JASA ARTICLE

Normalization of nonlinearly time-dynamic vowels

Cesko C. Voeten,[a] Wilbert Heeringa, and Hans Van de Velde[b]
Fryske Akademy, Doelestraat 8, Leeuwarden, 8911 DX, The Netherlands

ABSTRACT:
This study compares 16 vowel-normalization methods for purposes of sociophonetic research. Most of the previous work in this domain has focused on the performance of normalization methods on steady-state vowels. By contrast, this study explicitly considers dynamic formant trajectories, using generalized additive models to model these nonlinearly. Normalization methods were compared using a hand-corrected dataset from the Flemish-Dutch Teacher Corpus, which contains 160 speakers from 8 geographical regions, who spoke regionally accented versions of Netherlandic/Flemish Standard Dutch. Normalization performance was assessed by comparing the methods' abilities to remove anatomical variation, retain vowel distinctions, and explain variation in the normalized F0–F3. In addition, it was established whether normalization *competes* with by-speaker random effects or *supplements* it, by comparing how much between-speaker variance remained to be apportioned to random effects after normalization. The results partly reproduce the good performance of Lobanov, Gerstman, and Nearey 1 found earlier and generally favor log-mean and centroid methods. However, newer methods achieve higher effect sizes (i.e., explain more variance) at only marginally worse performances. Random effects were found to be equally useful before and after normalization, showing that they complement it. The findings are interpreted in light of the way that the different methods handle formant dynamics.

https://doi.org/10.1121/10.0015025

(Received 29 April 2022; revised 12 October 2022; accepted 13 October 2022; published online 4 November 2022)

[Editor: Ewa Jacewicz] Pages: 2692–2710

I. INTRODUCTION

A moderate number of studies have compared different vowel-normalization methods for phonetic research (e.g., Adank *et al.*, 2004a; Fabricius *et al.*, 2009; Flynn, 2011; Flynn and Foulkes, 2011; Johnson, 2020; Morrison and Nearey, 2006; van der Harst, 2011). Contrasting with older studies (e.g., Disner, 1980; Hindle, 1978), it is since Adank *et al.* (2004a) that these comparisons have explicitly addressed the *socio*phonetic dimension of regional variation as one of the goals of normalization. However, with the notable exception of van der Harst (2011), prior studies comparing normalization methods have included only *static* information, i.e., measuring all of the vowels at a single time point, usually the midpoint. In recent times, concerns have arisen that comparisons of normalization methods based only on static information do not adequately represent vowels that are time-dynamic (Clopper *et al.*, 2005; Hillenbrand *et al.*, 2001). Because phonetic microvariation in formant trajectories is a major component of sociolinguistic variation (Fox and Jacewicz, 2009; Jacewicz and Fox, 2013; Van de Velde, 1996; van der Harst, 2011; van Hout *et al.*, 1999), and this information is used by human listeners (Jacewicz *et al.*, 2003; Milroy and Gordon, 2003; Peeters, 1991; Strange *et al.*, 1983; Voeten, 2021a,b), neglecting this source of variation might affect the conclusions of these prior studies that are comparing normalization methods. Thus, the present paper aims to update our knowledge of normalization techniques by explicitly taking account of *time-dynamic* information. In addition, this paper investigates the extent to which speaker normalization is still a methodological necessity given that nowadays the same task might be employed by by-speaker random effects in the context of a mixed-effects model or generalized additive mixed model. We investigate these issues using 16 currently used normalization methods, which are implemented in the *R* package *Visible Vowels* (`visvow`; VV; Heeringa and Van de Velde, 2018[1]).

A comparison of normalization methods taking explicit account of dynamic information in vowels was conducted by van der Harst (2011). He developed and compared three different approaches. The *target approach* represents phonological monophthongs by their midpoint and phonological diphthongs by the difference between two points measured at 25% and 75% of their production, respectively. This is one example of a "steady-state" method, which reduces trajectorial information to a single data point; we refer to the different measures in Fox and Jacewicz (2009) for alternative ways to compress trajectorial information into a unidimensional measure of vowel dynamicity. Contrasting with steady-state approaches, van der Harst (2011) also included two dynamic approaches with respect to time. The *time-points approach* takes samples at 25%, 37.5%, 50%, 67.5%,

[a] Electronic mail: cvoeten@fryske-akademy.nl
[b] Also at: Department of Language, Literature and Communication, Utrecht University, Trans 10, 3512 JK Utrecht, The Netherlands.

and 75% of their production (excluding the expected 12.5% and 87.5% as a result of coarticulation being too strong at those time points; see van der Harst, 2011) for monophthongs and diphthongs and fits a linear regression through these seven time points. The *regression approach* takes the same seven samples as the time-points approach but then fits a cubic polynomial regression to them. Using the same criteria as Adank *et al.* (2004a), van der Harst (2011) evaluated a set of normalization techniques for each of these three approaches. He found a clear advantage of the dynamic approaches over the target approach, especially with the time-points approach being superior in revealing regional variation. However, this advantage of time-dynamic approaches over the steady-state approach did not result in a different recommendation in the comparison of normalization techniques: just like in Adank *et al.* (2004a), Lobanov (1971) came out as the best-performing normalization method, in achieving the best balance between normalizing anatomical differences while retaining vowel distinctions and sociolinguistic differences.

Building on this prior work, this paper revisits the issue of vowel normalization for sociolinguistic research with a specific focus on formant dynamics. There have been two major developments in the past decade that warrant this revisit. First, new normalization methods have been developed, which the present paper includes in its investigation. Second, the past decades have seen significant methodological advances in phonetics, particularly with respect to the statistical analysis of time-dynamic data. One important innovation for our study is the rise of generalized additive models (GAMs; Wood, 2017) in phonetic research. These are regression-based models that allow the incorporation of nonlinear smoothing splines into the predictor structure, whose degree of smoothness (i.e., shrinkage from a wiggly curve to a straight line) is determined automatically. This means that GAMs model dynamic formant trajectories in a potentially nonlinear way and in doing so, find the optimal balance between overfitting and underfitting that is warranted by the data. This paints an important contrast with van der Harst's (2011) time-points approach and regression approach: the time-points approach is likely to have underfitted the data while the regression approach was shown by van der Harst (2011) to have produced an overfit (of the 84 coefficients produced by the regression approach, only 11 turned out to be needed in the linear-discriminant analyses by van der Harst, 2011).

To facilitate a comparison with GAMs, the regression approach and time-points approach can be conceptualized in terms of (unpenalized) basis functions. The time-points approach, which draws a single straight line through the time-point predictor, uses a single basis function, x; per van der Harst (2011), this underfits. The regression approach, which fits a cubic polynomial, uses three basis functions, x, x^2, x^3; per van der Harst (2011), this overfits. GAMs use basis functions that incorporate an automatically determined penalty term, which enables them to strike a balance between these under- and overfits. Beyond simple exponential-family models, GAMs also support multinomial logistic regression—a more modern alternative to linear-discriminant analysis[2] that can easily be fitted within the GAM framework—and multivariate analysis of variance, making them directly comparable to the linear models run by van der Harst (2011) and Adank *et al.* (2004a), with the added ability to nonlinearly model the vowels' dynamic formant trajectories in an optimal way.

A second methodological innovation and an important additional component to the impetus of this paper has been the ability to incorporate random effects into statistical models. While such mixed-effects models have been around since Henderson (1950), it is only in the past two decades that they have gained traction in the phonetic sciences (having entered the field via psycholinguistics, e.g., Baayen *et al.*, 2008). The ability to incorporate speakers as a random factor has important consequences for the field of normalization because this opens up an alternative means by which to capture anatomical differences. Thus, from a methodological point of view, fitting a full-random-effects model might render speaker normalization obsolete. Alternatively, however, the separate step of speaker normalization may make the random effects more identifiable by separating the anatomical variation from phonological and social variation and thereby allowing the latter sources to be estimated with more precision (for an example, see Voeten, 2021b). There is evidence from the domain of speech perception that human listeners take the latter "double" approach: Barreda (2020) presents results supporting a view in which the listener estimates a speaker-specific spectral scaling parameter ψ (as in the shared log-mean model by Nearey, 1978; see Sec. II B 14), which is shown to be strongly correlated with speaker anatomy, operationalized in Barreda (2020) through their physical height. Clearly, this is a type of normalization, and one that does not take into account sociolinguistic and idiosyncratic variation in the speaker's vowel productions; this kind of linguistic variation, Barreda (2020) argues, is estimated separately (but not independently) from ψ. In a follow-up paper, Barreda (2021) demonstrates that in perception, normalization methods that operate using such uniform scaling factors outperform those that incorporate additional processing. The latter includes models such as, for example, centroid normalization by Lobanov (1971); according to the perceptual results by Barreda (2021), such methods overnormalize, i.e., remove phonetic variation that is not strictly anatomical in nature. Thus, given this possible tension between methodological considerations (by-speaker random effects should subsume speaker normalization) and psycholinguistic considerations (normalization should be an independent step before estimating any subsequent speaker-specific variation), the present paper explicitly investigates the extent to which by-speaker random effects provide additional added value in addition to speaker normalization.

Our approach is described in Sec. II. We compare 16 normalization methods on 6 evaluation criteria. Four of our evaluation criteria are based on those by van der Harst (2011) and Adank *et al.* (2004a) with appropriate modifications to fit within the GAM framework. These criteria are the preservation of phonemic variation, the normalization of anatomical variation due to speaker sex, the preservation of regional variation, and the identification of the main sources of variation after normalization. In Adank *et al.* (2004a) and van der Harst (2011), the first three of these criteria were operationalized by performing linear-discriminant analyses of their four acoustic measures F0, F1, F2, and F3 onto each of these three categories and testing whether they were separated by the analysis above chance level. For the fourth criterion, both sources used multivariate analysis of variance to demonstrate the sources of variation (vowel, speaker sex, and speaker regional origin) containing speaker-discriminative information. As outlined in Sec. II C, we extend these approaches to GAMs and, in addition, we contribute two novel evaluation criteria. The first concerns the normalization of differences between individual speakers, of which we have 160 in our dataset. The second new criterion considers the explained variance in an analysis of known regional differences in the trajectory of the first formant. For this criterion, we run an analysis as a sociolinguist would do it, and consider the proportion of variance explained by an analysis that incorporates by-speaker random effects, relative to an analysis that does not include these.

It is prudent to reflect on the goals that these evaluation criteria aim to set. This is particularly relevant for our qualitative criteria, which are explained in Sect. II C 1. For these, each evaluation metric returns a single score, which is a percentage of tokens correctly recovered after normalization. In contrast to the quantitative criteria in Secs. II C 2 and II C 3, these need to be interpreted: is a high percentage of correct classification an indicator of a "good" normalization method or a "bad" normalization method? A particularly illustrative circumstance of this problem is the case of speaker sex. In our comparison, we follow Adank *et al.* (2004a), Broadbent *et al.* (1956), Ladefoged and Broadbent (1957), and van der Harst (2011) in considering information on speaker sex to be anatomical in nature, which, hence, should be subject to normalization. Thus, a normalization method is considered good if it normalizes away the information that keeps the two sexes in our dataset apart. However, van der Harst (2011, p. 101) rightly points out that "it might of course be the case that some of these differences are (partially) sociolinguistic," in other words, that they are indexically representative of the social construct of *gender* (on which there is ample sociolinguistic literature; for an overview, see e.g., Eckert and McConnell-Ginet, 2013). Similarly, while we again follow the aforementioned literature in considering a normalization method to be well-performing if it maintains phonemic distinctions between vowels, this does not match the full breadth of phonetic reality: casual speakers coarticulate and even outright merge vowel categories all of the time. Therefore, a normalization method that separates vowels well, particularly if it does so better than the unnormalized baseline, might have succeeded at drawing out the correct vowel categories of the data, but it might just as well have introduced an artefactual separation of vowels that are actually more variable in reality. In this hypothetical case, the data have been overnormalized, which has caused legitimate phonetic variation to be lost.

