Laboratory Phonology
Journal of the Association for Laboratory Phonology

Johnson, K. 2020 The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 10, pp. 1–16. DOI: https://doi.org/10.5334/labphon.196

JOURNAL ARTICLE

The ΔF method of vocal tract length normalization for vowels

Keith Johnson
Department of Linguistics, University of California, Berkeley, CA, US
keithjohnson@berkeley.edu

Given the acoustic consequences of physiological differences between talkers, there is a practical need for effective and theoretically motivated procedures of vowel normalization to facilitate comparison of speech produced by people who differ by dialect or language. In addition, there is a question whether listeners might utilize a normalization procedure during speech perception. This paper reports the results of two studies that explore these questions—with particular focus on vocal tract length normalization. Drawing on research in speech engineering, where accurate estimates of vocal tract length are needed in some approaches to automatic speech recognition and speaker verification, a new model of vowel normalization is introduced. The model uses a direct measure of average formant spacing (the ΔF) which can be used to measure vocal tract length. The acoustic consequences of vocal tract length differences are removed from vowel measurements by scaling vowel formant measurements by ΔF. Study 1 found that this method is comparable to Nearey's (1978) uniform normalization method, while providing an explicit vocal tract length interpretation, and a rationalized unit of measure. Study 2 found that uniform normalization measures (which let each formant serve as a noisy estimator of ΔF) improve vowel classification even with only a couple of randomly selected vowel tokens. This suggests that vocal tract length normalization could be involved in speech perception.

Keywords: Speech Perception; Talker Normalization; Vowel Normalization; Vocal Tract Length

1. Introduction

The acoustic properties of speech are shaped in large part by the movements of the upper vocal tract and the vibrations and turbulence set up by the flow of air through the larynx and vocal tract constrictions. In addition to the controlled movements of the upper vocal tract, speech acoustics are determined by the overall size of the vocal tract and larynx.

Vowels differ primarily in the two lowest resonances of the vocal tract—the first and second 'formants' (F1 and F2). For instance, the vowel [i] has a low F1 and a high F2, while [a] has a high F1 and a low F2. The acoustic vowel space that is defined by a talker's range of F1 and F2 values is shifted up or down in average F1 and F2 frequency as a function of vocal tract length. Longer vocal tracts have lower average resonant frequencies. This vocal tract length effect creates a practical challenge for automatic speech recognition, and for the phonetic comparison of speakers, dialects, and languages, as well as a 'normalization' problem for listeners who must recognize vowels produced by a range of different talkers.

This paper will introduce the ΔF method of vocal tract length normalization. In two studies, the paper will evaluate ΔF vowel normalization as a practical technique for acoustic analysis of language data, and also gauge the feasibility of vocal tract length normalization as a perceptual mechanism.

1.1. Vocal tract length

According to the acoustic theory of speech production (Fant, 1960) it is theoretically possible to remove vocal tract length differences from our descriptions of speech acoustics (Nordström & Lindblom, 1975). Once we have measurements of the vocal tract resonant frequencies (F1-F4) we can use the talker-specific vowel formant distributions to estimate normalization factors. Several normalization methods have been proposed, with more or less reference to acoustic theory (Gerstman, 1968; Lobanov, 1971; Nordström & Lindblom, 1975; Nearey, 1978; Watt & Fabricius, 2002). The aim of acoustic vowel normalization is to make it possible to express the formant frequencies using a measurement scale that is independent of talker differences, for use in comparing dialects and languages with each other. Vowel normalization may also be a part of the cognitive process of speech perception.

Whether listeners make use of vocal tract length in speech perception is an open question (Johnson, 1997). It has been argued that perceptual compensation for talker differences must be more than simply a vocal tract length normalization, because acoustic differences between men and women (who tend to differ in vocal tract length) is language-specific; male/female differences depend in part on the language or dialect that they speak (Johnson, 2005). This language-specificity of gender differences suggests that there is more involved than just vocal tract length difference. Instead, there appears to be a performative aspect of gender that is overlaid on physical sex differences in vocal tract length. Secondly, it has been observed that perceptual talker normalization effects are conditioned to some degree on higher-level factors that are given by prior phonetic context (Ladefoged & Broadbent, 1957; Johnson, 1990), visual context (Strand & Johnson, 1996), or even experimenter suggestion (Johnson, Strand, & D'Imperio, 1999).

Despite these observations about speech perception, there are a couple of reasons to suppose that vocal tract length normalization might be a component of the speech perception process. First, the perception of a conspecific individual's body size is important for social organization in many species (e.g., Harrington & Mech, 1979), and there is evidence that vocalizations are used to convey individual characteristics such as size (Reby & McComb, 2003). This suggests that the perception of vocal tract length may be evolutionarily prior to linguistic communication, so language processing may be overlaid on a cognitive structure that already included vocal tract length perception. Second, there are regions of the brain involved in talker perception that do not overlap with regions involved in speech perception (e.g., Van Lanker, Kreiman, & Cummings, 1989; Johnson & Sjerps, to appear), suggesting that the perceptual system may include processes that compute talker information that can be mixed with phonetic information in a stream that produces 'talker neutral' phonetic information.

1.2. Extrinsic normalization

One reason to believe that vocal tract length normalization is not a viable mechanism for speech perception is that, as usually formulated, it relies on information that is extrinsic to the vowel. That is, it is usually assumed that the estimation of vocal tract length is computed over a collection of vowel tokens—information beyond what is available intrinsically in the vowel to be classified. Common experience, and controlled experimentation confirm that isolated vowels are accurately recognized (Nearey, 1989; Strange, Jenkins, & Johnson, 1983). This suggests that each vowel contains the information that is needed for its own recognition, and this intrinsic information is usually sufficient for vowel recognition.

Lammert and Narayanan's (2015) finding is relevant for this discussion. Building on earlier work in speech engineering (Paige & Zue, 1970; Kirlin, 1978; Wakita, 1977), they devised a regression method using the frequencies of formants F1-F4 to estimate

vocal tract length over short stretches of speech. Vocal tract length normalization using four formants F1-F4 should be less reliant on the specific vowel tokens that are used to calculate normalization parameters than methods like *z*-score normalization (Lobanov, 1971) that normalize F1 (for example) on the basis of information about the distribution of F1 in a set of vowel measurements. Obviously, you cannot accurately estimate the distribution of a formant's frequencies from a single token, so the Lobanov normalization method cannot be a model of how listeners identify (and normalize) isolated vowels. On the other hand, in vocal tract length normalization, information from F1-F4 determines a normalization scale factor, so each vowel contains a formant pattern within which to evaluate (and normalize) the formants.

1.3. Descriptive normalization

Regardless of whether vocal tract length normalization is done in speech perception, descriptive studies of language phonetic systems rely on vowel normalization algorithms to compare speech produced by different talkers or groups of talkers (Disner, 1980; Adank, Smits, & van Hout, 2004). Methods used in these studies are sometimes based on general-purpose statistical normalization techniques such as range normalization (Gerstman, 1968), or *z*-score normalization (Lobanov, 1971), but more specialized methods using what could be called 'mean normalization' (the ratio of x to the mean of x) are also widely used (Nearey, 1978; Watt & Fabricius, 2002).

