3

TITLE: Perceptual validation of vowel normalization methods for variationist research

6

AUTHOR: Santiago Barreda, University of California, Davis

9 Department of Linguistics, University of California, Davis

12 One Shields Avenue, Davis,
California 95616, USA
530-754-0995

15

SHORT TITLE: Perceptual validation of normalization methods

ABSTRACT:

The evaluation of normalization methods sometimes focuses on the maximization of vowel-space similarity. This focus can lead to the adoption of methods that erase legitimate phonetic variation from our data–that is, overnormalization. First, a production corpus is presented that highlights three types of variation in formant patterns: uniform scaling, non-uniform scaling, and centralization. Then the results of two perceptual experiments are presented, both suggesting that listeners tend to ignore variation according to uniform scaling, while associating non-uniform scaling and centralization with phonetic differences. Overall, results suggest that normalization methods that remove variation not according to uniform scaling can remove legitimate phonetic variation from vowel formant data. As a result, although these methods can provide more similar vowel spaces, they do so by erasing phonetic variation from vowel data that may be socially and linguistically meaningful, including a potential male-female difference in the low vowels in our corpus.

KEYWORDS: normalization, vowel space, variation

INTRODUCTION

In recent years, evaluations of normalization methods for vowel formant data have prioritized the maximization of vowel-space similarity (e.g., Adank, Smits, & Van Hout, 2004; Flynn & Foulkes, 2011). For example, Fabricius, Watt, and Johnson (2009) evaluated methods in part based on "the degree of intersection of individual vowel spaces achieved by the algorithms, because an optimal method would achieve the highest possible degree of overlap" (414). Since the vertices of vowel-space polygons are defined by tokens of individual vowel categories (e.g., /i a u/), in order to make vowel spaces as similar as possible, the vertex vowels must be as close together as possible in the normalized space (i.e., minimize within-category variance). However, production data is typically labelled using broad phonemic, rather than narrow phonetic labels. As a result, there is no reason to expect that all tokens labelled in the same way share identical phonetic properties. For example, Hillenbrand, Getty, Clark, & Wheeler (1995) noted of their own dataset: "It should not be concluded that all utterances that were assigned the same phonemic label are phonetically equivalent […]. Even a casual listening by an experienced phonetician shows clearly that there is a range of phonetic qualities within the vowel categories" (3108).

Problems with desiring maximal similarity

Two general problems arise if we take the position that we only care about obtaining maximally similar normalized data. First, speakers of a single dialect can produce the same phoneme in phonetically different ways based on social or ideological differences, or based on different rates of participation in sound changes in progress (Podesva, D'Onofrio, Hofwegen, & Kim, 2015; Tamminga, 2019). This suggests that we can reasonably expect two speakers in a single household, let alone an entire city or province, to exhibit interesting phonetic and structural differences in their vowel systems. Further, even if two speakers can in some situations produce tokens that are phonetically similar to each other, that does not mean that they always do so, or that they did so when the data was collected. As a result, the phonetic homogeneity of tokens in a dataset can never be known a priori, and can only be established empirically after productions are observed. Therefore, the a priori preference for normalization methods that maximize the similarity of normalized vowel spaces begs the question of the similarity of the vowel spaces, and may erase important phonetic, within-category variation from our data.

A second problem with preferring maximal similarity is that this desire does not offer a clear endpoint. In general, adding more parameters to a model will decrease the residual error and increase the variance explained. Analogously, normalization methods that feature more operations and parameters will generally lead to greater reductions in within-category, subphonemic variation. As a result, it is not surprising that methods with more parameters usually provide the 'best' performance when this is defined primarily in terms of variance reduction. However, if we truly desire methods that maximize the similarity of normalized data, it is not clear when we should stop adding more operations to our normalization methods, removing ever more variation from our data. For example, in addition to controlling for the mean and range of vowels along a formant (the first and second moments), perhaps between-speaker

variation in the skewness (the third moment) of vowel distributions could also be controlled for. If the only goal is maximal similarity, there is no clear basis to rule out additional operations so long as they 'improve' the performance of the algorithm. The limiting case would be a 'saturated model' that completely erases within-category variation, collapsing all productions of each phoneme into a single point in the normalized space. Clearly, the output of such a method would not be useful for most researchers, suggesting that some constraints on the power of normalization methods are necessary in practice.

Phonetic constraints on normalization

Disner (1980) suggested caution regarding the desire for maximal vowel-space similarity:

> "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" (253).

So, rather than wanting all normalized vowel spaces in a sample to be identical, Disner suggested that normalized vowel spaces should be identical if and only if their constituent vowels are phonetically identical. When this occurs, the normalized data will reflect the 'linguistic facts' represented by the productions.

Labov, Ash, & Boberg (2005) justified their use of normalization in the Atlas of North American English by saying that "men, women, and children have very different physical realizations of vowels that sound 'the same' to a listener. The task of normalization is to find a mathematical function that does the same work as the normalizing ear of the listener" (39). This suggests that there should be *perceptual* constraints on normalization methods: the ideal method will not remove *all* within-category variation, but just the variation that is perceptually removed by listeners. From this perspective, it is possible to 'overnormalize' (Barreda & Nearey, 2018) vowel spaces by removing variation that listeners *do not* remove in perception, resulting in the removal of legitimate phonetic variation from a dataset.

Nearey (1983) distinguished two types of variation in formant patterns: phone-preserving variation, and variation that is not phone-preserving. Phone-preserving variation is removed by the 'normalizing ear' of the listener, and so does not affect the phonetic content of a vowel sound. Sounds that differ only in terms of phone-preserving variation are not 'different' from the perspective of the linguistic system: they will convey the same linguistic and social information between speakers despite being acoustically different. Barreda and Nearey (2018) argued that the ideal normalization method removes only linguistically meaningless, phone-preserving variation, regardless of its source, leaving only variability associated with differences in the phonetic properties of vowel sounds. In this view, we should sometimes accept normalization methods that result in more within-category variance (and more vowel space heterogeneity) when this variation represents legitimate phonetic variation in a set of productions.

Phone-preserving variation in formant patterns

There are several theories of vowel perception suggesting that uniform scaling of formant patterns is phone preserving, and substantial experimental evidence that this type of variation in formant patterns tends to preserve phonetic structure (Barreda, 2020; Nearey, 1978; Smith & Patterson, 2005). When formant patterns vary according to uniform scaling, all formants increase in equal proportion, on average, between speakers. This sort of acoustic variation is expected when speakers vary strictly in terms of vocal-tract length, and is often interpreted by listeners as indicating a difference in speaker size rather than a difference in the phonetic content of the signal (Smith & Patterson, 2005). To my knowledge there is no theory of speech perception that suggests that uniform scaling is *not* generally phone preserving. In fact, it is typically *deviations* from uniform scaling that are used to convey linguistic and social meaning between speakers (see Barreda [2020] for an elaboration of this idea).

We may compare this to two other sorts of manipulations of formant patterns: non-uniform scaling and dispersion/centralization. Scaling formant frequencies using a different factor for each formant (i.e. non-uniform scaling) is generally not considered to be phone-preserving. Listeners are quite sensitive to the independent manipulation of formant frequencies, often associating relatively small changes in individual formants with large changes in vowel quality. Further, differences in the position of individual phonemes in the $F1 \times F2$ plane due to dialectal differences (or any other factor) will necessarily feature non-uniform scaling differences between speakers. For example, if average productions of /u/ in two dialects have approximately the same height but differ in frontness, this implies an average difference in F2 (but not F1) in the productions of /u/ between the dialects. As a result, both synchronic variation and diachronic changes in vowel systems may manifest as non-uniform differences in individual formants, and potential differences in vowel-space shapes.

Finally, we consider the changes in formant patterns associated with variation in the dispersion/centralization of vowel spaces between speakers. Differences in the centralization of vowels can be quantified using the distance of constituent phonemes to some internal reference point (e.g., the vowel-space centroid), and results in the expansion/contraction of a fixed vowel-space shape. To my knowledge there is no theory of vowel perception that suggests that vowel-space dispersion *cannot* or generally does not affect phonetic content. In fact, vowel-space dispersion relates to the 'clarity' of speech (Ferguson & Kewley-Port, 2007), can be affected by phonological and lexical factors (Munson, 2007), and potentially conveys linguistic and social meaning between speakers and listeners (e.g., D'Onofrio, Pratt, & Van Hofwegen, 2019).

NORMALIZATION OPERATIONS AND PHONE-PRESERVING VARIATION

If we are interested in determining which normalization methods tend to remove *only* phone-preserving variation, we must consider two questions. First, which kinds of variation are likely to be phone preserving? Second, which kinds of variation are erased by different normalization methods? To investigate these questions, we will outline the behavior of three classes of normalization methods, to be described below: 1) single parameter scaling methods, 2) formantwise scaling methods, and 3) formantwise standardization methods.

The test corpus

The behavior of normalization methods will be highlighted using a corpus of production data collected from 30 speakers of California English (14 males and 16 females), ranging in age from 18 to 25 (mean = 20.3, standard deviation = 1.7). All speakers lived in California since at least 5 years of age and reported English as their strongest language. This data comprises 15 repetitions of 11 English vowel phonemes (/ɪ i ʊ u e ɪ ɛ ʌ æ ɑ o ɝ/) in an /hVd/ context. Words were collected in a sound attenuated booth, in a single 30-minute session. Productions were collected in isolation, using single-word prompts presented on a computer monitor in a random order, but blocked by repetition. The first three formant frequencies were measured for each token by averaging measurements from 20-40% of the vowel duration, sampled every 3 ms. The high number of repetitions of each token and the phonologically and acoustically controlled conditions were intended to provide information about the idiosyncratic, but consistent, between-speaker variation present in the sample.

