A regression approach to vowel normalization for missing and unbalanced data

Santiago Barreda[a)]
Department of Linguistics, University of California, Davis, Davis, California 95616, USA

Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton T6G 2E7, Canada

(Received 26 September 2017; revised 4 July 2018; accepted 6 July 2018; published online 31 July 2018)

Researchers investigating the vowel systems of languages or dialects frequently employ normalization methods to minimize between-speaker variability in formant patterns while preserving between-phoneme separation and (socio-)dialectal variation. Here two methods are considered: log-mean and Lobanov normalization. Although both of these methods express formants in a speaker-dependent space, the methods differ in their complexity and in their implied models of human vowel-perception. Typical implementations of these methods rely on balanced data across speakers so that researchers may have to reduce the data available in the analyses in missing-data situations. Here, an alternative method is proposed for the normalization of vowels using the log-mean method in a linear-regression framework. The performance of the traditional approaches to log-mean and Lobanov normalization against the regression approach to the log-mean method using naturalistic, simulated vowel-data was investigated. The results indicate that the Lobanov method likely removes legitimate linguistic variation from vowel data and often provides very noisy estimates of the actual vowel quality associated with individual tokens. The authors further argue that the Lobanov method is too complex to represent a plausible model of human vowel perception, and so is unlikely to provide results that reflect the true perceptual organization of linguistic data. © *2018 Acoustical Society of America*. https://doi.org/10.1121/1.5047742

[TCB] Pages: 500–520

I. INTRODUCTION

The frequencies of the lowest two or three formants are widely viewed as the primary determinants of perceived vowel quality (for a review, see Kiefte *et al.*, 2012). These formants, and the plots representing them, have for decades been regarded as indispensable tools in studying the vowel systems of dialects and languages. Although the formant frequencies of a vowel sound contain crucial information about vowel systems, absolute formant frequencies are also strongly influenced by physiological differences between speakers. As a result, vowel formant-frequencies (FFs) cannot be unambiguously associated with a given perceived vowel quality.

Researchers investigating vowel systems are primarily interested in differences in formant patterns associated with variation in perceived vowel-quality, rather than phonetically-irrelevant variation (e.g., variation associated with typical physiological differences between speakers). As a result, researchers will often use vowel normalization methods that seek to remove phonetically-irrelevant variation in vowel formant data, so that the normalized formant patterns will more closely reflect perceived vowel-quality. For example, Hindle (1978) states that the ideal normalization method "will minimize formant differences between individuals due to inherent physiological factors, but will preserve distinctions that correspond to perceptibly different vowels" (p. 167). This conceptualization of a desirable normalization suggests that the ideal method provides a close approximation to the perception of human listeners: Two vowels should lie close together in the normalized vowel space if—and only if—they have a similar perceived vowel quality.

It is important to keep in mind that normalization methods are meant to reflect differences in perceived vowel quality and are not simply scatter-reduction algorithms. A normalization method that systematically disagrees with the judgments of human listeners will not be of much use for linguistic research. As noted by Disner (1980), "it is not enough that [a normalization method] reduce the variance while maintaining the separation in any given data set; caution should be exercised to ensure that the trends which remain in the normalized data are truly linguistic trends and not artifacts of the normalization technique itself. It cannot be overemphasized that the output of any adequate normalization procedure must be a correct representation of linguistic fact" (p. 253). From the perspective of the researcher investigating variation in vowel quality as a function of dialectal or sociological differences between speakers, the only relevant "linguistic fact" associated with a formant pattern is the perceived vowel quality associated with it. For example, variation in $F2$ is of interest to the linguist primarily to the extent that it signals differences in a perceptual property such as fronting.

The above suggests that the "ground truth" against which the output of normalization methods should be judged is human vowel-perception. As a result, the desired behavior

[a)]Electronic mail: sbarreda@ucdavis.edu of normalization algorithms should be considered in terms of the behavior of human listeners. This position runs counter to the treatment of vowel normalization methods primarily as methodological tools without a necessary theoretical component. For example, in discussing the comparison of normalization methods, several researchers (Fabricius *et al.*, 2009; Flynn and Foulkes, 2011; Thomas and Kendall, 2007) have suggested that these methods have four distinct motivations (Fabricius *et al.*, 2009, p. 415):

(a) to eliminate variation caused by physiological differences among speakers;
(b) to preserve sociolinguistic/dialectal/cross-linguistic differences in vowel quality;
(c) to preserve phonological distinctions among vowels; and
(d) to model the cognitive processes that allow human listeners to normalize vowels uttered by different speakers.

However, in a discussion of the relative merits of different normalization methods, Fabricius *et al.* (2009) state that "we focus on goals (a) and (b); moreover, it is the balance between these two that we see as crucial […] we do not enter into a discussion of point (d), because it is not relevant for our purposes in this article" (p. 415). Similarly, Thomas and Kendall (2007) state that "For sociolinguists and dialectologists, however, goal [d] is the least important of the four. […] it is the first two goals that matter."

However, note that goals (a) and (b) above result in potentially opposing goals for normalization methods. For example, if some speaker has a physiological idiosyncrasy that results in their produced vowels sounding more open to listeners, a method cannot remove the physiologically-motivated variation in the formant patterns while also maintaining the differences in the perceived vowel-qualities. In these situations, it seems clear that researchers interested in patterns in perceived vowel-quality must favor goal (b) at the expense of goal (a). As a result, normalization methods should not be approached as models of variation in formant patterns between speakers, but rather as models of the human perceptual accommodation to this variation.

A. Normalization methods as models of vowel perception

We may consider a 3-vector of formant frequencies f specifying $F1$, $F2$, and $F3$ for a vowel and corresponding to a specific location in the 3-formant vowel space. Because of phonetically-irrelevant between-speaker variation in the formant space, the position denoted by f cannot be unambiguously associated with a specific perceived vowel-quality. However, human vowel perception involves a process whereby listeners are easily able to associate a given formant pattern with a specific perceived vowel-quality with a high degree of consistency and accuracy. This can be thought of as a process which takes in the formant pattern f and assigns this to a specific location on a perceptual space that is unambiguously related to perceived vowel quality (the human perceptual space). The perceived vowel-quality associated with a specific f would then be determined by its position in the perceptual space rather than by its position in the formant space. A single set of formant values are associated with different vowel qualities to the extent that they have different locations in the perceptual space, and two different formant patterns (f and f') can correspond to the same perceived vowel quality to the extent that they correspond to similar locations in the perceptual space.

Vowel normalization methods also perform mappings of formant patterns f to normalized (rather than perceptual) spaces. Normalization methods must include operations that have analogues (in process or outcome) to those of human vowel perception. If not, the normalized space is unlikely to match the human perceptual space, which will lead to errors in the apparent vowel-quality associated with any given vowel token. Such errors are problematic because researchers rely on the locations of individual tokens in the normalized space being associated with specific perceived vowel-qualities. For example, if a speaker produces a token with a higher normalized $F2$ than the category mean, a researcher may say this speaker produces a "fronted" variant of the vowel. In this case, a specific location in the normalized space is associated with a specific linguistic fact (fronting). If a group of speakers produce this phoneme in roughly the same location in the normalized space, a researcher may conclude that the group of speakers is all fronting their vowel. However, it is important to keep in mind that this conclusion is only valid to the extent that human listeners agree that these vowels are perceptually fronted. If listeners considered that only about half of the vowels sounded fronted, it would not be appropriate to claim that all speakers in the group front this vowel.

Given that different normalization methods apply different transformation to the formant patterns, they suggest differing perceptual organization for the same set of tokens. However, if we accept that the "true" vowel quality is the one determined by human listeners, then to the extent that different normalization methods suggest different patterns of vowel-quality variation they cannot all be right. As a result, the theoretical assumptions embedded within the structure of a normalization method have serious practical consequences for the research carried out by linguists using these methods. By selecting a normalization method, researchers are committing to a specific mapping of formant frequencies to perceived vowel-qualities; and hence, we would argue, they are implicitly adopting a class of perceptual models consistent with that mapping.

1. Overnormalization

As a first approximation we can imagine that since normalization methods are a response to between-speaker variation, the ideal normalization method is the one which minimizes differences between the speakers of a dialect in the normalized space by minimizing within-category variation. However, there will always be within-category variability from token to token when produced by a single speaker. From the researcher's (and listener's) perspective, within-category variability represents the inseparable conflation of unintentional repetition error and intentional sub-phonemic linguistic variation. As a result, "error" in the perceptual space represents subphonemic variation of potential linguistic import. Thus, in some cases a normalization method could remove "too much" variation, e.g., by removing variability that was not phonetically-irrelevant. When this occurs, the method has "overnormalized" the data in question.

In the extreme case of overnormalization we can imagine a method that identifies the category of a vowel and then assigns it the mean normalized formant values for that category. Such a method would maximize the similarity of vowel spaces between speakers and minimize within-category scatter in the normalized space, however, no researcher would seriously consider using such a method. We may reject the above method because it is not a perceptually-plausible method of normalization and it will substantially overnormalize vowel data. The above method will not reflect any subphonemic variation between vowel tokens, making them artificially more similar than they ought to be given the vowel qualities associated with the sounds.

Assessing the degree of overnormalization can only be done with reference to the "correct" (i.e., phonetically irrelevant) variation to remove, which is ultimately determined by human vowel-perception. As a result, although exact processes of human vowel perception are not known, discussion of the behavior of normalization methods must involve a consideration of the theoretical and empirical support for the outcomes of the transformations carried out by these methods. A normalization algorithm that maximizes some performance metric but does not reflect the true structure of the perceptual space may seriously impede the ability of researchers to make reliable inferences based on normalized vowel data.

In this investigation we will focus on two normalization methods: single parameter log-mean normalization (Nearey, 1978) and Lobanov (1971) normalization. Both of these methods represent vowels within a speaker-dependent space but differ substantially in terms of their complexity. In the remainder of this section we will introduce each method, and consider the implications of each method for theories of vowel perception.

B. Single parameter log-mean normalization

Single parameter log-mean normalization (henceforth log-mean normalization) is motivated by the constant ratio hypothesis of vowel perception, one of the earliest accounts of the relation between the vowel systems of speakers of the same dialect, dating back to at least Lloyd (1890) [see Miller (1989) for additional references]. In its broadest form, it states that within a dialect the formant pattern produced by a speaker for a given vowel is relatable to the formant pattern produced for that vowel by any other speaker of the dialect by a single ratio or scale factor. As a result, the formants produced by a speaker s for a vowel v and formant number k (F_{kvs}) are relatable to a dialect-specific reference formant-pattern (F^*_{kv}) by a single speaker-dependent parameter (ρ_s) as in Eq. (1),

$$\mathrm{F}_{kvs} = \mathrm{F}^*_{kv} \times \rho_s \,. \tag{1}$$

In its log-formant version [the "constant log interval" hypothesis (Nearey, 1978)] presented in Eq. (2), each component of Eq. (1) is replaced by the logarithm of the corresponding term, and ψ_s is now a speaker-dependent displacement term rather than a multiplicative constant,[1]

$$\mathrm{G}_{vks} = \mathrm{N}^*_{\mathrm{vk}} + \psi_s \,. \tag{2}$$

Log-mean normalization attempts to estimate the speaker-dependent parameter (ψ_s) that relates the formants of different speakers of a dialect, so that it can be removed from the observed log-transformed formant frequency (G_{vks}) to reveal the dialectal reference pattern (N^*_{vk}). The first step is the estimation of the speaker parameter ψ_s.We adopt the convention that the reference-pattern terms (G^*_{vk}) are constrained to sum to zero across all vowels and formants. If this is done, the speaker parameter (ψ_{s}) represents the mean logarithmic formant frequency (the log of the geometric mean) produced by a speaker across all the vowels in their system. Although the exact value of the speaker parameter for any given speaker cannot be exactly known, it may be estimated for balanced complete data using the log-mean FF produced by that speaker (denoted using a bar over the variable, $\bar{\mathrm{G}}_{\mathrm{s}}$) for some set of observed vowels, using the formula provided in Eq. (3),

$$\bar{\mathrm{G}}_{\mathrm{s}} = \hat{\psi}_{\mathrm{s}} = \frac{1}{\mathrm{V}\cdot\mathrm{K}\cdot\mathrm{T}}\sum_{\mathrm{v}=1}^{\mathrm{V}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\sum_{\mathrm{t}=1}^{\mathrm{T}}\mathrm{G}_{vkst} \,, \tag{3}$$

where v is the vowel category, k is the formant number, s is an index for the speaker in question, and t is the token (i.e., repetition, or in recording studio jargon "take") number. The capitals V, K, and T are the number of vowel categories, formants, and tokens, respectively.

Formant frequencies are then normalized by subtracting the estimated speaker parameter ($\bar{\mathrm{G}}_{\mathrm{s}}$) from the logarithm of the observed formant frequencies[2] as in Eq. (4), which has effectively re-arranged the terms in Eq. (2),

$$\mathrm{N}_{vk} = \hat{\mathrm{N}}^*_{vk} = \mathrm{G}_{vks} - \bar{\mathrm{G}}_{\mathrm{s}} \,. \tag{4}$$

Note that the individual values of N_{vk} provide estimates of the dialectal reference pattern ($\mathrm{N}^*_{\mathrm{vk}}$) but are not expected to equal it exactly for any given observation due to repetition error.

