Identification of synthetic vowels based on selected vocal tract area functions (L)

Kate Bunton[a] and Brad H. Story[b]
Speech Acoustics Laboratory, Department of Speech, Language, and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 3 June 2008; revised 14 October 2008; accepted 16 October 2008)

The purpose of this study was to determine the degree to which synthetic vowel samples based on previously reported vocal tract area functions of eight speakers could be accurately identified by listeners. Vowels were synthesized with a wave-reflection type of vocal tract model coupled to a voice source. A particular vowel was generated by specifying an area function that had been derived from previous magnetic resonance imaging based measurements. The vowel samples were presented to ten listeners in a forced choice paradigm in which they were asked to identify the vowel. Results indicated that the vowels [i], [æ], and [u] were identified most accurately for all of speakers. The identification errors of the other vowels were typically due to confusions with adjacent vowels. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3033740]

PACS number(s): 43.70.Bk, 43.70.Mn, 43.70.Aj [AL] Pages: 19–22

I. INTRODUCTION

Magnetic resonance imaging (MRI) has been widely used to acquire volumetric image sets of the head and neck from which vocal tract area functions can be directly measured. These collections of area functions, which are assumed representative of an individual speaker's production of a target vowel or consonant, have then been used in the development of speech production models and speech synthesizers (e.g., Ciocea, 1997; Story, 2005a, 2005b; Mullen *et al.*, 2007).

The similarity of speech sounds produced by area-function-based synthesis to natural speech has been typically assessed by comparing calculated formant frequencies to formant frequencies extracted from recorded speech (Story *et al.*, 1996, 1998; Story, 2005a). Reasonable similarity has been demonstrated; however, stimuli generated based on measured area functions have rarely been evaluated perceptually. This step is important before stimuli generated by simulation of the speech production process are used to answer questions about the perceptual relevance of various types of kinematic and structural variations of the vocal tract (Carré *et al.*, 2001).

Collections of volumetric image sets based on MRI and their analyses have been reported by Story (2005a) and Story *et al.* (1996, 1998) for eight speakers (four females and four males). A second set of data obtained from the speaker presented in Story *et al.*, 1996 has been published as well (Story, 2008). The inventories include area functions (area as a function of distance from the glottis) of a set of 10 or 11 American English vowels ([i,ɪ,e,ɛ,æ,ʌ,ɑ,ɔ,o,ʊ,u]), depending on the particular speaker. Across speakers, vocal tract area functions varied in vocal tract length and other idiosyncratic differences, but were similar with regard to gross shape for each of the target vowels and the location of major constrictions and expansions.

The measured area functions were subsequently used as input to a computer model of one-dimensional acoustic wave propagation in the vocal tract. The synthetic speech samples were then compared, in terms of location of the first three formant frequencies, to recorded natural speech from each speaker. The natural speech samples were recorded with the subject in a supine position with ear plugs in an attempt to simulate, as closely as possible, the conditions experienced in the MRI sessions. Subjects produced speech sounds that corresponded to the static shapes that were acquired with MRI. Percent error based on comparisons of measured and calculated formant frequencies from natural and simulated speech across speakers and formants (F1-F2-F3) ranged from 0.1% to 39%. Errors larger than 30% were calculated for only seven instances and were limited to two speakers. The majority (defined here to be 95%) of the calculated formants differed from those of natural speech by less than 10%. Overall, results indicated that formant locations of the synthesized samples were reasonably well represented compared to the natural productions that were recorded. These comparisons quantify the success of measurement of area functions from MRI images and speech modeling efforts. However, since one aim of developing a speech production model is to understand the relation between area functions and changes in the vocal tract shape result in acoustic characteristics indicative of a phonetic category, perceptual testing of simulated samples based on these area functions is needed. The purpose of the present study was to determine the vowel identification accuracy for simulated vowel samples of eight speakers based on previously reported vocal tract area functions derived from MRI image sets.

