Vowel space density as an indicator of speech performance

Brad H. Story[a] **and Kate Bunton**
Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721, USA
bstory@email.arizona.edu, bunton@email.arizona.edu

Abstract: The purpose of this study was to develop a method for visualizing and assessing the characteristics of vowel production by measuring the local density of normalized F_1 and F_2 formant frequencies. The result is a three-dimensional plot called the vowel space density (VSD) and indicates the regions in the vowel space most heavily used by a talker during speech production. The area of a convex hull enclosing the vowel space at specific threshold density values was proposed as a means of quantifying the VSD.

[DDO'S]
Date Received: February 7, 2017 **Date Accepted:** April 22, 2017

1. Introduction

Measurement of the first two formant frequencies in a speech signal, F_1 and F_2, can be plotted as points in a two-dimensional plane. These measurements can be based on a single spectrum at a specific point in time, such as the nucleus of a syllable, or tracked over many consecutive spectra to produce time-dependent, $[F_1(t), F_2(t)]$ trajectories that may reveal information regarding the temporal characteristics of speech segments. In either case, the plotted formant frequencies generate a "vowel space" that can be used to visualize and measure the location of particular vowels relative to others.

Vowel space area (VSA) has been used by many researchers as a means of quantifying the degree of centralization or hyper-articulation of a given talker's speech production (cf. Turner *et al.*, 1995; Weismer *et al.*, 2000; Tjaden *et al.*, 2005). Typically, F_1 and F_2 frequencies are measured in the vowel portion of content words containing the corner vowels /i, ɑ, u/ or /i, æ, ɑ, u/; formants may also be determined from sustained productions of target vowels. In either case, the VSA is the area enclosed by line segments that connect the set of $[F_1, F_2]$ pairs (Kent and Kim, 2003), where the result is given in units of Hz^2. VSA can also be calculated after converting the formant values from frequency to a more psychoacoustically relevant scale. For example, Ferguson and Kewley-Port (2007) transformed all formant measurements to Barks (Syrdal and Gopal, 1986; Traunmüller, 1990) prior to calculating the VSA.

A different approach called the "formant centralization ratio" (FCR) was proposed by Sapir *et al.* (2010). Formant frequencies are measured in the same manner as for VSA calculations, but a ratio of formant sums is calculated such that $\mathrm{FCR} = (F_{2u} + F_{2ɑ} + F_{1i} + F_{1u})/(F_{2i} + F_{1ɑ})$. The idea is that the FCR has greater sensitivity to centralization of the vowels than the VSA, and also that it is less sensitive to variability across speakers.

A disadvantage of both the VSA and FCR approaches is that they rely on making formant measurements from specified content words or target productions at selected instants of time. As a result, these measures represent snapshots of the corner vowels but do not provide a view or assessment of vowel production during any significant duration of actual speaking.

The purpose of this study was to develop a method for visualizing and assessing the characteristics of the $[F_1, F_2]$ vowel space produced by a talker over a long duration (i.e., several minutes or more) of speech production. The method involves first tracking the time-dependent $[F_1(t), F_2(t)]$ values over the voiced segments of the speech signal, and then determining their density across the entire $[F_1, F_2]$ plane to provide a third dimension to the vowel space. Thus, "vowel space density" or VSD can be displayed as either a three-dimensional plot or color map that shows where in the vowel space a talker spent the most amount of time. For example, a talker with centralized vowel production would be expected to produce a VSD whose distribution would tend

[a]Author to whom correspondence should be addressed.

toward the middle of the space, whereas the VSD of hyperarticulated speech may indicate high densities in various locations along the perimeter. The vowel space is then quantified by measuring the area enclosed by a convex hull at a specific threshold value of density. Such area measurements can be used to compare different speaking "performances" of one talker, or similar performances of different talkers.

