Journal of Phonetics (1998) **26,** 223–260
Article ID: jp980076

Parameterization of vocal tract area functions by empirical orthogonal modes

Brad H. Story

National Center for Voice and Speech, WJ Gould Research Center, Denver Center for the Performing Arts, 1245 Champa St, Denver CO 80204, USA

Ingo R. Titze

National Center for Voice and Speech, WJ Gould Research Center, Denver Center for the Performing Arts, Denver USA and Dept. Speech Pathology and Audiology, University of Iowa, USA

Received 8 September 1997, revised 30 March 1998, accepted 22 April 1998

A set of ten vowel area functions, based on MRI measurements, has been parameterized by an "empirical orthogonal mode decomposition" which accurately represents each area function as the sum of the mean area function and proportional amounts of a series of orthogonal basis functions. The mean area function was found to possess a formant structure similar to that of a uniform tube (i.e., nearly equally spaced formants) suggesting that empirical orthogonal modes are perturbations on the mean (~neutral) vowel shape much like past vocal tract analyses have considered perturbations on a uniform tube. The acoustic characteristics of the two most significant empirical orthogonal modes were examined, showing that both modes tend to increase the first formant as the modal amplitude coefficients are both increased from negative to positive values. However, the second formant was found to decrease in frequency for increasing values of the first modal coefficient and to increase for increasing values of the second mode coefficient. Next, a mapping between F1-F2 formant pairs and vocal tract area functions is proposed which is largely one-to-one but was initially limited by a constant vocal tract length. A possible method to include variable vocal tract length and higher ordered orthogonal modes in the mapping is given. The mode-to-formant mapping suggested the possibility of an inverse mapping to determine physiologically realistic area functions from a speech waveform and a simple example is presented. Finally, empirical orthogonal modes for a collection of ten vowels and eight consonants were derived and showed many similarities to those for the vowel-only case.

1. Introduction

Models of the vocal tract have long been used to transform articulatory parameters, such as the positions of the tongue, lips and velum, to an area function; i.e., the cross-sectional area of the vocal tract as a function of the distance from the glottis. These models have typically been defined with reference to the midsagittal plane, which is a convenient reference because of the large body of x-ray films of speech production that are available for

0095–4470/98/030207 + 24 $30.00/0

analysis (e.g., Munhall, Vatikiotis-Bateson & Tohkura, 1994), and also the physiological correlation between model parameters and human articulatory structures. Such parametric models provide a simple, compressed representation of the state of the vocal tract at a given point in time. Articulatory parameters typically are of lower dimension than a full area function representation, but they are dependent on an accurate transformation from midsagittal distance to cross-sectional area. Examples of such midsagittally based models can be found in Lindblom and Sundberg (1971), Mermelstein (1973), Coker (1976), and Browman and Goldstein (1990).

Other highly compact articulatory models are those of Stevens and House (1955) and Fant (1960), both of which represented the vocal tract with only three parameters: the place and cross-sectional area of the main vocal tract constriction and a ratio of lip protrusion to lip open area. With these parameters, the entire area function from just above the glottis to the lip can be constructed by empirically-based rules.

Models such as these all depend heavily on the intuition and experience of the researcher to decide which features of the vocal tract shape are most significant and how to provide a numerical description of those features. This approach has produced valuable tools for synthesizing speech and for explaining many phenomena in both speech production and perception. However, it would be useful to have a more objective parameterization of the vocal tract shape. An example of such an approach is found in Liljencrants (1971), where he sought to explain the midsagittal profile of the tongue for 10 vowel shapes with a Fourier series representation. He made three key observations regarding his collection of tongue profiles: (1) the mean displacement of the tongue from a neutral position did not significantly change across vowels, implying a conservation of mass of the tongue body; (2) the fine structure of each profile was much smaller than that of the overall shape variation; and (3) many of the tongue profiles showed a strong resemblance to a sinusoid. These particular features suggested that the shape of the tongue for each vowel could be described by proportional amounts of a standard set of orthogonal basis functions: in this case, a Fourier series. Liljencrants found that the tongue shape could be reconstructed with small error using only a DC term and the first significant Fourier component. This representation produced about a 9 to 1 compression of the original data.

A similar study was performed by Harshman, Ladefoged and Goldstein (1977) where midsagittal tongue profiles for 10 English vowels were analyzed. However, instead of using an *a priori* set of the basis functions (i.e., Fourier series), a specialized 3-way factor analysis was developed to decompose the tongue shapes into a set of empirical factors that characterized the displacement of the tongue during production of the 10 vowels. Their analysis revealed two underlying displacement patterns, various proportions of which could be used to reconstruct the original tongue shapes. Much like the Fourier description proposed by Liljencrants (1971), the two displacement patterns uncovered by the factor analysis allow the tongue profile of all the vowel shapes to be represented by a set of basic features (or patterns) and a set of amplitude (weighting) coefficients that define each individual vowel.

The primary difference between the two studies is that the basis functions are empirically derived by the analysis in Harshman *et al.* (1977) rather then chosen *a priori* as in Liljencrants (1977). Lagefoged, Harshman, Goldstein and Rice (1978) used a multiple regression analysis to generate a model in which the weighting coefficients of the derived tongue shape factors from Harshman *et al.* (1977) were represented as a function of the formants (F1, F2 and F3) measured from acoustic recordings of their subjects.

Thus, given the first three formant frequencies, a tongue shape profile could be generated. This approach significantly reduced the difficulties of deriving a tongue shape from the acoustic speech signal because the factors (i.e., basis functions) were derived from physiologic data. However, this model could only generate two-dimensional tongue profiles rather than complete vocal tract area functions.

Using a factor analysis similar to that of Harshman *et al.* (1977), Jackson (1988) attempted to parameterize Icelandic vowels. Jackson found that three factors were needed to describe these vowels, with the second factor having a significantly different shape than the second factor given by Harshman *et al.* (1977). This led to the suggestion that factor shapes are not universal across languages but are language specific. However, Nix, Papcun, Hogden and Zlokarnik (1996) have re-analyzed Jackson's data and compared the results to Harshman *et al.*'s (1977) results and found that two factors are actually adequate in describing the Icelandic vowels and that the shape of each factor in the re-analysis is remarkably similar to Harshman *et al.*'s (1977) original factor shapes.

Meyer, Wilhelms and Strube (1989) have used a similar approach to generate articulatory parameters for a speech synthesizer. Using the data from Harshman *et al.* (1977), they computed ten-section vocal tract area functions based on midsagittal-to-area transformations. Each area function was assumed to have a length of 17.5 cm, giving a spatial resolution of 1.75 cm. Data from Fant (1960) was also used to supplement their collection. The area functions were then subjected to an eigenfunction decomposition that yielded three eigenvectors capable of explaining 93% of the variance in the area function set.

More recently, Yehia, Takeda and Itakura (1996) also used an eigenfunction decomposition to parameterize a set of area functions (also based on midsagittal-to-area transformations of x-ray sagittal projection images) of one female speaker of French. Their analysis yielded a five eigenvector set that accounted for 92% of the variance. After using their parameterization to create several thousand new area functions and computing formant frequencies for each one, they subsequently employed a series of additional transformations to generate a mapping between articulatory and acoustic parameters. This system showed moderate success at reconstructing vowel-like area functions from the recorded speech of the original female subject.

The quest for a vocal tract parameterization of this type is analogous to a Fourier-based spectral analysis of the acoustic speech waveform in which the sound wave is described by the amplitudes of the Fourier coefficients (Harshman *et al.*, 1977). In Liljencrants (1971), Harshman *et al.* (1977) and Meyer *et al.* (1989), the vocal tract shape for each vowel is described by the amplitudes of a descriptive set of orthogonal basis functions; in Liljencrants (1971) a standard Fourier series formed the set of basis functions and in Harshman *et al.* (1977) and Meyer *et al.* (1989) the basis function set was empirically derived. A similar representation of the vocal tract area function was reported by Schroeder (1966) and Mermelstein (1967) based on purely acoustic considerations of perturbing the shape of a closed-end tube of constant cross-sectional area. They both showed that the area function could be represented as the sum of a Fourier series and a constant area. This representation was developed in an effort to find a possible mapping from the vocal tract resonance peaks (poles) in a frequency spectrum to a specific vocal tract shape. A limitation of this approach is that the formants can only be used to determine the odd components of the Fourier series; the even components were set to zero. However, various sets of even Fourier series components can produce widely varying area functions while maintaining the same formant (pole) locations in the

frequency spectrum. This is the classic many-to-one mapping problem. The problem of unknown even Fourier components is equivalent to not knowing the location of the zeroes for a given vocal tract configuration. This lack of information about the zeroes is also the primary cause of difficulties in extracting vocal tract area functions based on LPC analysis.

Unfortunately, all of the studies cited above have suffered from the incomplete nature of area functions derived from sagittal x-ray projection images. Magnetic resonance imaging (MRI), which is both safe for the subject and allows for three-dimensional imaging, is now becoming an attractive alternative to x-ray techniques for studying the spatial detail of the vocal tract. Recently, Story, Titze and Hoffman (1996) have reported a set of MRI-based area functions corresponding to 10 vowels, 2 liquids, 3 plosives, and 3 nasals for one adult male speaker. The area functions were obtained from direct measurements of 3-D vocal tract reconstructions. It is the purpose of this paper to first develop a speaker-specific parameterization of a subset of those vocal tract shapes (ten vowels: i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, o, u) using a technique that decomposes them into empirical orthogonal modes, similar to the method described by Meyer *et al.* (1989). This parameterization is then used to explore possible connections between mode shape, articulation, and acoustic characteristics. Finally, the method is extended to all 18 area functions given in Story *et al.* (1996) and results are compared with the vowel only case. It is recognized that the use of data from a single subject does not allow for any statistically significant statements to be made about a general population of speakers. However, it is of interest to understand the general phonetic characteristics *as well as* those qualities which make a speaker unique. Thus, a method which decomposes a set of vocal tract shapes for one speaker at a time retains the detail that is unique to that person. This paper is intended to lay out a method that can be used for additional sets of vocal tract shapes that will hopefully be acquired from many more subjects in the future.

2. Empirical orthogonal mode decomposition

Statistical techniques that utilize a linear orthogonal transformation to extract a low-dimensional set of prominent features from a high-dimensional input set have been used in many fields, and as a result go by several different names. Principal components analysis, Karhunen-Loeve transform, empirical orthogonal functions, or singular value decomposition are a few of the names used to identify the method. In this paper, the term "empirical orthogonal modes" will be used since it implies that basis functions are derived purely from empirical data and the term "mode" is used to emphasize a similarity to a modal decomposition of a dynamical system into natural modes. Occasionally, the terms "orthogonal modes" or simply "modes" will be used; these should be assumed to mean the same as "empirical orthogonal modes".

Decomposition of a data set into orthogonal modes transforms a high dimensional input into a low-dimensional output consisting of significant, uncorrelated features where a small number of the features account for most of the variance in the original data set. The particular method used in this study to extract modes is given in general terms in Herzel *et al.* (1995) and specifically applied to vocal fold vibratory patterns in Berry *et al.* (1994). The formulation of the method will now be given in terms that are specific to the analysis of the area function set in Story *et al.* (1996). The notation will be similar to that used by Herzel *et al.* (1995) except that the temporal dimension will be replaced by

a phoneme dimension (the term 'phoneme' is used here only to conveniently denote a collection of both vowel and consonant area functions).

In this paper, the term "area vector" will refer to the collection of vocal tract cross-sectional area values represented in an ordered vector. This contrasts with the "area function", which is defined here as the combination of the area vector and a corresponding length vector which indicates each area element's distance from the glottis. Area functions will typically be shown in figures (that is, area vectors plotted against length vectors) but the area vectors are used in the analysis.

Each area function in Story *et al.* (1996) was reported as a set of cross-sectional areas with a constant length interval of 0.396 cm; the space between data points was assumed to be represented by a cylindrical tube section. The number of sections comprising each area function was chosen to most closely represent the measured length of the vocal tract during production of a given vowel or consonant. However, to extract empirical orthogonal functions, all of the area functions must be represented as vectors of equal length. Hence, each area function was fitted with a cubic spline and resampled to be 44 sections long. In order to maintain the original vocal tract lengths, the length intervals between cross-sectional areas have now been made phoneme dependent. The resulting area vectors and length intervals (δ_l) for the ten vowels are given in Table I; the length of any given area function is now $44\delta_l$.

To perform the decomposition of the area vectors into empirical orthogonal modes, assume that any area vector in the set can be represented by a mean and a variable part,

$$A(\mathbf{x},\ p)\ =\ A_0(\mathbf{x}) + \alpha(\mathbf{x}, p) \tag{1}$$

where $A(\mathbf{x}, p)$ is the area vector for a given phoneme p, $A_0(\mathbf{x})$ is the mean area vector across the data set, and $\alpha(\mathbf{x}, p)$ is the variation that is superimposed on $A_0(\mathbf{x})$ to produce a specific area vector. The $\mathbf{x}$ denotes the index vector $[1, 2, \ldots, 44]$, where 1 represents the first section above the glottis and section 44 is at the lips. The *phoneme-dimension* is denoted by p and for the vowel only case, which will be the main focus of this paper, $p = [1, 2, \ldots, 10]$, representing the vowels ordered as [i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, o, u] (see Table I). If consonants are added to the analysis, p would be increased accordingly; e.g., for all 18 of the vowels and consonants given in Story *et al.* (1996), $p = [1, 2, \ldots, 18]$. The mean area vector $A_0(\mathbf{x})$ is computed by

$$A_0(\mathbf{x}) = \frac{1}{M}\sum_{p=1}^{M} A(\mathbf{x}, p) \quad (M = \text{number of phonemes}). \tag{2}$$

The superimposed variations are computed by subtracting the mean area vector from all the area vectors in the input data set,

$$\alpha(\mathbf{x}, p) = A(\mathbf{x}, p) - A_0(\mathbf{x}). \tag{3}$$

The elements of a real, symmetric, covariance matrix $\mathbf{R}$ of the $\alpha(\mathbf{x}, p)$'s can now be computed as

$$R_{ij} = \frac{1}{M-1}\sum_{p=1}^{M} \alpha(\mathbf{x}_i, p)\alpha(\mathbf{x}_j, p) \quad (i, j = 1, 2, \ldots, N) \tag{4}$$

where M is again the number of area vectors in the input set and N is the number of elements in each area vector ($N = 44$). A matrix of N-element normalized eigenvectors ϕ and an N-element vector of eigenvalues λ can now be computed from

TABLE I. Area vectors for 10 vowels based on Story *et al.* (1996). Each original area function has been discretized to consist of 44 area sections; the length of each section is given by δ_l. The glottal end of each area vector is at section 1 and the lip end at section 44

