Re-introducing the probabilistic sliding template model of vowel perception

Santiago Barreda and T. Florian Jaeger

Abstract: Research in phonetics and phonology relies on strong assumptions about phonetic features (e.g., relevance of formants to vowel perception). This includes assumptions about how speech signal acoustics are normalized into phonetic representations that are comparatively invariant to talker physiology. Despite the fundamental importance of such normalization to speech perception and linguistic research, almost all popular normalization accounts make the obviously wrong assumptions that the relevant normalization parameters are already 'known'—e.g., because the researcher can estimate them from a fully balanced set of recordings. Listeners, however, have to incrementally infer these parameters from the speech input. A seminal, but still underappreciated, exception is the Probabilistic Sliding Template Model of vowel normalization and perception (PSTM, Nearey and Assmann, 2007). We introduce a fully Bayesian variant of the PSTM, along with the R package `STM`, which makes it trivial to apply the model to new data using only a few lines of R code. The PSTM integrates social and linguistic inferences into a joint inference process. This allows investigation of connections between speech perception, social knowledge, and the estimation of speaker indexical characteristics. A new bootstrap validation shows that the PSTM explains listener behavior much better than the hugely popular Lobanov normalization.

1. Introduction

Cross-talker variability poses a computational challenge for speech perception: since different talkers can realize the same phonetic information with different acoustics, robust perception across talkers can only be achieved by somehow adjusting to these differences. At the same time, cross-talker differences in pronunciation provide information about the language background and social identity of speakers. One way to approach these issues has been to describe both linguistic and social perception as inference over the joint distribution of auditory percepts (e.g., in the form of phonetic cues), linguistic, and social categories. This can be conceptualized as similarity-based inferences over exemplars (Johnson 1997; Sumner 2011) or as Bayesian inferences (Kleinschmidt and Jaeger 2015; Kleinschmidt, Weatherholtz, and Jaeger 2018). However, there is of yet no widely available model that implements these ideas and makes them testable. This motivates the present work.

Focusing on vowel perception, we describe the *Probabilistic Sliding Template Model (PSTM)*, first presented in Nearey and Assmann (2007), and based on Terrance Nearey's sliding template model (STM, Nearey 1978). The PSTM is a Bayesian model of *vowel normalization*, the perceptual mapping between acoustic information and some linguistic representation. It integrates social and linguistic inferences into a joint process, simultaneously estimating speaker characteristics and categorizing vowels. In doing so, the PSTM allows for the empirical investigation of the connection between speech perception, social knowledge, and the estimation of speaker indexical characteristics. Further, unlike most existing models of normalization, the PSTM infers the speaker characteristics required for normalization *incrementally* from the speech input, doing so—as we show below—with high accuracy even from a single vowel input. In doing so, the PSTM helps address the 'bootstrap' problem of speech normalization: listeners must know some speaker characteristics in order to normalize (and identify) speech but must accurately identify speech in order to accurately estimate the speaker characteristics required to normalize. Speaker characteristics and speech acoustics enter into the same many-to-many mapping relationship as the linguistic signal and speech acoustics. Consequently, determining both simultaneously from the speech signal is difficult, and the PSTM is one of the few models that addresses this problem in an empirically testable manner.

This incremental inference mechanism stands out even when compared to the type of modern models of automatic speech recognition (ASR; e.g., HuBERT, Wav2vec, Whipser). These models have made tremendous progress in understanding human speech, including when faced with inter-talker variability (e.g., Kim et al. 2024). Recent work further suggests that the latent perceptual representations learned by these ASR models capture some aspects of human speech perception, such as the ability to generalize to unfamiliar accents (Chernyak et al. 2024; Jin, Zhu, and Jaeger 2025). Unlike the PSTM, none of these models yet exhibits the incremental changes in perception that have been shown to occur in human listeners when exposed to different talkers of familiar accents (Magnuson and Nusbaum 2007; Barreda 2012; Persson, Barreda, and Jaeger 2025).

Notably, the PSTM is rooted in principled considerations about the biology of auditory perception in the mammalian brain, validated by cross-species comparisons. The uniform scaling of spectral patterns—the formant normalization procedure proposed by (P)STM—is closely related to size variation due to the physics of acoustic resonators. Indeed, several animal species, such as koalas (Charlton et al. 2012), dogs (Taylor, Reby, and McComb 2010), and red deer (Reby et al. 2005), have been found to separate uniform scaling (size information) from the communicative information present in the signal (for review, Barreda 2020).

1.1 Why we wrote this paper, and why you might want to read it

We believe that Nearey and Assmann (2007) is one of the most ground-breaking papers in speech perception that 'nobody has ever read'. As of the writing of this article, it had been cited less than 40 times over 17 years. During the same time, there have been 314,000 research articles on vowels. Over 17,000 of these mention vowel formants, vowel identification/categorization/recognition, and/or vowel/formant perception; over 5,000 additionally mention normalization. Among these articles, Lobanov normalization (Lobanov 1971) is cited three- to four-times as often as Nearey's STM or "uniform scaling" (Nearey 1978) and about 100-times more often than the PSTM (Nearey and Assmann 2007)—despite the fact that the latter is the only the only published article that describes *any* incremental inference model for vowel perception. So, here is a ground-breaking idea and promising candidate model of formant/spectral talker-normalization, with far-reaching consequences for research from typology to sociolinguistics to speech perception, … and few seem to know of it.

Why? Nearey and Assmann (2007) is a technical paper, written by experts for experts. It presents seven different models, without much explanation as to how they are derived, and no code to execute them. Even with access to additional notes kindly provided by the late Terrence Nearey, it took us days to verify some of the core theoretical results presented in Nearey and Assmann (2007). We imagine that the model is so underutilized in large parts because nobody knows how to implement it. Here we describe the PSTM and present a more fully Bayesian extension—the *Bayesian Sliding Template Model (BSTM)*—envisioned in Nearey and Assmann (2007) but not previously available. We walk through the logic of these models step by step and provide visualizations of core concepts. Our supplementary information (SI) contains detailed derivations for the most relevant models. The content of the present paper is still somewhat technical but sufficiently detailed to make it reproducible. To further facilitate adoption of the model by other researchers, we have implemented all models describe here in the R package `STM`. `STM` is designed to be user-friendly and flexible (constructive user feedback is welcome). The present article uses the `STM` package and is written in Quarto markdown. This markdown document, and all R code it sources, are available as part of the OSF repository at https://osf.io/tpwmv/. This allows interested researchers to re-create all of our analyses with the press of a button in RStudio (R Core Team 2023; RStudio Team 2024), and to apply similar analyses to their own data.

A second reason as to why the PSTM has not received more attention is that it has not previously been tested against human perception. Nearey and Assmann (2007) primarily evaluated the PSTM in terms of its ability to recognize the vowel intended by the talker. However, as we also demonstrate below, a model of human speech perception should be evaluated in terms of its ability to predict *listeners' perception*, not talkers' intentions (for discussion, see Barreda 2020; Persson, Barreda, and Jaeger 2025). While Nearey and Assmann (2007) also presented initial qualitative comparisons of PSTM's predictions and listeners' perception, those comparisons leave open how good a model the PSTM is of human perception. Here, we address this question. We find that the incremental inference procedure proposed for the PSTM explains listeners' perception very well, and *substantially* better than the most frequently used normalization account (Lobanov).

2. Perceptual normalization through uniform scaling

A core assumption of normalization accounts is that dialect-specific vowel representations are learned and represented in a *normalized* formant space. We start by motivating and describing the normalized formant space assumed by the models we present below.