In vocal tract length normalization (Nordström & Lindblom, 1975), a single normalization scale factor is derived from an estimate of the length of the speaker's vocal tract. An acoustic factor related to vocal tract length is ϕ, the fundamental frequency of vocal tract resonances in an unconstricted vocal tract (Paige & Zue, 1970; Kirlin, 1978; Lammert & Narayanan, 2015). The factor ϕ is equal to F1 and the other formants are on odd harmonics of this value: F2 = 3ϕ, F3 = 5ϕ, F4 = 7ϕ. ΔF is simply 2ϕ, and is an estimate of formant spacing in an unconstricted vocal tract (Reby & McComb, 2003), where F1 = ½*ΔF; F2 = 1½*ΔF; F3 = 2½*ΔF, etc. Because a single normalization scale factor is used to scale the formants (rather than a separate factor for each formant), this method is called a 'uniform' normalization method.

Despite its basis in the acoustic theory of speech production and the interpretation of the normalization scale factor in terms of a physical characteristic of the talker, vocal tract length normalization was rejected almost as soon as Nordström and Lindblom (1975) proposed it, because it did not seem to work very well. However, there has been significant progress in deriving more accurate vocal tract length estimates from the acoustic vowel spectrum, so a reconsideration of vocal tract length normalization is due.

1.4. Organization

This paper will introduce the ΔF method of vocal tract length normalization in Section 2. Section 3 will describe the methods used in two studies of vowel normalization techniques. The results of the studies are presented in Section 4, and the paper concludes with recommendations in Section 5. The aims of the paper are to evaluate ΔF vowel normalization as a practical technique for acoustic analysis of language data, and also to gauge the feasibility of vocal tract length normalization as a perceptual mechanism.

2. Methods for vocal tract length normalization

The earliest use of vocal tract length as a normalization factor was by Nordström and Lindblom (1975). They calculated the average third formant (F3) frequency in open vowels (where the frequency of F3 is easily distinguished from F2) and used this to estimate the talker's vocal tract length. Vocal tract length was then used to scale a talker's

vowel formant measurements onto a 'standard' (i.e., male) vocal tract. We can avoid the male-centric bias of this approach by using a talker-independent measurement scale—an estimate of formant spacing in an unconstricted vocal tract, the ΔF.

Speech engineers have devised measures of vocal tract length from vowel acoustics (Paige & Zue, 1970; Wakita, 1977; Kirlin, 1978) for use in automatic speech recognition and speaker identification (Eide & Gish, 1996; Lee & Rose, 1998). For example, Kirlin (1978) used information from the four lowest formants and estimates of formant variances to weight the contributions of formants. He treated "each formant as a noisy estimator" of vocal tract length, so in his calculation, because F2 has large variance, it contributes less to the vocal tract length estimate. Lammert and Narayanan (2015) published an important study in this line. They compared the vocal tract length predicted from acoustic measures to estimates of vocal tract length measured from magnetic resonant imaging (MRI) of the vocal tract, as well as computer simulated vocal tracts of known length. They found a family of regression formulas that predict vocal tract length from the first four formant frequencies of vowels, weighting formants as Kirlin did but finding the weights by regression to known vocal tract length.

Following the analysis outlined by Lammert and Narayanan (2015) and building on the line-fitting approach of Reby and McComb (2003; see **Figure 1**), a talker's ΔF is directly estimated by scaling formant frequencies by their formant number (F1, F2, F3, F4, etc.) as

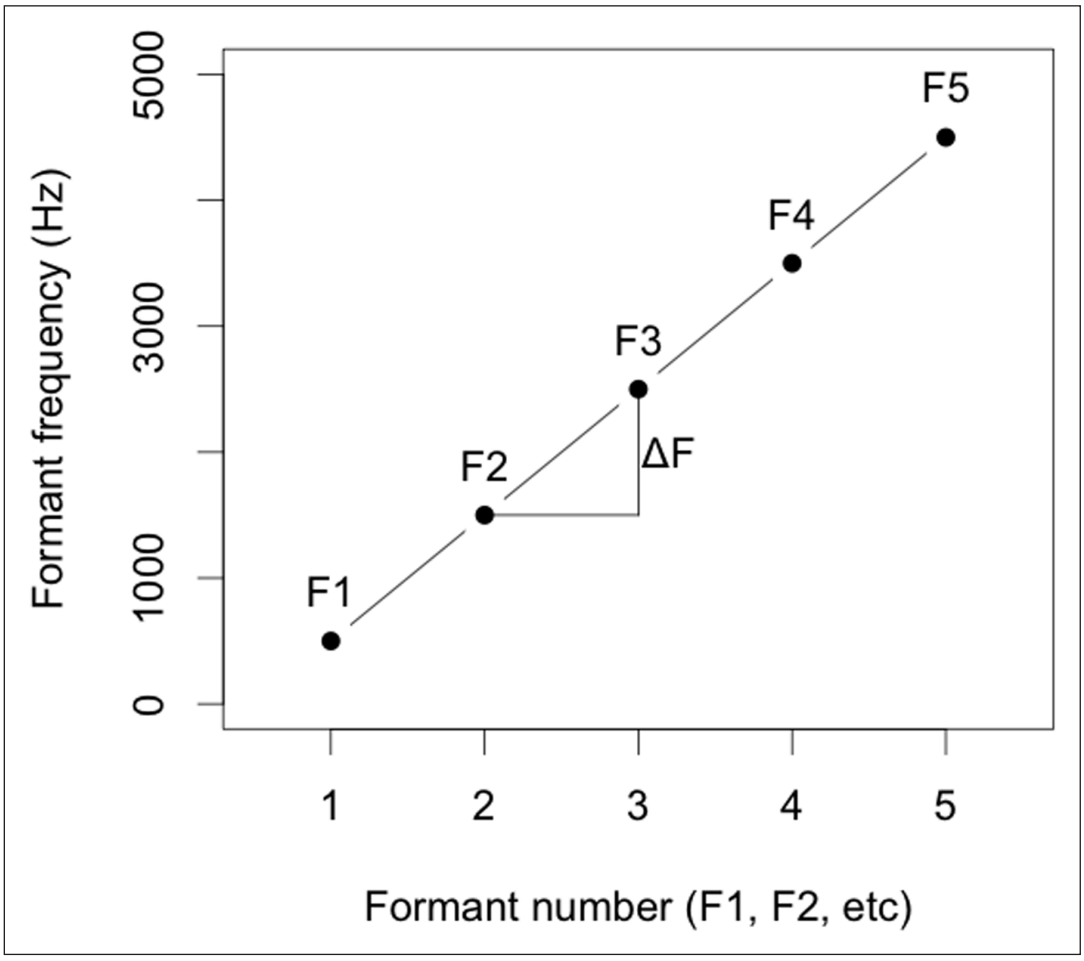

Figure 1: An illustration of Reby and McComb's (2003) direct estimation approach for finding ΔF from formant measurements. The average spacing between vowel formants (ΔF) is the slope of the line that relates formant number to formant frequency. The figure shows the formant frequencies of a vocal tract that is 17.5 cm long with no constrictions (a uniform tube). ΔF is 1000 Hz, and F1 is 500 Hz, F2 is 1500 Hz, etc.