1. Single-parameter log-mean normalization as a perceptual model

All other things being equal, proportional changes to the dimensions of each section of tube in multi-tube resonators (such as the vocal tract) will result in increases/decreases to their resonant frequencies according to a single, multiplicative scale factor (Markel and Gray, 1976). So, when speakers who differ primarily in vocal-tract length adopt similar articulatory gestures, they will tend to produce formant patterns that differ according to a single factor related to the difference in their vocal-tract lengths. On account of this, the speaker parameter ψ_s can be thought of as encoding information primarily related to speaker vocal-tract length, and the constant-ratio hypothesis can be thought of as modeling variation between speakers that is attributable primarily to differences in the vocal-tract lengths of different speakers.

It has been suggested that differences in vocal-tract geometry between speakers should result in deviations from constant ratios in formant patterns (Fant, 1966, 1975). However, several investigations have found that the influence of specific vocal tract geometry on formant patterns has been overstated and that the primary determinant of produced output is simply overall vocal-tract length (Goldstein, 1980; Nordström and Lindblom, 1975; Turner *et al.*, 2009). It has also been argued that different speakers adopt compensatory gestures to produce outputs that differ from the dialectal-reference pattern by a single factor, despite potentially different vocal-tract geometries. For example, Turner *et al.* (2009) report that the oral/pharyngeal cavity ratio varies continuously as a function of speaker height (and vocal-tract length). However, in an analysis of vowel formant patterns, the authors find that "once the measurement noise has been properly modeled, it is observed that the formant patterns of the vowel sounds do not vary systematically, either with the size or the sex of the speaker, despite the obvious non-uniformity in the growth of the anatomical cavities oral and pharyngeal" (p. 2375). Turner *et al.* (2009) conclude that a notational variant of constant-ratio hypothesis, which they call the constant formant-pattern model, is essentially correct for the data they analyzed, stating that "the anatomical distinction between the oral and pharyngeal divisions of the vocal tract is immaterial to the acoustic result of speech production. For a given vowel, the tongue constriction is simply positioned where it produces the appropriate ratio of front-cavity length to back-cavity length, independent of the location of the oral-pharyngeal junction" (p. 2379).

The above suggests that speakers attempt to produce outputs that differ primarily in ways consistent with the constant-ratio hypothesis. The constant-ratio hypothesis, or something very close to it, is also well-supported in perception. In particular, modification of formant patterns via vocoding that comports with the constant-ratio hypothesis (e.g., uniform scaling of the spectral envelope shape) generally preserves vowel quality and intelligibility (Assmann and Nearey, 2008). This fact has allowed uniform-scaling of speech sounds to be used in experimental research in order to simulate size differences between speakers (Assmann *et al.*, 2006; Barreda, 2017; Smith *et al.*, 2007). Furthermore, uniformly-scaled vocoding has been used for decades in the entertainment industry to manipulate the apparent size of voice actors (Lawson and Persons, 2004; Winer, 2012). Other factors such as *f*0 and the higher formants (*F*4 and above) can sometimes alter perception of lower-formant patterns in ways not explicable via the constant-ratio hypothesis, especially when the statistical relations among all these measures are drastically altered from typical values (Nearey, 1989; Barreda and Nearey, 2012). However, we know of no experiments showing that the relations implied by the constant-ratio hypothesis can be substantially altered in a vowel system without "damaging" phonetic identity.

In addition, the estimation of the single speaker-parameter used in log-mean normalization likely represents a tractable problem for the listener. As noted by Nearey (1978), given that a dialect only has a limited number of vowel phonemes (and corresponding acceptable formant-patterns), the speaker-dependent scaling parameter (ψ_s or an analogous value) can be estimated from a single vowel token whose category is known. At least two pattern-recognition studies have shown that given dialect constraints on formant patterns, the accurate estimation of ψ_s is also feasible from a single token even when the vowel quality of that token is unknown (Nearey and Assmann, 2007; Turner *et al.*, 2009). This is an important consideration given that listeners are able to accurately and consistently associate individual vowel tokens with a perceived vowel quality, even when presented with isolated vowels or CVC words from multiple speakers (Hillenbrand *et al.*, 1995; Peterson and Barney, 1952).

Taken together, the above suggests that speakers attempt to produce formant patterns that vary within-category according to a single multiplicative parameter, and that vowels that vary in such a manner are typically judged to have "the same" vowel quality according to listeners. Further, the accurate estimation of this parameter is feasible even from very limited numbers of tokens from a speaker, suggesting that listeners could reasonably be expected to estimate such a parameter in speech perception. As a result, log-mean normalization has a good deal of empirical and theoretical support as a plausible model of human vowel perception.

C. Lobanov normalization

As an alternative to log-mean normalization, we will consider the formant-wise standardization method proposed by Lobanov (1971). To normalize vowel data using the Lobanov method, the mean and standard deviation [referred to by Lobanov as the root-mean square (RMS)] are calculated (in Hertz) for each of formant (*k*) and speaker (*s*), across all vowels and tokens produced by that speaker [Eqs. (5) and (6)]. Then, formant frequencies are standardized using these estimated speaker parameters [Eq. (7)],

$$\hat{\mu}_{\rm ks} = \frac{1}{\rm V \cdot T}\sum_{\rm v=1}^{\rm V}\sum_{\rm t=1}^{\rm T} {\rm F}_{vkst}\,, \tag{5}$$

$$\hat{\sigma}_{\rm ks} = \sqrt{\frac{1}{\rm V \cdot T}\sum_{\rm v=1}^{\rm V}\sum_{\rm t=1}^{\rm T} ({\rm F}_{vkst} - \hat{\mu}_{\rm ks})^2}\,, \tag{6}$$

$$\mathrm{z}^*_{vk} = \frac{{\rm F}_{vks} - \hat{\mu}_{ks}}{\hat{\sigma}_{ks}}. \tag{7}$$

According to Lobanov, $\hat{\mu}_{\rm ks}$ and $\hat{\sigma}_{\rm ks}$ represent the "personal information" imparted onto the formant pattern by the speaker. So, after standardizing formant patterns with respect to these parameters, Lobanov normalization seeks to produce normalized formant patterns that are "[statistically] independent of [the] exact parallel shift and exact linear compression or expansion" (Lobanov 1971, p. 607) associated with each speaker. Thus, Lobanov normalization was explicitly designed to remove speaker-specific variation in productions with no regard for the potential perceptual consequences of the transformations.

1. Lobanov normalization methods as a perceptual model

Lobanov normalization was proposed as a tool to minimize mistakes in the automatic classification performance of Russian vowels when produced by more than one speaker. As a result, the objective of Lobanov normalization is to maximize the performance of statistical classifiers and not to reflect or preserve any linguistic fact (i.e., perceived vowel-quality). To our knowledge, no one has proposed a model of human vowel perception that is compatible with the assumptions of Lobanov normalization, though McMurray and Jongman (2011) have proposed a general theory of speech perception that is largely compatible with the assumptions of the Lobanov model. In addition, there are two main problems with considering Lobanov normalization as a model that is at least analogous to human vowel perception: the tractability of the problem and the normalization of vowel-space dispersion. Each of these issues has the potential to result in misalignments between the Lobanov normalized space and the human perceptual space, which can lead to incorrect conclusions drawn from normalized formant patterns.

Lobanov normalization requires the estimation of two parameters for every formant, meaning it would likely involve the estimation of at least six parameters for human vowel perception. It is not clear how so many independent parameters, in particular those related to formant dispersion, could be estimated from a single vowel with a high degree of accuracy. For example, consider two speakers of a vowel system with /a i u/ that have identical Lobanov parameters except for the standard deviation of $F2$. Assume also that the mean of $F2$ coincides with that of /a/ for both speakers. These speakers have identical positions for their /a/ in Hertz, but one speaker would have more peripheral /i/ and /u/. If a listener heard both speakers produce /a/ we would expect them to have the same vowel quality. If the listener then heard the same speakers produce /u/, would they also sound the same despite the fact that one speaker would have a higher $F2$? Lobanov normalization would control for the increased peripherality of the high vowels, suggesting that these should also sound "the same" to listeners. On what basis could the listener know that this apparent difference in fronting is actually a result of a difference in $F2$ dispersion without knowing that one speaker also has a more peripheral /i/? The apparent intractability of the estimation of Lobanov parameters from limited numbers of observations suggests that human listeners are likely doing something fundamentally different in vowel perception, either because human listeners control for fewer parameters in perception, or because the estimation of parameters is constrained in important, but as yet unspecified, ways.

The second problem is that although physiological differences will result in differences in vowel-space dispersion between speakers, articulatory variation not related to physiological differences can also affect vowel-space dispersion. Furthermore, it does not appear that vowel space dispersion always represents "phonetically-irrelevant" information that should be normalized away in all situations. For example, vowel space area and related measures of degree of deviation from a central neutral formant are widely viewed as useful for understanding aspects of speech intelligibility. Speakers showing "clear speech" formant patterns typically have more expanded vowel spaces than those in more relaxed or conversational speech (Ferguson and Kewley-Port, 2007). Foreigner- and infant-directed speech has also been shown to sometimes show formant space area expansion (Uther *et al.*, 2007). There also appears to be a relationship between vowel-space dispersion and the perception of indexical characteristics related to speaker gender and sexual orientation (Heffernan, 2010; Munson, 2007). Finally, it has been suggested that some aspects of apparent non-uniform scaling in formant averages between male and female speakers may be due to the fact that in some socio-cultural contexts, female speakers may tend to produce "clearer," more dispersed speech (Goldstein, 1980; Diehl *et al.*, 1996).

The perceptual salience of vowel-space dispersion suggests that listeners may have some expectations regarding the amount of vowel space dispersion to expect for a given speaker, so that deviations from this may be phonetically meaningful. For example, we can imagine two speakers who have identical physiological characteristics and who, all other things being equal would produce identical acoustic outputs (e.g., identical twins). Consider a situation where these speakers differed noticeably in their vowel-space dispersion so that one speaker produced a larger vowel space than the second. Given the results outlined above, it would be expected that the enlarged vowel space could result in phonetically-meaningful differences: the larger space will result in clearer speech, and may be associated with specific indexical speaker-characteristics related to gender or sexual orientation (among other things). However, Lobanov normalization will equalize the vowel spaces of these two speakers, completely removing any differences between the productions of the two speakers in the normalized space.

Log-mean normalization does not include speaker-dependent parameters for vowel space dispersion. However, the constant ratio hypothesis on which it is based entails a tight relationship between the speaker scale-factor and the dispersion of the vowel formants when frequencies are measured in Hertz. For example, consider the area of a triangle outlined by the /i a u/ produced by one speaker with a scaling factor of ρ_s. We can imagine that this is an equilateral triangle whose sides equal d_s, and whose area ($A = (\sqrt{3}/4)d_s^2$) provides a measure of formant dispersion. If another speaker with a shorter vocal tract (a higher scale factor $\rho_{s'}$) produces the same vowels, each of these vowels will have higher formant frequencies by a factor of $\Delta_\rho = \rho_{s'}/\rho_s$, where $\Delta_\rho > 1$. This has the effect of increasing the distances between the vertices of the triangle by a factor of Δ_ρ, which results in an increase of the area of the vowel space proportional to the square of the same factor (i.e., $A' = (\sqrt{3}/4)d_s^2\Delta_\rho^2$). As a result, if a speaker has a standard deviation for a formant frequency that is 20% higher overall than another, their vowel space area in a two-dimensional formant space will be approximately 44% larger when measured in Hertz, all other things being equal. Thus, we may state that between-speaker differences in average formant frequency due to vocal-tract length are necessarily related to expansions in vowel dispersion when formant measurements are measured in Hertz.

In contrast, when considering variation between speakers in a log space the multiplicative speaker parameter becomes an additive constant [Eq. (2)]. Unlike multiplication, addition will not result in expansions or contraction of the vowel spaces of different speakers. Instead, variation in formant patterns of the kind associated with vocal-tract length differences will only result in translations of the vowel spaces of different speakers along a given dimension. As a result, considering formant patterns in a log-space will automatically control for differences in dispersion in the linear Hertz space that can be directly attributable to changes in vocal-tract length. Importantly, this control is constrained by the relationship between resonator size and resonance frequencies and does not involve the estimation of independent parameters. In the absence of a better understanding of the relationship between vowel-space dispersion and perceived vowel-quality, normalization methods that use independent parameters to equalize vowel-space dispersion between speakers may leave themselves particularly susceptible to overnormalization by making the vowel spaces of different speakers "too equal" with respect to dispersion.

In summary, there is little to no theoretical or empirical support for Lobanov normalization as a plausible model of human speech perception. In particular, Lobanov normalization may represent an unrealistically-complicated model in light of rapid listener adaptation to between-speaker differences. Lobanov normalization may also result in removal of legitimate phonetic variation between tokens by independently normalizing for vowel-space dispersion along each formant. As a result, Lobanov normalization may not result in an accurate representation of the linguistic variation in the vowel productions of different speakers.

II. NORMALIZING MISSING OR UNBALANCED DATA USING A REGRESSION FRAMEWORK

It has been noted that normalization may not be appropriate for cross-language, or cross-dialectal, comparisons because of the differing vocalic inventories across languages (Adank *et al.*, 2004; Disner, 1980). In fact, the problems associated with cross-language normalization may also arise when listeners are compared within-dialect for unbalanced subsamples of vowels and tokens for each speaker. Estimating speaker parameters using different vowel inventories for different subsets of speakers can result in artificial shifts between the normalized vowel-spaces of these speakers. These shifts can cause specific locations in the normalized space to be associated with different perceived vowel qualities for different speakers, a situation which runs contrary to the purpose of normalization. Figure 1 presents an example of how normalization using different vowel subsets may incorrectly suggest differences in vowel quality between speakers.

