UC Berkeley

UC Berkeley PhonLab Annual Report

Title

Vocal Tract Length Normalization

Permalink

<https://escholarship.org/uc/item/16c753jz>

Journal

UC Berkeley PhonLab Annual Report, 14(1)

Author

Johnson, Keith

Publication Date

2018

DOI

10.5070/P7141042475

Copyright Information

Copyright 2018 by the author(s). All rights reserved unless otherwise indicated. Contact the author(s) for any necessary permissions. Learn more at <https://escholarship.org/terms>

Peer reviewed

Vocal tract length normalization

Keith Johnson

Department of Linguistics
University of California Berkeley
USA

keithjohnson@berkeley.edu

Abstract: The resonant frequencies of the vocal tract during vowel production convey information about the linguistic vowel intended by the talker - whether they mean to say 'hey' or 'hoe', for example - while also conveying information about the talker. One particularly salient bit of talker information that partially determines the frequencies of the vowel formants is the length of the talker's vocal tract. Vowel formant normalization aims to remove the effects of talker differences without also removing important linguistic information. This paper presents a study of vocal tract length normalization using a new ΔF method, and compares this method to other vowel normalization methods. A key point of comparison in this study is the number of vowel tokens that are needed in order to derive a stable estimate of vocal tract length. Several of the vowel normalization methods that are most commonly used in phonetic studies are shown to need a full set of vowels in order to be reliable, while methods that derive vocal tract length information from the full acoustic spectrum are much more stable and may even provide a length-normalized representation that could be cognitively computed and used in human speech perception.

key words: 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 flow of air through the larynx and vocal tract. In addition to these controlled actions, speech acoustics are determined by the overall size of the vocal tract and larynx. This paper revisits a central topic in the study of speech perception, perceptual talker normalization (Johnson, 2005), with the goal of exploring an improved approach to vocal tract length normalization.

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; Lammert & Narayanan, 2015). What is needed in order to do this is a good estimate of the vocal tract resonance frequencies -- the vowel formants -- for a talker, and then the talker-specific vowel formant distributions can be used to estimate a normalization factor. Several methods to do this have been proposed, with more or less reference to acoustic theory (Gerstman, 1968; Lobanov, 1971; Nordström & Lindblom, 1975; Nearey, 1978; Watt & Fabricius, 2002). Normalization makes it possible to express the formant frequencies using a measurement scale that is independent of vocal tract length differences, which is of practical use in comparing dialects and languages with each other, and may also be a part of the cognitive process of speech perception.

Whether listeners make use of vocal tract length normalization 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 talker differences suggests that there is more going on 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 responds to 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 posit a perceptual process of vocal tract length normalization. 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 et al., 1989; Johnson & Sjerps, to appear), suggesting that the perceptual system may include processes that compute talker information

which can be mixed with phonetic information in a stream that produces ‘talker neutral’ phonetic information.

Regardless of the cognitive status of vocal tract length normalization, 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 et al., 2004). Methods used in these studies are sometimes based on general-purpose statistical normalization techniques such as range normalization (Gerstner, 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 very popular (Nearey, 1978; Watt & Fabricius, 2002). In vocal tract length normalization the normalization scale factor is derived from an estimate of the length of the speaker’s vocal tract. 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 it was proposed (Nordström & Lindblom, 1975) because it didn’t seem to work very well. However, there has been significant progress in deriving accurate vocal tract length estimates from the acoustic vowel spectrum, so a reconsideration of vocal tract length normalization is due.

This paper will introduce the ΔF method of vocal tract length normalization, and compare that method with the existing state of the art (section 2). An additional goal of the paper is to determine the feasibility of vocal tract length normalization as a perceptual mechanism, so in section 3 simulations using talker information derived from samples of various sizes will test the applicability of the ΔF method as a possible perceptual mechanism. After a discussion of general conclusions in section 4, the methods of the study are presented in some detail in section 5.