Formant inputs from different talkers tend to be perceived as phonetically similar if they differ only according to a single multiplicative scaling parameter (for a review, see Barreda 2020). For example, in order to preserve phonetic similarity, a speaker who produces an F1 that is 10% higher for /a/ than another should also produce F2 and F3 that are 10% higher for that vowel—i.e., all formants are *uniformly scaled*. Consider a dialect-specific formant target $\vec{F}_{\mathrm{v}}^{*} := [F1_{\mathrm{v}}^{*}, F2_{\mathrm{v}}^{*}, F3_{\mathrm{v}}^{*}, \dots]$ for vowel v. A speaker s of that dialect, should target formants $\vec{F}_{\mathrm{v}}^{s*} := [F1_{\mathrm{v}}^{s*}, F2_{\mathrm{v}}^{s*}, F3_{\mathrm{v}}^{s*}, \dots]$ that are *uniformly scaled* by a speaker-specific scaling parameter ρ_s, as in Equation 1.

$$\vec{F}_{\mathrm{v}}^{s*} := [F1_{\mathrm{v}}^{s*}, F2_{\mathrm{v}}^{s*}, F3_{\mathrm{v}}^{s*}, \dots] = [F1_{\mathrm{v}}^{*}, F2_{\mathrm{v}}^{*}, F3_{\mathrm{v}}^{*}, \dots] \cdot \rho_s \qquad (1)$$

Equation 1 can be re-expressed as the sum of log-transformed formant targets $\vec{G}_{\mathrm{v}}^{*} := [G1_{\mathrm{v}}^{*}, G2_{\mathrm{v}}^{*}, G3_{\mathrm{v}}^{*}] = \log(\vec{F}_{\mathrm{v}}^{*})$, where $\psi_s = \log(\rho_s)$, as in Equation 2. In log-transformed Hz, uniform scaling thus results in additive, rather than multiplicative, changes. This relation between multiplicative scaling in Hz and additive scaling in log-Hz is illustrated in Figure 1.

$$\vec{G}_{\mathrm{v}}^{s*} := [\ G1_{\mathrm{v}}^{s*},\ G2_{\mathrm{v}}^{s*},\ G3_{\mathrm{v}}^{s*}, \ldots] = [\ G1_{\mathrm{v}}^{*},\ G2_{\mathrm{v}}^{*},\ G3_{\mathrm{v}}^{*}, \ldots] + [\psi_s, \psi_s, \psi_s, \ldots] \qquad (2)$$

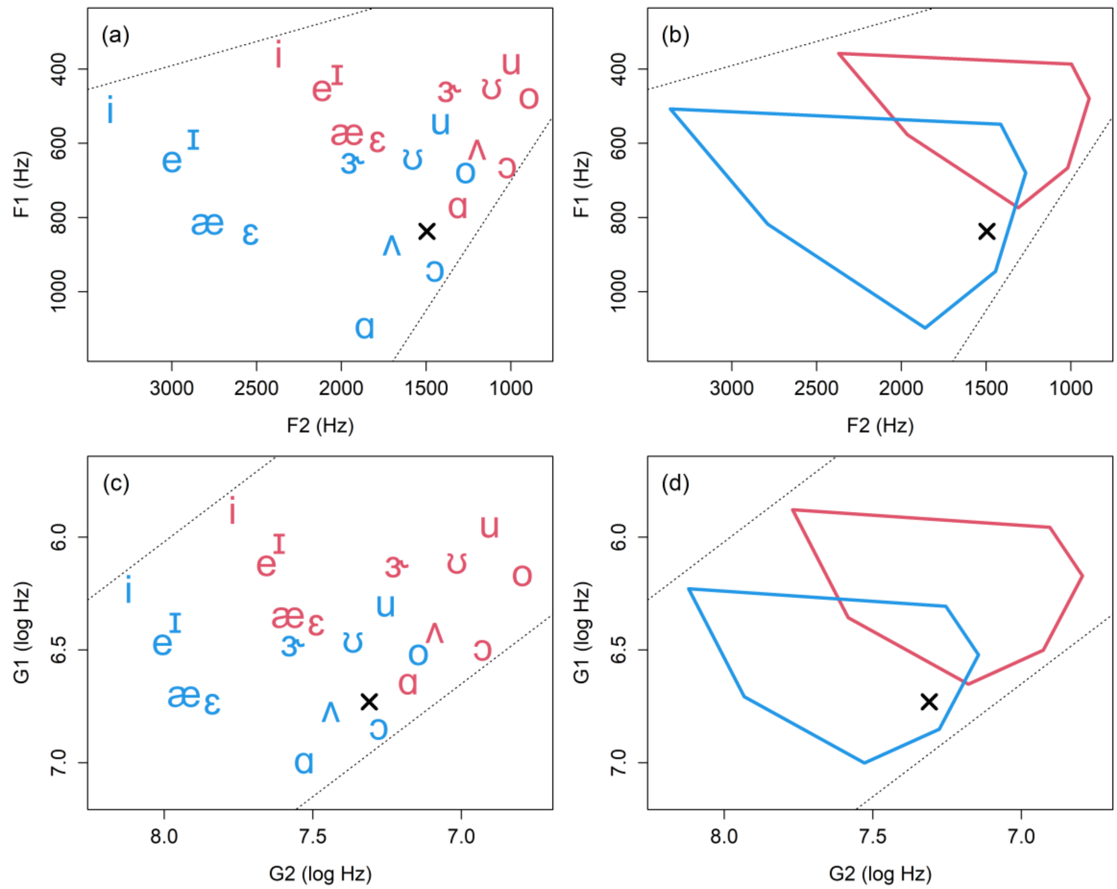

Figure 1: (a) A comparison of two vowel spaces (based on data in Hillenbrand et al., 1995) differing according to a single multiplicative scaling parameter (ρ_s). Each IPA symbols shows a speaker's vowel target $\vec{F}_{\mathrm{v}}^{s*}$ in an F1-F2 space. The "x" marks an ambiguous vowel token whose interpretations depends on the assumed vowel space. (b) The outline of the same vowel spaces presented as polygons. Uniform scaling expands or shrinks the polygons but preserves their shape. (c) The same comparison expressed in log-transformed Hz. Each IPA symbols shows a speaker's vowel target $\vec{G}_{\mathrm{v}}^{s*}$ in a log-F1-F2 space. The two vowel systems now differ in ψ_s. (d) When expressed in log-Hz, uniform scaling results in shifts of the polygons without changing their shape and size.

The interpretation of any formant input—e.g., the "x" in Figure 1 —will thus depend on the scaling parameter ψ_s. Specifically, to transform a formant input $\vec{F}$ into the uniformly scaled—i.e., normalized—dialect-specific formant space, listeners only need to subtract the speaker-specific scaling parameter $\vec{\psi}_s := [\psi_s, \psi_s, \psi_s, \ldots]$ from the log-transformed formant input $\vec{G} := [G1, G2, G3, \ldots]$ as in Equation 3. We refer to the resulting normalized formants as $\vec{N}$:

$$\vec{N} := [N1,\ N2,\ N3,\ \ldots] = \vec{G} - \vec{\psi}_s \qquad (3)$$

Such uniform scaling provides a comparatively parsimonious approach, assuming that listeners infer and store a single parameter. To put this into perspective, the perhaps the most commonly used normalization method in speech research—Lobanov normalization—assumes that listeners infer and store two parameters per feature. For instance, to model the effects of the first three formants at two time points in the vowel (as we do below), listeners would need to estimate twelve independent parameters. This linearly scaling complexity of Lobanov normalization is one of the reasons it might not be an adequate model for aspects of human speech perception that are thought to be evolutionary quite old [see above; for related discussion, see Barreda (2020); Barreda (2021); Persson, Barreda, and Jaeger (2025)].

3. The Probabilistic Sliding Template Model of vowel perception (PSTM)

As anticipated above, listeners are assumed to learn dialect-specific vowel representations over the *normalized* vowel formants. These representations are assumed to capture the formant distributions that result for each vowel due to noise during the articulation and perception of formants. Nearey and Assmann (2007) referred to listeners' dialect-specific implicit knowledge about each vowel's formant distributions as "probabilistic templates".

