S., Miller, J. L., and DeSteno, D. (**2003**). “Individual talker differences in voice-onset-time,” Journal of the Acoustical Society of America **113**(1), 544–552, doi: 10.1121/1.1528172.

Apfelbaum, K., and McMurray, B. (**2015**). “Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization,” Psychonomic Bulletin and Review **22**(4), 916–943, doi: 10.3758/s13423-014-0783-2.

Assmann, P. F., and Katz, W. F. (**2005**). “Synthesis fidelity and time-varying spectral change in vowels,” The Journal of the Acoustical Society of America **117**(2), 886–895, doi: 10.1121/1.1852549.

Assmann, P. F., Nearey, T. M., and Bharadwaj, S. (**2008**). “Analysis of a vowel database,” Canadian Acoustics **36**(3), 148–149.

Baese-Berk, M. M., Walker, K., and Bradlow, A. (**2018**). “Variability in speaking rate of native and non-native speakers,” The Journal of the Acoustical Society of America **144**(3), 1717–1717, doi: 10.1121/1.5067612.

Balzano, G. J. (**1982**). “The pitch set as a level of description for studying musical pitch perception,” in *Music, mind, and brain: The neuropsychology of music*, edited by M. Clynes (Springer), pp. 321–351.

Barreda, S. (**2020**). “Vowel normalization as perceptual constancy,” Language **96**(2), 224–254, doi: 10.1353/lan.2020.0018.

Barreda, S. (**2021**). “Perceptual validation of vowel normalization methods for variationist research,” Language Variation and Change **33**(1), 27–53, doi: 10.1017/S0954394521000016.

Barreda, S., and Jaeger, T. F. (**submitted**). “Re-introducing the probabilistic sliding template model of vowel perception,” Linguistic Vanguard .

Barreda, S., and Nearey, T. M. (**2012**). “The direct and indirect roles of fundamental frequency in vowel perception,” The Journal of the Acoustical Society of America **131**(1), 466–477, doi: 10.1121/1.3662068.

Barreda, S., and Nearey, T. M. (**2018**). “A regression approach to vowel normalization for missing and unbalanced data,” The Journal of the Acoustical Society of America **144**(1), 500–520, doi: 10.1121/1.5047742.

Bladon, A., Henton, C., and Pickering, J. (**1984**). “Towards an auditory theory of speaker normalization,” Language and Communication **4**, 59–69.

Boersma, P., and Weenink, D. (**2022**). “Praat: Doing phonetics by computer [Computer program]” .

Buz, E., and Jaeger, T. F. (**2016**). “The (in) dependence of articulation and lexical planning during isolated word production,” Language, Cognition and Neuroscience **31**(3), 404–424.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (**1995**). “A limited memory algorithm for bound constrained optimization,” SIAM Journal on Scientific Computing **16**(5), 1190–1208, doi: 10.1137/0916069.

Carpenter, G. A., and Govindarajan, K. K. (**1993**). “Neural Network and Nearest Neighbor Comparison of Speaker Normalization Methods for Vowel Recognition,” in *ICANN ’93. Proceedings of the International Conference on Artificial Neural Networks, Amsterdam, the Netherlands, 13-16 September*, edited by S. Gielen and B. Kappen (Springer London, London), pp. 412–415, doi: 10.1007/978-1-4471-2063-6_98.

Chládková, K., Podlipský, V. J., and Chionidou, A. (**2017**). "Perceptual adaptation of vowels generalizes across the phonology and does not require local context.," Journal of Experimental Psychology: Human Perception and Performance **43**(2), 414.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (**2008**). "Perception of speech reflects optimal use of probabilistic speech cues," Cognition **108**(3), 804–809, doi: 10.1016/j.cognition.2008.04.004.

Colby, S., Clayards, M., and Baum, S. (**2018**). "The role of lexical status and individual differences for perceptual learning in younger and older adults," Journal of Speech, Language, and Hearing Research **61**(8), 1855–1874, doi: 10.1044/2018_JSLHR-S-17-0392.

Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (**2010**). "Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach," Journal of Phonetics **38**(2), 167–184, doi: 10.1016/j.wocn.2009.08.004.