An additional criterion, which we added on top of those proposed by Adank *et al.* (2004a), presents a similar question. Each cell in the corpus design sampled five individuals from the same region, age, and sex. For the corpus's purposes of analyzing regional variation, the argument could be made that the variation within design cells (i.e., speakers matched on these sociolinguistic variables) should necessarily be smaller than the variation between them, such that speakers from the same cell should be rendered less distinct by the normalization process. Conversely, however, it has been shown that even speakers that are perfectly matched on sociolinguistic variables may differ from one another for reasons that are *not* purely anatomical (see Kendall and Fridland, 2012, for a particularly illustrative example of individual speakers from the same family). Thus, while it might be considered the main goal of speaker normalization to normalize differences between individual speakers from the same design cell, it is not realistic to require this expectation to be met perfectly—if it is, this might just as well be a sign of, again, overnormalization. Finally, an anonymous reviewer raises a similar concern about the separation of vowel categories: is it good to achieve a wide dispersion of the different vowels after normalization or is that an unrealistic representation of actual language use? We follow previous work in considering the separation (or rather, the non-merging) of vowel categories desirable. We believe that this choice is warranted by the nature of our data: recordings of individuals aiming to speak Standard Dutch on a word-list reading task. This is the most formal of speech styles and, hence, should achieve the highest degree of phoneme dispersion. However, again, we acknowledge that this consideration need not apply to other kinds of data for which a researcher might have different requirements.

In settling these issues, we believe that perceptual comparisons of normalization methods, like the one carried out by Barreda (2021), stand to add a lot of value: the ultimate benchmark for a good or bad normalization method is the language users themselves. The present paper, however, does not take a psycholinguistic stance but rather takes a quantitative-variationist stance. This means that our goal is not to identify the most *psychologically real* normalization method but rather to inventory the extent to which the different methods preserve and draw out variation that is meaningful for *sociophonetics*. Since it is well known that listeners cannot be probed on all aspects of linguistic variation (see, e.g., indicator variables; Labov, 1972), this paper makes the following choices. We consider a normalization method to perform well if it *maintains* differences between vowels and regions, *normalizes* differences between the two speaker sexes, and similarly *normalizes* individual speaker differences. However, our goals might be different from the reader's; we, therefore, warmly recommend readers to evaluate our results in light of their own research concerns, and then choose a normalization method that performs best on the criteria that they find important.

II. METHOD

A. The data

We used the dataset by van der Harst (2011), which is included with the tutorial to VV[3] (see the supplementary material[4]) and comprises the word-list component of the Dutch Teacher Corpus (Adank, 2003; van Hout *et al.*, 1999). The speakers in this dataset consist of 160 secondary-school teachers of the Dutch language, who each speak their respective variety of Standard Dutch (Netherlandic Standard Dutch or Flemish Standard Dutch). The speakers come from a total of eight major dialect regions in the Dutch language continuum: four in The Netherlands and four in Flanders. Each of these are subdivided into central (NL, Randstad; FL, Brabant), middle (NL, Gelderland; FL, East-Flanders), and two geographically peripheral regions (NL, Groningen and Netherlandic Limburg; FL, West-Flanders and Flemish Limburg). These regions' local interpretations of the supraregional standard are not homogeneous such that differences between these regions show up in objective phonetic measurements (van der Harst, 2011, Chap. 4; Voeten, 2021c) and subjective perception (Pinget *et al.*, 2014). We refer to van der Harst (2011) for more details about the data, including their recording and manual measurement.

The data were collected by van Hout *et al.* (1999) and measured by van der Harst (2011) in the following way. Each speaker produced two mono- or disyllabic isolated words containing one of the vowels /i,u,ɪ,ʏ,ɔ,eː,øː,oː,ɛi,œy,ɑu,ɛ,aː,ɑ/ and one monosyllabic word containing the vowel /y/. All of the vowels were followed by /s/ or /t/ and in the stressed position. F0–F3 of the vowels were extracted at 25%, 37.5%, 50%, 67.5%, and 75% production using Praat (Boersma and Weenink, 2007; Burg algorithm, time step 10 ms, F0 range 50–300 Hz for men and 100–500 Hz for women, cutoff point for formants 5000 Hz for men and 5500 Hz for women, window length 25 ms, pre-emphasis from 50 Hz) with the number of linear predictive coding (LPC) coefficients determined manually for every token (specified indirectly by setting the number of formants to half the number of LPC coefficients); we refer to van der Harst (2011) for further details. van der Harst (2011) manually checked the entire dataset for outliers or measurement errors and corrected these by hand. In our analyses of these data, we excluded time points one and seven based on findings by van der Harst (2011) that these are too strongly influenced by coarticulation. Thus, for each speaker and each vowel, we have five time points ranging from 25% production to 75% production. Figure 1 shows a vowel-space plot of all of the vowel types in the dataset sampled at the aforementioned five time points.

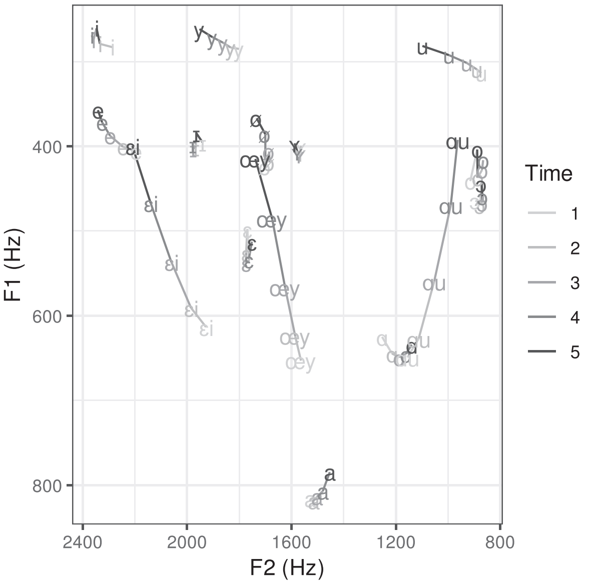

FIG. 1. A vowel-space plot of the raw data used in our comparisons of normalization methods, averaged over the 160 speakers per time point and vowel.

An important feature of the dataset is that there are no missing cells: all of the speakers produced all of the vowels. The dataset is not completely balanced because for the vowel /y/, we have only one word per speaker as opposed to two words per speaker for the other vowels. Some normalization methods are sensitive to imbalances in the data, whether these are due to missing cells or due to different numbers of observations per cell, as in our case of the vowel /y/. We refer to Barreda and Nearey (2018) for a comparison of some normalization methods' sensitivity to missing data as well as a proposed approach to remove this sensitivity. Since their proposal was originally applied to the method of Nearey 2, we discuss it in Sec. II B 14. VV also safeguards against imbalances in the design affecting the outcomes of normalization. VV does so by first computing the cell means and only using those to derive normalization factors. This prior averaging removes bias that may arise due to differences in sample size among the different speakers and vowels. This means that our present comparison of normalization methods is not affected by differences in sensitivity to imbalances between the normalization methods. Thus, our results are based only on the methods' performance on the task of normalization.

B. The normalization methods

We evaluate 16 speaker-normalization procedures. All of them are available in VV. In VV, the user can also choose from different scales: Hz, Bark, ERB (equivalent rectangular bandwidth), ln, mel, and ST (semitones).[5] Scales and speaker-normalization methods can be combined, except for six normalization methods in which a logarithmic transformation is included. The present paper focuses on normalization methods and draws its comparisons on the original scales for which each method was proposed. For a comparison of different scale transformations, we refer to Schützler (2015).

Most of the 16 normalization methods are described in at least 1 of the following publications: Adank (2003); Adank *et al.* (2004a); Barreda and Nearey (2018);

Esfandiaria and Alinezhadb (2014); Flynn (2011), and van der Harst (2011). Below we describe how we implemented them in the current study. Some normalization methods use descriptive statistics like the minimum, maximum, mean, or standard deviation. When F0 and formants are measured at multiple time points, the measurements at all of the time points are included when the descriptive statistics are calculated.

The 16 normalization methods can be classified into four types: formant-ratio normalization, range normalization, centroid normalization, and log-mean normalization. For an overview, see Table I.

1. *Peterson (1951)*

Peterson (1951) plotted F1/F3 ratios against F2/F3 ratios. For time point t, the measurements of the vowel productions are normalized as

$$F_{t1}^{\text{Peterson}} = \frac{F_{t1}}{F_{t3}}, \tag{1}$$

$$F_{t2}^{\text{Peterson}} = \frac{F_{t2}}{F_{t3}}, \tag{2}$$

Peterson (1951) calculated the ratios on the basis of mel-transformed measurements. In VV, the ratios can be calculated on the basis of any scale that is available in the application. The method by Peterson (1951) uses the mel scale, for which we use the formula by O'Shaughnessy (1987).

2. *Sussman (1986)*

Using the normalization method of Sussman (1986), each formant value is expressed relative to the geometric mean of F1, F2, and F3. For time point t and variable i, the measurements of the vowel productions are normalized as

$$F_{ti}^{\text{Sussman}} = \ln\left(\frac{F_{ti}}{\hat{F}_t}\right), \tag{3}$$

where $\hat{F}_t$ is the geometric mean, which is calculated from the values of the three formants of the vowel production that is being normalized, and ln is the natural logarithm. The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers $x_1, x_2, \ldots, x_n$, the geometric mean is

$$\sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}. \tag{4}$$

3. *Syrdal and Gopal (1986)*

The normalization procedure of Syrdal and Gopal (1986) is based on their observation that the distances F0 and F1 and between F2 and F3 are similar across speakers. In their model, F1 minus F0 corresponds to the height dimension and F3 minus F2 corresponds to the front-back dimension. They found that F1–F0 differences of high vowels and F3–F2 differences of front vowels are both less than 3 Bark. F1 and F2 frequencies of vowels are normalized at time point t by the following formulas:

$$F_{t1}^{S\&G} = F_{t1} - F_{t0}, \tag{5}$$

$$F_{t2}^{S\&G} = F_{t3} - F_{t2}. \tag{6}$$

Syrdal and Gopal (1986) applied the two formulas to frequencies that were scaled to Bark and, therefore, this

TABLE I. An overview of formant normalization methods. When no scale is given under "base scale," any scale is possible. A checkmark in the column "use descriptive statistics" indicates that the procedure uses descriptive statistics such as minimum, mean, etc.

	Applied to				Requires		Base	Use
	F0	F1	F2	F3	F0	F3	scale	descriptive statistics
Formant-ratio normalization								
Peterson (1951)	✓	✓	✓			✓		
Sussman (1986)		✓	✓	✓		✓	Hz	
Syrdal and Gopal (1986)		✓	✓		✓	✓		
Miller (1989)		✓	✓	✓	✓		Hz	✓
Thomas and Kendall (2007)	✓	✓	✓			✓		
Range normalization								
Gerstman (1968)	✓	✓	✓	✓				✓
Centroid normalization								
Lobanov (1971)	✓	✓	✓	✓				✓
Watt and Fabricius (2002)		✓	✓					✓
Fabricius *et al.* (2009)		✓	✓					✓
Bigham (2008)		✓	✓					✓
Heeringa and Van de Velde (2021) 1		✓	✓					✓
Heeringa and Van de Velde (2021) 2		✓	✓					✓
Log-mean normalization								
Nearey (1978) 1	✓	✓	✓	✓			Hz	✓
Nearey (1978) 2		✓	✓	✓		✓	Hz	✓
Labov (2006) 1		✓	✓				Hz	✓
Labov (2006) 2		✓	✓	✓		✓	Hz	✓

method is known as the "Bark-distance method." We use the formula by Traunmüller (1990) to compute the Bark transformation.

4. Miller (1989)

Using the normalization method of Miller (1989), formants are compared with their lower neighbours, i.e., F3 with F2 and F2 with F1. The first formant is normalized against a sensory reference (SR), which is calculated for each vowel type using geometric mean F0 (μ_{F0}), which is corrected for a constant c.

In our implementation, the geometric mean, μ_{F0}, is calculated as follows. First, the F0 values are averaged per combination of speaker, vowel type, and time point, where "time point" is one of the time points that were chosen by the user to be considered when calculating descriptive statistics. Using the averages in the dataset obtained in this way, the geometric mean, μ_{F0}, is calculated per speaker and descriptive time point.