Section	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	0.33	0.20	0.23	0.23	0.33	0.45	0.61	0.18	0.32	0.40
2	0.30	0.18	0.13	0.26	0.28	0.20	0.28	0.17	0.39	0.36
3	0.36	0.16	0.14	0.27	0.23	0.26	0.19	0.23	0.39	0.29
4	0.33	0.19	0.19	0.17	0.15	0.21	0.10	0.28	0.43	0.44
5	0.64	0.11	0.04	0.15	0.17	0.32	0.07	0.59	0.56	0.69
6	0.46	0.67	0.26	0.14	0.33	0.30	0.30	1.46	1.46	2.15
7	1.70	1.70	1.08	0.59	0.39	0.33	0.18	1.60	2.20	3.00
8	3.14	1.64	1.26	1.31	1.02	1.05	1.13	1.11	2.06	2.72
9	2.89	1.45	1.21	1.54	1.22	1.12	1.42	0.82	1.58	2.15
10	2.45	0.97	0.96	1.06	1.14	0.85	1.21	1.01	1.11	2.48
11	2.87	0.84	0.72	0.93	0.82	0.63	0.69	2.72	1.11	4.95
12	3.71	1.90	0.74	0.67	0.76	0.39	0.51	2.71	1.26	5.91
13	3.77	2.35	0.91	1.98	0.66	0.26	0.43	1.96	1.30	5.49
14	3.92	2.97	1.64	2.25	0.80	0.28	0.66	1.92	0.98	5.05
15	4.50	3.21	1.91	2.08	0.72	0.23	0.57	1.70	0.93	4.60
16	4.44	3.37	2.20	1.90	0.66	0.32	0.32	1.66	0.83	4.41
17	4.47	3.33	2.62	2.35	1.08	0.29	0.43	1.52	0.61	3.77
18	4.71	3.61	2.77	2.92	0.91	0.28	0.45	1.28	0.97	3.39
19	4.44	3.91	2.98	3.33	1.09	0.40	0.53	1.44	0.75	3.18
20	4.15	3.82	3.00	3.76	1.06	0.66	0.60	1.28	0.93	3.29
21	4.07	3.86	2.83	3.80	1.09	1.20	0.77	0.89	0.53	3.24
22	3.51	3.47	2.84	3.69	1.17	1.05	0.65	1.25	0.65	2.33
23	2.98	3.00	2.86	3.87	1.39	1.62	0.58	1.38	0.95	2.08
24	2.10	2.65	2.44	3.73	1.55	2.09	0.94	1.09	0.99	2.04
25	1.69	2.41	2.15	3.23	1.89	2.56	2.02	0.71	1.07	1.42
26	1.44	2.07	2.10	3.24	2.17	2.78	2.50	0.46	1.39	0.62
27	1.13	1.85	1.84	3.30	2.46	2.86	2.41	0.39	1.47	0.18
28	0.72	1.82	1.77	3.21	2.65	3.02	2.62	0.32	1.79	0.17
29	0.39	1.47	1.84	3.21	3.13	3.75	3.29	0.57	2.34	0.22
30	0.33	1.49	1.72	3.24	3.81	4.60	4.34	1.06	2.68	0.25
31	0.21	1.23	1.45	3.28	4.30	5.09	4.78	1.38	3.36	0.46
32	0.10	0.91	1.37	3.62	4.57	6.02	5.24	2.29	3.98	0.71
33	0.08	0.79	1.36	3.86	4.94	6.55	6.07	2.99	4.74	0.75
34	0.27	0.88	1.43	3.86	5.58	6.29	7.08	3.74	5.48	1.33
35	0.21	1.14	1.72	4.15	5.79	6.27	6.81	4.39	5.69	2.23
36	0.26	1.48	2.08	4.52	5.51	5.94	6.20	5.38	5.57	2.45
37	0.45	1.75	2.36	4.59	5.49	5.28	5.89	7.25	4.99	3.16
38	0.21	1.95	2.66	4.77	4.69	4.70	5.04	7.00	4.48	5.16
39	0.43	1.57	2.38	4.36	4.50	3.87	4.29	4.57	3.07	4.92
40	0.77	2.09	1.95	4.36	3.21	4.13	2.49	2.75	1.67	2.73
41	1.69	1.86	2.68	4.30	2.79	4.25	1.84	1.48	1.13	1.21
42	2.06	1.60	2.61	4.55	2.11	4.27	1.33	0.68	0.64	0.79
43	2.01	1.35	2.19	4.30	1.98	4.69	1.19	0.39	0.15	0.42
44	1.58	1.18	1.60	3.94	1.17	5.03	0.88	0.14	0.22	0.86
δ_l	0.366	0.373	0.362	0.371	0.386	0.393	0.402	0.395	0.388	0.410

the covariance matrix $\mathbf{R}$ such that $\mathbf{R}\phi = \phi\mathbf{I}\lambda$ where $\mathbf{I}$ is the identity matrix. The eigenvectors can be thought of as basic "building blocks" of the original set of area vectors and henceforth will be called *empirical orthogonal modes* (or simply *modes*). The eigenvalues λ quantify the significance of each corresponding orthogonal

mode by indicating how much of the total variance it accounts for in the input data set (i.e., the set of original area vectors); in particular, each individual eigenvalue λ_i divided by the sum of all the eigenvalues will yield the percentage of variance accounted for by its corresponding mode $\phi_i(\mathbf{x})$.

A result of the decomposition of area vectors as stated above is that the resulting orthogonal modes are ordered from least significant to most significant; i.e., the 44th mode accounts for the largest amount of variance. However, it is intuitively more desirable for the most significant mode to be ordered such that it is "first" and the second most significant mode to be ordered "second", etc. Thus, the eigenvector (modal) matrix ϕ has been reordered to accommodate this desire and the remainder of the analysis shown below reflects this reordering.

Defining $C_i(p)$ as the amplitude coefficient of the i^{th} mode corresponding to the p^{th} area function, the superimposed variation characterizing each area vector can be computed by

$$\alpha(\mathbf{x}, p) = \sum_{i=1}^{M} C_i(p)\,\phi_i(\mathbf{x}). \tag{5}$$

The amplitude coefficients are obtained by projecting the original data set onto the set of modes (normalized eigenvectors, $\phi(\mathbf{x})$),

$$C_i(p) = \sum_{j=1}^{N} \alpha(x_j, p)\,\phi_i(x_j) \quad (i = 1, 2, \ldots, N). \tag{6}$$

Once the coefficients have been computed, area vectors can be reconstructed by combining Equations 5 and 1 to get

$$A(\mathbf{x}, p) = A_0(\mathbf{x}) + \sum_{i=1}^{N} C_i(p)\,\phi_i(\mathbf{x}) \quad (N = 44). \tag{7}$$

The corresponding length vector, required to reconstruct the area function, is generated by

$$L(\mathbf{x}) = \mathbf{x} \cdot \delta_l(p). \tag{8}$$

Equation 7 will "perfectly" reconstruct all of the original area vectors because all of the derived modes are used (i.e., $N = 44$). However, if just a few of the low-ordered modes account for nearly all of the variance, the reconstruction of the area vectors can be performed by summing over less than N modes with a loss of only a small amount of the original fidelity. This provides a compression of the original area functions from N cross-sectional areas to much less than N modal amplitude coefficients. Thus, N in Equation 7 may be replaced with an N' where $N' \ll N$, assuming that N' modes account for a sufficient amount of the variance.

Experimentation with area vectors reconstructed with a small number of modes (e.g., $N' = 4$) showed them to be closely matched to the originals (the next section will discuss this further). However, due to the incomplete nature of the reconstruction (i.e., use of only four modes), areas in the highly constricted regions of the vocal tract often dipped slightly below zero, which is obviously unrealistic. Just as sharp temporal features in an acoustic waveform are composed of many harmonics, sharp spatial features in an area function, such as a tight constriction, must be represented by many modes.

To avert this problem, a square root pre-processing operation was applied to each original area vector prior to performing the modal decomposition laid out in Equations 1–7. This effectively converts each area into the radius of a circular cross-section, except

with a scaling factor of $\sqrt{\pi}$, and has the effect of expanding the highly constricted regions and compressing those that are more open. Thus, taking the square root of each area element transforms an area vector into a "radius"-like vector and the modal decomposition is performed in this $\sqrt{\text{area}}$ domain. Area vectors are still reconstructed as prescribed by Equation 7 except that the final result must be squared to convert it back to area. The final squaring operation also guarantees that negative areas will not be produced. The square root operation will be used throughout the remainder of this paper. It should be noted that a base ten logarithm was also tested as possible pre-processing operation on the area vectors. The area vectors were accurately reconstructed in the highly constricted portions of the vocal tract and also showed no negative areas. There was, however, more reconstruction error in the larger area portions of the tract.

2.1. *Decomposition and reconstruction of area functions*

As mentioned previously, the eigenvalues resulting from the decomposition of area vectors into orthogonal modes specify the amount of variance accounted for by each mode. However, since the decomposition was performed after taking the square root of each cross-sectional area, these variances correspond to the $\sqrt{\text{area}}$ domain. Of greater interest is to know how much variance each mode accounts for after the *area* vectors are created by the final squaring operation. To do this, the variance around the mean of each element of the original ten area vectors was computed as

$$\sigma^2(x_i) = \sum_{p=1}^{M} \alpha(\mathbf{x}_i, p)^2 \quad (i = 1, 2, \ldots, N) \tag{9}$$

(α was defined in Equation 3) so that the total variance is

$$\sigma_{tot}^2 = \sum_{i=1}^{N} \sigma^2(\mathbf{x}_i). \tag{10}$$

Now the area vectors are reconstructed with only the first mode (i.e., Equation 7 is used with $N = 1$ and the final result squared) and the total variance around the mean is computed just as was done with Equations 9 and 10, thus generating a σ_1^2 (analogous to σ_{tot}^2 in Equation 10). The percentage of the variance accounted for by the first mode is

$$\text{one mode \% of variance} = 100\,\frac{\sigma_1^2}{\sigma_{tot}^2}. \tag{11}$$

Similarly, the cumulative variance accounted for by any number of modes is determined by using Equation 7 with $N = 1, 2, 3, \ldots, 44$ and computing the total variance around the mean (Equations 9 and 10). This gives a cumulative σ_N^2 from which the percentage of variance accounted for by N modes is

$$N \text{ mode \% of variance} = 100\,\frac{\sigma_N^2}{\sigma_{tot}^2}. \tag{12}$$

Table II shows the cumulative percentages of variance accounted for by the first ten modes. Percentages are shown from both the eigenvalues of the original decomposition

TABLE II. Cumulative percentage of variance for the first ten empirical orthogonal modes derived from 10 vowel area vectors. The first column is the mode number while the second and third columns give percentage of variance based on the $\sqrt{\text{area}}$ decomposition and the reconstructed area vectors, respectively

Mode	% of var. ($\sqrt{\text{area}}$)	% of var. (area)
1	66.90	74.67
2	87.90	92.63
3	94.31	95.41
4	96.81	96.32
5	98.39	98.97
6	99.23	99.83
7	99.71	99.83
8	99.81	99.99
9	100.00	100.00
10	100.00	100.00

in the $\sqrt{\text{area}}$ domain and reconstructed area vectors (using Equations 9–12). Interestingly, the conversion to area from $\sqrt{\text{area}}$ increases the percentage of variance accounted for by the first two modes; however, the third through the tenth modes are about the same for both cases. It is also observed that nine modes effectively account for all of the variance in the original ten vowels. This means that only nine modal amplitude coefficients (C_i's) would be required to completely reconstruct each area vector. However, since just four modes capture over 96% and only two modes still account for over 92% of the total variance, highly accurate area vector reconstructions may be achieved with 2–4 modal amplitude coefficients. This would provide a compression of the original 44-element area vectors ranging from 11 : 1 for 4 modes to 22 : 1 for 2 modes. The four most prominent orthogonal modes (which are 44-element vectors) and the mean vector ($\sqrt{\text{area}}$) of the ten vowels in Table I are shown with solid lines in Fig. 1; the dotted lines represent the reflection of each mode across the zero axis. Vowel-dependent modal amplitude coefficients (C_i's) corresponding to these four modes are shown in Table III. Any given vowel can be reconstructed by multiplying each mode shape vector by the appropriate modal coefficient, summing them element-by-element with the mean vector, and finally squaring each element to generate the area vector. The length vector is produced by multiplying the δ_l for a particular vowel (from Table I) by the vector $[1, 2, \ldots, 44]$. The area function is represented as the area vector plotted against the length vector.

Fig. 2 shows reconstructions of the ten vowels using four mode shapes and the coefficients given in Table III. Each reconstructed vowel is shown with a thick line, while the original is shown with a dashed. A measure of the error between the reconstructed and original vowels is shown in the upper left corner of each plot and was calculated as $\varepsilon = \sum_{i=1}^{44} |A(i)^{orig} - A(i)^{recon}|$. In all cases the reconstructed area functions appear to be closely matched to the originals, with the exception of some fine detail. Interestingly, as indicated by error values, not all of the area functions are reconstructed with the same fidelity relative to their original counterpart. The [i] and [ʌ] have the least error while [ʊ] and [o] have the most.

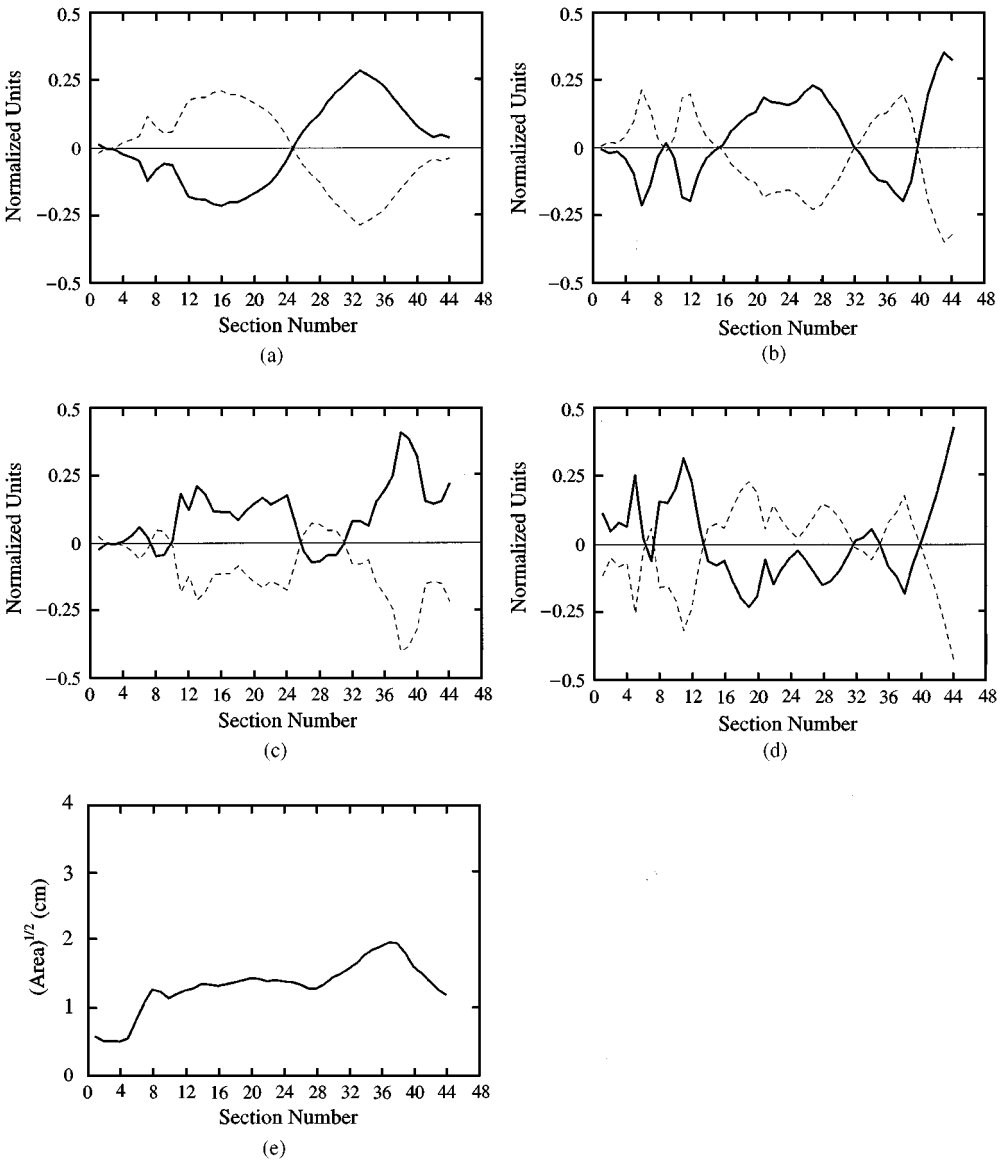


Figure 1. The four most prominent empirical orthogonal modes and the square root of the mean area vector derived from the 10 vowel set listed in Table I (the dashed lines are the reflection of each mode about the zero axis): (a) mode 1, (b) mode 2, (c) mode 3, (d) mode 4 and (e) square root of the mean area vector.