An outline of how unbalanced data lead to artifactual shifts in normalized vowel spaces will be given with reference to the operations of log-mean normalization, though the general reasoning will apply to most of the speaker-dependent parameters used by different normalization methods. A more complete account of what is being estimated in Eq. (3) is presented in Eq. (8),

$$\bar{G}_s = \hat{\psi}_s^{\{V\}} = \psi_s + C^{\{V\}} + \eta_s = \frac{1}{V \cdot K \cdot T}\sum_{v=1}^{V}\sum_{k=1}^{K}\sum_{t=1}^{T} G_{vkst}\,, \tag{8}$$

where ψ_s is the underlying scale factor for the speaker s, $\{V\}$ is the set of vowels used to perform the estimation, $\psi_{s\{V\}}$ is the apparent scale-factor when estimated based on vowel inventory $\{V\}$, $C^{\{V\}}$ is an inventory displacement factor that depends on the average formant-patterns of the exact vowel inventory$\{V\}$, and η_s is an estimation error for that speaker. Thus, in practice, $\bar{G}_s$ could be viewed as an

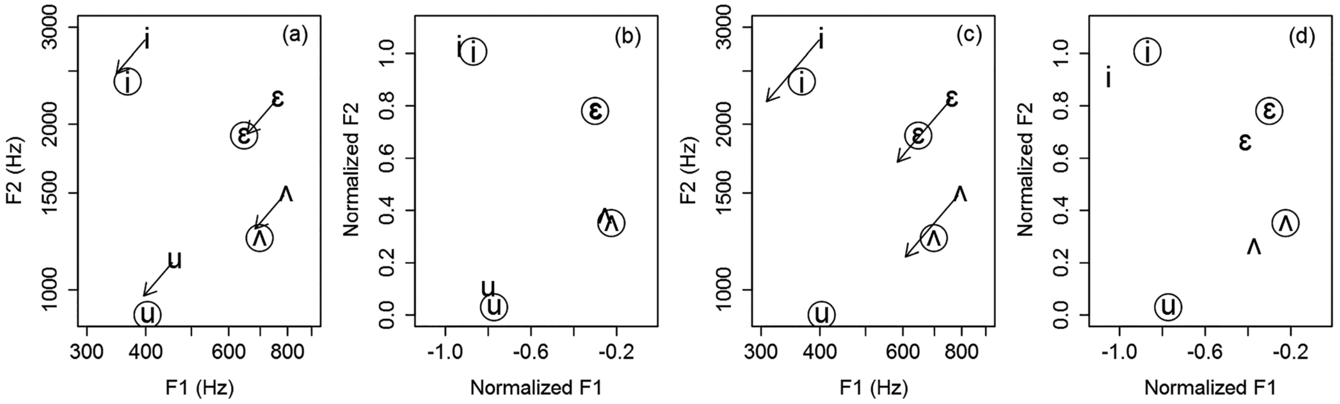

FIG. 1. Vowels produced by two speakers with different overall log-mean FFs. Circled symbols represent the speaker with the lower FFs. (a) Arrows indicate the change in the high FF vowels relative to the low FF vowels after log-mean normalization with full inventories for both speakers. (b) Systems have been aligned with respect to their overall log-mean FFs, resulting in tighter clustering of individual categories. (c) /u/ has been omitted for the high FF voice, and $\bar{G}_s$ has been calculated for each speaker using Eq. (3) and the available vowels. This results in a higher centroid for the high FF voice and an apparent larger difference between the two speakers. (d) When the systems in (c) are brought into alignment based on the biased $\bar{G}_s$ estimate there will be artefactual differences between the normalized vowel spaces of the two speakers. Compare the large discrepancies in panel (d) with the small ones in (b) based on the same underlying vowel systems.

estimate of the underlying ψ_s plus an inventory-specific constant $C^{\{V\}}$.

Under a fixed inventory $\{V\}$ for all speakers of interest $\{S\}$, between-subject differences in $\bar{G}_s$ as calculated by Eq. (3) will be due to $\hat{\psi}_s^{\{V\}}$ and estimation error, but not $C^{\{V\}}$ since this will be the same for all speakers in $\{S\}$. Normalizing vowels by $\bar{G}_s$ will remove variation due to speaker the single speaker parameter, as well as a fixed inventory constant $C^{\{V\}}$. This is not problematic provided the vowel inventory, and $C^{\{V\}}$, is the same for all speakers. However, in the event that different speakers have different vowel inventories, estimates of $\bar{G}_s$ will also include what will be referred to as an "inventory bias." More explicitly,

$$\begin{aligned}\bar{G}_s &= \hat{\psi}_s^{\{V_s\}} = \psi_s + C^{\{V_s\}} + \eta_s \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{V_s}\sum_{v\in\{V_s\}}\frac{1}{T_{vks}}\sum_{t=1}^{T_{vks}} G_{vkst},\end{aligned} \qquad (9)$$

where $\{V_s\}$ is a speaker-dependent inventory of vowels. Thus, in the unbalanced case, differences in $\bar{G}_s$ for two speakers, s and s' ($\bar{G}_s - \bar{G}_{s'}$) will include the difference in their scale factors ($\psi_s - \psi_{s'}$), but also inventory-dependent differences ($C^{\{V_s\}} - C^{\{V_{s'}\}}$). As a result, single-parameter log-mean normalization will only eliminate speaker effects effectively when inventories $\{V_s\}$ and $\{V_{s'}\}$ are the same across speakers.[3]

The problems with cross-language normalization will also arise when applying log-mean normalization to a single dialect with missing or unbalanced data across speakers. This is reflected in Eq. (10), which updates Eq. (9) to include differing vowels, and differing formants for different vowels, across speakers,

$$\begin{aligned}\bar{G}_s &= \psi_s^{\{V_s K_{v_s}\}} = \psi_s + C^{\{V_s K_{v_s}\}} + \eta_s \\ &= \bar{G}_{s_a} = \frac{1}{K_{v_s}}\sum_{k=1}^{K}\frac{1}{V_s}\sum_{v\in\{V_s\}}\frac{1}{T_{vks}}\sum_{t=1}^{T_{vks}} G_{vkst}\,.\end{aligned} \qquad (10)$$

In the case of unbalanced number of tokens, the researcher can calculate the within-category average for each vowel, for each speaker, before proceeding to calculation of $\bar{G}_s$, effectively using a weighted-mean to estimate the parameter. This approach is formalized in Eq. (11), where the notation T_{vks} indicates there may be different numbers of measured tokens for each formant of each vowel, and $T_{vks} > 0$ for all v, k, s,

$$\bar{G}_s = \frac{1}{V}\sum_{v=1}^{V}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{T_{vks}}\sum_{t_{vks}=1}^{T_{vks}} G_{vkst}\,. \qquad (11)$$

A more difficult problem arises when no observations exist at all for one or more vowel categories for subsets of the speakers in the dataset. This may be due to loss of data, noisy recording conditions, or simply the inability to collect a full vocalic inventory for all speakers. A related problem arises in the case where one or more formants cannot be tracked for any given reason, a situation that is not uncommon. In either of these cases, estimates of $\bar{G}_s$ using Eq. (11) will result in different inventory biases for each speaker with missing data.

A. Approaches to missing data in normalization

The simplest approach to dealing with missing data is to proceed with normalization as usual and estimate speaker parameters based on unbalanced samples for different speakers with no controls. This approach will result in biases in the estimation of the different speaker parameters and will result in artificial asymmetries in normalized vowel-spaces across speakers (as in Fig. 1).

One way to avoid introducing such biases in the estimation of $\bar{G}_s$ is to perform a "complete case" analysis (Rubin and Little, 2002, Chap. 3) by removing data from some speakers to produce a balanced subsample across all speakers. There are two ways this could be done. First, any speaker who is missing any vowel category could be omitted from the analysis. This option is undesirable due to the loss of data; a method that allows this data to be used is preferable. Second, some vowel-categories could be omitted for all speakers when any speakers are missing those categories (or formants from those categories). Speaker parameters could be estimated based on the balanced subsample, and normalization could proceed as normal.

Although the second approach will not introduce artificial biases into the vowel spaces of different speakers, it will result in an increase in the variance of estimated speaker parameters. For example, the standard error of $\bar{G}_s$ will increase by a factor of $\sqrt{n}/\sqrt{n-m}$ where n is the total number of categories, and m is the number of missing categories. This fact has important consequences for the performance of normalization methods since increased error in the estimation of speaker parameters may lead to overnormalization, which will directly translate to error in the apparent vowel quality associated with different tokens. In effect, by decreasing the sample size to accommodate missing data we are trading off biases in the data of some speakers for increased noise in the normalized data of all speakers.

Either approach outlined above may be acceptable under certain circumstances, however some researchers (e.g., those working with speech corpora, endangered languages, or special populations) may not have full or balanced vowel inventories for any given speaker or may be missing large numbers of vowels from some speakers. It is difficult to know how widespread the issue of missing formant data is, since in some cases researchers may be omitting speakers or performing complete-case analyses in order to restore balance to their data. However, two-large publicly-available datasets (Assmann and Katz, 2000; Hillenbrand *et al.*, 1995) feature many instances of missing data, untrackable formants, and apparent pronunciation errors. Each of these datasets were recorded in ideal laboratory conditions suggesting that a lack of balance in formant datasets is likely a common occurrence. For example, the Hillenbrand *et al.* (1995) dataset includes productions of 12 vowels for 138 speakers for a total of 1668 vowel tokens. However, only 98 speakers have a complete set of $F1$, $F2$, and $F3$ observations for their "steady-state" vowel tokens, and only 3 vowel categories are represented fully across all speakers. In cases with unbalanced data, researchers may wish to apply normalization without omitting data or unnecessarily increasing the error in the estimates of their speaker parameters. To address this methodological limitation, a regression-based approach to normalization that is tolerant to certain missing-data situations will be described.

B. Speaker parameter estimation as a regression problem

We will outline the estimation of the log-mean parameter used in log-mean normalization using a regression framework. The possibility of extending this approach to some other normalization methods is discussed in Sec. III D.

If all vowel categories are present and balanced across all speakers, estimating $\bar{G}_s$ using the regression approach to be outlined will yield identical results to Eq. (3). However, if vowel categories are unbalanced, or more importantly, if some measurements are not present at all, this method will provide estimates of $\bar{G}_s$ for each speaker that are not strongly or predictably affected by the missing data in most cases. This alternative approach to log-mean normalization is an instance of a standard method of estimation in certain missing data situations from the applied statistics literature (Rubin and Little, 2002, p. 27). This approach is generally appropriate if the "missing data mechanism" is "ignorable," which means that:

(1a) The data are MAR (missing at random), "… the reason for the occurrence of missing data in Y does not depend on any Y values."

(2b) Distinctness: "the parameters of the missing data mechanism do not depend on the value of the analysis of variance parameters."

One could run into problems with assumption (2a), for example, if measurements of extremely high or low formant frequencies were missing because of *a priori* limits of a formant tracker. A problem involving (2b) might arise if certain vowels were for some reason more likely to be missing from certain speakers. This situation is quite likely to arise if for example certain speakers' voices are more difficult to analyze, resulting in more missing values. However, this is of lesser consequence, pertaining largely to the variance of the estimates, rather than bias. Rubin and Little further note that: "MAR [(2a)] is typically regarded as the more important condition here, in the sense that if the data are MAR but distinctness [(2b)] does not hold, inference about the ignorable likelihood is still valid from the frequency perspective, but not fully efficient" (Rubin and Little, 2002, p.120).

The relationship presented in Eq. (2) can be updated to include estimated values for the dialectal reference pattern (N^*_{vk}), the speaker displacement terms (ψ_s), and an error term, as in Eq. (12),

$$G_{vkst} = \bar{N}_{vk} + \bar{G}_s + \varepsilon_t\,. \tag{12}$$

When presented as in Eq. (12), it is clear that estimation of the reference formant-pattern ($\bar{N}_{vk}$) and the speaker-displacement terms ($\bar{G}_s$) may be treated as a linear regression problem. The model equation presented in Eq. (12) contains only categorical predictors, with a total of $S + (V \times K) - 1$ predictor coefficients: S $\bar{G}_s$ terms and $(V \times K) - 1$ $\bar{N}_{vk}$ terms for a dataset with productions from S speakers for V vowels specified on K formants. The key advantage to estimating $\bar{G}_s$ via regression is that the inclusion of the $\bar{N}_{vk}$ terms allows each observed formant frequency to provide information regarding $\bar{G}_s$ by controlling for expected vowel- and formant-specific deviations from this value [as in Eq. (12)]. As a result, one or more missing data points simply result in fewer total observations, instead of introducing inventory difference biases as would be the case when using Eq. (3).

The standard "pure error term" in Eq. (12) is assumed to be the same (in log frequency units) for every speaker, every vowel, and every formant, and the errors are assumed uncorrelated with those of any other speaker, vowel, or formant (Nearey, 1978). In the widely-used terminology of analysis of variance, this effectively treats speaker and formant-pattern related terms as "fixed effects." More complex situations involving more sources of variation (e.g., by treating subject as a random effect in a mixed-effect model) can be imagined, but it is not clear whether such models are necessary, or appropriate, for the normalization of vowel data.