The constant, c, is the geometric mean of the F0 average of the male speakers and the F0 average of the female speakers. Miller (1989) suggested to use $c = 168$ Hz, a value he found on the basis of F0 measurements in the Peterson and Barney database (Peterson and Barney, 1952). The F0 average of the male speakers was 125 Hz, and the F0 average of the female speakers was 225 Hz. The geometric mean is $\sqrt{(125 \times 225)} = 168$ Hz. In VV, rather than recalculating the constant, c, from the input data, $c = 168$ Hz is used. Thus, the procedure can still be used when the input table contains measurements from only male speakers or only female speakers or when the number of speakers is small. Using this constant, SR is computed as

$$\mathrm{SR} = 168 \times \left(\frac{\mu_{\mathrm{F0}}}{168}\right)^{1/3}. \tag{7}$$

Now formant frequencies in Hz of all of the vowel productions are normalized for each time point, t, by the following formulas:

$$F_{t1}^{\mathrm{Miller}} = \ln\left(\frac{F_{t1}}{\mathrm{SR}}\right), \tag{8}$$

$$F_{t2}^{\mathrm{Miller}} = \ln\left(\frac{F_{t2}}{F_{t1}}\right), \tag{9}$$

$$F_{t3}^{\mathrm{Miller}} = \ln\left(\frac{F_{t3}}{F_{t2}}\right), \tag{10}$$

where ln is the natural logarithm.[6]

5. Thomas and Kendall (2007)

In the procedure by Thomas and Kendall (2007), F0, F1, and F2 are normalized by subtracting them from F3. Formant frequencies of vowel productions are normalized for a time point t as

$$F_{t1}^{T\&K} = F_{t3} - F_{t1}, \tag{11}$$

$$F_{t2}^{T\&K} = F_{t3} - F_{t2}. \tag{12}$$

6. Gerstman (1968)

Gerstman (1968) normalizes the frequencies of a formant on the basis of the lowest and highest values found per speaker and across the selected descriptive time points. The frequencies are scaled so that they range from 0 to 999.

As a first step, for each of the variables F0, F1, F2, and F3, the values are averaged per combination of speaker, vowel type, and descriptive time point, which is a time point that was chosen by the user to be considered when calculating descriptive statistics. This prevents vowel types that occur frequently from being weighted more heavily than those that occur less frequently.

Next, per speaker, the formant values are averaged across the descriptive time points, and the minima and maxima of the averaged F1, F2, and F3 values are found. Then, for the vowel productions of speaker k, time point t, and variable i, we calculate

$$F_{kti}^{\mathrm{Gerstman}} = 999 \times \frac{F_{kti} - F_{ki}^{\min}}{F_{ki}^{\max} - F_{ki}^{\min}}. \tag{13}$$

7. Lobanov (1971)

A normalization procedure that expresses values relative to the hypothetical center of a speaker's vowel space is the one developed by Lobanov (1971). Using this method, a speaker's mean formant frequency is subtracted from a specific formant value and then divided by the standard deviation for that formant. In the normalized F1–F2 plot, Lobanov's centroid lies at (0,0).

First, for each of the variables F1, F2, and F3, the mean and standard deviation are calculated per speaker and per the descriptive time points that were chosen by the user to be considered when calculating descriptive statistics. We have to ensure that vowel types are weighted equally rather than by their token frequencies in the data. This is solved by averaging the F1, F2, and F3 measurements per combination of speaker, vowel type, and descriptive time point. Then, per speaker, the mean and standard deviation of those averaged values are calculated.

Using the mean and standard deviation, the formant frequencies of all of the vowel productions of speaker k, time point t, and variable i, are normalized as follows:

$$F_{kti}^{\mathrm{Lobanov}} = \frac{F_{kti} - \mu_{ki}}{\sigma_{ki}}. \tag{14}$$

8. Watt and Fabricius (2002)

The procedure of Watt and Fabricius (2002) expresses frequency values relative to a constructed centroid that is based on points that trace the edges of a speaker's vowel space. After normalization, the centroid lies at (1,1) in the F1–F2 plot.

As a first step, for each of the variables F1 and F2, the values are averaged per combination of speaker, vowel type, and descriptive time point, which is a time point that was chosen by the user to be considered when calculating descriptive statistics. This prevents more frequently occurring vowels from being weighted more strongly than those that occur less frequently.

Next, per speaker, the formant values are averaged across the descriptive time points. Thus, we get a dataset that contains average F1 and F2 frequencies for each vowel type per speaker. Using this dataset, for each speaker, we find the corners of the vowel space, which we call [i], [a], and [u′].[7] The coordinates of [i] are the minimum F1 and maximum F2. The coordinates of [a] are the maximum F1 and the F2 of the vowel type that has the maximum F1. The minimum F1 is also assigned to the F1 *and* the F2 of [u′]. Now, the coordinate of formant, i, of the centroid for speaker, k, is calculated as

$$S_{ki} = \frac{F_{ki[\text{i}]} + F_{ki[\text{a}]} + F_{ki[\text{u}']}}{3}. \tag{15}$$

Formant values of the vowel productions of speaker k, formant i, and time point t are normalized as

$$F_{kti}^{W\&F} = \frac{F_{kti}}{S_{ki}}. \tag{16}$$

9. Fabricius et al. (2009)

A weakness of the normalization method of Watt and Fabricius (2002), which was noticed by Thomas and Kendall (2007), is that the F2 of [a] might differ considerably from the ideal F2 midpoint of the vowel space—which they propose to obtain by averaging the F2 value of [i] and the F2 value of [u′]—and, thus, distort the lower part of the vowel space. Therefore, Fabricius *et al.* (2009) proposed an alternative with a modified formula for calculation of the coordinates of the centroid such that

$$S_{ki} = \begin{cases} \dfrac{F_{ki[\text{i}]} + F_{ki[\text{a}]} + F_{ki[\text{u}']}}{3}, & i = 1, \\ \dfrac{F_{ki[\text{i}]} + F_{ki[\text{u}']}}{2}, & i = 2. \end{cases} \tag{17}$$

10. Bigham (2008)

When using the procedures of Watt and Fabricius (2002) and Fabricius *et al.* (2009), it is assumed that the vowel space has the shape of a triangle. The normalization method of Bigham (2008) is another derivation of the procedure of Watt and Fabricius (2002), but its centroid is obtained on the basis of the corners of a quadrilateral. As corners, Bigham chose the American English vowels [ɪ], [u], [æ], and the average of [ɑ] and [ɔ] with tokens taken from word-list items of the form /hVd/.

We implemented a modified version of the method of Bigham (2008) that was proposed by Flynn (2011) (see also Flynn and Foulkes, 2011). When using this approach, the centroid is obtained on the basis of the corners of the vowel space, which are called [i′], [a′], [o′], and [u′]. The coordinates of [i′] are minimum F1 and maximum F2. Minimum F1 is also assigned to [u′]. Minimum F2 is assigned to [u′] and [o′]. Maximum F1 is assigned to [o′] and [a′].

In Bigham (2008), the F2 of [a′] was set equal to the F2 of the vowel [æ]. The implementation in VV first tries to find [æ] in the dataset. If the vowel is not found, the procedure searches for [æˑ]; if that vowel is not found, the procedure searches for [æː], then for [a], then for [aˑ], then for [aː], then for [ɛ], then for [ɛˑ], and then for [ɛː]. For the data in our comparison of normalization methods, this will cause the vowel [aː] to be used to compute [a′].

Now, the coordinate of formant, i, of the centroid for speaker, k, is calculated as

$$S_{ki} = \frac{F_{ki[\text{i}']} + F_{ki[\text{a}']} + F_{ki[\text{o}']} + F_{ki[\text{u}']}}{4}. \tag{18}$$

11. Heeringa and Van de Velde (2021) 1

The last three methods that we discussed all assume that vowel spaces have a specific shape. Such assumptions have been criticized in the literature as not adequately representing the full space that is available to speakers (Fox and Jacewicz, 2008, 2017; Jacewicz *et al.*, 2007). In response to this, Heeringa and Van de Velde (2021) developed a normalization method that does not assume a particular shape. It calculates a speaker's centroid on the basis of all of the points that constitute the convex hull that encloses the speaker's vowel space.

Just as for the procedures of Watt and Fabricius (2002), Fabricius *et al.* (2009), and Bigham (2008), per speaker, the average F1 and F2 is calculated for each combination of vowel type and descriptive time point such that each vowel is equally weighed in the normalization procedure.

Next, per speaker, the averaged formant values are averaged across the descriptive time points. Thus, per speaker, averaged F1 and F2 frequencies are obtained for all of the vowel types. Then, for each speaker, the vowels that constitute the convex hull are obtained on the basis of the speaker's frequencies of the vowel types. To this end, the *R* function `chull` from the `grDevices` package is used. This function uses an algorithm that is given by Eddy (1977). On the basis of the F1,F2 coordinates of the vowels that constitute the convex hull, the coordinates of a speaker's centroid are found with the R function `poly_center` of the `pracma` package (Borchers, 2021). This function calculates the centroid as the center (of mass) of the convex hull.

If S_{ki} is the coordinate of formant i of the centroid of the vowel space of speaker k, then the vowel productions of speaker k, formant i, and time point t are normalized as follows:

$$F_{kti}^{\text{convex hull}} = \frac{F_{kti}}{S_{ki}}. \tag{19}$$

12. Heeringa and Van de Velde (2021) 2

Using this method, a speaker's centroid is calculated in the same way as done in the previous method. However, it differs from that method because it integrates the centroid values in Lobanov's *z*-score formula.

Characteristic for Lobanov's normalization is that it does not only center the vowels around (F1 = 0, F2 = 0) but also scales them by dividing them by the standard deviation of the vowel formants, which makes the sizes of the vowel spaces of the speakers more comparable. However, a weakness of the method is that μ and σ depend on the distribution of the vowels in the vowel space. Vowel spaces with the same shapes but with different distributions within the vowel spaces will have different μ's and σ's. To solve this, Heeringa and Van de Velde (2021) made two changes to Lobanov's *z*-score formula. First, they replaced the μ of formant i and speaker k by the centroid coordinate S_{ki}. Second, the σ is calculated on the basis of only the formant values, i, of the vowels that constitute the convex hull. The result is a variant of Lobanov's method that does not depend on the distribution of the vowels within the vowel spaces of the speakers.

The vowels that constitute the convex hull may be irregularly distributed, i.e., the Euclidean distances of pairs of consecutive vowels may vary (strongly). To solve this, the number of points on the convex hull is interpolated up to 1000 points. Next, the points are classified in ten classes of equal width, both on the basis of F1 and F2. It appeared that by using ten classes, there is an equilibrium between the even distribution of the points on the convex hull that provides sufficient detail. The ten F1 classes do not exactly correspond with the F2 classes as F1 and F2 differences of the pairs of two successive points do not correlate exactly. Therefore, points may have the same F1 class and different F2 classes or the other way around. For each F1 class–F2 class combination, points that have an F1 within the F1 class and an F2 within the F2 class are averaged. Then, the number of points becomes equal to the number of F1-class–F2-class combinations. σ is calculated as the standard deviation of the formant values, i, of these points.

13. Nearey (1978) 1

The normalization methods of Nearey (1978) transform Hz measurements to logarithms and, subsequently, subtract a reference value from the log-transformed frequencies. The reference value is a log-mean. In the version explained in this section, the log-mean is calculated for each variable individually. Therefore, van der Harst (2011) refers to this procedure as *Nearey's individual log-mean model*.

Nearey (1978) originally proposed his methods to normalize *formant* values, specifically F1 and F2. In the later NORM software suite (Thomas and Kendall, 2007), versions of the methods were added that also normalize F3. In addition to these three formants, the formulas used by Adank *et al.* (2004a) and van der Harst (2011) also normalize F0; the same is possible in VV.