[a]Electronic mail: bunton@email.arizona.edu
[b]Electronic mail: bstory@u.arizona.edu

II. METHOD

A. Area function sets

Previously published area functions for eight speakers were used to synthesize vowel samples in the present study (Story *et al.*, 1996, 1998; Story, 2005a). This included four male (range 29–40 years) and four female (range 23–39 years) speakers. Speakers in Story's (2005a) article were identified as SF1, SF2, SF3, SM1, SM2, and SM3, where "F" denotes female and "M" denotes male. The two speakers presented in Story *et al.*, 1996, 1998 will be identified as SM0 and SF0, respectively. Finally, data for a second set of area functions obtained from speaker SM0 in 2002 will be identified as SM0-2.

B. Synthetic vowel samples

A synthetic vowel sample was generated for each area function of each speaker's inventory. Following Hillenbrand and Gayvert (1993), the duration of all samples was set at 0.3 s, and the fundamental frequency (F0) contour varied from 25% above an F0 target to 25% below that same target. The F0 targets for males and females were set at 110 and 220 Hz, respectively. The sample duration was chosen so that it would not be a primary cue in vowel identification (Hillenbrand *et al.*, 2000); that is, 0.3 s is on average shorter than long vowels and longer than short vowels. The samples were generated with a wave-reflection model of the trachea and vocal tract (Liljencrants, 1985; Story, 1995) that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips. The tracheal portion extended from the glottis to the bronchial termination. Its shape was idealized as a tube that is tapered from 0.3 cm^2 to just below the glottis to a constant area of 1.5 cm^2. All synthesized vowels were based on coupling this tracheal configuration to the respective measured area functions, which included their measured vocal tract lengths. The synthesis was driven by the respiratory pressure (P_R) assumed to exist at the bronchial termination of the trachea. In generating each sample for this study, P_R was ramped from 0 to 6000 dyn/cm^2 in 20 ms with a cosine function, similar to Hillenbrand and Gayvert's (1993) ramping of peak amplitude. The voice source was generated by a model of the time-varying glottal area for which wave shape parameters such as F0, amplitude, pulse skewing (skewing quotient), and duty cycle (open quotient) can be varied over the duration of the synthesized speech sound or held constant. The glottal area model was based on the glottal flow pulse model of Rosenberg (1971) but scaled in amplitude for glottal area. For each sample, the F0 followed either the male or female contour detailed above, the maximum glottal opening was set at 0.08 cm^2, the skewing quotient was held at a value of 2.4, and the open quotient was set to 0.6. The appropriateness of these values for both male and female speech might be questioned; however, they were chosen so that the energy in the harmonic components of the glottal flow wave would be similar for all samples. Although these parameters may reduce the femalelike quality of the samples produced with the SF-area functions, this was not considered to be problematic since the listening task was concerned only with phonetic identification.

In addition to the synthetic samples based on the original measured vowel area functions, a sample was also generated from each speaker's mean area function. That is, the mean of all 10 or 11 vowels measured for each speaker. These samples are effectively neutral vowels and were used as precursors to the other samples in the listening tests to provide a context for extrinsic normalization of each speaker (e.g., Ladefoged and Broadbent, 1957).

C. Listening Task

Ten listeners (mean age 26 years) participated in the present study. Listeners were native English speakers and native to Arizona and passed a hearing screening. All procedures were approved by the Institutional Review Board at the University of Arizona.

An ALVIN interface (Hillenbrand and Gayvert, 2005) was used to present samples via loudspeakers to listeners seated in a sound treated room. Samples were presented in pairs with the first sample being the mean vowel of a particular speaker followed by a target vowel from the same speaker. The computer screen displayed buttons for 11 English vowels that were labeled with both the phonetic symbol and an example "hVd" word. Listeners were asked to identify the second vowel in the pair. Vowel samples were blocked by speaker and each listener heard five repetitions of each vowel. The order of presentation for speaker and vowel samples was randomized. Each listening session lasted no longer than 30 min. A confusion matrix based on listener identification of the vowel samples was calculated separately for each speaker. Listeners also completed a training task with vowel samples recorded by a male speaker (second author) to assure they could identify all 11 English vowels. Accuracy was greater than 98% across vowels and listeners. Errors were limited to a confusion of [ɔ] and [ɑ].