Single parameter scaling methods

The simplest normalization methods we will consider involve the division of all formant frequencies by a single, speaker-specific scaling parameter (Labov et al., 2005; Nearey, 1978). Division by a single parameter means these methods can only erase differences in vowel-space expansion or contraction that is uniform across all formants. This constraint imposes two important limitations on the sort of vowel-space variation these methods can erase from data. First, uniform scaling cannot affect the basic 'shape' of vowel spaces as defined by polygons in the $F1 \times F2$ space. This means that, for example, these methods cannot equate a vowel space in the shape of an equilateral triangle and a vowel space in the shape of an isosceles triangle. Second, uniform scaling can only equate differences in vowel-space dispersion (area) that are predictable from differences in average formant frequency. For example, a speaker who produces formants that are 15% higher also produces vowel phonemes that are 15% further apart in the formant space (since all distances have increased by 15%), leading to a predictable increase in vowel-space area. All of the methods we will consider here can remove variation according to uniform scaling, providing roughly equivalent outputs when vowel spaces differ mostly in this manner (see Figure 1).

$$\sigma_s = \sum_{v=1}^{V} \sum_{j=1}^{J} \ln\left(F_{vjs}\right) \; / \; (V * J) \quad (1)$$

$$N_{vjs} = \ln\left(F_{vjs}\right) - \sigma_s \quad (2)$$

$$\exp\left(N_{vjs}\right) = F_{vjs} / \exp\left(\sigma_s\right) \quad (3)$$

A representative of this class of methods is the single-parameter log-mean normalization method first proposed by Nearey (1978), henceforth LM. This method finds the mean log-transformed formant frequency for each speaker (*s*) across all *V* vowels and *J* formants (Equation 1). This single value is then subtracted from the logarithm of each observed formant frequency (Equation 2). This method is mathematically equivalent to calculating the geometric mean formant frequency and then dividing formant frequencies by this value (Labov et al., 2005), as

shown in Equation 3. The estimated parameter is represented by σ to underscore the fact that this is a scaling parameter: it affects normalized data only by affecting the scaling of formant patterns in a multiplicative manner.

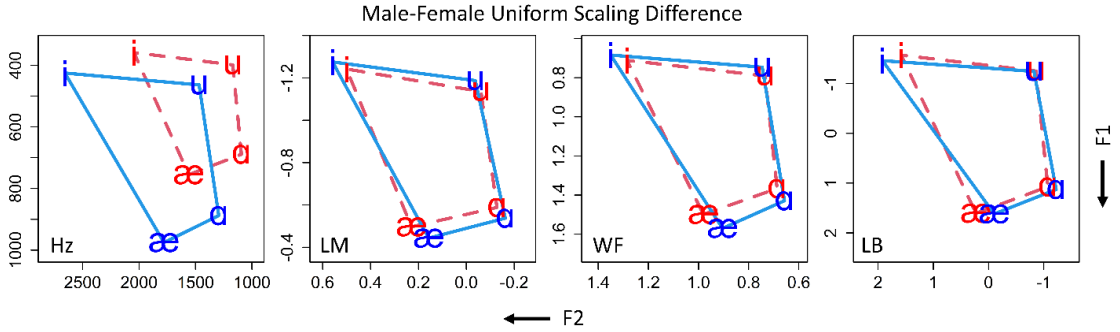


FIGURE 1. Vowel spaces of one male and one female speaker who differ primarily according to uniform scaling, presented in Hertz and normalized using the log-mean (LM), Watt and Fabricius (WF) and Lobanov (LB) methods.

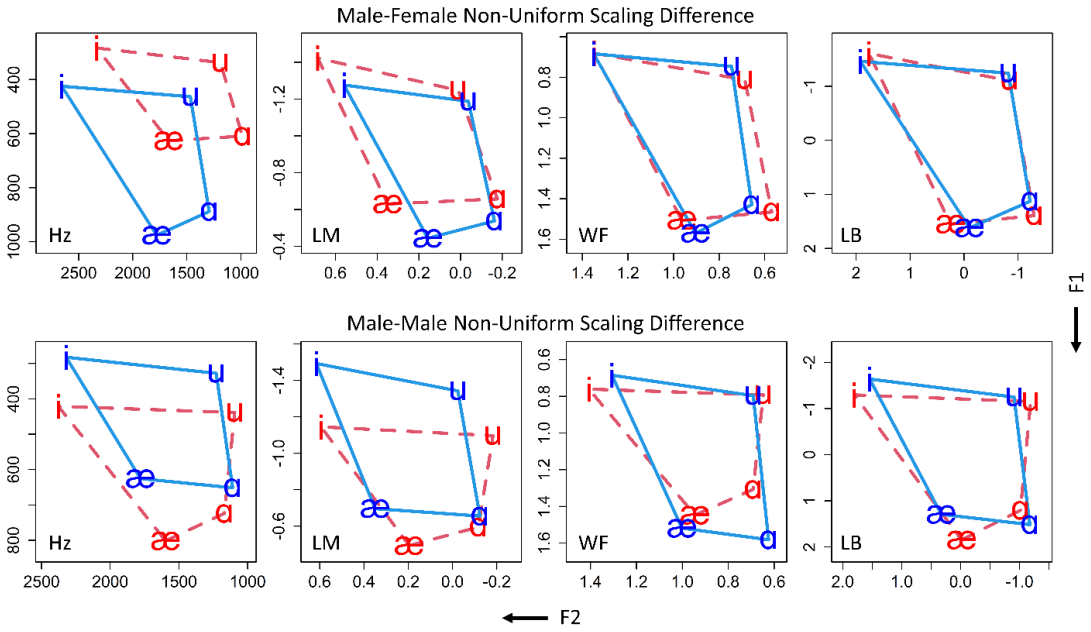
Formantwise scaling methods

Rather than using a single parameter for all formants, formantwise scaling methods use an independent scaling parameter to normalize each formant (Nearey, 1978; Watt & Fabricius, 2002). The method proposed by Watt and Fabricius (2002), henceforth WF, will be used to represent this class of methods. The WF method calculates a scaling parameter for each speaker (s) for each formant (j), as in Equation 4. Formant frequencies are then divided by the formant-specific scaling parameters (σ_{js}), as in Equation 5. Note that unlike in Equation 2, the scaling parameters feature a formant-specific subscript so that values of σ_{js} will differ across the J formants.

$$(4)\ \sigma_{js} = (F_{js}^{/i/} + F_{js}^{/a/} + F_{js}^{/u/})/3$$

$$(5)\ N_{vjs} = F_{vjs}/\sigma_{js}$$

Although formantwise scaling methods differ in how they define their scaling parameters (σ_{js}), the use of an independent scaling parameter for each formant means that these methods will erase 'non-uniform' *shape* variation in vowel spaces. As a result, these methods *can* relate a vowel space that looks like an equilateral triangle and one that looks like an isosceles triangle, potentially making them identical after normalization. For example, although the female speaker in Figure 2 has a larger F1 and F2 range than the male speaker, her F1 is larger than expected given her F2 range, resulting in differences in vowel-space shape between the speakers. As seen in Figure 2, these differences can be reduced by formantwise scaling methods such as WF, but not by single parameter scaling methods such as LM.



231 FIGURE 2. The vowel spaces of one male and one female speaker (top row), and two males speakers (bottom row) who differ in non-uniform scaling. Vowel spaces are presented in Hertz and normalized using the log-mean (LM), Watt and Fabricius (WF) and Lobanov (LB) methods.

234

As noted above, non-uniform differences between speakers can potentially convey important linguistic information. For example, the male speakers in Figure 2 produce
237 approximately the same F2 values but differ in average F1 across all tokens. It is possible that these differences represent anatomic variation that should be erased by a useful normalization method. However, it is also possible that these represent legitimate phonetic, dialectal variation
240 that should be maintained in our data. Unfortunately, formantwise scaling methods do not make any such distinction and erase such variation indiscriminately.

Formantwise standardization methods

243 The most complex class of methods we will consider control for dispersion and central location independently for each formant (Gerstman, 1968; Hindle, 1978; Lobanov, 1971). This last class of methods features two operations per formant: division and subtraction. Formantwise
246 standardization methods will be represented by the method proposed in Lobanov (1971). To use the Lobanov (henceforth LB) method, researchers calculate the mean (Equation 6) and standard deviation (Equation 7) independently for each of *J* formants, across all the *V* vowels produced by
249 a speaker. Formant frequencies are then standardized by subtracting the formantwise mean for that formant and dividing by the standard deviation for that formant (Equation 8).