2. Vocal Tract Length Normalization

The earliest use of vocal tract length as a normalization factor was by Nordström & Lindblom (1975). They calculated the average F3 frequency in open vowels (where the frequency of F3 is not influenced by F2), and used this to estimate the talker’s vocal tract length. Vocal tract length was then be used to scale a talker’s vowel formant measurements onto a “standard” (i.e. male) vocal tract. We can avoid the sexism of this approach by using a talker-independent measurement scale -- the mean distance or spacing between the formants, the ΔF .

Lammert & Narayanan (2015) published an important study that compared predicted versus actual vocal tract length using MRI estimates of vocal tract length, as well as computer simulated vocal tracts of known length. They found a family of regression formulas that predict vocal tract length from all of the formant frequencies of vowels, not just the F3. Note also, in this regard, Barreda & Nearey’s (in press) regression approach.

Figure 1. An illustration of Reby & McComb's (2003) line fitting 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.

Following the analysis outlined by Lammert & Narayanan (2015) and building on the line-fitting approach of Reby & McComb (2003, see Figure 1), a talker's ΔF is calculated by scaling formant frequencies by their formant number (F1, F2, F3, F4, etc.) as in formula (1). Note that each formant (F1-F4) of each vowel provides an estimate of ΔF and the sum in (1) can be taken over all of the vowels for a talker that are available in a corpus. The talker's ΔF value is related to their vocal tract length by formula (2), and the normalized vowel formants are calculated using ΔF as in (3).

- 1) $\Delta F = 1/mn \sum_j \sum_i [F_{ij}/(i - 0.5)]$, where i = formant number (1...4), and j is token number.
- 2) $VTL = c/2\Delta F$, $c = 35000$ cm/s, the speed of sound
- 3) $norm(F_{ij}) = F_{ij}/\Delta F$

Figure 2 indicates that the line-fitting method of estimating the length of the vocal tract is comparable to Lammert & Narayanan's (2015) estimated vocal tract length (using their 'no intercept' formula, see the detailed methods in section 5 below), and both of these are much more in agreement with each other than they are with the method given by Nordström & Lindblom (1975) which was based on F3 alone.

As mentioned above, ΔF can also be used as a unit of measure for vowel formant frequencies, putting all vocal tracts on the same measurement scale. The normalized F1 value is given as $F1/\Delta F$ and is expected to be equal to 0.5 for a uniform tube of any length, the normalized F2 is $F2/\Delta F$ and is expected to be equal to 1.5 for a uniform tube, and so on (formula 3). This is ΔF vocal tract length normalization, where ΔF can be estimated in several different ways - from F3 alone, using the line-fitting approach, or using Lammert & Narayanan's regression fits.

Figure 2. Comparison of three different methods for calculating vocal tract length (VTL) for the talkers in the Hillenbrand et al., (1995) corpus of vowel formant measurements. The line-fitting method is given in formulas (1) and (2) of this paper. The method specified by Lammert & Narayanan (2015) was optimized over MRI measurements of actual vocal tract length. This figure used their ‘no intercept’ formula, see section 5. The dashed line is the identity line, $y=x$.

As shown in Table 1, when F1 and F2 are normalized by the line-fitting ΔF method we have vowel classification and talker normalization accuracy that is close to the state of the art achieved by the non-uniform normalization methods proposed by Nearey (1978), Lobanov (1971), and Watt & Fabricius (2002).

Vowel classification with line-fitting ΔF normalization, using Support Vector Machine classification (see section 5, “Methods”), 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 line-fitting ΔF method is also quite comparable with the other models in removing much of the talker group information (eg. 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 & 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, in press, 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 a huge problem. This highlights the need for a principled approach, grounded in acoustic theory. This 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.

Table 1. 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 scaling factor (such as vocal tract length) for all of the formants produced by a person.

Method	Type	PB52 (vowels)	H95 (vowels)	PB52 (MWC)	H95 (MWBG)
No normalization (F1 and F2)	NONE	77.3	62.9	66.7	53.2
ΔF (Nordström & Lindblom)	Uniform	82.5	72.7	49.8	41.9
Mean log Fs (Nearey 2)	Uniform	87.9	77.8	52.0	43.2
ΔF (line-fitting)	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

Figure 3 shows plots of normalized and unnormalized vowel formants for the Hillenbrand et al. (1995) data set. Visual inspection of these plots indicates 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 -- the average formant spacing for a talker) 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, F4, etc), 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.