in formula (1). Note that each formant (F1-F4) of each vowel provides an estimate of ΔF. The sum in (1) can be taken over all of the vowels for a talker that are available in a dataset so missing values have a limited impact on the estimate. It is important to note that even though F1 and F2 vary substantially for different vowels, averaged over a representative dataset they approximate the frequencies of an unconstricted vocal tract—a uniform tube. Section 3 will compare the direct estimation of ΔF using the average formant spacing method in (1) with the optimized empirical method presented by Lammert and Narayanan (2015). Kirlin's (1978) a posteriori method was also implemented and results from this method of vocal tract length estimation will also be reported. Once we have calculated ΔF, then the talker's vocal tract length can be calculated from ΔF by formula (2), and the normalized vowel formants are calculated using ΔF as in (3).

(1) $$\Delta F = \frac{1}{mn}\sum_{j}^{m}\sum_{i}^{n}\left[\frac{F_{ij}}{i-0.5}\right],$$ where $i =$ formant number (1…4), and j is token number

(2) $VTL = c/2\Delta F$, $c =$ 34000 cm/s, the speed of sound, warm moist air

(3) $F'_{ij} = F_{ij}/\Delta F$, normalized formant frequencies

This average formant spacing method of estimating vocal tract length (VTL) is essentially perfectly correlated with Lammert and Narayanan's (2015; henceforth L&N) estimated VTL on the Hillenbrand, Getty, Clark, and Wheeler (1995) dataset ($r^2 = 0.998$, **Figure 2** below), while the Nordström and Lindblom (1975) estimate which was based on F3 alone, is much less strongly correlated with L&N ($r^2 = 0.829$). Kirlin's (1978) method was also closely correlated with the L&N method ($r^2 = 0.963$). The strong correlations between the average spacing (formula 1) and L&N methods is an indication that with a full sample of vowels for a talker, the average spacing method is a valid measure of apparent vocal tract length. The key difference between L&N's data-driven approach and the average spacing approach is that the average spacing approach assumes that the formant values will, on average, be equally spaced (as in **Figure 1**). The high correlation between the methods can be seen as a validation of this assumption for a typical phonetic dataset. As seen in **Figure 2**, the length estimated by the average formant spacing method is consistently about 0.8 cm longer than the length estimated by the L&N method.

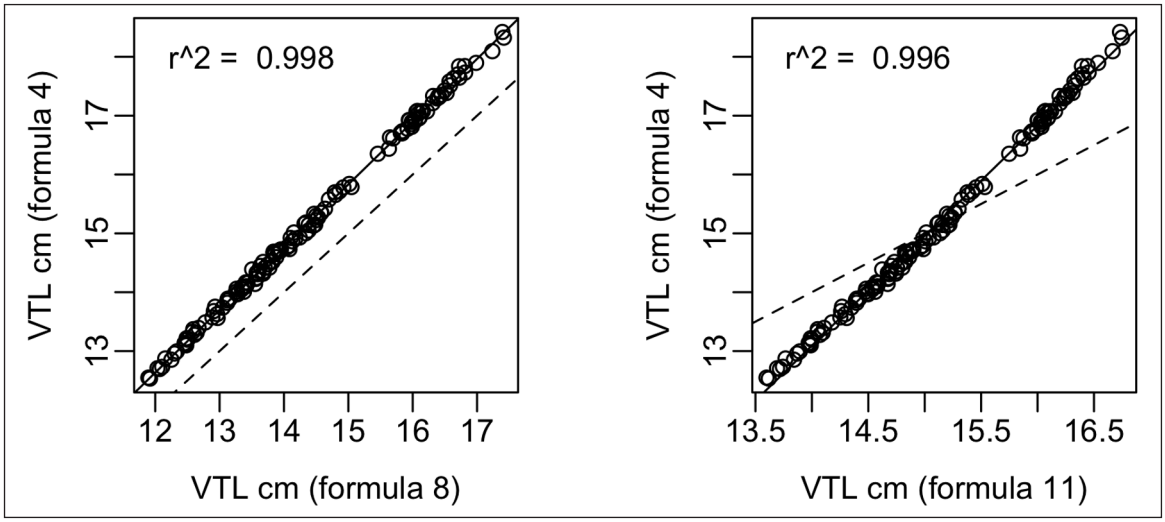

Figure 2: Vocal tract length estimates for the talkers in Hillenbrand et al. (1995), as calculated using the average spacing formula (4) and the two formulas calculated by Lammert and Narayanan (2015): formula (8) with no zero intercept, and formula (11), which does have a zero intercept. The dashed line is the identity line, y = x.

Formula (3) shows that ΔF can be used as a unit of measure for vowel formant frequencies, putting vowel formants from all vocal tracts on the same measurement scale. The normalized F1' value is given as F1/ΔF and is expected to be equal to 0.5 for a uniform tube of any length, the normalized F2' is F2/ΔF and is expected to be equal to 1.5 for any uniform tube, and so on. This is *ΔF vocal tract length normalization*, where ΔF can be estimated in several different ways—from F3 alone (Nordström & Lindblom, 1975), using the average formant spacing approach (formula 1), or using Lammert and Narayanan's (2015) regression fits, or Kirlin's (1978) weight by variance method. This use of ΔF as a vowel normalization factor, which has not been proposed before, is the main practical contribution of this paper.

3. Methods

3.1. Data sets

The data analyzed in this paper are the published American English vowel production data from Peterson and Barney (1952) and Hillenbrand et al. (1995), as distributed by Santiago Barreda in his 'phonTools' package for the R statistical programming language.

3.2. Normalization formulas

The data were analyzed in the R statistical programming language, and the normalization algorithms were implemented as illustrated in **Table 1**. As the variable names in the last four rows of the table make clear, the normalizing factor ΔF can be calculated by any method that provides an estimated vocal tract length from the acoustic vowel formants. The Watt & Fabricius method prescribes taking the mean of formants from particular judiciously selected vowel qualities to ensure that the center of the talker's acoustic vowel space is adequately captured, to provide a cross-linguistically consistent scale. The implementation here used the mean of the entire sample as the estimate of the center of the acoustic vowel space. This works just as well for vowel classification within language, and avoids subjectivity or other mistakes in the selection of the exemplary vowel tokens.

Table 1: Some details, in R, of how the normalization algorithms were implemented. In these code snippets 'f1,' 'f2,' 'f3' are arrays of vowel formant measurements for a particular talker. The first three rows are non-uniform methods and the last four rows are uniform methods.