252 (6) $\mu_{js} = \sum_{\mathrm{v}=1}^{V} F_{vjs} \,/\, \mathrm{V}$

(7) $\sigma_{js} = \sqrt{\sum_{\mathrm{v}=1}^{V} \left(F_{vjs} - \mu_{js}\right)^2 / \mathrm{V}}$

(8) $N_{vjs} = (F_{vjs} - \mu_{js})/\sigma_{js}$

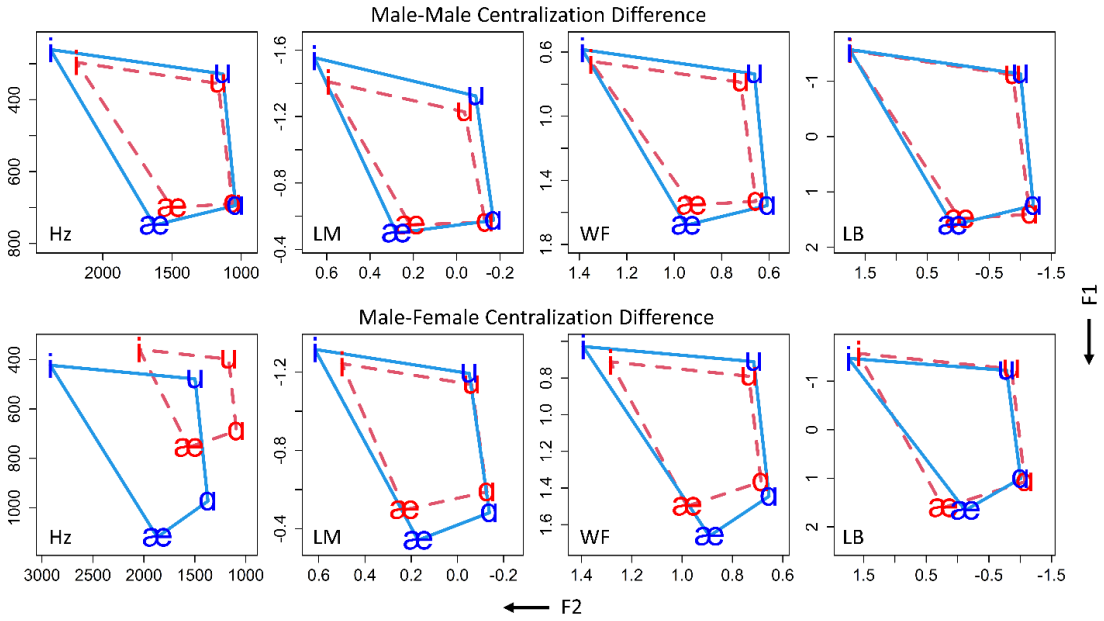


FIGURE 3. The vowel spaces of two male speakers (top row), and one male and one female speaker (bottom row) who differ in centralization. Vowel spaces are presented in Hertz and normalized using the log-mean (LM), Watt and Fabricius (WF) and Lobanov (LB) methods.

Since formantwise standardization methods independently control for both the mean and range of each formant, these methods will erase differences in formant dispersion (hyper/hyporarticulation or centralization) between speakers. For example, in Figure 3 we see two pairs of speakers who differ in vowel-space dispersion above and beyond their difference in average formants. As noted above, there is evidence to suggest that these differences may be phonetically salient and linguistically meaningful to listeners. In both cases the differences are erased by the LB method but maintained by the WF and LM methods.

Clopper, Pisoni, & de Jong (2005) reported that the LB method resulted in 'artifacts' in their data consistent with overnormalization: "the back-vowel fronting that is found in the speech of Southern talkers led to a higher mean F2 and a smaller F2 standard deviation. The [Lobanov] z-score transformation then produced artificially backed low back vowels as a result of the larger numerator and smaller denominator" (1674). In this case, the normalized data produced by the LB method placed back vowels from different dialects in the same location in the normalized space despite differences in their phonetic properties (i.e., overnormalization). Clopper et al. (2005) went on to state that, "in cases where vowel systems are being compared that differ in their overall shape, the *z*-score transform should be used with caution" (1674). However, this advice is problematic for the LB method as there may *always* be differences in vowel-space shapes in a sample, and normalization may be desirable precisely to establish the existence of such differences.

NORMALIZED SPACES AS PHONETIC MAPS

Lobanov normalization (LB) was intended to maximize the accurate statistical classification of speech sounds, a task for which it is demonstrably well-suited (Adank et al., 2004). The Watt and Fabricius (WF) method was designed to aid the visual comparison of vowel spaces of speakers of the same dialect (Watt & Fabricius, 2002:169), explicitly in cases where vowel spaces of roughly the same shapes are expected. Generally speaking, methods are neither good nor bad, but are instead suitable for specific purposes. Neither the LB nor the WF method was meant to preserve or relate information regarding the perceived vowel qualities (i.e., the *phonetic* properties) of a set of vowel sounds in a broad range of circumstances.

Plots of normalized vowels (and the data used to generate them) are often treated as if they provided 'phonetic maps', broadly analogous to geographic maps. Geographic maps can often be interpreted as metric spaces where the Euclidean 'straight line' distance in the map reflects the geographic distance between real-world locations. If a user of the map sees that points A and B lie closer together than points A and C, they may infer that A and B are closer together in real life. The consistent relationships between distances on the map and the geographic distances of real-world locations makes the map useful, as this allows the user to make reliable inferences based on the map.

When researchers use normalized vowel spaces as phonetic maps, the Euclidean distance (or some other distance metric) between two points is used to quantify the phonetic differentness of two tokens. A researcher may infer that vowels that cluster together in one location of a normalized space have similar phonetic properties because the distances between them are small, and that vowels in distant locations have different phonetic properties because the distances between them are large. Such inferential practices rely on there being consistent relationships between distance in the normalized space and the phonetic 'distance' of speech sounds. Thus, researchers engaging in these practices will benefit from selecting methods that provide reliable phonetic organizations ('phonetic maps') for their data.

Overnormalization and the reliability of phonetic maps

An example of how overnormalization can harm the reliability of 'phonetic maps' is presented here. Suppose we are interested in whether the men and women in our sample of California speakers have 'the same' vowel spaces or not. The men and women in our sample exhibit a non-uniform difference in their formant patterns. While the female speakers have an F1 mean for /i æ u/ that is 37% higher than the male speakers, their F2 mean is only 21% higher, meaning their F1 frequencies have increased 13% more than expected according to uniform scaling. This scaling difference results in variation in the shape of the male and female vowel spaces, with the female vowel space being relatively more elongated along F1 (seen in Figure 4a). The question is, is this difference phonetically, linguistically, or socially meaningful? Or is it simply meaningless between-speaker variation that we should erase from our data?

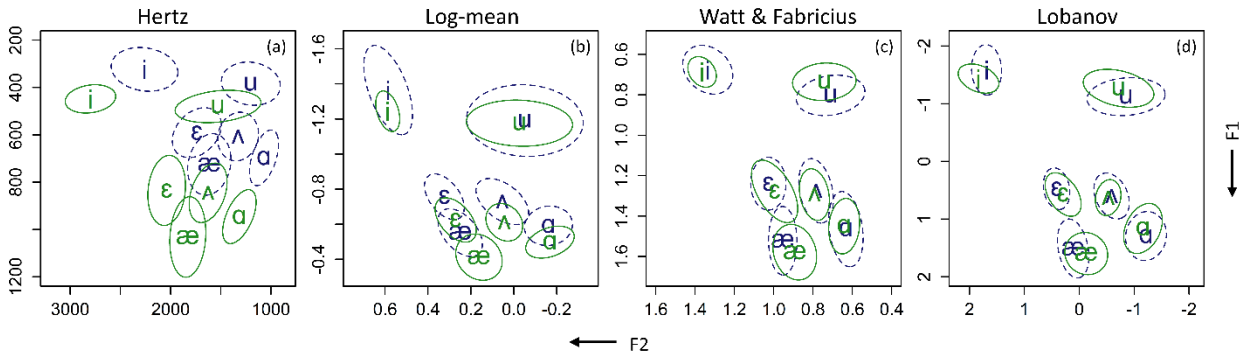


FIGURE 4. Mean productions of a subset of vowels produced by 30 male (dashed line) and female (solid line) speakers of California English, presented in Hertz and normalized using three methods. Ellipses enclose two standard deviations.

If relying on the organization provided in Figure 4b, the researcher would likely infer from the distance between the tokens that productions of /æ/ differ in height between the male and female speakers, meaning these vowels would be able to convey social and linguistic differences. For example, if representative tokens were played for a room full of linguists (e.g., at a conference), the audience would be expected to 'hear' a difference in height between the vowels and react accordingly. On the other hand, based on the proximity between the tokens, the researcher would conclude that /u/ does not differ in height between men and women. When played for the same room full of linguists, this researcher would expect the audience to think that these vowels did not differ substantially in openness. Now, suppose that the organization presented in Figure 4b represented the true linguistic facts in the data (i.e., /æ/ differs but /u/ does not). In this case, a researcher could not make reliable inferences about phonetic characteristics using the representation in Figure 4d because distance in the normalized space would not be a reliable metric for phonetic difference. In some cases, a lack of distance would represent phonetic similarity (/u/), while in others it would not (/æ/).

If a researcher is interested in using distances in normalized data to make inferences about the phonetic similarity of tokens, they must ensure that distance in the normalized space functions as a reliable metric for phonetic difference. The only way to ensure this is to favor normalization methods whose outputs accurately reflect the phonetic information in vowel sounds. Such methods will cluster sounds that are phonetically similar, and separate sounds that are phonetically different, even if this results in relatively more within-category variation in normalized data.