2. Method

2.1 Audio recordings

For purposes of demonstrating the construction of a VSD plot and its potential use, a recording was obtained of a male talker producing several minutes of speech in three conditions. The first consisted of repetitions of the syllables /hid/, /hæd/, /hɑd/, and /hud/ over the course of approximately 120 s (30 s per each hVd). This highly artificial test case was intended to assure that the talker spent most of the speaking time producing just the corner vowels. In the two other conditions, the talker read from a prepared script that included the "Farm Script" and "Hunter Script" from Crystal and House (1982) and the "Caterpillar Passage" from Patel *et al.* (2013). Each script was first read in a hypoarticulated or mumbled quality, and then a second time with a hyperarticulated quality. Digital recordings for all three speaking conditions were collected with a CSL 4500 (Kay Pentax) in a sound booth using a high-quality microphone (C410, AKG Acoustics, Vienna, Austria), resulting in a collection of audio files with sampling frequency of $f_s = 44\,100$ Hz and 16 bit amplitude resolution. It is noted that, although exactly the same material was read by the talker, the duration of the hypoarticulated speech signal was 90 s shorter than the recording of the hyperarticulated condition. This was due to the nature of the rate of speech production in the two conditions.

2.2 Formant tracking

To generate a VSD plot, formant frequencies must first be tracked over the duration of a speech signal. This can be accomplished with a variety of algorithms or signal analysis packages (e.g., PRAAT, CSpeech). For this study, the formants were tracked with custom analysis algorithms written in MATLAB (Mathworks, Natick, MA, 2016). The analysis proceeded by using a periodicity detector based on autocorrelation to find the time points that bracket all voiced segments in the recorded speech signal (Xie and Niyogi, 2006). Subsequent analyses were limited to these segments. After downsampling the speech signal to a 10 kHz sampling frequency, an autocorrelation LPC algorithm (Mathworks, Natick, MA, 2016) was applied to subsequent 0.04 s windows, tapered at both the left and right sides with a Gaussian ($\alpha = 2.5$) window, to generate an estimate of the vocal tract frequency response. The first two formant frequencies, F_1 and F_2, were estimated with a parabolic peak-picking algorithm (Titze *et al.*, 1987) applied to each frequency response estimate. The analysis windows were overlapped to generate a frequency response, and hence formant values, at 0.005 s increments. The time-varying formant values associated with each voiced segment were processed, independent of all other segments, with a three-point median filter followed by a five-point smoothing filter, and then concatenated to form a two column matrix representing the entire duration of the recording.

2.3 VSD

With n representing the time sample index, the $[F_1(n), F_2(n)]$ pairs measured over the 120 s duration of the hVd recording are shown in Fig. 1(a) as a vowel space plot. The regions with high density of points correspond to the expected locations of the corner vowels. The /hæd/ and /hɑd/ segments do, however, produce lower density "tails" that extend toward the median formant pair $[F_1^{\text{median}}, F_2^{\text{median}}] = [617, 1266]$ Hz shown near the center of the space.

To transform these data into a VSD plot, each $[F_1(n), F_2(n)]$ pair was first normalized by the median value such that

$$F_j^*(n) = \frac{F_j(n) - F_j^{\text{median}}}{F_j^{\text{median}}} \qquad j = \{1, 2\}, n = \{1 \ldots N\}, \tag{1}$$

where N is the total number of formant measurements, and j is the formant number.

Applying this process to the data in Fig. 1(a) results in the normalized vowel space shown in Fig. 1(b). This step is not specifically necessary because the density could be determined based on a linear frequency scale, or with other types of transformations (e.g., Barks) and normalizations (e.g., Disner, 1980). The particular