The probabilistic templates could be represented, for instance, as exemplar clouds (Johnson 1997) or, in a more compact form, as parametric distributions (Nearey and Assmann 2007). For computational tractability, we follow the latter approach and represent vowel categories as multivariate Normal distributions over normalized formants, with mean vectors $\vec{\mu}_{\mathrm{v}}$ and covariance matrices $\vec{\Sigma}_{\mathrm{v}}$. We may then use the multivariate normal density (MVN) of each vowel to calculate the likelihood of any observed log-transformed formant input $\vec{G}$ given that vowel and the speaker-specific scaling parameter ψ_s. This can be done by either sliding the template as in Figure 1(d) above or, equivalently, by scaling the formants into the normalized template space:

$$P(\vec{G}|\mathrm{v}, \psi_s) := MVN(\vec{G}|\vec{\mu}_{\mathrm{v}} + \vec{\psi}_s, \vec{\Sigma}_{\mathrm{v}}) = MVN(\vec{G} - \vec{\psi}_s|\vec{\mu}_{\mathrm{v}}, \vec{\Sigma}_{\mathrm{v}}) = MVN(\vec{N}|\vec{\mu}_{\mathrm{v}}, \vec{\Sigma}_{\mathrm{v}}) \qquad (4)$$

The likelihoods in Equation 4 can be combined with the prior probability of each vowel category in the current context to find the *posterior probability* of each vowel category v_i, as in Equation 5 (where V is the number of unique vowel categories).

$$P(\mathrm{v}_i|\vec{G}, \psi_s) = \frac{P(\vec{G}|\mathrm{v}_i, \psi_s) \cdot P(\mathrm{v}_i)}{\sum_{j=1}^{V} P(\vec{G}|\mathrm{v}_j, \psi_s) \cdot P(\mathrm{v}_j)} \qquad (5)$$

This posterior probability provides a gradient measure of category membership (Nearey and Assmann 2007; Luce and Pisoni 1998; Norris and McQueen 2008; Xie, Jaeger, and Kurumada 2023)—conceptually paralleling the gradient activation of categories in connectionist, neural network, or exemplar theories of speech perception. Categorization—and thus listeners' responses in a n-alternative forced-choice categorization task—can be modeled via decision rules based on the posterior probability. For instance, listener might always respond with the vowel that has the highest posterior probability (criterion choice rule, as assumed in Nearey and Assmann 2007) or respond by sampling from the posterior (Luce's choice rule, cf. discussion in Massaro and Friedman 1990).

Given an estimate of the speaker-specific scaling parameter ψ_s, listeners can thus categorize formant inputs under any dialect-specific template. Similarly, researchers can use speaker-specific ψ_s estimates to project formant data from various talkers into a common normalized space. This can be useful when comparing vowel productions across different types of speakers—for instance, to assess effects of social identity, language background, or to compare clinical and neurotypical populations. However, a speaker's ψ_s is a latent variable not directly present in their speech. As a result, both researchers and listeners must *estimate* ψ_s.

4. Probabilistic estimation of ψ_s

For researchers who are only interested in projecting formant estimates $\vec{F}$ from different talkers into a common normalized space, estimation of ψ_s can be relatively straightforward. Provided access to a data set with sufficiently many formant estimates per speaker, and the same number of formant estimates per vowel for each speaker, researchers can obtain speaker-specific estimates $\hat{\psi}_s$ by simply averaging all log-transformed formant inputs $\vec{G}$ (method 1 from Nearey and Assmann 2007):

$$\hat{\psi}_s := \bar{G} = \frac{1}{(K \cdot N)} \sum_{j=1}^{N} \sum_{k=1}^{K} G_{j,k,s} \qquad (6)$$

where N is the number of observations per talker, K is the number of formants per input (e.g. $K = 2$ if only F1 and F2 are considered), and $G_{j,k,s}$ is the kth formant of the j vowel observation of speaker s. Under this formulation, uniform scaling is a form of extrinsic normalization, relying on information that is collected across observations.

This simple approach quickly becomes unfeasible even for researchers. First, estimates based on Equation 6 can be highly sensitive to formant measurement errors when the number of recordings per vowel is small. Second, even for large data sets, the approach in Equation 6 runs into substantial problems when comparing within or across data sets with different numbers of observations per vowel and talker (for discussion, Barreda and Nearey 2018; Nearey and Assmann 2007; Xie, Jaeger, and Kurumada 2023, SI 2.1). This problem arises for the same reasons that make formants relevant to vowel perception in the first place: the distribution of formants—and thus their mean—differs between vowels. This means that Equation 6 will yield systematically different estimates $\hat{\psi}_s$, even for the same speaker with the same underlying ψ_s, depending on the specific set of vowel instances over which ψ_s is estimated.

The same considerations make the approach in Equation 6 unsuitable for *listeners*. And, unlike researchers, listeners who encounter an unfamiliar talker do not have the luxury of waiting until they have observed a large amount of formant inputs from that talker: if normalization via uniform scaling is to aid robust speech perception, listeners must quickly arrive at adequate estimates of ψ_s—ideally within the very first instance of a vowel that they heard from the unfamiliar talker. If listeners were using something like Equation 6 to estimate ψ_s for individual vowel tokens, their estimates of ψ_s would vary substantially from trial to trial—depending also on which vowels the talker has produced. This would result in both highly inaccurate speech perception and very large changes to inferred speaker characteristics that themselves depend on ψ_s (e.g., size and gender). Neither of these outcomes are observed in the literature.

Nearey and Assmann (2007) addressed this issue by proposing several probabilistic approaches to ψ_s estimation that do not rely on a balanced sample of the speaker's entire vowel system. These models share two key methodological insights. First, they constrain estimates of ψ_s given the listener's assumed phonological knowledge (i.e. the template). Second, listeners are assumed to have prior expectations about the distribution of ψ_s depending on the acoustic and indexical characteristics of the talker. Here, we focus on three of the methods proposed by Nearey and Assmann (methods 2, 3, and 6). These methods differ only in their assumptions about listeners' prior expectations about ψ_s. Specifically, the three methods make increasingly stronger assumptions about the implicit knowledge that listeners have learned and stored about the distribution of ψ_s. By comparing how well these methods fit listeners' behavior, one can therefore test hypotheses about the type of implicit knowledge listeners have. Before we turn to the differences between the three methods, we describe their shared characteristics.

4.1 Estimating ψ_s using prior expectations about vowel templates and ψ_s

Methods 2, 3, and 6 share that they estimate the *maximum a posteriori* (MAP) value of ψ_s. Nearey and Assmann (2007) did not provide the derivation of these MAP estimates, but they are based on the posterior distribution of ψ_s given the observed formant pattern, the vowel category, and the listener's prior expectations about the distributions of ψ_s and vowel categories:

$$P\left(\psi_s|\vec{G},\mathrm{v}_j\right) = \frac{P\left(\vec{G}|\mathrm{v}_j,\psi_s\right)\cdot P(\psi_s)\cdot P(\mathrm{v})_j}{\sum_{i=1}^{V} P\left(\vec{G}|\mathrm{v}_i,\psi_s\right)\cdot P(\psi_s)\cdot P(\mathrm{v}_i)} \qquad (7)$$

As per Equation 4, the likelihood $P\left(\vec{G}|\mathrm{v},\psi_s\right)$ given the vowel category and ψ_s is assumed to be a multivariate normal distribution. Researchers can estimate the mean $\vec{\mu}_v$ and covariance matrix $\vec{\Sigma}_v$ for each vowel category based on any reasonably sized database of vowel formants from the dialect(s) that the listener is assumed to have learned their template(s) from.[1] Similarly, $P(\mathrm{v})$, can be calculated from relevant speech corpora or, for many experimental contexts, assumed to be equal across all categories and ignored in the calculation.