PERCEPTUAL VALIDATION OF NORMALIZATION METHODS

There are not many published evaluations of normalization methods that directly consider how well these maintain the phonetic properties of speech sounds. Both Hindle (1978) and Labov (1994) found that Sankoff (1978) normalization (a formantwise standardization method) reduced more within-category variation in formant patterns than the LM method. However, it also removed some of the socially conditioned variation in vowel quality between speakers (i.e.,

overnormalization). Kohn and Farrington (2012) found a slight advantage for WF over LB normalization in the maintenance of perceptually salient sociolinguistic differences, though the structure of their statistical models makes their results somewhat difficult to interpret for our purposes. Kohn and Farrington did not test the LM method because it was found to perform poorly by Adank et al. (2004). However, the poor performance of the LM method in Adank et al. is likely due to an error in the implementation of the method, which included f0 in the calculation of the formant-scaling parameter (3103, Eq. 8). This would have caused large deviations in speakers' formant-scaling estimates that were not reflective of changes in vocal-tract length and could be extremely deleterious to the performance of the algorithm. More recently, Rankinen & de Jong (2020) report a comparison of the LB and LM methods, finding that the LM method better preserves the vowel-space differences associated with ethnic heritage in their sample.

Although there has not been much direct perceptual validation of normalized outputs, the operations employed by the normalization methods have different levels of support as 'phone preserving' in the literature on speech perception. As noted in the introduction, there is general agreement that uniform scaling tends to be phone preserving. What is less certain is which additional sorts of between-speaker variation can be phone preserving, and under what conditions. Here, we seek evidence that the additional operations employed by more complex normalization methods (e.g., formantwise scaling, formantwise standardization) are supported by listener judgments.

We wish to investigate gradient changes in the sub-phonemic, phonetic properties of vowel sounds. One way to do this is by considering the varying classification rates of ambiguous vowel tokens into two or more phonemic categories (Nearey, 1998). For example, consider a continuum spanning between two vowels varying primarily in height (e.g., from /i/ to /ɪ/). Imagine we begin with a good exemplar of /i/ that we expect to be classified as /i/ nearly 100% of the time. As F1 increases, the phonetic 'height' of the vowel will decrease, together with the probability that it will be identified as /i/. Thus, the increasing probability of an /ɪ/ response reflects gradual changes in the phonetic properties of the vowel, providing insight into gradient variation in the phonetic properties of vowel sounds: acoustically different vowels with similar classifications are likely to share phonetic properties.

Classification rates will be used to investigate which sorts of transformations to formant patterns are 'phone preserving', meaning they can be removed from our data without erasing phonetic information. Rather than preferring normalization methods that provide maximal normalized similarity, we will prefer methods that cluster vowel sounds that are classified in the same ways, and separate vowel sounds that are classified in different ways. If a normalization method clusters vowels that are classified in substantially different ways, this will be suggestive of overnormalization, the artificial grouping of phonetically dissimilar vowels due to the removal of legitimate phonetic variation from vowel data.

Simulating between-speaker differences

The experiments described below feature vowels produced by six artificial speakers. These speakers share dialectal information but feature systematic between-speaker variation in the realization of their tokens. Each speaker was represented by two types of tokens, training stimuli

and testing stimuli. Training stimuli consisted of the vowels /i u æ ɑ o/. These vowels were intended to provide information about speaker vowel spaces, and listeners were not asked to classify these vowels. The testing stimuli were a seven-step continuum from the average F1, F2, and F3, of /i/ to /ɪ/ in six equal steps (Figure 5a), for each unique voice (Figure 5d). Formant frequencies for each phoneme were based on average values calculated from our sample of California speakers, a subset of which was presented in Figure 4.

Between-speaker variation existed along two dimensions: vowel-space type (standard, high F1, and centralized), and size (large, medium and small). Size differences were implemented by modifying all formant frequencies in equal proportions (uniform scaling, see Figure 5b), and by changing the fundamental frequency (f0) of the vowels. Mean F1, F2, and F3 frequencies across all male and female speakers were decreased by 15% to create the formant values for the large standard speaker. Large speakers had F4 values of 3500 Hz, F5 values of 4500 Hz for all vowels, and f0s that decreased linearly from 120 to 110 Hz during the vowel. Medium voices were created by increasing all large formant frequencies (F1 to F5) by 14%, and by increasing f0 so that it went from 170 Hz to 156 Hz, an increase of a half-octave over the large condition. Small voices were created by increasing all medium formant frequencies (F1 to F5) by a further 14% (30% relative to the large condition), and by increasing f0 so that it began at 240 Hz and decreased to 220 Hz over the course of the vowel. This is an increase of a half-octave over the medium condition, and an octave over the large condition.

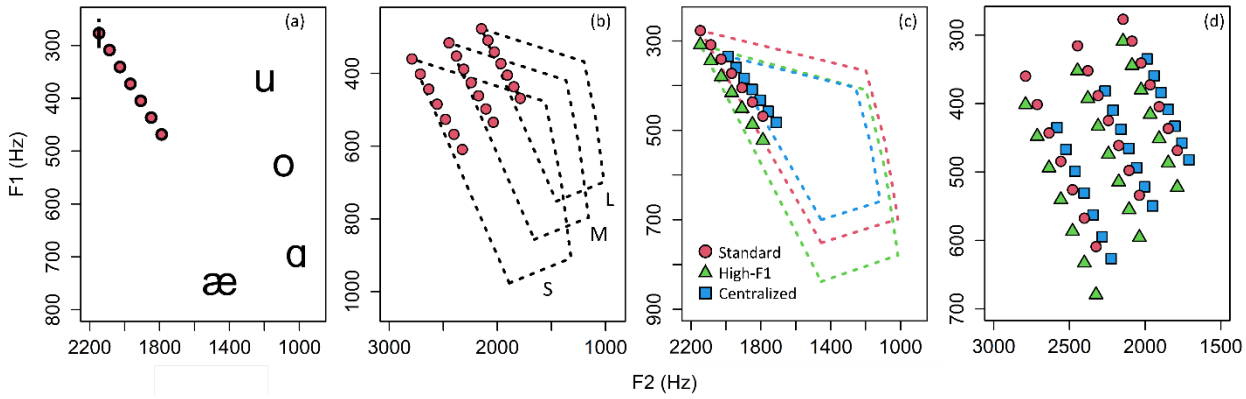


FIGURE 5. (a) Locations of training and testing stimuli for the standard voice. Points indicate the steps along the /i/-/ɪ/ continuum. (b) Comparison of testing stimuli for the large (L), medium (M) and small (S) standard speaker. Smaller speakers produce higher formant-frequencies. Polygons outline vowel spaces implied by training vowels. (c) Comparison of testing stimuli for the large voices by voice type. (d) Comparison of all testing stimuli for standard (circle), high-F1 (triangle), and centralized (square) voice types across large, medium, and small speakers.

Type differences were implemented by varying voices in ways that deviated from uniform scaling, within each size level (Figure 5c). The high-F1 speaker was created by increasing the standard speaker's F1 frequencies by 11.5% relative to the standard speaker, but not modifying any other formant frequencies. This resulted in a non-uniform scaling difference

with respect to the standard speaker. The centralized speaker was created by centering formant frequencies about their mean values, multiplying centered values by 0.85, and then adding the mean values back again. This resulted in a vowel space with the same centroid as the standard speaker but with reduced vowel space dispersion (i.e. centralization). Formant values for the large speaker are provided in Table 1, and all testing vowels for all voices are compared in Figure 5d. All vowels were 250 ms in duration with steady-state formants, and were synthesized using a Klatt-style parametric synthesizer (Klatt, 1980).

TABLE 1. *Formant frequencies for large-speaker stimuli. Tokens whose vowel labels are numbers are steps along the /i/-/ɪ/ continua for the voices, with /i/ being the first step*

Standard			
F1	F2	F3	Vowel
751	1454	2268	æ
698	1015	2272	ɑ
527	1089	2202	o
367	1198	2129	u
277	2147	2648	1
309	2087	2613	2
341	2027	2577	3
373	1967	2542	4
405	1908	2507	5
437	1848	2471	6
468	1788	2436	7

High F1			
F1	F2	F3	Vowel
838	1454	2268	æ
779	1015	2272	ɑ
588	1089	2202	o
410	1198	2129	u
277	2147	2648	1
309	2087	2613	2
341	2027	2577	3
373	1967	2542	4
405	1908	2507	5
437	1848	2471	6
468	1788	2436	7

Centralized			
F1	F2	F3	Vowel
700	1454	2267	æ
659	1117	2270	ɑ
527	1174	2216	o
404	1257	2160	u
277	2147	2648	1
309	2087	2613	2
341	2027	2577	3
373	1967	2542	4
405	1908	2507	5
437	1848	2471	6
468	1788	2436	7

Differential predictions made by normalization methods

Consider a researcher who is interested in the phonetic properties of the /i/-/ɪ/ continuum as produced by our artificial speakers. We may ask a very basic question: which, if any, of the tokens in Figure 6a share phonetic properties? Since the speakers differ substantially in their acoustics, the researcher cannot compare tokens directly by using formant frequencies measured in Hertz (i.e., their positions in 6a). A typical approach would be to normalize the continuum steps for each speaker, and then compare the normalized data. Figure 6 presents the organization of the continuum steps produced by the six voices in normalized spaces. In each case, normalization was carried out using statistics calculated from the peripheral vowels (/i u æ ɑ o/) for each voice, since these provide information about vowel space shape and size.

