Automatic assessment of vowel space area

Steven Sandoval
School of Electrical, Computer and Energy Engineering, Sensor, Signal & Information Processing Center, Arizona State University, Tempe, Arizona 85287
spsandov@asu.edu

Visar Berisha, Rene L. Utianski, and Julie M. Liss
Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287
visar@asu.edu, rutiansk@asu.edu, julie.liss@asu.edu

Andreas Spanias
School of Electrical, Computer and Energy Engineering, Sensor, Signal & Information Processing Center, Arizona State University, Tempe, Arizona 85287
spanias@asu.edu

Abstract: Vowel space area (VSA) is an attractive metric for the study of speech production deficits and reductions in intelligibility, in addition to the traditional study of vowel distinctiveness. Traditional VSA estimates are not currently sufficiently sensitive to map to production deficits. The present report describes an automated algorithm using healthy, connected speech rather than single syllables and estimates the entire vowel working space rather than corner vowels. Analyses reveal a strong correlation between the traditional VSA and automated estimates. When the two methods diverge, the automated method seems to provide a more accurate area since it accounts for all vowels.

PACS numbers: 43.72.Ar, 43.71.Gv, 43.71.Sy [DDO]
Date Received: July 22, 2013 **Date Accepted:** September 26, 2013

1. Introduction

Vowel space area (VSA) refers to the two-dimensional area bounded by lines connecting first and second formant frequency coordinates ($F1/F2$) of vowels.[1] Estimation of VSA has a long history in the study of vowel identity, speaker characteristics, speech development, speaking style and sociolinguistic factors that influence vowel production.[2–10] Traditional VSA computation methodology is shown in Fig. 1(a). A typical computation involves making static measurements of the $F1/F2$ values for each of the four corner vowels (or three point vowels, /a, i, u/ for triangle) at 50% vowel duration, for several productions of each vowel. The mean $F1/F2$ value for each of the four corner vowels is then used to compute the area of the quadrilateral formed by the corner vowels. Since 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 ($F2$), the VSA is an acoustic proxy for the kinematic displacements of the articulators.[11] In general, studies have shown that VSA is larger in speech that is clearer and more intelligible than speech associated with smaller VSAs.[12] This is interpreted as corresponding to greater articulatory excursions and more distinct acoustic-articulatory vowel targets. Thus, the VSA and other derived vowel metrics related to distinctiveness have been quite successful in the study of speaking style, dialects, and languages.[6,7,9]

Because abnormal vowel formant reduction (centralization) is a common feature of speech production deficits, there has been a longstanding interest in using VSA estimations for characterizing speech motor control, including speech development,[10,13] speech disorders,[14–17] and speech interventions.[18] Despite the intuitive appeal of using

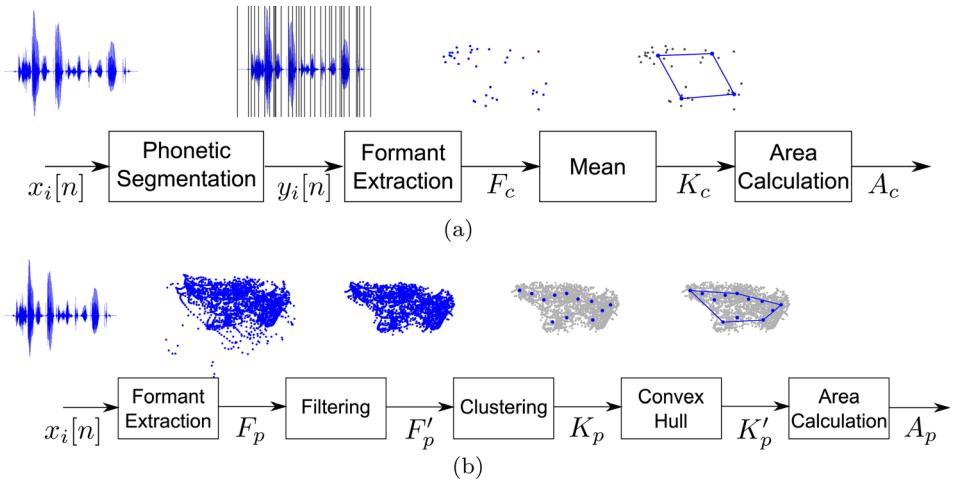

Fig. 1. (Color online) Block diagrams for (a) the typical steps taken in the manual computation of the vowel space area. Speech samples are phonetically segmented, formants for the corner vowels are estimated, the mean value of each corner vowel is computed, and finally the area bounded by the mean of corner vowels is computed. (b) The proposed method.