Method	R code
Lobanov, *z*-score	F1' = scale(f1);
Watt & Fabricius	F1' = f1/mean(f1);
Nearey 1, non-uniform	F1' = exp(log(f1) – mean(log(f1)));
Nearey 2, uniform	mf = mean(c(log(f1),log(f2),log(f3)));
	F1' = exp(log(f1)-mf);
Nordström & Lindblom	ΔF = mean(f3[f1>600]/2.5);
	F1' = f1/ΔF;
Kirlin	x = mean(f1/146^2) + 3*mean(f2/485^2) + 5*mean(f3/322^2) + mean(f1)/40^2;
	ΔF = 2(x*1051.2);
	F1' = f1/ΔF;
Lammert & Narayanan	ΔF = 2*(262 + mean(0.14*f1) + mean((0.16*f2)/3) + mean((0.25*f3)/5));
	F1' = f1/ΔF;
average spacing ΔF	ΔF = mean(c(f1/0.5, f2/1.5, f3/2.5));
	F1' = f1/ΔF;

3.3. Vocal tract length coefficients

The regression coefficients used in this paper for the Lammert & Narayanan (2015) method are different from the ones they published because the datasets used here only include the first three formants, while L&N used F1–4 to estimate vocal tract lengths. Dr. Lammert was kind enough to provide two sets of coefficients for regressions fitted from F1–3 for simulated data. Without an intercept ($\beta_1 = 0.28$, $\beta_2 = 0.31$, $\beta_3 = 0.47$) the RMS error in estimated vocal tract length is 1.91 cm. A regression formula that includes an intercept term ($\beta_0 = 262$, $\beta_1 = 0.14$, $\beta_2 = 0.16$, $\beta_3 = 0.25$) leads to a smaller error of estimated vocal tract length (1.22 cm).[1] The classification studies in the next section used the second formula, the one with an intercept term. The ΔF estimates of the intercept formula tend to be regulated, drawn away from extremely short or long estimates, by the β_0 constant, as further discussed below.[2]

The comparison of the Lammert and Narayanan (2015) coefficients to the average spacing coefficients, formula (1), is complicated by a slight difference in how they are presented and calculated. Formula (1) is reproduced here as (4) and can be expanded as (5) in the case where we have three formants per vowel. Simplifying (5) into an expression similar to the form used by L&N, we get (7).

$$\Delta F = \frac{1}{mn} \sum_j^m \sum_i^n \left[\frac{F_{ij}}{i - 0.5} \right], \text{ where } i = \text{formant number} \tag{4}$$

$$\Delta F = 1/3\ (F1/0.5 + F2/1.5 + F3/2.5) \tag{5}$$

$$\Delta F = 1/3\ (2 * F1 + 0.666 * F2 + 0.4 * F3) \tag{6}$$

$$\Delta F = 0.6667 * F1 + 0.222 * F2 + 0.1333 * F3 \tag{7}$$

Lammert and Narayanan's no-intercept expression for ΔF with three vowel formant measurements is in (8). Simplifying, we get equation (10).

$$\Delta F = 2(0.28 * F1 + (0.31 * F2)/3 + (0.47 * F3)/5) \tag{8}$$

$$\Delta F = 2(0.28 * F1 + 0.10333 * F2 + 0.094 * F3) \tag{9}$$

$$\Delta F = 0.56 * F1 + 0.20666 * F2 + 0.188 * F3 \tag{10}$$

Now we can compare the coefficients used in the Lammert and Narayanan (2015) calculation for three formant vowels (10) with the average spacing formula (7). The L&N formula has a larger coefficient for F3 and smaller coefficients for F1 and F2, than are found in the average spacing expression. This is likely to lead to better vocal tract length estimates when only one or two vowel tokens are available because the frequency of F3 is less variable across vowels than are F1 and F2. However, as seen in the left panel of **Figure 2**, with a full set of data, this method (formula 8) produces vocal tract length estimates that differ from the estimates given by the average-spacing formula (4) by only

[1] The measurement errors reported for these estimates of vocal tract length are about 10% of the length of a typical vocal tract, which may seem a bit large. It is, however, a small enough error to be able to correctly classify speakers in the Hillenbrand et al. (1995) data set as 'male,' 'female,' or 'child' with 90% accuracy based on the L&N estimate of vocal tract length (formula 8). This was measured using the methods for support vector machine classification that are described below for study 1.

[2] Though the coefficients for four formant data using Lammert & Narayanan's (2015) method are published in their paper, for completeness and ease of reference they are repeated here: with intercept (formula 11): $\beta_0 = 52$, $\beta_1 = 0.078$, $\beta_2 = 0.099$, $\beta_3 = 0.101$, $\beta_4 = 0.609$, and without intercept (formula 8): $\beta_1 = 0.089$, $\beta_2 = 0.102$, $\beta_3 = 0.121$, $\beta_4 = 0.669$.

a constant (probably due to the difference between the acoustic effective length of the vocal tract versus the actual length of the vocal tract; Johnson, 2011, p. 42).

Lammert and Narayanan (2015) also fit a formula for estimating vocal tract length that includes an intercept term. This provides a slightly more accurate measure of vocal tract length with a small set of vowel tokens, by regularizing the estimated vocal tract length. This is illustrated in the right panel of **Figure 2**, which compares the Lammert and Narayanan VTL estimates (formula 2) using a regression formula with an intercept term (11), to the VTL estimates produced by average spacing (4) for the Hillenbrand et al. data. The main thing to notice is that the range of VTL estimates for formula (11) is smaller (min = 13.6 cm, max = 16.75cm) than the range given by the non-regularized formula (8).

$$\Delta F = 2(262 + 0.14 * F1 + (0.16 * F2)/3 + (0.25 * F3)/5) \quad (11)$$

$$\Delta F = 524 + 0.28 * F1 + 0.10666 * F2 + 0.1 * F3 \quad (12)$$

The regression formula is simplified as (12). Notice that the coefficients for F1-3 are about half as large as in the no-intercept formula (10). If we enter F1 = 500, F2 = 1500, and F3 = 2500 into formula (12) we get 1074 Hz. So, about one half of the value of ΔF is determined by the formant frequencies and the rest is determined by the intercept ($\beta_0 = 524$). This will tend to shrink the range of ΔF, and with it the range of estimated vocal tract lengths. The formula in (11) results in better vowel classification when vocal tract length is estimated from only a few vowel tokens and was used in the analyses reported here. Unsurprisingly, using Lammert and Narayanan's no-intercept estimate of ΔF (formula 8) resulted in classification accuracy that was almost identical to that reported for the average-spacing ΔF method.

3.4. Study 1: Classification performance

For each normalization method, normalization factors (ΔF, log(mean), *SD*, etc.) were calculated over all vowel tokens for a talker, and then the talker's vowel formant frequencies were normalized using the code in **Table 1**. After all of the vowel formant frequencies in the dataset were normalized, support vector machines (SVM, Cortes & Vapnik, 1995) were used to classify the vowels. SVM is a supervised machine learning technique that, in this case, can be used to find optimal classification boundaries between the vowel categories given the data. The results of SVM classification can be used to infer an objective estimate of the level of vowel separation in the data. The classifiers built in this study used a radial basis function with gamma = 0.5. These are the default parameters for the R function svm(). Classification performance was evaluated by using the trained model to predict the category membership of each item in the training set and the percent correct vowel classification was computed from these predictions.