C. Implementation

The use of ordinary least-squares regression to estimate $\bar{G}_s$ for a set of speakers will be outlined with reference to the statistical software R (R Core Team, 2017). However, this approach may be extended to any other method that allows for regression to be carried out. The method will be outlined with respect to a vowel system with V vowels, specified on K formant frequencies, representing data from S speakers. Equation (13) presents a regression equation that predicts the log-transformed FFs for a given vowel, formant and speaker (G_{vks}) as the sum of a speaker displacement term (S_s), and normalized vowel-formant effects (N_{vk}), which are estimates of the dialectal reference pattern N^*_{vk}.

$$G_{vks} = S_s + N_{vk} + \varepsilon. \tag{13}$$

To implement this analysis in R, it is first assumed that the data are available in a data frame object in a "long" format with only one log-formant measurement per row. Further, it is assumed that each row of the data frame has (at least) four columns, labeled: G for the single formant measurement, V indicating vowel, K indicating formant number, and S indicating the speaker. It is important that categorical variables coded with numbers (e.g., subject numbers, formant numbers) be represented as categorical variables ("factors" in R) so that they are not treated as continuous predictors in the model. Given this data, an additional variable (N) may be created to represent the N_{vk} terms, using the R command: `N=factor[interaction (V,K)]`. Equation (14) presents the R syntax for implementing

Eq. (13) using the vectors G (log-transformed FFs), and the factor variables S (Speaker), and N (Formant × Vowel),

$$\mathtt{M} = \mathtt{lm}(\mathtt{G} \sim 0 + \mathtt{S} + \mathtt{N}, \mathtt{contrasts} = \mathtt{list}(\mathtt{N} = \mathtt{contr.\ sum})). \tag{14}$$

The model specified in Eq. (14) specifically omits an intercept (by including a zero in the right-hand side of the equation), and uses treatment coding (sometimes referred to as "indicator" or "dummy" coding) for the *S* term, which means that each level is compared to a reference level. "Sum coding" (sometimes referred to as "deviation" coding) is used for the *N* term, which means that all levels of the term are constrained to sum to zero. Usually, all the levels of a factor cannot be estimated because of collinearity with the intercept term. However, in this case the intercept term is not of interest while estimating all *S* Speaker coefficients is a key goal of the analysis. By default, R uses the degree of freedom gained by omitting the intercept to estimate all levels of the first factor entered in the model equation. As a result, if the variable representing Speakers is in first position in the model equation, all *S* Speaker levels will be directly estimated (rather than being compared to a reference level). The estimates of the Speaker terms $\bar{\mathrm{G}}_s$ may be accessed conveniently by the R command: `S = dummy.coef(M)$S`, and those of the Vowel-Formant terms by `N = dummy.coef(M)$N`. Unfortunately, the documentation for the features of R exploited in the code above is somewhat scattered. See Venables and Ripley (2002, p.165 ff.) or Hastie and Chambers (1992) for more details, as well as the help files for the R functions *contr.sum*, *model.matrix*, and *dummy.coefs*.

D. Extending the regression approach to other normalization methods

The difficulty of extending the regression approach to the estimation of parameters for other normalization methods will depend on expected distribution of the necessary parameters and the ease with which these parameters can be estimated using a regression framework.

It is straightforward to adapt the approach presented in Sec. II B to formant-wise log-mean normalization, sometimes referred to as "formant intrinsic" log-mean normalization (Nearey, 1978). This method calculates an independent log-mean parameter for each formant for each speaker, and then carries out normalization by subtracting the formant-wise log-mean parameter from the log-transformed frequencies for each formant. Although this method may seem substantially different from single-parameter log-mean normalization, the high correlations between formant-mean parameters within-speaker (see Table II in the Appendix) mean that formant-wise and single-parameter log-mean normalization tend to yield similar results. However, the increased complexity of the formant-wise log-mean method comes at the expense of increased errors in parameter estimates (due to the reduction of data used to calculate each formant-wise parameter) and associated increases in sampling overnormalization (see Sec. III). Furthermore, the increased complexity does not appear to be theoretically or empirically justified since vowel productions do not seem to vary in ways that would require fully independent scaling estimates for each formant, and listeners do not seem to independently control for the scaling of each formant in perception. For these reasons, although we present an extension of the regression approach to this normalization method, we recommend the use of single-parameter log-mean normalization for linguistic research.

The easiest way to do this is to simplify the approach presented in Eq. (14) so that the equation for each formant is carried out independently. For each formant, *k*, we first select as *G* the data for that formant only (rather than stacking multiple formants) and eliminate the *k* subscript. An additional variable (N) may be created to represent the N_v terms, using the R command: `N = factor(V)`. Equation (15) presents the R syntax for implementing Eq. (14) using the vectors G (log-transformed FFs), and the factor variables S (Speaker) and N (Vowel),

$$\mathtt{M} = \mathtt{lm}(\mathtt{G} \sim 0 + \mathtt{S} + \mathtt{N}, \mathtt{contrasts} = \mathtt{list}(\mathtt{N} = \mathtt{contr.\ sum})). \tag{15}$$

The estimated speaker means formant *k* can be extracted as `S = dummy.coef(M)$S`, and can then be used to normalize FF by subtracting each subjects S coefficient from the log formant frequencies produced by that speaker for that formant.

The relationship between the Lobanov parameters, the dialectal template, and observed formant frequencies assumed by Lobanov normalization was presented in Eq. (7). These terms can be rearranged to resemble a regression form as in Eq. (16),

$$\mathrm{F}_{vks} = \hat{\sigma}_{ks}\mathrm{z}^*_{vk} + \hat{\mu}_{ks} + \varepsilon. \tag{16}$$

Equation (16) expresses the observed formant frequencies (F_{vks}) for a formant as a function of the formant standard deviation for a speaker ($\hat{\sigma}_{ks}$) multiplied by the dialectal target for the phoneme (z^*_{vk}), plus the mean value for the speaker, plus error. Equation (16) represents a non-linear regression problem which cannot be estimated using the ordinary least squares method outlined above. There are several ways that this regression could be approached, including nonlinear regression or alternating least squares, both requiring either constraints or penalty functions to ensure z^*_{vk} remains standardized in keeping with the spirit of Lobanov normalization. Alternatively, a multilevel Bayesian approach might be feasible (Gelman and Hill, 2006). The robustness of a regression approach to Lobanov normalization in the face of missing data will depend on the nature of the estimation and the missing-data pattern. Indeed, in light of difficulties with the identifiability and convergence of alternating least squares methods, which involve problems similar to the factoring of $\hat{\sigma}_{ks}$ and z^*_{vk} in Eq. (16), it may prove difficult to find a fully satisfactory solution (Uschmajew, 2012).

Finally, some normalization methods may not be suited to reformulation in a regression framework at all. For example, Gerstman (1968) normalization expresses formant frequencies relative to the range for each speaker for each formant. Unfortunately, this procedure involves the estimation of maximum and minimum values for each speaker for each formant, which are not values that can be easily estimated using a regression framework.

III. EVALUATING NORMALIZATION PERFORMANCE USING SIMULATED VOWEL DATA

Most previous comparisons of vowel normalization procedures have focused on the degree to which these methods minimize some aspect of between-speaker variation (e.g., sex differences), and maximize the similarity of normalized vowel spaces, within dialect (Fabricius *et al.*, 2009; Flynn and Foulkes, 2011; Adank *et al.*, 2004). Overall, these studies focus on maximizing metrics relating to the similarity (or dissimilarity) of vowel systems as a function of the amount of within-category scatter after normalization. In general, previous reports indicate that Lobanov normalization exhibits superior within-category scatter-reduction in the normalized space, though log-mean normalization is typically not substantially worse.

According to this general approach, the best normalization method would be one that makes the vowel spaces of different speakers of the same dialect identical in the normalized space. However, the "scatter" in the normalized space represents subphonemic variation of the very sort that interests linguistic researchers. Further, repetition error and legitimate subphonemic variation mean that some amount of within-category variation will, and should, always be present in the productions of real speakers. If a normalization method is to preserve the linguistic facts with respect to vowel formant data, within-category variation in the normalized space should only be removed when it is phonetically-irrelevant. As a result, if two vowels appear the same after normalization they should have the same perceived vowel quality. If a normalization method makes the productions and vowel spaces of different speakers more similar than they ought to be, it is overnormalizing the data in question leading to a loss of legitimate within-category variation in vowel quality.

To our knowledge, there are three published comparisons of normalization methods that, in addition to scatter reduction, consider explicitly how well these maintain impressionistic differences in perceived vowel quality between speakers (Hindle, 1978; Labov, 1994; Kohn and Farrington, 2012). Both Hindle and Labov found that Sankoff (1978) normalization (a method of similar complexity to the Lobanov method, which was not tested) was found to be more effective than log-mean normalization at reducing within-category variation in formant patterns, but importantly, it also removed some of the sociolinguistic variation in perceived vowel quality between speakers. Kohn and Farrington (2012) found a very slight advantage for log-mean over Lobanov normalization in the ability to preserve perceptually-salient sociolinguistic differences; however, two issues affect the interpretation of their results. First, the authors sought to model perceived vowel quality as a function of normalized formant frequencies but also included several indexical predictors (e.g., speaker gender, age) in their models. This would have the effect of inappropriately controlling for systematic gender and age differences on vowel quality in the normalized space. Second, the researchers used nearly 100 tokens per speaker per time point to carry out their normalizations. This very large amount of data is not standard in most research situations and may reflect a situation where the performance of different normalization methods converge as estimated speaker parameters become increasingly accurate (see Sec. III C). Adank *et al.* (2004) address the issue of preserving sociolinguistic variation less directly in the form of unspecified regional dialect differences, but in our analysis, the results are inconclusive.[4]

There are two potential sources of overnormalization, which we will refer to as inherent overnormalization and sampling overnormalization. Inherent overnormalization occurs when a method controls for variation that is not controlled for (perceptually accommodated) by human listeners. Overnormalization of this kind is inherent to the structure and operation of a normalization method and so is likely to arise whenever the offending method is used. In the long run, we believe that this is the more pernicious source of overnormalization. However, it is also the more difficult type of overnormalization to study as it requires collecting independent perceptual information about the stimuli. For that reason, it is typically not considered in investigations regarding the appropriateness of different normalization methods for linguistic research.

Sampling overnormalization occurs due to noise in the estimated speaker parameters used for normalization in real-world situations. Errors in the estimation of speaker parameters required for normalization will result in artificial shifts between the vowel systems of different speakers of the same kind as shown in Fig. 1. However, in the case of error in estimated speaker parameters, these shifts will tend to bias the vowel spaces of different speakers toward more similarity. To understand why this is the case we can imagine that, given a normalization method and a dialect, each speaker can be associated with a set of true normalization parameters. These parameters reflect the exact distributional characteristics of the formants produced by that speaker given the structure of the normalization method. However, researchers do not have access to the true parameters for a given speaker, and so must estimate these parameters using the same data that is being normalized. A reliance on estimated parameters has the effect of making vowel spaces more similar than they should be by equating estimated parameters between speakers, rather than true speaker parameters.

For example, consider the case of Lobanov normalization for two vowels produced by two speakers. This example is unrealistically simple, but instructive as to the problem of sampling overnormalization. Imagine that it were somehow known that the two speakers produce identical acoustic output except for random production and measurement errors, and had true Lobanov $F2$ parameters of $\mu = 1200$ Hz and $\sigma = 600$ Hz. If speaker A produced two vowel tokens with $F2$ values of 600 and 1800 Hz, respectively, they would appear to have Lobanov $F2$ parameters of $\hat{\mu} = 1200$ Hz and

$\hat{\sigma} = 600$, which in this case would be correct. Lobanov normalization would make the speaker's standard deviation (600) equal to 1 so that the normalized $F2$ values for the vowels would equal -1 and 1. In this case the estimated speaker parameters equal the true speaker parameters so that the vowels fall on their true positions in the normalized space. In reality, the estimated and true speaker parameters will almost never exactly match.

If speaker B produced tokens of the same vowels with $F2$'s of 550 and 1850 Hz, they would have apparent Lobanov $F2$ parameters of $\hat{\mu} = 1200$ Hz and $\hat{\sigma} = 650$ Hz. However, although the standard deviation parameter has been overestimated, Lobanov normalization will still make this speaker's estimated standard deviation (650) equal to 1, and the normalized vowels will also have $F2$ values of -1 and 1. In this case, using the estimated standard deviation to normalize the productions is artificially equalizing the peripherality of the vowels produced by speakers A and B, and leading to incorrect positions in the normalized space for the vowels produced by speaker B. Had the true Lobanov parameters been used, the vowels produced by speaker B would have appeared at -1.08 and 1.08, reflecting the more peripheral productions of that speaker. Although all normalization methods will exhibit some degree of sampling overnormalization, situations involving small amounts of data and normalization methods that use parameters with relatively larger estimation errors are particularly susceptible to this. Just as with inherent overnormalization, sampling overnormalization is difficult to investigate quantitatively in typical situations when researchers only have access to a limited number of productions from a small group of speakers.

In the remainder of this section we will present an evaluation of the Lobanov and log-mean methods using simulated vowel-formant data. The use of simulated data allows us to consider the performance of these normalization methods in new ways that are not practical with real speakers. In most real-world situations researchers only have access to estimated speaker parameters, usually based on small numbers of tokens per speaker. This means that normalization methods can only be compared in terms of their superficial performance, without a consideration of sampling overnormalization related to errors in speaker-parameter estimation. In contrast, the use of synthetic speakers allows us to establish a ground truth regarding speaker parameters, true vowel quality, the true amount of scatter reduction that should be observed in a dataset, and so on. In Sec. III A we discuss how normalization methods can be compared using simulated data, and then present details about the simulation conditions (Sec. III B). In Sec. III C we investigate errors in the estimation of the normalization parameters used in Lobanov and log-mean normalization in different situations. In Sec. III D we investigate error in the estimation of the true locations of normalized vowel tokens. These errors are considered in two ways: errors due to the estimation of normalization parameters, and errors that are associated with the use of the "wrong" perceptual model. Finally, in Sec. III E we compare the scatter reduction observed when using estimated speaker parameters to true or expected scatter reduction, the amount of scatter reduction obtained when normalizing using the true speaker parameters.