Crinnion, A. M., Malmskog, B., and Toscano, J. C. (**2020**). "A graph-theoretic approach to identifying acoustic cues for speech sound categorization," Psychonomic Bulletin & Review **27**(6), 1104–1125, doi: 10.3758/s13423-020-01748-1.

Disner, S. F. (**1980**). "Evaluation of vowel normalization procedures," The Journal of the Acoustical Society of America **67**(1), 253–261, doi: 10.1121/1.383734.

Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., and Shafto, P. (**2016**). "Infant-directed speech is consistent with teaching.," Psychological Review **123**(6), 758.

Escudero, P., and Bion, R. A. H. (**2007**). "Modeling vowel normalization and sound perception as sequential processes," **XVI**, pp. 1413–1416.

Fant, G. (**1975**). "Non-uniform vowel normalization," STL-QPSR **16**(2–3), 001–019.

Fant, G., Kruckenberg, A., Gustafson, K., and Liljencrants, J. (**2002**). "A New Approach to Intonation Analysis and Synthesis of Swedish," Proceedings of Fonetik, TMH-QPSR **44**(1), 161–164.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (**2009**). "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference," Psychological Review **116**(4), 752–782, doi: 10.1037/a0017196.

Flemming, E. (**2010**). "Modeling listeners: Comments on pluymaekers et al. and scarborough," in *Laboratory Phonology*, edited by C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, **10**, pp. 587–606.

Gahl, S., Yao, Y., and Johnson, K. (**2012**). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," Journal of Memory and Language **66**(4), 789–806, doi: 10.1016/j.jml.2011.11.006.

Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Transactions on Audio and Electroacoustics **16**(1), 78–80, doi: 10.1109/TAU.1968.1161953.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hearing Research **47**(1), 103–138, doi: 10.1016/0378-5955(90)90170-T.

Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory.," Journal of Experimental Psychology: Learning Memory and Cognition **22**(5), 1166–1183, doi: 10.1037/0278-7393.22.5.1166.

Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (**2018**). "The role of predictability in shaping phonological patterns," Linguistics Vanguard **4**(s2), 20170027, doi: 10.1515/

lingvan-2017-0027.

Hay, J., Podlubny, R., Drager, K., and McAuliffe, M. (**2017**). "Car-talk: Location-specific speech production and perception," Journal of Phonetics **65**, 94–109, doi: 10.1016/j.wocn.2017.06.005.

Hay, J., Walker, A., Sanchez, K., and Thompson, K. (**2019**). "Abstract social categories facilitate access to socially skewed words," PLoS ONE **14**(2), 1–29, doi: 10.1371/journal.pone.0210793.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," Journal of the Acoustical Society of America **97**(5), 3099–3111, doi: 10.1121/1.411872.

Hillenbrand, J. M., Houde, R. A., and Gayvert, R. T. (**2006**). "Speech perception based on spectral peaks versus spectral shape," The Journal of the Acoustical Society of America **119**(6), 4041–4054, doi: 10.1121/1.2188369.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hvd/ utterances: Effects of formant contour," Journal of the Acoustical Society of America **105**(6), 3509–3523, doi: 10.1121/1.424676.

Hindle, D. (**1978**). "Approaches to Vowel Normalization in the Study of Natural Speech," in *Linguistic Variation: Models and Methods*, edited by D. Sankoff (Academic Press, New York), pp. 161–171.

Jaeger, T. F. (**2024**). *MVBeliefUpdatr: Fitting, Summarizing, and Visualizing of Multivariate Gaussian Ideal Observers and Adaptors*, https://github.com/hlplab/MVBeliefUpdatr, r package version 0.0.1.0010.

Johnson, K. (**1997**). “Speech perception without speaker normalization,” in *Talker Variability in Speech Processing*, edited by K. Johnson and W. Mullennix (CA: Academic Press, San Diego), pp. 146–165.

Johnson, K. (**2020**). “The ΔF method of vocal tract length normalization for vowels,” Laboratory Phonology **11**(1), doi: 10.5334/labphon.196.