The normalized vowels were also used to build SVM classifiers to identify the talker group (male, female, child) of each token. Assuming that the goal in normalization is to remove talker differences, then a lower percent correct talker classification is an indication of better normalization performance.

The studies in this paper neglect two sources of vowel intrinsic information that have been shown to be useful both in automatic speech recognition and human speech perception—fundamental frequency of voicing (Fujisaki & Kawashima, 1968), and vowel inherent spectral change (Nearey & Assman, 1986). Thus, the classification results here are a minimal baseline.

3.5. Study 2: Effects of sample size

As in study 1, the vowel formant data to be classified in the second study were the full Hillenbrand et al. (1995) or the Peterson and Barney (1952) datasets. However, where in

study 1 the normalization factors were calculated over all of the tokens for a talker, in this study the vowels were normalized with scale factors that were computed from randomly selected subsets of the available vowel data.

Different tests were conducted with normalization based on different numbers of vowel tokens per talker. Subsets of size 1, 2, 4, 6, or 9 tokens per talker were tested in the Hillenbrand data set, and subsets of size 1, 2, 5, 10, 15, and 20 tokens per talker were tested in the Peterson and Barney data set. In each test, a random subset of vowel tokens was selected and then normalization scale factors for each of the several normalization methods were computed from the randomly selected tokens. These normalization scale factors were then used to normalize the full set of vowel tokens (all of the tokens in the corpus), and as in study 1, SVM vowel classifiers were then built and the classification accuracy over the test data set was noted. This process was repeated 50 times in each test to give an estimate of the variability of the accuracy scores for a given normalization subset size.

When the subset was a single token, the formant 'mean' that was used in the non-uniform normalization methods was set equal to the value of that formant in the selected token. The standard deviation in the Lobanov method was also set equal to the formant value when only a single token was the basis for normalization parameters.

Two additional tests were conducted. In one, a single token of the vowel [ə] was used to calculate normalization factors, and in the second, formants of the corner vowels [i, u, ɑ] were used to define the normalization statistics.

4. Results

4.1. Study 1: Classification accuracy

As shown in **Table 2**, vowel classification with ΔF (average-spacing or Kirlin's VTL method) is close to state of the art, with vowel classification accuracy approaching 90% correct for the Peterson and Barney (1952) data set, and about 80% correct for the Hillenbrand et al. (1995) data set. Also, reduction of talker information in vowels was comparable to the state of the art achieved by the non-uniform normalization methods proposed by Nearey (1978), Lobanov (1971), and Watt & Fabricius (2002). The Lammert & Narayanan (2015) method

Table 2: SVM percent correct identification of vowels and talker group (man, woman, child [MWC], or man, woman, boy, girl [MWBG]) in the vowel formant data reported by Peterson and Barney (1952; PB52) and Hillenbrand et al. (1995; H95) by different vowel normalization methods. Non-uniform vowel normalization refers to methods that use a different normalization factor for each formant. Uniform normalization refers to methods that use a single uniform scaling factor (such as vocal tract length) for all of the formants produced by a person.

Method	Type	Vowels		Talker Group	
		PB52	H95	PB52	H95
Chance	—	12.5%	10%	40%	30%
No normalization (F1 & F2)	NONE	77.3	62.9	66.7	53.2
ΔF (Nordström & Lindblom)	Uniform	82.5	72.7	49.8	41.9
ΔF (Lammert & Narayanan)	Uniform	86.3	73.4	59.6	47
Mean log Fs (Nearey 2)	Uniform	87.9	77.8	52.0	43.2
ΔF (Kirlin)	Uniform	88	78.4	52.2	41.7
ΔF (average-spacing)	Uniform	88.2	78.1	50.9	42.9
Mean log F (Nearey 1)	Non-uniform	90.9	80.1	51.6	42.4
Mean F, ratio (Watt & Fabricius)	Non-uniform	90.8	80.7	50.8	41.4
Z-score normalization (Lobanov)	Non-uniform	92.6	84.4	49.3	39.8

of VTL estimation (with intercept) is less accurate. The Lammert & Narayanan no-intercept method (not shown in the table) is virtually the same as the average spacing method.

Vowel classification with average-spacing ΔF normalization is much better than with Nordström & Lindblom's (1975) method of VTL estimation, but not quite as good as the best non-uniform normalization methods. The ΔF method is also quite comparable with the other models in removing much of the talker group information (e.g., man, woman, child). It is worth noting, however, that chance 'talker group' identification (calculated with 1000 random permutations of the datasets) is 40% correct in the Peterson and Barney (1952) corpus, and 31% correct in the Hillenbrand et al. (1995) corpus, so none of the methods fully 'normalize out' this information. This is a neglected point in vowel normalization studies (but see the recent insightful discussion in Barreda & Nearey, 2018, who reference a discussion in Hindle, 1978). We 'normalize' vowel formant measurements, but gender and age differences are not fully removed. Knowing when a method 'over-normalizes' and removes talker variability that actually 'should' remain because it is sociolinguistically significant, is an important problem. This highlights the need for a principled approach, grounded in acoustic theory, and also means that regardless of the normalization method used, the residue of talker variation left behind by normalization should be statistically modeled, because we can be sure that some talker information remains.

Figure 3 shows plots of normalized and unnormalized vowel formants for the Hillenbrand et al. (1995) data set. These plots show that vocal tract length normalization using the ΔF method results in a normalized vowel space that is remarkably similar to spaces obtained with non-uniform methods that require more, and less interpretable, parameters. Note that the center of the vowel space is marked by the horizontal and vertical lines. For the Nearey, Lobanov, and Watt & Fabricius methods these lines mark the mean F1 and F2; for the ΔF methods they mark the resonances of the uniform tube.

This analysis indicates that vocal tract length normalization using a single interpretable normalization parameter (ΔF—an estimate of formant spacing in a vocal tract with no constrictions) is comparable to other vowel normalization methods. The talker-independent dimension that is used in this normalized representation is derived from both the formants being normalized (F1 and F2), as well as higher formants (F3 in this case, and also F4; Lammert & Narayanan, 2015), which are less likely to vary as a function of the particular inventory of vowels found in a language. Higher formants, F3 and F4, are sometimes not measured. This study suggests that they should be, and that this additional information about the talker's vocal tract could be extremely valuable in interpreting the F1/F2 vowel space.

Because the normalization factor ΔF is directly interpretable in terms of a physical property of the talker, vocal tract length normalization is valid for cross-linguistic comparison of vowel spaces. In addition, it is remarkable that the state of the art in vowel formant normalization is almost entirely reducible to normalization in terms of the talker's vocal tract length. This has not been observed before because phoneticians have not used a measure calibrated to vocal tract length in our vowel normalization schemes. Arguably, Nearey's (1978) uniform normalization technique, with mean logF*, uses a measure that reflects vocal tract length, although because it is not calibrated to vocal tract length it is a problematic measure, giving incomparable measurement scales for values normalized over mean logF* (F1..F4) versus values normalized over mean logF* (F1..F3). The ΔF measurement scale is the same whether ΔF is estimated from three formants or four.