The posterior distribution in Equation 7 can be used to obtain separate MAP estimates of ψ_s for each vowel category, $\hat{\psi}_{s,\mathrm{v}}$. A vowel system with V vowels would result in V MAP estimates $\hat{\psi}_{s,\mathrm{v}}$:

$$\begin{aligned}
\hat{\psi}_{s,\mathrm{v}=1} &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v}{=}1, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v} = 1)\right] \\
\hat{\psi}_{s,\mathrm{v}=2} &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v}{=}2, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v} = 2)\right] \\
&\vdots \\
\hat{\psi}_{s,\mathrm{v}=\mathrm{V}} &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v}{=}\mathrm{V}, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v} = V)\right]
\end{aligned} \tag{8}$$

Nearey and Assmann (2007) use the best $\hat{\psi}_{s,\mathrm{v}}$ for each vowel to calculate the posterior probability of each category given the formant input, as in Equation 9. Note that this is an update of Equation 5 using vowel-specific estimates of ψ_s, and the prior probability of that estimate.

$$P\left(\mathrm{v}_j|\vec{G}, \hat{\psi}_{s,\mathrm{v}=j}\right) = \frac{P\left(\vec{G}|\mathrm{v}_j, \hat{\psi}_{s,\mathrm{v}=j}\right) \cdot P\left(\psi_s = \hat{\psi}_{s,\mathrm{v}=j}\right) \cdot P\left(\mathrm{v}_j\right)}{\sum_{i=1}^{V} P\left(\vec{G}|\mathrm{v}_i, \hat{\psi}_{s,\mathrm{v}=i}\right) \cdot P\left(\psi_s = \hat{\psi}_{s,\mathrm{v}=i}\right) \cdot P(\mathrm{v}_i)} \tag{9}$$

These $\psi_{s,\mathrm{v}}$ are then used to categorize the formant input into a winning category, c. Specifically, listeners are assumed to infer the $\hat{\psi}_s$ with the highest MAP, and to use this $\hat{\psi}_s$ to categorize the input as the vowel category with the maximum posterior density (Equation 10). Nearey and Assmann describe this process as “choose the vowel that looks best when it tries to look its best” (p. 253).

$$\begin{aligned}
\hat{\psi}_s &:= \underset{\psi_s}{\operatorname{argmax}}\left[\hat{\psi}_{s,\mathrm{v}}\right] \\
c &:= \underset{\mathrm{v}}{\operatorname{argmax}}\left[\hat{\psi}_{s,\mathrm{v}}\right]
\end{aligned} \tag{10}$$

4.2 Method 2: Unrestricted optimization of ψ_s

Method 2 spells out Equation 7 as Equation 11, assuming that listeners’ prior expectations about ψ_s correspond to a uniform distribution (U) such that any value of ψ_s is equally plausible.

---

[1] Nearey and Assmann (2007) calculated a single shared covariance matrix across all vowel categories. Here, we use separate covariance matrices for each vowel category. The `STM` library supports both options.

$$\begin{aligned}\hat{\psi}_{s,\mathrm{v}} \quad &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v},\psi_s\right)\cdot P(\psi_s)\cdot P(\mathrm{v})\right] \\ &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v},\psi_s\right)\cdot \mathrm{U}(\psi_s)\cdot P(\mathrm{v})\right]\end{aligned} \qquad (11)$$

Using a uniform prior on ψ_s means that our 'best' estimate of ψ_s can be rather implausible, as long as it provides a good fit to one of our categories. For example, Figure 2 presents the likelihood functions $P\left(\vec{G}_x|\mathrm{v}_i,\hat{\psi}_s\right)$ from Equation 11 for four vowel hypotheses, along with different interpretations of an ambiguous vowel according the MAP estimates $\hat{\psi}_{s,\mathrm{v}}$ for the four vowels. Because the prior exerts no influence on the posterior, the likelihoods and posteriors share the same relative shapes in this case. As a result, method 2 suggests that /o/, /ɑ/, and /ʌ/ are all among the probable solutions. As we show next, however, some of these vowels would result in estimates of ψ that are *a priori* highly improbable (e.g., for /o/, $\hat{\psi}_s = 7.64$). Method 3 takes this into account.

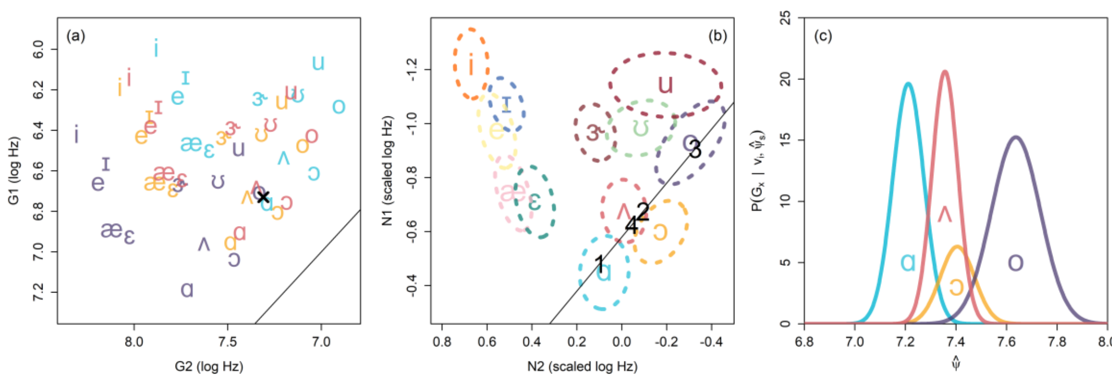

Figure 2: (a) A formant input "x", relative to four possible ways of sliding the dialect template, corresponding to four different values of $\hat{\psi}_s$. The template 'slides' along lines parallel to $G1 = G2$, indicated on the figure. (b) Alternatively, the vowel can be thought of 'sliding' across the normalized template. The line indicates the possible interpretations of the point in (a), ellipses enclose one standard deviation. Points indicate the best possible locations for /a/ (1), /c/ (2), /o/ (3), and /ʌ/ (4), according to method 2. These interpretations correspond to the different vowel spaces in (a). (c) Posterior distributions of ψ_s given different vowel categories. Because the prior in method 2 exerts no influence, these posteriors are proportional to the likelihood: $P\left(\vec{G}_x|v_i,\hat{\psi}_s\right)$. Thus, these curves represent the values of the densities of the distributions in (b) along the line. These likelihoods highlight the relationship between categorization and ψ_s estimation.

4.3 Method 3: Informative expectations about ψ_s

Method 3 (Equation 12) introduces stronger constraints on listeners' prior expectations for ψ_s by assuming that ψ_s is normally distributed across speakers, with a mean μ_{ψ_s} and standard deviation σ_{ψ_s} based on the ψ_s that the listener previously observed across speakers. This means that values of $\hat{\psi}_{s,\mathrm{v}}$ that are closer to the population mean will be considered more generally plausible by that listener.

$$
\begin{aligned}
\hat{\psi}_{s,\mathrm{v}} \quad &:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v},\psi_s\right)\cdot P(\psi_s)\cdot P(\mathrm{v})\right] \\
&:= \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v},\psi_s\right)\cdot \mathrm{Normal}\left(\psi_s|\hat{\mu}_{\psi_s},\hat{\sigma}_{\psi_s}\right)\cdot P(\mathrm{v})\right]
\end{aligned}
\tag{12}
$$

Just like the multivariate Normal distributions for the vowel templates, estimates of μ_{ψ_s} and σ_{ψ_s} can be obtained from a reasonably sized database. Nearey and Assmann (2007) suggest $\hat{\mu}_{\psi_s} = 7.23$ and $\hat{\sigma}_{\psi_s} = 0.128$ based a database of 265 speakers of US English (86 adult females, 88 adult males, 91 children) from three dialect regions (James M. Hillenbrand et al. 1995; Peterson and Barney 1952; Assmann and Katz 2000).[2] The two leftmost columns of Figure 3 compare methods 2 and 3, showing that posterior probabilities can be affected even by comparatively weak constraints on the prior of ψ_s. This includes the fact that a value of $\hat{\psi}_s = 7.64$ is *a priori* improbable because it is more than 3 standard deviations higher than the mean of the prior distribution.