Figure 2. Reconstructions of the ten vowel area functions using the first *four* modes (—) along with the original area functions (- - -); a measure of error relative to the original area function is shown in the upper left forner of each plot and was calculated by $\varepsilon = \sum_{i=1}^{44} |A(i)^{orig} - A(i)^{recon}|$.

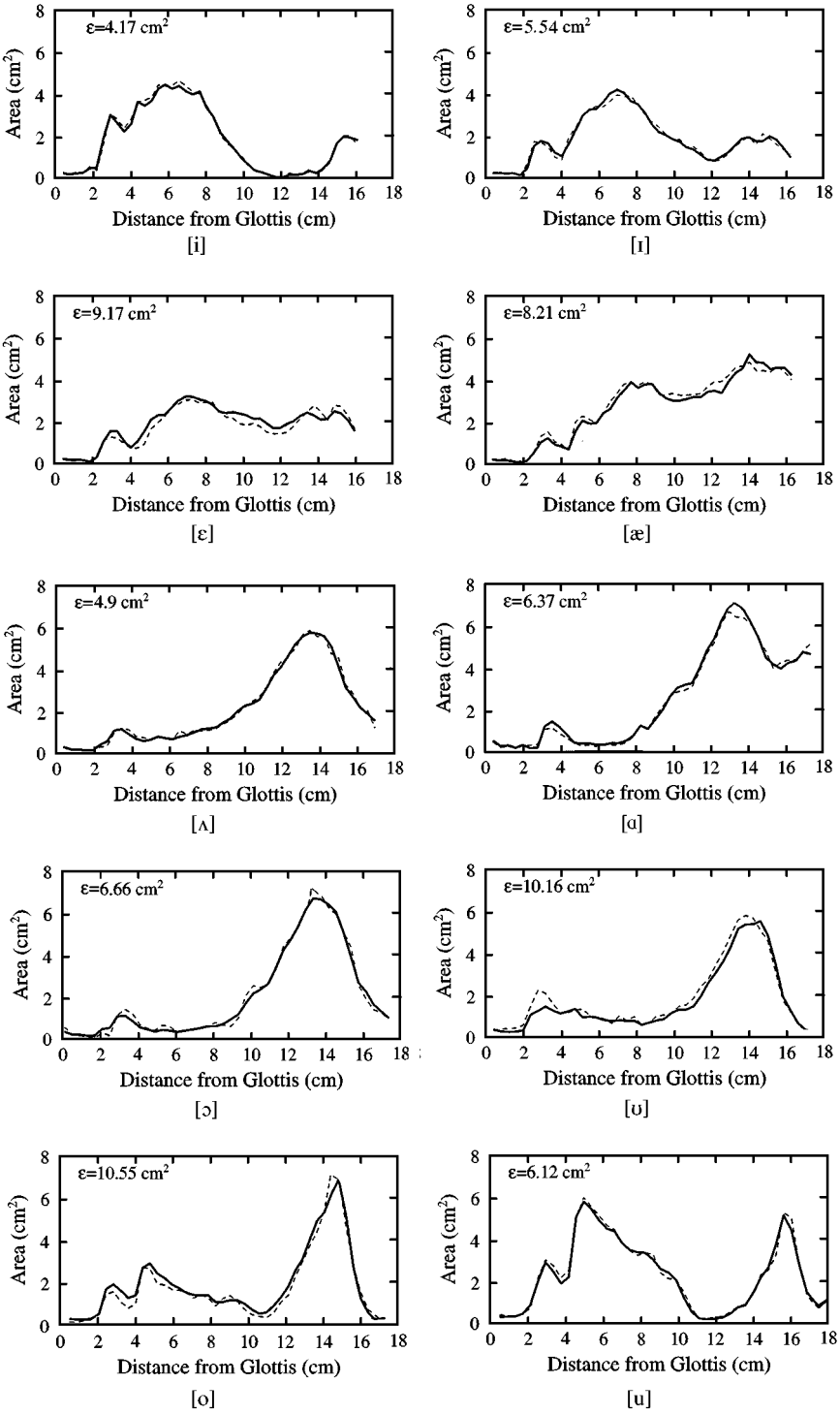

ε=4.17 cm²
Area (cm²)
Distance from Glottis (cm)
[i]
ε=5.54 cm²
[ɪ]
ε=9.17 cm²
[ɛ]
ε=8.21 cm²
[æ]
ε=4.9 cm²
[ʌ]
ε=6.37 cm²
[ɑ]
ε=6.66 cm²
[ɔ]
ε=10.16 cm²
[ʊ]
ε=10.55 cm²
[o]
ε=6.12 cm²
[u]

TABLE III. Modal amplitude coefficients corresponding to ten vowels

Mode	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	− 4.5868	− 2.2662	− 0.9671	0.6019	2.2705	3.4088	3.0793	1.5314	0.0892	− 3.1610
2	0.8701	0.9242	1.1817	2.0505	0.1111	1.2017	− 0.4369	− 1.6899	− 2.3804	− 1.8320
3	− 1.0200	− 0.3415	− 0.3772	1.4780	− 0.0643	0.1532	− 0.6305	− 0.8610	0.4413	1.2219
4	0.7094	− 0.7028	− 0.4845	− 0.3044	− 0.1079	0.9708	− 0.0853	− 0.2056	− 0.1372	0.3477

Vowel reconstructions using only two mode shapes are shown in Fig. 3. As expected, the reconstructed area functions are not matched as closely to the originals as with four modes but they still reasonably approximate the original vowels. The error for [ɪ, ɛ, ʌ, o] has increased only slightly over the previous case, while more significant error increases are observed for the other vowels.

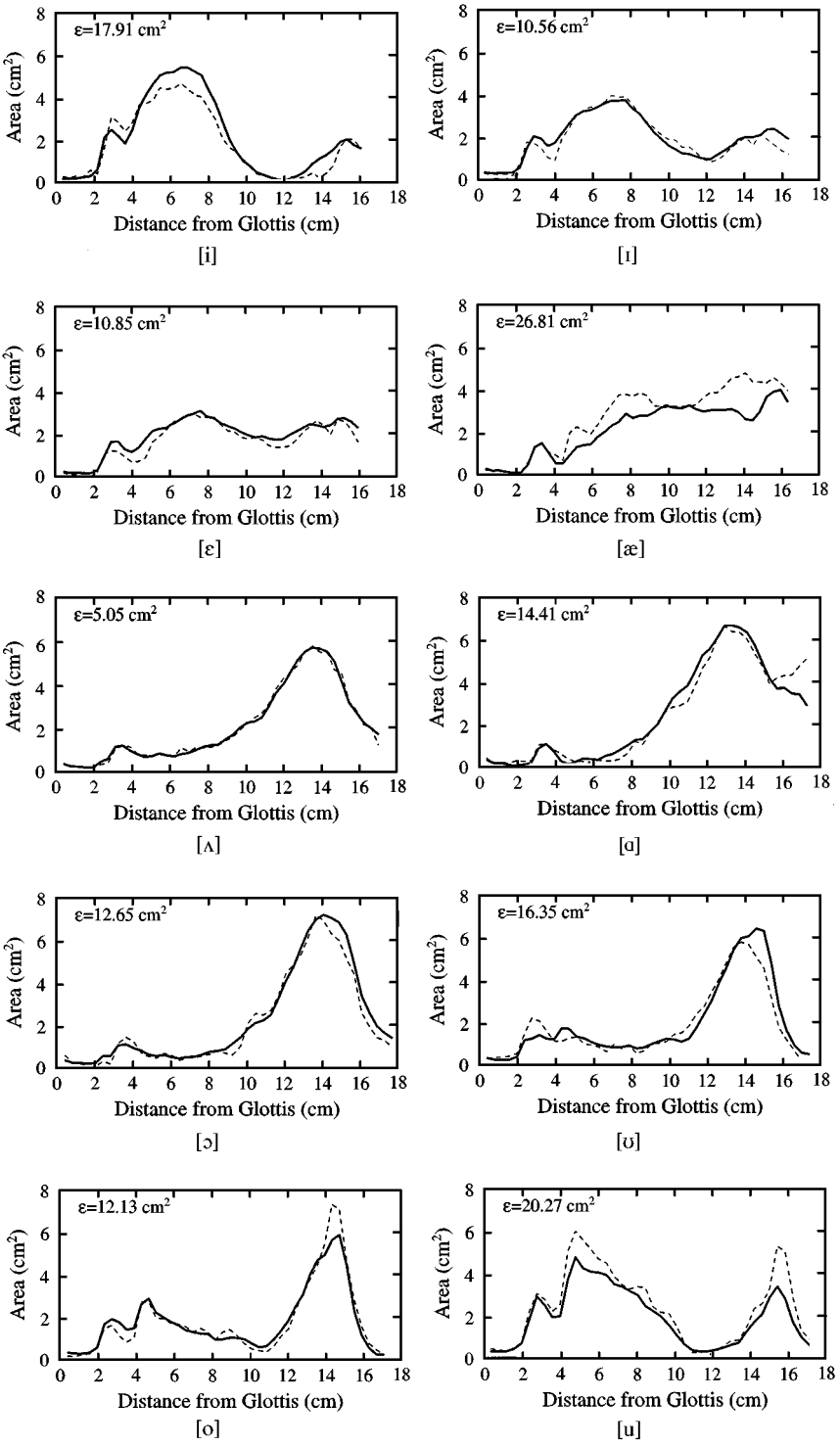
2.2. *Articulatory interpretation of mode shapes*

While the previous section discussed the modal decomposition simply as a method of parameterizing and compressing the area functions, it is also of interest to investigate a possible articulatory interpretation of the empirical orthogonal modes. Since the modal decomposition extracts the most prominent features or patterns from the input data set, it is likely that the most prominent modes contain significant articulatory information.

The first mode shape in Fig. 1 accounts for nearly 75% of the total variance in the area vector set (see Table II), which makes it, by far, the most prominent mode. It has a back-to-front asymmetry that, for a negative amplitude coefficient, would largely replicate the forward and upward movement of the tongue and some upward jaw movement (recall that the area vectors are the sum of the mean area vector and the mode shapes; thus, a negatively valued portion of a mode shape reduces the cross-sectional area). Equivalently, a positive modal amplitude coefficient would signify a backward and downward tongue movement and a dropping of the jaw. Thus, it is not surprising that the coefficients for [ɑ] and [i] (the most extreme front and back vowels) given in Table III have the largest positive and negative values, respectively, for the first mode in the set. However, because each mode encompasses the entire length of the vocal tract, the structure of the tongue and jaw cannot account for the complete shape of the first mode. For example, the shape of the region between 0 and 5 cm (from the glottis) is due to the lower pharyngeal structure such as the epilaryngeal tube and epiglottis.

The second mode (Fig. 1b), which accounts for 18% of the total variance, crosses the zero axis several times and allows for tract variations in areas where the first mode has diminished amplitude, as would be expected for orthogonal modes. Because it can affect a large region in the middle of the vocal tract, this mode may capture the up-down motion and possible arching of the tongue. Note that the coefficients for the second mode in Table III have large negative values for the vowels [ʊ, o, u], all of which have a mid-tract constriction. The [æ] has a large positive coefficient for the second mode combined with a low-valued first mode coefficient to create a expansion that is farther back than for [ɑ]. Additionally, the region of the second mode between 0 and 5 cm defines the shaping of the epilarynx and the lower pharynx with more detail than mode 1. At the lip end of the vocal tract, mode 2 can exert much more influence than can mode 1, indicating that it may also contain shaping of the vocal tract corresponding to lip rounding and spreading. Again, note that the large negative coefficient values of mode 2 for the vowels [o, ʊ, u], also generate lip rounding as well as a mid-tract constriction.

Any region in which the amplitude of the mode shape is near zero represents a portion of the vocal tract that changes very little across vowels. With regard to modes 1 and 2, such a region exists from about 0–1.5 cm above the glottis. Both modes have amplitudes close to zero indicating that the epilaryngeal section does not change much across the vowels. This can also be seen in Figs. 2 and 3 where all ten vowels are plotted.


ε=17.91 cm²
Area (cm²)
Distance from Glottis (cm)
[i]
ε=10.56 cm²
[ɪ]
ε=10.85 cm²
[ɛ]
ε=26.81 cm²
[æ]
ε=5.05 cm²
[ʌ]
ε=14.41 cm²
[ɑ]
ε=12.65 cm²
[ɔ]
ε=16.35 cm²
[ʊ]
ε=12.13 cm²
[o]
ε=20.27 cm²
[u]

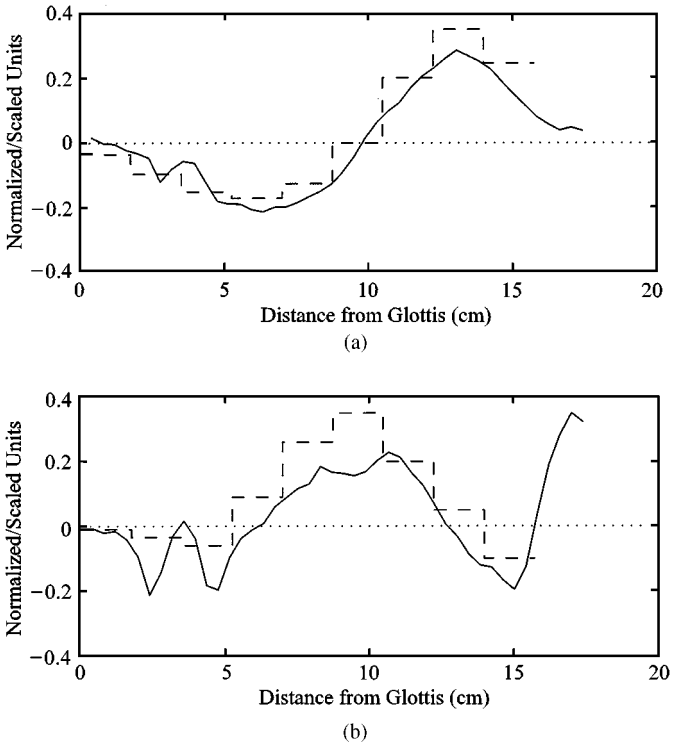


Figure 4. Comparison of modes 1 and 2 (—) with rescaled versions of those reported by Meyer *et al.* (1989) (- - -): (a) mode 1, and (b) mode 2.

It is of interest to compare the first two modes in the present analysis to the two modes (or as they called them, components) derived by Meyer *et al.* (1989). The first two components from Meyer *et al.* (1989) have been reconstructed from their Fig. 1 (p. 524) and are shown with modes 1 and 2 from this study in Fig. 4. They subjected only nine segments of each ten-segment area function to their eigenfunction decomposition; the lip section was defined as a separate, isolated parameter. Hence, only nine sections of their components can be shown in each plot; the section length for each of the Meyer *et al.* (1989) modes was 1.75 cm, giving a total vocal tract length of 17.5 cm. To maintain a similar length, the two modes shown from the present study are plotted with length vector **L** based on $\delta_l = 0.397$ cm (i.e., 0.397*44 = 17.5 cm). Also, the Meyer *et al.* (1989) components have been multiplied by -1 and scaled to have similar amplitudes to those from the present study. In both studies, modes 1 and 2 are similarly shaped. The zero crossings for each mode occur in nearly the same location across studies, but with the obvious difference in spatial resolution. It is also apparent that the large positive portion

Figure 3. Reconstructions of the ten vowel area functions using the first *two* modes (—) along with the original area functions (- - -); a measure of error relative to the original area function is shown in the upper left corner of each plot and was calculated by $\varepsilon = \sum_{i=1}^{44} |A(i)^{orig} - A(i)^{recon}|$.

at the lip end of mode 2 (approximately 16 cm to 17.5 cm) from this study is due to lip motion, since the Meyer *et al.* (1989) components did not include the lips.