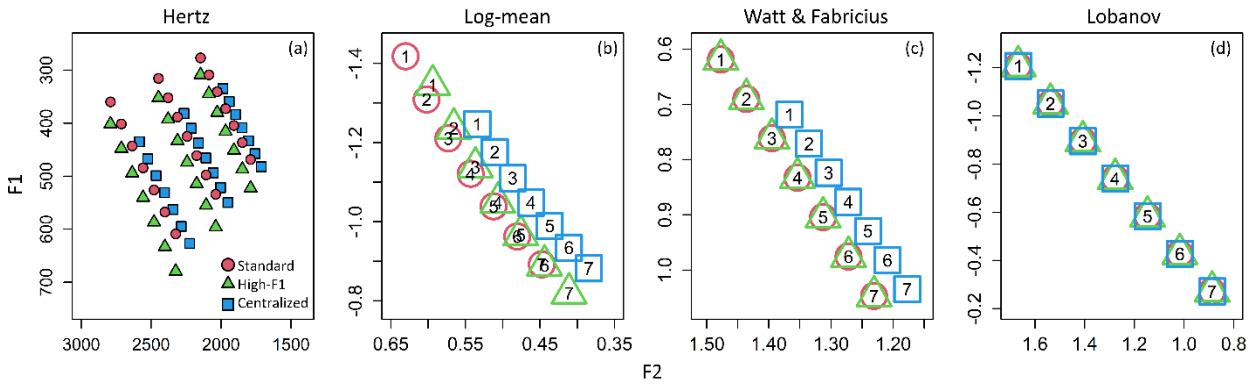


FIGURE 6. (a) Testing vowels plotted according to F1 and F2 values in Hertz. The same vowels are presented when normalized according different normalization methods (b-d).

As noted above, normalized spaces are often treated as metric spaces where Euclidean distance is expected to relate to phonetic differentness. Therefore, by proposing different organizations for the vowel sounds produced by our artificial speakers, the normalization methods effectively suggest different phonetic properties for the vowels. The LM method erases variation according to uniform scaling, so it will erase differences in formant patterns associated with the size manipulation. However, all differences in type (standard, high-F1, centralized) are predicted to result in phonetic differences (Figure 6b). The WF method can erase variation according to non-uniform scaling, and thus it will equate the vowel spaces (and vowels) of the standard and high-F1 speakers, in addition to removing size differences (Figure 6c). As a result, this approach suggests that the differences in F1 range used by the standard and high-F1 speakers will not be 'heard' by listeners, and will not result in phonetic differences. Finally, the LB method will erase all differences in size and type, equating the vowels produced by all six speakers under consideration (Figure 6d). Although this method results in the tightest clustering for each continuum step (i.e., the least 'within-category' variance), it also predicts a complete lack of between-speaker phonetic variation in the productions of the artificial speakers.

EXPERIMENT 1

In Experiment 1, listeners were presented with vowels from the /i/-/ɪ/ continuum as produced by the synthetic speakers described above, in isolation and randomized by speaker. Since listeners were asked to classify isolated vowels, they did not have information about the speaker's vowel-space dispersion along F1 or F2 when classifying tokens. As a result, it is not expected that listeners could perform perceptual operations analogous to those required by WF and LB normalization in this listening situation. Thus, this experiment is meant to verify that in the absence of knowledge about a speaker's vowel system listeners will: 1) associate vowels relatable by uniform scaling (size differences) with similar phonetic properties, and 2) will 'hear' differences that deviate from uniform scaling, associating these with phonetic differences.

Listeners

Listeners were 32 Native speakers of California English (8 men, 24 women). All listeners lived in California since at least two years of age and indicated that English is their strongest language. Listeners ranged in age from 18 to 35 years, with a mean of 20 and a standard deviation of 3 years. All listeners were students at the University of California, Davis, who participated in 30-minute experimental sessions in exchange for partial course credit.

Stimuli

Stimuli consisted of the seven-step testing /i/-/ɪ/ continuum as produced by the standard, high-F1, and centralized voices for large and small speakers. The experiment consisted of 42 unique vowel sounds (3 voice types × 2 sizes × 7 vowels per voice). A description of the stimuli is provided in Figure 5 and Table 1.

Procedure

Listeners were presented with stimuli over headphones in a sound-attenuated booth. Stimuli were presented one at a time, randomized across all stimulus dimensions but blocked by repetition. Listeners responded on a graphical user interface with three buttons that read ‘Heed’, ‘Hid’, ‘Head’. Listeners were asked to click on the word “containing the vowel that 'sounds' most like the sound you hear”. Listeners were presented with each stimulus up to 6 times for a total maximum of 252 responses per participant. All but 3 subjects completed the full experiment, with the fewest responses per subject being 204.

Results and discussion

Figure 7a presents the results of experiment 1. The comparisons presented in Figure 7b suggest that size differences barely affect classification functions, while voice-type differences have a large effect. To investigate this, the following bootstrap analysis was carried out. For each bootstrap sample (10,000 total samples), the following process was carried out. Data was selected from 32 subjects with repetition at the subject level. Classification rates for each stimulus into each of the response categories was then calculated. The coefficient of determination (squared correlation, R^2) between classification rates (for all categories) across panels in Figure 7a was then found and recorded. This value allows us to quantify the similarity of classifications (and phonetic properties) across different formant-pattern manipulations. For example, the correlation between the functions in the top and bottom row of the leftmost column in Figure 7a reveal the similarity of classifications across size manipulations for the standard voice. In addition, R^2 was calculated for the classification functions of each stimulus across successive random samples in order to get an estimate of the sampling error for the classification functions.

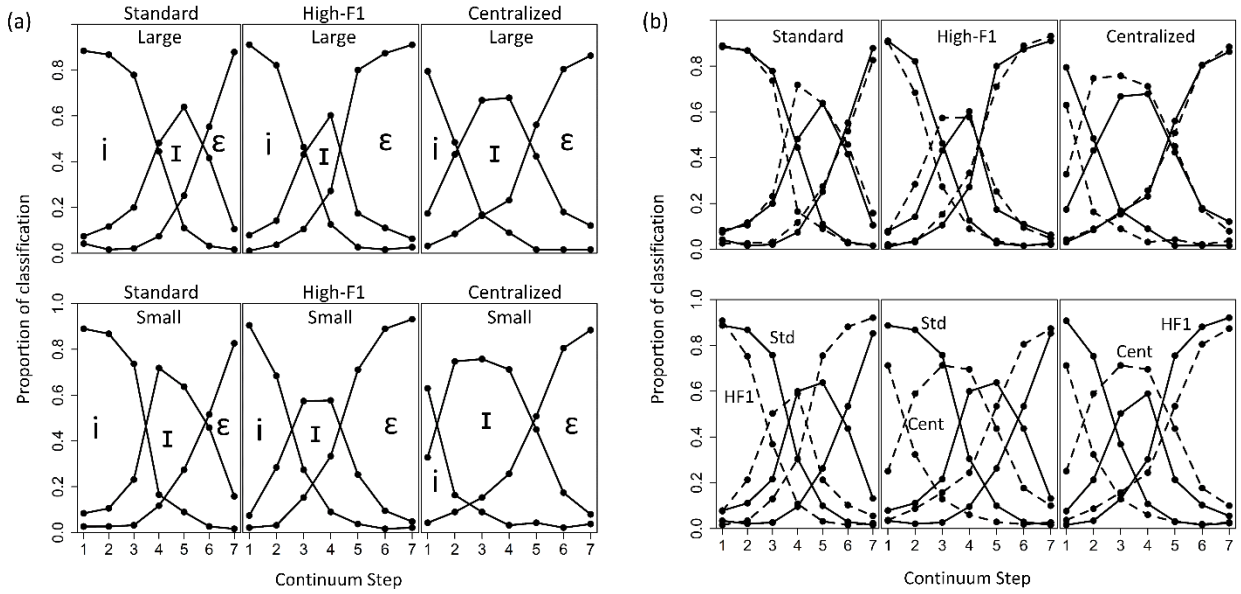


FIGURE 7. (a) Proportion of classifications into response categories for each continuum step, by voice type and size. (b) The top row contrasts classifications across size for each voice type (small voices in broken lines). The bottom row contrasts average classification rates for standard (Std), high-F1 (HF1) and centralized (Cent) voices, averaged across sizes.

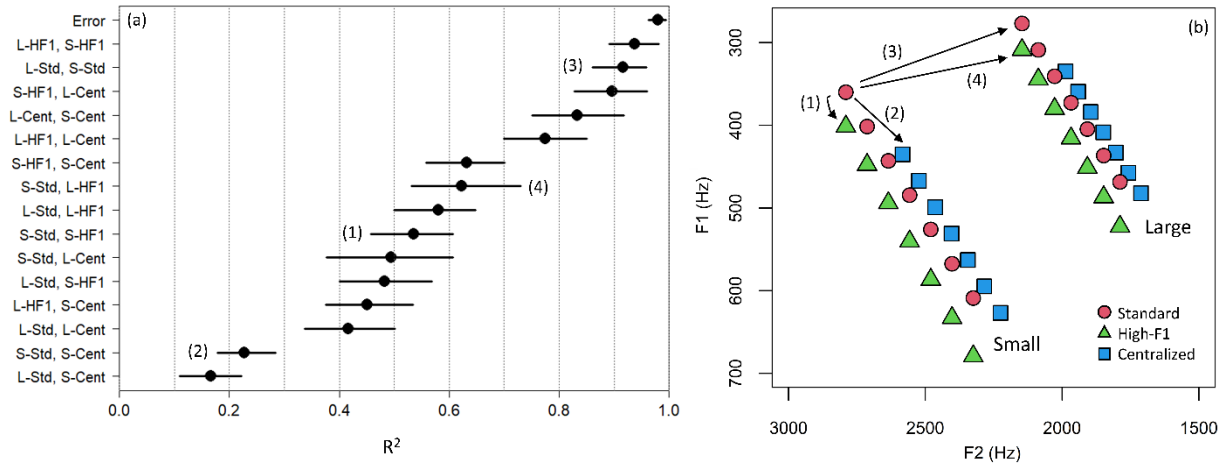


FIGURE 8. (a) Distribution of squared correlation (R^2) between classification functions for pairs of voices in the bootstrap analysis. Lines indicate 95% highest-density intervals, points indicate means. Numbers indicate corresponding lines in the right panel. (b) Testing continua used in this experiment. Lines indicate specific differences highlighted in the left panel. For example, although line (1) indicates a small acoustic difference, the members of these continua are phonetically dissimilar. In contrast, line (3) compares continua that are acoustically dissimilar yet phonetically similar.