Johnson, K., and Sjerps, M. J. (**2021**). “Speaker normalization in speech perception,” in *The Handbook of Speech Perception*, edited by J. S. Pardo, L. C. Nygaard, R. E. Remez, and D. B. Pisoni (John Wiley & Sons, Inc), pp. 145–176, doi: 10.1002/9781119184096.ch6.

Johnson, K., Strand, E. A., and D’Imperio, M. (**1999**). “Auditory–visual integration of talker gender in vowel perception,” Journal of Phonetics **27**(4), 359–384, doi: 10.1006/jpho.1999.0100.

Joos, M. (**1948**). “Acoustic Phonetics,” Language **24**(2), 5–136, doi: 10.2307/522229.

Kleinschmidt, D. (**2020**). “What constrains distributional learning in adults?,” .

Kleinschmidt, D., and Jaeger, T. F. (**2015**). “Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel,” Psychological Review **122**(2), 148–203, doi: 10.1037/a0038695.

Kleinschmidt, D., and Jaeger, T. F. (**2016**). “What do you expect from an unfamiliar talker?,” Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016 2351–2356.

Kleinschmidt, D., Liu, L., Bushong, W., Burchill, Z., Xie, X., Tan, M., Karboga, G., and Jaeger, F. (**2021**). “JSEXP” https://github.com/hlplab/JSEXP.

Kronrod, Y., Coppess, E., and Feldman, N. H. (**2016**). “A unified model of categorical effects in consonant and vowel perception,” Psychological Bulletin and Review 1681–1712, doi: 10.3758/s13423-016-1049-y.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). “Cross-language analysis of phonetic units in language addressed to infants,” Science **277**(5326), 684–686, doi: 10.1126/science.277.5326.684.

Labov, W., Ash, S., and Boberg, C. (**2006**). *The atlas of North American English: Phonetics, phonology, and sound change* (De Gruyter Mouton, Berlin; New York).

Ladefoged, P., and Broadbent, D. E. (**1957**). “Information conveyed by vowels,” Journal of the Acoustical Society of America **29**, 98–104, doi: 10.1121/1.1908694.

Lee, C.-Y. (**2009**). “Identifying isolated, multispeaker mandarin tones from brief acoustic input: A perceptual and acoustic study,” The Journal of the Acoustical Society of America **125**(2), 0001–4966, doi: 10.1121/1.3050322.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). “Perception of the speech code,” Psychological review **74**(6), 431–461, doi: 10.1037/h0020279.

Lindblom, B. (**1986**). “Phonetic universals in vowel systems,” in *Experimental Phonology*, edited by J. J. Ohala and J. J. Jaeger (Academic Press, Orlando), pp. 13–44.

Lindblom, B. (**1990**). “Explaining phonetic variation: A sketch of the H&H theory,” in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Dordrecht: Kluwer), pp. 403–439.

Lobanov, B. M. (**1971**). “Classification of Russian vowels spoken by different speakers,” The Journal of the Acoustical Society of America **49**(2B), 606–608, doi: 10.1121/1.1912396.

Luce, P. A., and Pisoni, D. B. (**1998**). “Recognizing spoken words: The neighborhood activation model,” Ear and Hearing **19**(1), 1–36, doi: 10.1097/00003446-199802000-00001.

Luce, R. D. (**1959**). *Individual Choice Behavior* (John Wiley, Oxford).

Magnuson, J. S., and Nusbaum, H. C. (**2007**). “Acoustic differences, listener expectations, and the perceptual accommodation of talker variability,” Journal of Experimental Psychology: Human Perception and Performance **33**(2), 391–409, doi: 10.1037/0096-1523.33.2.391.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R., Monto, N., and Rueckl, J. G. (**2020**). “EARSHOT: A minimal neural network model of incremental human speech recognition,” Cognitive Science **44**(4), 1–17, doi: 10.1111/cogs.12823.

Massaro, D. W., and Friedman, D. (**1990**). “Models of integration given multiple sources of information.,” Psychological Review **97**(2), 225–252, doi: 10.1037/0033-295X.97.2.225.

McClelland, J. L., and Elman, J. L. (**1986**). “The TRACE model of speech perception,” Cognitive Psychology **18**(1), 1–86, doi: 10.1016/0010-0285(86)90015-0.