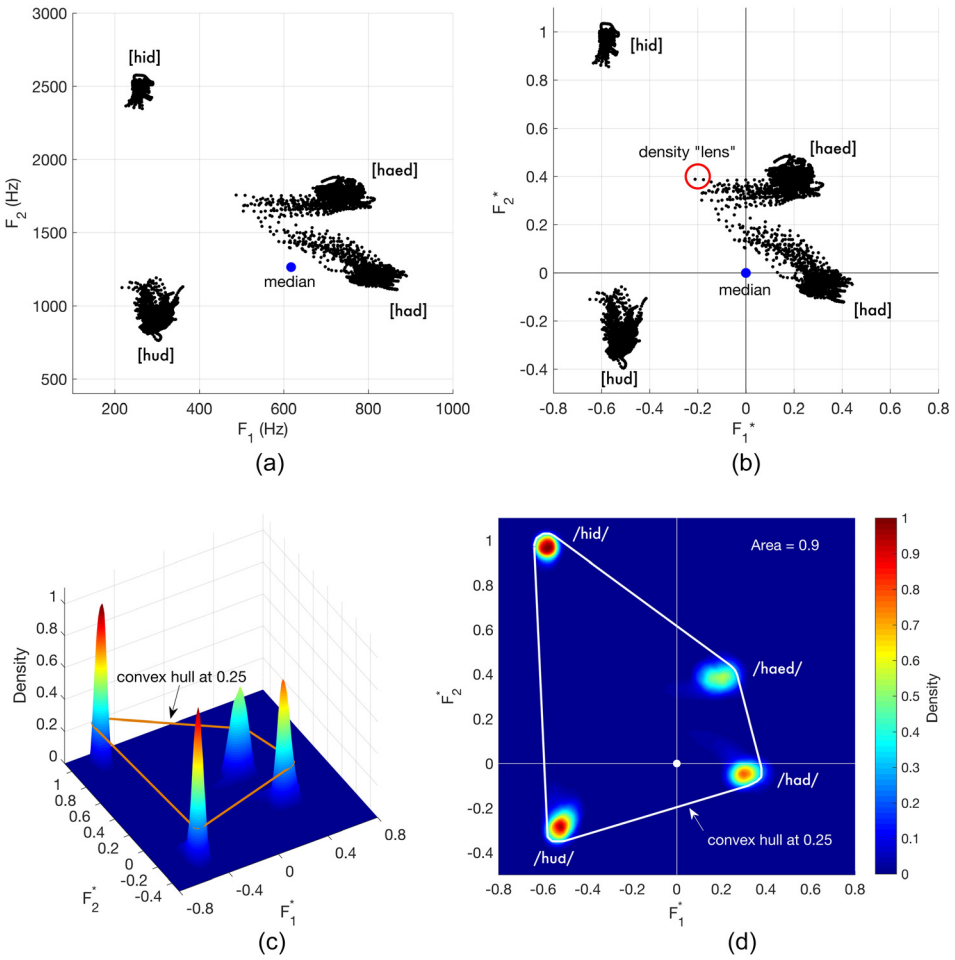

Fig. 1. (Color online) Demonstration of generating a VSD plot and convex hull measurement based on analysis hVd syllables. (a) [F_1, F_2] vowel space in hertz; median value shown as the dot at [617, 1266] Hz. (b) Normalized vowel space based on Eq. (1) where the origin of this coordinate system is the median value from (a); the circle located at [−0.2, 0.4] is an example of the density lens. (c) Perspective view of VSD; the band wrapped around the four vowel pillars is the convex hull at a level of 0.25. (d) Superior view of the VSD and convex hull; area of the convex hull is 0.9.

normalization scheme used here was chosen so that the origin of the new vowel space will be equivalent to the median value of the formants, and the range of both formant axes is roughly constrained to be within the interval [−1, 1]. It is noted, however, that the only true constraint is the value $F_j^*(n) = -1$ which occurs at the asymptotic limit of $F_j(n) = 0$ Hz. In contrast, $F_j^*(n) = 1$ means only that $F_j(n) = 2F_j^{\text{median}}$, a value which could possibly be exceeded during production of some extreme vowels; an example can be seen in Fig. 1(b) where some values of F_2^* calculated for /hid/ are greater than 1.0.

The density of the normalized vowel space [Fig. 1(b)] was determined by first generating a grid that discretized the unitless F_1^* and F_2^* dimensions along the range [−1, 1.5], with an increment of 0.01. The increment value represents 1% of the median formant frequencies and was found to provide sufficient resolution for visualization and analysis of VSD. Smaller increments could be used to enhance resolution but would increase the computation time required to complete the analysis. The next step was to position a circular field of view, or "lens," of a specified radius at every point in the grid and count the number of $[F_1^*, F_2^*]$ pairs contained within the view of the lens. Specifically, the Euclidean distance of each $[F_1^*, F_2^*]$ pair from a given grid point was calculated and the number of pairs whose distance was less than the radius was logged as the density at that grid point. The lens radius was set to 0.05 for all density calculations in this study. As an example, the lens is shown in Fig. 1(b) at a location [−0.2, 0.4] where there would be only two points counted; as the lens is moved toward the right, the number of points that fall within the circle will greatly increase. At this stage, the radius value was set simply by considering that it represents a circular diameter that is 10% of the median values of F_1^* and F_2^* along the horizontal and vertical dimensions, respectively. Different values of the radius could be used to enhance or diminish the density resolution along a continuum. At one end, reducing the radius to

values approaching the discretization of the grid (i.e., 0.01) would not be useful since only a few formant pairs would ever fall within the lens; this would produce a density plot not much different than the raw vowel space plot itself. At the other end, increasing the radius value to be higher than about 0.1 allows for too many formant pairs to be located within the lens at each grid point and results in blurring of the density across the vowel space. It is recommended that the radius setting be limited to the interval [0.025, 0.075].