4.2. Study 2: Effects of sample size

Beyond the practicalities of having a workable vowel normalization scheme for comparing dialects and languages, the discovery that vocal tract length normalization is a powerful method for reducing some of the talker variability found in speech leads one to wonder

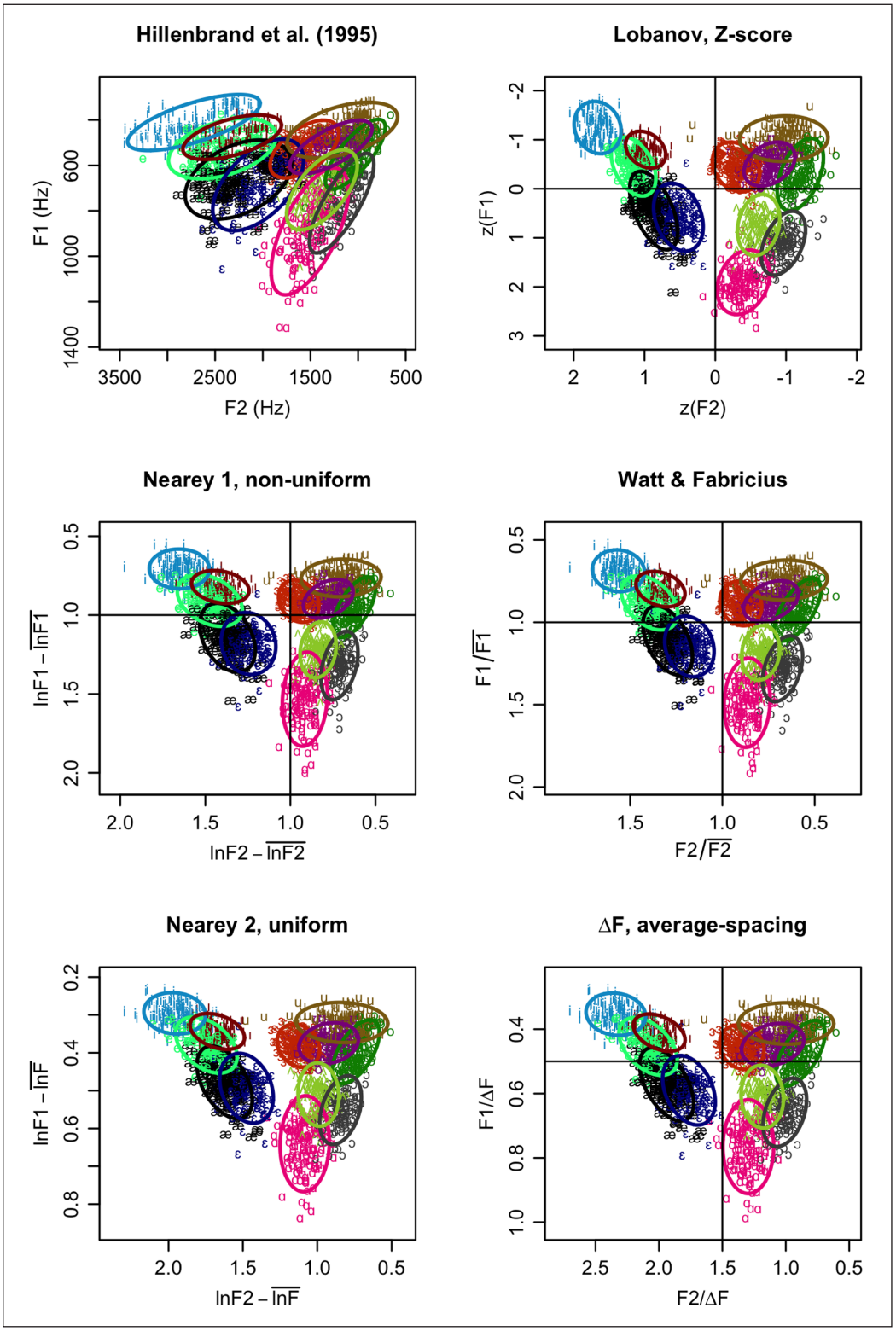

Figure 3: Upper left: F1 and F2 vowel formant frequencies from Hillenbrand et al. (1995). Other panels: the same data normalized by several of the methods identified in the text.

whether vocal tract length normalization might play a role in speech perception. Study 2 tested this by limiting the amount of information available to the normalization algorithms in tests of vowel classification using the Hillenbrand et al. (1995) and Peterson and Barney (1952) datasets. Limiting the information available for normalization creates a situation that Barreda and Nearey (2018) call 'type B' over-normalization which is "due to noise in the estimated speaker parameters used for normalization."

Figure 4 shows the results of study 2 on the Hillenbrand et al. (1995) data set, with normalization based on different numbers of randomly selected vowels (1, 2, 4, 6, or 9 tokens), or basing vowel normalization on the corner vowels [i], [u], and [ɑ], on schwa [ə], or on the entire set of observations from each talker (12 tokens). For these later models (12 tokens, schwa, or corner vowels) there was no repeated random selection of a basis for extrinsic normalization, so only one SVM was fit for each normalization method.

Figure 5 shows analogous results for the Peterson and Barney (1952) dataset. With both the HIllenbrand et al. (1995) data and the Peterson and Barney (1952) data, the results show that uniform scaling techniques (Nearey's uniform scaling method, the ΔF methods—Kirlin, Ave-spacing, and Lammert & Narayanan) improve vowel classification accuracy over unnormalized classification even with a very small random sample of speech. While non-uniform methods, i.e., those that scale each formant using information in the corpus about that formant (*z*-score normalization [Lobanov], mean ratio [Watt & Fabricius], or log mean difference [Nearey], non-uniform), are more dependent upon the particular vowel tokens that are used to calculate the vowel normalization factors, and need either a carefully chosen sample, or a large sample. Random selection of vowel tokens, as done here, has a catastrophic effect on the non-uniform methods if only a few vowel tokens are taken to represent the talker. In practice, where a large corpus of vowel measurements is available, this does not matter, and in fact the non-uniform methods may reduce talker differences better than the uniform methods in corpus analysis. But these methods are brittle and are generally inappropriate as models of the perceptual process.