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

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 hasn't been observed before because phoneticians haven't used an accurate measure of vocal tract length in our vowel normalization schemes. Arguably, Nearey's (1978) uniform normalization technique, with mean $\log F^*$ uses a measure that reflects vocal tract length, although because it is not calibrated to vocal tract length it is a problematic measure, giving an incomparable measurement scale for values normalized over mean $\log F^*$ ($F1..F4$) versus values normalized over ($F1..F3$), for example.

3. Vocal tract length normalization for speech perception

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 whether vocal tract length normalization might play a role in speech perception.

One reason to believe that vocal tract length normalization, like any extrinsic vowel normalization method, isn't a viable mechanism for speech perception is that it relies on information that is extrinsic to the vowel. That is, the estimation of ΔF is derived over a collection of vowel tokens - information beyond what is available intrinsically in the vowel to be classified.

There is evidence that listeners use extrinsic information in vowel normalization. For example, Ladefoged & Broadbent (1957) showed that contextual vowel formant range influences vowel perception, which suggests that an estimate of vocal tract length might develop over the course of an utterance. Nonetheless, isolated vowels are accurately recognized (Nearey, 1989; Strange, Jenkins, & Johnson, 1983), which suggests that each vowel contains the information that is needed for its own recognition, and this intrinsic information is usually sufficient for vowel recognition. The studies in this section neglect two sources of vowel intrinsic information that have been shown to be useful - $F0$ (Fujisaki & Kawashima, 1968), and vowel inherent spectral change (Nearey & Assman, 1986). Thus, the classification results here are a minimal baseline.

Lammert & Narayanan's (2015) finding is relevant for this discussion. They found that estimation of vocal tract length over short stretches of speech can be quite accurate when the $F3$ and $F4$ are included in the estimation of vocal tract length. In the line-fitting method as well, information from all of the formants is used to estimate the vocal tract length, because the line-fitting method (as shown in Figure 1) can make use of as many formants as can be reliably measured (Reby & McComb, 2003 used up to $F8$ to measure vocal tract length in red deer). Note that the regression formulas reported by Lammert & Narayanan (2015) are not as flexible as this because separate regression models must be fit for a specific number of formants.

By using of all of the formants in a vowel, vocal tract length normalization should be less reliant on extrinsic context 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 can't accurately estimate the distribution from only a few instances, so, the Lobanov normalization method can't be a model of how listeners identify (and normalize) isolated vowels. On the other hand, in ΔF normalization, information about all of the formants can be used to determine a normalization scale factor, and thus each vowel contains a formant pattern within which to evaluate (and normalize) the formants.

I conducted a test of these ideas by limiting the amount of information available to the normalization algorithms in tests of vowel classification using the Hillenbrand et al. (1995) and Peterson & Barney (1952) datasets. Limiting the information available for normalization creates a situation that Barreda and Nearey (in press) call "type B" over-normalization which is "due to noise in the estimated speaker parameters used for normalization".

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 maximally 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 - eg. Lobanov normalization is the leftmost bar in each panel. Classification with no normalization results in 62.9% correct vowel identification (see Table 1), which is indicated with the dashed horizontal line.

For example, in one test with the Hillenbrand et al. data, shown in the panel labeled "9" in Figure 4, nine vowel tokens (of the 12 vowel tokens for each talker) were randomly selected as a basis for calculating normalization statistics. Thus, for instance, nine randomly selected tokens were used to calculate the mean and standard deviation of F1 and F2 for each talker and then these

statistics were used to z-score normalize the talker's vowels. Similarly, the formants from the nine randomly selected tokens were used to calculate normalization scale factors in the other methods listed in Figure 4 and these were then used to normalize all of the vowels produced by the talker. As in the analyses in section 2 above, I used the normalized data to build support vector machine (SVM) classifiers and calculated the percent correct vowel classifications for several normalization algorithms. The random selection was repeated 50 times and the box and whisker plots in Figures 4 and 5 show the distributions of obtained vowel identification performance.