McGowan, K. B. (**2015**). “Social expectation improves speech perception in noise,” Language and Speech **58**(4), 502–521, doi: 10.1177/0023830914565191.

McMurray, B., and Jongman, A. (**2011**). “What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations,” Psychological Review **118**(2), 219–246, doi: 10.1037/a0022325.What.

Merzenich, M. M., Knight, P. L., and Roth, G. L. (**1975**). “Representation of cochlea within primary auditory cortex in the cat,” Journal of Neurophysiology **38**(2), 231–249, doi: 10.1152/jn.1975.38.2.231.

Miller, J. D. (**1989**). “Auditory-perceptual interpretation of the vowel,” The Journal of Acoustical Society of America **85**(5), 2114–2134, doi: 10.1121/1.397862.

Moore, B. C. (**2012**). *An Introduction to the Psychology of Hearing* (Brill, Bingley).

Moulin-Frier, C., Diard, J., Schwartz, J.-L., and Bessière, P. (**2015**). “Cosmo (“communicating about objects using sensory–motor operations”): A bayesian modeling framework for studying speech communication and the emergence of phonological systems,” Journal of Phonetics **53**, 5–41, doi: 10.1016/j.wocn.2015.06.001.

Nearey, T. M. (**1978**). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Indiana).

Nearey, T. M. (**1989**). “Static, dynamic, and relational properties in vowel perception,” The Journal of the Acoustical Society of America **85**(5), 2088–2113, doi: 10.1121/1.397861.

Nearey, T. M. (**1990**). “The segment as a unit of speech perception,” Journal of Phonetics **18**(3), 347–373, doi: 10.1016/S0095-4470(19)30379-1.

Nearey, T. M., and Assmann, P. F. (**1986**). “Modeling the role of inherent spectral change in vowel identification,” The Journal of the Acoustical Society of America **80**(5), 1297–1308, doi: 10.1121/1.394433.

Nearey, T. M., and Assmann, P. F. (**2007**). “Probabilistic ’sliding template’ models for indirect vowel normalization,” in *Experimental approaches to phonology*, edited by J.-J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press), pp. 246–270.

Nearey, T. M., and Hogan, J. (**1986**). “Phonological contrast in experimental phonetics: Relating distributions of measurements production data to perceptual categorization curves,” in *Experimental Phonology*, edited by J. J. Ohala and J. Jaeger (Academic Press, New York), pp. 141–161.

Newman, R. S., Clouse, S. A., and Burnham, J. L. (**2001**). “The perceptual consequences of within-talker variability in fricative production,” The Journal of the Acoustical Society of America **109**(3), 1181–1196, doi: 10.1121/1.1348009.

Nordström, P., and Lindblom, B. (**1975**). “A normalization procedure for vowel formant data,” Proceedings of the 8th international congress of phonetic sciences, Leeds 212.

Norris, D., and McQueen, J. M. (**2008**). “Shortlist B: A Bayesian model of continuous speech recognition.,” Psychological review **115**(2), 357–95, doi: 10.1037/0033-295X.115.2.357.

Oganian, Y., Bhaya-Grossman, I., Johnson, K., and Chang, E. F. (**2023**). “Vowel and formant representation in the human auditory speech cortex,” Neuron **111**(13), 2105–2118.

Patterson, R. D., and Irino, T. (**2014**). “Size matters in hearing: How the auditory system normalizes the sounds of speech and music for source size,” in *Perspectives on auditory research* (Springer), pp. 417–440.

Persson, A., and Jaeger, T. F. (**2023**). “Evaluating normalization accounts against the dense vowel space of Central Swedish,” Frontiers in Psychology **14**, doi: 10.3389/fpsyg.2023.1165742.

Peterson, G. E. (**1961**). “Parameters of vowel quality,” Journal of Speech and Hearing Research **4**(1), 10–29, doi: 10.1044/jshr.0401.10.

R Core Team (**2023**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Repp, B. H., and Crowder, R. G. (**1990**). "Stimulus order effects in vowel discrimination," The Journal of the Acoustical Society of America **88**(5), 2080–2090, doi: 10.1121/1.400105.