A. Investigating overnormalization

Overnormalization by Lobanov and log-mean normalization is investigated below using simulated vowel data from simulated speakers. We may distinguish models that are meant to explain between-speaker variation in formant patterns in production from models that account for how listeners perceptually accommodate to this variation. Although related, these models are logically distinct and it is not necessary that they exactly correspond. Empirical data regarding variation in the formant patterns produced by different speakers suggest that speakers may vary in consistent idiosyncratic ways that cannot be captured by a single multiplicative parameter (see the Appendix). For this reason, simulated speakers used in this investigation were made to vary on the basis of the formant-wise means and standard deviation parameters as used in Lobanov normalization. This approach means that speakers can vary systematically in formant-wise location and dispersion in ways that cannot be equalized by log-mean normalization (but can be with Lobanov normalization).

The information presented in the Appendix suggests that between-speaker variability in production may be less constrained than what is suggested by the constant-ratio hypothesis (though perhaps not by much), however, it is not clear to what extent listeners control for this additional variation in perception. For example, if human listeners are "Lobanov listeners," they would estimate and control for all of the variation in Lobanov parameters between speakers in perception. This means that, for example, listeners would equate differences in formant-wise vowel-space dispersion between speakers and not perceive distinctions due to vowel space dispersion. On the other hand, despite the fact that variation in production may involve small systematic deviations from the constant-ratio hypothesis, listeners may still be operating under a perceptual model broadly consistent with the constant-ratio hypothesis (i.e., humans are "log-mean listeners"). This would mean that speaker-dependent deviations from the constant-ratio hypothesis in production are not normalized away and may be perceptually salient to human listeners. If this were the case, for example, differences in formant-wise dispersion between speakers may very well result in differences in perceived vowel-quality.

Since inherent overnormalization arises in the event of a mismatch between the normalization and perceptual models, it cannot be directly investigated with simulation studies. Instead we will consider the practical implications of this type of overnormalization by highlighting differences in scatter reduction and the apparent perceptual structure of vowel tokens on the basis of the normalization method employed. We will also consider sampling overnormalization, which can easily be investigated using simulations since the exact characteristics of the simulated speakers and repetition error are known.

B. Simulation of vowel formant data

The generation of naturalistic simulated vowel data is described in detail in the Appendix. Simulated vowels were based on the mean locations of the vowel phonemes in the Hillenbrand *et al.* (1995) data in a Lobanov-normalized space. First, 20 simulated speakers were generated, represented by a set of Lobanov parameters ($\hat{\mu}_{ks}$, $\hat{\sigma}_{ks}$) for $F1$, $F2$, and $F3$. Then, a random subset of the Hillenbrand vowels was selected, representing the vowel phonemes of a simulated pseudolanguage. Each language had either five or nine vowels, and these always contained /i ɑ u/, the "point" vowels in the Hillenbrand data. Based on the selected vowel phonemes, the normalized template, and the simulated-speaker coefficients, a given number of tokens were generated for 20 simulated speakers of the pseudolanguage.

Pseudolanguages were simulated in six different conditions varying according to the number of repetitions and the number of vowel categories featuring missing data, presented in Table I. For the conditions with missing data, missing observations were determined by randomly selecting between 1 and 14 tokens for each category with missing data and designating these "missing." This process was repeated with independent randomly-selected tokens for each vowel category with missing data. Ten thousand pseudolanguages were simulated for each condition, for each vowel-system size, for a total of 120 000 pseudolanguage datasets.

C. Error in normalization parameter estimates

Sampling overnormalization arises as a result of error in parameter estimates when based on a limited sample of data. To investigate errors in parameter estimates for the different parameters, the true Lobanov parameters for each speaker from each pseudolanguage dataset were recorded.

Although simulated speakers varied in more ways than is explainable by a single log-mean parameter, any given speaker can still be thought of as being associated with a true log-mean parameter given their exact expected phoneme targets. A true log-mean parameter for each speaker was calculated by finding phoneme targets for each speaker (in Hertz) and calculating the speaker log-mean parameter using Eq. (3). Phoneme targets in Hertz were found for each speaker using their Lobanov parameters and the dialectal vowel template (see the Appendix). In addition to the true parameters for each speaker, Lobanov and log-mean parameters were estimated from the simulated vowel data (including error), resulting in parameter estimates that tend to differ from the true parameters.

TABLE I. Information regarding number of repetitions (reps.), vowel-system sizes, and number of vowel categories with missing data (missing) in simulated pseudolanguage datasets.

	5 vowels		9 vowels	
Condition	Reps.	Missing	Reps.	Missing
1	10	0	10	0
2	5	0	5	0
3	2	0	2	0
4	1	0	1	0
5	1	1	1	2
6	1	2	1	4

The amount of error between estimated and true parameters was quantified in two ways. First, the RMS error between the true and estimated parameter (in Hertz) was calculated. Log-mean parameters were exponentiated for all comparisons. Second, the percent RMS error was calculated by dividing errors by the value of the true parameter and expressing this as a percentage. The percent RMS error is intended to give a clearer picture of error magnitude relative to parameter values. In addition to error in parameter estimates, the amount of error in the underlying data (e^2) was estimated by finding the RMS error between the simulated vowel tokens and the exact phoneme targets for each speaker. In missing data situations, all data were estimated based on the complete-case data, meaning all categories with missing data were excluded from the calculation of the true and estimated parameters. A subset of these results is presented in Fig. 2.

As seen in Fig. 2, the log-mean parameter has the lowest percent RMS error in all situations, usually below around 1% of the parameter value. The formant-wise mean parameters used by the Lobanov method feature larger errors, which is to be expected given that they can only be calculated using 1/3 of the available data for 3-formant data. However, errors in the formant-wise mean parameters are not substantially larger than for the log-mean parameter. On the other hand, the standard deviation parameters used in Lobanov normalization exhibit large errors in all conditions. In fact, this error is about as large as the underlying repetition error even when 10 repetitions are available for each speaker, and is expected to be 5%–20% as large as the value of the parameter even when a full set of nine vowels is available per speaker. This

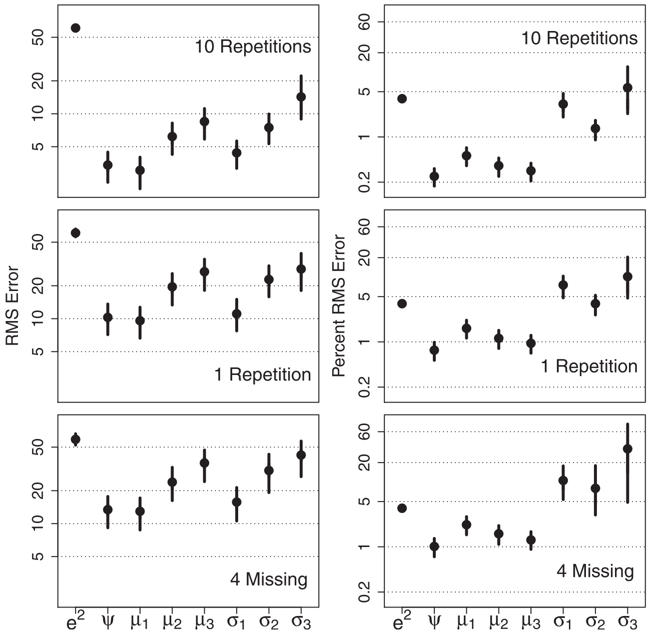

FIG. 2. Highest-density intervals of RMS error (95%) for parameter estimates (e^2 = noise, ψ = log-mean parameter, μ_k = mean for formant k, σ_k = standard deviation for formant k) for 9-vowel systems based on 10 repetitions, 1 repetition, and with 4 missing categories, meaning only 5 tokens were available for each speaker.

suggests that log-mean normalization will exhibit very little type-B overnormalization, while Lobanov normalization will be associated with moderate to large amounts of type-B overnormalization even when multiple repetitions are available for each speaker.

The errors for the standard deviation parameters calculated for Lobanov normalization may seem unusually large since these parameters appear to be estimates of the standard deviation of productions about the mean produced by a speaker for a formant. However, the formant values produced by a speaker will be distributed around the means for their respective vowel categories rather than about the overall mean for the speaker. Thus, the standard deviation parameters calculated as in Eq. (6) are analogous to the square root of the expected mean square (EMS) for treatments in a one-way analysis of variance. The EMS is given by Keppel [1991, p 94, Eq. (5-3)] as

$$\mathrm{EMS(A)} = \sigma^2_{\mathrm{error}} + \mathrm{n} \cdot \sum_{i=1}^{a} \frac{\alpha_i^2}{a-1}. \tag{17}$$

In Keppel's example there is one treatment factor A with a levels and n replicates per level and α_i represents the true (but unknown) value of the ith treatment mean from a true (but unknown) population mean.

We translate this now into the current framework, where the single "treatment" is the vowel category V. We take as the observations in the analysis of variance the sample vowel means over repetitions (not the raw individual measurements themselves), so that v_i represents the true difference between the ith vowel mean and the true overall speaker mean. From the perspective of the analysis of variance, the number of replicates is 1 and the σ_{error} corresponds to the vowel mean measurement error τ_{vowel}, which is equal to repetition error divided by the square root of the number of tokens used to calculate the vowel means. The updated expression is presented in Eq. (18),

$$\mathrm{EMS(V)} = \tau^2_{\mathrm{vowel}} + 1 \ \cdot \ \sum_{i=1}^{v} \frac{v_i^2}{v-1}. \tag{18}$$

The relationship expressed in Eq. (18) states that the EMS of the vowel factor for a formant (i.e., the standard deviation parameter $\hat{\sigma}_{ks}$) is equal to the true mean-square for the factor $\left[\sum(v_i^2/v-1)\right]$, plus τ^2_{vowel}. Thus, we can see that the expected error in this estimator will itself have a variance of τ^2_{vowel} and a standard deviation of τ_{vowel}, corresponding to a repetition error divided by the square root of the number of repetitions. This means that the error in the Lobanov standard deviation parameter estimates is expected to be quite large in situations with few replicates, as indicated in Fig. 2.

D. Vowel quality error

Locations in the normalized space are intended to reflect specific vowel qualities. As a result, error in identifying the true location of vowel tokens in the normalized space is equivalent to misidentifying the vowel-quality associated with a token. In the ideal situation, a researcher is able to identify the true locations of vowels in the normalized space such that the structure of normalized tokens closely reflects their perceptual organization. In practice, there are two potential sources of error in identifying the correct location of a token in the normalized space, resulting in error in the apparent vowel-quality of a token.

First, a researcher may introduce vowel-quality error into their data because they employ a method that looks for the wrong target location in the normalized space. Although thus far our focus has been on overnormalization, we can consider "misnormalization" more generally. Owing to the different transformations employed by the two methods considered here (and different normalization methods more generally), Lobanov and log-mean normalization assign different true locations in the normalized space to each token. Roughly speaking, these target locations correspond to true perceptual vowel-qualities from the perspective of two ideal classes of listeners: The true Lobanov targets represent the perceptual structure of the data according to the perceptual model adopted by Lobanov normalization (i.e., according to Lobanov listeners), while the true log-mean targets represent the perceptual structure of the tokens according to the log-mean perceptual model (i.e., according to log-mean listeners). As a result, each method "misnormalizes" the data from the perspective of the perceptual model implied by the other method. From the perspective of Lobanov normalization, log-mean normalization misnormalizes to the extent that it preserves phonetically-irrelevant (i.e., imperceptible) variation in formant-wise dispersion in the normalized space. From the perspective of log-mean normalization, systematic between-speaker variation in formant means and dispersions may result in perceived phonetic differences, and the removal of this variation will constitute misnormalization (in this case, overnormalization). In sum, using log-mean normalization when the true perceptual mapping is in line with a Lobanov model (and vice versa) will introduce error in the apparent vowel quality for a set of tokens, which will remain no matter how much evidence is available.

Second, vowel-quality error will arise from error in the estimation of the speaker parameters relevant for normalization (related to sampling overnormalization). Given a normalization method and the relevant exact true speaker parameters, any vowel token can be associated with a target location in the normalized space that reflects its exact true vowel quality from the perspective of the model. However, a researcher does not have access to the true speaker parameters for a speaker and must estimate these from a sample of tokens. As shown in Sec. III C, there can be substantial error in the estimated speaker parameters necessary for normalization (particularly for the Lobanov standard deviation parameters). Error in the speaker parameter estimates will directly translate into shifts in the vowel spaces of different speakers of the same general kind as those shown in Fig. 1. This sort of vowel quality error should decrease as more data are available for the estimation of the relevant speaker parameters; however, Sec. III C indicates that Lobanov parameters can have large errors even in cases when many tokens are available for every speaker.

We will quantify vowel-quality error by considering the distance between target locations in the normalized space (the target vowel-quality) and the estimated location in the normalized space (the estimated vowel-quality) for vowel tokens under different conditions. The following process was carried out for each pseudolanguage dataset. First, vowels were normalized using the true speaker parameters using both Lobanov and log-mean normalization. Recall that the true speaker parameters were fixed in advance and served as the basis of all simulations. After this, the log-mean normalized vowels were exponentiated and globally-standardized within-formant, so that they would represent vowel qualities in a Lobanov-compatible space. The Lobanov-normalized vowels were also globally standardized within formant so that both sets of normalized vowel tokens would have formant-wise means of zero and standard deviations of one. These two sets of normalized vowel tokens represent target locations in the normalized space and associated true vowel qualities from the perspective of each normalization method.