Figure 4 shows the results of normalizing based on different numbers of randomly selected vowels (1, 2, 4, 6, 9), or basing vowel normalization on the corner vowels [i], [u] and [a], on schwa [ʌ], or on the entire set of observations from a talker (12). For these later models (the ones with no error bars in figure 3) there was no repeated random selection of a basis for extrinsic normalization, so only one SVM was fit for each normalization method.

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 1), which is indicated with the dashed horizontal line.

Figure 5 shows results from an analogous study of the Peterson & Barney (1952) dataset. With both the Hillenbrand et al. (1995) data and the Peterson & Barney (1952) data, the results show that uniform scaling techniques (Nearey's uniform scaling method, the line-fitting ΔF method, and ΔF with Lammert & Narayanan's method of vocal tract length estimation) improve vowel

classification accuracy over unnormalized classification even with a very small random sample of speech. While non-uniform methods that scale each formant based on 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 chosen as the basis of the vowel normalization algorithm, 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 vowels 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 it is important to recognize the brittleness of these methods, and their inappropriateness as models of the perceptual process.

Testing the Watt and Fabricius (2002) method in this way, with randomly selected vowel tokens, especially goes against the spirit of that method because its main feature is a judicious selection of vowel tokens to represent the full possible ranges of F1 and F2 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.

4. 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 an accurate measurement of 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) method of vocal tract length estimation is more robust when only a few tokens are available for a talker.

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 would obviously suggest that these methods are not plausible as models of the cognitive processes involved in speech perception, but even in practice for 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 a language has no vowel that uses a very low F2, for

example, we may need to try to guess how low F2 could in theory go for the talker (Fabricius, Watt & Johnson, 2009) in order to have normalized values that can be cross-linguistically compared. Vocal tract length normalization does not have this problem because it is not sensitive to the composition of the vowel inventory, and not very sensitive to the particular vowel tokens that represent a talker.

The practical conclusion for speech researchers is that ΔF normalization is preferable to other methods. 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), has at least a starting point for vocal tract length normalization even within an isolated vowel. Future research may bear this out.

5. Methods

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

Normalization formulas. The data were analyzed in the R statistical programming language, and the normalization algorithms were implemented as illustrated in Table 2. As the variable names in the last three rows of the table make clear, the normalizing factor ΔF can be calculated by any method that estimates 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, and avoids subjectivity or other mistakes in the selection of the exemplary vowel tokens.

Vocal tract length coefficients. The regression coefficients used in this paper for the Lammert & Narayanan (2015) method are different from those that 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 a zero intercept term ($\beta_0 = 262$, $\beta_1 = 0.14$, $\beta_2 = 0.16$, $\beta_3 = 0.25$). This formula (see Table 2, and formula 11 below) leads to a smaller error of estimated vocal tract length (1.22 cm). The classification

studies in section 3 used this more accurate regression formula. The VTL 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.

Table 2. *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. In section 3, normalization statistics were calculated from subsets of a talker’s vowel tokens, as illustrated for the Lobanov method. The first four rows are non-uniform methods and the last four rows are uniform methods.*

Method	R code
Lobanov, z-score	<code>normF1 = scale(f1);</code>
Lobanov, with a subset	<code>normF1 = (f1-mean(subset_f1)) / sd(subset_f1);</code>
Nearey non-uniform	<code>normF1 = exp(log(f1) - mean(log(f1)));</code>
Watt & Fabricius	<code>normF1 = f1/mean(f1);</code>
Nearey uniform	<code>mf = mean(c(log(f1),log(f2),log(f3))) normF1 = exp(log(f1)-mf);</code>
Nordstrom & Lindblom	<code>deltaf = mean(f3[f1>600]/2.5) normF1 = f1/deltaf;</code>
line-fitting ΔF	<code>deltaf = mean(c(f1/0.5, f2/1.5, f3/2.5)); normF1 = f1/deltaf;</code>
Lammert & Narayanan	<code>deltaf = 2*(262 + mean(0.14*f1) + mean((0.16*f2)/3) + mean((0.25*f3)/5)); normF1 = f1/deltaf;</code>

The comparison of the Lammert & Narayanan (2015) coefficients to the line-fitting 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).