Figure 8a presents the distribution of estimated R^2 values for all 15 unique voice comparisons, and the estimate of the sampling error. In general, size variation tends to result in small changes in classification functions that are not much larger than the sampling error, while type variation can result in very large differences in classification rates. Figure 8b highlights the discrepancy that can exist between the acoustic and phonetic properties of vowel sounds. For example, the average R^2 between the large and small standard voices was 0.93, while the R^2 between the standard and high-F1 small voices was 0.54. This means that an increase of 11.5% to F1, with all other formants and f0 held constant, caused a substantially greater phonetic change than a 30% shift to all formant frequencies, combined with an octave difference in f0.

Results indicate that when little is known about the speaker (e.g., in mixed-speaker listening conditions, when playing an isolated example at a conference) data normalized using a single parameter scaling method (e.g., LM) is most likely to reflect the phonetic organization of vowel sounds. In these situations, normalization methods that perform formantwise scaling or standardization (e.g., WF and LB) can overnormalize vowels, eraseing perceptible phonetic variation from vowel formant data.

EXPERIMENT 2

Listeners may exhibit behaviors that justify the use of formantwise scaling and standardization methods in situations where they have more information about a speaker's vowel system. In the second experiment vowels produced by different synthetic voices are presented in blocked conditions, after familiarization with a speaker's vowel space.

Participants

Listeners were 49 Native speakers of California English (29 men, 20 women). All listeners lived in California from at least six years of age and indicate that English is their strongest language. Listeners ranged in age from 18 to 22 years, with a mean of 19 and a standard deviation of 1.2 years. All listeners were students at the University of California, Davis, who participated in 30-minute experimental sessions in exchange for partial course credit.

Stimuli

Stimuli consisted of the vowels /i u æ ɑ o/, and the seven step /i/-/ɪ/ continuum produced by the large standard voice, and the medium and small high-F1 and centralized voices. Stimulus information is presented in Table 1 and in Figure 5. The vowels /i u æ ɑ o/ served as training stimuli, which were meant to provide listeners with information about the location and dispersion of the speaker's vowel space. The seven step /i/-/ɪ/ continuum were the testing stimuli, the vowels sounds that listeners would be asked to classify during the experiment.

Procedure

Each listener heard vowels produced by three speakers: a single speaker for each voice type, each paired with a different voice size. All listeners heard the standard voice paired with the large size. Half of listeners heard a medium high-F1 voice and a small centralized voice, and the other half heard the opposite combination: the medium centralized voice and the small high-F1 voice. So, each listener heard three different speakers, and these speakers differed in their voice

size *and* in their voice type.

The experiment consisted of five rounds. Round 1 was an initial mixed-speaker condition where listeners were presented with testing stimuli (7-step /i/-/ɪ/ continuum) for each of the three voices 5 times each, blocked by repetition. Listeners were not given information about how many voices they would be hearing, when any given speaker was speaking, or what any of these speakers sounded like. In all rounds, classification was carried out using the same instructions and general procedure as in Experiment 1.

Rounds 2-4 were blocked-speaker rounds. First, listeners were informed that they would listen to some vowels produced by a single speaker, and that they would then be asked to classify vowels produced by that same speaker. Listeners heard the training vowels (/i u æ ɑ o/) repeated 5 times each, blocked by repetition, with 500 ms of silence in between each vowel (25 total sounds). After this, listeners were asked to classify the testing stimuli (/i/-/ɪ/ continuum), 4 times each blocked by repetition (28 classifications per round). During these rounds, a large label displayed the speaker number (e.g., 'Speaker 1') using labels 1, 2, and 3 for the speakers in rounds 2, 3, and 4 respectively. The second round always featured the large standard voice. Half of participants heard the medium voice in round 3 and the small voice in round 4, and the other half heard the small voice in round 3 and the medium voice in round 4.

Round five was a final mixed-speaker round, after familiarization with the talkers. As in round 1, listeners were presented with testing stimuli for three voices 5 times each, blocked by repetition. However, in this round a label told listeners which speaker was producing each token (as in rounds 2-4). Listeners were also instructed before beginning the round that these labels would tell them which speaker produced the vowel, and that these speakers would be the same ones they just heard in the previous rounds.

Results and discussion

Classification differences between blocks

We are primarily interested in the blocked-voice rounds as these represent situations where listeners had the most information about vowel-space characteristics. However, we were also interested in comparing listener behavior across different presentation types. It may be the case that listeners are operating in a sort of LM-compatible listening mode in the initial mixed-voice round (as in Experiment 1). In blocked-voice rounds after familiarization, listeners may shift to a listening mode more in line with the outputs of the WF and LB normalization methods. In the final mixed-voice round, listeners could revert to an LM listening mode or they may 'remember' the voice of speakers in the final round, behaving in a similar manner to blocked-voice rounds.

If listeners were changing their behavior across presentation type, we would see differences in the classification functions of voices across rounds. The three speaker types would be differentiated in the initial mixed-voice round (as in experiment 1), then look more similar (or identical) in the blocked-voice round, with unclear expectations for the final mixed-voice round. As seen in Figure 9, listeners are not exhibiting substantially different behavior across the presentation types. Classification functions differ across voice types in all presentation

conditions, indicating that listeners are not perceptually 'erasing' the variation in formant patterns that distinguishes these voice types, even when given more information about the idiosyncratic differences between speakers. A more complete analysis of the results in the blocked rounds is provided in the following section.

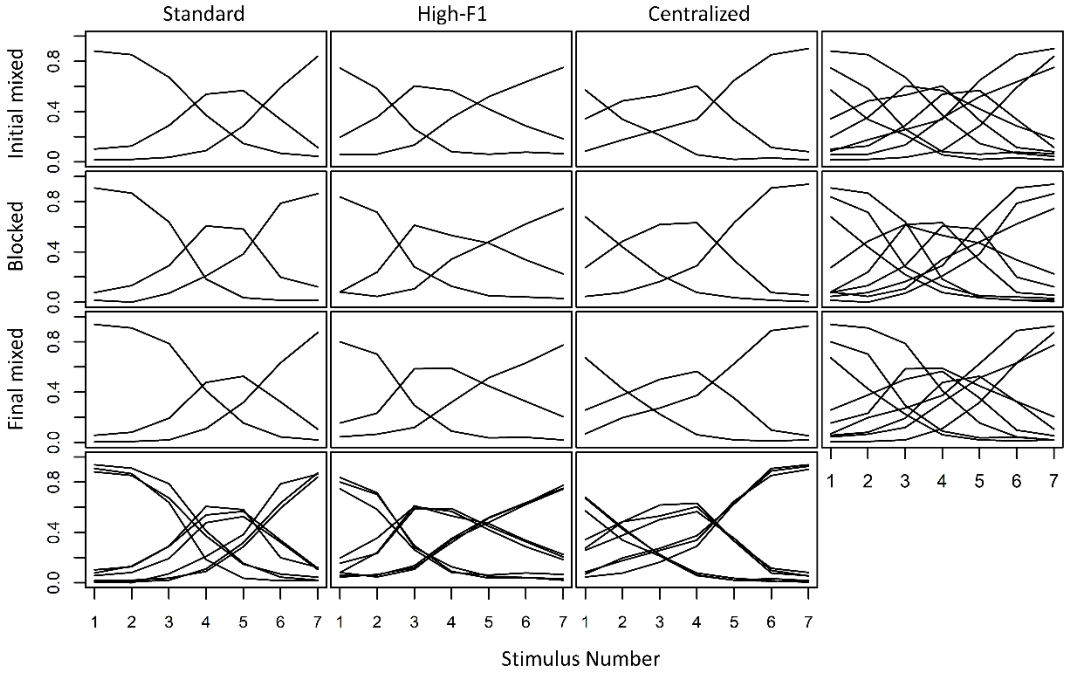


FIGURE 9, Proportion of classifications into /i/ (left distribution), /ɪ/ (middle distribution), and /ɛ/ (right distribution) by continuum step, presented across voice and presentation type. The final row compares the classification functions of the panels in each column. The final column compares the classification functions for the panels in that row.

Blocked-Speaker Rounds

In blocked speaker rounds (2-4), listeners are asked to classify continuum vowels after being presented with examples of a speaker's peripheral vowels (/i u æ ɑ o/). In these rounds, listeners could potentially use the familiarization vowels to adapt to the speaker, and classify the continuum vowels relative to knowledge of the speaker's vowel space. In other words, in these rounds it is clear that listeners *could* exhibit perceptual behavior in line with WF and LB normalization if they were so inclined, as they have all the information required by these normalization methods.