VSA as an index of speech motor control and intelligibility, its success has been limited and modest.[19] For instance, VSA was minimally predictive of overall intelligibility for individuals with dysarthria, secondary to Parkinson's disease and multiple sclerosis (between 6 and 13%).[20,21] More optimistic relationships (over 40%) were reported when examining the same relationship for speakers with dysarthria, secondary to amyotrophic lateral sclerosis (ALS).[22,23] The most promising predictive relationship of VSA and intelligibility was demonstrated by Higgins and Hodge,[24] in an assessment of a heterogeneous sample of children with dysarthria. Attempts to modify the VSA estimate to more sensitively account for differences in the front-back and high-low dimensions have offered some benefit.[25] Such modifications may be preferable for mapping VSA to perceptual measures and speaker classification.[14,26–28] However, it is likely that more extensive modifications are required to obtain VSA estimates that hold clinical utility for speech production deficits and the resulting decrements in speech intelligibility. Such information, particularly if fully automated and robust to speech sample, would provide an important objective assessment to augment and support clinical practice.

There are several significant limitations associated with existing VSA estimates in the context of speech production disorders. The first two limitations are that VSA calculations are based only on point (triangle) or corner (quadrilateral) vowels, rather than all vowels; and these vowels are produced in isolation (typically hVd). This methodology was borrowed from the study of vowel production in healthy speech to examine vowel distinctiveness as described above.[8] This makes good sense from the standpoint of defining the most disparate regions of the vowel space (and, by extension, the maximal articulatory excursions) in a way that is free of extraneous coarticulatory influences. However, previous research using VSA estimations on disordered speech has not shown the ability to robustly elicit and/or capture speech production deficits and intelligibility in a clinically meaningful way. There is every reason to believe that when the VSA is globally reduced, as in speech production disorders, more sensitive methodology is required. One possibility is to sample the entire articulatory working space, and characterize its shape, to fully account for the extent of articulatory displacements and their acoustic consequences. It also may be useful to extract vowel formant information from productions in connected speech rather than single word productions to magnify the impact of the underlying movement disorder. Finally, the third, and perhaps most important limitation from an applied stand-point, is that the traditional VSA estimation process is cumbersome, requiring phonetic segmentation of input speech.

In an effort to overcome these limitations and move closer to a clinical tool, the present report describes a novel alternative for VSA estimation that (1) is fully automated, (2) can be collected from any length or variety of speech material that contains a range of vowels, and (3) considers all vowels produced rather than estimating the shape of the VSA with a triangle or quadrilateral. The algorithm relies on a series of automated tools for extracting all formants from voiced sections of speech, thereby removing the need for hand segmentation. This is followed by a clustering and area calculation algorithm based on the convex hull of the cluster centers to estimate the final VSA. The proposed algorithm is applied to healthy speech and then compared against an estimate of the vowel space quadrilateral area formed from hand-segmented speech of the same sample.[29] Results show that the automated estimate exhibits a strong correlation with the hand-segmented estimate, and often yields a more accurate estimate of the VSA.

2. Methods

Figure 1(b) shows a block diagram of the proposed method for the automated estimation of the VSA. The algorithm can operate on any incoming speech signal that contains a range of vowels. The signal is analyzed on a frame-by-frame basis and, for each voiced frame, the first and second formants are estimated. Following, outliers are removed and the remaining points are clustered. The convex hull of the cluster centers is determined and the area of the resulting convex hull is calculated. In the following sections the details of each of the required steps is discussed.

2.1 Formant extraction

A PRAAT script[30] is used to automatically extract all $F1/F2$ pairs corresponding to voiced frames. The PRAAT script assesses voicing on a frame-by-frame basis by estimating periodicity using an autocorrelation-based method. In this study we only consider the first two formants, however, using the recommended PRAAT values, five formants were extracted per frame below a ceiling value (5000 male, 5500 female) in Hz. Other settings were as follows: 1 ms frame advance; 50 ms analysis window; pre-emphasis starting from 50 Hz. Internally, PRAAT computes estimates of the formants by resampling to twice the ceiling of the formant search range, then applying a pre-emphasis filter, windowing the speech in the time domain using a Gaussian window, and estimating the LPC coefficients using the algorithm by Burg.[31,32] Processing all input speech results in an $N \times 2$ matrix, F_p, that stores all $F1/F2$ pairs for a particular speaker, where N is the number of formant observations for a particular speaker.