In our implementation, we first calculate the natural logarithms of the acoustic variables F0–F3, which should be given in Hz. Next, we calculate means, μ, for each combination of speaker, vowel type, descriptive time point, and variable, thus preventing vowel types that occur more frequently in the data from being weighted more strongly than vowel types that occur less frequently. Using these means, for each speaker, k, and variable, i, we calculate the average frequency, $\mu_{ki}^{\ln}$. Then, for each speaker k, each time point t, and each variable I, we calculate

$$F_{kti}^{\text{Nearey 1}} = F_{kti}^{\ln} - \mu_{ki}^{\ln}. \tag{20}$$

14. Nearey (1978) 2 / Barreda and Nearey (2018)

The method presented in this section is similar to the method explained in Sec. II B 13 except that the reference value is calculated by taking a speaker's mean of the log-means of the variables F1, F2, and F3. Because the same reference value is used for normalizing F1, F2, and F3 frequencies, van der Harst (2011) refers to this method as *Nearey's shared log-mean model*.

The formula used by Adank *et al.* (2004a) and van der Harst (2011) also included F0 in the speaker-specific reference value. However, it was pointed out by Barreda (2021) that the inclusion of F0 is detrimental to the performance of the Nearey 2 method. For this reason, in our comparison of normalization methods, we do not include F0 in the reference value and, therefore, also do not attempt to compute a normalization of F0. Thus, our version of the method only works with F1, F2, and F3.

In our implementation, we first compute the aforementioned speaker-specific reference value. This scaling factor is called ψ_k:

$$\psi_k = \frac{\mu_{k1}^{\ln} + \mu_{k2}^{\ln} + \mu_{k3}^{\ln}}{3}. \tag{21}$$

Normalized log frequencies of the F1–F3 of speaker k, time point t, and variable i are then calculated by subtracting the per-speaker ψ_k values,

$$F_{kti}^{\text{Nearey 2}} = F_{kti}^{\ln} - \psi_k. \tag{22}$$

It was noted by Barreda and Nearey (2018) that ψ_k will be biased if a dataset does not contain the same number of productions for all of the vowels for all of the speakers. They propose a modification (which can also be generalized to certain other methods that use speaker-specific scaling factors, e.g., Lobanov, 1971) that avoids this bias by replacing the direct averages computed by Nearey 2 with the corresponding marginal effects of a linear-regression model. Given each speaker k, vowel type v, time point t, and formant i, they fit

$$F_{kvti}^{\ln} = \beta_k X_k + \beta_{vti} X_{vti} + \epsilon, \tag{23}$$

where X_k is an indicator matrix coding for the individual speakers, and X_{vti} is a sum-coded matrix indicating the combinations of vowels, time points, and formants in the data. Because the model does not include an intercept term, the speaker-specific scaling factors, S_k, are given directly by the estimates for β_k, such that $S_k = \hat{\beta}_k$. These are then used as robust replacements for ψ_k in the Nearey 2 method:

$$F_{kti}^{\text{Barreda \& Nearey}} = F_{kti}^{\ln} - S_k. \tag{24}$$

The benefit of this approach lies in the sum-coding scheme used in the factor X_{vti}. While the estimates for the β_{vti} coefficients will still be influenced by the number of tokens in the data, the β_k coefficients will *not* be, because the sum-coding scheme is an orthogonal coding scheme. Thus, any imbalance in the number of productions per speaker, vowel, time point, and formant will *not* carry through into the speaker-specific S_k values.

As mentioned in Sec. II A, VV derives normalization factors like ψ_k after first averaging the data over the different productions per speaker, vowel type, time point, and formant. This is mathematically equivalent to the linear-regression approach by Barreda and Nearey (2018). As such, for purposes of the present comparison of normalization methods, the method of Barreda and Nearey (2018) is equivalent to the method of Nearey 2. In our results, we, therefore, refer to both methods as "Nearey 2/Barreda and Nearey."

15. Labov's ANAE method 1

The normalization procedure of Labov *et al.* (2006) was designed for the *Atlas of North American English* (ANAE).

First, the natural logarithms of the formant values (F1, F2) given in Hz are calculated. Using these values, the grand mean, G, is calculated, which is the geometric mean of the values of F1 and F2 of all of the speakers.

To avoid any bias toward vowel types that are more frequently found in the data, we generate a table that contains the geometric mean of the productions for each combination of speaker, vowel type, descriptive time point, and formant (F1, F2).

Given n_k speakers, n_v vowel types, n_t time points, and $n_i(=2)$ formants, we calculate G as the geometric mean of the $n_k \times n_v \times n_t \times n_i$ geometric means.

In addition, the mean, S_k, per speaker is calculated as the geometric mean of the $n_v \times n_t \times n_i$ geometric means of speaker k.

Subsequently, the anti-log (with base e, i.e., the exponent) of the difference between the two means ($G - S_k$) is calculated, which results in a speaker-specific scaling factor, d_k,

$$d_k = \exp(G - S_k). \tag{25}$$

Next, the formant values of the vowel productions of speaker k, time point t, and formant i are multiplied by this scaling factor such that

$$F_{kti}^{\text{Labov}} = d_k \times F_{kti}. \tag{26}$$

Labov *et al.* (2006) pointed out that G, using F1 and F2, stabilizes when the number of speakers exceeds 345. Thomas and Kendall (2007) mention that this may indicate "that this method (and perhaps speaker-extrinsic methods in general) are best only when a study has an exceptionally high subject count."

16. Labov's ANAE method 2

Labov's ANAE method 2 uses the same procedure as used in Labov's ANAE method 1 except that F3 measurements are also included when G and S_k are calculated. Therefore, when using this method in VV, F3 formant frequencies are normalized as well.

C. Evaluating normalization performance

We evaluated normalization performance using criteria inspired by those of Adank *et al.* (2004a) but modified to be able to incorporate nonlinear trajectories. For example, the linear-discriminant analyses by Adank *et al.* (2004a) cannot naturally include trajectorial information, whereas this *is* possible when using multinomial logistic GAMs instead.

For ease of exposition, we divide our evaluation criteria into three components:

(1) Classification, based on the similar evaluation criteria used by Adank *et al.* (2004a). Here, we established the extent to which the different normalization methods were able to distinguish sociolinguistically relevant categories, such as maintaining distinctions between vowel phonemes;
(2) sources of variation, based on the eponymous evaluation criterion by Adank *et al.* (2004a). Here, we established the amount of sociolinguistic information that the different normalization methods were able to capture; and
(3) explained variance. This is a novel criterion, where we established how well a sociolinguistic analysis would be able to explain sociolinguistically relevant variance in the data. We use this to investigate, based on model comparisons, the interplay between speaker normalization and the presence of by-speaker random effects in a statistical analysis.

1. Classification

Our classification component builds on the approach by Adank *et al.* (2004a). The following questions are addressed here:

(1) How well do the different techniques maintain distinctions between different vowels?
(2a) How well do the different techniques remove speaker-sex-related information in the F0?
(2b) How well do the different techniques remove speaker-sex-related information in the formants?
(3) How well do the different techniques maintain regional differences?

These questions coincide with the three sources of information mentioned by Broadbent *et al.* (1956) and Ladefoged and Broadbent (1957).

We also added a fourth question, which considers the extent to which *individual speakers* are indistinguishable from one another after normalization. As discussed in the Introduction, a complete neutralization of interspeaker differences is almost certainly a case of overnormalization. However, for an investigation of regional variation, it is still useful to know the extent to which differences between speakers from the same region–age–sex combination indeed do and do not remain in the data after normalization; by definition, such information reflects individual differences that cannot be regional in nature and, hence, might be undesirable to researchers specifically interested in regional differences. This is formalized by question (4):

(4) How well do the different techniques remove information that can separate individual speakers in the same design cell?

We used multinomial logistic GAMs to answer these questions by taking the category enquired about for each question (respectively, "vowel," "sex," "region," and "speaker") as a dependent variable and the normalized F0–F3 as predictors (or subsets thereof for normalization methods that do not retain F0 and/or F3). For region, one model was fitted for each vowel and the results from these 15 models were pooled (cf. Adank *et al.*, 2004a); for the other three dependent variables, a single model was run per normalization method. In all of the models, the F0 and formant values were assumed to interact with a nonlinear trajectory along the five time points. We modeled this through an approach called "varying-coefficient regression," whereby we modeled time using a thin-plate regression spline and then multiplied this spline by the relevant formant values. For example, the model for question (3) would predict sex out of three thin-plate regression splines along the predictor "time," each of which is multiplied by one of the three formants. The splines were afforded at most five basis functions such that, in theory, they were able to fit the five time points exactly, but the extent to which the data warranted the use of these degrees of freedom was determined automatically using the restricted maximum likelihood (REML)-based smoothness-selection methods in *R* package `mgcv` (Wood, 2017). Estimation was performed using the extended Fellner-Schall optimizer by Wood and Fasiolo (2017). After fitting each model and assessing them for successful convergence, we took the models' classification accuracies as our criteria of interest. These represent the *quality* of the normalization methods: for every token in the dataset, its corresponding vowel, speaker sex, region, and speaker individual are each either classified correctly or confused with something else, with nothing in between.

2. Sources of variation

We used multivariate normal GAMs to investigate what time-dynamic sources of variation (vowel, sex, region, and their interactions) could explain the joint variation in the normalized F0 and formants for each normalization method. We took F0–F3 as dependent variables, or the relevant subset of these for normalization methods that do not include F0 or F3. We set up sum contrasts for the predictors "vowel," "sex," and "region," and modeled the nonlinear formant dynamics in the data using a reference smooth (a thin-plate regression spline with, maximally, five basis functions) along time and difference smooths along time for each of the contrasts. For each normalization method, we fitted a GAM using function `gam` from *R* package `mgcv`, using REML as the smoothness-selection criterion and, again, fitting via the extended Fellner-Schall optimizer. All of the models converged successfully, with the single exception of the model for Sussman; for that model, we had to fix the smoothing parameters at effective infinity to get the model to converge.[8] This reduces the model complexity to a straight line through the five time points, which may, hence, underexplain the variance in those formant trajectories.

After fitting each model, we computed Wald *F*-values for each set of contrasts over all of the acoustic measures simultaneously; thus, we obtained *F*-values for the total nonlinear effects of vowel, sex, region, and each of their interactions. The Wald statistics were computed using the approach by Wood (2003). We then converted these *F*-values into partial-eta-squared values using the identity $\eta^2_{\text{partial}} = (F \times \text{edf})/(F \times \text{edf} + \text{rdf})$, where "edf" are the factor's estimated degrees of freedom and "rdf" are the corresponding residual degrees of freedom. Partial eta squared was also used by Adank *et al.* (2004a) here, and has the advantage that it is a scale-independent measure of effect size, meaning that its values can be compared across the different normalization methods. These values represent the *quantity* of the information that is reproduced by the normalization methods: the size of the effect corresponds to the amount of information present in the normalized data.

3. Random effects and explained variance

We investigated the extent to which the different normalization methods were able to capture regional variation between the eight regions. We focused on regional differences in the vowels' F1 trajectories, which are known to be subject to extensive regional variation. This concerns at least the tense mid vowels /eː,øː,oː/ and diphthongs /ɛi,œy,ɑu/ (Jacobi, 2009; Van de Velde, 1996; van der Harst, 2011; Voeten, 2021a), but Fig. 1 suggests that other vowels (e.g., /aː/) also vary in F1 across regions. We therefore included all 15 vowels in our comparison of the normalization methods, operationalizing the capture of regional variation by means of the R^2 statistic, which represents the total explained variance by a statistical model. Additionally, we investigated whether this explained variance improved significantly by the inclusion of a full by-speaker random-effects structure. A sociolinguistic analysis of this type would normally include such random effects, which serve to absorb any systematic inter-speaker differences, be they anatomical or sociolinguistic. If random effects serve this job well, then a separate speaker-normalization procedure might be unnecessary. In that case, we would expect a significant difference in the baseline model with versus without random effects but not much change in the normalized datasets. We tested this explicitly by comparing two models per normalization method: one with random effects (i.e., a generalized additive *mixed* model, or GAMM) and one without (i.e., a GAM).

