TR－H－020

From EMG to formant patterns of vowels: the implication of vowel systems and spaces

Shinji Maeda
Kiyoshi Honda

1993. 8. 10

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2 ☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

From EMG to formant patterns of vowels: the implication of vowel systems and spaces

Shinji Maeda
(Ecole Nationale Supérieure des Télécommunications, Paris
and Centre National de la Recherche Scientifique)
Kiyoshi Honda
(ATR Human Information Processing Research Laboratories)

From EMG to formant patterns of vowels: the implication of vowel systems spaces

Shinji Maeda
(Ecole Nationale Supérieure des Télécommunications
and Centre National de la Recherche Scientifique, Paris)

Kiyoshi Honda
(ATR Human Information Processing Research Laboratories)

abstract. With a few exceptions [e.g., Kakita et al., 1985], EMG data are interpreted with reference to the intended output, such as the phonetic description of utterances spoken by speakers. For a more rigorous interpretation, the data should also be analyzed in terms of the displacement of the articulators and the acoustic patterns. In this paper, we describe our attempts to calculate the formant patterns from recorded EMG activities via an articulatory model [Maeda, 1990]. The value of the model parameters, such as the tongue-body position, tongue-body shape, is derived from the EMG activities of the specific pairs of antagonistic tongue muscles. The model-calculated F1-F2 patterns for 11 American English vowels correspond rather well with those measured from the acoustic signals. What strikes us is the simplicity of the mappings from the muscle activities, to vocal-tract configurations, and to the formant patterns. We speculate that the brain optimally exploits the morphology of the vocal tract and the kinematic functions of the tongue muscles so that the mappings from the muscle activities (production) to the acoustic patterns (perception) are simple and robust.

1. Introduction

Recently, Honda [1992], Honda et al. [1992], and Kusakawa et al. [1993] have shown that vowels plotted on a two-dimensional EMG space form a classical vowel triangle. One axis is **HG** - **GGp** (the hyoglossus activity minus the genioglossus posterior activity). The other axis, which is perpendicular to the above, corresponds to **SG** - **GGa** (the styloglossus minus the genioglossus anterior). The paired muscles function antagonistically. Thus **HG** and **GGp** contribute to a back-low/front-high tongue body movement, while **SG** and **GGa** contribute to a back-high/front-low movement [Honda, 1991]. But do those movements suffice to specify the acoustic characteristics, specifically F1 and F2, of the vowels? This question motivated us to carry out simulation experiments where an articulatory model is driven by a set of measured EMG activities and then the corresponding F1 and F2 frequencies of the vowels are calculated. The EMG-derived F1-F2 scatter plot is compared with that measured on the spectrograms of the audio signal simultaneously recorded with the EMG activities.

Attempting to calculate formant frequencies from EMG activities is not new. A few researchers [e.g., Kakita and Fujimura, 1984] have already done such calculations using a tongue model based on a finite element method. The EMG patterns that are assumed to represent directly the muscular force drive the finite element model. In our simulation experiment, we employ an articulatory model based on a factor analysis of X-ray film data [Maeda, 1979; Maeda, 1990]. The model is purely a descriptive one. The model consists of three parts: the lip opening tube, the tongue profile, and the larynx. The region of the hard and soft palates, the velum and the rear pharyngeal walls are assumed to be fixed. The geometry of the oral and pharyngeal cavities, therefore, are determined by the time-varying tongue shape. Each part is specified as the sum of the influences of the individual articulatory parameters. The midsagittal tongue shape is determined by specifying the values of an extrinsic parameter, lower jaw position (**jp**), and of three intrinsic parameters, tongue-body position (**tp**), tongue-body shape (**ts**), and tongue tip position (**tt**). The lip-opening tube is approximated by a uniform tube having an elliptic cross-section. The lip tube is, thus, determined by its three dimensions, height, width, and length. Each of these dimensions is calculated by a linear function of the common extrinsic parameter, **jp**, and two intrinsic parameters, height (**lh**) and protrusion (**lp**). The larynx tube is specified by **jp** and the intrinsic larynx height parameter (**lx**). Our model, therefore, can be stated as jaw based. From model-specified vocal-tract configuration, the corresponding area function and then acoustic characteristics, such as the transfer function and formant frequencies, and the stationary vowel signal are computed .

The factor analysis of vocal-tract data extracts the causes of variances observed in the time-varying vocal-tract shapes. The causes are nothing but the results of the systematic muscular activities that arise during speech production. The parameters could represent, therefore, the strength of the influence of each muscular activity upon the tongue and lip configurations. It is not overly unreasonable, then, to formulate an interface in which the averaged EMG activities representing muscular forces are converted into the values of positional model parameters.

We cannot expect, however, an exact prediction of the absolute vowel positions in the F1-F2 space, because our model was derived from X-ray film data of one subject and the EMG data were recorded from another subject. It is certain that due to the speaker difference in the vocal-tract geometry, the same vowels are produced with varying patterns of articulatory maneuvers depending on the individual speaker [Johnson et al., 1993]. Nevertheless, because the basic human vocal-tract morphology and kinematic functions are similar, it is rather natural to expect that a coherent vowel formant pattern could be derived from the EMG activity patterns via the articulatory model. Indeed, with relatively simple EMG-to-parameter conversion, we are able to obtain reasonable F1-F2 patterns for 11 American English vowels. We shall discuss the implication of the simulation results in order to shed some light on the interpretation of vowel systems and spaces.

2. Haskins Laboratories' EMG data and normalization

The Haskins Laboratories' EMG data [Baer et al., 1988] include the activities of the six extrinsic tongue muscles corresponding to an utterance type, /əpVp/, where V is one of 11 English vowels. Four principal extrinsic tongue muscles, GGa, GGp, HG, SG, as described earlier, and the geniohyoid (GH) that acts as a tongue fronting muscle are considered. They participate in the control of the tongue-body positions and can be directly related to the positional articulatory parameters. In addition, the EMG activities of the orbicularis oris superior (OOs) that contributes to the lip rounding and protrusion is used. These EMG patterns were obtained from ensemble average over 10 tokens of rectified and smoothed EMG signals for each utterance. Data resolution is 12 bits (4096 points) and the sampling rate is 5 ms per frame. The signal from each muscle is time-aligned at the target vowel onset. The data also include vertical (JawY) and horizontal (JawX) mandible movements in the midsagittal plane.

The data are accompanied by an audio tape which contains speech signals simultaneously recorded with the EMG and jaw movement data. The audio signal is used to measure the formant frequencies of the 11 vowels, which serve as a reference in order to judge the adequacy of the EMG-derived formant pattern.

The maximum EMG activity value of each muscle over 11 utterances is detected and used for normalization. For JawY data, the mean value over the 11 utterances is calculated and then this mean value is subtracted from the original jaw movement. The absolute maximum value, then, is determined and used for the normalization of the vertical jaw movement patterns. As the result, the values of all normalized EMG patterns are always positive and less than or equal to one. The normalised jaw position varies between plus and minus one. In the calculation of the jaw parameter (jp) of the model, this normalized vertical jaw position data (JawY) are translated into jp by a simple scaling with a coefficient, c_1, as

$$jp(n) = c_1 JawY(n),$$

where n is the frame number corresponding to time. The values of the remaining parameters are determined from the EMG activity patterns, as described in the following section.

3. Conversion from EMG to articulatory parameters

Our conversion from EMG to positional model-parameters is based on the empirical observation that tongue deformations inferred from the EMG activity patterns compare well with the effects of specific parameters of the articulatory model. Honda [1991] has examined the effects of muscle contraction upon the tongue shape with a 2D finite element tongue model. In Figure 1a, the tongue deformations due to the activities of HG and GGp calculated with the finite element model are shown. Due to the volume incompressibility of the tongue, the forward pull of GGp (indicated by the corresponding arrow) results not only in the forward movement as seen in the pharyngeal region, but also in the tongue raising movement in the palato-alveolar region. These tongue deformations from the equilibrium state, toward the front-high position with the activation of GGp and the back-low position with the activation of HG, correspond well to the

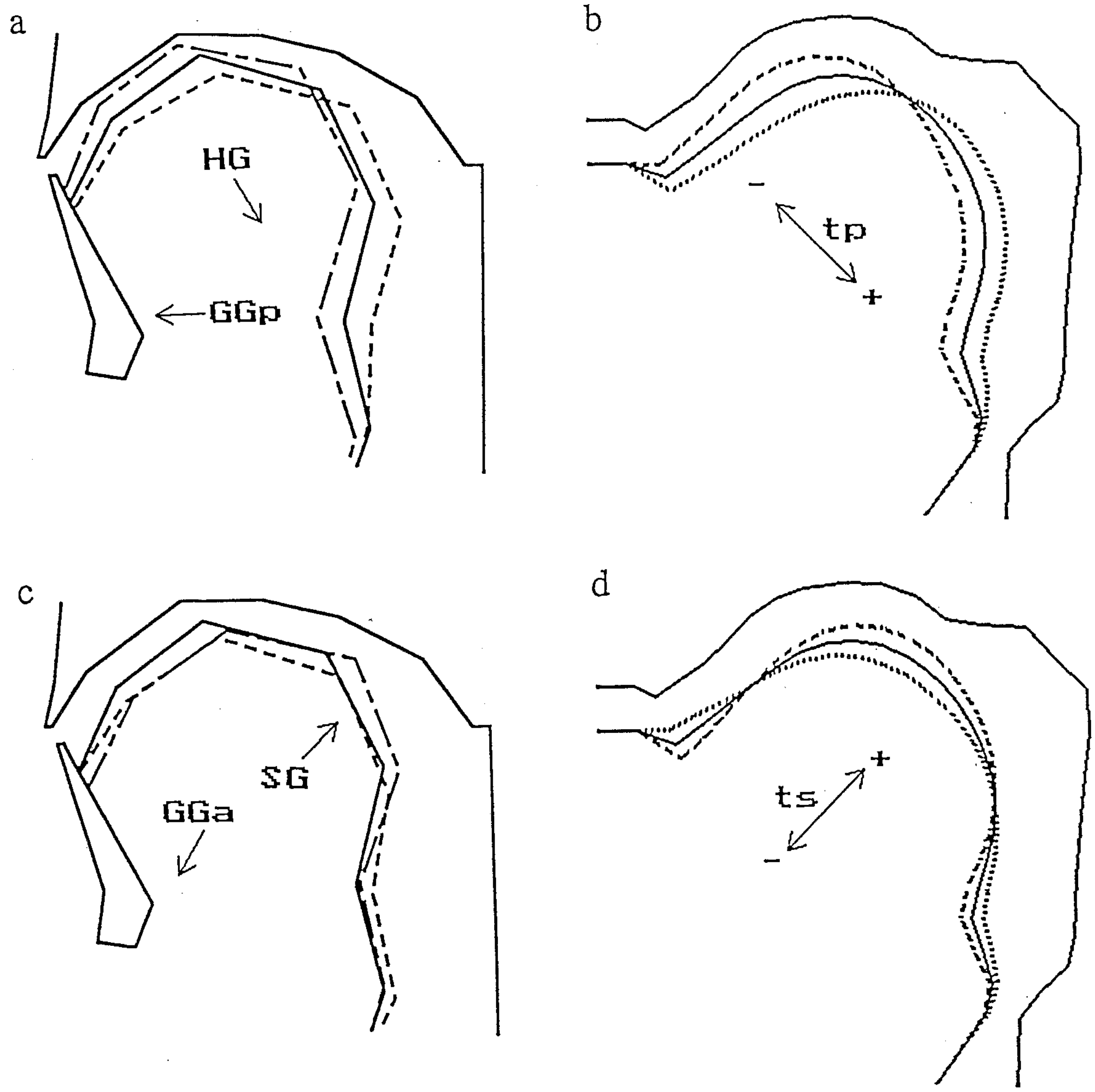

Figure 1. (a) The effects of HG (dashed line) and GGp contraction (dashed-dotted line) on the tongue shapes calculated with a 2D finite element model. The equilibrium shape is indicated by the solid line. The arrows schematically indicate the orientation of the contraction force of the corresponding muscles. (b) The effects of changes in **tp** (tongue-body position parameter) on the tongue profile calculated with an articulatory model based on a factor analysis of X-ray film data. The **tp** value varies between -1 (dashed line), 0 (solid line) and +1 standard deviation (dotted line). (c) The effects of SG (dashed-dotted line) and GGa contraction (dashed line). (d) The effects of change in **ts** (tongue-body shape parameter) on the tongue profile. The **ts** value varies between -1 (dotted line), 0 (solid line), and +1 standard deviation (dashed line).

tongue profiles of the present articulatory model shown in Figure 1b. These profiles are calculated by varying the value of **tp** between -1 (a front-high position), 0 (the neutral position), and 1 standard deviation (a back-low position). The effects of the other pair, **SG** and **GGa**, are shown in Figure 1c. The activation of **SG** moves the tongue-body toward the back-high position resulting in a bulged tongue body shape. This movement roughly corresponds to the variation of the **ts** parameter in the current model as shown in Figure 1d. In this figure, the value of **ts** is varied between 1 (a bulged tongue shape), 0 (the neutral position), and -1 standard deviation (a flattened tongue shape).

At least qualitatively, the tongue deformations due to the activities of **HG** and **GGp** correspond to those due to **tp,** and **SG** and **GGa** activities correspond to those of **ts.** It is not overly unrealistic to assume, then, that the value of the articulatory parameters, **tp** and **ts,** are determined by a function of the antagonistically combined EMG patterns. In an experiment, the articulatory parameter values were calculated as the sum of linearly scaled EMG activity patterns, assuming a proportional relation between the force (EMG response) and the displacement (model parameter). But this resulted in a poor prediction of F2 frequencies for certain vowels. We have introduced, therefore at the expense of simplicity, a non-linear scaling of EMG patterns, specifically for **HG** and **SG.** Moreover, it was effective to include **GH** , which, in the prediction of **tp,** can affect the tongue position in a way similar to GGp, as mentioned earlier. Thus, we have formulated the following set of EMG-to-parameter conversion equations to determine the two major tongue parameters.

$$\mathrm{tp}(n) = c_2\mathrm{HG}(n - \tau)^{\alpha} + \{c_3\mathrm{GGp}(n - \tau) + c_4\mathrm{GH}(n - \tau)\}/2$$

and

$$\mathbf{ts}(n) = c_5\mathrm{SG}(n - \tau)^{\alpha} + c_6\mathrm{GGa}(n - \tau),$$

where c_2, c_3, c_4, c_5 and c_6 are fixed coefficients and their values must be determined empirically. The coefficients c_2 and c_5 are positive and the remaining coefficients are negative to implement the antagonistic relations. The power constant, α, controls a linear scaling ($\alpha = 1$) or a non-linear scaling ($\alpha < 1$) of the EMG patterns. τ accounts for the latency from EMG response to the consequent movement, which will be described in more detail later.

The EMG data include the activity of **OOs,** which contributes to the lip rounding gesture. The **OOs** activity, therefore, can be related to the lip-opening height parameter (**lh**) as

$$lh(n) = c_7 OOs(n - \tau),$$

where c_7 is a negative fixed coefficient. Note that without the EMG data of the muscles antagonistic to **OOs**, **lh** always has a negative or zero value, thus the lip aperture can only decrease from a neutral value depending on the degree of **OOs** activities. The absence of antagonistic muscle activities in the **lh** specification must be corrected by an ad hoc manner, at least for some vowels, in formant calculation as described later.

How can such a linear or non-linear scaling of the EMG patterns be interpreted? The EMG activities represent the muscular force and the articulatory parameters are essentially positional. In the case of $\alpha = 1$, the linear (proportional) relationships between the force and position mean that a spring system obeying Hooke's law is assumed, which is described as

$$\Delta \mathbf{x}(n) = cF(n)^{\alpha},$$

where $\Delta \mathbf{x}$ is a displacement from the equilibrium position (corresponding to the articulatory parameter value) and F is a muscular force measured by the EMG activity patterns. Thus c_N (N = 2, 3, .., 7) in the above equations can be regarded as a "spring" constant. The spring model is particularly convenient, since each linear component in the articulatory model describes the "deviation" from the neutral (actually the mean) vocal-tract configuration and its magnitude is specified by the corresponding parameter value. When α is less than one, the scaling becomes non-linear such that the stiffness of the spring increases with higher degrees of contraction force.

Note that in Hooke's law (a spring), there is no time lag between force and displacement. Then the delay element, τ, which appears in the above equations, could be interpreted as a time lag between the EMG activity and the consequent force in our extremely simplified modeling. In order to align an **OOs** activity peak (which creates the lip closure) with the silence just before the intervocalic /p/ release in the audio signals, it is necessary to delay the **OOs** response by about 20 frames (100 ms). We assume the EMG pattern of every muscle is advanced by 100 ms relative to the movement.

It may be noted here that the EMG advance roughly corresponds to syllable length. This implies that the articulatory commands to the muscles must be issued during the preceding syllable and that articulatory movements could be purely ballistic without any feedback. Articulatory maneuvers such as compensation [e.g., Lindblom et al., 1979; Maeda 1990] and anticipation, must be programmed far in advance.

In addition to these four articulatory parameters, there are three others: **tt** (tongue tip position), **lx** (larynx height), and **lp** (lip protrusion). Since the EMG data are not available for the muscles controlling these parameters, the values of these three parameters are kept at zero corresponding to their equilibrium positions. However, the effect of **lp** can be compensated for by **lh**, because what counts acoustically is the ratio between the lip aperture (which is a function of **lh**) and the lip tube length (**lp**). The remaining two parameters, **tt** and **lx**, have little acoustic effect, particularly on the vowels and the consonant, /p/ treated here. The formant calculations would not be greatly affected by our assumption of their neutrality.

The values of the seven coefficients are determined using the following empirical procedure. The EMG and jaw position data are normalized as described already. Considering the antagonistic combination of the paired muscles, the maximum range of combined EMG activities becomes from -1 to 1. Since the maximum value of any articulatory parameters is generally within the range between -3 and 3 standard deviations, it is reasonable to let the initial guess value of all coefficients be +3 for the agonists and -3 for the antagonists. At the initial step, therefore, the range of **jp**, **tp** and **ts** becomes -3 to +3 and that of lh 0 to -3. The coefficient values are then empirically adjusted so that a complete closure or a strong constriction anywhere inside the vocal tract does not occur during any vowel segment. The lip closure, however, does occur during every /p/ before the release. The result of this empirical procedure is shown in Table 1.

Table 1. Values of scaling coefficients for converting the EMG response to the values of positional articulatory parameters.

c_1	c_2	c_3	c_4	c_5	c_6	c_7
-2.0	3.0	-4.0	-4.0	3.0	-2.0	-2.0

4. Simulation experiments: From EMG to formant frequencies

Formant frequencies are calculated, frame-by-frame, from the determined midsagittal vocal-tract shape and frontal lip shapes. First we estimate the corresponding area function from the EMG-specified midsagittal dimension of the vocal tract [Maeda, 1990]. Then, the transfer ratio between the volume velocity at the glottis and the radiated sound pressure is calculated as a function of frequency. The transfer calculation takes into account the effects of the yielding vocal-tract walls and of the radiation impedance at the lips [Maeda, 1982]. The formant frequencies are determined using a peak-picking algorithm from the smooth transfer function calculated. The formant frequencies at 75 ms after the vowel onset are taken as the formant values of the target vowels, V's, in the utterances [əpVp]. The calculated F1-F2 pattern for the 11 vowels are compared with those measured on the simultaneously recorded speech signals.

In order to obtain a reference F1-F2 plot for the 11 vowels, the F1 and F2 frequencies of each vowel are visually determined from the spectrogram and the spectrum slice of each utterance at the same sampling point as the model calculations. The resultant F1-F2 scatter of the 11 vowels is plotted on every graph in Figure 2 using closed circles. The measured F1-F2 plot roughly corresponds to the standard vowel pattern of American English. The relative positioning of the low vowels, /æ, a, and U/ does not conform to the typical vowel pattern. This discrepancy may be, at least in part, due to the fact that we have measured F1 and F2 from only a single token for each vowel. It should be noted that the two vowels /e and o/ were diphthongized. This deviation from the standard vowel pattern is of little consequence, however, since the objective here is to compare the acoustically measured F1-F2 frequencies with those calculated from the EMG activities sampled simultaneously (with delay τ). We describe three simulation experiments in which the conditions for calculating the principal articulatory parameters are varied.