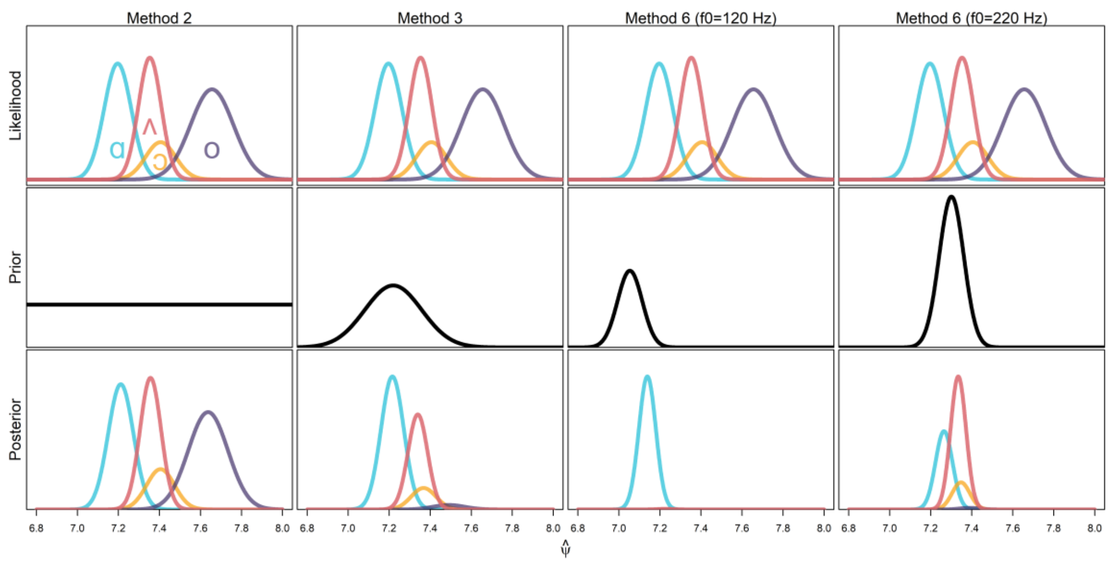

Figure 3: Comparison of different methods (columns) to estimating the prior probability of ψ_s. (top) Likelihood of ψ_s values given different vowel categories in Figure 2. (middle) Prior distribution of ψ_s. (bottom) Posterior probabilities of ψ_s conditional on vowel category resulting from combining the likelihood and prior in each column. The most probable $\hat{\psi}_s$ (location along x-axis of highest peak), and vowel category (color of curve with highest peak), changes as a function of the selected prior, and the vowel's $f0$.

4.4 Method 6: Informative expectations about ψ_s conditional on $f0$

Method 6 further strengthens the constraints on listeners' prior expectations for ψ_s by conditioning those expectations on talkers' $f0$. Specifically, method 6 finds the MAP

---

[2] Alternatively, researchers may use other data sources to replace or refine these estimates.

estimate of ψ_s while considering both the likelihood of the observed log-transformed formant input $\vec{G}$ and the likelihood of the observed $g0 := \log(f0)$, as in Equation 13.

$$\hat{\psi}_{s,\mathrm{v}} := \underset{\psi_s}{\operatorname{argmax}}\left[P\left(\vec{G}|\mathrm{v},\psi_s\right) \cdot P(g0|\psi_s) \cdot P(\psi_s) \cdot P(\mathrm{v})\right] \qquad (13)$$

Just like the mean and standard deviation of the normal prior for method 3, researchers can estimate $P(g0|\psi_s)$ in Equation 13 from a reasonably-sized database. Specifically, assuming a linear relationship between $g0$ and ψ_s with normally distributed residual uncertainty about $g0$, researchers can use linear regression to predict observed $g0_s$s from estimates of $\hat{\psi}_s = \bar{G}_s$ as in (Equation 6).[3] The intercept $\hat{\alpha}_{g0}$ and slope $\hat{\beta}_{g0}$ can then be used to predict the mean $\hat{\mu}_{g0}$ around which $g0$ is expected to be distributed given ψ_s with a standard deviation equal to the residual standard deviation of the linear regression, $\hat{\sigma}_{g0}$:

$$P(g0|\psi_s) := N\left(\hat{\mu}_{g0}, \hat{\sigma}_{g0}\right) = N\left(\hat{\alpha}_{g0} + \hat{\beta}_{g0} \cdot \psi_s, \hat{\sigma}_{g0}\right) \qquad (14)$$

Using the same database of speakers from method 3, Nearey and Assmann obtained $\hat{\alpha}_{g0} = -10.3$, $\hat{\beta}_{g0} = 2.14$, and $\hat{\sigma}_{g0} = 0.133$. All other steps required to obtain the MAP estimate of ψ_s parallel method 3. The two rightmost columns of Figure 3 illustrate how $f0$ can affect both the estimation of ψ_s and the posterior probabilities of different vowel categories.

4.5 Implementation

Nearey and Assmann (2007) note that, for the models they provide, "analytic solutions to the optimizations are available and no search is necessary" (p. 252). However, they did not provide details regarding these analytic solutions. Here, we provide the general approach to finding analytic solutions to the maximization of the posterior density of ψ_s for each vowel category. We focus on the derivation for method 6, as methods 3 and 2 comprise a subset of the necessary calculations. To find the MAP estimates $\hat{\psi}_{s,\mathrm{v}}$ for each vowel, we must calculate the product of the densities in Equation 15 for each vowel and find the maximum. The complete derivation is provided in the supplemental online materials.

$$\begin{aligned} P\left(\vec{G}|\mathrm{v},\psi_s\right) \cdot P(g0|\psi_s) \cdot P(\psi_s) \cdot P(\mathrm{v}) = \\ \mathrm{MVN}\left(\vec{N}_\psi \mid \vec{\hat{\mu}}_\mathrm{v}, \hat{\Sigma}_\mathrm{v}\right) \cdot \\ N\left(\hat{\alpha}_{g0} + \hat{\beta}_{g0} \cdot \psi_s, \hat{\sigma}_{g0}\right) \cdot \\ N\left(\hat{\mu}_\psi, \hat{\sigma}_\psi\right) \cdot \\ P(\mathrm{v}) \end{aligned} \qquad (15)$$

[3] Note that the database should have formant and $f0$ measurements for all vowels, and should be balanced with equally many instances of each vowel both within and across recorded talkers. Otherwise the mean of the log-formants, $\bar{G}_s$, will be systematically biased by differences in the distribution of vowel instances across talkers.

4.6 The Bayesian Sliding Template Model

The PSTM calculates MAP values of $\hat{\psi}_s$ for each vowel, and then uses the largest of these point estimates to categorize the observed formant input (Equation 10). Nearey and Assmann (2007) noted that "[i]t would also be possible [...] to use a fuller Bayesian approach by integrating over the values of ψ_s, rather than selecting the maximum. We leave that possibility for future research" (p. 254). In order to make the methods outlined above more fully Bayesian, we can instead focus on the joint posterior distribution of ψ_s and the vowel category v:

$$P\left(\mathrm{v}, \psi_s | \vec{G}\right) = \frac{P\left(\vec{G}|\mathrm{v}, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v})}{\sum_{i=i}^{V} \int P\left(\vec{G}|\mathrm{v}_i, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v}_i)\, d\psi} \qquad (16)$$

To find the posterior distribution of ψ_s, we marginalize over v:

$$P\left(\psi_s | \vec{G}\right) = \frac{\sum_{i=1}^{V} P\left(\vec{G}|\mathrm{v}_i, \psi_s\right) \cdot P(\mathrm{v}_i) \cdot P(\psi_s)}{\sum_{i=1}^{V} \int P\left(\vec{G}|\mathrm{v}_i, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v}_i)\, d\psi} \qquad (17)$$