The third and fourth modes correspond mainly to higher spatial frequency detail or the fine structure of the area vector shape, which makes it difficult to speculate on the possible articulatory connection to these modes. Their respective coefficients in Table III indicate their diminished significance for reconstructing the vowel shapes.

3. Acoustic considerations

In this section, the acoustic characteristics of the mean area function will be first examined. Then the effect of adding increasing amplitudes of the mode shapes to the mean area function will be studied in terms of formant perturbations.

3.1. *Significance of the mean area function*

It has often been the case that theoretical analyses of the vocal tract formant structure begin by considering a uniform tube, closed at the glottis and open at the mouth, with formants spaced according to

$$F_n = \frac{(2n-1)c}{4L} \quad (n = 1, 2, 3, \ldots)$$

which for an ideal, lossless tube of length $L = 17.5$ cm and a speed of sound $c = 350$ m/s will generate formants located at 500, 1500, 2500, … Hz. The uniform tube is assumed to produce the acoustic characteristics of a neutral [ə] vowel. The tube is then systematically perturbed to shift the formants up or down in frequency to generate other formant structures (Fant, 1960; Schroeder, 1966; Mermelstein, 1967; Mrayati *et al.*, 1988). Fig. 5a shows a 1.0 cm^2 uniform tube along with the mean area function of the ten vowels in Table I (this is obtained by squaring the mean vector in Fig. 1e and plotting it against a length vector where $\delta_l = 0.397$ cm). The frequency responses of both area functions, computed with a frequency domain transmission line technique (Sondhi & Schroeter, 1987), are shown in Fig. 5b. The formant peaks for the mean area function occur at similar locations to those of the uniform tube. The first four formants generated by the mean area function show an upward shift in frequency relative to those of the uniform tube, while the fifth formant is shifted down in frequency. Formant frequencies and percent differences are given in Table IV for the two cases. Note that the uniform tube formants do not occur at exactly 500, 1500, 2500, … Hz, but are shifted because of the inclusion of frequency dependent losses (e.g., yielding walls and radiation impedance). The first formant is the worst match, with an 18.9% difference. Both F2 and F3 show about a 2% difference, while F4 differs by 4.8%.

The characteristic of the mean area function to produce formants similar to a uniform tube is a demonstration of how two (and theoretically more) different tube shapes can produce essentially the same formant structure. The constriction of the mean area function in the short region above the larynx is countered with an expansion at the opposite end of the tract, keeping the formants in nearly the same locations as the uniform tube. Thus, due to anatomical constraints, the physiological neutral vowel shape is a deformed tube, but still produces the appropriate neutral vowel formant structure. In light of this, the mean area function can be considered an approximate neutral vowel

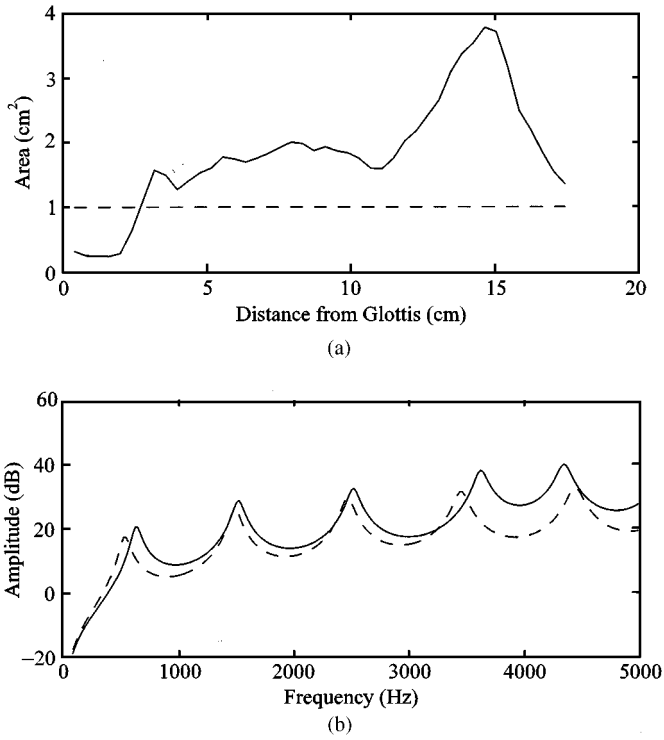


Figure 5. Comparison of the mean area function with a uniform tube; the vocal tract length was 17.5 cm for both cases: (a) mean area function (—) and 1 cm^2 uniform tube (- - -), and (b) frequency response of the mean area function (—) and 1 cm^2 uniform tube (- - -).

TABLE IV. First four formant frequencies of a 1 cm^2 uniform tube and the mean area function based on 10 vowels; both had a length of 17.5 cm. Percent differences between the two sets of formant frequencies are also shown

Tract shape	F1	F2	F3	F4
uniform tube	528	1482	2456	3436
mean area function	628	1510	2506	3606
% diff	18.9	2.0	2.0	4.9

configuration and other vowels are produced by imposing perturbations on it. These perturbations were quantified (for one speaker) in Section 2.1 with the decomposition of the area vector set into orthogonal modes.

A comparison of the Fourier series representation of the area function suggested by Schroeder (1966) and the modal representation presented in this paper show an interesting similarity. Following Schroeder (1966), the area function can be represented by

$$A(\mathbf{x}) = A_0 + A_0 \sum_{m=1}^{M} p_m \cos\left(\frac{(2m-1)\pi \mathbf{x}}{l}\right), \tag{13}$$

in which p_m is the coefficient determining the magnitude of the m^{th} odd Fourier component, M is the number of formants, and l is the vocal tract length. In this equation, the area function $A(\mathbf{x})$ is the sum of the uniform tube cross-sectional area A_0 and the scaled (by A_0) sum of the odd Fourier cosine series (a standard set of orthogonal basis functions) over the number of formants extracted from a speech waveform spectrum; A_0 is a scalar value, since the area along the length of the tube was constant. Inclusion of even cosine terms in Equation 13 will not significantly effect the locations of the formants but will greatly alter the resulting area function. Thus, in theory, an infinite number of area function shapes could generate the same formant structure. By comparison, the empirical orthogonal mode representation given by Equation 7 (recall $A(\mathbf{x}, p) = A_0(\mathbf{x}) + \sum_{i=1}^{N} C_i(p)\,\phi_i(\mathbf{x})$), also generates an area function by the sum of a constant and the sum of a set of scaled basis functions. The difference is that the constant, $A_0(\mathbf{x})$, is a spatially varying area vector and the basis functions are the empirically-based orthogonal modes.

In summary, Equation 13 (from Schroeder) is based on the *theoretically-derived acoustical* possibilities of deforming a uniform tube, whereas Equation 7 is based on the *empirically-derived physiological* possibilities of deforming the neutral vocal tract shape. When reconstructing the original ten vowels, Equations 7 and 8 are automatically constrained to generate physiologically realistic area functions. While it is not possible to say whether area functions generated by arbitrary combinations of modal coefficients (i.e., not corresponding to any of the ten original vowels) would be part of a speaker's physiologic repertoire, it is likely that the use of empirical orthogonal modes would give more realistic area functions than a mathematical basis function set (e.g., Fourier series), especially if the coefficients C_i are not extended beyond the maximum and minimum values obtained for the original ten vowels. It is noted, however, that the epilaryngeal region of the vocal tract (1.5–2 cm above the glottis) is definitely constrained by the subject's physiology, since the amplitude of all the mode shapes in this region is near zero (see Fig. 1). This ensures that this region will have small cross-sectional areas (similar to those of the mean area vector in this region) regardless of the modal coefficient values.

3.2. *Acoustic effects of the mode shapes*

In the previous section, the mean area function was shown to have a formant spectrum similar to that of a uniform tube. In this section we will examine the displacement of the mean vowel formant locations due to the superposition of varying amplitudes of modes 1 and 2 upon the mean area function.

First, the effect of increasing and decreasing the amplitude coefficients of mode 1 and 2 in isolation was studied; mode 1 was varied while mode 2 was held at zero amplitude and vice versa. Additionally, all other modes were set to zero amplitude and the vocal tract length was held constant at 17.5 cm (i.e., $L(\mathbf{x}) = 0.397\mathbf{x}$). For each increment of the coefficient values, the area vector was generated by Equation 7 and its frequency response function was again computed with a transmission line method (Sondhi & Schroeter, 1987). The first three peaks in each frequency response function were extracted by peak-picking and formant frequencies were determined with a parabolic interpolation (Titze, Horii & Scherer, 1987). Fig. 6a shows the frequency locations of F1, F2, and F3 as a function of amplitude coefficient for mode 1. The value of mode 1 ranges from the most negative value in Table III up to the most positive. F1 and F2 change in a nearly monotonic fashion, but in different directions, as functions of the first modal

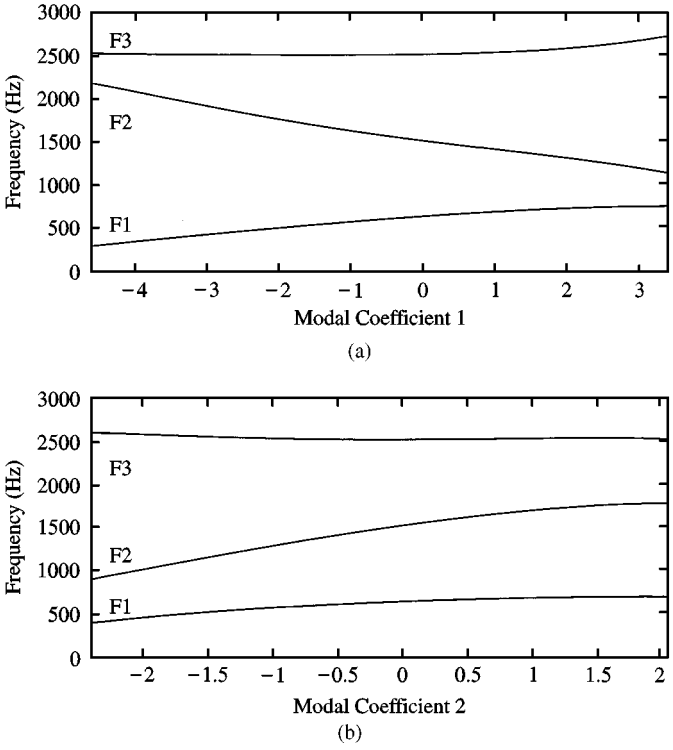


Figure 6. Formant frequencies (F1, F2, and F3) as a function of the modal coefficients: (a) modal coefficient 1 was varied while modal coefficient 2 = 0, and (b) modal coefficient 2 was varied while modal coefficient 1 = 0.

coefficient. F1 begins at a value of 300 Hz and rises to 746 Hz, while F2 begins at 2195 Hz and decreases to 1130 Hz. The third formant remains reasonably flat between coefficient values of -4.58 to $+1$ and then shows a slight increase in frequency out to the final value. Results of varying the mode 2 coefficients, while holding mode 1 at zero amplitude, are shown in Fig. 6b. Again, the coefficients range from their most negative value to their most positive value. The first formant shows a similar trend of increase from negative to positive coefficient values as in Fig. 6a. However, the second formant has almost exactly the opposite trend for the varying mode 2 coefficient as for the mode 1 coefficient. The third formant also shows nearly an opposite trend to that seen in Fig. 6a, but the effect is more subtle than for F2. These tests show that both modes 1 and 2 similarly affect the location of F1, but oppositely affect F2 (and to some degree F3).

To further investigate the effect of each mode on the resulting formant frequencies, sensitivity functions for the first three formants of the mean area function were computed. The sensitivity of a particular formant is defined as the difference between the kinetic energy (KE) and potential energy (PE) divided by the total energy (Fant & Pauli, 1975)

$$S_i(\mathbf{x}) = \frac{KE_i(\mathbf{x}) - PE_i(\mathbf{x})}{TE_i(\mathbf{x})} \quad (i = 1, 2, 3, \ldots) \tag{14}$$

where $\mathbf{x}$ is the index vector $[1, 2, \ldots, 44]$ and i is the formant number. The sensitivity function can then be used to compute the change in a particular formant frequency (F_i) due to perturbation of the area function (ΔA) with the relation,

$$\frac{\Delta F_i}{F_i} = \sum_{n=1}^{N} S_i(\mathbf{x}) \frac{\Delta A(\mathbf{x})}{A(\mathbf{x})} . \tag{15}$$

This equation says that, if the sensitivity function is positive valued and the area perturbation is also positive (area is increased), the change in formant frequency will be upward (positive). If the area change is negative (area decreased), the formant frequency will decrease. When the sensitivity function is negative, the opposite effect occurs for positive or negative area perturbations.

The sensitivity function for the first formant is shown superimposed onto the first two mode shapes in Fig. 7a. From 0 cm to 11 cm, the sensitivity function is negatively valued, meaning that expanding this portion of the mean area function would move F1 downward in frequency. Conversely, from 11 cm to the lips, an expansion in the mean area function will raise F1. In the same two portions of the vocal tract (i.e., 0 to 11 cm and 11 cm to the lips), mode 1 maintains nearly the same polarity as the F1 sensitivity function. This means that a positive valued modal amplitude coefficient for mode 1 will decrease the cross-sectional area where the sensitivity is negative, and increase it where the sensitivity is positive, thus raising F1. Recall that [ɑ], which typically has a high F1 around 700 Hz, also has a large positive coefficient for mode 1 (see Table III) and [i] has a large negative mode 1 coefficient to generate its characteristically low F1 of about 300 Hz. The second mode maintains the same polarity as the F1 sensitivity function from 0 cm to about 6.5 cm (except for a small positive value at 3.5 cm) and then has opposite polarity out to 10 cm. From 10 cm to the end of the vocal tract, mode 2's polarity with respect to the F1 sensitivity function oscillates. The net effect on F1 due to a positively valued mode 2 coefficient would be a raising of F1 but with less effectiveness than mode 1. This trend can be seen in Fig. 6b where F1 rises with increasing values of the second modal coefficient.

The F2 sensitivity function is given in Fig. 7b along with the shapes of the first and second modes. Except for the region from 0 cm to 5 cm, mode 1 has mostly opposite polarity to the sensitivity function, while mode 2 is primarily of the same polarity. Thus, positive valued coefficients for mode 1 tend to lower F2, while positive coefficients for mode 2 will raise F2; the same trend was observed in Figs. 6a and b.

To condense these observations, the normalized correlation (at zero lag) of each mode with each sensitivity function was computed and results are shown graphically in Fig. 8. Mode 1 has a positive correlation of 0.53 with the F1 sensitivity function and a negative correlation of -0.56 with the sensitivity function for F2. The F3 sensitivity was also computed but found to be nearly uncorrelated with either modes 1 or 2. Mode 2 is positively correlated to F1 sensitivity with a value of 0.36, and also to F2 sensitivity, with a value of 0.65. The largest correlation value for both modes occurred for F2, even though they have opposite signs. Thus, with nearly equal valued modal amplitude coefficients, mode 1 and 2 can have almost the same effect on F2, but in opposite directions.