2.2 Filtering

Automated formant estimation algorithms can result in outliers. In order to identify the extrema, the probability distribution of each speaker's formants, F_p, is modeled using a Gaussian mixture model (GMM) and low-likelihood points are identified and removed. The use of GMMs is common in speech processing applications.[33] The weight, mean, and (full) covariance matrix for each of the four component densities in the Gaussian mixture are learned using the expectation maximization (EM) algorithm. For each formant in F_p, the log-likelihood is calculated and components with a likelihood less than $0.3\overline{L(F_p)}$ are identified as outliers and removed from downstream processing. $\overline{L(F_p)}$ denotes the mean likelihood of all observations in F_p. The filtered parameter set is denoted by F_p'. The outlier filtering rejected approximately 15% of the total number of formant observations for a particular speaker.

2.3 Clustering

Following outlier rejection, the remaining points F_p' are clustered using the k-means algo-rithm.[34] Twelve cluster centers (one corresponding to each of the 12 English vowels) were initialized using the mean $F1/F2$ values as reported by Hillenbrand[5] at 50%

vowel duration. The cluster centers were initialized for adult males and females, using the respective reported values, and returned values are denoted by K_p.

2.4 Convex hulllarea calculation

Using the Quick-hull[35] algorithm in MATLAB,[36] the convex hull of the set of points in K_p is found. The clockwise ordered endpoints (beginning and ending with the same point) of the resulting convex polygon is denoted by K'_p. The area of the polygon with m corners is then given, with slight abuse of the determinant notation, by

$$A_p = \frac{1}{2}|K'_p| = \frac{1}{2}\begin{vmatrix} F1_1 & F2_1 \\ F1_2 & F2_2 \\ \vdots & \vdots \\ F1_m & F2_m \\ F1_1 & F2_1 \end{vmatrix} = \frac{1}{2}\sum_{i=1}^{m}(F1_iF2_{i+1} - F2_iF1_{i+1}). \quad (1)$$

2.5 Stimuli

Speech samples were drawn from the TIMIT corpus commissioned by DARPA.[30] The TIMIT corpus consists of 6300 sentences, 10 sentences spoken by 630 speakers from 1 of 8 major dialect regions[37] of the United States. The TIMIT corpus includes hand verified, time-aligned orthographic, phonetic, and word transcriptions as well as 16-bit, 16 kHz speech waveform files for each utterance. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI), and Texas Instruments, Inc. (TI). The speech material consists of phonetically diverse sentences intended to expose dialectal variants of the speech.

3. Results and discussion

The output of the automated metric is compared to a traditional VSA metric computed from hand-segmented speech, including several derivations of the Pearson correlation coefficient.

3.1 Performance analysis

In order to assess the performance of the proposed method several comparisons were made between the proposed method and a control method. The control method uses the traditional VSA computation paradigm by utilizing the meta-data provided with

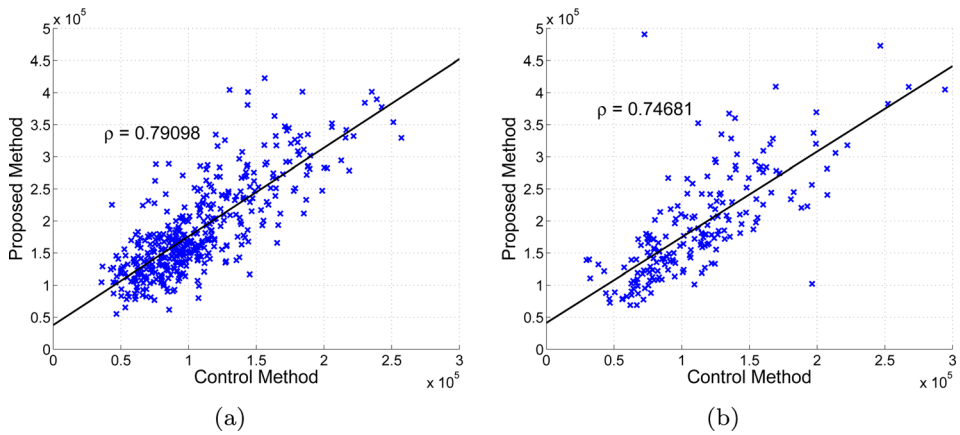

Fig. 2. (Color online) A scatter plot showing the estimated VSA obtained using the proposed and control methods for each of the 630 speakers in the TIMIT corpus for (a) male speakers, (b) female speakers. Male and female speakers yield correlation coefficients of $\rho = 0.790980$ and 0.74681, respectively. The proposed method yields a correlation coefficient of $\rho = 0.77553$ over all speakers.