To find the posterior probabilities of different vowel categories, we marginalize over ψ_s:

$$P\left(\mathrm{v}_j | \vec{G}\right) = \frac{\int P\left(\vec{G}|\mathrm{v}_j, \psi_s\right) \cdot P(\psi_s) \cdot P\left(\mathrm{v}_j\right) d\psi}{\sum_{i=1}^{V} \int P\left(\vec{G}|\mathrm{v}_i, \psi_s\right) \cdot P(\psi_s) \cdot P(\mathrm{v}_i)\, d\psi} \qquad (18)$$

Compared to focusing on MAP estimates of ψ_s, consideration of the posterior distribution in Equation 16 takes into account the uncertainty about the $\hat{\psi}_{s,\mathrm{v}}$ for different vowel categories. This results in a model of vowel perception that is more in line with Bayesian principles, suggesting the name Bayesian Sliding Template Model (BSTM) for this version of the PSTM. By calculating the full posterior distribution of $\hat{\psi}_s$ rather than just MAP estimates, the BSTM allows optimal information integration across multiple vowel tokens. This will allow future work to model the iterative estimation and refinement of $\hat{\psi}_s$ across vowel observations, something which is not possible with the original PSTM. It is, of course, an empirical question whether *listeners' behavior* is better explained by this more fully Bayesian model—a question to which we turn next.

5. Bootstrap Evaluation

To compare the performance of different implementations of the STM in understanding perceptual vowel normalization, we conduct a bootstrap analysis using data on the perception of US English vowels (James M. Hillenbrand et al. 1995). A step-by-step walk-through of the analysis is available as an R Markdown document as part of the OSF repo for this paper. As demonstrated in that vignette, the `STM` package allows researchers to conduct the type of analysis we present here on their own data, with just a handful of R commands.

The James M. Hillenbrand et al. (1995) data include acoustic measures from 139 speakers producing twelve vowels each, including formant measures at multiple time slices. This data also contains 12-alternative forced-choice categorization responses for each vowel token, aggregated across 20 listeners of the same dialect as the speakers (average accuracy = 95%).[4] Crucially, these responses were elicited over stimuli from the different speakers in the database, presented in randomized order—i.e., precisely the type of input for which the naive estimation of ψ_s (Equation 6) would yield utterly unreliable results for listeners (as we confirm below).

5.1 Approach

For each bootstrap data set, we compare approaches based on three performance metrics:

A) The **likelihood of *listeners'* categorization responses** $\Lambda := \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} \cdot \log(P_{ij})$, where C_{ij} is the number of times token i was classified as vowel category j, and $\log(P_{ij})$ is the log-transformed posterior probability for token i and category j. This is the critical measure of how well a method describes listeners' categorization behavior.

B) The **likelihood of the vowel category that the *talker intended to produce** (or, specifically, that the experimenter asked the talker to produce). This metric is closer to what a speech engineer would choose to evaluate normalization methods, as it measures how well the method performs in recognizing the *intended* vowel category. While this metric does *not* assess how well a method explains speech perception, it provides an important comparison, as we explain below.

C) The **root-mean square (RMS) error in estimating** ψ_s compared to an estimate of $\hat{\psi}$ obtained from a balanced sample of each speaker's entire vowel system, and Equation 6. Ultimately, researchers interested in assessing how well a method describes listeners' estimation of ψ_s should compare each method's predictions for $\hat{\psi}_s$ against listeners' estimates.

Of note, the use of likelihood as a metric to assess A) the fit against listeners' perception and B) the intended category differs from the approach in Nearey and Assmann (2007). Nearey and Assman instead used the accuracy (under the criterion choice rule) of predicting the intended category label for B) and, critically, did not evaluate against listeners' perception. The use of accuracy as a metric is now understood to be problematic, as it does not capture how well a model predicts the *distribution* of responses (for discussion, Persson, Barreda, and Jaeger 2025). Indeed, we find that our choice of a more adequate evaluation metric leads to a clearer result.

We used the following process, repeated over 1000 bootstrap iterations for each method (all functions referred to are from the `STM` package):

---

[4] The data used to estimate vowel templates should be carefully chosen to reflect the sort of listener of interest to the researcher. For example, a template trained on a database of Canadian English will not be suitable to model the perception of an Irish English listener (cf. discussion in Persson, Barreda, and Jaeger 2025).

1. Randomly divide the data into a 79-speaker training set and a 60-speaker testing set. Resample the training data (with replacement) at the speaker level. Sixty-three vowel tokens (4%) had one or two missing formant measurements (out of 6), representing 0.09% of the total formant measurements in the data. The missing values were imputed using stochastic regression imputation, independently for each iteration (using the `impute_NA` function).

2. Normalize the training data using log-mean uniform scaling normalization (the `normalize` function). This means estimating ψ_s for each speaker using their complete vowel system as in Equation 6, and normalizing as in Equation 3.

3. Estimate the dialect-specific category means and covariance matrices of all vowels using the normalized training data (with `create_template`). We follow Nearey and Assmann (2007) in estimating a single pooled covariance matrix across all vowel categories.[5] **Each iteration of the bootstrap thus simulates a 'listener' of the dialect with slightly different speech experience.**

4. For each token in the testing data, estimate $\hat{\psi}_{s,\mathrm{v}}$ for each vowel category using each method. Use these values of $\hat{\psi}_{s,\mathrm{v}}$ to obtain the winning estimate $\hat{\psi}_s$ and calculate the posterior probability of each vowel category.

5. Calculate performance metrics A)-C) described above.

We compare 11 different methods that differ in terms of how they estimate ψ_s, and whether they estimate ψ_s at all. Some of these were included in Nearey and Assmann (2007), others were not (incl. the popular Lobanov method):

- **No normalization**
  - input is $\vec{F}$ (Hz)
  - input is $\vec{G}$ (log Hz), allowing an assessment of whether log-transformation in and off itself helps.
- **Normalization with $\hat{\psi}_s$ estimated using the 'classic approach'** in Equation 6
  - over the entire training data (*PSTM1 (balanced data)*)
  - based on the individual token (*PSTM1 (single trial)*), as a more direct baseline for the remaining methods (all of which are based on individual tokens)
- **Normalization using the most commonly used normalization method (Lobanov 1971)**
  - using the entire training data (Hz). This method involves standardization of each speaker's formants by subtracting the mean and dividing by the standard deviation of each formant, independently for each time point.

---

[5] It is an interesting questions whether category-specific covariance matrices, as in Equation 4, provide a better fit to listener's responses. The `STM` package supports either option.

- **Normalization with $\hat{\psi}_s$ estimated using *PSTM***
  - method 2
  - method 3
  - method 6
- **Normalization with $\hat{\psi}_s$ estimated using *BSTM***
  - method 2
  - method 3
  - method 6

We used formant measurements at 20% and 80% of the vowel duration. This meant each vowel token was represented by a vector of length six, two measurements for each of three formants. We used two time points to parallel the analyses presented in Nearey and Assmann (2007), and because formant dynamics are known to affect vowel perception for the dialect recorded in the database (James M. Hillenbrand and Nearey 1999). By fitting multivariate normal distributions in this six-dimensional space, we assume that listeners have learned expectations about the (co)variance between formants within and across time points.

5.2 Results

The results of the bootstrap analysis are presented in Figure 4. Method 6 offers the best fit against listeners' categorization responses for the James M. Hillenbrand et al. (1995) data. These differences hold regardless of whether the PSTM or BSTM is used. Based on the bootstrap, all differences—except for the contrast between method 6 and unnormalized Hz—are significant, i.e., 95% quantiles do not overlap 0 in Figure 4 (a). However, method 6 performed better than unnormalized Hz in 94.4% of all bootstrap iterations. This shows the importance of considering prior information about ψ_s during speech perception, and also the potential usefulness of considering the joint distribution of ψ_s and $f0$ in perception.