The statistical analyses took the F1 as the dependent variable. As predictors, we included a thin-plate regression spline along time by the combination of region and vowel plus appropriate parametric terms. The splines were afforded, at most, five basis functions. For the models that included random effects, we also added a random intercept by speakers per vowel and a random smooth along time by speakers per vowel. The models were fitted using function `bam` from *R* package `mgcv` with discretization enabled to make it possible to fit even the full-random-effects models. All of the models converged successfully except for the model for Gerstman with random effects included. This model did converge, however, when we forced the random smooths by participants to be straight lines, similarly to the Sussman model in Sec. II C 2 (except only applied to the random smooths rather than to all of the terms in the model).[9]

For each model, we extracted its R^2 value. This represents the goodness of fit with values of zero indicating a total lack of fit and values of one indicating a perfect fit. These R^2 statistics are our metric of interest for comparing the different normalization methods. In addition, for each individual normalization method, we computed F-values for the difference in R^2 between the (reduced) model without random effects and the (full) model with random effects using the identity $F = \text{rdf}_{\text{full}} \times (R^2_{\text{full}} - R^2_{\text{reduced}})/(1 - R^2_{\text{full}})$, where "rdf" are the residual degrees of freedom.[10] The resulting statistic is F-distributed with $\text{edf}_{\text{full}} - \text{edf}_{\text{reduced}}$ numerator degrees of freedom and rdf_{full} denominator degrees of freedom, where "rdf" is the same as before and "edf" are the model degrees of freedom.

III. RESULTS

A. Classification

Figure 2 shows the extent to which the different normalization techniques were able to preserve the different vowel identities. Higher percentages indicate better retention of this important information. Figure 2 is sorted such that the methods that achieve better vowel separation are closer to the top. The best-performing method is Lobanov (77% correct), followed closely by Nearey 1 (75% correct). These methods plus those by Gerstman, Nearey 2 / Barreda and Nearey, Labov 2, and Heeringa and Van de Velde 2 separate the vowels better than the baseline of no normalization—the latter only achieves 65% correct. Most of the remaining methods score only marginally worse than this with the notable exception of Peterson (56% correct), Syrdal and Gopal (53% correct), and Sussman (51% correct). These latter three methods score far below the baseline and, hence, confuse noticably more vowels.

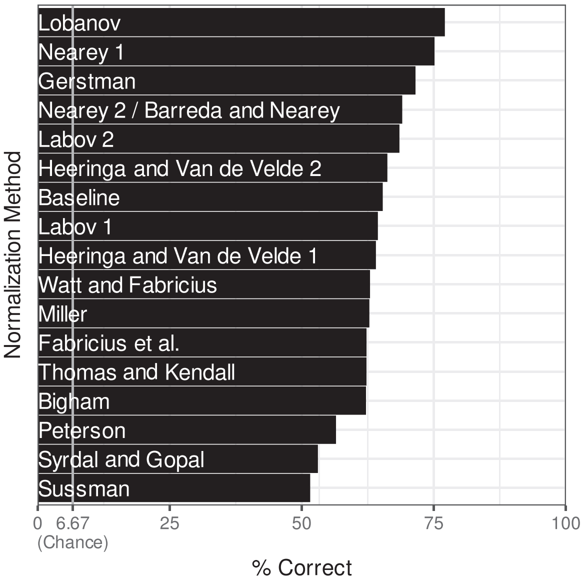

FIG. 2. Classification accuracies of recovering vowel identities from the different normalized datasets, with 100% correct indicating that all of the vowels are identified correctly and 0% indicating total confusion. Higher accuracies are taken as better performance. The chance level (indicated by the gray line) is at 11.1%.

Figure 3 shows the different normalization methods' performances at normalizing speaker sex. Figure 3 visualizes this by means of a stacked-bars plot that splits out the results by F0 information only, formant information only, and the combination of both sources of information (cf. Adank *et al.*, 2004a; van der Harst, 2011). Note that not all of the normalization methods include F0, and some also do

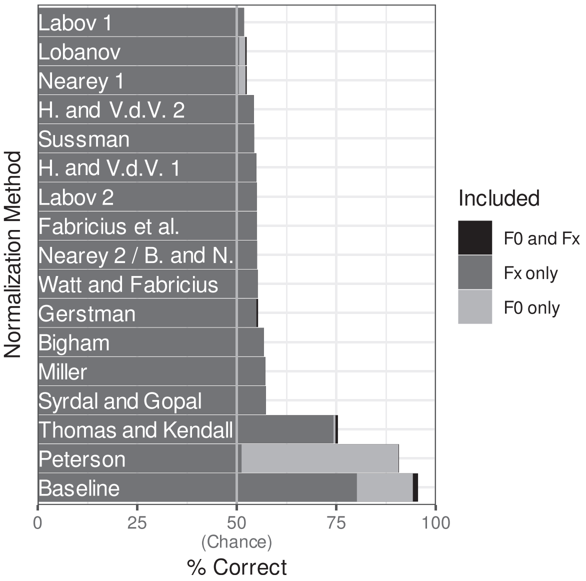

FIG. 3. Classification accuracies of recovering speaker sex from the different normalized datasets. Lower scores indicate better performance. The chance level (indicated by the gray line) is at 50%. "Fx" is an abbreviation for the three terms of F1, F2, and (if returned by the normalization method) F3.

not return F3 values. In both cases, the absent information is simply unused, such that some methods have no F0-only model and some of the "Fx" models include only F1 and F2, whereas others include F1, F2, and F3. Figure 3 is, again, ranked by classification accuracy with a lower percentage indicating more homogeneization and, thus, better normalization. A perfect normalization method would cause both sexes to be equally likely for all of the tokens; because the dataset is totally balanced, this would yield a classification accuracy of 50%. The best-performing methods are ranked closest to the top of Fig. 3. It turns out that the top three approaches—those being Labov 1 (no F0, 52% correct based on formants), Lobanov (52% F0 only, 50% formants only, 52% both), and Nearey 1 (52% F0 only, 50% formants only, 52% both)—are nearly indistinguishable from one another. The methods from Heeringa and Van de Velde 2 to Syrdal and Gopal all perform rather similarly to each other (55% correct, on average). The remaining methods—Thomas and Kendall, Peterson, and the baseline—all score substantially worse, ranging from 75% correct (Thomas and Kendall, with F0 and formant information combined) to as high as 95% correct (baseline, F0, and formants combined).

The ability of the different normalization techniques to preserve sociolinguistically meaningful regional differences is shown in Fig. 4. Following Adank *et al.* (2004a) and van der Harst (2011), the analyses were run separately for each vowel; the resulting classification accuracies are pooled across these different models. Again, the models are ranked by performance. We observe that Gerstman performs best, followed closely by Lobanov, the baseline, and Nearey 1—all of these score 27% correct after rounding. Then, the results naturally drop off until reaching Labov 1 at 22% correct. All of these results well exceed the chance level.

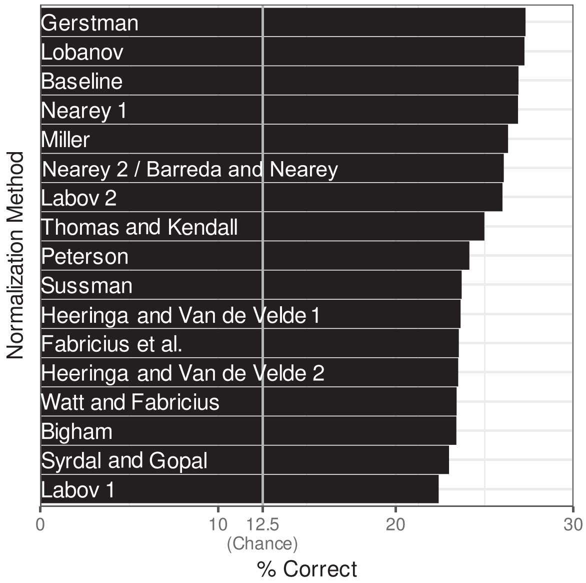

FIG. 4. Classification accuracies of recovering the speakers' regional origins from the different normalized datasets. Higher scores indicate better performance. The chance level is at 12.5%.

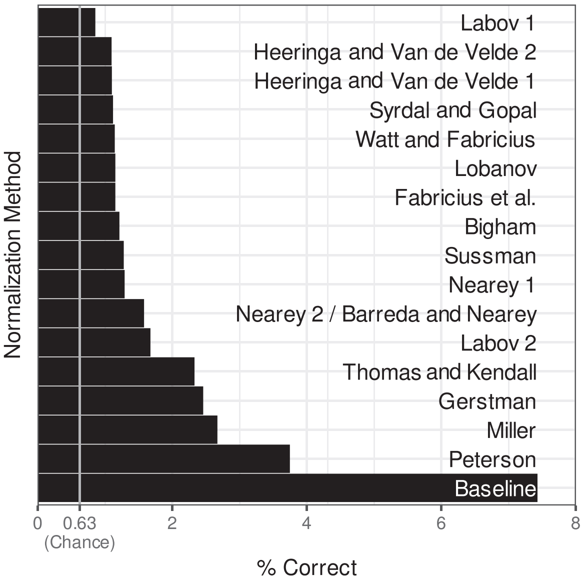

FIG. 5. Classification accuracies of recovering the individual speakers from the different normalized datasets. Lower scores indicate better performance. The chance level is at 0.63%.

Figure 5 shows the extent to which the different approaches were able to normalize differences between individual speakers. As discussed in the Introduction, this measure could be considered a double-edged sword: on the one hand, merging *anatomically* similar speakers could be considered the main goal of normalization, but on the other hand, no two speakers are exactly the same from a sociolinguistic point of view (cf. Kendall and Fridland, 2012). Thus, while the results in Fig. 5 provide more precise information than the coarse category of speaker sex from Fig. 3, the method that achieves the highest normalization here is not necessarily the *best* method *per se*. Nonetheless, Fig. 5 does rank the methods from most-normalizing to least-normalizing, leaving it up to the researcher to determine their own trade-off between overnormalization and undernormalization. As expected, we observe that in the baseline condition, i.e., in the absence of normalization, the highest number of speakers can be recovered from the data: 7.41%, well above chance. The method that normalizes these interspeaker differences the strongest is Labov 1, which allows our classification model to recover only 0.84% of speakers. Most of the other methods score only a bit above this, starting from Heeringa and Van de Velde 2 at 1.08% correct, but after reaching Labov 2 (1.66% correct), the remaining two methods recover well over 2.31% (Thomas and Kendall, 2007) of individual speakers.

B. Sources of variation

Figure 6 shows the partial-eta-squared values for the nonlinear sources of variation that we tried to account for in the data. Because partial eta squared is equivalent to a squared partial correlation coefficient, we can interpret the numbers in Fig. 6 as proportions of explained variance due to the individual factor in each facet. For the same reason,

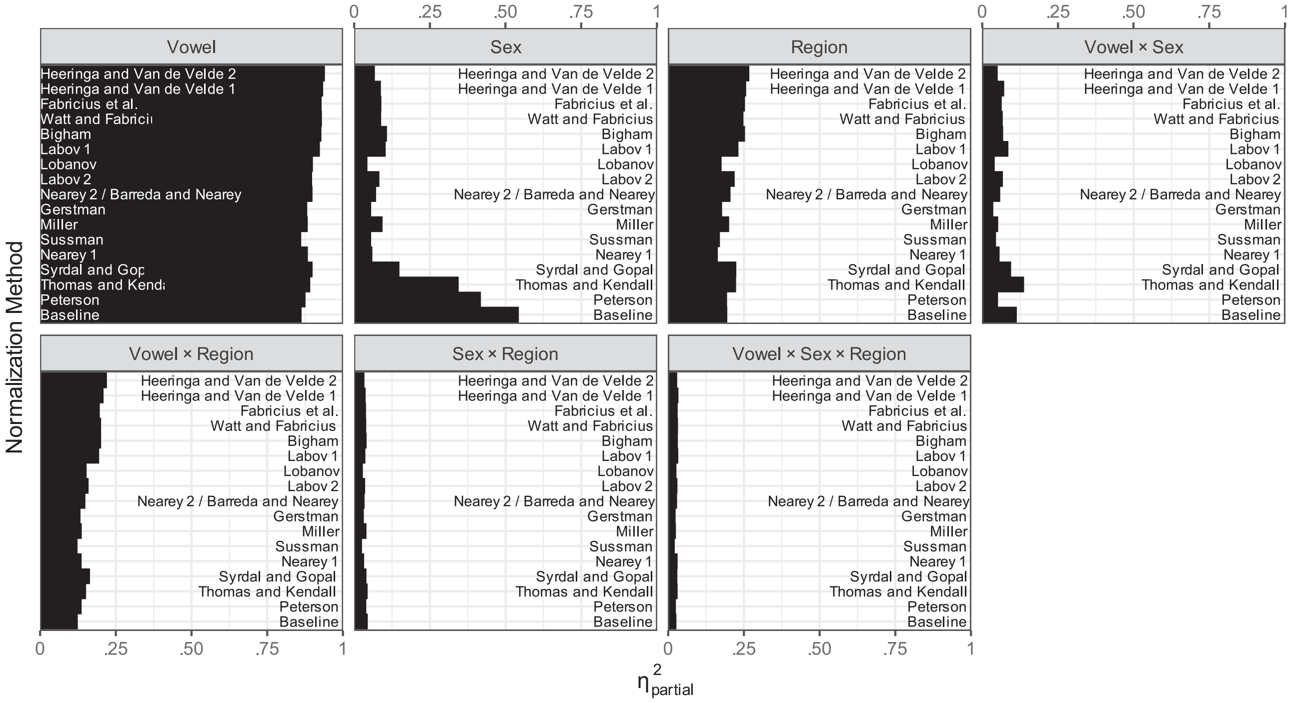

FIG. 6. Partial-eta-squared values for sources of variation.

the sum of a model's partial-eta-squared values may exceed one, particularly when the residual variance is small. Figure 6 is sorted such that larger effect sizes (i.e., more variation retained) are favored for the factors "vowel," "region," and their interaction, and lower effect sizes (i.e., more normalization) are favored for the facets involving sex or its interactions. The methods performing best according to these criteria over the seven facets are sorted closest to the top.