Table 1. Correlation between the proposed and control methods.

Case	By speaker	By dialect region
Male	0.79098	0.50937
Female	0.74681	0.52836
All	0.77553	0.60118

the TIMIT corpus. More specifically, for each occurrence of the corner vowels, estimates of the means of the formant frequencies are calculated and the area of the resulting quadrilateral is computed. An estimate of the VSA for each of the 630 speakers utilizing all ten sentences per speaker ($x_i[n]$, $i=1,\ldots,10$) for both the proposed and control methods were computed. When divided by sex, male and female speakers yield correlation coefficients of $\rho=0.790980$ and 0.74681, respectively. The proposed method yields a correlation coefficient of $\rho=0.77553$ when computed over all 630 speakers. A scatter plot of the data is show in Fig. 2 and the results are summarized in Table 1.

Similar analyses comparing estimates of the VSA corresponding to an entire dialect region were performed. When estimating the VSA for the eight dialect regions by sex, estimates yield a correlation coefficients of $\rho=0.50937$ and 0.52836, for male and female speakers, respectively. The proposed method yields a correlation coefficient of $\rho=0.60118$ when estimating the VSA for a dialect region using both male and female speakers. Again, the results are summarized in Table 1.

Overall the proposed method has high correlation to the control method. However, the proposed method may actually yield a more accurate result than the conventional method, because the conventional method limits the definition of the vowel space area to the space interior of only four of the twelve English vowels (the corner vowels). In reality, there are many occurrences of $F1/F2$ pairs that occur outside of this space and contribute to the overall shape of the vowel space. This is readily seen in Fig. 3, by comparing the VSA as bounded using the proposed and control methods. The proposed metric results in consistently larger VSA estimates, but also more accurately accounts for the actual shape of the VSA. This may provide a more complete assessment of the contribution of VSA to intelligibility and subsequent decrements.

It is important to note that a key requirement of the algorithm is that the vowel space is adequately sampled. This means that the analyzed content must be phonetically balanced or consistent across individuals for comparison. By design, the TIMIT corpus indeed satisfied this requirement. For clinical applications of this work, clinicians will have the option of specifying the spoken text, ensuring that the incoming speech stream is balanced.

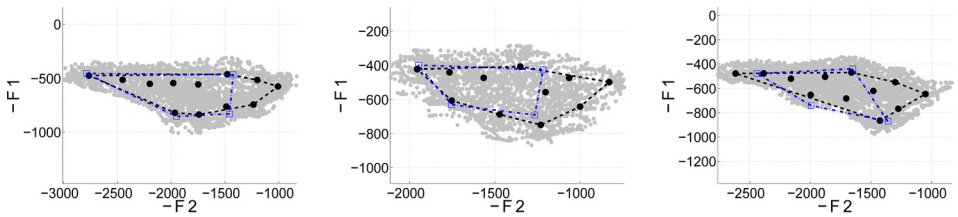

Fig. 3. (Color online) The VSA for three speakers as bounded using the proposed (dashed line) and control (dash-dot line) methods overlaid on the filtered points F_p' (small gray dots). The mean corner vowels K_c (large squares) and the cluster centers K_p (large dots) are also shown. The proposed method better accounts for the actual shape of the VSA. The axes have been chosen so that the plots have the same orientation as the standard IPA vowel trapezium.

4. Conclusion

The assessment of speech intelligibility is the cornerstone of clinical practice in speech-language pathology, as it indexes a patient's communicative handicap. There has been a desire to develop efficient, objective, and reliable measures that can be added to the clinical repertoire. Given the relationship of VSA and intelligibility decrements[20–22,24,38] it is critical to have a sensitive and efficient assessment of VSA; this includes the exploration of a more complete assessment of the vowel space, by including the complete range of vowels in spoken language. In the current investigation, an automated assessment of the VSA demonstrated a strong relationship with the traditional methods of VSA derivation.

Moreover, the proposed method is fully automated and was demonstrated to capture a more complete assessment of the VSA by allowing for arbitrary VSA shapes, rather than only triangle or quadrilateral shaped VSAs. Moving forward, the relationship between the proposed calculation of VSA will be related to intelligibility ratings to understand its relationship with intelligibility decrements. The success with which the automated procedure estimated the VSA along with the ease of computation, makes the proposed an attractive metric for characterizing speech motor control.

