A comparison of four vowel overlap measures

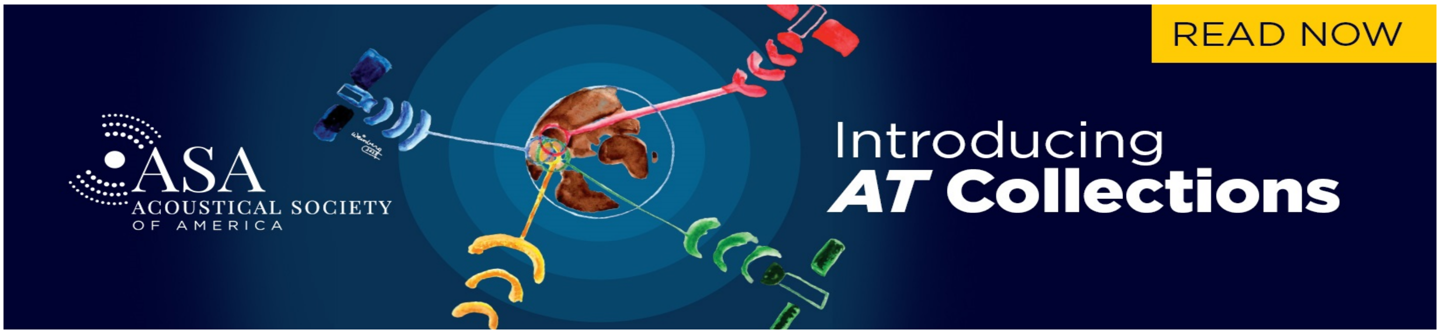
Matthew C. Kelley and Benjamin V. Tucker

Citation: The Journal of the Acoustical Society of America **147**, 137 (2020); doi: 10.1121/10.0000494

View online: https://doi.org/10.1121/10.0000494

View Table of Contents: https://asa.scitation.org/toc/jas/147/1

Published by the Acoustical Society of America



JASA ARTICLE

A comparison of four vowel overlap measures

Matthew C. Kelley[a] and Benjamin V. Tucker[b]
Department of Linguistics, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

ABSTRACT:
Multiple measures of vowel overlap have been proposed that use F1, F2, and duration to calculate the degree of overlap between vowel categories. The present study assesses four of these measures: the spectral overlap assessment metric [SOAM; Wassink (2006). J. Acoust. Soc. Am. **119**(4), 2334–2350], the *a posteriori* probability (APP)-based metric [Morrison (2008). J. Acoust. Soc. Am. **123**(1), 37–40], the vowel overlap analysis with convex hulls method [VOACH; Haynes and Taylor, (2014). J. Acoust. Soc. Am. **136**(2), 883–891], and the Pillai score as first used for vowel overlap by Hay, Warren, and Drager [(2006). J. Phonetics **34**(4), 458–484]. Summaries of the measures are presented, and theoretical critiques of them are performed, concluding that the APP-based metric and Pillai score are theoretically preferable to SOAM and VOACH. The measures are empirically assessed using accuracy and precision criteria with Monte Carlo simulations. The Pillai score demonstrates the best overall performance in these tests. The potential applications of vowel overlap measures to research scenarios are discussed, including comparisons of vowel productions between different social groups, as well as acoustic investigations into vowel formant trajectories. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0000494

(Received 16 September 2019; revised 26 November 2019; accepted 2 December 2019; published online 17 January 2020)

[Editor: Ewa Jacewicz] Pages: 137–145

I. INTRODUCTION

Vowel category overlap is a phenomenon where vowel categories have some degree of acoustic similarity. Measuring vowel category overlap has applications across a wide variety of research that deals with speech acoustics. In dialectology, for example, measures of vowel overlap have been used to assess vowel mergers, such as by Nycz and Hall-Lew (2015), Freeman (2014), and Wong and Hall-Lew (2014). A prominent historical example in certain dialects of North American English is the *cot–caught* merger, where the historic /ɑ/ and /ɔ/ phonemes are produced as one phoneme /ɑ/. A high measured degree of overlap between two putative vowel categories would suggest to a researcher that a merger is in progress or has been completed. In language documentation scenarios, vowel overlap measures have been used to help determine the overall contribution of duration to vowel category distinctiveness (Haynes and Taylor, 2014). The underlying idea is to measure how much two vowel categories overlap in terms of F1 and F2, and then see how much they overlap in terms of F1, F2, and duration. If the latter measure of overlap is substantially smaller, duration could be said to be contributing to the distinction of the two vowel categories. Having such knowledge could be especially useful if a researcher is attempting to determine whether there is a phonemic length contrast in a language.

In second language acquisition, measuring vowel overlap could be used to quantify how far apart a learner's productions of different vowel categories are. Perry and Tucker (2019) perform this kind of analysis, examining the extent to which different vowel categories in function words overlap in L2 and L1 speakers of English. Similarly, Mairano *et al.* (2019) investigate how much overlap is observed in L2 English speakers' productions of vowel categories that are acoustically distinct for L1 speakers but may not be for L2 speakers. And in clinical settings, such measures have the potential to be used to quantify how acoustically separable a patient's vowel categories are, as suggested by Kain *et al.* (2017). Similar to the aforementioned L2 speech studies, a researcher or clinician could examine a patient's productions of vowels across separate phonemic categories and quantify the extent to which the categories are being separated acoustically. In the present paper, we test and critique a selection of vowel category overlap measures for their performance on accuracy and precision.