- 4)
$$\Delta F = 1/mn \sum_j^m \sum_i^n [F_{ij}/(i - 0.5)], \text{ where } i = \text{formant number}$$
- 5)
$$\Delta F = 1/3 (F1/0.5 + F2/1.5 + F3/2.5)$$
- 6)
$$\Delta F = 1/3 (2 * F1 + 0.666 * F2 + 0.4 * F3)$$
- 7)
$$\Delta F = 0.6667 * F1 + 0.222 * F2 + 0.1333 * F3$$

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

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

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

$$10) \quad \Delta F = 0.56 * F1 + 0.20666 * F2 + 0.188 * F3$$

Now we can compare the coefficients used in the Lammert & Narayanan (2015) calculation for three formant vowels (10) with the line-fitting formula (7). The Lammert and Narayanan formula has a larger coefficient for F3 and smaller coefficients for F1 and F2, than are found in the line-fitting 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. Recall that Nordström and Lindblom (1975) disregarded F1 and F2 entirely and calculated vocal tract length from F3 alone.

Figure 6. Vocal tract length estimates for the talkers in Hillenbrand et al. (1995), as calculated using the line-fitting formula (4) and Lammert & Narayanan's (2015) zero intercept formula (11). The dashed line is the identity line, $y=x$.

Lammert & Narayanan (2015) also fit a formula for estimating vocal tract length that includes an intercept term which provides a slightly more accurate measure of vocal tract length, by regularizing the estimated vocal tract length. This is illustrated in Figure 6, which compares the Lammert & Narayanan VTL estimates using a regression formula with a zero intercept (11), to the VTL estimates produced by line-fitting (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 of the line-fitting estimates (min=12.5, max=18.4). Compare Figure 6 to Figure 2 above.

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

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

The regression formula is simplified as (12). Notice that the coefficients 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) was used in the analyses presented in sections 2 and 3, and results in better vowel classification when vocal tract length is estimated from only a few vowel tokens. Using Lammert & Narayanan's no-intercept estimate of ΔF (formula 8) resulted in classification accuracy that was almost identical to that reported for the line-fitting ΔF method.

Classification performance. After all of the vowel formant frequencies in a dataset were normalized, using each of the normalization methods, support vector machines (Cortes & Vapnik, 1995) were used to classify the vowels (or talkers). SVM is a supervised machine learning technique that finds a set of boundaries that will separate the training data into the categories that are specified in the training set. 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 classification was computed from these predictions.

Sampling. SVM classifiers were built for data sets that were normalized in different ways. The data sets were always the same size and included the same vowel tokens - all of the tokens in the Hillenbrand, et al., (1995) or the Peterson & Barney (1952) datasets. In section 2, the normalization factors were calculated over all of the tokens for a talker. In section 3, the vowels were normalized with scale factors that had been computed from subsets of the available vowel data, and then these normalized data were used to train classifiers. For example, in the 'schwa' subset (Figures 4 & 5), a token of the vowel [ʌ] by a particular talker was used to set the "mean" $F1$ and $F2$ in the Lobanov method for that talker (and in this case where only one token was available the "standard deviation" was set equal to the mean). Similarly, in the vocal tract length normalization methods, ΔF for the talker was calculated from the formants of this single token. Statistics for the other vowel normalization methods were also calculated from this single token. Then these parameters were used to compute the normalized vowel formants for all of the vowels for the talker, and this process was repeated for each of the talkers in the data set. In cases where a limited number of vowel tokens was selected, a random sample of the target size (1, 2, 4, 6, or 9 tokens per talker in the Hillenbrand data set) was selected, and then normalization scale factors were computed from those randomly selected tokens. SVM classifiers were then built using these normalized tokens and the classification accuracy over the entire data set was noted. This process was repeated 50 times and the distribution of accuracy measures is reported in the box and whisker plots in figures 4 and 5.

Acknowledgements

Thanks to Adam Lammert for fitting VTL estimation coefficients for data with measurements for only the three lowest formants, and to Santiago Barreda 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. Thanks also to two anonymous reviewers for their comments on an earlier version.