We observe that for all of the normalization methods, the largest source of variation is the factor "vowel" with the different vowels' formant trajectories, on average, explaining 90% of the partial variance associated with it. This means that the different vowels are well distinguishable from one another. For sex, most normalization methods have successfully normalized the variation: most normalization methods score well below the baseline's 54% explained partial variance. Thomas and Kendall and Peterson are noteworthy in this context, having noticeably undernormalized this variance, with an average η^2_{partial} of .38 between them. The two- and three-way interactions in which sex is involved all explain very little variance across all of the normalization methods. An anonymous reviewer asks whether the poor normalization of speaker sex by Thomas and Kendall and Peterson is different from the way these methods perform on single-point data—i.e., whether these two methods are particularly ill-adapted to trajectory data. We checked this by rerunning the same analysis with only the vowel midpoints and changing the thin-plate regression splines to ordinary factors. The results are nearly the same with, respectively, η^2_{partial} values of .34 and .42 for the midpoint-only models versus η^2_{partial} values of .36, and .44 for the time-dynamic models. Thus, this result is not specific to trajectory data.

On average, the factor region explains 21% partial variance. It is not necessarily informative to split this out across the different normalization methods; this is not only because the different methods seem relatively homogeneous on this factor but, moreover, because the main effect for region reflects across-the-board regional differences. However, the primary regional differentiator in this corpus is not the average F0–F3 measures altogether but rather their differences across the different vowels. The main effect is, therefore, a statistical nuisance term, and it is the interaction "vowel × region" that is of sociolinguistic interest. On that factor, the normalization method of Heeringa and Van de Velde 2 retains the largest amount of sociolinguistically relevant information as it is able to account for 22% of the partial variance through this interaction. The methods of Heeringa and Van de Velde 1, Fabricius *et al.*, Watt and Fabricius, Bigham, and Labov 1 follow quite closely, explaining between 21% and 19% of the partial variance. The other methods all ascribe less than 17% of the variance to the vowel × region interaction, which suggests that they have not been able to draw out all of the regional variation in vowel production to be found in the data.

C. Random effects and explained variance

Figure 7 presents the goodness-of-fit values for our sociolinguistic models of differences in F1 trajectories between vowels and regions. "−RE" indicates that the model did not include random effects, "+RE" indicates that the model did include random effects; Fig. 7 is sorted by the performance of the former model. When not considering random effects, we observe that all of the models perform very similarly, with perhaps Thomas and Kendall as an

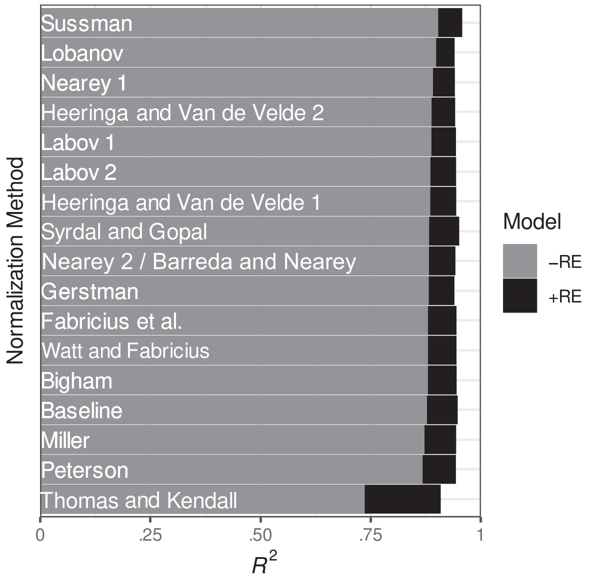

FIG. 7. R^2 values of the amount of regional variation in the different vowels' F1 trajectories that could be explained by the models. For all of the normalization methods as well as the baseline, the random-effects model represents a significant improvement over the fixed-effects-only model (all $p < 0.001$).

exception (the latter reaches an R^2 of .73, compared to the other models, which average out on $R^2 = .88$). When random effects are included, the methods other than Thomas and Kendall gain a boost of, on average, +.06 points on the R^2 metric or a 6% increase in total explained variance. For Thomas and Kendall, the gain from including random effects is noticably larger with a difference of +.17 points or 17% explained variance. This suggests that Thomas and Kendall's method, in principle, undernormalizes but the missed anatomical variation can be captured by including random effects, after which it performs on par with the other random-effects models.

It is important to observe that the baseline model, which had been fitted to the unnormalized data, does not accrue a noticeably larger gain between the −RE and +RE models than the normalized datasets do (the R^2 increase for the baseline model is +.07). This is important because it shows that the increase in R^2 for the random-effects models cannot be explained by anatomical factors alone: had that been the case, the unnormalized data should have seen a much larger R^2 increase after including random effects than the normalized datasets. We believe that an important factor in explaining this is the fact that our design was perfectly balanced (we have data from all of the participants on all of the items). In this case, a model with random effects will—by definition—give the same fit (i.e., coefficient estimates) as a model without random effects, just with a larger proportion of explained variance (since some variance will be reallocated from the error to the random effects). As such, the relatively systematic and modest contribution of random effects might be more visible in an unbalanced dataset, in which the grand mean and average of the per-speaker means no longer coincide. The result for purposes of sociophonetic research is that we observe a clear benefit to including random effects, even when data have been normalized. These results refute the hypothesis that normalization might be unnecessary as long as a proper random-effects structure is used; instead, normalization and random effects can work in tandem. In addition, there do not turn out to be substantial differences between the different normalization methods in this regard: all perform similarly, perhaps with the single exception of Thomas and Kendall, which without and with random effects scores a few percent points lower than the other methods. For this method, however, the advantage of including normalization and random effects also is the clearest: any putative undernormalization performed by this method has been correctly picked up by the random effects.

IV. DISCUSSION

There are two goals to speaker normalization: retaining (socio)linguistic information (here, vowel identity and regional origin) and normalizing anatomical differences (here, speaker sex and individual differences between speakers matched on sex, age, and region). On both of these goals, the selection of normalization methods by Adank *et al.* (2004a) showed the best performance for Lobanov, Nearey 1, and Gerstman. Our results, which include more methods, reproduce the findings by Adank *et al.* (2004a) for the first goal: Lobanov, Nearey 1, and Gerstman perform best. These are the top three methods for retaining distinctions between the vowels and in the top four (together with the baseline) for revealing regional differences. When it comes to normalizing anatomical differences, our results diverge from those by Adank *et al.* (2004a). Contrary to their findings of, again, Lobanov, Nearey 1, and Gerstman performing best, we observe an advantage for *log-mean* and *centroid* normalization methods. Speaker sex and individual differences between matched speakers were normalized best by Labov 1, a log-mean method. This is followed by Lobanov's centroid method, Nearey 1's log-mean method, and Heeringa and Van de Velde 2's centroid method; only then does Sussman's formant-ratio method appear. However, for the individual speaker, Labov 1 (a log-mean method) is followed by *all* of the centroid methods, starting from Heeringa and Van de Velde 2, after which the remaining three log-mean methods follow. We also observe that of the plots in Fig. 7, the five top-performing methods (i.e., those achieving high effect sizes on variables we want to maintain and low effect sizes on variables to be normalized) are all centroid methods as is the seventh, where methods six, eight, and nine are log-mean methods.

Why do log-mean and centroid methods perform so well? We believe that part of the reason is to be found in the way in which they respond to time-dynamic information. Figure 8 shows how four normalization methods, one of each type, use the first formant to distinguish between productions of /ɛi/ in two different regions: Netherlands-Randstad and Flemish Brabant. These two regions are the centers of, respectively, the Netherlandic and Flemish

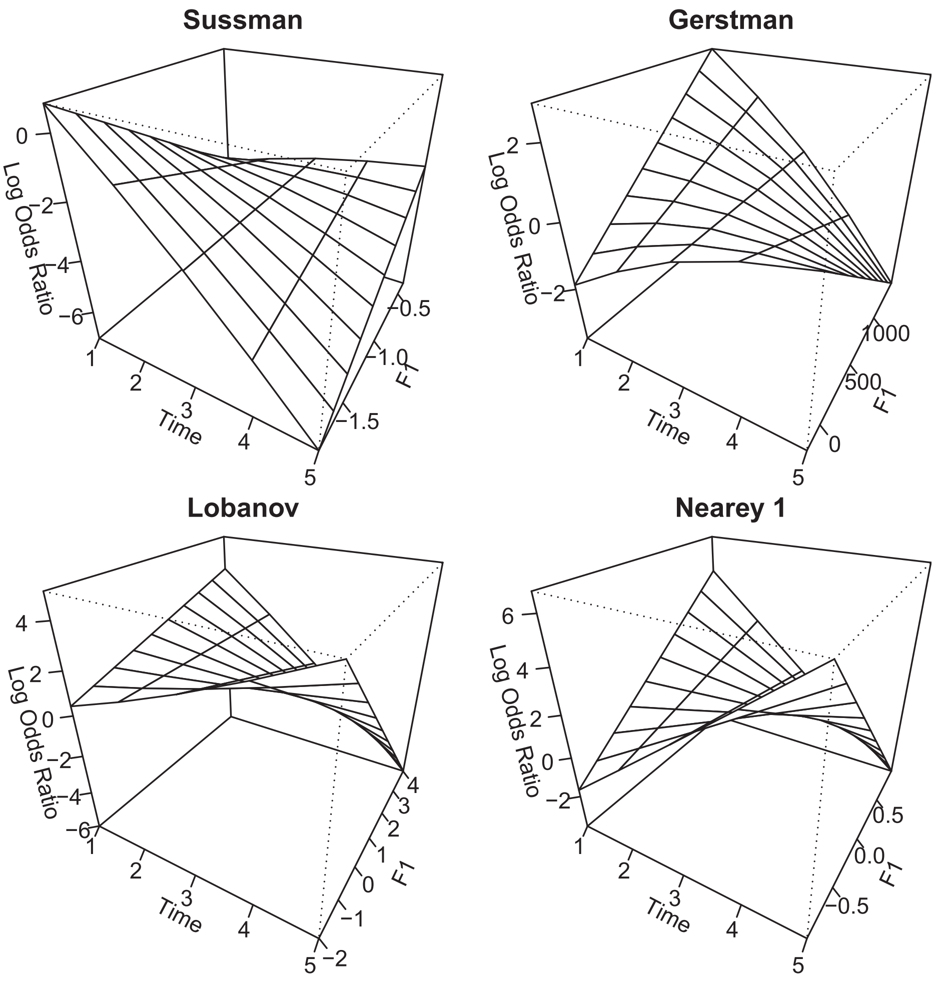

FIG. 8. The decision boundaries for four normalization methods in their regional classification of productions of the vowel /ɛi/. Higher log odds ratios denote higher probabilities of the region being classified as Netherlands-Randstad, whereas lower log odds ratios correspond to higher probabilities of a classification as Flemish Brabant.

varieties of Standard Dutch. We know from prior research (Adank *et al.*, 2004b; Adank *et al.*, 2007; Van de Velde, 1996; van der Harst, 2011; Voeten, 2021a) that there are two key differences in the productions of /ɛi/ between these varieties: The Netherlands has a lower nucleus and a higher target, both of which are reflected in the first formant. To visualize the use of this information in Fig. 8, we took the relevant classification models from Fig. 4 and generated predictions from them onto a grid of time point × F1 values (mapped to the *x* and *y* axes, respectively) while holding the other variables constant at their means. The *z* axis represents the fitted probability (on the logit scale) between the two regions of interest: positive values indicate a higher probability of the region being classified as Netherlands-Randstad while negative values represent Flemish Brabant. We observe that Lobanov and Nearey 1 accurately reproduce the expected regional differences. At early time points (i.e., in the nucleus), a low F1 is associated with Flemish Brabant while a high F1 is associated with Netherlands-Randstad. At late time points (i.e., in the target), a low F1 is associated with Netherlands-Randstad while a high F1 is associated with Flemish Brabant. When we turn to Gerstman,[11] however, we see that it has successfully captured the former pattern in the nuclei but fails to reproduce the latter observation on the targets. Sussman's method, in turn, not only fails to reproduce *both* observations, but the pattern that it *does* come up with is actually inverted, using the time-dynamic information in the signal in an incorrect way. We recommend that future work investigate the extent to which these observed correlations—log-mean and centroid methods making fuller use of time-dynamic information and achieving better performance—can be derived causally from the mathematical specifications of these normalization methods.