III. RESULTS

Identification errors made for the vowels based on each speaker are indicated in the confusion matrices displayed in Tables I–III. In each matrix, the target vowel is listed in the leftmost column and the vowel identified is listed across the top of the columns. Accurate identification of target tokens can be seen along the diagonal in the boldface cells.

Accuracy across vowels varied from a low of 21% for female [ɛ] to a high of 98% for male [i]. Vowels with the highest accuracy rates across speakers (>89%) included three English corner vowels [i, æ, u]. Accuracies for the three vowels [ɛ, ɔ, ɑ] were greater than 50% for the male speakers and less than 50% for the female speakers. For the vowels [ɪ, e, o] and [ʌ] identification accuracy was less than 50% for both male and female speakers.

Although there was considerable variability in the identification accuracy, vowel confusions were typically between adjacent vowel categories in the vowel space. For example, the target vowel [ɪ] was identified as either [e] or [ɛ] for all of the speakers except SF2 whose [ɪ] targets were identified

TABLE I. Confusion matrices for synthesized vowels of speakers SM0 and SM0–2.

		Listener's identification											
		i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u	Total
Vowel intended by speaker SM0	i	**50**	0	0	0	0	0	0	0	0	0	0	50
	ɪ	0	**1**	32	17	0	0	0	0	0	0	0	50
	e	0	0	**0**	0	0	0	0	0	0	0	0	0
	ɛ	0	0	1	**25**	23	0	1	0	0	0	0	50
	æ	0	0	0	1	**48**	0	1	0	0	0	0	50
	ʌ	0	0	0	0	0	**22**	22	6	0	0	0	50
	ɑ	0	0	0	0	1	0	**29**	20	0	0	0	50
	ɔ	0	0	0	0	0	0	8	**28**	14	0	0	50
	o	0	0	0	0	0	0	1	1	**9**	6	33	50
	ʊ	0	0	0	0	0	0	0	0	0	**9**	41	50
	u	0	0	0	0	0	0	0	0	1	6	**43**	50
Vowel intended by speaker SM0-2	i	**50**	0	0	0	0	0	0	0	0	0	0	50
	ɪ	6	**23**	16	5	0	0	0	0	0	0	0	50
	e	0	8	**24**	18	0	0	0	0	0	0	0	50
	ɛ	0	0	2	**35**	13	0	0	0	0	0	0	50
	æ	0	0	0	0	**49**	0	1	0	0	0	0	50
	ʌ	0	0	0	0	0	**0**	10	39	0	1	0	50
	ɑ	0	0	0	0	0	0	**27**	23	0	0	0	50
	ɔ	0	0	0	0	0	0	5	**27**	16	2	0	50
	o	0	0	0	0	0	0	1	6	**43**	0	0	50
	ʊ	0	0	0	0	0	0	2	5	43	**0**	0	50
	u	0	0	0	0	0	0	0	0	0	3	**47**	50

TABLE II. Confusion matrices for synthesized vowels of speakers SM1, SM2, and SM3.