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, S. & Nearey, T.M. (in press) A regression approach to vowel normalization for missing and unbalanced data.
- Cortes, Corinna & Vapnik, Vladimir N. (1995). "Support-vector networks". *Machine Learning*. **20** (3): 273–297. doi:10.1007/BF00994018
- Bladon, R.A.W. & Henton, C.G. & Pickering, J.B. (1984) Towards an auditory theory of speaker normalization. *Language & Communication* **4**, 59-69.
- Disner, Sandra. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America* **67**, 253–261.
- Fabricsius, 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.
- Fant, G. (1960) *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.
- Fujisaki, H., and Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels. *IEEE Trans. Audio Electroacoust.* AU-16, 73-77.
- Gerstman, L. (1968) Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoustics* 16, 78-80. doi:10.1109/TAU.1968.1161953
- Harrington, F. H. & Mech, L. D. (1979) Wolf howling and its role in territory maintenance. *Behaviour*, **68**, 207–249.
- Hillenbrand, J.; Getty, L.A.; Clark, M.J. & Wheeler, K. (1995) Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* **97**(5), 3099-3111.
- Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech, *Linguistic Variation, Models and Methods*, edited by D. Sankoff (Academic Press, New York).
- Johnson, K. (1990) The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am.* **88**, 642-654.
- Johnson, K. (1997) Speech perception without speaker normalization: An exemplar model. In K. Johnson & J.W. Mullennix (eds.) *Talker Variability in Speech Processing*. (Academic Press, San Diego, pp. 145-166).
- Johnson, K. (2005) Speaker Normalization in speech perception. In Pisoni, D.B. & Remez, R. (eds) *The Handbook of Speech Perception*. (Oxford, Blackwell Publishers. pp. 363-389).
- Johnson, K. & Sjerps, M. (2018) Speaker Normalization in Speech Perception.
- Johnson, K., Strand, E.A. and D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *J. of Phonetics* **27**, 359-384.

- Ladefoged, P., and Broadbent, D. E. (1957). "Information conveyed by vowels," *J. Acoust. Soc. Am.* **39**, 98–104.
- Lammert, A.C. & Narayanan, S.S. (2015) On Short-Time Estimation of Vocal Tract Length from Formant Frequencies. *PLoS ONE* **10**(7): e0132193.
<https://doi.org/10.1371/journal.pone.0132193>
- Lee, L. & Rose, R.C. (1998) A frequency warping approach to speaker normalization. *IEEE Trans. Speech & Audio Proc.* **6**(1), 49-60.
- Lobanov, B. M. (1971) 'Classification of Russian vowels spoken by different speakers, *J. Acoust. Soc. Am.* **49**, 606–608.
- Nearey, T. M. (1978) *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club. Bloomington, Indiana.
- Nearey, T.M. (1989) Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* **85**, 2088-2113.
- Nearey, Terrance M. and Assmann, Peter F. (1986) Modeling the role of inherent spectral change in vowel identification. *J. Acoust. Soc. Am.* **80**, 1297-1308.
- Nordström, P.E. and Lindblom, B. (1975): A normalization procedure for vowel formant data. *Proceedings of the International Congress of Phonetic Sciences*. Leeds, England.
- Peterson, G.E. & Barney, H.L. (1952) Control methods used in the study of vowels. *J. Acoust. Soc. Am.* **24**, 175-184.
- 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.
- Strange, W., Jenkins, J.J. & Johnson, T.L. (1983) Dynamic specification of coarticulated vowels. *J. Acoust. Soc. Am.* **74**(3), 695-705.
- Strand, E.A. & Johnson, K. (1996) Gradient and visual speaker normalization in the perception of fricatives. In Gibbon, Dafydd (ed.) *Natural Language Processing and Speech Technology. Results of the 3rd KOVENS Conference, Bielefeld, October, 1996*. Berlin: Mouton de Gruyter, pp 14-26.
- 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: 10.1080/01688638908400923
- Wakita H. (1977) Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans Acoust Speech Sig Proc.* **25**, 183–192. doi: 10.1109/TASSP.1977.1162929
- 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. Available at: http://www.leeds.ac.uk/linguistics/WPL/WP2002/Watt_Fab.pdf.