Measuring vowel category overlap requires that two decisions be made. The first decision is how the concept of overlap should be defined. The second decision is how to operationalize the definition of overlap. Researchers have made a variety of decisions on each point, resulting in a number of potential approaches for measuring overlap. For example, some researchers use Euclidean distance from centroids to quantify degree of overlap or merger. Treating overlap in this way is not particularly satisfying, as it does not account for the distributional properties in the data, such as variability of the data or how densely populated a region of the vowel space is. However, some measures of vowel overlap do take advantage of the fact that vowel tokens have underlying distributions. The measures that are assessed in the present paper are the spectral overlap assessment metric (SOAM; Wassink, 2006), the *a posteriori* probability (APP)-based metric (Morrison, 2008), the vowel

[a] Electronic mail: mckelley@ualberta.ca. ORCID: 0000-0002-7218-5599.
[b] ORCID: 0000-0001-8965-7890.

overlap analysis with convex hulls (VOACH, Haynes and Taylor, 2014) method, and the application of Pillai scores to vowel overlap (Hay *et al.*, 2006).

We employ two criteria to assess the measures: (1) accuracy and (2) precision. We take accuracy to mean that the result of running a vowel overlap measure deviates as little as possible from true overlap values. For a measure to be precise, its outputs on a series of random samples taken from known vowel distributions must have little spread.

To our knowledge, Nycz and Hall-Lew (2015) is the only other study to compare measures of vowel overlap. Nycz and Hall-Lew compared Euclidean distance, linear mixed-effects regression modeling, SOAM, and the Pillai score. Their qualitative comparisons were based on their stated criteria of how well the measures capture distance, how well the measures capture overlap, how well they deal with unbalanced data, and whether they allow speaker comparisons. These criteria are important for assessments of which measure to use in practice, although they do not address theoretical questions regarding the measures' approaches, in general, nor do they address the issue of distributional properties as raised by Morrison (2008). Not all of the criteria lend themselves easily to rigorous testing either. As well, there are other proposed measures that have not yet been assessed in great detail. As such, the present study expands on Nycz and Hall-Lew's work by examining additional measures that have not yet been analyzed with more rigorously defined comparison criteria.

In the remainder of the present paper, we summarize and comment on the measures being compared. Next, we present theoretical critiques of the measures. We then perform two tests using simulated data. The tests are designed to compare the performance of the different metrics with respect to the criteria stated above. Using simulated data permits the calculation of true values against which to compare the measures. The first test uses Monte Carlo simulations to assess the measures' performance on two-dimensional (2D) data using F1 and F2. The second is similar to the first, but it assesses the measures' performance on three-dimensional (3D) data using F1, F2, and duration. (In theory, it is also possible to use other acoustic variables as input to these measures, such as F3 in place of duration.) In both tests, the simulated data are sampled from distributions based on steady-state formant values from all 139 speakers—men, women, and children—in the vowel data of Hillenbrand *et al.* (1995). We make use of the Lobanov (1971) normalization method to transform the formant data to reduce the effects of anatomical and physiological differences. Finally, we discuss the overall results from the simulations and make recommendations on which measures are applicable for general use.

As a terminological note, we employ the term "measure" to refer to each assessed method for evaluating overlap, similar to measures used in other fields, such as the Kullback-Leibler divergence or Jenson-Shannon divergence. Many researchers reserve the term "metric" for a specific class of dissimilarity measures that meet certain mathematical properties. Not all of the methods assessed in the present paper satisfy these properties. So, to maintain compatibility with terminology for similar concepts across different fields of research, we do not use metric to refer to the assessed methods as a group.

A. SOAM

The SOAM of Wassink (2006) uses elliptical representations of vowel categories to calculate its overlap measure. Three different normalization routines are described that can be used with SOAM: Nearey vowel-extrinsic normalization (Nearey, 1978), known-extremes vowel-extrinsic normalization (Shirai, 2005), and Lobanov vowel-extrinsic normalization (Lobanov, 1971). Within the analysis itself, Wassink appears to use Nearey vowel-extrinsic normalization routines. The ellipses/ellipsoids are calculated as centered at the origin in a rotated and translated space derived from the data. The process to fit an ellipse is as follows. Note that it does not use a standard least-squares ellipse fitting process.

(1) The data are centered around the origin by subtracting the mean F1 value from the F1 values and the mean F2 value from the F2 values. (2) The angle of rotation of the ellipse from the *x* axis is determined by fitting a linear regression to the data. The angle of rotation is taken as the angle between the line of best fit and the *x* axis. (3) The data are rotated by that angle toward the *x* axis, and the lengths of the ellipse's radii are determined by the standard deviations of F1 and F2 in the rotated space. The ellipse is then defined in the rotated and translated space as centered at the origin with principal axes extending along the *x* and *y* axes corresponding to the previously calculated standard deviations. Fitting an ellipsoid is analogous, except that there is an additional angle of rotation between the *x* and *z* axes.

Once the ellipses/ellipsoids have been calculated, their areas/volumes are approximated by means of a uniform grid of points in the 2D or 3D space. Each point is projected into the rotated and translated space and determined to be in one ellipse/ellipsoid, both, or neither. The area/volume of each ellipse/ellipsoid is taken to be represented by the number of points contained within it, and the overlap is then taken to be the larger of the ratios between the number of points in both figures and the number of points in one of the categories. The resulting number gives the shared area or volume between the base of a slice of normal distributions containing 87% of the probability density in the 2D case or 74% of the density in the 3D case (Wang *et al.*, 2015).

Figure 1(a) is a visualization of SOAM in two dimensions. The measure was run using the vowel data of Hillenbrand *et al.* (1995) for /u/ (orange triangles) and /ʊ/ (blue circles). Values presented are Lobanov normalized. All 139 speakers are used in the analysis. In the plot, the original data from all speakers are visualized and surrounded by the ellipses determined during SOAM's calculation. For SOAM and all other measures tested, the /u/ and /ʊ/ pairing was selected as an example of two vowels that

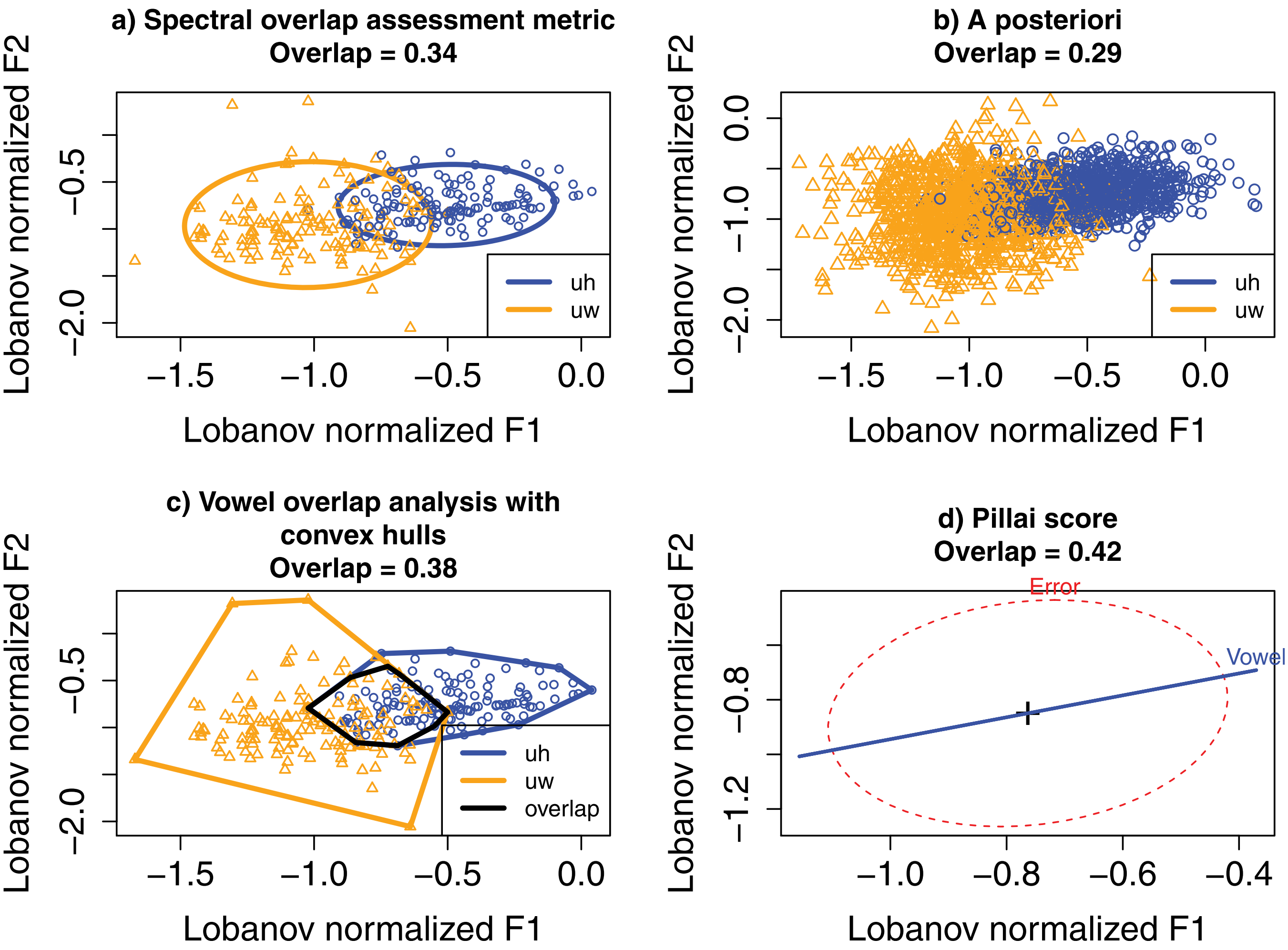


FIG. 1. (Color online) Visualizations for each measure for the vowel data of Hillenbrand *et al.* (1995) for /u/ (uw) and /ʊ/ (uh). The /u/ and /ʊ/ pairing was selected as an example of two vowels that should present some degree of acoustic overlap based on the similarity of their associated formant values. Formant values have been Lobanov normalized, and all 139 speakers in the data set were used in the calculation. Some speakers had missing formant values, which were denoted with a “0” in the data set. We included these speakers and the 0 values to help visualize how outliers affect the measures. Note that the scale differs between some of the plots, notably the plot for the Pillai score.

should present some degree of acoustic overlap based on the similarity of their associated formant values.

B. APP-based metric

The APP-based metric by Morrison (2008) uses maximum likelihood estimates of the underlying probability distributions for the vowel categories. In its original presentation, the APP-based metric used a log-interval normalization routine (Nearey and Assmann, 2007) on the vowel formants.

When calculating the measure on two vowels A and B, the procedure is as follows: (1) The sample mean vectors and covariance matrices for both vowels A and B are determined. (2) A large number of points (specified by the researcher) for each vowel are sampled from a multivariate normal distribution using the sample mean vectors and covariance matrices as parameters. (3) A quadratic discriminant analysis is applied to the generated data. The quadratic discriminant analysis is then used to calculate the posterior probability distribution over vowels A and B for each token. (4) Calculate the mean of the posterior probabilities of the tokens in vowel category A being in vowel category B. (5) Calculate the mean of the posterior probabilities of the tokens from vowel B being from vowel A. (6) Add the two means to yield the overall overlap measure, resulting in an overall mean probability that a token from one vowel could also be a token of the other vowel.

Figure 1(b) is a visualization of the APP-based metric in two dimensions. It also uses the vowel data of Hillenbrand *et al.* (1995) for /u/ (orange triangles) and /ʊ/ (blue circles). Formant values were Lobanov normalized before calculating the APP-based metric. All 139 speakers are used in the analysis. For the plot, the random number generation seed was set to a value of 9 before calculating the metric with 1000 generated tokens for each vowel category. The plot visualizes the generated data used in the quadratic discriminant analysis. These data points indicate the degree to which the two vowel distributions overlap, which should be relatively high when the two point clouds overlap with one another.

C. VOACH

The VOACH method by Haynes and Taylor (2014) is similar to SOAM, but it represents the vowels with convex hulls instead of ellipses/ellipsoids. In its original presentation, VOACH used a Nearey vowel-extrinsic routine (Nearey, 1978) on the analyzed vowel tokens. For a given set of points, a convex hull is the smallest geometric shape that will contain or cross through all of the points, like fitting a rubber band around all the points in two dimensions. In effect, it connects the points in the set along the perimeter.

Unlike the previous two measures, Haynes and Taylor (2014) claim that this procedure is designed to be more robust in researching under-documented languages or analyzing data sets with few data points. However, it is unclear what the motivation for their claims are, and as we will see later in the present study, these claims appear to be unfounded. Convex hulls have previously been used for analyzing vowel spaces, such as for speaker intelligibility estimates (Luan *et al.*, 2014), quantifying and visualizing how much a speaker uses different regions of the vowel space (Story and Bunton, 2017), and determining vowel space area (Sandoval *et al.*, 2013). An important distinction between these latter uses of convex hulls is that there is some kind of thresholding to exclude outlying data (Sandoval *et al.*, 2013; Story and Bunton, 2017) and look at the entire vowel space (Luan *et al.*, 2014; Sandoval *et al.*, 2013; Story and Bunton, 2017). As defined, VOACH maximizes the area for only a particular vowel category, which may not give the best representation of the category.

Haynes and Taylor (2014) describe the following steps for calculating the measure: (1) Fit convex hulls around the data points representing each vowel being assessed. (2) For each vowel token in the data set, determine whether it is contained in one or both hulls. (3) Fit a third convex hull around the points contained in both initial hulls. (4) The overlap is taken as the larger of the ratios of the area/volume between the hull containing the points in the overlapping region and hulls for each vowel category.

Figure 1(c) is a 2D example of VOACH. Again, we use the vowel data of Hillenbrand *et al.* (1995) for /u/ (orange triangles) and /ʊ/ (blue circles). Formant data were Lobanov normalized prior to calculation. All 139 speakers are used in the analysis. The plot displays a scatter of the original data as well as the convex hulls that the data are bounded by. Within the hulls, there is the black convex hull, which contains the overlapping data. Note that the hull for /u/ is oversized due to outlying data. In this case, the overlap value is not affected by this poor representation of the /u/ distribution because the hull for /ʊ/ is not distorted by outlying data and has a smaller area than the hull for /u/. As such, the hull for /ʊ/ will be the hull used in the final overlap calculation.

D. Pillai score

Hay *et al.* (2006) introduced the use of the Pillai trace statistic from the multivariate analysis of variance (also known as a MANOVA) technique to quantify vowel overlap. They refer to this statistic as the "Pillai score," which is a convention that has been adopted by other researchers such as Hall-Lew (2010) and Nycz and Hall-Lew (2015). They do not appear to have used any particular normalization routine on the vowel formants prior to calculating the Pillai score, but the raw formant values seem to have been converted to a Bark scale before analysis.

The Pillai score is calculated using the eigenvalues of the matrix formed by multiplying the hypothesis cross-product sum of squares matrix and the inverted error matrix together. The exact mathematical definition of the Pillai score can be found in Pillai (1954) and Bray and Maxwell (1985).

At a higher level, the Pillai score is a multivariate analog of the F-ratio test statistic from an analysis of variance (ANOVA). For a two vowel scenario, when a large amount of the variation in the data is due to vowel categories differing from each other, the Pillai score will be near one, corresponding to a small amount of vowel overlap. When the variation is more likely to be due to random variation in the data, the Pillai score will be near zero, indicating high vowel overlap.

Because the other three measures considered in the present study define values around one as indicating high overlap and values around zero as indicating low overlap, we use the inverse of the Pillai score, which is one minus the Pillai score. This modification will allow the same tests that are used for the other three measures to be used for the Pillai score as well.

Figure 1(d) visualizes the MANOVA behind the Pillai score for a 2D comparison, using what is known as an HE plot—or hypothesis-error plot—from the HEPLOTS package in R (Fox *et al.*, 2018; Friendly, 2007). In this plot, both the hypothesis and error terms of the MANOVA analysis are displayed, comparing the effect of the between-group variation to the within-group variation. Linear models like MANOVA have a hypothesis term and an error term, where the hypothesis term represents the variation explained by the model, and the error term represents the residual variation left unexplained. The size of the hypothesis and error ellipses in Fig. 1(d) represent how much variation the hypothesis term and error term explain in the model. The larger the error ellipse is in comparison to the vowel ellipse or line, the greater the degree of overlap indicated by the Pillai score. Note that before creating the plot, the formant values were Lobanov normalized, and all 139 speakers are used in the analysis.

II. THEORETICAL ANALYSIS

Among the four approaches being analyzed, there are three conceptions of overlap and four different procedures for calculating overlap. The four procedures have a variety of strengths and weaknesses.

The first conception of overlap is shared by SOAM and VOACH. Overlap is taken as the common space shared between two probability distributions. Inman (1984) provides the analytical solution to this concept of overlap through integration. Procedurally, though, neither measure makes use of the analytical solution, opting instead to resolve a simpler but related problem. They flatten the probability density function to be uniform over a given region delimited by ellipses/ellipsoids or convex hulls. They then approximate, rather than solve for, the amount of space shared between the two regions. This simplification is peculiar because it removes the influence of probability density on the output, even though probability density is central to the concept of the shared space between two probability density functions. There are thus two sources for potential deviation from the analytical solution: (1) the simplification of the overlap problem and (2) the approximation of the space shared between the ellipses/ellipsoids or convex hulls. The first source of deviation implies that these procedures will never converge on the theoretical overlap value.

The APP-based metric takes overlap as a sum of mean probabilities that a vowel in one category belongs to the other category. The analytical solution for this conception of overlap is the sum of the mean probability values for each analyzed population. The procedure provides an estimate that will converge on the theoretical value as (1) the sample mean and covariance approach the population values with larger samples and (2) the number of generated samples increases.

The Pillai score represents overlap as the amount of variation in the data that is explained by group differences as opposed to random variation in the data. The theoretical value would be the Pillai score as calculated on the population distributions, and its estimate is simply the Pillai score as calculated on a random sample. As the size of the samples grows, the estimated Pillai score will converge on the true value.

There are two theoretical advantages to using the APP-based measure or the Pillai score as opposed to SOAM or VOACH. The first is that the APP-based metric and the Pillai score as calculated on samples will converge on their population values in appropriate conditions. SOAM and VOACH will not converge on the correct answer, rendering them categorically more biased estimators of overlap. The second advantage is that the sources of additional variation in SOAM and the VOACH method will drive up their sampling variability. On these grounds, SOAM and VOACH are theoretically both less accurate and less precise than the Pillai score and the APP-based metric.

In the remainder of this paper, we empirically demonstrate that SOAM and VOACH have lower accuracy and precision than the APP-based metric and Pillai score. We use Monte Carlo simulations to test the measures' performance in three different potential vowel overlap conditions: little overlap, partial overlap, and full overlap. We test both the 2D and 3D versions of these measures. In line with our theoretical analysis, we expect the Pillai score and the APP-based metric to be both more accurate and more precise than SOAM and VOACH.

III. TEST 1: 2D OVERLAP CALCULATIONS

The first aspect tested is the accuracy and precision of the measures in the 2D condition. In this case, the two dimensions are the F1 and F2 of the vowels being analyzed.

A. Assesing the measures

The measures are assessed using two statistics. The first statistic is the mean absolute error from the desired output. The mean absolute error represents, on average, how far away the measure's output is from the target. Small values indicate that the output is close to the target, indicating high accuracy. Larger values indicate that the output is further from the target, indicating lower accuracy. Thus, an accurate measure of vowel overlap has low mean absolute error.

The second statistic used is the standard deviation of each measure's outputs. It is a measure of spread, which represents the measures' precision.

B. Methods

The measures were assessed using Monte Carlo simulations with steady-state F1 and F2 values from all 139 speakers in the data set. The little overlap condition compares /i/ and /ɑ/ because the acoustic distance between them is high, so there should be little overlap between them. The partial overlap condition compares /u/ and /ʊ/ as an example of a vowel pair that has some degree of acoustic overlap. The full overlap condition compares two separate random samples of /i/ data. Before beginning any of the simulations, the vowel data set of Hillenbrand *et al.* (1995) was normalized using the Lobanov normalization procedure (Lobanov, 1971) in the PHONTOOLS (Barreda, 2015) package in R (*R* Core Team, 2017). Other normalization procedures could have been selected at this stage, but the Lobanov method was chosen to reflect current trends and advice in the sociophonetic literature (e.g., Adank *et al.*, 2004; Fridland and Kendall, 2017; Hall-Lew *et al.*, 2017). However, see Barreda and Nearey (2018) for a discussion of why the Lobanov normalization routine may not reflect listener vowel perception as well as log-mean normalization routines.

The use of the Lobanov normalization routine only matches up with the description of SOAM, where the Lobanov normalization routine was described as a possible option. However, insofar as the goal of vowel formant normalization is to reduce variation in the data due to anatomical and physiological differences among speakers while also retaining variation due to sociolinguistic differences (Adank *et al.*, 2004), any such normalization routine that is appropriately applied and reduces undesired variation should function equally well as a preprocessing step for each of the measures being analyzed.

A simulation scenario was run 1000 times on the normalized data, and the results of each run were recorded. For each run of a simulation, the data were sampled from a bivariate normal distribution using the mvrnorm function from the MASS (Venables and Ripley, 2002) R package. Thirty data points were generated for both vowel categories being analyzed for a total of 60 data points in each trial of a simulation. The sample size of 30 per category was chosen as a stress test to assess how the measures perform with relatively small samples.

Bivariate normal distributions were chosen because SOAM, the APP-based metric, and the Pillai score assume normally distributed data, and VOACH makes no assumptions about the statistical distribution of the data. It stands to reason that if the measures perform poorly on data that match their ideal assumptions, they would not perform any better in conditions where those assumptions are violated. As for VOACH, its results are designed to be invariant to the way that the data are distributed. Furthermore, Whalen and Chen (2019) found preliminary evidence that F1 and F2 follow normal distributions. Thus, drawing simulated data from a bivariate normal distribution has some degree of ecological validity.

In the little and partial overlap tests, the ground truth values were determined in three different ways to reflect the three different conceptions of overlap that the measures employ and not favor one conception of overlap over another. These different procedures result in different target values for each measure. For each category, 100,000 points were generated from a bivariate random number generator to simulate a large population based on the mean and covariance statistics of each tested vowel category. The overlapping coefficient in Inman (1984) is used as the theoretical target for SOAM and VOACH. For the APP-based metric, the classification rule from quadratic discriminant analysis was applied to each class to get the posterior probabilities, and the probabilities were averaged and summed as in the procedure definition. The Pillai score target is calculated on the 100 000 point simulated populations.

In each little and partial overlap simulation run, the mean and covariance statistics from the population sets are used as the seeds for the random samples of each vowel. In the full overlap test, the mean and covariance statistics for /i/ directly from the Hillenbrand *et al.* (1995) data are used as the seed for generating all random samples in each Monte Carlo simulation round. Because the underlying distribution is the same in the full overlap test, the target value for all the vowel overlap measures is one.

C. Results

The results of the 2D simulations may be found in Table I.

In the no overlap scenario, only the APP-based metric and Pillai score produce usable results with the APP-based metric performing the best in terms of both accuracy and precision. The value of exactly zero that SOAM and VOACH output is due to them assigning a probability density of zero to any formant configuration that falls outside the perimeter of the region bounded by their shapes. The overall magnitude of the errors for all of the measures is not large, however.

As for the partial overlap condition, the Pillai score had the lowest error and the lowest spread, followed by the APP-based metric. The Pillai score also performed the best in the full overlap case. Notably, VOACH had a substantially higher mean absolute error than the other measures, showing that it is far less accurate than the other measures.

D. Discussion

These results confirm our theoretical argument that the Pillai score and the APP-based metric are more accurate and precise than SOAM and VOACH. The Pillai score shows the best overall performance. The APP-based metric is better in the little overlap condition, but the Pillai score's average output of 0.02 is still readily interpretable as an indicator of little to no overlap. For the 2D case of comparing F1 and F2, the Pillai score generally provides the closest fit to the tested target scores.

It should be noted that SOAM did outperform the APP-based metric in the full overlap condition. This pattern is not expected to be the case with larger samples and larger numbers of points generated in the APP-based metric.

IV. TEST 2: 3D OVERLAP CALCULATIONS

This test set of simulations was analogous to the 2D simulations except that the 3D versions of the measures were used with F1, F2, and duration.

A. Methodology

Each of these simulation conditions was analogous to the 2D simulations. We use the Lobanov normalization routine from the PHONTOOLS package on the duration values to place the duration values on the same relative scale as the formant values. In doing so, we minimize errors induced from the variables being on different scales.

TABLE I. Summary results for the 2 D versions of SOAM, the APP-based metric, VOACH, and the Pillai score. Presented for each condition are the target value, each measure's mean output, the standard deviation (SD) of each measure's output, and the mean absolute error (MAE) of each measure's output from the target. The target and mean are merely presented to help contextualize the results; only the SD and the MAE are used when analyzing the measures' relative performance. The MAE is used to assess accuracy, and the standard deviation is used to assess precision. The best results are in boldface, and results that stem from errors induced by a particular measure's implementation are in italics.

	No overlap: /i/ vs /ɑ/				Partial overlap: /u/ vs /ʊ/				Full overlap: Two samples of /i/			
Measure	Target	Mean	MAE	SD	Target	Mean	MAE	SD	Target	Mean	MAE	SD
SOAM	4.9e-14	0	*4.9e-14*	*0*	0.22	0.32	0.13	0.12	1	0.93	0.067	0.06
APP	9.2e-17	1.5E-16	**1.9E-16**	**3.2e-15**	0.32	0.30	0.07	0.08	1	0.93	0.075	0.04
VOACH	4.9e-14	0	*4.9e-14*	*0*	0.22	0.24	0.11	0.14	1	0.72	0.28	0.10
Pillai	0.02	0.02	0.003	0.003	0.42	0.41	**0.06**	**0.07**	1	0.97	**0.03**	**0.03**

B. Results

The results of the 3D simulations can be seen in Table II. These results mirror those of the 2D simulations. SOAM and VOACH consistently output values of zero in the little overlap condition. And, the APP-based metric performs better than the Pillai score in the little overlap condition with the Pillai score showing practically competitive performance regardless.

In the partial and full overlap simulations, the Pillai score again shows greater accuracy and precision with the APP-based metric coming in as the second-best performer. Particular to the full overlap simulations, only the Pillai score was reasonably close to the true value of one. The APP-based metric and SOAM are still high enough that a researcher could reasonably interpret the values as indicating a high degree of overlap. On the other hand, VOACH is so far away from the target value in the full overlap condition that it could not be reasonably interpreted as indicating a high degree of overlap. However, its spread is the second lowest, meaning that it is consistent, but it consistently shows high degrees of error.

C. Discussion

In the 3D tests, the same pattern as seen in the 2D tests is present. In the little overlap scenario, the APP-based metric and Pillai score perform the best and output small positive values. Otherwise, in both the partial and full overlap scenarios, the Pillai score performs the best. Overall, these results further confirm the theoretical analysis that the Pillai score and the APP-based metric are the most accurate and precise measures.

V. GENERAL DISCUSSION

Both the 2D and 3D results serve as empirical evidence for our theoretical analysis that the Pillai score and APP-based metric are more accurate and precise. And, between the two, the Pillai score provided the highest level of overall accuracy and precision in our tests.

The estimate of the Pillai score on a small sample is already performing well. For that reason, we argue that it is the preferred measure of vowel overlap to use when samples are small. As the sample size grows, all the measures will benefit from better estimates of the population parameters. However, as discussed in the theoretical analysis, only the Pillai score and the APP-based metric will converge on their population values. As such, the Pillai score or the APP-based metric are the most appropriate measures for calculating vowel overlap on larger samples. When the covariance is believed to be different between the two groups, the APP-based metric is theoretically preferred because the Pillai score assumes the covariance is the same between all tested groups.

Note that the conception of vowel overlap as the shared space between two probability density functions has not yet been represented in our findings. We believe that either a modified version of SOAM should be developed for this use that will converge on the correct value, or else the sample-based estimate of the overlapping coefficient (Inman, 1984) should be used to assess vowel overlap. A modified version of SOAM would need to account for density to converge on the theoretical value it estimates because density is intrinsic to SOAM's concept of overlap.

As for VOACH, there does not seem to be a compelling reason for its use. It performed poorly overall in our tests. And, even if its performance was due solely to a small sample size, it would not converge on the correct value with larger sample sizes. When it is unreasonable to assume that the vowel data are normally distributed, a nonparametric estimate of the overlapping coefficient, as Schmid and Schmidt (2006) describe, would be theoretically preferable.

Future work should also endeavor to find empirically grounded definitions of little overlap, partial overlap, and full overlap so as to contextualize the calculations of these vowel overlap measures. This process may also involve finding confidence intervals for the measures based on the sample size being analyzed. Additionally, it would be insightful to see how the measures perform in more real-world scenarios to assess their utility in situations that are closer to what a researcher may encounter, including data that are not normally distributed. Other potential measures of vowel overlap should be considered, such as the overlapping coefficient calculated on sampled data rather than populations (Inman, 1984), or the cross-entropy of the two distributions as used by Ghorshi *et al.* (2008), neither of which have yet been applied to the investigation of vowel overlap.

Having found the Pillai score and APP-based metric to be appropriate for measuring vowel overlap, we would like

TABLE II. Summary results for the 3 D versions of SOAM, the APP-based metric, VOACH, and the Pillai score. Presented for each condition are the target value, each measure's mean output, the standard deviation (SD) of each measure's output, and the mean absolute error (MAE) of each measure's output from the target. The target and mean are merely presented to help contextualize the results; only the SD and the MAE are used when analyzing the measures' relative performance. The MAE is used to assess accuracy, and the SD is used to assess precision. The best results are in boldface, and results that stem from errors induced by a particular measure's implementation are in italics.

	No overlap: /i/ vs /ɑ/				Partial overlap: /u/ vs /ʊ/				Full overlap: Two samples of /i/			
Measure	Target	Mean	MAE	SD	Target	Mean	MAE	SD	Target	Mean	MAE	SD
SOAM	4.4e-14	0	*4.4E-14*	*0*	0.14	0.10	0.06	0.05	1	0.83	0.17	0.10
APP	1.9e-24	2.7e-15	**2.7E-15**	**8.5e-14**	0.21	0.20	0.04	0.05	1	0.87	0.13	0.05
VOACH	4.4e-14	0	*4.4E-14*	*0*	0.14	0.07	0.09	0.05	1	0.36	0.64	0.09
Pillai	0.02	0.02	0.003	0.003	0.34	0.33	**0.03**	**0.04**	1	0.95	**0.05**	**0.04**

to suggest some lines of research that may benefit from using quantitative measures of overlap. As mentioned in the Introduction, examining the contrastiveness of two different vowel categories could benefit from using vowel overlap measures. One relevant question could be whether duration, for example, is contributing much to the acoustic separation of two vowel categories. Measuring how much the categories overlap when using only F1 and F2 and, then, comparing how much that overlap level does or does not change when duration is added into the mix can quantify just how much duration is actually contributing to the separation. Wassink (2006) and Haynes and Taylor (2014) perform such an analysis, and Morrison (2008) and Wassink (2006) provide some guidelines about how best to interpret the differences between the F1-by-F2 and F1-by-F2-by-duration overlap values, such as how it is appropriate to directly compare those overlap values for the APP-based metric but not for SOAM. Similarly, studies that examine vowel production across different social variables may gain additional insight into group differences in production by measuring the degree of vowel category overlap. A simple research question may be whether certain dialect groups front /u/ more than others, which could be addressed by comparing / u/ productions across dialect groups, as well as perhaps how close each group's /u/ is to /i/.

Some aspects of vowel inherent spectral change (Nearey and Assmann, 1986) could also be studied using a measure of vowel overlap. For example, Nearey and Assmann (1986) find that /ɛ/ in Canadian English is backed and lowered during the process of articulation. Taking formant measurements at the beginning and end of multiple productions of a particular vowel category and, then, measuring the overlap between the initial and final sections of the vowel productions would yield a quantification of how much the formant configurations change over time. Such an analysis could be made more fine-grained by taking measurements at more incremental steps through the vowel and comparing those incremental measurements to each other. The time-course of the distributional change in the formants could then be observed. And to that end, any study that is comparing vowel productions across time points, linguistic categories, and/or social groups could benefit from a measure of vowel overlap because acoustic parameters are treated distributionally (or as bundles) instead of in isolation from each other. There is also evidence that distributions of acoustic features are relevant to infant speech learning (Wanrooij *et al.*, 2014), so studies may be more ecological to cognitive reality when treating vowels as distributions.

A limitation of the current analysis is that the simulations were run on idealized data, whereas samples of real data have more noise in the measurements and are less likely to be balanced. It could be the case that certain measures of vowel overlap are more robust to noise or imbalance in the data, which our simulations do not address. Future research building on the work in the present paper should examine how measures of overlap fare in more realistic scenarios, as well as how to make informed choices about the methods with regard to statistical properties like the bias-variance trade-off.

Readers may view a bundle containing the R implementation of these measures, the script used to run the simulations, and the markup for this document online (Kelley and Tucker, 2019).

VI. CONCLUSION

In this study, four different measures of vowel overlap were examined: SOAM (Wassink, 2006), the APP-based metric (Morrison, 2008), VOACH (Haynes and Taylor, 2014), and a modified version of the Pillai score (Hay *et al.*, 2006). They were tested using a Monte Carlo simulation technique to examine their performance in terms of accuracy and precision for 2D (F1 and F2) and 3D (F1, F2, and duration) cases.

Overall, the Pillai score performed best in the Monte Carlo simulations. The APP-based metric performed the second best, and its performance should increase with larger samples and greater numbers of generated points. For the present moment, the Pillai score and APP-based metric are the theoretically preferred options. Thus, researchers have more options for calculating vowel overlap than Euclidean distance, which better account for the distributional properties of their data.

ACKNOWLEDGMENTS

We would like to thank Geoffrey Stewart Morrison for his numerous comments and corrections on earlier drafts of this work, as well as for giving us the idea for this study. We are also grateful to Alicia Beckford Wassink for graciously assisting in the process of implementing SOAM in R. As well, the authors would like to thank the attendees of the NorthWest Phonetics and Phonology 2016 Conference, the Alberta Conference on Linguistics 2017, and the 177th Meeting of the Acoustical Society of America for their helpful comments and feedback.

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

Barreda, S. (**2015**). "phonTools: Functions for phonetics in R" (R package version 0.2-2.1) [computer program], available at https://cran.r-project.org/web/packages/phonTools/index.html (Last viewed January 6, 2020).

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.

Bray, J. H., and Maxwell, S. E. (**1985**). *Multivariate Analysis of Variance* (Sage, Beverly Hills, CA), Quantitative Applications in the Social Sciences, Vol. 54.

Fox, J., Friendly, M., and Monette, G. (**2018**). "heplots: Visualizing tests in multivariate linear models (R package version 1.3-5) [computer program]," available at https://CRAN.R-project.org/package=heplots (Last viewed January 6, 2020).

Freeman, V. (**2014**). "Bag, beg, bagel: Prevelar raising and merger in Pacific Northwest English," Univ. Wash. Work. Pap. Linguist. **32**, 1–23.

Fridland, V., and Kendall, T. (**2017**). "Speech in the silver state," Publ. Am. Dialect Soc. **102**(1), 139–164.

Friendly, M. (**2007**). "HE plots for multivariate general linear models," J. Comput. Graphic. Stat. **16**(4), 421–444.

Ghorshi, S., Vaseghi, S., and Yan, Q. (**2008**). "Cross-entropic comparison of formants of British, Australian and American English accents," Speech Commun. **50**(7), 564–579.

Hall-Lew, L. (**2010**). "Improved representation of variance in measures of vowel merger," Proc. Meet. Acoust. **9**, 060002.

Hall-Lew, L., Eiswirth, M., Valentinsson, M.-C., and Cotter, W. (**2017**). "Northern Arizona vowels," Publ. Am. Dialect Soc. **102**(1), 59–82.

Hay, J., Warren, P., and Drager, K. (**2006**). "Factors influencing speech perception in the context of a merger-in-progress," J. Phonetics **34**(4), 458–484.

Haynes, E. F., and Taylor, M. (**2014**). "An assessment of acoustic contrast between long and short vowels using convex hulls," J. Acoust. Soc. Am. **136**(2), 883–891.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.

Inman, H. F. (**1984**). "Behavior and properties of the overlapping coefficient as a measure of agreement between distributions (association, dissimilarity)," Ph.D. dissertation, The University of Alabama at Birmingham, available at http://search.proquest.com/docview/303295844/abstract/6A03AD442D6C405BPQ/1 (Last viewed January 6, 2020).

Kain, A., Giudice, M. D., and Tjaden, K. (**2017**). "A comparison of sentence-level speech intelligibility metrics," in *Interspeech 2017, ISCA*, pp. 1148–1152, available at http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0567.html (Last viewed January 6, 2020).

Kelley, M. C., and Tucker, B. V. (**2019**). "Supplementary files for 'A comparison of four vowel overlap measures," available at https://doi.org/10.7939/r3-ayh1-5x98 (Last viewed January 6, 2020).

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

Luan, Y., Wright, R., Ostendorf, M., and Levow, G.-A. (**2014**). "Relating automatic vowel space estimates to talker intelligibility," in *INTERSPEECH-2014*, available at https://www.isca-speech.org/archive/interspeech_2014/i14_2238.html (Last viewed January 6, 2020).

Mairano, P., Bouzon, C., Capliez, M., and De Iacovo, V. (**2019**). "Acoustic distance, Pillai scores, and LDA classification scores as metrics of L2 comprehensibility and nativelikeness," in *Proceedings of the 19th International Congress of Phonetic Sciences*, edited by S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Melbourne, Australia, pp. 1104–1108.

Morrison, G. S. (**2008**). "Comment on 'A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties,' J. Acoust. Soc. Am. **119**, 2334–2350 (2006)]," J. Acoust. Soc. Am. **123**(1), 37–40.

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

Nearey, T. M., and Assmann, P. F. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**(5), 1297–1308.

Nearey, T. M., and Assmann, P. F. (**2007**). "Probabilistic 'sliding template' models for indirect vowel normalization," in *Experimental Approaches to Phonology*, edited by M. J. Solé, P. S. Beddor, and M. Ohala (Oxford University Press, Oxford), pp. 246–269.

Nycz, J., and Hall-Lew, L. (**2015**). "Best practices in measuring vowel merger," Proc. Meet. Acoust. **20**, 060008.

Perry, S. J., and Tucker, B. V. (**2019**). "L2 production of American English vowels in function words by Spanish L1 speakers," in *Acoustics Week in Canada 2019 Conference Proceedings*, Edmonton, Alberta, Canada.

Pillai, K. C. S. (**1954**). "On some distribution problems in multivariate analysis," Technical Report 88 (Institute of Statistics, University of North Carolina, Chapel Hill, NC).

R Core Team (**2017**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/ (Last viewed January 6, 2020).

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., and Spanias, A. (**2013**). "Automatic assessment of vowel space area," J. Acoust. Soc. Am. **134**(5), EL477–EL483.

Schmid, F., and Schmidt, A. (**2006**). "Nonparametric estimation of the coefficient of overlapping–theory and empirical application," Comput. Stat. Data Anal. **50**(6), 1583–1596.

Shirai, S. (**2005**). "Lexical effects in Japanese vowel reduction," Ph.D. dissertation, University of Washington, available at https://digital.lib.washington.edu:443/researchworks/handle/1773/8381 (Last viewed January 6, 2020).

Story, B. H., and Bunton, K. (**2017**). "Vowel space density as an indicator of speech performance," J. Acoust. Soc. Am. **141**(5), EL458–EL464.

Venables, W. N., and Ripley, B. D. (**2002**). *Modern Applied Statistics with S*, fourth ed. (Springer, New York).

Wang, B., Shi, W., and Miao, Z. (**2015**). "Confidence analysis of standard deviational ellipse and its extension into higher dimensional Euclidean space," PLoS One **10**(3), e0118537.

Wanrooij, K., Boersma, P., and Zuijen, T. L. v. (**2014**). "Distributional vowel training is less effective for adults than for infants. A study using the mismatch response," PLoS One **9**(10), e109806.

Wassink, A. B. (**2006**). "A geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties," J. Acoust. Soc. Am. **119**(4), 2334–2350.

Whalen, D. H., and Chen, W.-R. (**2019**). "Variability and central tendencies in speech production," Front. Commun. **4**, 1–9.

Wong, A. W.-M., and Hall-Lew, L. (**2014**). "Regional variability and ethnic identity: Chinese Americans in New York City and San Francisco," Lang. Commun. **35**, 27–42.