The absolute density values assigned to each point in the grid depend on the total number of formant pairs obtained during the formant tracking procedure (i.e., the number of data points in the raw and normalized vowel spaces). Thus, analysis of a fairly long audio recording (or collection of recordings) will produce higher absolute densities than a shorter length signal. To normalize the density distribution across talkers or conditions so they can be compared, the density values at every grid point were divided by the maximum density value found within the entire grid. This results in a density range of [0, 1] regardless of the total number of formant pairs plotted in the vowel space.

With the density values providing an additional dimension of information, the two-dimensional vowel space of Fig. 1(b) can now be transformed to the three-dimensional VSD shown in Fig. 1(c). The peaks correspond to regions of vowel space with the highest density of formant values, and hence indicate where in the vowel space the talker spent most of his voiced speaking time. In this test case, where four different hVd syllables were repeated many times, it is not surprising that the VSD contains four fairly sharp peaks. It can be observed, however, that the peaks associated with /hæd/ and /hɑd/ syllables are lower in amplitude and slightly wider near their base. This is a reflection of greater spread of the formant values for these two syllables in the original vowel space plot.

The three-dimensional view is useful for visualizing how the $[F_1^*, F_2^*]$ plane is utilized by a talker, but it may also be useful to quantify a given VSD with a number that can be compared across speaking conditions or across talkers. Much like the VSA described in Sec. 1, the area enclosed in a VSD by the most extreme $[F_1^*, F_2^*]$ values can be measured, but at a selected density threshold within the range between 0 and 1. This area would be largest when measured at a threshold of zero since it would include all formant pairs, and then would decrease as the threshold is increased. For this study, a value of 0.25 was selected as the density level at which an enclosed area was measured so that low density values representing infrequent or spurious formant pairs at the extremes of the vowel space would not contribute to the VSD area. Measurement of the area was accomplished by first finding the set of all points in the $[F_1^*, F_2^*]$ grid at which the density value was greater than or equal to 0.25. A convex hull algorithm (cf. Graham, 1972; Andrew, 1979) was then used to geometrically wrap a "band" around the perimeter of the set to generate the enclosed area. The built-in MATLAB function called "*convhull*" was implemented for this purpose in the current study. It accepts as input the set of *x*-*y* grid pairs from which it finds the convex hull and enclosed area. As an example, the band that wraps around the four peaks in Fig. 1(c) is a convex hull calculated for the VSD at a density level of 0.25.

Rotating the VSD, such that the viewing position is directly from above as shown in Fig. 1(d), allows both the density of the peaks and the shape of the convex hull to be more easily observed. Because this case consisted only of repetitions of hVd syllables containing the corner vowels, the VSD and the resulting convex hull clearly delineate a vowel quadrilateral. The tails projecting toward the median value that were observed in both Figs. 1(a) and 1(b) are still visible in the VSD plot, but their visual prominence is greatly diminished when density is considered. The area enclosed by the convex hull for this case is equal to 0.9, as indicated in the upper right corner of Fig. 1(d). This suggests that an area equal to 1.0 could perhaps be considered a mark of extremely articulated speech.

3. Results

Using the process described in Sec. 2, VSD plots were generated for the audio recordings of the hypo- and hyper-articulated speaking conditions, and are shown in Fig. 2. The VSD for the hypoarticulated (mumbled) condition [Fig. 2(a)] indicates that the density is highest (dark red) at, and within close proximity to, the median value (i.e., [0,0]). This high density region extends along the F_1^* axis to a value of about 0.2 and then becomes vertically oriented in a negative direction along the F_2^* axis. Lower density regions are clearly visible too and expand the vowel space around the median value. The convex hull, determined at a density threshold of 0.25,