To investigate support for phone-preserving variation in line with WF and LB normalization in blocked-speaker rounds, the testing vowels were normalized using the three methods (LM, WF, and LB) based on vowel-space statistics calculated using the training stimuli, as in Figure 6. An ordinal logistic multilevel regression model was fit to listener classification

data for the blocked-speaker rounds (2-4), for each normalization method. Each model had a single predictor (position in the normalized space) with random slopes and intercepts for listener. Since variation in the testing continuum was almost entirely unidimensional (99-100% of variance along the first principal component), the position of each token was specified along the axis of the first principal component of stimulus variation for each normalized space (corresponding primarily to F1).

The above models were used to predict classification rates along the different normalized spaces. The classification rates predicted by each model represent our best estimates of the phonetic properties associated with each location in the normalized space, given our data. For example, we could say that a token at coordinate $\langle x, y \rangle$ in the LB normalized space is 78% likely to be classified as /i/, 13% likely to be classified as /ɪ/, and 9% likely to be classified as /ɛ/. Ideally, a normalization method will offer a tight clustering of tokens along predicted classification rates so that, for example, all stimuli near $\langle x, y \rangle$ in the LB space are more likely to be classified as /i/ than /ɪ/. When this does not occur, tokens in one location of the normalized space will have diverse phonetic properties, meaning that distance in the normalized space will not be a reliable metric for phonetic differences. Figure 10a compares predicted classification functions (solid lines) to the observed classification rates of testing vowels (points) in blocked-speaker conditions. Each row presents the same data, with differing alignments between predicted and observed classifications arising from the differing arrangement of tokens in the normalized space (seen in Figure 6).

In Figure 10a we see that the decreased similarity of normalized vowel spaces provided by the LM method directly leads to a greater degree of clustering of tokens in the LM-normalized space. The LM method groups the first step of the centralized continuum, the second step of the high-F1 continuum, and the third step of the standard continuum because these are all being classified as /i/ roughly 70% of the time. In contrast, the increased vowel-space similarity afforded by the LB and (to a lesser extent) the WF methods is obtained by grouping phonetically dissimilar vowels, and leads to less certainty regarding the vowel quality associated with any given sound. For example, as seen in the right column of Figure 10a, the third continuum step is classified as /i/ for the standard voice (circles) and as /ɪ/ for the high-F1 and centralized voices (triangle, square), despite being placed in the same location in the LB-normalized space. Thus, we see that the LB method is placing tokens that correspond to different vowel phonemes in a single location in the normalized space.

The following bootstrap analysis was carried out to investigate the reliability of the differences seen in Figure 10a. For each of 10,000 iterations, pooled classification rates were found for each stimulus for data from 49 listeners, resampled with replacement at the listener level. Then, the square of the correlation (R^2) between the predicted and observed classification rates was calculated and recorded, independently for each normalized space. When the classification of tokens is highly predictable from their position in the normalized space, R^2 will be high. As a result, a high R^2 indicates that distance is a good metric for phonetic differences in a given normalized space. In contrast, a lower R^2 indicates that there is more variation around any given location, meaning there can be less certainty regarding the phonetic properties of any

given vowel. The correlation between all tokens across successive samples was also recorded in order to get an idea of the amount of random error expected across samples.

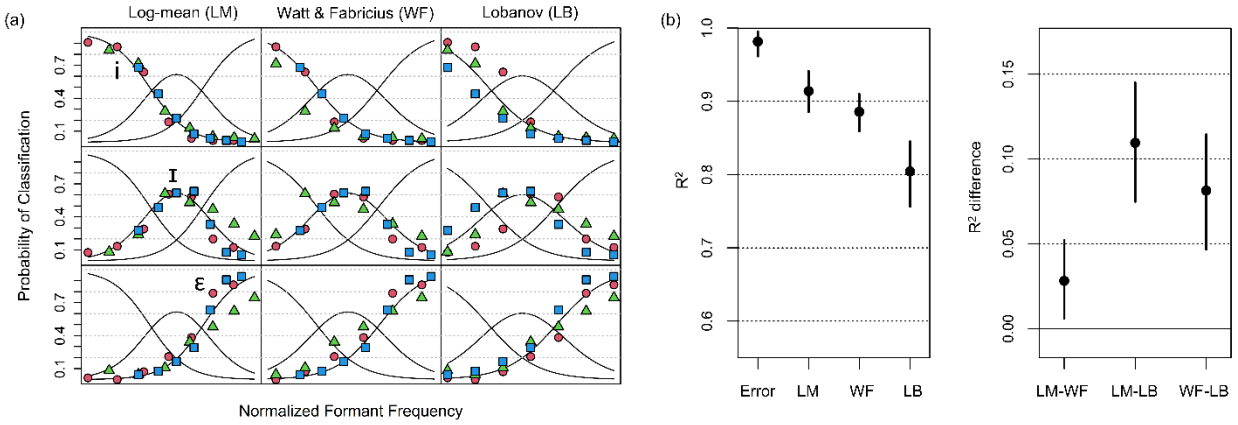


FIGURE 10. (a) Points indicate classification rates for all testing stimuli into different categories, organized along each of the different normalized spaces. Continuum steps increase left to right. Point types indicate standard (circles), high-F1 (triangle) and centralized (square) speakers. Lines indicate predicted classification rates at each location. (b) Distribution of R^2 for each normalization method, and the differences in R^2 between each method resulting from the bootstrap analysis.

The distribution of recorded R^2 values for each normalization method is presented in Figure 10b, as are distributions of differences in R^2 between the methods. Results indicate a slight advantage in R^2 for LM normalization over the WF method, while both the LM and WF methods show a large advantage over the LB method. It is important to note that although the differences between the LM and WF methods were small, they were very consistent, with the LM method showing an advantage in 99.2% of samples. Further, the acoustic differences between the standard and high F1 voices were not large: an 11.5% difference to a single formant frequency. If the acoustic difference had been doubled, it is reasonable to think that listeners would have heard a larger phonetic difference between the standard and high-F1 voices, leading to a larger advantage for the LM method. However, in such a situation the WF (and LB) methods would have erased this variation just the same.

GENERAL DISCUSSION

The goal of the experiments outlined above was to investigate which sorts of normalization operations are perceptually justified, meaning they tend to remove only phone-preserving variation from formant patterns. Experiment 1 showed that in situations with no information about speakers, large differences in uniform scaling were mostly ignored, while relatively smaller differences that deviate from uniform scaling resulted in large shifts in classification functions, suggesting differences in perceived vowel quality. In experiment 2, we saw that this tendency is maintained even when listeners have information regarding the geometry of the

speaker's vowel space, allowing them an opportunity to perceptually 'erase' systematic deviations from uniform scaling of formant patterns. Overall, results indicate that variation in formant patterns according to uniform scaling tends to be phone preserving, while deviations from uniform scaling tend to be 'heard' by listeners, resulting in variation in the phonetic properties of vowel sounds.

We may return to the possible differences in the low vowels produced by the male and female speakers in our corpus (Figure 4), presented in more detail in Figure 11. Often, a researcher wishes to use distance in the normalized space in order to quantify some linguistic difference between productions (e.g., Clopper et al., 2005; Podesva et al., 2015). As an example of this, we will investigate the normalized productions of the California speakers described above. It is expected that differences in normalized F1 should relate to variation in phonetic height and differences in normalized F2 should relate to variation in phonetic frontness. A series of two-sample t-tests were carried out comparing average productions of each phoneme by male and female speakers. Differences were tested along F1 and F2 independently, for each normalization method (summarized in Figure 11).

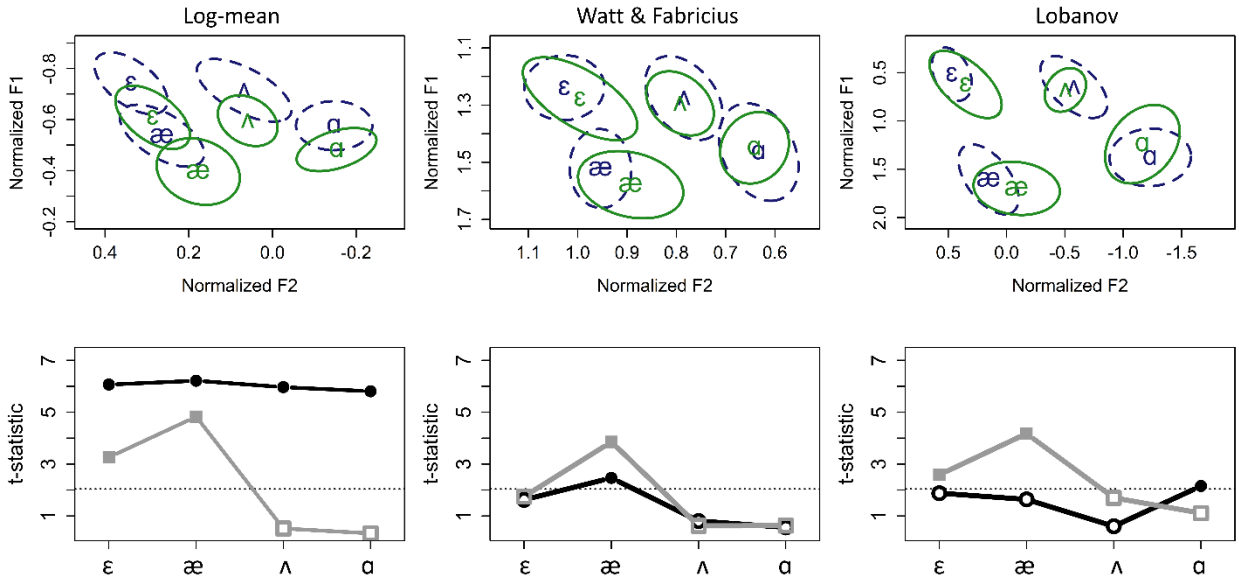


FIGURE 11. In the top row, productions of 30 male (dashed line) and female (solid line) California speakers are compared for low and mid vowels. Ellipses enclose two standard deviations. In the bottom row, lines indicate values of t-statistics comparing means for each vowel along F1 (circles) and F2 (squares). The horizontal dotted lines indicate the level at which values reach significance, and filled points indicate significant comparisons.