Richter, C., Feldman, N. H., Salgado, H., and Jansen, A. (**2017**). "Evaluating low-level speech features against human perceptual data," Transactions of the Association for Computational Linguistics **5**, 425–440, doi: 10.1162/tacl_a_00071.

RStudio Team (**2020**). *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA.

Saenz, M., and Langers, D. R. (**2014**). "Tonotopic mapping of human auditory cortex," Hearing Research **307**, 42–52, doi: 10.1016/j.heares.2013.07.016 human Auditory NeuroImaging.

Scarborough, R. (**2010**). "Lexical and contextual predictability: Confluent effects on the production of vowels," in *Laboratory Phonology*, edited by C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, **10** (De Gruyter Mouton Berlin), pp. 557–586.

Schertz, J., and Clare, E. J. (**2020**). "Phonetic cue weighting in perception and production," Wiley Interdisciplinary Reviews: Cognitive Science **11**(2), doi: 10.1002/wcs.1521.

Shankweiler, D., Verbrugge, R. R., and Studdert-Kennedy, M. (**1978**). "Insufficiency of the target for vowel perception," The Journal of the Acoustical Society of America **63**(S1), S4–S4, doi: 10.1121/1.2016686.

Shannon, C. E. (**1948**). "A mathematical theory of communication," The Bell System Technical Journal **27**(3), 379–423, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Siegel, R. J. (**1965**). "A replication of the mel scale of pitch," The American Journal of Psychology **78**(4), 615–620, doi: 10.2307/1420924.

Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (**2019**). "Speaker-normalized sound representations in the human auditory cortex," Nature Communications **10**(1), 01–09, doi: 10.1038/s41467-019-10365-z.

Skoe, E., Krizman, J., Spitzer, E. R., and Kraus, N. (**2021**). "Auditory cortical changes precede brainstem changes during rapid implicit learning: Evidence from human EEG," Frontiers in Neuroscience **15**, 01–09, doi: 10.3389/fnins.2021.718230.

Smith, B. L., Johnson, E., and Hayes-Harb, R. (**2019**). "ESL learners' intra-speaker variability in producing American English tense and lax vowels," Journal of Second Language Pronunciation **5**(1), 139–164, doi: 10.1075/jslp.15050.smi.

Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," The Journal of the Acoustical Society of America **117**(1), 305–318, doi: 10.1121/1.1828637.

Steriade, D. (**2008**). "The phonology of perceptibility effects: the P-map and its consequences for constraint organization," in *The Nature of the Word: Studies in Honor of Paul Kiparsky*, edited by K. Hanson and S. Inkelas (MIT Press, UCLA), doi: 10.7551/mitpress/9780262083799.001.0001.

Stevens, K. N. (**1972**). "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human communication: a unified view* (McGraHill, New York), pp. 51–66.

Stevens, K. N. (**1989**). "On the quantal nature of speech," Journal of phonetics **17**(1-2), 3–45.

Stevens, S. S., and Volkmann, J. (**1940**). "The Relation of Pitch to Frequency: A Revised Scale," The American Journal of Psychology **53**(3), 329–353, doi: 10.2307/1417526.

Stilp, C. (**2020**). "Acoustic context effects in speech perception," WIREs Cognitive Science **11**(1), 1–18, doi: 10.1002/wcs.1517.

Strange, W., and Jenkins, J. J. (**2012**). "Dynamic specification of coarticulated vowels: Research chronology, theory, and hypotheses," in *Vowel Inherent Spectral Change* (Springer), pp. 87–115.

Sumner, M. (**2011**). "The role of variation in the perception of accented speech," Cognition **119**(1), 131–136, doi: 10.1016/j.cognition.2010.10.018.

Syrdal, A. K. (**1985**). "Aspects of a model of the auditory representation of American English vowels," Speech Communication **4**(1-3), 121–135, doi: 10.1016/0167-6393(85)90040-8.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," The Journal of the Acoustical Society of America **79**(4), 1086–1100, doi: 10.1121/1.393381.