Testing the Watt & Fabricius method with randomly selected vowel tokens, as done here, especially goes against the spirit of that method because its main feature is a judicious selection of vowel tokens to represent the full possible range of F1 and F2

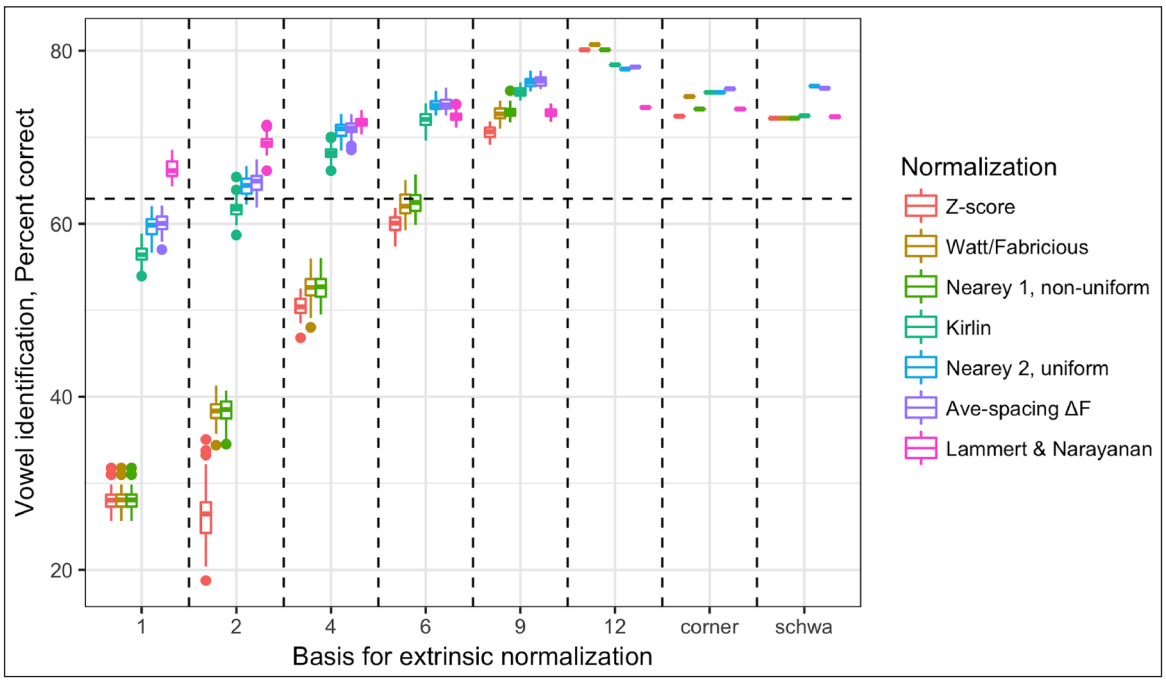

Figure 4: Results of SVM vowel classification of the Hillenbrand et al. (1995) vowels when the vowel normalization statistics are calculated over different numbers of randomly selected vowel tokens, or sets designed to be uniquely informative about the vowel space (the corner vowels) or the vocal tract (schwa). The order of the bars within each panel is indicated in the legend, reading top to bottom for the bars from left to right—e.g., Lobanov normalization is the leftmost bar in each panel. Classification with no normalization results in 62.9% correct vowel identification (see Table 2), which is indicated by the dashed horizontal line.

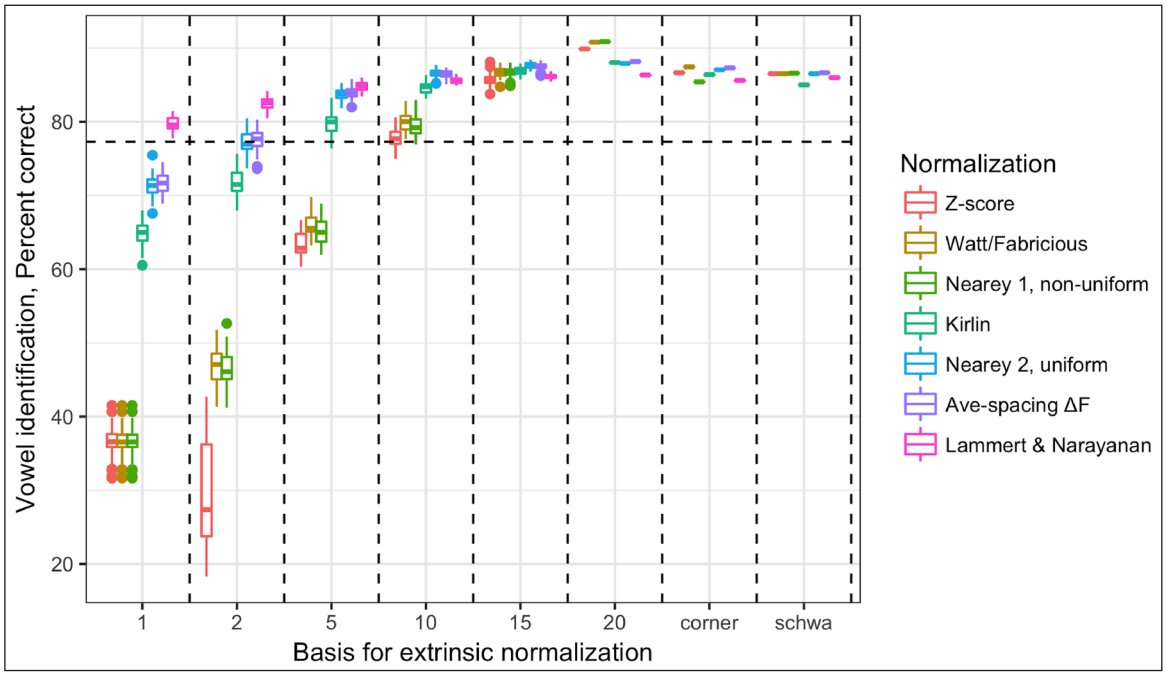

Figure 5: Results of SVM vowel classification of the Peterson and Barney (1952) vowels when the vowel normalization statistics are calculated over different numbers of randomly selected vowel tokens, or sets designed to be maximally informative about the vowel space (the corner vowels) or the vocal tract (schwa). Classification with no normalization results in 77.3% correct vowel identification (see Table 2), which is indicated by the dashed horizontal line.

for a talker. However, random selection of tokens is justified in this study because it is designed to evaluate the plausibility of extrinsic normalization as a component of speech perception. Listeners are not presented with a judicious selection of vowel tokens, but have to deal with whatever the talker says. It is worth noting that even without judicious selection of vowel tokens, Watt & Fabricius' mean ratio representation has very good vowel classification performance when a large sample of tokens is taken; however the normalization scale may depend on the specific vowel inventory. Regardless, the Watt & Fabricius method was not designed as a model of perceptual vowel normalization and is clearly not a feasible one.

5. Conclusion

This study introduced a new method of vowel normalization, the ΔF method. This is explicitly a vocal tract length normalization method, and represents vowels on a talker-independent measurement scale—the average formant spacing of the talker, their ΔF. Using a metric that is closely related to the length of the talker's vocal tract, we are able to produce a vowel space in which vocal tract length effects have been removed. The resulting vowel space is largely equivalent to Nearey's (1978) log-mean uniform vowel normalization method, but is explicitly rationalized in terms of vocal tract length and has a consistent unit of measure whether 3 formant or 4 formant measurements are used. ΔF normalization based on Lammert and Narayanan's (2015) intercept method of vocal tract length estimation is more robust when only a few tokens are available for a talker, but provides less vowel category separation in models built over a larger corpus.