Up to this point we have ignored the effect of vocal tract length, which has been held constant at 17.5 cm throughout this section. Repeating the above tests with a shortened or lengthened vocal tract produces similar relationships between formant patterns and

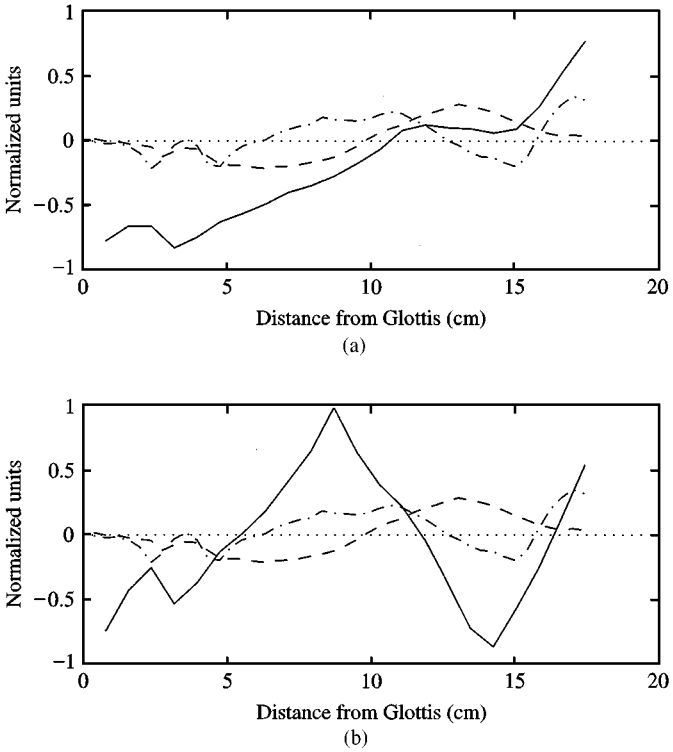


Figure 7. Comparison of sensitivity functions with mode shapes: (a) F1 sensitivity with modes 1 and 2, and (b) F2 sensitivity with modes 1 and 2. S_1 (—); mode 1 (- - -); mode 2 (·-·-·-).

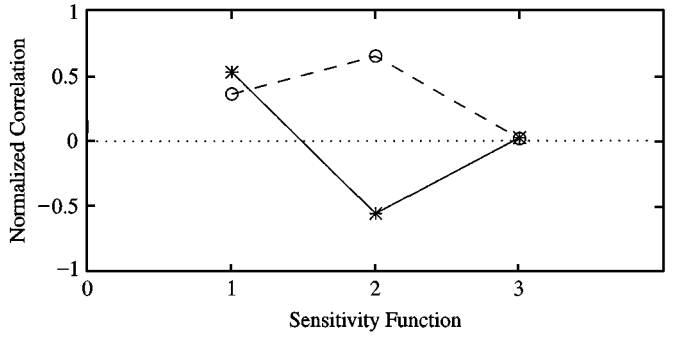


Figure 8. Normalized correlations of mode shapes with sensitivity functions: mode 1 (—) and mode 2 (- - -).

modal coefficients, except that the absolute values of the formants are raised or lowered accordingly.

4. Mapping from formants to modal coefficients

It was shown in Section 2.1 that the first two modes contain the majority of the variance (~92%) in the original area vector set (Table II) and that reconstructed area vectors

based on only the first two modes generate reasonable approximations to the original vowels (Fig. 3). Additionally, the previous section has indicated a strong correlation between these first two modes and the first two formant frequencies. The results suggest that the mode/formant relationship could be exploited to create a mapping between modal coefficients and formant frequencies. Toward this goal, a two-dimensional grid of mode 1 and 2 amplitude coefficients was generated by choosing the same maximum and minimum bounding values for each mode as for Fig. 6, and pairing 48 incremental values between them while all other modal amplitude coefficients are set to zero. This is shown mathematically by

$$\Delta c_1 = \frac{C_1^{max} - C_1^{min}}{M - 1} \tag{16a}$$

$$\Delta c_2 = \frac{C_2^{max} - C_2^{min}}{M - 1} \tag{16b}$$

$$c_1(i) = C_1^{min} + i\Delta c_1 \qquad i = 1, \ldots, M - 1 \tag{17a}$$

$$c_2(i) = C_2^{min} + j\Delta c_2 \qquad j = 1, \ldots, M - 1 \tag{17b}$$

where Δc_1 and Δc_2 are the coefficient increments, C_1^{max}, C_1^{min}, C_2^{max}, and C_2^{min}, are the maxima and minima of the mode 1 and mode 2 coefficients derived for the original ten vowels, and $M = 50$. The index i represents the increment of the mode 1 coefficient, while j is the increment of the mode 2 coefficient. Vocal tract length is again held constant, but at the mean length of the original vowels which is 16.90 cm ($\delta_i = 0.384$ cm). The result is a 50×50 (2500 point) grid shown as a 2-D mesh in Fig. 9a; the intersecting points of each horizontal and vertical line represent a modal amplitude coefficient pair.

The coefficient pairs corresponding to each of the original ten vowels are shown with solid dots and are connected to indicate their clockwise order (i.e., the upper left dot represents [i] while the dot at the lower left represents [u]). Based on Equations 7 and 8, each pair of coefficients in the grid was used to generate an area function with

$$A(\mathbf{x}, i, j) = A_0(\mathbf{x}) + c_1(i)\,\phi_1(\mathbf{x}) + c_2(j)\,\phi_2(\mathbf{x}). \tag{18}$$

and

$$L(\mathbf{x}) = 0.384\mathbf{x}. \tag{19}$$

The frequency response function of each area function was computed by the same frequency domain method used in Section 3 and the locations of F1 and F2 were similarly extracted from the frequency spectrum. The F1–F2 pairs corresponding to each modal coefficient pair were plotted in the F2 versus F1 plane, generating the deformed grid shown in Fig. 9b. As in the coefficient grid, the F1–F2 pairs computed for the original ten vowels are shown with solid dots. Each line connecting formant pairs in this grid represents a constant value of either the first or second modal coefficient; i.e., each line is an "iso-coefficient" line. For example, the bottom horizontal line in Fig. 9a, representing a c_2 value of -2.38 and varying values of c_1, corresponds to formant pairs along the bottom, horizontally-oriented, curve in Fig. 9b. Thus, the leftmost coefficient pair along this line $(c_1, c_2) = (-4.58, -2.38)$ leads to the leftmost formant pair along its respective curve $(F1, F2) = (233, 650)$ Hz and at the right, $(c_1, c_2) = (3.41, -2.38)$ gives $(F1, F2) = (469, 747)$ Hz.

The effect of the mapping is that many (2490, in this case) area functions and consequent formant combinations have been created that were not in the original ten

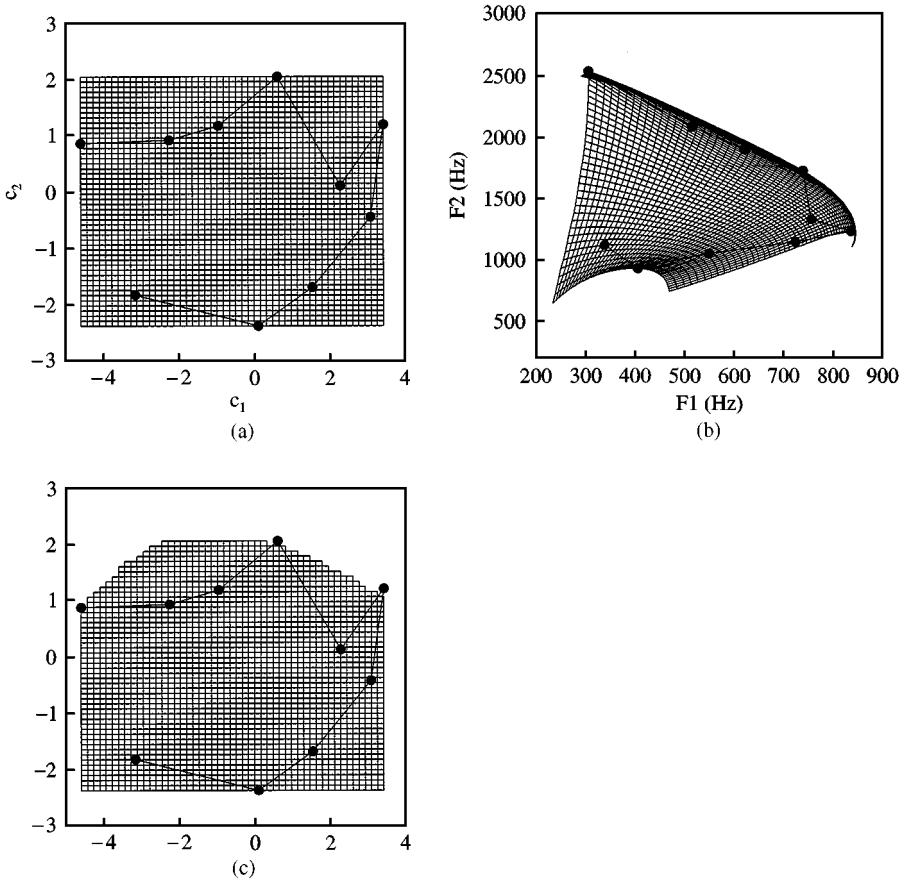


Figure 9. Mapping of mode 1 and mode 2 coefficient pairs to F1–F2 pairs, where the dots represent the coefficients and formant frequencies of the original ten vowels: (a) modal coefficient grid (b) corresponding F1–F2 grid and (c) modified coefficient grid when those coefficients generating the saturated portion of the F1–F2 grid are removed.

vowel set. In a visual, qualitative sense, the F1–F2 plane is created by deforming the modal coefficient grid such that the upper right-hand corner is pulled down and to the right, which stretches the upper portion of the grid, and at the same time a compression pushes the lower right-hand corner upward and to the left. The upper boundary of the grid shows a saturation in the form of an apparent folding of the F1–F2 pairs, so that several pairs of coefficients corresponding to the boundary would produce nearly the same formant locations for F1 and F2. This saturation region corresponds to large positive c_2 coefficients and is most prevalent when c_1 is close to its most negative or most positive values. Fig. 9c shows a reformatted coefficient grid where the coefficient pairs in the saturation region have been removed, suggesting that the formant saturation could be eliminated by plotting only those formant pairs which have a correspondence in this reformatted grid.

With the exception of the saturation region, the F1–F2 grid represents a one-to-one mapping between F1 and F2 formant frequencies and modal coefficients (or equivalently

an area function created by the coefficients). This suggests that an utterance consisting of connected vowels could be analyzed to extract F1 and F2 as functions of time, and then could be mapped back to the modal coefficient grid, and consequently to area functions. Hence, a time-dependent series of physiologically realistic area functions could be obtained from the acoustic speech waveform. However, a limitation of such a mapping would be the constant vocal tract length that was used to generate the area functions and consequent formant grid. In fact, this nontrivial problem is typically encountered with most inverse mappings of speech-to-area function. As examples, Schroeder (1966), Mermelstein (1967), Wakita (1972) all specified a constant vocal tract length prior to performing their respective inverse mapping schemes. As Mermelstein (1967) notes, "*tracts of different length can be distorted to yield the same formant frequencies. Hence, differences in the length of speakers' tracts do not necessarily manifest themselves in systematic formant differences*", suggesting that the actual vocal tract length would be very difficult, if not impossible, to obtain. However, Atal, Chang, Mathews and Tukey (1978) attempted to include vocal tract length in an inverse mapping as a recoverable parameter. They first generated a forward mapping from incremented articulatory model parameters such as place of maximum constriction, area of the maximum constriction and mouth opening, lip protrusion, and total vocal tract length to formant frequencies for F1, F2, and F3. The inverse mapping was then realized by correspondences between the formants and the articulatory parameters. They found, however, that three formant frequencies were unable to resolve a set of articulatory parameters uniquely.

To include vocal tract length in the present mapping approach, the δ_l in Equation 8 could be incrementally swept through its full range (see Table I) similar to c_1 and c_2; this would also be similar to the Atal *et al.* (1978) approach. If as many values of δ_l were combined with all of the previous values of c_1 and c_2, the number of coefficient "triples" would increase to 125000 (50^3). This could be reduced somewhat by using coarser increments of δ_l than the other coefficients but still presents a more complicated mapping than that presented previously. In addition, it would almost certainly generate a non-uniqueness between formants and modal/length coefficients. Such nonuniqueness may be an advantage with a more sophisticated system of matching formants to coefficients, but at this time we prefer to maintain the simplicity of the initial approach, that is to control only the mode 1 and 2 coefficients. An alternative method to include vocal tract length is to modify the mapping shown in Fig. 9 so that δ_l varies as a function of c_1 and c_2; define this as δ_{lvar}. Thus, the mapping is generated just as it was before, except that δ_{lvar} is calculated for every coefficient pair based on their values. The δ_{lvar}'s are calculated as a weighted sum of the original ten $\delta_l(p)$'s, where the weights are based on the squared distances between a given pair of coefficients $c_1(i)$, $c_2(j)$ and the coefficients derived for the original ten vowels $C_1(p)$, $C_2(p)$. This is accomplished by first computing the reciprocal of the squared distances

$$d_p(i,j) = \{[C_1(p) - c_1(i,j)]^2 + [C_2(p) - c_2(i,j)]^2 + \varepsilon\}^{-1} \quad (p = 1, 2, 3 \ldots 10) \quad (20)$$

where $\varepsilon = 1 \times 10^{-16}$ and is included to ensure that $d_p(i,j)$ does not become infinite under the condition $c_1(i) = C_1(p)$ and $c_2(j) = C_2(p)$. Next define the weights $w_p(i,j)$ to be the $d_p(i,j)$'s normalized to their sum across the ten vowels

$$w_p(i,j) = \frac{d_p(i,j)}{\sum_{p=1}^{10} d_p(i,j)} \quad (p = 1, 2, 3 \ldots 10) \quad (21)$$

so that each weight has a value between 0 and 1 and the sum of the weights $\sum_{p=1}^{10} w_p(i,j)$ is equal to 1. Now $\delta_{lvar}(i,j)$ can be calculated as

$$\delta_{lvar}(i,j) = \sum_{p=1}^{10} w_p(i,j)\,\delta_l(p) \tag{22}$$

where the $\delta_l(p)$'s are derived from the original vowel area functions and were given in Table I. Thus, an area vector is generated as before with Equation 18 but the length vector is now given by

$$L(\mathbf{x}) = \delta_{lvar}(i,j)\,\mathbf{x} \tag{23}$$

For cases when $c_1(i)$ and $c_2(j)$ are equivalent to one of the original vowels (i.e., $c_1(i) = C_1(p)$ and $c_2(j) = C_2(p)$), the weight corresponding to that vowel would be 1 while all others would be 0, thus equating $\delta_{lvar}(i,j)$ to $\delta_l(p)$.

Fig. 10 shows the coefficient-to-formant mapping modified by the variable vocal tract length. The coefficient grid (Fig. 10a) is shown now in three dimensions: the c_1 and c_2 along the x and y axes, respectively, and δ_{lvar} as a function of c_1 and c_2 along the z-axis. In Fig. 10a, the connected solid dots again denote the coefficients, and now also the vocal tract length intervals, of the original ten vowels, while the solid dots in Fig. 10b are the computed F1–F2 formant pairs based on these "triples" of coefficients and length intervals. The formant grid in Fig. 10b shows that the upper and lower borders attain higher and lower F2 frequencies, respectively, than in the constant length version of Fig. 9b. This is the expected result since the formants for front vowels are now computed with more appropriate shorter lengths and the mid-vowel formants are computed with their more typical longer lengths. The left border shows little change compared to the constant length version while the right border is slightly shifted to the left, indicating a lowering of F1. In addition to the overall increases and decreases in formant frequencies, the grid in Fig. 10b also indicates that some of the iso-coefficient lines have been warped relative to those for constant tract length. This warping most notably occurs in regions near (F1, F2) = (650, 2100) Hz and (F1, F2) = (750, 1200) Hz, while more a subtle warping is observed around (F1, F2) = (320, 1000) Hz. All three of these formant space regions correspond to a region in the coefficient grid where the δ_{lvar} is either at a peak or a valley.