Tan, M., and Jaeger, T. F. (**2024**). "Incremental adaptation to an unfamiliar talker," Manuscript, Stockholm University .

Tang, C., Hamilton, L. S., and Chang, E. F. (**2017**). "Intonational speech prosody encoding in the human auditory cortex," Science **357**(6353), 797–801, doi: 10.1126/science.aam8577.

ten Bosch, L., Boves, L., Tucker, B., and Ernestus, M. (**2015**). "DIANA: towards computational modeling reaction times in lexical decision in north American English," in *Proc. Interspeech 2015*, pp. 1576–1580, doi: 10.21437/Interspeech.2015-366.

Traunmüller, H. (**1981**). "Perceptual dimension of openness in vowels," The Journal of the Acoustical Society of America **69**(5), 1465–1475, doi: 10.1121/1.385780.

Traunmüller, H. (**1990**). "Analytical expressions for the tonotopic sensory scale," The Journal of the Acoustical Society of America **88**(1), 97–100, doi: 10.1121/1.399849.

Vaughn, C., Baese-Berk, M., and Idemaru, K. (**2019**). "Re-examining phonetic variability in native and non-native speech," Phonetica **76**(5), 327–358, doi: 10.1159/000487269.

Vorperian, H. K., and Kent, R. D. (**2007**). "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," Journal of Speech, Language & Hearing Research **50**(6), 1510–1545, doi: 10.1044/1092-4388(2007/104).

Wade, T., Jongman, A., and Sereno, J. (**2007**). "Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds," Phonetica **64**(2-3), 122–144, doi: 10.1159/000107913.

Walker, A., and Hay, J. (**2011**). "Congruence between 'word age' and 'voice age' facilitates lexical access," Laboratory Phonology **2**(1), 219–237, doi: 10.1515/labphon.2011.007.

Watt, D., and Fabricius, A. (**2002**). "Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 ~ F2 plane," in *Leeds Working Papers in Linguistics and Phonetics*, edited by D. Nelson, 9, pp. 159–173.

Weatherholtz, K., and Jaeger, T. F. (**2016**). "Speech perception and generalization across talkers and accents," Oxford Research Encyclopedia of Linguistics doi: 10.1093/

acrefore/9780199384655.013.95.

Wedel, A., Nelson, N., and Sharp, R. (**2018**). "The phonetic specificity of contrastive hyperarticulation in natural speech," Journal of Memory and Language **100**, 61–88, doi: 10.1016/j.jml.2018.01.001.

Whalen, D. H. (**2016**). "A double-Nearey theory of vowel normalization: Approaching consensus," The Journal of the Acoustical Society of America **140**(4_Supplement), 3163–3164, doi: 10.1121/1.4969932.

Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Perception & psychophysics **63**(8), 1293–1313, doi: 10.3758/BF03194544.

Winn, M. (**2018**). "Speech: It's not as acoustic as you think," Acoustics Today **12**(2), 43–49.

Xie, X., Buxó-Lugo, A., and Kurumada, C. (**2021**). "Encoding and decoding of meaning through structured variability in speech prosody," Cognition **211**, 1–27, doi: 10.1016/j.cognition.2021.104619.

Xie, X., and Jaeger, T. F. (**2020**). "Comparing non-native and native speech: Are L2 productions more variable?," The Journal of the Acoustical Society of America **147**(5), 3322–3347, doi: 10.1121/10.0001141.

Xie, X., Jaeger, T. F., and Kurumada, C. (**2023**). "What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review," Cortex **166**, 377–424, doi: 10.1016/j.cortex.2023.05.003.

Zahorian, S. A., and Jagharghi, A. J. (**1991**). "Speaker normalization of static and dynamic vowel spectral features," The Journal of the Acoustical Society of America **90**(1), 67–75, doi: 10.1121/1.402350.

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," The Journal of the Acoustical Society of America **33**(2), 248–248, doi: 10.1121/1.1908630.

Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). "Critical band width in loudness summation," The Journal of the Acoustical Society of America **29**(5), 548–557, doi: 10.1121/1.1908963.

Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," The Journal of the Acoustical Society of America **68**(5), 1523–1525, doi: 10.1121/1.385079.