A reliance on significance testing to infer 'real' phonetic differences between the men and women in our sample suggests different patterns of results based on the normalization method chosen, making the choice of method of high practical importance. The different results presented in Figure 11 primarily arise from the non-uniform scaling differences between the

male and female speakers seen in the original data (see Figure 4). This variation is directly analogous to the difference between the large standard voice and the small high-F1 voice. The results of the experiments above indicate that listeners are likely to 'hear' these non-uniform differences in formant scaling, suggesting likely differences in the phonetic properties of the low vowels produced by these speakers. Based on this, the organization presented by the LM method is most likely to reflect the phonetic properties, and the linguistic facts, of the data in question.

As noted in Barreda and Nearey (2018: Appendix) there is good evidence that speakers vary in production in ways that can only be captured by formantwise standardization methods. So, it is a fact that speakers differ from each other in the range and location of each formant (somewhat) independently, in a manner consistent with the complexity of Lobanov normalization (as seen in Figures 2, and 3). However, there is no evidence that listeners exhibit this level of complexity in their perceptual normalization. Basically, it appears as though we are Lobanov speakers but log-mean listeners: our variation in production is as complex as the Lobanov model, but our adaptation to this variation in perception is only as complex as the log-mean model. As a result, the many subtle, idiosyncratic between-speaker differences in production that cannot be captured by single parameter scaling methods may result in perceivable phonetic differences that potentially transmit linguistic and social information between speakers (Barreda, 2020). Thus, a focus on finding a normalization method that "does the same work as the normalizing ear of the listener", as Labov and colleagues put it, will focus on modeling the judgments of human listeners in response to between-speaker variation in production, rather than on modeling the variation in production itself.

Limitations and future directions

Although we did not observe any perceptual behavior consistent with WF and LB normalization, it is possible that the design of the experiment (the stimuli, the training, etc.) was such that we did not 'trigger' whatever perceptual processes are necessary to result in outputs like those of WF and LB normalization. It is difficult to prove that perceptual processes consistent with WF and LB normalization *never* occur. Instead, we looked for positive support for the use of these methods in different listening conditions and failed to find any. Further work is needed to investigate whether formantwise scaling and formantwise standardization methods are perhaps appropriate in other listening situations. However, there is reason to be somewhat skeptical of this possibility. As noted earlier, formantwise scaling and standardization methods do not conform to any well-known theory of vowel perception, nor do they have any empirical support in the literature on speech perception. Further, it bears noting that if listeners only perceive speech in modes consistent with WF and LB in a very restricted set of conditions, then the output of these normalization methods will also only be valid in those conditions.

Although uniform-scaling methods were the best testing in our comparison, there is room for improvement. For example, as seen in Figure 7, there is systematic variation across sizes within voice types that cannot be explained by single parameter scaling methods. This suggests the possibility of a more complicated relationship between the lower formants (F1 and F2), the higher formants (F3 and above), and the fundamental frequency of vowel sounds. A perception-centric perspective on normalization suggests that we should adapt these methods as required so

that they reflect listener judgments of the phonetic properties of vowel sounds. Thus, a better understanding of the nature of 'phone-preserving' variation in speech perception can only benefit empirical research investigating variation and change in vowel systems across speakers.

Finally, it bears noting that the search for the 'perfect' normalization method from a perception-centric perspective is complicated by the fact that vowel-quality judgments are inherently 'fuzzy', varying probabilistically within-listener and systematically between listeners. Listeners of the same dialect, and even trained phoneticians, can have differences of opinion regarding the vowel quality associated with a given sound. Further, the development of the 'perfect' normalization method is limited by our knowledge of human speech perception, which is as yet incomplete. As a result, the outputs of normalization methods should perhaps be thought of as estimates of the judgments of some 'average' listener, to within some degree of certainty. Of course, employing a normalization method that focuses primarily on maximizing vowel-space similarity does not make any of this uncertainty go away, and potentially obscures the truth further. Thus, if researchers intend to use normalized data to make inferences about the vowel quality of a set of tokens, they will benefit from employing methods whose outputs more-closely reflect the perceptual organization of those tokens. However, it is useful to always keep in mind the limitations inherent in representing complex perceptual events such as vowel sounds using points in a low-dimensional space.

Conclusion

Instead of desiring maximally similar normalized data, researchers may benefit by focusing on obtaining normalized data that reflects the phonetic properties of a set of vowel sounds. When this is the case, normalized tokens will lie together if they 'sound' similar and lie apart if they 'sound' dissimilar. Speakers with vowel systems whose tokens largely lie together in the normalized space will tend to sound 'the same' to listeners, therefore likely constituting speakers of the same dialect. So, by attempting to obtain data that reflects the perceptual and phonetic structure of speech sounds, we can allow for a 'bottom-up' approach to investigating the homogeneity of groups of speakers rather than imposing homogeneity through the selection of our methods.

The results of two perceptual experiments suggest that single parameter scaling methods (e.g., log-mean normalization) most faithfully reflect the phonetic structure of vowel sounds, and that formantwise scaling or standardization methods (e.g., Watt and Fabricius, Lobanov) can both remove legitimate phonetic variation from vowel formant data. As a result, the log-mean method is likely preferable in cases where researchers wish to use distances in the normalized space to infer the phonetic properties of a set of vowel sounds. However, although the log-mean method performed best of the methods we tested, there is room for improvement, and it is crucial to continue to investigate and refine the normalization methods so often relied upon in quantitative variationist research.

References

Adank, Patti, Smits, Roel, &Van Hout, Roeland. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America* 116: 3099–3107.

Barreda, Santiago. (2020). Vowel normalization as perceptual constancy. *Language* 96: 224-254.

Barreda, Santiago & Nearey, Terrance. (2018). A regression approach to vowel normalization for missing and unbalanced data. *Journal of the Acoustical Society of America:* 144, 500–520.

Clopper, Cynthia, Pisoni, David & de Jong, Kenneth. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America* 118: 1661–1676.

Disner, Sandra. (1980). Evaluation of Vowel Normalization Procedures. *The Journal of the Acoustical Society of America:* 67: 253–261.

D'Onofrio, Annette, Pratt, Teresa, & Van Hofwegen, Jannekke. (2019). Compression in the California Vowel Shift: Tracking generational sound change in California's Central Valley. *Language Variation and Change* 31: 193-217.

Fabricius, Anne, Watt, Dominic, and Johnson, Daniel. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change* 21: 413–435.

Ferguson, Sarah, and Kewley-Port, Diane. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research* 50: 1241–1255.

Flynn, Nicholas, and Foulkes, Paul. (2011). Comparing vowel formant normalization methods. *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong*: 683–686.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*: 16, 78–80.

Hillenbrand, James, Getty, Laura, Clark, Michael & Wheeler, Kimberlee. (1995). Acoustic Characteristics of American English Vowels. *The Journal of the Acoustical Society of America* 97: 3099–3111.

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. Linguistic variation: Models and methods, 161-171.

Klatt, Dennis. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America* 67: 971–995.

Kohn, Mary & Farrington, Charlie. (2012). Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. *Journal of the Acoustical Society of America* 131: 2237–2248.

Labov, William. (1994). Principles of Linguistic Change. Vol 1: Internal Factors. Vol. 2: Social Factors, Oxford: Blackwell.

Labov, William, Ash, Sharon & Boberg, Charles. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. De Gruyter Mouton.

Lobanov, Boris. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America* 49: 606–608.

Munson, Benjamin. (2007). Lexical Access, Lexical Representation, and Vowel Production. *Laboratory Phonology 9: 201–228*.

Nearey, Terrance. (1978). *Phonetic feature systems for vowels*. Indiana University Linguistics Club.

Nearey, Terrance. (1983). Vowel-space normalization procedures and phone-preserving transformations of synthetic vowels. *The Journal of the Acoustical Society of America* 74: S17–S17.

Nearey, Terrance. (1998). Selection of a tonotopic scale for vowels. *Proceedings of the 16th International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*: 2001–2002.

Podesva, Robert, D'Onofrio, Annette, Hofwegen, Jannekke & Kim, Seung. (2015). Country ideology and the California Vowel Shift. *Language Variation and Change* 27: 157–186.

Rankinen, Wil & de Jong, Kenneth. (2020). The Entanglement of Dialectal Variation and Speaker Normalization. *Language and Speech*, June.

Sankoff, David. (1978). *Linguistic variation: Models and methods*. Academic Press.

Smith, David & Patterson, Roy. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America* 118: 3177–3186.

Tamminga, Meredith. (2019). Interspeaker covariation in Philadelphia vowel changes. *Language Variation and Change* 31: 119–133.

Watt, Dominic & Fabricius, Anne. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1~F2 plane. *Leeds Working Papers in Linguistics and Phonetics:* 9, 159–173.