As was shown in Figs 2 and 3, some of the fine detail of the area vectors is lost when only two modes are used for reconstruction instead of four. An attempt was made to bring the contribution of the third and fourth modes into the coefficient-to-formant mapping in much the same was as the variable tract length. In fact, the weights calculated by Equations 20–22 can be used to determine values of $c_3(i,j)$ and $c_4(i,j)$ as

$$c_3(i,j) = \sum_{p=1}^{10} w_p(i,j)\,C_3(p) \tag{24a}$$

$$c_4(i,j) = \sum_{p=1}^{10} w_p(i,j)\,C_4(p) \tag{24b}$$

so that an area vector can be computed by

$$A(\mathbf{x},i,j) = A_0(\mathbf{x}) + c_1(i)\,\phi_1(\mathbf{x}) + c_2(j)\,\phi_2(\mathbf{x}) + c_3(i,j)\,\phi_3(\mathbf{x}) + c_4(i,j)\,\phi_4(\mathbf{x}) \tag{25}$$

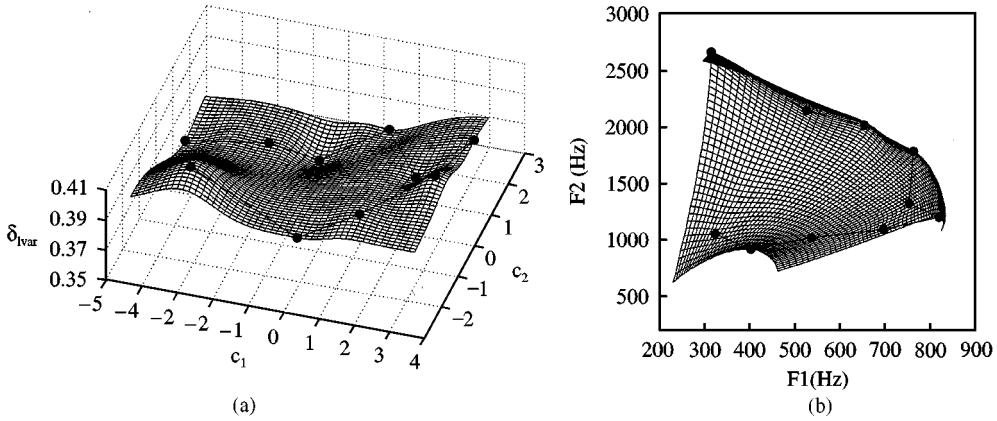


Figure 10. Mapping of mode 1 and mode 2 coefficient pairs to F1–F2 pairs when the vocal tract length interval is varied as a function of the mode 1 and mode 2 coefficients. The dots represent the coefficients, length intervals, and formant frequencies of the original ten vowels: (a) modal coefficient grid in three dimensions where δ_{lvar} is represented by the z-axis, the x and y axes represent the values of the mode 1 and mode 2 coefficients; (b) corresponding F1–F2 grid.

and the length vector is the same as Equation 23. Thus, the coefficients for modes 1 and 2 are still the only "controlled" parameters but the coefficients for modes 3 and 4 as well as the tract length interval are varied as functions of the mode 1 and 2 coefficients.

Two coefficient grids are shown in Fig. 11. The first (Fig. 11a) is the variation of the mode 3 coefficient as a function of c_1 and c_2, while the mode 4 variation is similarly shown in the second grid; the length is also varied but its corresponding grid is identical to that in Fig. 10a. The formant grid in Fig. 11c exhibits further deformation relative to the previous two cases. In particular, F2 values have increased in the right portion of the upper border but have decreased at the left. Some further warping of iso-coefficient lines is noted in the same three regions as discussed for the previous case ((F1, F2) = (650, 2100), (750, 1200) and (320, 1000) Hz), but a severe warping or folding is observed to extend diagonally from (F1, F2) = (340, 800) Hz, to (F1, F2) = (550, 1200) Hz. This region is characterized by a rapid decrease in both F1 and F2; beyond this region, F1 increases while F2 continues to decrease. A close study of Fig. 11 indicates that the folded formant region approximately corresponds to a portion of the coefficient grids bounded by $0 < c_1 < 2$ and $-2.38 < c_2 < -1.38$. Along the c_1 axis in this part of Fig. 11a, c_3 rapidly switches from a large positive value to large negative value which is mirrored by the behavior of F1 in the formant grid (i.e. F1 rapidly changes from a high value to a low value). No such correlation is observed for c_4.

4.1. *An example of the speech-to-area transform*

To test the possibility of mapping F1–F2 pairs back to a modal coefficient grid, two utterances recorded from the same subject who was imaged for the original MRI-acquired area functions (see Story *et al.*, 1996) were selected to be analyzed. The utterances both consisted of a continuous articulation of the vowels [i ɑ u i] such that the standard F1–F2 vowel chart was maximally traversed. For the first utterance the subject was asked to produce the [i ɑ u i] in a natural "conversational" mode, while the second

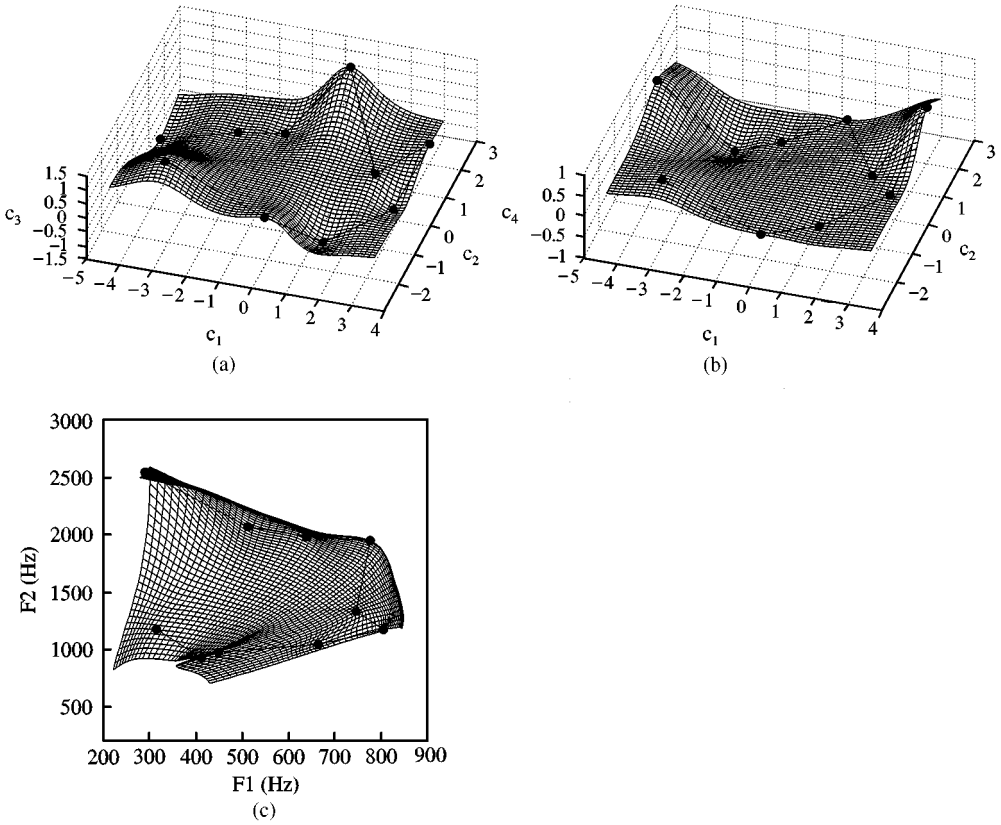


Figure 11. Mapping of mode 1 and mode 2 coefficient pairs to F1–F2 pairs when the vocal tract length interval and the coefficients for the third and fourth modes are varied as functions of the mode 1 and mode 2 coefficients. The dots represent the coefficients, and formant frequencies of the original ten vowels: (a) modal coefficient grid in three dimensions where the mode 3 coefficient is represented by the z-axis; the x and y axes represent the values of the mode 1 and mode 2 coefficients; (b) same as (a) except that the mode 4 coefficient is represented by the z-axis and (c) corresponding F1–F2 grid.

production was performed in an over-articulated style. Each utterance was recorded at a sampling frequency of 44.1 kHz in an anechoic chamber with a Panasonic SV-3700 DAT recorder and an AKG CK22 microphone. After recording, the utterances were downloaded digitally via a Digidesign Audiomedia board installed in a MacIntosh Quadra 950. The audio files were later read into MATLAB where a 50 coefficient LPC analysis coupled with a peak-picking algorithm (see Section 3) extracted the first two formants at 5 ms intervals.

Figs 12a, 12c, and 12e show the F1–F2 trajectory of the conversational [i ɑ u i] superimposed on the F1–F2 grid meshes (now shown with dotted lines so that the F1–F2 trajectory can be better seen) for constant tract length, variable tract length, and variable mode 3 and 4 coefficients (as well as variable tract length), respectively. The '1' and '140' symbols represent the beginning and the end of the utterance, respectively, and the other numerical symbols represent specific analysis frames that are discussed later. The

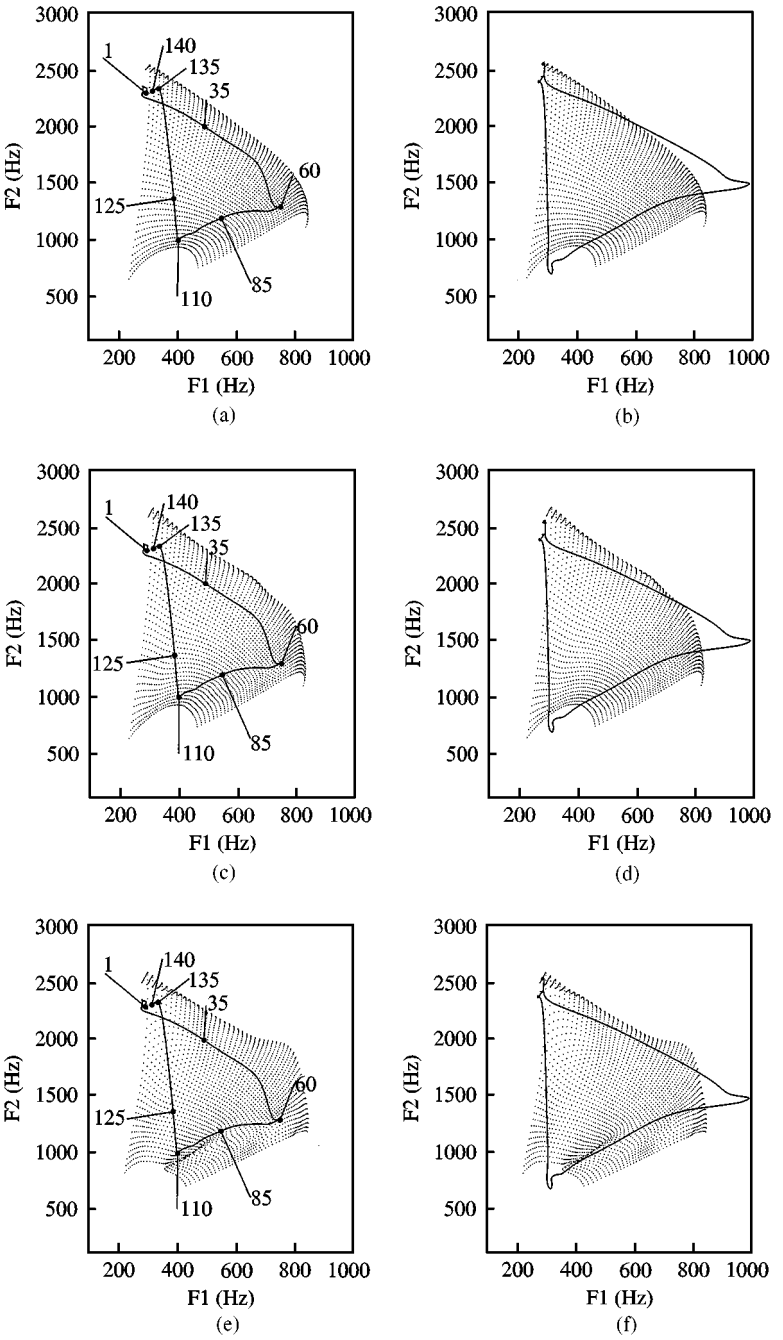


Figure 12. F1–F2 trajectories for the utterance [i ɑ u i] superimposed on the F1–F2 grids of Figures 9b, 10b, and 11c, the numerical symbols represent specific analysis frames: (a) conversational manner – constant vocal tract length, (b) over-articulated manner – constant vocal tract length, (c) conversational manner – variable vocal tract length, (d) over-articulated manner – variable vocal tract length, (e) conversational manner – variable vocal tract length and variable mode 3 and mode 4 coefficients (d) over-articulated manner – variable vocal tract length, and variable mode 3 and mode 4 coefficients.

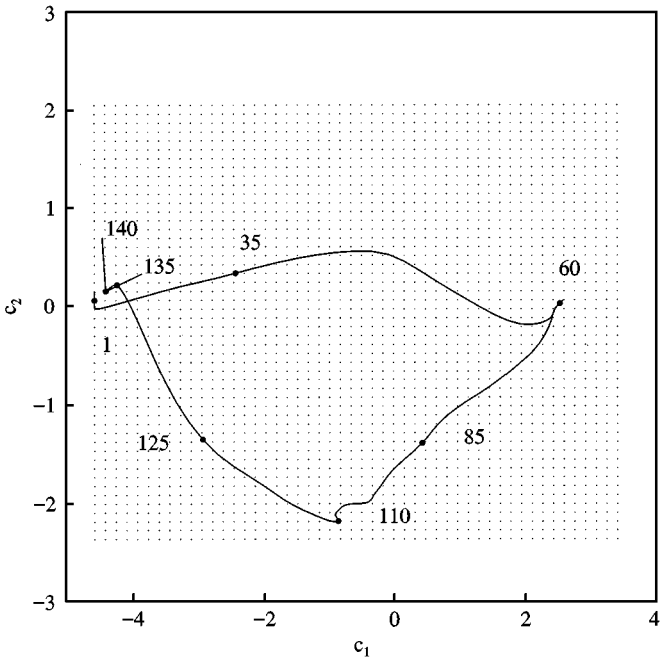


Figure 13. Modal coefficient trajectory corresponding to the F1–F2 trajectory shown in Fig. 12a. The numerical symbols correspond to the analysis frames indicated in Fig. 12a.

over-articulated version is similarly shown in Figs 12b, 12d, and 12f. The trajectory for the conversational version lies comfortably within all three F1–F2 grids and thus can easily be mapped into corresponding mode 1 and 2 coefficients as well as δ_{lvar}, c_3, and c_4. However, the over-articulated version produced a trajectory that extends outside the boundaries of the each F1–F2 grid. For the portions of the trajectory outside the mesh, any mapping back to modal coefficients is necessary deficient because no combination of the two modal coefficients would produce F1–F2 pairs in those regions. This is, however, expected since the original area functions (acquired from MRI) represented typical articulatory configurations and did not include extreme articulatory maneuvers such as excessive lip-rounding/spreading or larynx raising/lowering that may be required to produce overarticulated speech. Thus, the empirical orthogonal modes derived from the MRI-based area functions can only be expected to represent the boundaries of comfortable or conversational type speech. It may, however, be interesting to manipulate the mode shapes, mean area vector, and vocal tract length intervals to generate an F1–F2 grid that fits the overarticulated speech, but this is beyond the scope of the present study.

Toward a goal of mapping formant locations to area functions, each F1–F2 pair on the trajectory in Fig. 12a was first matched to the closest F1–F2 pair in the formant grid using a minimum squared error criterion from which the corresponding pair of mode 1 and 2 coefficients was determined. The resulting coefficient trajectory is shown in Fig. 13. Again the '1' represents the beginning of the utterance, '140' the end, and the numerical symbols and adjacent dots are the coefficient pairs that correspond to the frame numbers in Fig. 12a. The coefficient trajectory had a slightly jagged characteristic

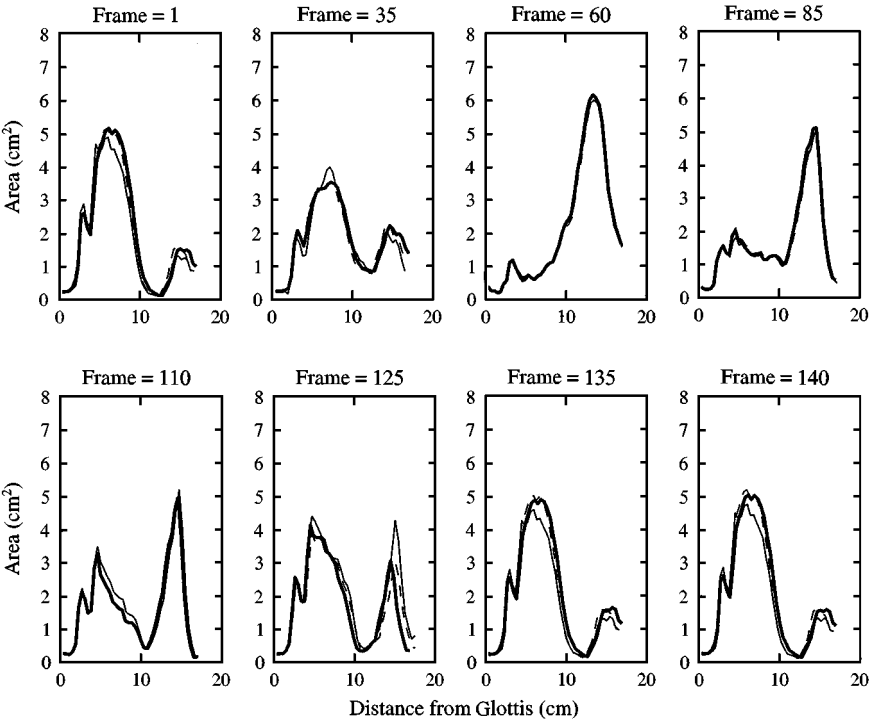


Figure 14. Series of eight area functions sampled from the 140 frame sequence of [i ɑ u i]. Thick solid lines denote the use of a constant tract length mapping (see Fig. 12a), while dashed and thin solid lines represent the variable tract length mapping and the variable tract length, mode 3 and mode 4 coefficient mapping, respectively.

due to forcing each F1–F2 pair extracted from the speech utterance to a discrete point in the formant grid. However, the trajectory shown in Fig. 13 has been smoothed with a fifth order low-pass (cutoff frequency of 5 Hz) FIR filter. A smooth trajectory could also be realized by interpolating within grid cells or by generating a finer grid initially, but for this preliminary study a simple minimum distance criterion was assumed to be adequate. The general shape of the coefficient trajectory is similar to that of the F1–F2 trajectory but rotated by approximately $+45$ degrees. For the grids in Fig. 12c and 12e, the $c_1 - c_2$ trajectory is similarly determined but with the addition of the tract length factor δ_{lvar} for Fig. 12c and the δ_{lvar}, c_3, and c_4 for Fig. 12e. These coefficient trajectories can now be used to generate area functions with the same time interval that formants were extracted from LPC spectra of the original speech.

Fig. 14 shows a series of eight area functions sampled from the 140 frame sequence of [i ɑ u i]. Area functions with a thick solid line were determined from the constant tract length mapping, while the dashed and thin lines represent area functions based on the variable length and variable length, c_3 and c_4 mappings, respectively. The numerical symbols shown in Figs 12a, 12c, and 12e are the sampled frames and their frame number is indicated at the top of each area function plot. Frame 1 is the starting [i] -like vowel which evolves into an [ɑ] -like shape by frame 60. The shape then changes into more of an [u] by frame 110. Frames 125 to 140 represent the transition back to the [i] -like

TABLE V. Area vectors for 8 consonants based on Story *et al.* (1996). Each original area function has been discretized to consist of 44 area sections; the length of each section is given by δ_l. The glottal end of each area vector is at section 1 and the lip end at section 44

Section	ɹ	l	m	n	ŋ	p	t	k
1	0.41	0.54	0.57	0.27	0.50	0.31	0.38	0.34
2	0.38	0.61	0.57	0.24	0.48	0.39	0.50	0.35
3	0.40	0.85	0.20	0.17	0.45	0.42	0.40	0.49
4	0.29	2.03	0.58	0.21	0.30	0.71	1.07	0.78
5	0.13	3.18	2.18	0.15	0.48	1.28	1.38	1.31
6	0.53	3.58	3.15	0.36	0.67	1.80	1.65	1.34
7	1.58	3.35	2.96	1.37	0.83	1.70	1.29	1.19
8	1.56	3.28	2.89	1.66	0.96	1.43	1.01	0.94
9	1.22	3.00	3.70	1.35	1.43	1.25	0.92	0.69
10	1.19	2.61	4.21	0.90	1.14	0.90	0.86	0.92
11	1.00	3.38	3.57	0.71	0.84	2.06	1.03	1.45
12	0.77	3.71	3.59	0.93	0.69	2.77	1.60	1.73
13	0.92	3.63	2.97	1.41	0.82	2.19	2.46	1.67
14	1.19	2.97	3.17	2.07	0.86	2.35	2.24	2.13
15	1.27	2.41	3.25	2.12	0.57	2.67	2.47	1.61
16	1.35	2.29	2.58	2.04	0.81	2.17	2.86	1.56
17	1.48	2.49	2.74	2.16	1.00	1.77	2.74	1.54
18	1.56	2.13	2.77	2.36	0.66	2.09	3.32	1.18
19	1.61	2.25	2.49	2.52	0.80	2.16	3.83	1.44
20	1.87	2.28	2.93	2.88	0.97	2.26	3.97	1.12
21	2.10	2.26	3.33	2.30	0.78	2.26	4.16	0.76
22	2.01	2.33	2.27	1.93	0.58	2.29	4.41	0.96
23	2.62	2.43	2.57	1.77	0.46	2.17	4.11	1.09
24	2.96	2.44	2.17	0.96	0.44	2.13	3.95	0.79
25	3.07	2.56	1.84	0.89	0.47	2.64	3.64	0.25
26	3.11	2.63	1.98	1.22	0.41	2.65	3.37	0.00
27	2.77	2.75	1.73	1.30	0.11	2.30	2.89	0.06
28	2.67	3.32	1.43	1.30	0.00	2.12	2.61	0.03
29	2.47	3.88	1.73	1.14	0.00	1.67	2.69	0.09
30	2.34	4.52	2.08	0.77	0.00	1.44	2.32	0.10
31	2.25	5.57	2.32	0.34	0.00	1.16	2.04	0.06
32	1.90	6.24	2.84	0.15	0.00	1.51	1.64	0.03
33	1.32	6.39	3.51	0.22	0.00	1.76	1.39	0.48
34	0.76	6.75	4.25	0.21	0.00	1.93	1.26	1.27
35	0.44	6.99	4.79	0.00	2.18	1.98	0.87	2.28
36	0.45	6.27	4.61	0.00	4.72	2.21	0.60	2.35
37	0.92	4.50	4.07	0.00	6.86	2.35	0.10	2.40
38	2.05	3.11	3.64	0.00	8.58	2.45	0.00	2.41
39	3.25	0.59	2.84	0.00	8.76	2.37	0.00	4.21
40	3.63	0.83	1.42	0.00	7.21	2.47	0.00	3.37
41	3.59	3.83	0.28	1.59	4.92	1.75	0.18	2.46
42	3.08	4.95	0.00	1.60	3.21	1.09	1.48	2.46
43	2.25	4.57	0.00	1.69	2.17	0.70	1.60	2.14
44	1.20	3.70	0.00	1.17	1.41	0.00	1.43	1.50
δ_l	0.391	0.422	0.374	0.348	0.374	0.339	0.368	0.392

vowel. The area functions derived from grid mappings with variable tract length (dashed and thin lines) show that for the most of the frames, the length is nearly the same as for the constant length case. The exception is for the [u] -like shape in frame 125 where the length is increased by almost a centimeter over the constant length version. Greater

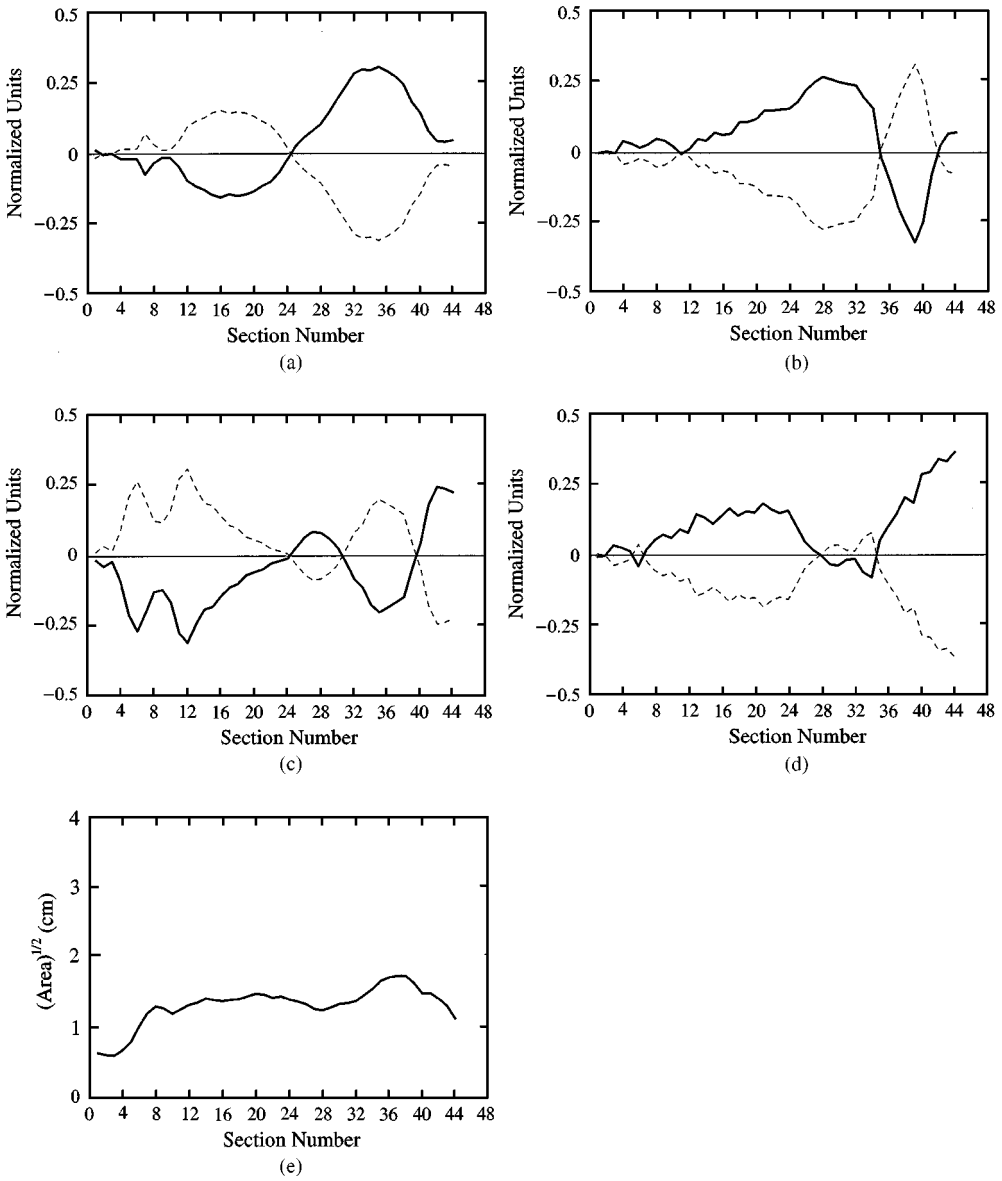


Figure 15. The four most prominent empirical orthogonal modes and the square root of the mean area vector derived from the 18 vowel–consonant set listed in Table V (the dashed lines are the reflection of each mode about the zero axis): (a) mode 1, (b) mode 2, (c) mode 3, (d) mode 4 and (e) square root of the mean area vector.

length differences throughout across the utterance would be expected if the F1–F2 trajectory followed more closely to the boundaries of the formant grid. Including the mode 3 and 4 coefficients does add some fine detail to the area functions, as can be seen in frames 1, 35, 125, 135, and 140. Frames 60, 85, and 110 show little advantage in using the extra coefficients.

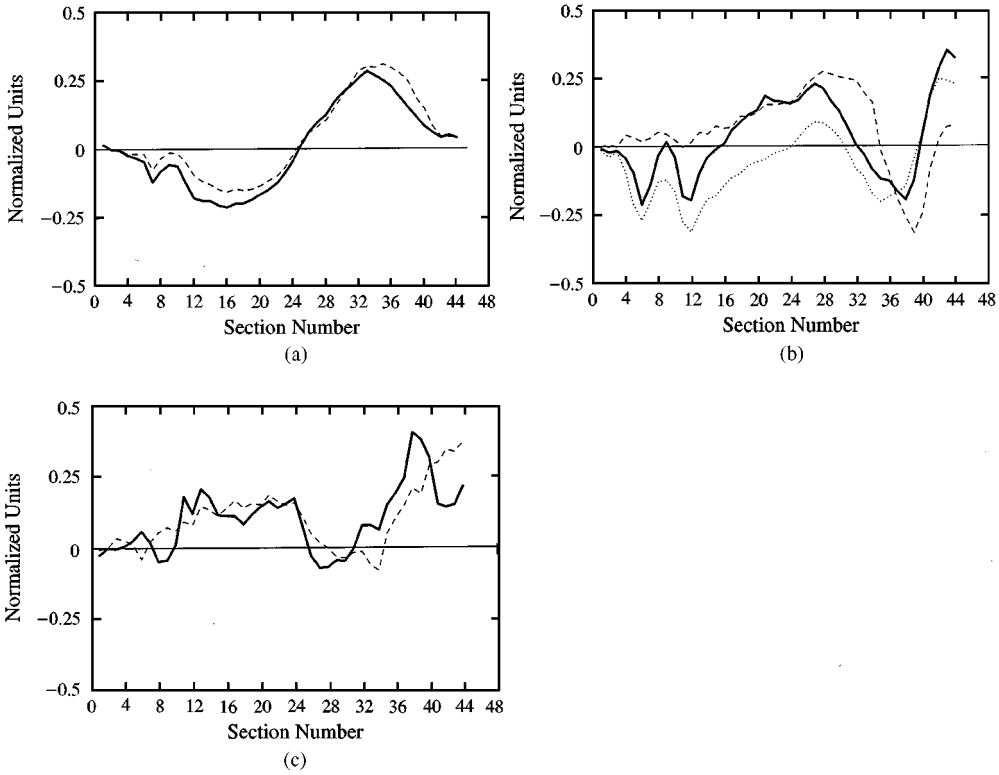


Figure 16. Comparison of the four most prominent modes derived from the 10 vowel set (solid lines) and the 18 vowel-consonant set (dashed and dotted): (a) mode 1 from both sets, (b) mode 2 from the 10 vowel set and modes 2 and 3 from the 18 vowel-consonant set, (c) mode 3 from the ten vowel set and mode 4 from the 18 vowel-consonant set.