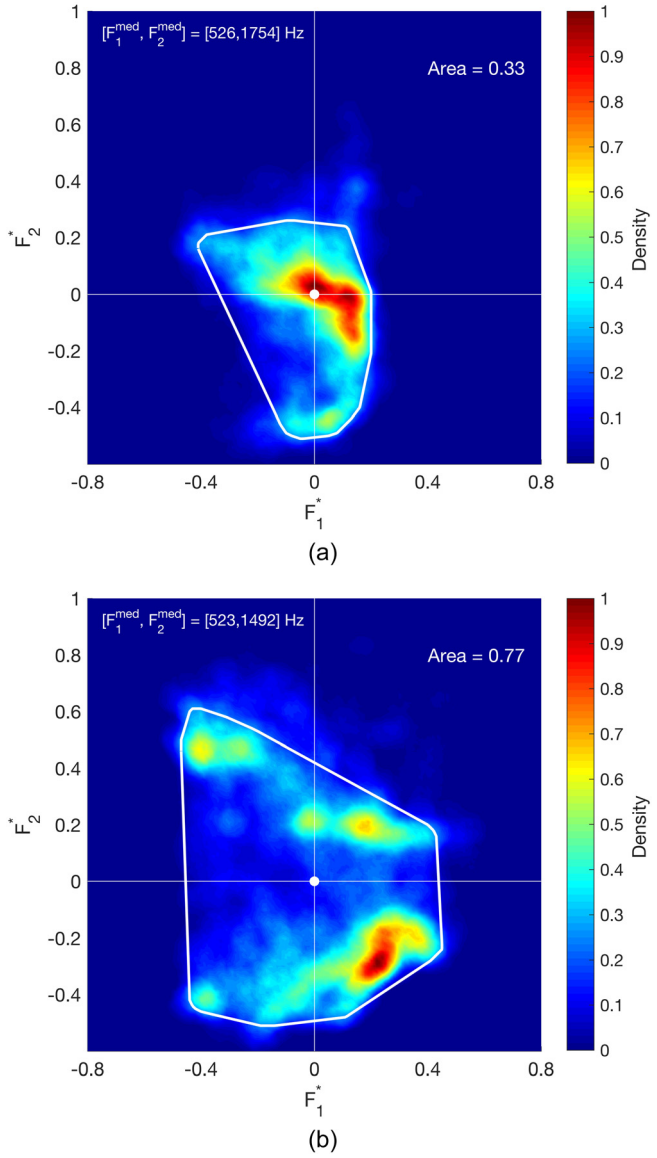

Fig. 2. (Color online) VSD plots of a male talker producing several minutes of (a) hypoarticulated speech and (b) hyperarticulated speech. Convex hulls based on threshold value of 0.25 are shown on each plot as well as the measurement of the enclosed area (unitless).

is shown as the white line and encloses an area measured to be 0.33, only about one-third of the area measured for the hVd condition.

The VSD for the hyperarticulated version of exactly the same material is shown in Fig. 2(b). In this case, the high density regions are located on the edges of the vowel space rather than near the center as in the previous condition. This suggests that the talker emphasized the vowel targets and reduced the amount of time spent producing centralized vowels. In fact, the density near the [0,0] origin is quite low. The convex hull area is 0.77 which is smaller than the area measured for the hVd recording analyzed in Sec. 2, but is 2.3 times larger than the hypoarticulated condition.

In the upper left corner of both VSD plots in Fig. 2 are the median formant values (in Hz) that were used to perform the normalization. The median F_1 value is nearly the same in both the hypo- and hyper-articulated cases (i.e., 526 and 523 Hz, respectively), but for F_2 the median is higher in the hypo-articulated case. This is not unexpected considering the difference in the speaking conditions, but it is important to be aware that normalized VSDs are based on the median values of a specific analyzed recording.

4. Discussion

The VSD plots developed in this study can be used as a visualization of how a talker utilizes the vowel space over several minutes (or more) of speaking. In the deliberately

hypoarticulated condition, the VSD [Fig. 2(a)] clearly showed that the talker spent most of his voiced speaking time producing vowels that were near the center of the vowel space. The entirety of the vowel space used, however, is also apparent in the plot and indicates that there were some brief instances in which the formants were shifted well away from the center. The VSD plot of the hyperarticulated condition [Fig. 2(b)] provides an interesting contrast to the hypoarticulated case in that the dramatic outward expansion of the vowel space is easily observed. Perhaps more importantly is that the VSD shows *how* the talker modified his speech with respect to vowel production. In particular, the high density regions appear to correspond to specific vowels, suggesting that the talker limited most of his vowel production to focused targets in the space, and nearly eliminated central vowels.