Thus, overall, in choosing a normalization method for sociophonetic research, we recommend giving strong consideration to log-mean methods (particularly Nearey 1) and centroid methods (particularly Lobanov). On the tasks of retaining vowel identities (Fig. 2) and normalizing anatomical factors (Figs. 3 and 5), these methods excel *qualitatively*: they produce relatively few vowel confusions and produce relatively many sex and speaker confusions (as would behoove a well-performing speaker-normalization technique). However, Fig. 6 showed that there are other options that retain higher *quantities* of sociolinguistically relevant information. We emphasize the difference between quality and quantity in this context: Figure 4 shows that the Gerstman-Lobanov-Nearey 1 triad is excellently able to find the qualitatively *right* information to separate regions based on these regions' vowel productions, such that a classifier like the one we used in Sec. III A is able to separate the regions on the basis of this information. However, when additional information is also present, as is common in analyses of sociolinguistic data, Fig. 6 shows that there are normalization methods other than those by Lobanov that obtain larger effect sizes, i.e., a higher signal-to-noise ratio and more statistical power, on the crucial interaction of vowel × region. If retaining the maximum amount of this information is of importance and the cost of slightly higher proportions of vowel confusion and speaker-sex retention is acceptable, we recommend the method of Heeringa and Van de Velde 2. This method obtained the highest partial-eta-squared value on the vowel × region interaction of all of the methods. This means that it has a high signal-to-noise ratio—at the cost of having only a slightly weaker signal than the other methods—translating into smaller residuals and, hence, higher power.

An anonymous reviewer asks to what extent the qualitatively different nature of F0 versus formants might play a role in the performance of the different normalization methods. From the perspective of speech production, F0 differs from formants in, among others, two fundamental ways. First, F0 is strongly influenced by sex and gender; formants are influenced by these as well (cf. sex-based formant shifts between male and female speakers, sex-based differences in vowel-space sizes, and gender-based differences in vowel dispersion) but not as dramatically and obviously. Second, F0 is used to communicate paralinguistic features (such as the speaker's emotional state) in addition to linguistic information (such as phrasal groupings or lexical tone) while formants do not inherently represent such information. For purposes of investigating the role of F0, the normalization methods could be divided into three types:

(1) Methods that ignore F0. For example, most of the centroid methods (see Table I) only involve F1 and F2. The methods that ignore F0 are Sussman, Watt and Fabricius, Fabricius *et al.*, Bigham, Heeringa and Van de Velde 1 and 2, Nearey 1, and Labov 1 and 2;
(2) methods that *return* F0 but do not *use* it (e.g., in calculating some kind of reference value). One such example is Lobanov: F0 is not intrinsically involved in the calculation of the means and standard deviations used as references except when the variable being normalized happens to be F0 itself. The full list of these methods is the baseline, Peterson, Thomas and Kendall, Gerstman, Lobanov, and Nearey 1; and
(3) methods that *use* F0 but do not *return* it. These are Syrdal and Gopal and Miller, which normalize F1 by subtracting F0 itself (Syrdal and Gopal) or the F0-based SR (Miller) from it but do not return F0 itself.

Type-1 and type-2 methods are useful control cases: Their juxtaposition makes it possible to ascertain the role of F0 in and of itself. Type-3 methods are particularly interesting for the present question because they introduce F0 in a Trojan-horse-like way: F0 influences the calculation of the formants, but the statistical models we use in our evaluations cannot control for this influence due to F0 itself being absent from the output of the normalization procedure.

We conducted an exploratory investigation of the role of F0 on the basis of the multivariate models from Sec. III B. Due to their having the acoustic measures as multiple dependent variables at once, these models make it possible to evaluate simultaneously the contribution of F0 *and* the way F0 influences the formants. We tested the role of the 3 different F0 types by means of linear regression on the 119 partial-eta-squared values from Sec. III B. We converted these into partial correlation coefficients by taking their square roots and then applied the z-transformation of Fisher (1921) to these to yield a Gaussian random variable, on which linear-regression analysis is valid. We used function `lme` from *R* package `nlme` to fit a linear model taking these transformed effect sizes as the dependent variable. We included fixed effects for the three F0 types ("F0 type"), the seven factors from Fig. 6 ("factor"), and their interaction (all coded using deviation coding). Also, we added a random intercept by normalization methods. The fixed-effects results showed significant main effects of "factor," which reflects that some effects (e.g., vowel) from Fig. 6 had larger effect sizes than others (e.g., sex) across the board. We did not find any significant main effect of F0 type. Most importantly, of the 12 possible interactions between F0 type and "factor," we found 4 that were significant. Type-1 methods achieved lower effect sizes than the other methods on sex [$\hat{\beta} = -2.21$, SE (standard error) $= 0.57$, $t_{84} = -3.90$, $p < 0.001$], which confirms that methods that ignore F0 retain less information about speaker sex (which the results from Fig. 3 strongly corroborate). Type-1 methods achieved higher effect sizes on vowel ($\hat{\beta} = 2.09$, SE $= 0.57$, $t_{84} = 3.69$, $p < 0.001$), which means that excluding F0 results in a slightly better separation of the different vowels. We surmise that this is because F0 is essentially noise when it comes to vowel identity, such that models that exclude it simply profit from a higher signal-to-noise ratio in identifying the individual vowels. Type-2 methods, in contrast, show a completely inverted pattern: these methods achieved significantly higher effect sizes on sex ($\hat{\beta} = 2.64$, SE $= 0.61$, $t_{84} = 4.35$, $p < 0.001$), and significantly lower ones on vowel ($\hat{\beta} = -1.51$, SE $= 0.61$, $t_{84} = -2.49$, $p = 0.01$). Type-3 methods, then, are in between the other two methods. On the basis of this brief exploratory analysis, we conclude that F0 type *does* make a difference: Methods that completely exclude F0 retain less information on speaker sex and also return better-dispersed vowels compared to methods that retain F0. In addition, if a method does make use of F0, these differences are more pronounced if the F0 information remains explicitly available in the analysis (i.e., type-2 methods) than if it is withheld from the model (i.e., type-3 methods).

An additional issue that was explicitly investigated in this paper has been the possible relationship between (knowledge-based) speaker normalization and (knowledge-poor) by-speaker random effects. The question was whether random effects would be useful either *instead of* or *supplementary to* speaker normalization. Our results from Sec. II C 3 clearly supported the latter view and discounted the former. There was an overall benefit to including random effects that was the same for the unnormalized baseline and the different normalized datasets. As such, normalization and random effects were found to offer independent contributions rather than one supplanting the other. Furthermore, in one specific case (namely, the normalization method of Thomas and Kendall), random effects were found to effectively compensate for undernormalization. The method of Thomas and Kendall retains a relatively high amount of anatomical information (Fig. 3), which translated into a lower amount of explained variance in our sociolinguistic analysis in Sec. III C when random effects were not included. However, after random effects were added, the method performed on par with the others; this shows that part of the additional variation accounted for was the same as that for the other methods, whereas another part was due to anatomical differences that the method had undernormalized. In other words, Thomas and Kendall retained some between-speaker anatomical variation, which was able to distinguish speaker sexes and be captured by random effects. This shows that researchers can afford themselves a bit of freedom in their choices of normalization methods: Should the optimal method for a given problem be one that does not normalize to the fullest extent, random effects will compensate for this.

The previous paragraph demonstrated one instance of a more general point: A researcher's choice of normalization method depends on their research question. If there is one of the factors we investigated in Sec. III A that the researcher strongly cares about in isolation, it is a valid strategy to simply pick the normalization method that scores best on that evaluation criterion. If the researcher is more concerned about a particular factor in the presence of (possibly many) others, Sec. III B may be of more help. We note, furthermore, that our normalization criteria need to be viewed in the context of all of the speakers speaking one standardized variety in which there are small but notable regional differences. In a different setting, different properties might be desirable. For example, Lobanov excellently homogenizes tokens of the same vowel category produced by different speakers (Fig. 2), but in cases of stronger variation, where there may also be differences in the centroid location and contour shape of the relevant vowel category, such homogeneization might not be desired (and even within this language situation, one could raise objections to this goal in light of e.g., Kendall and Fridland, 2012). In a similar vein, we have considered information on speaker sex to be anatomical in nature (following Adank *et al.*, 2004a and van der Harst, 2011). However, while *sex* is indeed often not of sociolinguistic interest, the closely related notion of *gender* often is; in such a context, a method that we consider good in Fig. 3 might actually be considered to be bad. We, thus, cannot offer a one-size-fits-all "optimal normalization method"; rather, our results should be taken to guide a careful consideration of the relative advantages and disadvantages of the different options.

V. CONCLUSION

This paper considered 16 normalization methods in their performance on a well-known and validated dataset of regional variation, explicitly taking account of nonlinear formant dynamics. Our investigation partially reproduced the standing recommendation to use Lobanov, Nearey 1, or Gerstman (Adank *et al.*, 2004a) based on their excellent classification performances. However, if seeking to draw out the maximum amount of sociolinguistic variation in the presence of other factors, our results show that it is worthwhile to consider possible alternatives (e.g., Heeringa and Van de Velde 2). We also investigated the contribution of random effects and their interaction with the choice of normalization method (including no normalization) and found that normalization and random effects operate independently and complement one another. While a definitive choice of optimal normalization method will, by necessity, depend on research-specific considerations, we believe that our results offer a viable starting point for other researchers in sociolinguistics for making these considerations.

ACKNOWLEDGMENTS

We gratefully thank Vincent van Heuven, Roeland van Hout, and Toni Rietveld for their valuable feedback on draft versions of this article. Additionally, we thank the editor and two anonymous reviewers for their constructive criticism. The research reported in this study is part of the research project "Frisian Laboratory Sociolinguistics," sponsored by the Provincial Government of Fryslân (Grant No. PF-2019/194740).

[1] See https://www.visiblevowels.org/ (Last viewed October 27, 2022).

[2] An advantage of multinomial logistic regression over linear-discriminant analysis is that it does not require the latter's assumption that the classification boundaries must be linear functions of the predictors, or quadratic functions in the case of quadratic discriminant analysis.

[3] See https://fryske-akademy.nl/fa-apps/tutorial/#dataset (Last viewed October 27, 2022).

[4] See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0015025 for a copy of our dataset, our full R code, and summaries of our "models."

[5] van der Harst (2011) rightly points out that Adank (2003) and Adank *et al.* (2004a) do not display the formulas for ERB and mel correctly. In VV, the correct formulas are used.