While it is presumed that the addition of variable tract length and higher order mode generates more accurate area functions, the reader is reminded that all the three area functions in each analysis frame will produce the same formant frequencies for F1 and F2. This demonstrates the very issue discussed by Mermelstein (1967) (see previous section) concerning vocal tract length. A more thorough analysis and comparison of a larger number of formants (e.g., F1–F4) produced by these area functions with natural speech would be required to assess whether the added articulatory accuracy translates to increased acoustic accuracy.

5. Empirical orthogonal modes for vowels *and* consonants

So far, empirical orthogonal modes of only ten vowel area functions have been discussed. It is the intent of this section to briefly show mode shapes and the mean vector that result if the eight consonants given in Story *et al.* (1996) are included with the vowels in the modal decomposition. Table V gives the eight consonants in the same form as the vowels in Table I (i.e, 44 sections with a tract length interval δ_l).

TABLE VI. Cumulative percentage of variance for the first ten empirical orthogonal modes derived from 18 vowel and consonant area vectors. The first column is the mode number while the second and third columns give percentage of variance based on the $\sqrt{\text{area}}$ decomposition and the reconstructed area vectors, respectively

Mode	% of var. ($\sqrt{\text{area}}$)	% of var. (area)
1	41.45	69.94
2	66.14	77.72
3	81.58	83.16
4	89.36	93.10
5	93.93	94.71
6	95.71	95.42
7	96.99	96.44
8	97.91	97.10
9	98.57	97.12
10	99.02	97.57

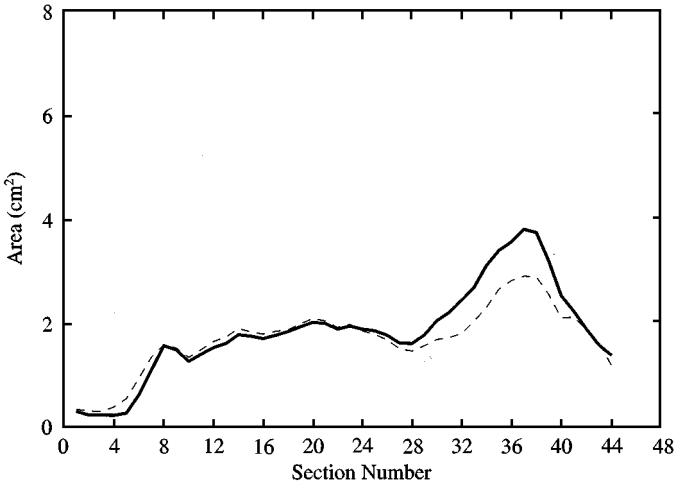


Figure 17. Mean area vectors (obtained by squaring the mean vectors from the modal decomposition) for the 10 vowel set (—) and the 18 vowel set (- - - -).

Fig. 15 shows the first four modes and the mean vector ($\sqrt{\text{area}}$) derived by using Equations 1–6, but with $p = 18$. The first mode is quite similar to that derived for the ten vowels (see Fig. 1). The second mode bears some resemblance to that of the ten vowel version in that it is similarly valued from sections 16 to 44, although the third mode is also similar to the *second* mode for the ten vowel case in the lower tract region of about sections 1 to 16. Furthermore, the fourth mode is most similar to the ten vowel *third* mode. These comparisons are displayed in Fig. 16, and suggest that the inclusion of the consonants has effectively "split" the second mode observed for the ten vowels, into two modes that give finer control of the velar, alveolar, dental, and labial regions of the vocal tract. This additional control would be expected due to the need to generate the consonantal constrictions and occlusions. The cumulative variances, given in Table VI,

TABLE VII. First four formant frequencies of a 1 cm^2 uniform tube and the mean area function based on 10 vowels and 8 consonants; both had a length of 17.5 cm. Percent differences between the two sets of formant frequencies are also shown

Tract shape	F1	F2	F3	F4
uniform tube	528	1482	2456	3436
mean area function	590	1568	2458	3637
% diff	11.8	5.9	0.1	5.8

indicate that several more modes are now required to account for high levels of variance. For example, in the ten vowel case two modes accounted for more than 92% of the total variance, while in the vowel-consonant case, four modes would be needed to account for this much variance.

The mean area vectors (obtained by squaring the mean vectors from the modal decomposition) for both cases are plotted together in Fig. 17. Interestingly, the mean area vector for 10 vowels and 8 consonants is nearly the same as that for only 10 vowels. The only significant difference between the two is in sections 28–40, where the 10 vowel mean area is 0.1–0.9 cm^2 larger than for the 18 vowel-consonant case. A calculation of the formant frequencies of the 18 vowel-consonant mean area function (with length = 17.5 cm) is given with those computed earlier for a 1 cm^2 uniform tube in Table VII. In comparison to the formants shown in Table IV for the 10 vowel mean area function, the F1 and F3 are now closer to the uniform tube formants, while F2 and F4 differ slightly more.

6. Discussion and conclusions

The primary goal at the outset of this study was to develop an efficient parameterization of the vowel area function set. However, finding the mean area function to have a formant structure similar to that of a uniform tube suggested that the empirical orthogonal modes could be considered as perturbations on a neutral vowel, that is, a physiologically-realistic neutral vowel. The area function for this neutral vowel, along with the empirical orthogonal modes (a set of orthogonal basis functions) can be considered a speaker-specific, empirically-based, physiological equivalent to the Fourier series representation (also a set of orthogonal basis functions) of the area function proposed by Schroeder (1966) and used by many researchers since.

With the data presented in this paper, it is not possible to quantitatively determine exact articulator positions that might be captured in each orthogonal mode shape. However, since the empirical orthogonal mode decomposition extracts the most prominent features from the input data set, it is useful to at least speculate on the articulatory characteristics that might be contained in the most significant modes. The asymmetrical shape of the first mode appears to result from front-back tongue positioning, while mode 2 seems to capture both the up-down tongue positions and lip opening/closing. The remaining higher, and less significant, modes fill in much of the fine detail in each area function, but their shapes do not lend themselves easily to any articulatory interpretation.

Since the first two modes account for over 92% of the total variance in the original area function set, the articulatory characteristics captured in those modes would likely be the primary mechanisms of perturbing the formants of the neutral area function. To study this, the amount (magnitude) of each of the first two orthogonal modes imposed on the mean area function was varied in isolation to demonstrate the acoustical effect of each mode by itself. It was found that F1 increased in frequency with an increase in the amplitude of mode 1 or mode 2, and F2 was moved down by increasing mode 1 and up by increasing mode 2. A calculation of the F1 and F2 sensitivity functions for the mean area function showed that both modes 1 and 2 were positively correlated with the F1 sensitivity function and oppositely correlated with F2 sensitivity, but with nearly the same absolute value. Modes 1 and 2 seem to be shaped so that they efficiently exploit the most acoustically sensitive regions of the neutral vowel, and the tendency for modes 1 and 2 to act cooperatively for F1 and in opposition for F2 allows an efficient coding of the first two formants by unique combinations of modal coefficients.

A mapping of a 2-dimensional grid of modal amplitude coefficient pairs for modes 1 and 2 to a deformed grid of F1–F2 pairs showed that each coefficient pair, within a large range of values, could be mapped to a unique F1–F2 pair. Thus, the selection of any combination of coefficient pairs is also a selection of a unique F1–F2 pair. This property leads to the possibility of mapping formants, extracted from a speech signal, to physiologically realistic area functions. Such a mapping was shown to be moderately successful with the vowel-only utterance [i ɑ u i]. This mapping was extended to also include variable vocal tract length and coefficients for the third and fourth modes as functions of the mode 1 and 2 coefficients.

Whether or not the mode shapes uncovered in this study will be similar for other speakers can only be answered with further vocal tract imaging studies. However, the similarity of the mode shapes to those derived by Meyer *et al.* (1989) (see Figure 4), and also the tongue factors determined by Harshman *et al.* (1977), suggest that similar modes might be expected for other speakers. In fact, Coker (1976) suggested that articulatory control might be organized around some collection of 'natural' modes:

> "Linguistic control appears to be organized around the modes of articulatory response. The reason is probably nothing more than the physical separation of articulators. But there would be tendencies for languages to align themselves around modes, even without physically separate articulators. A mode-oriented control strategy is simplest to learn in this domain, cause and effect are most directly associated."

As noted previously, the studies of Liljencrants (1971), Harshman *et al.* (1977), Jackson (1988), and Meyer *et al.* (1989) have all advocated mode like shapes as building blocks for vowels. It is nearly always the case that movement and/or vibration of physical systems is composed of fundamental modes. Classical modal analyses of strings, membranes, bars, air columns, etc. always show natural modes of vibrations. Berry *et al.* (1994) have found, using a similar decomposition of input data into empirical orthogonal modes, that the vocal folds, even though they possess complicated tissue layers and boundary conditions, typically vibrate with combinations of just a few fundamental vibratory modes. It would seem parsimonious that vowel production (and possibly speech production in general) would be organized around a few fundamental articulatory modes, especially if those modes efficiently exploit the acoustic modes (i.e., sensitivity functions) of the vocal tract. The results of Section 5 suggest that consonants as well as vowels can be represented by

a small set of modes; however, more modes are necessary to adequately describe the combined vowel-consonant set.

For some purpose, such as simulating connected speech segments, a vocal tract parameterization that utilizes mode shapes is an attractive alternative to controlling individual articulator positions. Since the typical articulatory specifiers such as tongue tip, tongue body, lip position, etc. are efficiently "packaged" by the mode shapes, extensive knowledge of how the articulators are coordinated is not needed. As an example, a simple model for simulation of speech production could be depicted as a time-dependent voice source specified by the desired fundamental frequency (F_0) and amplitude (A_0) (glottal flow pulse-shaping parameters such as skewing and open quotients could also be specified to control the voice quality). The time-dependent articulation could be governed by the choice of two coefficient values (c_1 and c_2) which scale the orthogonal modes. Also, based on the mapping between coefficients and formant locations, the selection of a given pair of coefficient values is equivalent to a selection of a unique pair of F1 and F2. Thus, formant values could also be used as control parameters. As stated, this simple modeling approach may not yield high quality speech simulation, but could serve as a starting point for further development.

Future work needs to focus on the limitations of this study. In particular, the coefficient-to-formant mapping should be extended to include higher formants, which may be essential for differentiating certain speech sounds (e.g., [l] and [ɹ]) and enhancing the general quality of any speech synthesis produced with the area functions. The effect of nasalization and consonantal constriction on the formant grid also needs to be explored in order to go beyond the realm of idealized vowel-like speech. Furthermore, the methods used for including variable tract length and higher ordered modes should be compared to other possible approaches, such as generating more extensive multi-dimensional coefficient/length grids similar to those proposed by Atal *et al.* (1978). It will also be of interest to apply the modal decomposition to area function sets for other speakers to see if their mode shapes are essentially the same as those shown here. It would appear that the modes capture the global qualities necessary to generate phonetically appropriate vowel sounds, but they likely also contain subtle variations that comprise an "acoustic signature" for a particular person.

This work was supported by Grant R01 DC02532 from the National Institute on Deafness and other Communication Disorders. The author would like to thank Dr. David Berry for fruitful discussions on empirical orthogonal mode decomposition. Three reviewers are also acknowledged for their helpful suggestions on improving the original manuscript.

References

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey, J. W. (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting-sorting technique, *Journal of Acoustical Society of America*, **63**(5), 1535–1555.

Berry, D. A., Herzel, H., Titze, I. R. & Krischer, K. (1994) Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions, *Journal of Acoustical Society of America*, **95**(6), 3595–3604.

Browman, C. & Goldstein, L. (1990) Gestural specification using dynamically-defined articulatory structures, *Haskins Lab. Stat. Rep. on Speech Res.*, SR-103/104, 95–110.

Coker, C. H. (1976) A model of articulatory dynamics and control, *Proceedings of IEEE*, **64**(4), 452–460.

Fant, G. (1960) *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G. & Pauli, S. (1975) Spatial characteristics of vocal tract resonance modes; In Fant, *Proc. Speech Comm. Sem.* 74., Stockholm, Sweden, Aug 1–3, 121–132.
Harshman, R., Ladefoged, P. & Goldstein, L. (1977) Factor analysis of tongue shapes, *Journal of Acoustical Society of America*, **62**(3), 693–707.
Herzel, H., Krischer, K., Berry, D. A. & Titze, I. R. (1995) Analysis of spatio-temporal patterns by means of empirical orthogonal functions, In *Spatio-Temporal Patterns in Nonequilibrium Complex Systems* (P. E. Cladis and P. Palffy-Muhoray, editors), pp. 505–518. Reading, MA: Addison-Wesley.
Jackson, M. T. T. (1988) Analysis of tongue positions: Language-specific and cross-lignuistic models, *Journal of Acoustical Society of America*, **84**(1), 124–143.
Ladefoged, P., Harshman, R., Goldstein, L. & Rice, L. (1978) Generating vocal tract shapes from formant frequencies, *Journal of Acoustical Society of America*, **64**(4), 1027–1035.
Liljencrants, J. (1985) Speech Synthesis with a Reflection-Type Line Analog, DS Dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden.
Liljencrants, J. (1971) Fourier series description of the tongue profile, *STL-QPSR*, **4**, 10–18.
Lindblom, B. & Sundberg, J. (1971) Acoustical consequences of lip, tongue, jaw and larynx movement, *Journal of Acoustical Society of America*, **4**(2), 1166–1179.
Mermelstein, P. (1973) Articulatory model for the study of speech production, *Journal of Acoustical Society of America*, **53**(4), 1070–1082.
Mermelstein, P. (1967) Determination of the vocal-tract shape from measured formant frequencies, *Journal of Acoustical Society of America*, **41**(5), 1283–1294.
Meyer, P., Wilhelms, R. & Strube, H. W. (1989) A quasiarticulatory speech synthesizer for German language running in real time, *Journal of Acoustical Society of America*, **86**(2), 523–539.
Mrayati, M., Carre, R. & Guerin, B. (1988) Distinctive regions and modes: A new theory of speech production, *Speech Comm.*, **7**, 257–286.
Munhall, K. G., Vatikiotis-Bateson, E. & Tohkura, Y. (1994) X-ray film database for speech research, ATR Technical Report, Human Information Processing Research Laboratories.
Nix, D. A., Papcun, G., Hogden, J. & Zlokarnik, I. (1996) Two cross-linguistic factors underlying tongue shapes for vowels, *Journal of Acoustical Society of America*, **99**(6), 3707–3717.
Schroeder, M. R. (1966) Determination of the geometry of the human vocal tract by acoustic measurements, *Journal of Acoustical Society of America*, **41**(4), 1002–1010.
Sondhi, M. M. & Schroeter, J. (1987) A hybrid time-frequency domain articulatory speech synthesizer, *IEEE Transactions* ASSP, ASSP-35(7), 955–967.
Stevens, K. N. & House, A. S. (1995) Development of a quantitative description of vowel articulation, *Journal of Acoustical Society of America*, **27**(3), 484–493.
Story, B. H., Titze, I. R. & Hoffman, E. A. (1996) Vocal tract area functions from magnetic resonance imaging, *Journal of Acoustical Society of America*, **100**(1), 537–554.
Story, B. H. (1995) Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract, Ph. D. Dissertation, University of Iowa.
Titze, I. R., Horii, Y. & Scherer, R. C. (1987) Some technical considerations in voice perturbation measurements. *Journal of Speech and Hearing Research*, **30**, 252–260.
Wakita, H. (1972) Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods, SCRL Monograph No. 9, Speech Communications Research Laboratory, Inc., Santa Barbara, CA.
Yehia, H. C., Takeda, K. & Itakura, F. (1996) An acoustically oriented vocal-tract model, *IEICE Trans. Inf & Syst.*, E79-D(8), 1198–1208.