		Listener's identification											
		i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u	Total
Vowel intended by speaker SM1	i	**46**	3	0	0	0	0	0	0	0	0	1	50
	ɪ	0	**37**	1	2	0	0	0	0	0	5	5	50
	e	0	14	**4**	13	2	5	0	0	0	9	3	50
	ɛ	0	2	2	**21**	3	11	0	0	0	10	1	50
	æ	0	0	0	1	**48**	0	1	0	0	0	0	50
	ʌ	0	0	0	0	0	**41**	2	0	0	7	0	50
	ɑ	0	0	0	0	1	0	**34**	15	0	0	0	50
	ɑ	0	0	0	0	0	0	19	**31**	0	0	0	50
	o	0	0	0	0	0	0	0	0	**7**	25	18	50
	ʊ	0	0	0	0	0	0	0	0	0	**17**	33	50
	u	0	0	0	0	0	0	0	0	0	5	**45**	50
Vowel intended by speaker SM2	i	**50**	0	0	0	0	0	0	0	0	0	0	50
	ɪ	0	**33**	5	12	0	0	0	0	0	0	0	50
	e	4	22	**14**	10	0	0	0	0	0	0	0	50
	ɛ	0	0	0	**0**	0	0	0	0	0	0	0	0
	æ	0	0	0	0	**50**	0	0	0	0	0	0	50
	ʌ	0	0	0	0	0	**42**	0	0	1	6	1	50
	ɑ	0	0	0	0	0	0	**43**	7	0	0	0	50
	ɔ	0	0	0	0	0	0	4	**46**	0	0	0	50
	o	0	0	0	0	0	2	2	3	**14**	23	6	50
	ʊ	0	0	0	0	0	7	0	0	2	**38**	3	50
	u	0	0	0	0	0	0	0	0	0	8	**42**	50
Vowel intended by speaker SM3	i	**49**	1	0	0	0	0	0	0	0	0	0	50
	ɪ	0	**1**	37	12	0	0	0	0	0	0	0	50
	e	0	0	**26**	23	1	0	0	0	0	0	0	50
	ɛ	0	1	21	**26**	2	0	0	0	0	0	0	50
	æ	0	0	1	0	**49**	0	0	0	0	0	0	50
	ʌ	0	0	0	0	0	**0**	0	0	10	36	4	50
	ɑ	0	0	0	0	0	0	**35**	15	0	0	0	50
	ɔ	0	0	0	0	0	0	8	**42**	0	0	0	50
	o	0	0	0	0	0	13	8	9	**19**	0	1	50
	ʊ	0	0	0	0	0	2	0	0	1	**31**	16	50
	u	0	0	0	0	0	0	0	0	1	7	**42**	50

TABLE III. Confusion matrix for synthesized vowels of speaker SF0.

		Listener's identification											
		i	ɪ	e	ɛ	æ	ʌ	ɑ	ɔ	o	ʊ	u	Total
Vowel intended by speaker SF0	i	**50**	0	0	0	0	0	0	0	0	0	0	50
	ɪ	0	**10**	9	26	5	0	0	0	0	0	0	50
	e	0	0	**0**	0	0	0	0	0	0	0	0	0
	ɛ	0	0	1	**8**	41	0	0	0	0	0	0	50
	æ	0	0	0	6	**44**	0	0	0	0	0	0	50
	ʌ	0	0	0	0	0	**0**	39	11	0	0	0	50
	ɑ	0	1	0	0	0	22	**19**	5	2	1	0	50
	ɔ	0	0	0	0	0	2	6	**23**	19	0	0	50
	o	0	0	0	0	0	12	5	3	**13**	17	0	50
	ʊ	0	0	0	0	0	0	25	25	0	**0**	0	50
	u	0	0	0	0	0	1	0	0	0	5	**44**	50
Vowel intended by speaker SF1	i	**49**	0	1	0	0	0	0	0	0	0	0	50
	ɪ	0	**23**	20	7	0	0	0	0	0	0	0	50
	e	0	1	**10**	36	2	1	0	0	0	0	0	50
	ɛ	0	3	0	**1**	0	5	0	0	0	33	8	50
	æ	0	0	0	0	**50**	0	0	0	0	0	0	50
	ʌ	0	0	0	0	0	**30**	3	0	5	11	1	50
	ɑ	0	0	0	0	46	0	**3**	1	0	0	0	50
	ɔ	0	0	0	0	0	0	16	**33**	1	0	0	50
	o	0	0	0	0	0	6	6	0	**24**	14	0	50
	ʊ	0	0	0	0	0	0	0	0	0	**29**	21	50
	u	0	0	0	0	0	0	0	0	0	6	**44**	50
Vowel intended by speaker SF2	i	**27**	22	0	1	0	0	0	0	0	0	0	50
	ɪ	0	**0**	5	0	0	35	0	0	0	10	0	50
	e	2	29	**7**	11	1	0	0	0	0	0	0	50
	ɛ	0	0	5	**1**	44	0	0	0	0	0	0	50
	æ	0	0	0	0	**49**	0	0	1	0	0	0	50
	ʌ	0	0	0	0	0	**39**	4	1	2	3	1	50
	ɑ	0	0	0	0	0	1	**12**	22	15	0	0	50
	ɔ	0	0	0	0	0	0	32	**18**	0	0	0	50
	o	0	0	0	0	0	2	0	0	**11**	34	3	50
	ʊ	0	0	0	0	0	0	0	1	0	**42**	7	50
	u	0	0	0	0	0	0	0	0	0	4	**46**	50
Vowel intended by speaker SF3	i	**48**	2	0	0	0	0	0	0	0	0	0	50
	ɪ	2	**9**	30	9	0	0	0	0	0	0	0	50
	e	1	21	**21**	7	0	0	0	0	0	0	0	50
	ɛ	0	4	11	**31**	0	3	0	0	0	1	0	50
	æ	0	0	0	0	**49**	0	0	1	0	0	0	50
	ʌ	0	0	0	0	0	**1**	0	0	3	38	8	50
	ɑ	0	0	0	0	0	0	**30**	20	0	0	0	50
	ɔ	0	0	0	0	0	0	22	**28**	0	0	0	50
	o	0	0	0	0	0	0	4	6	**22**	17	1	50
	ʊ	0	0	0	0	0	0	0	0	0	**10**	40	50
	u	0	0	0	0	0	0	0	0	0	6	**44**	50