This study has also shown that non-uniform methods, that rely on within-formant scale factors (Nearey's log-mean non-uniform method, Lobanov's *z*-score normalization, and Watt & Fabricius' mean ratio method) are highly sensitive to the vowel tokens that are chosen as the basis for calculating the normalization scale factors. This leads to the

conclusion that these methods are not plausible as models of the cognitive processes involved in speech perception. Also, in practical phonetic description of languages this dependence on particular vowel tokens for calculating normalization scale factors complicates cross-linguistic or cross-dialect comparisons of vowels spaces. If languages have different vowel inventories, then the unit of measure for normalized vowels will not be comparable across languages. This leads one to try to determine formant ranges indirectly (Fabricius, Watt, & Johnson, 2009) in order to have normalized values that can be cross-linguistically compared. Vocal tract length normalization does not have as much of a problem with this because it is less sensitive to the composition of the vowel inventory, and not very sensitive to the particular vowel tokens that represent a talker.

The practical conclusion is that ΔF normalization likely has advantages over other methods for speech researchers. It produces good classification accuracy, is robust to sample size variation across talkers, is independent of the vowel inventory or phonetic vowel realizations in the language or dialect studied, puts all talkers, regardless of language or dialect, on the same measurement scale, and is rationalized in terms of the acoustic theory of speech production.

Finally, the results of this study also encourage us to think that listeners may be able to employ a type of vocal tract length normalization with very little extrinsic context. Non-uniform normalization schemes are untenable when faced with a single isolated vowel, but a normalization scheme in which vocal tract length is estimated from the entire spectrum of a vowel (see e.g., Wakita, 1977; Bladon, Henton & Pickering, 1984; Lee & Rose, 1998), does seem to be a plausible perceptual mechanism even with isolated vowels. Future research may bear this out.

Acknowledgements

This paper is dedicated to Mary Beckman. Thanks to Adam Lammert for sharing VTL estimation coefficients for three-formant vowel data and for extensive comments on an earlier version of the manuscript. Thanks also to Santiago Barreda and two anonymous reviewers for comments on an earlier version of the manuscript. The work here was made possible in part by Barreda's collection of datasets and analysis/visualization tools that he has made available in the **phonTools** R package.

Competing Interests

The authors have no competing interests to declare.

References

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

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

Bladon, R. A. W., Henton, C. G., & **Pickering, J. B.** (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, *4*, 59–69. DOI: https://doi.org/10.1016/0271-5309(84)90019-3

Cortes, C., & **Vapnik, V. N.** (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. DOI: https://doi.org/10.1007/BF00994018

Disner, S. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, *67*, 253–261. DOI: https://doi.org/10.1121/1.383734

Eide, E., & **Gish, H.** (1996). A parametric approach to vocal tract length normalization. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*

Conference Proceedings, *1*, 346–348. Atlanta, GA, USA. DOI: https://doi.org/10.1109/ICASSP.1996.541103

Fabricius, A., Watt, D., & **Johnson, D. E.** (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change*, *21*, 413–35. DOI: https://doi.org/10.1017/S0954394509990160

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fujisaki, H., & **Kawashima, T.** (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*, 73–77. DOI: https://doi.org/10.1109/TAU.1968.1161952

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

Harrington, F. H., & **Mech, L. D.** (1979). Wolf howling and its role in territory maintenance. *Behaviour*, *68*, 207–249. DOI: https://doi.org/10.1163/156853979X00322

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

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In D. Sankoff (Ed.), *Linguistic variation: Models and methods* (pp. 161–171) New York: Academic Press.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, *88*, 642–654. DOI: https://doi.org/10.1121/1.399767

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing*. (pp. 145—166). San Diego, CA: Academic Press.

Johnson, K. (2005). Speaker Normalization in speech perception. In D. B. Pisoni & R. Remez, (Eds.), *The handbook of speech perception* (pp. 363—389). Oxford: Blackwell Publishers. DOI: https://doi.org/10.1002/9780470757024.ch15

Johnson, K. (2011). Acoustic and Auditory Phonetics, 3rd Edition. Boston: Wiley-Blackwell.

Johnson, K., & **Sjerps, M.** (to appear). Speaker Normalization in Speech Perception.

Johnson, K., Strand, E. A., & **D'Imperio, M.** (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384. DOI: https://doi.org/10.1006/jpho.1999.0100

Kirlin, R. L. (1978). *A posteriori* estimation of vocal tract length. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 571–574. DOI: https://doi.org/10.1109/TASSP.1978.1163151

Ladefoged, P., & **Broadbent, D. E.** (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *39*, 98–104. DOI: https://doi.org/10.1121/1.1908694

Lammert, A. C., & **Narayanan, S. S.** (2015). On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLoS ONE*, *10*(7): e0132193. DOI: https://doi.org/10.1371/journal.pone.0132193

Lee, L., & **Rose, R. C.** (1998). A frequency warping approach to speaker normalization. *IEEE Transactions on Speech & Audio Processing*, *6*(1), 49–60. DOI: https://doi.org/10.1109/89.650310

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

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

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

Nearey, T. M., & **Assmann, P. F.** (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*, 1297–1308. DOI: https://doi.org/10.1121/1.394433

Nordström, P. E., & **Lindblom, B.** (1975). A normalization procedure for vowel formant data. *Proceedings of the 8th international congress of phonetic sciences*. Leeds, England.

Paige, A., & **Zue, W. Z.** (1970). Calculation of vocal tract length. *IEEE Transactions on Audio and Electroacoustics*, *18*, 268–270. DOI: https://doi.org/10.1109/TAU.1970.1162113

Peterson, G. E., & **Barney, H. L.** (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, *24*, 175–184. DOI: https://doi.org/10.1121/1.1906875

Reby, D., & **McComb, K.** (2003). Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behavior*, *65*, 519–530. DOI: https://doi.org/10.1006/anbe.2003.2078

Strand, E. A., & **Johnson, K.** (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural Language Processing and Speech Technology. Results of the 3rd KOVENS Conference, Bielefeld, October, 1996* (pp. 14–26). Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110821895-003

Strange, W., Jenkins, J. J., & **Johnson, T. L.** (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*(3), 695–705. DOI: https://doi.org/10.1121/1.389855

Van Lanker, D. R., Kreiman, J., & **Cummings, J.** (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 665–674. DOI: https://doi.org/10.1080/01688638908400923

Wakita, H. (1977). Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*, 183–192. DOI: https://doi.org/10.1109/TASSP.1977.1162929

Watt, D., & **Fabricius, A.** (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. Retrieved from http://www.leeds.ac.uk/linguistics/WPL/WP2002/Watt_Fab.pdf

How to cite this article: Johnson, K. 2020 The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 10, pp. 1–16. DOI: https://doi.org/10.5334/labphon.196

Submitted: 12 March 2019 **Accepted:** 20 January 2020 **Published:** 22 July 2020

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS