Characterizing the distribution of the quadrilateral vowel space area

Visar Berisha[a)]
Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287

Steven Sandoval
School of ECEE, SenSIP Center, Arizona State University, Tempe, Arizona 85287

Rene Utianski and Julie Liss
Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287

Andreas Spanias
School of ECEE, SenSIP Center, Arizona State University, Tempe, Arizona 85287

(Received 22 March 2013; revised 7 October 2013; accepted 25 October 2013)

The vowel space area (VSA) has been studied as a quantitative index of intelligibility to the extent it captures articulatory working space and reductions therein. The majority of such studies have been empirical wherein measures of VSA are correlated with perceptual measures of intelligibility. However, the literature contains minimal mathematical analysis of the properties of this metric. This paper further develops the theoretical underpinnings of this metric by presenting a detailed analysis of the statistical properties of the VSA and characterizing its distribution through the moment generating function. The theoretical analysis is confirmed by a series of experiments where empirically estimated and theoretically predicted statistics of this function are compared. The results show that on the Hillenbrand and TIMIT data, the theoretically predicted values of the higher-order statistics of the VSA match very well with the empirical estimates of the same.
 [http://dx.doi.org/10.1121/1.4829528]

PACS number(s): 43.71.Gv, 43.72.Ar [MAH] Pages: 421–427

I. INTRODUCTION

The vowel space area (VSA), defined as the area of the quadrilateral formed by the four corner vowels when projected on the first two formant frequencies (F1 and F2), is often used to characterize speech motor control.[1–7] Frequencies of the first and second formants roughly relate to the size and shape of the cavities created by jaw opening (F1) and tongue position (F2). As such, VSAs are acoustic proxy for the kinematic displacements of the articulators. Figure 1 shows a sample quadrilateral that forms the VSA for a group of individuals. In particular, the figure demonstrates the distribution of the first two formant frequencies for the four corner vowels that define the area of the vowel space (data from Hillenbrand[8]). The area (in Hz^2) of the shape defined by these vowels serves as a quantitative measure of articulatory displacement for this group of speakers. This metric is interpreted as a measure of articulatory excursions and separability between distinct acoustic-articulatory vowel targets. This interpretation makes the VSA an attractive metric for characterizing speech motor control, including speech development,[1,2] speech disorders,[3,4] speech interventions,[5] dialects,[6] and speaking styles.[7]

A fundamental issue often unaccounted for in existing VSA studies is that vowel acoustics and the associated articulatory kinematics in connected speech are non-deterministic. This is largely the result of anticipatory and carry-over coarticulation, which directly influence vowel formant frequencies. Further, speaking effort (clear versus conversational), speaking rate, regional dialects, idiosyncratic speaking styles, and VSAs computed for groups of individuals (as in Fig. 1) also contribute to the stochastic nature of vowel acoustics and production. None of the commonly used methods for estimating VSA accommodates this fact. Indeed, the /hVd/ context commonly used for generating vowel samples for VSA estimation is designed to reduce the effects of coarticulation on vowel formant values. This results in incomplete characterizations of the VSA that focus on average values of the area formed by /hVd/ stimuli. It is hypothesized that to fully understand its utility as a measure of articulatory excursion (and intelligibility, by proxy), the VSA must be characterized stochastically through statistics that more completely describe the underlying distribution of the area.

Here the main contribution is to extend the mathematical analysis of the VSA by treating it as a random variable and characterizing its full distribution rather than only its average. It is important to note that the aim of this work is *not* to confirm or refute the utility of this metric as a measure of intelligibility. Rather, under reasonable assumptions on the distribution of the formant frequencies for the four corner vowels, the distribution of the vowel area is characterized by defining a closed-form expression for its moment generating function. From this, expressions for a series of higher-order statistics (variance, skewness, kurtosis, etc.) are derived, and their accuracy is confirmed using numerical experiments. The newly derived expressions can be used by researchers in

[a)]Author to whom correspondence should be addressed. Also at: School of ECEE, SenSIP Center, Arizona State University, Tempe, AZ 85287. Electronic mail: visar@asu.edu

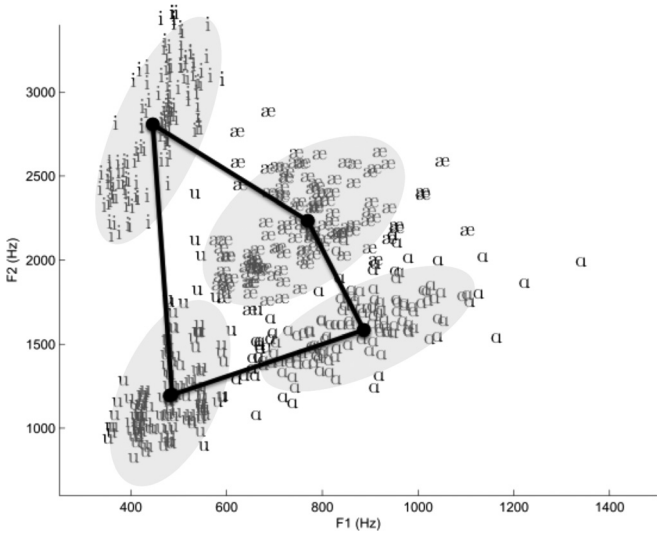

FIG. 1. The distribution of the first two formant frequencies for the four corner vowels that define the area of the vowel space. The mean of each distribution serves as an edge for the vowel space quadrilateral.

the field to more completely characterize the vowel area in future studies and to study the relationship between intelligibility (or other measures of articulator displacement) and this new characterization. In addition, from the calculated statistics (the variance in particular), confidence intervals can be computed for the area, which allow for a more accurate comparison of differences in VSA between individuals.

There are two principal contributions in this study. First, a mathematical analysis of the VSA yields a closed-form expression for its moment generating function and, as a result, all of its moments. Second, the formulae are validated through a series of numerical experiments on two speech databases.[8,9] The speech in the databases is processed and formants extracted for each of the corner vowels. From this, values of the sample mean, variance, skewness, and kurtosis are compared against the derived mathematical expressions of the same.

II. METHODS

The main theoretical result of this paper is a closed form mathematical expression for the moment generating function of the area of a quadrilateral that defines a person's vowel space. Before stating the main result, the notation is defined, a new form for the area of an arbitrary quadrilateral is derived, and necessary assumptions are outlined.

A. Notation

In the rest of this paper, a random variable is notated by a capital letter (e.g., F_1). A specific draw from a random variable is notated by a lowercase letter (f_1). Operations on random variables result in new random variables (e.g., for $T_1 = F_{\mathrm{u}}^1 - F_{\mathrm{ae}}^1$, T_1 is a new random variable). Vectors and matrices are notated by lowercase and uppercase, boldface variables, respectively (e.g., $\boldsymbol{\mu}_{\mathrm{u}}$, $\mathbf{C}_{\mathrm{u}}$). Indexing on vectors and matrices is notated by a parenthetic superscript index (e.g., $\boldsymbol{\mu}_{\mathrm{u}}^{(1)}$, $\mathbf{C}_{\mathrm{u}}^{(1,1)}$).

To write the area of the quadrilateral in Fig. 1 in closed form, a series of random variables must be defined. Let $(F_{\mathrm{u}}^1, F_{\mathrm{u}}^2) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{u}}, \mathbf{C}_{\mathrm{u}})$, $(F_{\mathrm{a}}^1, F_{\mathrm{a}}^2) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{a}}, \mathbf{C}_{\mathrm{a}})$, $(F_{\mathrm{ae}}^1, F_{\mathrm{ae}}^2) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{ae}}, \mathbf{C}_{\mathrm{ae}})$, and $(F_{\mathrm{iy}}^1, F_{\mathrm{iy}}^2) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{iy}}, \mathbf{C}_{\mathrm{iy}})$ denote the formant pairs (and their respective distributions) for each of the four vowels shown in Fig. 1. The following auxiliary random variables and distributions are required for the analysis in ensuing sections:

$$T_1 = F_{\mathrm{u}}^1 - F_{\mathrm{ae}}^1, \qquad T_1 \sim \mathcal{N}(\mu_1, \sigma_1^2),$$
$$T_2 = F_{\mathrm{a}}^2 - F_{\mathrm{iy}}^2, \qquad T_2 \sim \mathcal{N}(\mu_2, \sigma_2^2),$$
$$T_3 = F_{\mathrm{a}}^1 - F_{\mathrm{iy}}^1, \qquad T_3 \sim \mathcal{N}(\mu_3, \sigma_3^2),$$
$$T_4 = F_{\mathrm{u}}^2 - F_{\mathrm{ae}}^2, \qquad T_4 \sim \mathcal{N}(\mu_4, \sigma_4^2),$$

where $\mu_1 = \boldsymbol{\mu}_{\mathrm{u}}^{(1)} - \boldsymbol{\mu}_{\mathrm{ae}}^{(1)}$, $\mu_2 = \boldsymbol{\mu}_{\mathrm{a}}^{(2)} - \boldsymbol{\mu}_{\mathrm{iy}}^{(2)}$, $\mu_3 = \boldsymbol{\mu}_{\mathrm{a}}^{(1)} - \boldsymbol{\mu}_{\mathrm{iy}}^{(1)}$, $\mu_4 = \boldsymbol{\mu}_{\mathrm{u}}^{(2)} - \boldsymbol{\mu}_{\mathrm{ae}}^{(2)}$, $\sigma_1^2 = \mathbf{C}_{\mathrm{u}}^{(1,1)} + \mathbf{C}_{\mathrm{ae}}^{(1,1)}$, $\sigma_2^2 = \mathbf{C}_{\mathrm{a}}^{(2,2)} + \mathbf{C}_{\mathrm{iy}}^{(2,2)}$, $\sigma_3^2 = \mathbf{C}_{\mathrm{a}}^{(1,1)} + \mathbf{C}_{\mathrm{iy}}^{(1,1)}$, and $\sigma_4^2 = \mathbf{C}_{\mathrm{u}}^{(2,2)} + \mathbf{C}_{\mathrm{ae}}^{(2,2)}$. Because it is assumed that each formant pair for the corner vowels is drawn from a jointly Gaussian distribution, the distributions of T_i result directly from the fact that the difference of Gaussian random variables is also Gaussian.[9]

B. The area of a non-crossing quadrilateral

In Fig. 2, we show an arbitrary, non-crossing quadrilateral with endpoints drawn from the distributions $(F_{\mathrm{u}}^1, F_{\mathrm{u}}^2)$, $(F_{\mathrm{iy}}^1, F_{\mathrm{iy}}^2)$, $(F_{\mathrm{ae}}^1, F_{\mathrm{ae}}^2)$, and $(F_{\mathrm{a}}^1, F_{\mathrm{a}}^2)$ previously defined. The area of this quadrilateral can be split into two triangular regions with areas A_1 and A_2, as shown in the figure. We define the vectors $\vec{v}_1$, $\vec{v}_2$, and $\vec{v}_3$ from the endpoints of the quadrilateral as follows:

$$\vec{v}_1 = \langle f_{\mathrm{iy}}^1 - f_{\mathrm{u}}^1, f_{\mathrm{iy}}^2 - f_{\mathrm{u}}^2 \rangle, \tag{1}$$
$$\vec{v}_2 = \langle f_{\mathrm{ae}}^1 - f_{\mathrm{u}}^1, f_{\mathrm{ae}}^2 - f_{\mathrm{u}}^2 \rangle, \tag{2}$$
$$\vec{v}_3 = \langle f_{\mathrm{a}}^1 - f_{\mathrm{u}}^1, f_{\mathrm{a}}^2 - f_{\mathrm{u}}^2 \rangle. \tag{3}$$

Using vector notation, the areas of the two triangles are

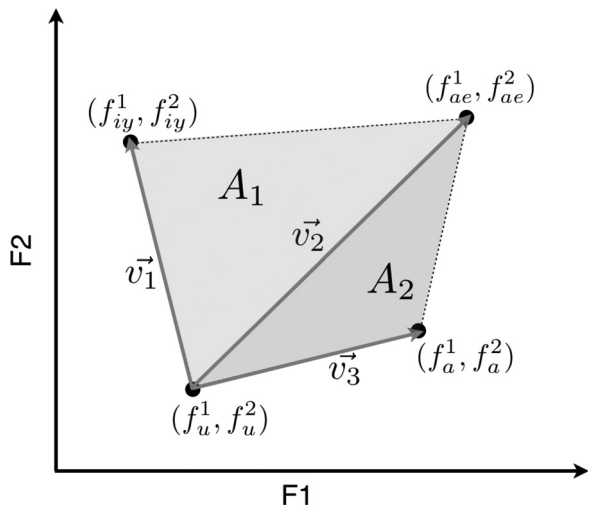

FIG. 2. A notional quadrilateral formed from one instance of each of the four corner vowels. The quadrilateral vowel space area, A, can be decomposed into the areas of two triangles, A_1 and A_2.

$$A_1 = \frac{1}{2}|\vec{v}_2 \times \vec{v}_1|, \tag{4}$$

$$A_2 = \frac{1}{2}|\vec{v}_3 \times \vec{v}_2|. \tag{5}$$

The total area of the quadrilateral is

$$A = A_1 + A_2 = \frac{1}{2}\vec{v}_2 \times \vec{v}_1 + \frac{1}{2}\vec{v}_3 \times \vec{v}_2. \tag{6}$$

Under the assumption that the quadrilateral is non-crossing, the quantities defined inside the vector norm are positive. As a result, the norms are removed, each vector replaced by the definitions in Eqs. (1) to (3), and the expression is simplified. The resulting area of a quadrilateral the vertices of which are the four points in Fig. 2 is given by

$$A = \frac{1}{2}T_1T_2 - \frac{1}{2}T_3T_4,$$

where $T_1, \ldots, T_4$ are defined in the previous section.

C. Assumptions

The goal of the rest of the paper is to completely characterize the distribution of A. Two key assumptions are made in this analysis: (1) The distribution of the formant values can be modeled by a jointly Normal distribution and (2) The random variable pairs $(T_1,\ T_2)$, $(T_3,\ T_4)$, and $(T_1T_2,\ T_3T_4)$ are independent.

Single Gaussian models and Gaussian mixture models have been used to successfully model formant distributions in the literature.[11,12] These distributions have been used in forensic speech analysis and have proved to be an adequate representation in that field. In the results, this is confirmed in the present application by comparing theoretical values of higher-order statistics with those empirically estimated from the same data set. The results show that the closed-form statistics derived using the Gaussian assumption match well with the empirical estimates of the same values.

To confirm the validity of the independence assumption, it is shown that the random variable pairs have low correlation coefficients. In general, random variables can be uncorrelated but dependent; however, for jointly Gaussian random variables, the components that are uncorrelated are independent. The independence assumption is empirically confirmed using data from the phonetically segmented TIMIT database.[9] TIMIT contains speech from 630 speakers from eight dialect regions. For each dialect region, formant pairs for each of the corner vowels are extracted and the correlation coefficient between the three pairs of random variables is calculated. The values are shown in Table I. Details on how the formants were extracted can be found in Sec. III. As the table shows, the correlation coefficient between the random variable pairs is low. For multivariate, normally distributed data, any two or more of its components that are uncorrelated are also independent. As such, the low values of the correlation coefficient, combined with the assumption of joint normality, implies the independence assumption is reasonable for the first two variable pairs in this representative data set. The correlation coefficient of the third variable pair is larger (and not normally distributed); however, in Sec. III, it is demonstrated that closed-form statistics are still able to follow empirical estimates of the same.

TABLE I. Correlation coefficients for random variable pairs.

RV pairs	Correlation coefficients
(T_1, T_2)	−0.0020
(T_3, T_4)	−0.0028
$(T_1T_2,\ T_3T_4)$	0.1327

D. Analytic expression for the moment generating function of the vowel space area

The main theoretical result of this paper is a closed form expression for the moment generating function of A, denoted by $M_A(s)$. From this, the central and non-central moments of the area are derived. From the preceding information, the area of the non-crossing quadrilateral is given by

$$A = \frac{1}{2}T_1T_2 - \frac{1}{2}T_3T_4. \tag{7}$$

If we denote $T_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $T_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $Z = T_1T_2$, then the moment generating function of Z, $M_Z(s)$, is given by

$$M_z(s) = E[e^{T_1T_2s}] \tag{8}$$

$$= \frac{1}{\sigma_1\sigma_2 2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(T_1-\mu_1)^2/2\sigma_1^2} e^{-(T_2-\mu_2)^2/2\sigma_2^2} \times e^{T_1T_2s}\mathrm{d}T_1\mathrm{d}T_2. \tag{9}$$

The integral in Eq. (9) can be solved in closed form, and the resulting MGF of Z is

$$M_Z(s) = \frac{1}{\sqrt{1-\sigma_1^2\sigma_2^2s^2}} \times \exp\left(\mu_1\mu_2 s + \frac{1}{2}\mu_1^2\sigma_2^2s^2 + \frac{1}{2}\mu_2^2\sigma_1^2s^2 \Big/ 1-\sigma_1^2\sigma_2^2s^2\right). \tag{10}$$

The closed form solution to the integral in Eq. (9) is derived in the Appendix.

Using the intermediate result in Eq. (10), the MGF of the area in Eq. (7) can be derived. The moment generating function of A is

$$M_A(s) = E[e^{[(1/2)T_1T_2 - (1/2)T_3T_4]s}] \tag{11}$$

$$= E[e^{T_1T_2s/2}]E[e^{-T_3T_4s/2}] \tag{12}$$

$$= M_{T_1T_2}\left(\frac{s}{2}\right)M_{T_3T_4}\left(\frac{-s}{2}\right), \tag{13}$$

where the expectations are split because of the independence assumption. Substituting the intermediate result of Eq. (10) in Eq. (13) and simplifying, yields the moment generating function of the area, $M_A(s)$:

$$M_A(s) = E[e^{T_1T_2s/2}]E[e^{-T_3T_4s/2}] = \frac{\exp\left(\frac{(\mu_1\mu_2 s/2 + \mu_1^2\sigma_2^2 s^2/8 + \mu_2^2\sigma_1^2 s^2/8)}{1 - s^2\sigma_1^2\sigma_2^2/4} + \frac{(-\mu_3\mu_4/2 + \mu_3^2\sigma_4^2 s^2/8 + \mu_4^2\sigma_3^2 s^2/8)}{1 - s^2\sigma_3^2\sigma_4^2/4}\right)}{\sqrt{1 - s^2\sigma_1^2\sigma_2^2/4}\sqrt{1 - s^2\sigma_3^2\sigma_4^2/4}}. \quad (14)$$

In the literature, oftentimes, only the average vowel space area is calculated empirically from the formant measurements. The closed form expression of the MGF, $M_A(s)$, allows us to derive expressions for other statistics. The non-central moments can be calculated directly from the moment generating function. To calculate the nth moment of the area, we use

$$E[A^n] = \left.\frac{d^n}{ds^n}\right|_{s=0} M_A(s). \quad (15)$$

Using Eq. (15), expressions for the central moments are calculated, using the definitions in Papoulis.[10] In particular, the mean, μ_A, and variance, σ_A^2, of the distribution of the area are given by

$$\mu_A = \left.\frac{d}{ds}\right|_{s=0} M_A(s) = \frac{\mu_1\mu_2}{2} - \frac{\mu_3\mu_4}{2}, \quad (16)$$

$$\sigma_A^2 = \left.\frac{d^2}{ds^2}\right|_{s=0} M_A(s) - \mu_A^2 = \frac{\mu_2^2\sigma_1^2}{4} + \frac{\mu_1^2\sigma_2^2}{4} + \frac{\sigma_1^2\sigma_2^2}{4} + \frac{\mu_4^2\sigma_3^2}{4} + \frac{\mu_3^2\sigma_4^2}{4} + \frac{\sigma_3^2\sigma_4^2}{4}. \quad (17)$$

The closed-form expressions for the higher-order statistics (e.g., skewness and kurtosis) calculated using Eq. (15) are omitted from the paper because of space constraints; however, it is shown in the next section that these expressions match well with empirical estimates of the same values.

III. NUMERICAL RESULTS AND DISCUSSION

The validity of the derived results are confirmed by comparing the theoretical, closed-form vowel space area statistics against empirical estimates of the same on two data sets—the Hillenbrand[8] data and the TIMIT[9] data.

A. Hillenbrand data

The Hillenbrand data are used to assess the validity of the newly derived analytic expressions. In the Hillenbrand study, speech samples were collected from speakers consisting of 45 men, 48 women, and 46 ten- to 12-yr-olds (27 boys, 19 girls). Eighty-seven percent of speakers were raised in Michigan's lower peninsula, primarily in the southeastern and southwestern parts of the state. After a screening process, audio recordings were taken of the 12 English vowels in /hVd/ syllables, then low-pass filtered at 7.2 kHz, and digitized at 16 kHz. Measurements were made of vowel duration, F0 contour, and formant frequency contours for all of the 1668 utterances. Vowel start and end times were obtained by hand using high resolution spectrographs by two experimenters. Formant frequencies were obtained by calculation of 14-pole, 128-point linear predictive coding (LPC) spectra with 16 ms (256-point) hamming windowed frames. Spectral peaks were estimated using three-point parabolic interpolation of the LPC spectrum. F0 contours were extracted using an autocorrelation pitch tracker.

For each of the four corner vowels in the Hillenbrand data set,[8] a bivariate Gaussian distribution is fit to the first and second formant. The covariance ellipse associated with these distributions is shown in Fig. 1. In an effort to evaluate the derived statistics on a number of underlying distributions, a scaling coefficient, α, is introduced to generate new distributions by modifying those learned from the Hillenbrand data. This parameter helps generate a set of underlying distributions with varying mean and variance to validate the expressions for statistics derived by Eq. (15). This is done by scaling the mean vector and covariance matrix of each corner vowel by $\alpha \in [0, 0.1, 0.2, \ldots, 1]$. For each value of α, 100 000 sets of four corner vowels are drawn from the resulting distribution, and the VSA for each set is empirically calculated.

The theoretical values of the same parameters are calculated by making use of Eq. (15) and the formulas for the sample skewness and kurtosis in Papoulis.[10] The results are overlaid in Fig. 3. As is apparent from the figure, there is very good agreement between the empirical and theoretical estimates. As the order of the statistic increases, the agreement between the theoretical and empirical estimates decreases (in particular the kurtosis estimate). The principal reason for this is that the non-zero correlation between T_1T_2 and T_3T_4 (see Table I) becomes more important for the higher-order terms, resulting in a slight difference between the empirical and theoretical estimates of kurtosis. Nonetheless the theoretical estimates capture the general trend in the data even for the kurtosis, where the agreement is not exact.

B. TIMIT data

In addition to the academic example using the Hillenbrand data, the validity of the theoretical estimates are further assessed on the TIMIT data set.[9] For each dialect region (DR) in TIMIT, all instances of the corner vowels are extracted using the meta-information in TIMIT, which provides phonetic segmentation. A PRAAT (Ref. 13) script is used to automatically extract the first and second formant at the midpoint of each vowel instantiation. The PRAAT formant extraction algorithm works by resampling the speech signal to a frequency of twice the maximum formant frequency (a user-defined parameter in the algorithm). Follow this, a pre-emphasis filter is applied, the signal is windowed with a Gaussian window, and the LPC spectrum is estimated. The

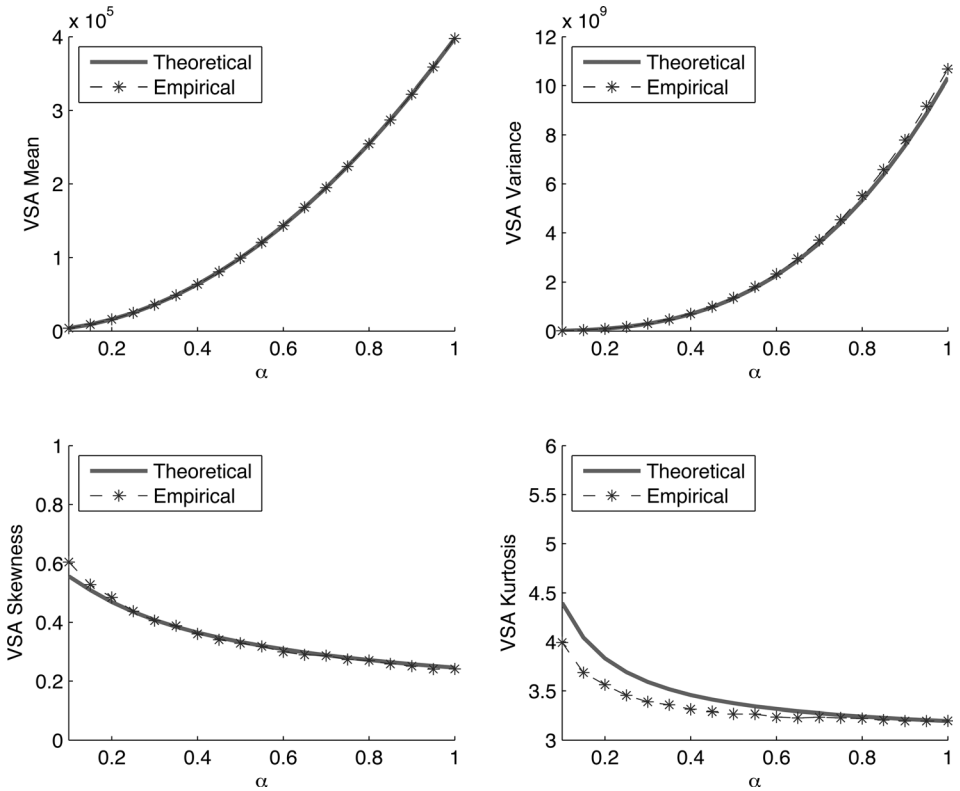

FIG. 3. Empirical and theoretical estimates of the vowel space area mean (top left), variance (top right), skewness (bottom left), and kurtosis (bottom right) computed for the distributions learned from the Hillenbrand (Ref. 8) data.

peaks in the result spectrum are used as estimates of the formant frequency.

The resulting formants are filtered such that only those within 3σ of each formant's mean are kept. The 3σ threshold was determined by visually inspecting the formants to ensure that outliers arising from visually obvious errors in the formant extraction algorithm were removed. For those that remain, the vowel space area for each set of four corner vowels is empirically estimated, followed by the mean and variance of the VSA. The theoretical values of the same parameters are calculated using Eqs. (16) and (17). The comparative results are shown in Fig. 4. As the figure shows, there is very good agreement between the empirical and the theoretical results, further confirming the validity of the derived results.

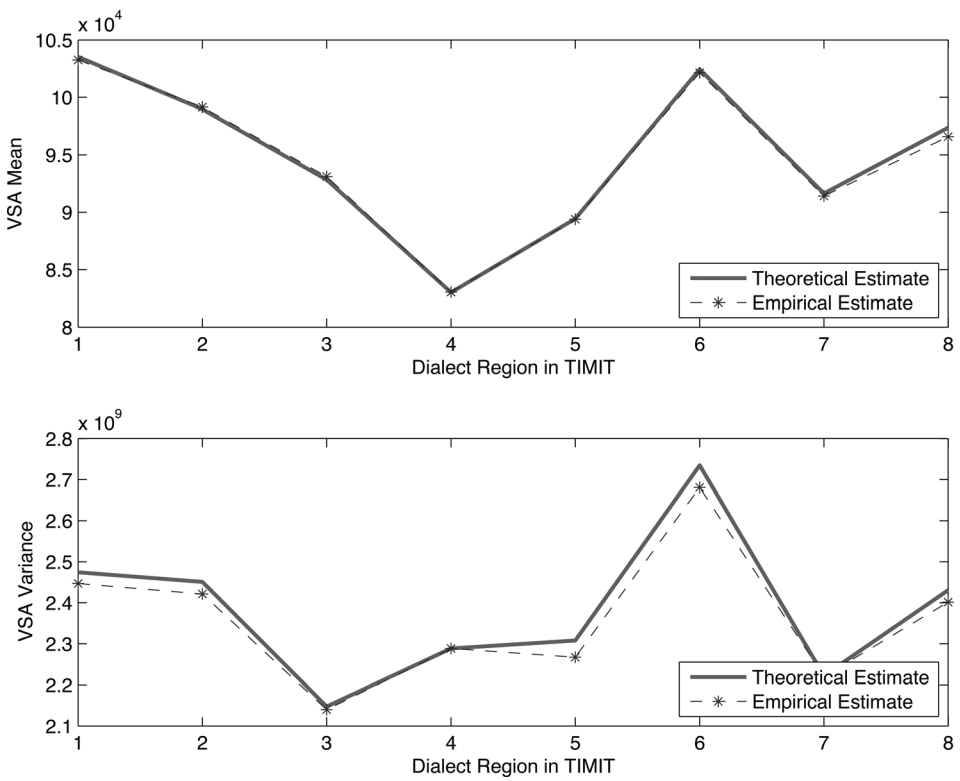

FIG. 4. Empirical and theoretical estimates of the vowel space area mean (top figure) and variance (bottom figure) computed for each dialect region in the TIMIT (Ref. 12) database.

Figure 4 provides further motivation for extending the analysis of the VSA to beyond the mean. As an example, consider the VSA of DR 6. The mean VSA for this dialect is comparatively high. In fact, in a ranking of the eight dialect regions by mean VSA, DR 6 has the second largest vowel area (behind DR 1). Of course, the mean estimate fails to capture the variation in the VSA of DR 6. Analyzing the VSA variance of DR 6, it is noted that it contains significantly more variation than the other dialect regions. With this additional statistic, ranking can be calculated using the inverse of the coefficient of variation (μ_A/σ_A)—a statistic that takes into consideration both the mean and variance of the VSA. Qualitatively, this statistic makes sense because it positively weights a large vowel space area but penalizes dialect regions with large variance in the VSA. This ranking is different and uses a metric that provides a more complete characterization of the VSA. Beyond ranking, from the closed form variance expression, confidence intervals can be computed for the area, which allow for a more accurate comparison of differences in VSA between individuals/groups.

IV. CONCLUSION

The distribution of the vowel space area is characterized under reasonable assumptions. From this, expressions are derived for a series of higher-order statistics, and their accuracy is confirmed using numerical experiments. The newly derived expressions can be used by researchers in the field to better characterize the robustness of the vowel area estimates by measuring not only its mean, but its variance, and potentially third- or fourth-order statistics like skewness and kurtosis. This provides a multi-dimensional, statistical representation of the VSA and, like the mean, these additional quantities (and their combinations) can be correlated against intelligibility to assess their predictive power. The higher-order statistics capture information about the shape of the distribution of the VSA by modeling the articulatory kinematics in a non-deterministic manner. In addition, a closed form expression for the variance means that we can define confidence intervals for the computed area and allows us to more accurately compare differences in VSA between individuals. Future work involves assessing the newly derived statistics on pathological speech to compare the additional information provided against intelligibility. Additionally, relaxing the independence assumption could be considered in an effort to yield even more accurate estimates of the higher-order statistics.

ACKNOWLEDGMENTS

This research was supported in part by National Institute of Health, National Institute on Deafness and Other Communicative Disorders Grant Nos. 2R01DC006859 (J.L.) and 1R21DC012558 (J.L. and V.B.).

APPENDIX: THE MGF OF THE PRODUCT OF TWO GAUSSIANS

Let $T_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $T_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $Z = T_1T_2$. The moment generating function of Z, $M_Z(s)$, is given by

$$M_Z(s) = \frac{1}{\sqrt{1-\sigma_1^2\sigma_2^2 s^2}} \exp\left[\mu_1\mu_2 s + (1/2)\mu_1^2\sigma_2^2 s^2 + (1/2)\mu_2^2\sigma_1^2 s^2/1 - \sigma_1^2\sigma_2^2 s^2\right].$$

1. Proof

Craig[14] and Ware[15] analyze the product of Gaussian random variables. The current problem is set up in similar fashion. Using the definition of the moment generating function,

$$M_z(s) = E[e^{T_1T_2s}] \tag{A1}$$

$$= \frac{1}{\sigma_1\sigma_2 2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(T_1-\mu_1)^2/2\sigma_1^2} e^{-(T_2-\mu_2)^2/2\sigma_2^2} \times e^{T_1T_2s}\mathrm{d}T_1\mathrm{d}T_2 \tag{A2}$$

$$= \frac{1}{\sqrt{2\pi\sigma_2^2}}\int_{-\infty}^{\infty} e^{-(T_2-\mu_2)^2/2\sigma_2^2} \times \left[\frac{1}{\sqrt{2\pi\sigma_1^2}}\int_{-\infty}^{\infty} e^{-(T_1-\mu_1)^2/2\sigma_1^2} e^{T_1T_2s}\mathrm{d}T_1\right]\mathrm{d}T_2. \tag{A3}$$

It is noted that that the quantity in the parenthesis is $M_{T_1}(T_2s)$, the MGF of T_1 evaluated at T_2s,

$$M_z(s) = \frac{1}{\sqrt{2\pi\sigma_2}}\int_{-\infty}^{\infty} e^{-(T_2-\mu_2)^2/2\sigma_2^2} e^{\mu_1T_2s} e^{-\sigma_1^2T_2^2s^2/2}\mathrm{d}T_2. \tag{A4}$$

The quadratic polynomial in the exponent is expanded (w.r.t. T_2) and like terms are combined. Letting $a = 1 - \sigma_1^2\sigma_2^2 s^2/2\sigma_2^2$ and $b = \mu_2 + \mu_1\sigma_2^2 s^2/\sigma_2^2$, the square in the exponent is completed by adding and subtracting b^2/a from the exponent to obtain

$$M_z(s) = \frac{e^{-\mu_2^2/2\sigma_2^2}}{\sigma_2\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{aT_2^2+bT_2+(b^2/4a)-(b^2/4a)}\mathrm{d}T_2 \tag{A5}$$

$$= \frac{e^{(-\mu_2^2/2\sigma_2^2)}e^{(b^2/4a)}}{\sigma_2\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-[T_2-(b/2a)]^2/1/a}\,\mathrm{d}T_2 \tag{A6}$$

$$= \frac{e^{-\mu_2^2/2\sigma_2^2}e^{b^2/4a}}{\sigma_2\sqrt{2\pi}}\frac{\sqrt{2\pi}}{\sqrt{2a}}\int_{-\infty}^{\infty}\frac{\sqrt{2a}}{\sqrt{2\pi}} e^{-[T_2-(b/2a)]^2/1/a}\mathrm{d}T_2 \tag{A7}$$

$$= \frac{e^{(-\mu_2^2/2\sigma_2^2)+(b^2/4a)}}{\sigma_2\sqrt{2a}}. \tag{A8}$$

Going from Eq. (A7) to (A8), it is seen that the integrand as $\mathcal{N}(b/2a, 1/2a)$, therefore integrating over $\mathbb{R}$ equals 1. Substituting for a and b and simplifying, the final MGF is obtained,

$$M_Z(s) = \frac{1}{\sqrt{1-\sigma_1^2\sigma_2^2 s^2}}\exp\left[\left(\mu_1\mu_2 s + \frac{1}{2}\mu_1^2\sigma_2^2 s^2 + \frac{1}{2}\mu_2^2\sigma_1^2 s^2\right)\Big/ 1 - \sigma_1^2\sigma_2^2 s^2\right]. \tag{A9}$$

[1]P. Flipsen and S. Lee, "Reference data for the American English acoustic vowel space," Clin. Linguist. Phonet. **26**(11–12), 926–933 (2012).
[2]H. K. Vorperian and R. D. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," J. Speech, Lang. Hear. Res. **50**(6), 1510–1545 (2007).
[3]S. Skodda, W. Grönheit, and U. Schlegel, "Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease," PLoS One **7**(2), e32132 (2012).
[4]L. B. Leonard, S. Ellis Weismer, C. A. Miller, D. J. Francis, J. B. Tomblin, and R. V. Kail, "Speed of processing, working memory, and language impairment in children," J. Speech, Lang. Hear. Res. **50**(2), 408–428 (2007).
[5]S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," J. Speech, Lang. Hear. Res. **53**(1), 114–125 (2010).
[6]E. Jacewicz and R. A. Fox, "Dialectal and age-related acoustic variation in vowels in spontaneous speech," J. Acoust. Soc. Am. **132**(3), 2002 (2012).
[7]J. Lam, K. Tjaden, and G. Wilding, "Acoustics of clear speech: Effect of instruction," J. Speech, Lang. Hear. Res. **55**(6), 1807–1821 (2012).
[8]J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995).
[9]J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," CDROM, 1993.
[10]A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes* (McGraw-Hill, New York, 2002), 852 p.
[11]T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 1505–1508.
[12]A. Moos, "Long-term formant distribution," Master's thesis, Universitat des Saarlandes, Saarbrcken, Germany, 2008, 92 pp.
[13]P. Boersma, "PRAAT, a system for doing phonetics by computer," Glot Int. **5**(9/10), 341–345 (2001).
[14]C. C. Craig, "On the frequency function of xy," Ann. Math. Stat. **7**(1), 1–15 (1936).
[15]R. Ware and F. Lad, "Approximating the distribution for sums of products of normal variables," Technical Report UCDMS 2003/15, University of Canterbury (2003).