Each pseudolanguage dataset was also log-mean and Lobanov normalized using parameters estimated from the simulated data. Lobanov normalization parameters were estimated using a complete-case analysis, using only tokens whose categories were present for all speakers. All nonmissing data were then normalized using the complete-case subset parameters. Log-mean parameters were estimated using both the complete-case analysis and the regression approach presented in Sec. II. The log-mean normalized data based on estimated parameters was exponentiated, and both the log-mean and Lobanov-normalized data were globally standardized within-formant, independently for each normalization method. The above process resulted in five sets of normalized vowels for each pseudolanguage dataset: two sets of true vowel-qualities (the Lobanov and log-mean target vowel qualities), and three sets of estimated vowel-qualities.

To quantify errors, we looked at the data from the perspective of each normalization model, effectively considering tokens from the perspective of each type of ideal listener (Lobanov and log-mean). For vowel-quality error associated with sampling overnormalization, we calculated the average Euclidean distance in the 3-formant space between the estimated vowel-quality for each normalization method and the true vowel-quality target for that method (e.g., true Lobanov vs estimated Lobanov). To quantify the vowel-quality error associated with inherent overnormalization, and misnormalization more generally, we calculated the Euclidean distance in the 3-formant space between true vowel-quality targets for log-mean and Lobanov normalization, and the vowel quality estimates provided by the other method (e.g., Lobanov estimates to log-mean targets).

1. Results

Figure 3 presents vowel-quality error for the complete-case analysis of each normalization method, relative to each perceptual model. Since data were standardized along each formant for each model, the average distances in the normalized space are expressed in z-scores. Although this value is difficult to interpret in an absolute sense, we can compare the magnitude of vowel quality errors relative to the range of magnitudes seen under different conditions below.

First, we will discuss the vowel quality errors made by each normalization method relative to the log-mean vowel-quality targets (top row of Fig. 3), that is, from the perspective of a log-mean perceptual model. Sample-based log-mean normalization is generally accurate in all conditions, producing vowel-quality errors that are reasonably small even with only a single repetition available per speaker. Lobanov normalization provides vowel-quality estimates that tend to be quite far from the true log-mean vowel qualities, always resulting in a larger error than the log-mean estimates. This indicates that Lobanov normalization provides an incorrect picture of the vowel qualities associated with the tokens if humans are log-mean listeners. As noted above, since this sort of vowel quality error is associated with inherent overnormalization, increasing numbers of repetitions cannot improve accuracy beyond a certain point, though decreasing the number of observations does increase the size of errors.

In contrast to log-mean normalization, the Lobanov method (bottom row of Fig. 3) is quite sensitive to the amount of information used to estimate the speaker parameters. Error in vowel-quality estimates by the Lobanov method relative to Lobanov vowel-quality targets increases rapidly as the number of observations available for each speaker decreases. In fact, Lobanov normalization produces relatively large vowel-quality errors even when there is a single full set of observations for each speaker. Just as with Lobanov estimates for log-mean perceptual targets, log-mean normalization provides poor estimates of the true Lobanov vowel-quality. However, Lobanov vowel-quality

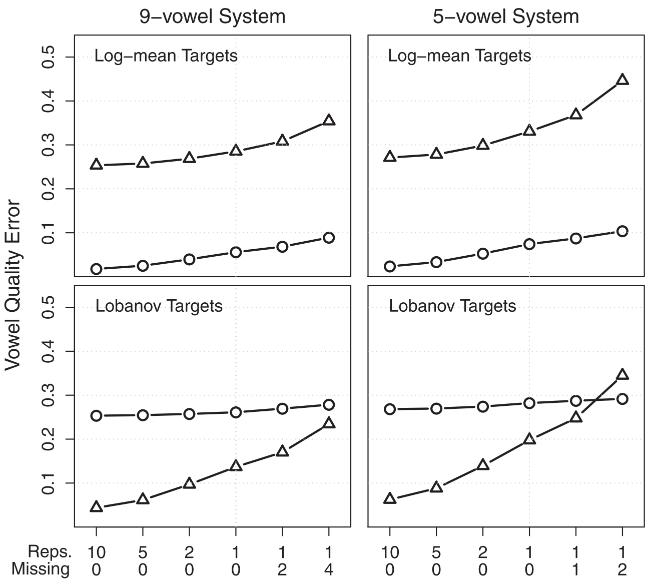

FIG. 3. Mean vowel quality error in different conditions varying by number of repetitions (reps.) and number of categories with missing data (missing). Vowel estimates based on Lobanov (triangle) and log-mean (circle) normalization are presented for to the log-mean and Lobanov vowel-quality targets. Vertical lines indicate the complete-case, single repetition condition.

estimates are so noisy and sensitive to decreases in the amount of data available that in some cases log-mean estimates of Lobanov vowel-quality targets are nearly as (or more) accurate as those of Lobanov normalization, even though the log-mean estimates are attempting to recover different locations in the normalized space (i.e., the log-mean vowel-quality targets).

a. Advantage for regression estimation in missing-data situations. The advantage for the regression approach to log-mean normalization relative to the complete-case approach was calculated by finding the difference in performance for the two methods for each simulation, and then finding the average across all simulations in each condition. In general, the advantage for the regression approach to normalization increases with the number of categories that must be omitted from the analysis due to missing data. For 9-vowel systems, vowel-quality error was 8% smaller for 2 missing categories, and 19% smaller for 4 missing categories. For 5-vowel systems, errors were 6% smaller for 1 missing category and 14% smaller for 2 missing categories. Overall, the use of regression to estimate the log-mean parameters resulted in smaller errors in 84% of cases, and this advantage is obtainable "for free" simply by adopting an alternative parameter estimation method.

E. Scatter reduction

In order to investigate scatter-reduction relative to overnormalization, the following process was carried out for each pseudolanguage dataset across each condition and vowel-system size. First, the data were normalized using the Lobanov and log-mean methods using the true speaker parameters for each speaker. A multivariate analysis of variance was carried out on each set of normalized formant values with vowel category as the only predictor, and the Pillai score was recorded for each normalization method. A higher Pillai score indicates that there is less within-category variation relative to between-category variation in a dataset, indicating more-similar normalized vowel spaces for the speakers in the dataset. In addition, the speaker parameters necessary for Lobanov and log-mean normalization were also estimated from the simulated data. In missing-data situations, Lobanov parameters were estimated using a complete-case analysis. Log-mean parameters were estimated in two ways: using a complete-case analysis and using the regression approach outlined in Sec. II. This process resulted in five sets of Pillai scores for each pseudolanguage: two sets of true scores (log-mean and Lobanov), a score for Lobanov normalization with estimated parameters, and two scores for log-mean normalization based on the complete-case and regression parameter estimates.

The Pillai score observed when a normalization method uses the true speaker parameters (i.e., the true Pillai score) represents the desired amount of scatter reduction for the data in question. When a normalization method is applied using the exact true parameters for a given speaker, it has no basis by which to appropriately eliminate further phonetically-irrelevant variation from a dataset. As a result, although in general a higher Pillai score is more desirable, an observed Pillai score higher than the true Pillai score indicates that too much variation has been removed from the data so that (sampling) overnormalization has occurred.

1. Results

The distribution of mean Pillai scores for the complete-case estimations of each normalization method across conditions is presented in Fig. 4, which shows the same general pattern for 5- and 9-vowel systems. First, Pillai scores are higher for Lobanov normalization, which is in line with previous reports that Lobanov normalization is the more successful method with respect to scatter reduction in the normalized space. Second, Lobanov and log-mean normalization show different patterns of overnormalization as a function of the amount of data available to estimate the speaker parameters. Log-mean normalization generally shows a close alignment between observed and true Pillai scores across all of the conditions tested. This indicates that log-mean normalization tends to remove an appropriate amount of variation when relying on estimated speaker parameters.

In contrast, the Pillai scores for the Lobanov method show more variation as a function of the amount of data available to the algorithm. Consider the performance of the Lobanov method in the case of one repetition per category, with no missing data. As can be seen in Fig. 4, this condition results in substantially higher Pillai scores relative to when the true speaker parameters are used to normalize. This indicates that in these situations, vowel data are being overnormalized and legitimate variation in vowel quality is being removed from the data. The fact that Pillai scores decrease as the number of repetitions increase suggests that the overnormalization of data seen in the single-repetition condition is related to error in the estimation of Lobanov parameters from small amounts of data (sampling overnormalization). As the number of observations increase, error in speaker parameters decreases resulting in a reduction in the amount

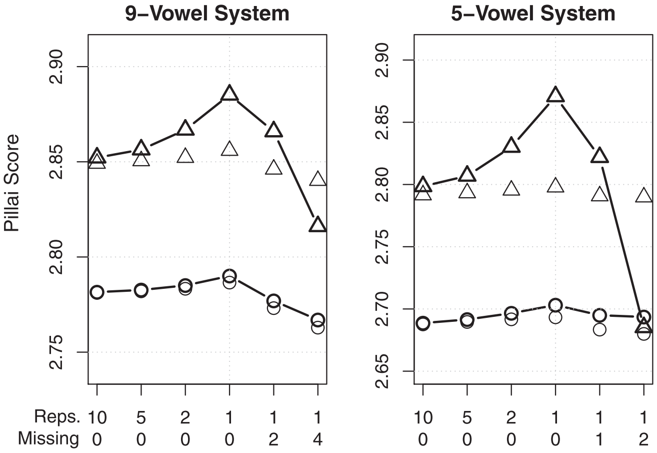

FIG. 4. Average Pillai scores for each condition using Lobanov (triangles) and log-mean (circles) normalization. Conditions differ in number of repetitions (reps.) and number of categories with missing data (missing). Points joined by lines indicate average Pillai scores based on estimated parameters, points indicate average true Pillai scores. Vertical lines indicate the complete-case, single repetition condition.

of scatter reduction (and overnormalization) produced by the Lobanov method. The reduction in Pillai scores in the missing-data conditions may appear to indicate the performance of Lobanov normalization is improving in these situations. However, these conditions involve fewer tokens and are associated with noisier parameter estimates (see Sec. III C). Thus, the decrease in Pillai scores in these situations likely reflects the fact that in missing-data conditions where parameters are estimated based on a complete-case analysis, parameter estimates were determined using a subset of the available data and so are only overfit with respect to some of the tokens produced by each speaker.

a. Advantage for regression in missing-data situations. The regression approach to estimation of the speaker parameters again resulted in a modest but consistent improvement relative to the complete-case approach. For 9-vowel systems, the average amount of overnormalization (measured by subtracting the observed Pillai score from the true Pillai score) in the missing-data conditions was 0.0039 for the complete case analysis as compared to 0.0007 for the regression analysis, representing a decrease of 87% in this value. Both of these values compare quite favorably to the average overnormalization of 0.029 when using the Lobanov method in complete-case situations with a single repetition per speaker. For 5-vowel systems, overnormalization decreases from 0.0125 to 0.0087, a decrease of 86%. Again, both of these values compare favorably to the 0.072 overnormalization by Lobanov in single-repetition complete-case situations for a 5-vowel system. Overall, overnormalization was reduced at least somewhat in 84% of missing-data cases when speaker parameters were estimated using the regression approach rather than the complete-case analysis.

F. General discussion

Examples of vowel quality errors for each normalization method relative to their own vowel-quality targets are given in Figs. 5(a) and 5(b). As can be seen, the average magnitude of the vowel-quality errors with Lobanov normalization are large enough to meaningfully affect the apparent vowel quality associated with many tokens, while those of the log-mean method tend to be quite small. The error magnitudes shown in Figs. 5(a) and 5(b) are typical for log-mean and Lobanov normalization with a complete set of nine observations per speaker, suggesting that researchers using the Lobanov method in many typical research situations run a high risk of introducing significant vowel-quality error into their data. The substantial difference in vowel-quality errors can be directly attributable to the large errors in the standard deviation parameters used for Lobanov normalization presented in Sec. III C.

Figure 5(c) shows the difference between log-mean and Lobanov vowel-quality targets for the same vowel tokens, arising as a result of the different operations carried out by the two normalization methods. Clearly, these perceptual targets (and the overarching perceptual organization of the tokens) cannot both be right: one is likely to correspond more closely to the true perceptual organization of the tokens in the opinion of human listeners. Using a normalization method that does not reflect the perceptual processes of human listeners will thus result in the wrong targets in the normalized space, potentially leading to large errors in apparent vowel quality relative to the true quality of each token.

The above highlights that the selection of a normalization method has practical consequences for the ability to make reliable inferences regarding patterns in vowel-quality between and within-speakers. The selection of a normalization method will determine the apparent structure of the data so that committing to a method necessarily means committing to a position regarding the perceptual organization of the tokens. In light of this, the selection of a normalization method cannot be viewed simply as a methodological tool without any theoretical implications. If a researcher uses a normalization method that suggests an inappropriate perceptual mapping, this may result in substantial errors in the apparent vowel quality associated with individual tokens. Although the true nature of the human perceptual space is not exactly known, we have reason to believe that it is broadly consistent with the assumptions underlying single-parameter log-mean normalization (see Secs. I B 1 and I C 1). Conversely, and perhaps more importantly, we know of no reason to believe that the range of variation in formant patterns allowed by Lobanov normalization is consistent with the tolerated range of phonetically-equivalent sounds in the opinion of human listeners.

In addition to its theoretical and empirical support as a model of human vowel perception, log-mean normalization is very resistant to changes in the amount of information available to the method, and resulted in low overnormalization and small vowel-quality and parameter-estimation errors in all conditions. The regression approach to estimation of the log-mean parameter leads to modest but consistent reductions in overnormalization and vowel quality error in missing-data situations relative to the complete-case analysis. Because of the ease of implementing this method and the expected improvement in performance, it seems prudent to carry out log-mean normalization using the approach outlined in Sec. II in missing-data situations.