Acknowledgments

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

References and links

[1]G. Fant, *Speech Sounds and Features* (MIT Press, Cambridge, 1973).
[2]A. Bladon, "Two-formant models of vowel perception: Shortcomings and enhancement," Speech Commun. **2**(4), 305–313 (1983).
[3]R. L. Diehl, B. Lindblom, K. A. Hoemeke, and R. P. Fahey, "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phonetics **24**(2), 187–208 (1996).
[4]R. A. Fox, "Perceptual structure of monophthongs and diphthongs in English," Lang. Speech **26**(1), 21–60 (1983).
[5]J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111 (1995).
[6]E. Jacewicz and R. A. Fox, "The effects of cross-generational and cross-dialectal variation on vowel identification and classification," J. Acoust. Soc. Am. **131**(2), 1413–1433 (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]G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184 (1952).
[9]C. G. Clopper, D. B. Pisoni, and K. de Jong, "Acoustic characteristics of the vowel systems of six regional varieties of American English," J. Acoust. Soc. Am. **118**(3) 1661–1676 (2005).
[10]P. Flipsen and S. Lee, "Reference data for the American English acoustic vowel space," Clin. Linguist. Phonetics **26**(11-12), 926–933 (2012).
[11]J. Lee and S. Shaiman, "Relationship between articulatory acoustic vowel space and articulatory kinematic vowel space," J. Acoust. Soc. Am. **132**(3), 2003 (2012).
[12]A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**(1), 272–284 (2002).
[13]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).
[14]A. T. Neel, "Vowel space characteristics and vowel identification accuracy," J. Speech Lang. Hear. Res. **51**(3), 574–585 (2008).
[15]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).
[16]L. B. Leonard, S. E. 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 (2007).

[17]H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy," J. Acoust. Soc. Am. **117**, 3879–3889 (2005).
[18]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 (2010).
[19]Y.-I. Bang, K. Min, Y. H. Sohn, and S.-R. Cho, "Acoustic characteristics of vowel sounds in patients with Parkinson disease," J. Speech, Lang. Hear. Res. **47**, 766–783 (2013).
[20]K. Tjaden and G. E. Wilding, "Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings," J. Speech Lang. Hear. Res. **47**(4), 766–783 (2004).
[21]P. A. McRae, K. Tjaden, and B. Schoonings, "Acoustic and perceptual consequences of articulatory rate change in Parkinson disease," J. Speech Lang. Hear. Res. **45**(1), 35–50 (2002).
[22]G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," J. Speech Hear. Res. **38**(5), 1001–1013 (1995).
[23]G. Weismer, J.-Y. Jeng, J. Laures, R. D. Kent, and J. F. Kent, "Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders," Folia Phoniatr. Logop. **53**, 1–18 (2001).
[24]C. M. Higgins and M. M. Hodge, "Vowel area and intelligibility in children with and without dysarthria," J. Med. Speech Lang. Pathol. **10**(4), 271–278 (2002).
[25]S. Sapir, J. L. Spielman, L. O. Ramig, B. H. Story, and C. Fox, "Effects of intensive voice treatment (the Lee Silverman voice treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings," J. Speech Lang. Hear. Res. **50**(4), 899–912 (2007).
[26]K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Speech disorder diagnosis and classification," J. Speech Lang. Hear. Sci., in press.
[27]K. L. Lansford and J. M. Liss, "Vowel acoustics in dysarthria: Mapping to perception," J. Speech Lang. Hear. Sci., in press.
[28]J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," J. Acoust. Soc. Am. **134**, 2171–2181 (2013).
[29]W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proceedings of DARPA Workshop on Speech Recognition* (1986), pp. 93–99.
[30]P. Boersma, "PRAAT, a system for doing phonetics by computer," Glot Int. **5**(9/10), 341–345 (2001).
[31]D. G. Childers, *Modern Spectrum Analysis*, *IEEE Press Selected Reprint Series* (IEEE, New York, 1978).
[32]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, *The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, New York, 1992).
[33]D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995).
[34]J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Neyman (University of California Press, Berkley, 1967), Vol. 1, pp. 281–297.
[35]F. P. Preparata and M. I. Shamos, "Introduction," in *Computational Geometry, Texts and Monographs in Computer Science* (Springer, New York, 1985), pp. 1–35.
[36]MATLAB, version 8.0.0.783 (R2012b), The MathWorks Inc., Natick, Massachusetts, 2012.
[37]*Language Files*, edited by C. J. Colby, R. Wallace, and C. Jolly (Ohio State University Press, Columbus, 1982).
[38]K. Bunton and G. Weismer, "The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria," J. Speech Lang. Hear. Res. **44**(6), 1215–1228 (2001).