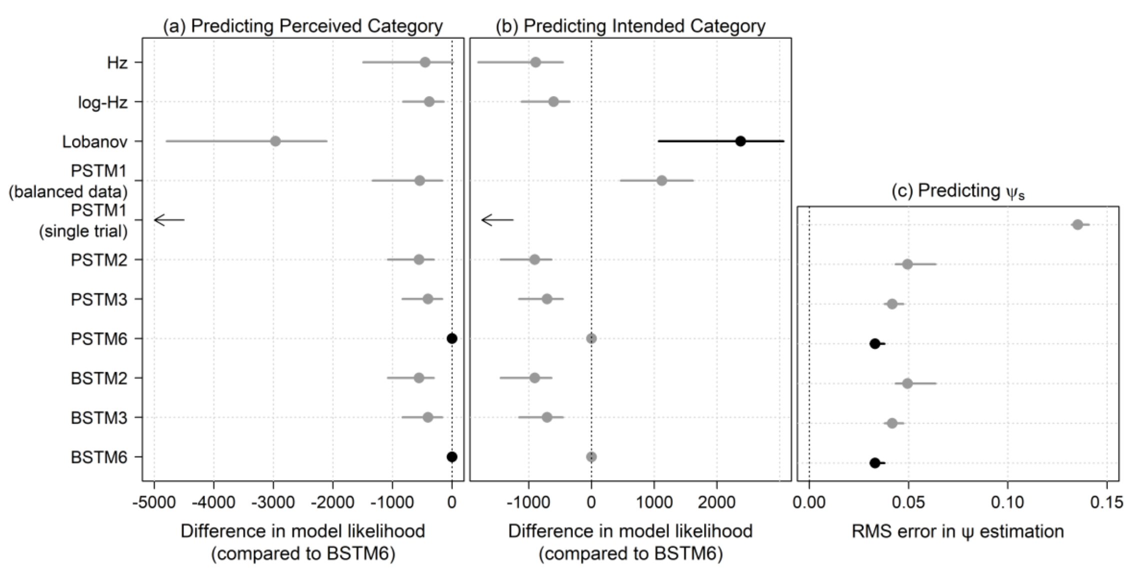

Figure 4: (a) Difference in model log-likelihoods of listeners' responses between each model and BSTM method 6 across bootstrap iterations. Points indicate bootstrap means; intervals indicate lower 95% quantiles. Arrows indicate the direction of log-likelihoods for the single-trial application of balanced-data method 1. These were more than an order of magnitude lower than all other log-likelihoods and thus do appear on the plot. (b) Same as (a) but for the vowel category that the talker intended to produce, rather than what listeners responded. (c) Root-mean-squared (RMS) error for ψ_s prediction for the approaches that estimate ψ_s. This leaves open which approach best predicts listeners' $\widehat{\psi}_s$ (to which we have no access here). Methods without normalization (Hz and log-Hz) are not shown since they do not provide estimates of ψ_s.

Of particular note, method 6 provides a better fit to listeners' responses than the 'classic' approach over balanced data (method 1 in Nearey and Assmann 2007). This is of note because the classic approach—which estimates ψ_s from the balanced and complete training data—has access to far more speaker-specific information than method 6. As a consequence, the classic approach achieves the best performance in predicting the vowel category that the *talker intended to produce* (see Figure 4(b)). Yet, the classic approach does *not* provide the best explanation for listeners' categorization responses. This highlights what should be obvious—and yet is often ignored in research on normalization: that listeners do not have access to a large and balanced set of vowel productions from a talker when they first encounter them. Effective speech perception thus requires mechanisms that can incrementally infer the relevant quantities (here ψ_s) based on prior expectations and other available information (e.g., $f0$, as in method 6).

As anticipated above, this finding—that method 6 outperforms the classic approach in predicting listeners' responses—differs from Nearey and Assmann (2007). This is evident from the fact that the classic approach excels at prediction of *intended* category (measure B), a metric that is closer to that used in Nearey and Assmann (2007). Just how well method 6 performs is further highlighted by the comparison against the normalization approach that arguably remains the most popular, Lobanov normalization. While Lobanov normalization provides the best fit to the intended vowel category, it is one of the worst models of listener responses. This lends additional emphasis to a point we have made elsewhere: just because a normalization procedure excels at reducing cross-talker formant variability this does *not* necessarily mean it is an adequate model of human perception (e.g., Barreda 2021; Persson, Barreda, and Jaeger 2025).

Method 6 also outperforms method 1 when the latter is applied trial-by-trial. We included this variant of method 1 (not considered in Nearey and Assmann 2007) as a more direct baseline to the single-trial methods 2, 3, and 6. The fact that the single-trial variant of method 1 performs so poorly that it is not even visible in Figure 4 (a) validates the issues we raised in section Section 4 about method 1: it would lead to highly unstable vowel perception, and listeners do not seem to use it.

Finally, another appealing feature of the STM is that it provides estimates of ψ_s 'for free', naturally tying together speech perception and the perception of speaker size. Here, we do not have access to listeners' estimates of ψ_s, and thus cannot directly compare the different methods against that ground truth. We can, however, compare how accurately the incremental P/BSTM methods estimate the ψ_s that would result from calculating ψ_s over the entire data from the talker (i.e., our best estimate of that ψ_s). As shown in Figure 4 (c), method 6 again performed best, and achieved the lowest RMS errors in $\hat{\psi}_s$ estimation compared to methods 2 and 3.

6. Conclusions and Future Directions

Almost two decades after its introduction, the PSTM remains the only published fully spelled-out model of incremental formant normalization and vowel perception. By making available the code for this model in the form of an R library, we hope to make the advantages of the PSTM and its extensions more accessible to other researchers. The PSTM's approach to normalization, uniform scaling, is rooted in principled considerations about the biology of auditory perception in the mammalian brain, validated by cross-species comparisons (reviewed in Barreda 2020). The log-mean method of uniform scaling has been consistently found to provide a better fit against listeners' perception than other influential methods, such as Lobanov normalization (Barreda 2021; Persson, Barreda, and Jaeger 2025; Richter et al. 2017). Here, we show that the predictive power of the PSTM improves further, and in non-subtle ways, once the *incremental* inferences listeners need to make are adequately modeled.

Our bootstrap simulations confirm that method 6—which assumes that listeners consider both intrinsic and extrinsic information in guiding their prior expectations about ψ_s—best explains listeners' behavior in the data from James M. Hillenbrand et al. (1995). The new BSTM, a more fully Bayesian extension of the PSTM, performed similarly to the original PSTM—both in terms of the ability to capture listeners' perception of vowels, and in terms of estimating the uniform scaling parameter ψ_s. However, the BSTM has the advantage of providing the full posterior distribution over ψ_s and the vowel category, allowing ideal information integration across multiple observations from the same talker. This is a critical feature for a model of speech perception as it will allow for the investigation of process of adaptation to speakers across increasing amounts of information.

In future work we plan to expand evaluation of the BSTM in three ways. First, here we evaluated 'single-shot' implementations of methods 2, 3, and 6. That is, we did not incrementally accumulate information about ψ_s across trials. For data like those in James M. Hillenbrand et al. (1995), where listeners rarely hear speech from the same talker for multiple trials in a row, this is likely an acceptable simplifying assumption. However, the single-shot implementations considered here might well underestimate listeners' ability to integrate information across observations from the same talker. In future work, we thus plan to extend these methods to incrementally update the prior based on observations from the same talker (see discussion in Xie, Jaeger, and Kurumada 2023). The Bayesian approach inherent in the P/BSTM is ideally suited for this purpose.