As with previously-reported findings, Lobanov normalization resulted in greater scatter-reduction across most conditions tested here. The increased scatter-reduction within category is evident when comparing the distribution of tokens in Figs. 5(a) and 5(b). However, the increased Pillai scores can be directly attributed to the independent dispersion parameters estimated for each formant as opposed to the more restricted dispersion reduction carried out by the log-mean method. To the extent that log-mean normalization is a more accurate reflection of the true perceptual organization, most of the increase in scatter-reduction seen for Lobanov normalization represents the removal of legitimate linguistic variation (inherent overnormalization) and is therefore undesirable.

In addition, Lobanov normalization has a tendency to overnormalize vowels due to noisy parameters when small numbers of repetitions are available for each speaker, a situation which is common in many research situations. The tendency of Lobanov normalization to exhibit sampling overnormalization is evident in Fig. 5(b) where the true vowel qualities (empty points) are more dispersed than the estimated vowel qualities (filled points). Overall, it seems

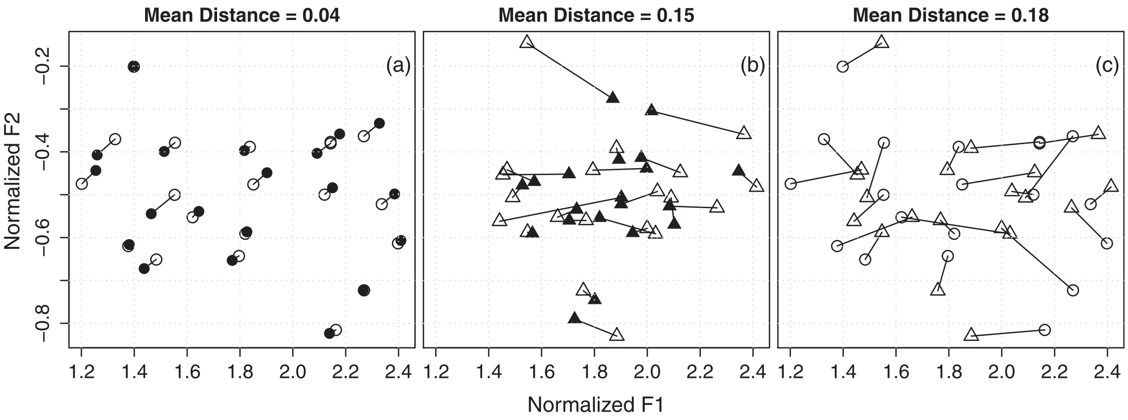

FIG. 5. (a) True (empty points) and estimated (filled points) vowel qualities for log-mean normalization relative to the log-mean perceptual model. Data from a single vowel category from a 9-vowel pseudolanguage is presented. (b) True (empty triangles) and estimated (filled triangles) vowel qualities for Lobanov normalization relative to the Lobanov perceptual model, for the same data in (a). (c) True vowel quality according to the log-mean (circle) and Lobanov (triangle) perceptual models, for the same data in (a) and (c).

that even if one adopts the Lobanov model of vowel perception and desires a normalization that equalizes for formant-wise dispersion, in practice, the large errors in the Lobanov standard deviation parameters will tend to result in overnormalization of data and large errors in apparent vowel-quality. As a result, it seems prudent to reserve use of the Lobanov method for situations when multiple repetition of complete vowel inventories are available for each speaker.

IV. CONCLUSION

This article presents a readily-applicable generalization of the log-mean normalization method that avoids certain obvious kinds of bias in the face of missing and unbalanced data. This approach shares the same important assumptions made by traditional log-mean normalization, namely, that the formant patterns produced by speakers of the same dialect differ from each other on the basis of a speaker-dependent scaling factor and repetition error. This regression-based approach provides more accurate estimates of the vowel quality associated with each token, and is less prone to overnormalization in missing-data situations.

In addition, we have argued that the selection of a normalization method for use in linguistic research cannot simply be treated as a methodological issue: if a normalization method is to accurately reflect the linguistic structure of a set of vowel tokens it must correspond, in process or outcome, to the mechanisms of human vowel perception. On this approach, there are several reasons to be skeptical of the normalized vowel qualities recovered by the Lobanov method. First, to our knowledge no current theory of vowel perception is compatible with the Lobanov model, and no researcher is advocating for its plausibility. The estimation of location and dispersion parameters for each formant for each speaker likely represents an intractable problem for the listener when only a limited amount of information is available for that speaker. In fact, Lobanov normalization includes large errors in apparent vowel-quality and overnormalizes data substantially even in cases where a full vowel inventory is available for a speaker. Second, it is not clear that variation in formant dispersion is phonetically-irrelevant in all situations, a precondition that must be satisfied for Lobanov normalization to not remove linguistically meaningful variation from formant patterns. As a result, Lobanov normalization may be prone to committing both inherent and sampling overnormalization.

Although the use of log-mean normalization has a good amount of empirical and theoretical support (see Sec. I B 1), further research is needed to understand the nature of the speaker-dependent normalized space. Furthermore, we suggest that the consideration of vowel normalization methods as models of human vowel perception can help guide the evaluation and consideration of methods that reflect the linguistic facts reflected in vowel formant data, rather than seeking to maximize metrics not directly related to human speech perception. Finally, we readily acknowledge the need for more direct perceptual testing of alternative "normalizable" vowel spaces. This approach may prove to be difficult for a number of reasons, including the fact that detailed perceptual judgments are relatively difficult to obtain and that they may be quite variable compared to often relatively small differences in predicted phonetic quality provided by alternate normalization models. We nonetheless think it is necessary for solid refinement of the theories involved, and for the selection of normalization methods that lead to reliable inferences regarding patterns in vowel-quality between and within speakers.

APPENDIX: SIMULATING NATURALISTIC VOWEL DATA

The generation of simulated vowel data requires that three aspects of variation in formant patterns be specified: (1) between-phoneme variation, (2) random within-speaker, within-phoneme variability, and (3) systematic between-speaker variation. Each of these aspects will be outlined for the simulation method used for the results in Sec. III.

1. Between-vowel variation

The dialectal template specifies the ideal locations of each phoneme of the dialect in a normalized space, representing the phonetic information shared by all speakers in the dialect. For the data used in Sec. III, the dialectal template was established using the vowels produced by the 98 speakers in the Hillenbrand *et al.* (1995) dataset will a full set of $F1$, $F2$, and $F3$ steady-state measurements for all 12 vowels. The dialectal template was established by Lobanov normalizing all formant frequencies in a dataset and finding the average normalized $F1$, $F2$, and $F3$ value for each vowel. These mean values were then standardized along each formant, resulting in the mean dialectal targets specified in a Lobanov-normalized space.

Given a dialectal template specified in the Lobanov-normalized space, the phoneme targets for a specific speaker in Hertz can be obtained by "unnormalizing" the dialect template values, shown in Eq. (A1),

$$\mathrm{F}_{vks} = (\mathrm{z}^*_{vk} \cdot \hat{\sigma}_{ks}) + \hat{\mu}_{ks}. \tag{A1}$$

This process is the inverse of Lobanov normalization and generates Hertz-valued formant frequencies (F_{vks}) based on the dialectal reference value for vowel v and formant k (z^*_{vk}), and a speaker-specific mean and standard deviation parameter for that formant ($\hat{\mu}_{ks}$, $\hat{\sigma}_{ks}$).

2. Random within-speaker within-phoneme variability

Within-phoneme, within-speaker variability in formant patterns was investigated using the North Texas Vowel Database (Assmann and Katz, 2000), which appears to be the only publicly available dataset of formant values that includes several repetitions of the same word by multiple speakers in controlled conditions. This data will be used to establish the characteristics of the repetition error that will be added to the phoneme targets for simulated speakers in the Hertz space.

The North Texas data include 12 vowels produced by 20 adult speakers, 10 males and 10 females. The data are highly unbalanced with different numbers of repetitions for each category for each speaker (between 0 and 14), including missing categories for several speakers. For each speaker, we included only those categories that had at least five repetitions. The mean and standard deviation for each formant, for each vowel, for each speaker were found, resulting in three mean and standard deviation measurements for each vowel produced by each speaker. The standard deviation values provide estimates of the repetition error associated with each average formant value.

A visual inspection of the relationship between formant standard deviations and formant means revealed that this relationship was non-linear, and also showed an increasing standard error as mean formant increased. To address both of these issues, we modeled the logarithm of the standard deviation of a formant as a function of the logarithm of the mean of that formant. We considered three models, one featuring only mean formant frequency as a predictor, a second that included additional predictors related to each speaker, and a third that also included predictors relating to the vowel category associated with the measurement. The model with only an intercept and formant frequency as a predictor explained 57% of the variance in the standard deviations of formant frequencies. The addition of speaker predictors increased this to only 59% despite the addition of 19 parameters to the model. The inclusion of vowel predictors only increased the variance explained to 61% despite adding a further 11 degrees of freedom to the model. As a result, although there may well be systematic variation in error magnitudes according to speaker (and perhaps vowel), for our purposes it seems reasonable to model this variation solely on the basis of the mean formant associated with the vowel as this more restricted model captures a large amount of the variance with only a single predictor.

Error was created by drawing from a normal distribution with a standard deviation was equal to $\sigma = 0.423 \times \mathrm{F}^{0.662}$, where F is equal to the Hertz value of the phoneme target for that speaker. The error was then added to the Hertz value of the phoneme target for that speaker such that no production for any given speaker will be expected to exactly equal their phoneme target. An example of the sort of error generated by this model is presented in Fig. 6. It should be noted that the Assmann and Katz data features more variability in the magnitude of error expected for a given formant frequency but, on average, the magnitude of errors predicted by this model is in line with that seen in natural productions.

3. Systematic between-speaker variation

Generating simulated speakers was done by generating vectors of Lobanov parameters that conformed to the covariance patterns of these parameters in real speakers. On account of the relationships between the size of a resonator, formant outputs, and vowel-space dispersion outlined in Sec. I C 1, it is expected that there will be a strong positive correlation between the mean frequencies of each formant, and between mean FFs and measures of formant dispersion (e.g., formant standard deviation). In fact, if speakers varied exactly according to a single multiplicative scaling-parameter (as suggested by the constant-ratio hypothesis) then all of the Lobanov parameters would be perfectly correlated with each other since a proportional increase to one formant would result in equal proportional increases to all formants (and their standard deviations). To investigate the

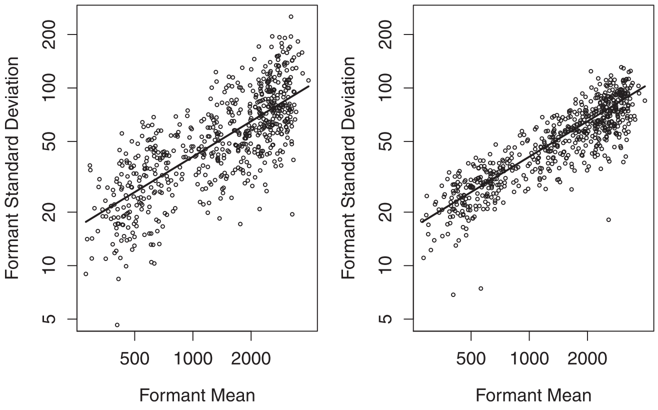

FIG. 6. A comparison of (a) repetition error observed in the Assmann and Katz (2001) data, and (b) simulated repetition error for the same mean formant frequencies and number of observations as the Assmann and Katz data. Lines in both panels show the relationship between formant means and standard deviation used by the model.

TABLE II. Correlations between Lobanov parameters for the speakers in the Hillenbrand *et al*. (1995) data with a full set of steady-state observations.

	μ_1					
μ_2	0.91	μ_2				
μ_3	0.91	0.94	μ_3			
σ_1	0.75	0.68	0.72	σ_1		
σ_2	0.68	0.66	0.70	0.72	σ_2	
σ_3	0.54	0.6	0.61	0.49	0.60	σ_3

empirical relationship between the Lobanov parameters across speakers, we found the mean and standard deviation for the first three formants for the speakers with complete data from the Hillenbrand *et al.* vowels. The correlations between the six calculated Lobanov parameters for these speakers are presented in Table II.

As seen in Table II, there are very high correlations between the formant-wise mean parameters and moderate to high correlations between the mean and standard deviation parameters for each formant and among the standard deviation parameters for each formant. Given that these correlations are being estimated, including repetition and measurement error, which can be quite large (see Sec. III C), the true correlations between these parameters are likely to be higher. To investigate the amount of independence in these parameters, the mean and standard deviation coefficients were standardized within-formant. A principal components analysis carried out on the standardized mean parameters indicates that 95% of the variance in these parameters falls along a single dimension. A similar analysis carried out on the standard deviation parameters indicates that 74% of the variance in these parameters falls along a single dimension. So, while the relationships between these parameters are not perfect, they are not as independent as they might be given the lack of constraint in variation of these parameters implicit in Lobanov normalization.

In order to create speakers that varied in accordance with the patterns seen in Table II, we found the average mean and standard deviation for each for the first three formants across all the Hillenbrand *et al.* speakers with complete data. This vector served as the starting point for all the simulated speakers. The overall mean speaker vector was log transformed to make new speakers so that the variability added would be proportional to the values of the parameters (equivalent to adding log-normal noise in the Hertz space). New speakers were generated by adding four distinct kinds of noise representing idiosyncratic but consistent between-speaker variation. The magnitude of each independent source of error was determined heuristically so that the correlations between the Lobanov parameters would be in line with empirical correlations after the addition of noise. Interpretation of the errors added at this stage is facilitated by the fact that for small values (<0.2), logarithmic changes are close to proportional changes. For example, an increase of 0.1 log-Hz corresponds to an increase of approximately 10% in formant frequencies [$\exp(0.1) = 1.105$].