as [ʌ]. A similar confusion was found for the target [e], which was identified as either [ɪ] or [ɛ] for all speakers. Identification of the target vowel [ɑ] included both [æ] and [ɔ] responses, and is the only vowel where listener identification was not an adjacent vowel category. The target [ɔ] was most commonly confused with [ɑ]. This confusion is not unexpected and the vowel [o] was confused with [ʊ] in a majority of cases. Confusion between [ɔ] and [ɑ] is not unexpected given that speakers in the southwest part of the United States tend to collapse these two categories (Labov, 1996).

Comparisons of vowel identification accuracy for the samples based on the data sets from the same speaker (SM0 and SM0-2) were similar, with overall accuracy slightly higher for the second data set (53% and 59%, respectively). The largest difference between samples was seen for the vowels [ɪ] and [o]. In both cases, confusions were between adjacent vowels.

IV. DISCUSSION

The confusion matrices suggest that most of the area functions from each speaker's inventory produce sound samples that can be expected to be identified as either the target vowel or as an adjacent vowel in the vowel space. Therefore, with a few exceptions, each area function is representative of the "neighborhood" of the target vowel. The modest accuracy rates for several vowels, however, also beg the question of why the identification accuracy is not better.

An obvious possibility is that some of the area functions are simply not good representations of the target vowels. In some cases, this is likely true. For example, the poor identification of SF0's [ʊ] vowel could have likely been predicted based on the fairly large errors found between the formant frequencies calculated from the [ʊ] area function and those measured from natural speech (Story *et al.*, 1998). In other cases, a presumably good area function representation of a particular vowel would not have predicted poor identification accuracy. SM2's [ɪ] area function produced formant frequencies with small error relative to natural speech and yet the identification responses indicated that listeners were correct only 66% of the time. Although area function quality is undoubtedly part of the problem, it would seem that other factors must also contribute.

The constant 0.3 s duration that was used to generate every sample may have affected some identification responses, especially for the "short" vowels. This duration was chosen as a compromise between short and long vowels (Hillenbrand and Gayvert, 1993), but may have been too long such that it inadvertently created a cue that conflicted with the typical duration of some of the shorter vowels.

Another possible reason for reduced identification accuracy is that each vowel sample was generated from a "static" area function. That is, each vowel was effectively produced without any change in vocal tract shape and, hence, no change in formant frequencies. In connected speech, vowels are typically embedded between consonants so that the formant frequencies are almost continuously in transition. Even productions of isolated vowels tend to have formant transitions over the course of the utterance (e.g., Story, 2007). There is much evidence that listeners use this dynamic spectral change for identification of vowels (Jenkins *et al.*, 1983; Strange *et al.*, 1983; Nearey, 1989; Hillenbrand and Gayvert, 1993; Nittrouer, 2007).

Finally, the listening paradigm, which consisted of presentations blocked by speaker and included a precursor mean vowel followed by the target, may have influenced the identification accuracy. This paradigm was implemented so that the precursor might allow for extrinsic normalization by the listener. Similar methods have been used with some success for vowel recognition algorithms (Pols and Weenink, 2005; Nearey and Assman, 2007).