Second, the BSTM can naturally integrate the perception of indexical speaker characteristics such as age (Barreda and Assmann 2018), height (Barreda 2017), or gender (Barreda and Assmann 2021), via their shared reliance on ψ_s (see also Kleinschmidt, Weatherholtz, and Jaeger 2018). For example, consider the perception of some speaker characteristic θ. To simultaneously estimate vowel category, ψ_s, *and* this new characteristic, we would add it to our model as in Equation 19. This requires adding θ to the likelihood and considering joint prior of θ, ψ_s and vowel category. Such an approach would allow for a direct empirical comparison of different models regarding the perceptual integration of speech and social perception with relatively modest extensions of the models described here.

$$P\left(\mathrm{v}, \psi_s, \theta | \vec{G}\right) \propto P\left(\vec{G} | \mathrm{v}, \psi_s, \theta\right) \cdot P(\mathrm{v}, \psi_s, \theta) \qquad (19)$$

Third and finally, the general framework for incremental normalization presented here can be extended to a wide range of other normalization methods. This will allow comparisons between different approaches to incremental formant normalization and beyond. In particular, future work can investigate how incremental normalization can be integrated into modern ASR architectures, to improve the suitability of these architectures as a model of human speech perception. This will also researchers to move beyond normalization of formant point estimates (e.g., F1-F3 at the start and end of the vowel), addressing longstanding concerns that human listeners use a much broader sets of acoustic features when identifying vowels Barreda and Nearey (2012).

7. References

Assmann, Peter F, and William F Katz. 2000. “Time-Varying Spectral Change in the Vowels of Children and Adults.” *The Journal of the Acoustical Society of America* 108 (4): 1856–66.

Barreda, Santiago. 2012. “Vowel Normalization and the Perception of Speaker Changes: An Exploration of the Contextual Tuning Hypothesis.” *The Journal of the Acoustical Society of America* 132 (5): 3453–64.

———. 2017. “An Investigation of the Systematic Use of Spectral Information in the Determination of Apparent-Talker Height.” *The Journal of the Acoustical Society of America* 141 (6): 4781–92.

———. 2020. “Vowel Normalization as Perceptual Constancy.” *Language* 96 (2).

———. 2021. “Perceptual Validation of Vowel Normalization Methods for Variationist Research.” *Language Variation and Change*, 1–27.

Barreda, Santiago, and Peter F Assmann. 2021. “Perception of Gender in Children’s Voices.” *The Journal of the Acoustical Society of America* 150 (5): 3949–63.

Barreda, Santiago, and Peter F. Assmann. 2018. “Modeling the Perception of Children’s Age from Speech Acoustics.” *The Journal of the Acoustical Society of America* 143 (5): EL361. https://doi.org/10.1121/1.5037614.

Barreda, Santiago, and Terrance M Nearey. 2012. “The Direct and Indirect Roles of Fundamental Frequency in Vowel Perception.” *The Journal of the Acoustical Society of America* 131 (1): 466–77.

———. 2018. “A Regression Approach to Vowel Normalization for Missing and Unbalanced Data.” *The Journal of the Acoustical Society of America* 144 (1): 500–520.

Charlton, Benjamin D, William AH Ellis, Rebecca Larkin, and W Tecumseh Fitch. 2012. “Perception of Size-Related Formant Information in Male Koalas (Phascolarctos Cinereus).” *Animal Cognition* 15: 999–1006.

Chernyak, Bronya R., Ann R. Bradlow, Joseph Keshet, and Matthew Goldrick. 2024. “A Perceptual Similarity Space for Speech Based on Self-Supervised Speech Representations.” *JASA* 155 (6): 3915–29.

Hillenbrand, James M, Michael J Clark, and Robert A Houde. 2000. “Some Effects of Duration on Vowel Recognition.” *The Journal of the Acoustical Society of America* 108 (6): 3013–22.

Hillenbrand, James M., Laura A. Getty, Michael J. Clark, and Kevin Wheeler. 1995. “Acoustic Characteristics of American English Vowels.” *The Journal of the Acoustical Society of America* 97 (5): 3099. https://doi.org/10.1121/1.413041.

Hillenbrand, James M, and Terrance M Nearey. 1999. “Identification of Resynthesized/hVd/Utterances: Effects of Formant Contour.” *The Journal of the Acoustical Society of America* 105 (6): 3509–23.

Jin, Zhengyang, Yuhao Zhu, and T. Florian Jaeger. 2025. “Latent Speech Representations Learned Through Self-Supervised Learning Predict Listeners’ Generalization of Adaptation Across Talkers.”

Johnson, K. 1997. “Speech Perception Without Speaker Normalization.” In *Talker Variability in Speech Processing*, edited by K Johnson and J W Mullennix, 145–46. San Diego: Academic Press.

Kim, Seung-Eun, Bronya R Chernyak, Olga Seleznova, Joseph Keshet, Matthew Goldrick, and Ann R Bradlow. 2024. “Automatic Recognition of Second Language Speech-in-Noise.” *JASA Express Letters* 4 (2).

Kleinschmidt, Dave F, and T Florian Jaeger. 2015. “Robust Speech Perception: Recognize the Familiar, Generalize to the Similar, and Adapt to the Novel.” *Psychological Review* 122 (2): 148.

Kleinschmidt, Dave F, Kodi Weatherholtz, and T Florian Jaeger. 2018. “Sociolinguistic Perception as Inference Under Uncertainty.” *Topics in Cognitive Science* 10 (4): 818–34.

Lobanov, Boris M. 1971. “Classification of Russian Vowels Spoken by Different Speakers.” *Journal of the Acoustical Society of America* 49 (2B): 606–8.

Luce, Paul A, and David B Pisoni. 1998. “Recognizing Spoken Words: The Neighborhood Activation Model.” *Ear and Hearing* 19 (1): 1–36.

Magnuson, James S, and Howard C Nusbaum. 2007. "Acoustic Differences, Listener Expectations, and the Perceptual Accommodation of Talker Variability." *Journal of Experimental Psychology: Human Perception and Performance* 33 (2): 391.

Massaro, Dominic W, and Daniel Friedman. 1990. "Models of Integration Given Multiple Sources of Information." *Psychological Review* 97 (2): 225.

Nearey, T M. 1978. *Phonetic Feature Systems for Vowels*. 177. Indiana University Linguistics Club.

Nearey, T M, and Peter F Assmann. 2007. "Probabilistic 'Sliding-Template' Models for Indirect Vowel Normalization." *Experimental Approaches to Phonology* 246.

Norris, Dennis, and James M McQueen. 2008. "Shortlist b: A Bayesian Model of Continuous Speech Recognition." *Psychological Review* 115 (2): 357.

Persson, Anna, Santiago Barreda, and T Florian Jaeger. 2025. "Comparing Accounts of Formant Normalization Against US English Listeners' Vowel Perception." *Journal of the Acoustical Society of America* 157: 1458–82.

Peterson, Gordon E., and Harold L. Barney. 1952. "Control Methods Used in a Study of the Vowels." *The Journal of the Acoustical Society of America* 24 (2): 175–84. https://doi.org/10.1121/1.1906875.

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

Reby, David, Karen McComb, Bruno Cargnelutti, Chris Darwin, W Tecumseh Fitch, and Tim Clutton-Brock. 2005. "Red Deer Stags Use Formants as Assessment Cues During Intrasexual Agonistic Interactions." *Proceedings of the Royal Society B: Biological Sciences* 272 (1566): 941–47.

Richter, Caitlin, Naomi H. Feldman, Harini Salgado, and Aren Jansen. 2017. "Evaluating Low-Level Speech Features Against Human Perceptual Data." *Transactions of the Association for Computational Linguistics* 5: 425–40. https://doi.org/10.1162/tacl_a_00071.

RStudio Team. 2024. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Sumner, Meghan. 2011. "The Role of Variation in the Perception of Accented Speech." *Cognition* 119 (1): 131–36.

Taylor, AM, David Reby, and Karen McComb. 2010. "Size Communication in Domestic Dog, Canis Familiaris, Growls." *Animal Behaviour* 79 (1): 205–10.

Xie, Xin, T Florian Jaeger, and Chigusa Kurumada. 2023. "What We Do (Not) Know about the Mechanisms Underlying Adaptive Speech Perception: A Computational Framework and Review." *Cortex* 166: 377–424.