First, a random variable with a mean of 0 and a standard deviation of 0.1 was drawn and added to the parameter vector. This represented differences between speakers that arise primarily from differences in vocal-tract length and is variation in strict accordance with the constant-ratio hypothesis. This resulted in unimodal variation in the average formant frequencies produced by simulated speakers, whereas variation in real speakers is multimodal with modes for adult males, adult females, and pre-pubescent children. However, this simplification should not have any effect on the performance of the normalization methods being considered here.

Second, a random variable was drawn with a mean of 0 and a standard deviation of 0.07 and added to all the standard deviation parameters. This allows for the overall vowel space dispersion to vary independently of the overall mean formants produced by the speaker. Third, three draws were taken from a normal distribution with mean 0 and standard deviation of 0.03. These values were centered and added to the mean parameters. These values allowed for formant means to vary independently of each other, and independently of their standard deviations. Finally, another three draws were taken from a normal distribution with a mean of 0 and a standard deviation of 0.07. These values were centered about zero and added to the standard deviation parameters. These values allowed the standard deviation for each formant to vary independently of each other, and of their respective formant mean.

To investigate the appropriateness of these values, the following verification was carried out. We generated 98 simulated speakers (the same *n* as the complete-case Hillenbrand dataset), generated a single repetition of each vowel, and added repetition error as outlined in section 2 of this appendix. After this, the Lobanov parameters for each speaker were calculated and recorded. This process was repeated 10 000 times resulting in a distribution of Lobanov parameter correlations for simulated datasets of the same size as the empirical Hillenbrand *et al.* data. These distributions are compared to the empirical correlations in Fig. 7, showing that the between-speaker variability generated by our simulations corresponds well to empirical estimates.

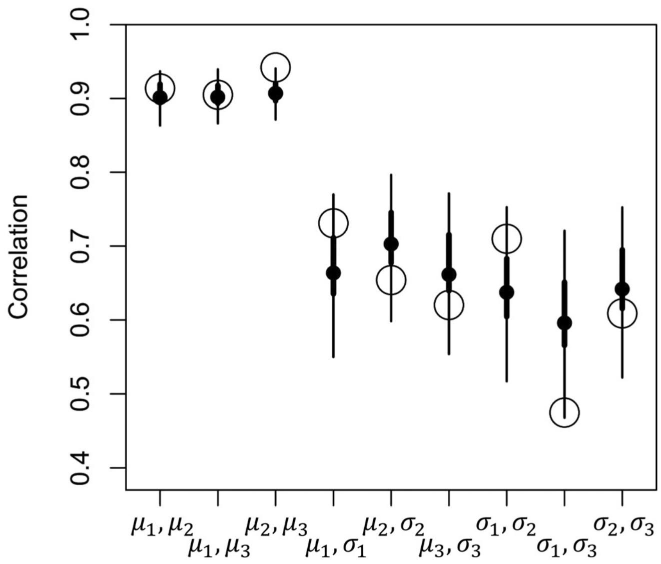

FIG. 7. Highest-density intervals of the correlations of Lobanov parameters for simulated speakers. Circles indicate the empirical correlations between parameters in the Hillenbrand *et al*. (1995) data.

[1] Using log-transformed formant frequencies also offer additional advantages, such as more nearly homogeneous variances for speaker groups within vowel within formant (Nearey, 1978, 1992).

[2] The normalization method used in the Atlas of North American English (Labov *et al.*, 2005, p. 39) is sometimes referred to as Labov or ANAE normalization. This normalization method is an implementation of single-parameter log-mean normalization with two small differences relative to the variant outlined in Sec. I B. First, rather than setting all $\bar{G}$ to 0, the $\bar{G}$'s are set to equal some pre-determined reference $\bar{G}$. In the Atlas of North American English, the reference $\bar{G}$ was set to the mean $\bar{G}$ for 345 speakers in the Telsur corpus: 989 Hz (6.897 log-Hz). Second, vowel spaces are normalized directly in Hertz by multiplying formant frequencies by a scale factor, rather than by subtracting in log-Hertz. These differences result in normalized formant values that have realistic Hertz values for an adult speaker, rather than the somewhat opaque values provided by log-mean normalization.

[3] The extra term $C^{\{V\}}$ is at the heart of the problem noted by Disner (1980). In cases of generally similar cross-language or cross listener comparisons, where a limited number of formant-vowel combinations are suspected of important pattern differences, the unmodeled "wobble" in $C^{\{V\}}$ may not be too distorting, since $\bar{G}_s$ estimates depend on many vowels and all formants. Alternatively, a small number of suspicious vowel-formants could be left out of the calculations of $\bar{G}_s$ to avoid possible artefacts. Finally, if additional assumptions can be made about (roughly speaking) the relation of average vocal-tract lengths of the samples of speakers in the language pairs under consideration, the method outlined by Morrison and Nearey (2006) might be beneficial.

[4] Adank *et al.* (2004) attempt to assess indirectly how well "sociolinguistic variation" is preserved using a factorial multivariate analysis of variance involving regional differences as a factor. We find this analysis inferior to the Hindle study for three reasons. First, unlike in Hindle's case, there is no independent evidence (in the form of impressionistic judgments) of specific vowel quality differences for a specific categories. Second, unlike Hindle, Adank *et al.* did not include a correct implementation of single-parameter log-mean normalization (which they call "Nearey2") in the analysis as they erroneously included *f*0 in the calculations of the parameter even though this parameter was only intended to include information related to formant (resonant) frequencies. As a result, only the results presented regarding their "Nearey1" formant-wise log-mean normalization (using an independent log-mean parameter for each formant) are interpretable. Third, Adank *et al.* chose a complex analysis technique, factorial multivariate analysis of variance to assess an unspecified difference in dialect patterns. Adank *et al.* conclude that Lobanov normalization is better than formant-wise log-mean normalization at preserving assumed dialect-related differences on the basis of the relative size of the eta-squared statistic of one selected interaction. However, Adank *et al.* look for preservation of regional differences in normalized data as indexed by the η^2 value for the Vowel-by-Region interaction effect, which the authors take as the only effect probative of dialect differences. Considering only the three formant case, Lobanov, Gerstman, and Nearey normalizations come out in the top 3 in that order on that criterion. However, this analysis ignores the main effect for Region, because "…it seems likely that large [main] effects for Region would only be found if the size and shape of the entire vowel system varies across regions" (p. 3104). However, they ignore the well-known caveat that main-effects and interactions can be interpreted separately only in very specific circumstances. Without a more compelling reason to the contrary, we believe the combined contributions of Region and Region-by-Vowel are more indicative of regional differences in vowel systems. Unfortunately the eta-square values for neither Lobanov nor Gerstman normalization are reported as those effects did not reach significance and were filtered out of their Table V, and so no definitive comparison is possible from the published article. On the basis of the published information, however, there is no reason to believe that "Nearey1" (formant-wise log-mean normalization) performs any worse at distinguishing dialects than the other two normalization methods.

Adank, P., Smits, R., and Van Hout, R. (**2004**). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**, 3099–3107.

Assmann, P. F., Dembling, S., and Nearey, T. M. (**2006**). "Effects of frequency shifts on perceived naturalness and gender information in speech," in *Proceedings of the Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA (September 17–21, 2006), pp. 889–892.

Assmann, P. F., and Katz, W. F. (**2000**). "Time-varying spectral change in the vowels of children and adults," J. Acoust. Soc. Am. **108**, 1856–1866.

Assmann, P. F., and Nearey, T. M. (**2008**). "Identification of frequency-shifted vowels," J. Acoust. Soc. Am. **124**, 3203–3212.

Barreda, S. (**2017**). "An investigation of the systematic use of spectral information in the determination of apparent-talker height," J. Acoust. Soc. Am. **141**, 4781–4792.

Barreda, S., and Nearey, T. M. (**2012**). "The direct and indirect roles of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **131**, 466–477.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (**1996**). "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phon. **24**, 187–208.

Disner, S. F. (**1980**). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am. **67**, 253–261.

Fabricius, A. H., Watt, D., and Johnson, D. E. (**2009**). "A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics," Lang. Var. Change **21**, 413–435.

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," Speech Transm. Lab. Q. Prog. Status Rep. **1**, 22–30.

Fant, G. (**1975**). "Non-uniform vowel normalization," STL-QPSR **16**, 1–19.

Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," J. Speech Lang. Hear. Res. **50**, 1241–1255.

Flynn, N., and Foulkes, P. (**2011**). "Comparing vowel formant normalization methods," in *Proceedings of the 17th International Congress on Phonetic Science*, City University of Hong Kong, Hong Kong, pp. 683–686.

Gelman, A., and Hill, J. (**2006**). *Data Analysis using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, London).

Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electroacoust. **16**, 78–80.

Goldstein, U. G. (**1980**). *An Articulatory Model for the Vocal Tracts of Growing Children* (Massachusetts Institute of Technology, Cambridge, MA).

Hastie, T. J., and Chambers, J. M. (**1992**). "Statistical models," in *Statistical Models in S* (Chapman & Hall/CRC, Boca Raton, FL), Chap. 2, pp. 13–44.

Heffernan, K. (**2010**). "Mumbling is macho: Phonetic distinctiveness in the speech of American radio DJs," Am. Speech **85**, 67–90.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hindle, D. (**1978**). "Approaches to vowel normalization in the study of natural speech," in *Language Variation: Models and Methods* (Academic Press, New York), pp. 161–171.

Keppel, G. (**1991**). *Design and Analysis: A Researcher's Handbook* (Prentice-Hall, Englewood Cliffs, NJ).

Kiefte, M., Nearey, T. M., and Assmann, P. F. (**2012**). "Vowel perception in normal speakers," in *Handbook of Vowels and Vowel Disorders*, edited by M. J. Ball and F. Gibbon (Psychology Press, New York), pp. 160–185.

Kohn, M. E., and Farrington, C. (**2012**). "Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data," J. Acoust. Soc. Am. **131**, 2237–2248.

Labov, W. (**1994**). *Principles of Linguistic Change*, Vol 1: Internal Factors and Vol. 2: Social Factors (Blackwell, Oxford, UK).

Labov, W., Ash, S., and Boberg, C. (**2005**). *The Atlas of North American English: Phonetics, in Phonology and Sound Change* (Walter de Gruyter, Berlin).

Lawson, T., and Persons, A. (**2004**). *The Magic Behind the Voices: A Who's Who of Cartoon Voice Actors* (University Press of Mississippi, Jackson, MS), 404 pp.

Lloyd, R. (**1890**). *Some Researches Into the Nature of Vowel-Sound* (Turner & Dunnett, Liverpool, UK).

Lobanov, B. M. (**1971**). "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am. **49**, 606–608.

Markel, J. D., and Gray, A. H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, Berlin).

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," Psychol. Rev. **118**(2), 219.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Morrison, G. S., and Nearey, T. M. (**2006**). "A cross-language vowel normalisation procedure," Can. Acoust. **34**, 94–95.

Munson, B. (**2007**). “The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation,” Lang. Speech **50**, 125–142.

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

Nearey, T. M. (**1989**). “Static, dynamic, and relational properties in vowel perception,” J. Acoust. Soc. Am. **85**, 2088–2113.

Nearey, T. (**1992**). “Applications of generalized linear modeling to vowel data,” in *Proceedings 1992 International Conference on Spoken Language Processing*, Vol. 1, pp. 583–586.

Nearey, T. M., and Assmann, P. F. (**2007**). “Probabilistic ‘sliding-template’ models for indirect vowel normalization,” in *Experimental Approaches to Phonology* (Oxford University Press, Oxford), pp. 246–269.

Nordström, P.-E., and Lindblom, B. (**1975**). *A Normalization Procedure for Vowel Formant Data* (University, Leeds, UK).

Peterson, G. E., and Barney, H. L. (**1952**). “Control methods used in a study of the vowels,” J. Acoust. Soc. Am. **24**, 175–184.

R Core Team (**2017**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, available at http://www.R-project.org (Last viewed April 2018).

Rubin, D. B., and Little, R. J. (**2002**). *Statistical Analysis with Missing Data* (J. Wiley & Sons, Hoboken, NJ).

Sankoff, D. (**1978**). *Linguistic Variation: Models and Methods* (Academic Press, Waltham, MA).

Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007**). “Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled,” J. Acoust. Soc. Am. **122**, 3628–3639.

Thomas, E. R., and Kendall, T. (**2007**). “NORM: The vowel normalization and plotting suite,” http://ncslaap.lib.ncsu.edu/tools/norm/ (Last viewed April 2018).

Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (**2009**). “A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data,” J. Acoust. Soc. Am. **125**, 2374–2386.

Uschmajew, A. (**2012**). “Local convergence of the alternating least squares algorithm for canonical tensor approximation,” SIAM J. Matrix Anal. Appl. **33**, 639–652.

Uther, M., Knoll, M. A., and Burnham, D. (**2007**). “Do you speak E-NG-L-I-SH? A comparison of foreigner-and infant-directed speech,” Speech Commun. **49**, 2–7.

Venables, W. N., and Ripley, B. D. (**2002**). “Random and mixed effects,” in *Modern Applied Statistics with S* (*Statistics and Computing*) (Springer, New York), pp. 271–300.

Winer, E. (**2012**). *The Audio Expert: Everything You Need to Know About Audio* (CRC Press, New York), 672 pp.