The measurement of the VSD area based on a convex hull provides a means of quantifying the size of the vowel space used by a talker. The density threshold of 0.25 was chosen so that low density formant values were excluded from the area measurement values. For example, a brief production of extreme vowels during an otherwise relatively hypoarticulated speaking condition would not affect the convex hull area measurement. It is only if those extreme vowels become more frequent that they would contribute. Future testing of the VSD approach may show that other threshold values are more useful, but at this proof-of-concept stage, 0.25 was deemed a reasonable value. In any case, the threshold must be set to the same value in order to compare different speaking conditions or talkers.

While the simplest means of quantifying the convex hull is by measuring its area at a single density threshold, there are other possible characterizations of a convex hull that may provide useful information as well. Areas could be measured at several density levels from low to high, and ratios of these areas could be calculated to indicate the degree to which the high density regions become localized in the normalized vowel space. A characterization of the shape of the convex hull may also be useful in understanding some features of speech production. For example, two convex hulls may enclose the same area but one might be elongated to a greater extent in the horizontal dimension than the vertical, whereas the other case may be oppositely shaped. Quantifying the shapes in terms of major and minor axis lengths or some other parametric approach might allow for further differentiation of vowel space features.

VSD plots provide a different, and possibly more comprehensive, view of vowel characteristics during speech production than other approaches that utilize formant measurements limited to content words or sustained vowels. The qualitative visualization as well as quantification based on convex hull measurements could be used to compare the speech production of a specific talker in different speaking conditions, whether they be based on different instructions, as demonstrated in Fig. 2, or perhaps before and after treatment for a speech disorder. The VSD and associated measurements are not necessarily intended to replace more established measurements such as VSA or FCRs, but rather to offer additional insight into production of vowels over the course of several minutes of connected speech. To fully understand the advantages and disadvantages of the VSD relative to other methods of characterizing vowel space will require studies in which several approaches are applied to the same case, such as a pre- and post-treatment situation. Information obtained from measurements of more typical VSA or FCRs could then be compared to that determined by the VSD approach.

Acknowledgments

The research was supported by NIH R01-DC011275 and NSF BCS-1145011.

References and links

Andrew, A. M. (**1979**). "Another efficient algorithm for convex hulls in two dimensions," Inf. Process. Lett. **9**(5), 216–219.

Crystal, T. H., and House, A. S. (**1982**). "Segmental durations in connected speech signals: Preliminary results," J. Acoust. Soc. Am. **72**(3), 705–716.

Disner, S. F. (**1980**). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am. **67**(1), 253–261.

Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," J. Speech, Lang., Hear. Res. **50**(5), 1241–1255.

Graham, R. L. (**1972**). "An efficient algorithm for determining the convex hull of a finite planar set," Inf. Process. Lett. **1**(4), 132–133.

Kent, R. D., and Kim, Y. J. (**2003**). "Toward an acoustic typology of motor speech disorders," Clin. Linguist. Phonetics **17**(6), 427–445.

Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E., and Russell, S. (**2013**). " 'The Caterpillar': A novel reading passage for assessment of motor speech disorders," Am. J. Speech-Lang. Pathol. **22**(1), 1–9.

Sapir, S., Ramig, L. O., Spielman, J. L., and Fox, C. (**2010**). "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," J. Speech, Lang., Hear. Res. **53**(1), 114–125.

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

Titze, I. R., Horii, Y., and Scherer, R. C. (**1987**). "Some technical considerations in voice perturbation measurements," J. Speech, Lang., Hear. Res. **30**(2), 252–260.

Tjaden, K., Rivera, D., Wilding, G., and Turner, G. S. (**2005**). "Characteristics of the lax vowel space in dysarthria," J. Speech, Lang., Hear. Res. **48**(3), 554–566.

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

Turner, G. S., Tjaden, K., and Weismer, G. (**1995**). "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," J. Speech Hear. Res. **38**, 1001–1003.

Weismer, G., Laures, J. S., Jeng, J.-Y., Kent, R. D., and Kent, J. F. (**2000**). "Effect of speaking rate manipulations of acoustic and perceptual aspects of the dysarthria in Amyotrophic Lateral Sclerosis," Folia Phoniatr. Logop. **52**, 201–219.

Xie, Z., and Niyogi, P. (**2006**). "Robust acoustic-based syllable detection," in INTERSPEECH, paper 1327-Wed1BuP. 6.