The next steps in this research are to explore some of these possible influences on vowel identification; specifically, use of area functions for each speaker that have been "tuned" to produce formant frequencies directly aligned with those of recorded speech (Story, 2006), use of an area function model that allows for time variation of the vocal tract shape (e.g., Story, 2005b), build in natural vowel durations, and use of a listening paradigm that does not include a precursor vowel for normalization.

ACKNOWLEDGMENTS

This research was supported by NIH Grant No. R01-DC04789.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S., and Pasdeloup, V. (**2001**). "Perception of vowel-to-vowel transitions with different formant trajectories," Phonetica **58**, 163–178.

Ciocea, S. (**1997**). "Semi-analytic formant-to-area mapping," Ph.D. thesis, Université Libre de Bruxelles, Brussels, Belgium.

Hillenbrand, J., and Gayvert, R. T. (**1993**). "Identification of steady-state vowels synthesized from the Peterson and Barney measurements," J. Acoust. Soc. Am. **94**, 668–674.

Hillenbrand, J., Clark, M., and Houde, R. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**, 3013–3022.

Hillenbrand, J., and Gayvert, R. T. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res. **48**, 45–60.

Jenkins, J. J., Strange, W., and Edman, T. R. (**1983**). "Identification of vowels in 'vowelless' syllables," Percept. Psychophys. **34**, 441–450.

Labov, W. (**1996**). "The organization of dialectic diversity in North America," presented at the Fourth International Conference on Spoken Language Proceeding, Philadelphia, 6 October; Available online at www.ling.upenn.edu/phono_atlas/ICSLP4.html (Last viewed 10/9/2008).

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

Liljencrants, J. (**1985**). "Speech synthesis with a reflection-type line analog," DS thesis, Department of Speech Communication and Music Acoustic, Royal Institute of Technology, Stockholm, Sweden.

Mullen, J., Howard, D. M., and Murphy, D. T. (**2007**). "Real-time dynamic articulations in the 2-D waveguide mesh vocal tract model," IEEE Trans. Audio, Speech, Lang. Process. **15**, 577–585.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nearey, T. M., and Assmann, P. (**2007**). "Probabilistic 'sliding-template' models for indirect vowel normalization," in *Experimental Approches to Phonology*, edited by M. Sole, P. Speeter Beddor, and M. Ohala (Oxford University Press, Oxford).

Nittrouer, S. (**2007**). "Dynamic spectral structure specifies vowels for children and adults," J. Acoust. Soc. Am. **122**, 2328–2339.

Pols, L., and Weenink, D. (**2005**). "Vowel recognition and (adaptive) speaker normalization," *Proceedings of the Tenth International Conference on Speech and Computer*, edited by G. Kokkinakis (University of Patras Press, Patras, Greece), Vol. **1**, 17–24.

Rosenberg, A. (**1971**). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.

Story, B. H. (**1995**). "Speech simulation with an enhanced wave-reflection model of the vocal tract," .Ph.D. thesis, University of Iowa, Iowa City, IA.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am. **100**, 537–554.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1998**). "Vocal tract area functions for an adult female speaker based on volumetric imaging," J. Acoust. Soc. Am. **104**, 471–487.

Story, B. H. (**2005a**). "Synergistic modes of vocal tract articulation for American English vowels," J. Acoust. Soc. Am. **118**, 3834–3859.

Story, B. H. (**2005b**). "A parametric model of the vocal tract area function for vowel and consonant simulation," J. Acoust. Soc. Am. **117**, 3231–3254.

Story, B. H. (**2006**). "A technique for 'tuning' vocal tract area functions based on acoustic sensitivity functions," J. Acoust. Soc. Am. **119**, 715–718.

Story, B. H. (**2007**). "Time-dependence of vocal tract modes during production of vowels and vowel sequences," J. Acoust. Soc. Am. **121**, 3770–3789.

Story, B. H. (**2008**). "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J. Acoust. Soc. Am. **123**, 327–335.

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