[6] Again van der Harst (2011) rightly writes that the formulas given by Adank (2003) and Adank *et al.* (2004a) are incorrect.

[7] The prime symbol following the [u] vowel should be interpreted in the mathematical sense of derivation, rather than as a stress diacritic. It indicates that the vowel [u′], having been constructed from two F1 values rather than from F1 and F2, only by analogy corresponds to a location in F1–F2 space. We refer to Watt and Fabricius (2002) for motivation.

[8] Allowing the smoothing parameters to vary freely resulted in the likelihood eventually becoming indefinite (i.e., diverging towards infinity), raising an error in the fitting process.

[9] As a result, the Gerstman model with random effects might have underfitted the random-effects structure in the data. However, as this model's results do not appear notably different from the others, we conclude that this possibility is not fatal to this model's inclusion in the comparisons.

[10] We also computed likelihood-ratio tests of the difference in REML between the two models. These yielded the same substantive outcomes. However, the F statistic is more appropriate, as it takes the uncertainty in the scale parameter into account, while the χ^2 statistic from the

likelihood-ratio test does not. (This is why $F_{\text{edf,rdf}} = X^2_{\text{edf}}$ in the large-sample limit, where the residual degrees of freedom become irrelevant, but not in finite samples.)

[11] The astute reader will note that the F1 range for Gerstman exceeds the scaled range of (0,999). This is because VV bases the extrema of the relevant scaling factors on the per-speaker, per-vowel averages of the data, rather than the raw extrema. As a consequence, it is possible for some of the raw data to have more extreme values than the computed scaling factors. As explained in Sec. 6, this was a design choice in VV that prevents vowel types that occur more frequently within the dataset from being assigned more weight.

Adank, P. (**2003**). "Vowel normalization: A perceptual-acoustic study of Dutch vowels," Ph.D. thesis, Katholieke Universiteit Nijmegen, Nijmegen.

Adank, P., Smits, R., and van Hout, R. (**2004a**). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**(5), 3099–3107.

Adank, P., van Hout, R., and Smits, R. (**2004b**). "An acoustic description of the vowels of Northern and Southern Standard Dutch," J. Acoust. Soc. Am. **116**(3), 1729–1738.

Adank, P., van Hout, R., and Van de Velde, H. (**2007**). "An acoustic description of the vowels of Northern and Southern Standard Dutch II: Regional varieties," J. Acoust. Soc. Am. **121**(2), 1130–1141.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**(4), 390–412.

Barreda, S. (**2020**). "Vowel normalization as perceptual constancy," Language **96**(2), 224–254.

Barreda, S. (**2021**). "Perceptual validation of vowel normalization methods for variationist research," Lang. Var. Change **33**(1), 27–53.

Barreda, S., and Nearey, T. M. (**2018**). "A regression approach to vowel normalization for missing and unbalanced data," J. Acoust. Soc. Am. **144**(1), 500–520.

Bigham, D. S. (**2008**). "Dialect contact and accommodation among emerging adults in a university setting," Ph.D. thesis, University of Texas at Austin, Austin, TX.

Boersma, P., and Weenink, D. (**2007**). "Praat: Doing phonetics by computer (version 4.5.11) [computer program]" available at http://www.praat.org/ (Last viewed January 29, 2007).

Borchers, H. W. (**2021**). "pracma: Practical Numerical Math Functions," *R* package version 2.3.3, available at https://CRAN.R-project.org/package=pracma (Last viewed October 27, 2022).

Broadbent, D. E., Ladefoged, P., and Lawrence, W. (**1956**). "Vowel sounds and perceptual constancy," Nature **178**(4537), 815–816.

Clopper, C. G., Pisoni, D. B., and De Jong, K. (**2005**). "Acoustic characteristics of the vowel systems of six regional varieties of American English," J. Acoust. Soc. Am. **118**(3), 1661–1676.

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

Eckert, P., and McConnell-Ginet, S. (**2013**). *Language and Gender* (Cambridge University Press, Cambridge, UK).

Eddy, W. F. (**1977**). "A new convex hull algorithm for planar sets," ACM Trans. Math. Softw. **3**(4), 398–403 and 411–412.

Esfandiaria, N., and Alinezhadb, B. (**2014**). "Evaluating normalization procedures on reducing the effect of gender in Persian vowel space," Int. J. Sci.: Basic Appl. Res. **13**(2), 303–316, avaiable at https://gssrr.org/index.php/JournalOfBasicAndApplied/article/view/2756.

Fabricius, A., Watt, D., and Johnson, D. E. (**2009**). "A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics," Lang. Var. Change **21**(3), 413–435.

Fisher, R. A. (**1921**). "On the 'probable error' of a coefficient of correlation deduced from a small sample," Metron **1**, 3–32.

Flynn, N. (**2011**). "Comparing vowel formant normalisation procedures," York Papers Linguist. **2**(11), 1–28.

Flynn, N., and Foulkes, P. (**2011**). "Comparing vowel formant normalization methods," in *Proceedings of the XVII ICPhS*, edited by W. S. Lee and E. Zee, City University of Hong Kong, pp. 683–686.

Fox, R. A., and Jacewicz, E. (**2008**). "Analysis of total vowel space areas in three regional dialects of American English," in *Proceedings of Acoustics 2008*, pp. 495–500.

Fox, R. A., and Jacewicz, E. (**2009**). "Cross-dialectal variation in formant dynamics of American English vowels," J. Acoust. Soc. Am. **126**(5), 2603–2618.

Fox, R. A., and Jacewicz, E. (**2017**). "Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English," J. Acoust. Soc. Am. **142**(1), 444–459.

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

Heeringa, W., and Van de Velde, H. (**2018**). "Visible Vowels: A tool for the visualization of vowel variation," in *Proceedings of the CLARIN Annual Conference 2018 (CLARIN ERIC)*, pp. 120–123, available at https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (Last viewed October 27, 2022).

Heeringa, W., and Van de Velde, H. (**2021**). "A new vowel normalization for sociophonetics," in *Proceedings of Interspeech 2021*, pp. 4024–4028.

Henderson, C. R. (**1950**). "Estimation of genetic parameters," Ann. Math. Stat. **21**, 309–310.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (**2001**). "Effects of consonant environment on vowel formant patterns," J. Acoust. Soc. Am. **109**(2), 748–763.

Hindle, D. (**1978**). *Approaches to Formant Normalization in the Study of Natural Speech* (Academic, Berlin, Heidelberg), pp. 161–171.

Jacewicz, E., and Fox, R. A. (**2013**). "Cross-dialectal differences in dynamic formant patterns in American English vowels," in *Vowel Inherent Spectral Change*, edited by G. Morrison and P. Assmann (Springer, Berlin, Heidelberg), pp. 177–198.

Jacewicz, E., Fox, R. A., and Salmons, J. (**2007**). "Vowel space areas across dialects and gender," in *Proceedings of XVI ICPhS*, pp. 1465–1468.

Jacewicz, E., Fujimura, O., and Fox, R. A. (**2003**). "Dynamics in diphthong perception," in *Proceedings of XV ICPhS*, pp. 993–996.

Jacobi, I. (**2009**). "On variation and change in diphthongs and long vowels of spoken Dutch," Ph.D. thesis, University of Amsterdam, Amsterdam, available at https://hdl.handle.net/11245/1.316951 (Last viewed October 27, 2022).

Johnson, K. (**2020**). "The Δ*F* method of vocal tract length normalization for vowels," Lab. Phon. **11**(1), 10.

Kendall, T., and Fridland, V. (**2012**). "Variation in perception and production of mid front vowels in the US Southern Vowel Shift," J. Phon. **40**(2), 289–306.

Labov, W. (**1972**). *Sociolinguistic Patterns* (University of Pennsylvania Press, Philadelphia, PA).

Labov, W., Ash, S., and Boberg, C. (**2006**). *The Atlas of North American English: Phonetics, Phonology and Sound Change* (Mouton de Gruyter, Berlin).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.

Lobanov, B. M. (**1971**). "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am.**49**(2), 606–608.

Miller, J. D. (**1989**). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**(5), 2114–2134.

Milroy, L., and Gordon, M. (**2003**). *Sociolinguistics: Methods and Interpretation* (Blackwell, Malden, MA).

Morrison, G. S., and Nearey, T. M. (**2006**). "A cross-language vowel normalisation procedure," Can. Acoust. **34**(3), 94–95, available at https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1838.

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

O'Shaughnessy, D. (**1987**). *Speech Communication: Human and Machine* (Addison-Wesley, Reading, MA).

Peeters, W. J. M. (**1991**). "Diphthong dynamics: A cross-linguistic perceptual analysis of temporal patterns in Dutch, English, and German," Ph.D. thesis, Rijksuniversiteit te Utrecht, Utrecht.

Peterson, G. E. (**1951**). "The phonetic value of vowels," Language **27**, 541–553.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184.

Pinget, A.-F., Rotteveel, M., and Van de Velde, H. (**2014**). "Standaardnederlands met een accent: Herkenning en evaluatie van regionaal gekleurd Standaardnederlands in Nederland" ("Standard Dutch with an accent: Recognition and evaluation of regionally colored Standard Dutch in The Netherlands"), Ned. Taalk. **19**(1), 3–45.

Schützler, O. (**2015**). "Transforming acoustic vowel data: A comparison of methods, using multi-dimensional scaling," in *Trends in Phonetics and*

Phonology: Studies from German-Speaking Europe, edited by V. Dellow, M.-J. Kolly, A. Leemann, and S. Schmid (Lang, Bern/New York), pp. 35–47.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). "Dynamic specification of coarticulated vowels," J. Acoust. Soc. Am. **74**(3), 695–705.

Sussman, H. M. (**1986**). "A neuronal model of vowel normalization and representation," Brain Lang. **28**(1), 12–23.

Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**(4), 1086–1100.

Thomas, E. R., and Kendall, T. (**2007**). "NORM: The vowel normalization and plotting suite," available at http://lingtools.uoregon.edu/norm/ (Last viewed 21 January 2018).

Traunmüller, H. (**1990**). "Analytical expressions for the tonotopic sensory scale," J. Acoust. Soc. Am. **88**(1), 97–100.

van der Harst, S. (**2011**). "The vowel space paradox: A sociophonetic study on Dutch," Ph.D. thesis, Radboud Universiteit Nijmegen, Nijmegen.

Van de Velde, H. (**1996**). "Variatie en verandering in het gesproken Standaard-Nederlands" ("Variation and change in spoken Standard Dutch"), Ph.D. thesis, Katholieke Universiteit Nijmegen, Nijmegen.

van Hout, R., de Schutter, G., de Crom, E., Huinck, W., Kloots, H., and Van de Velde, H. (**1999**). "De uitspraak van het Standaard-Nederlands. Variatie en varianten in Vlaanderen en Nederland" ("The pronunciation of Standard Dutch. Variation and variants in Flanders and The Netherlands"), in *Artikelen van de Derde Sociolinguïstische Conferentie (Articles of the Third Sociolinguistic Conference)*, edited by E. Huls and B. Weltens (Eburon, Delft), pp. 183–196.

Voeten, C. C. (**2021a**). "How long is 'a long term' for sound change? The effect of duration of immersion on the adoption of on-going sound change," Lang. Dyn. Change. **12**, 28–77.

Voeten, C. C. (**2021b**). "Individual differences in the adoption of sound change," Lang. Speech **64**, 705–741.

Voeten, C. C. (**2021c**). "Regional variation in ongoing sound change: The case of the Dutch diphthongs," J. Linguist. Geogr. **9**, 162–177.

Watt, D. J. L., and Fabricius, A. H. (**2002**). "Evaluation of a technique for improving the mapping of multiple speakers," Leeds Work. Papers Linguist. Phonetics **9**, 159–173, available at https://www.latl.leeds.ac.uk/wp-content/uploads/sites/49/2019/05/Watt_Fab_2002.pdf.

Wood, S. N. (**2003**). "Thin-plate regression splines," J. R. Stat. Soc. **65**(1), 95–114.

Wood, S. N. (**2017**). *Generalized Additive Models: An Introduction with R*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, FL).

Wood, S. N., and Fasiolo, M. (**2017**). "A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models," Biometrics **73**(4), 1071–1081.