Experiment 1: Linear EMG scaling without lip height correction

In this first experiment, we deliberately employ non-corrected lh values that are directly derived from OOs as specified by the formula. The value of α is equal to one, i.e., a linear scaling of the EMG responses. Figure 2a compares the EMG-derived and spectrally measured F1-F2 plots. The corresponding vowels are connected to each other by straight lines.

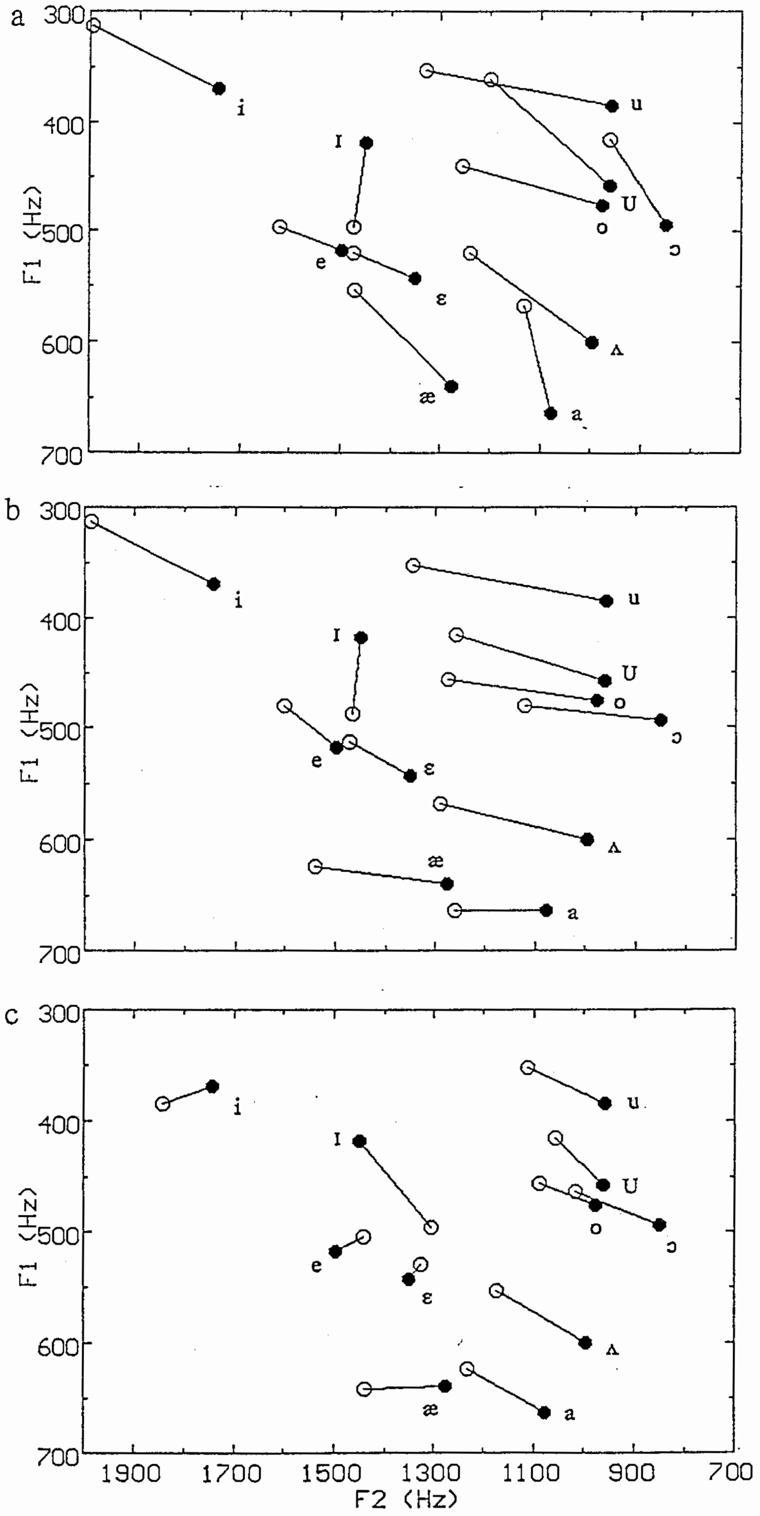

Figure 2. F1-F2 plot of 11 American English vowels measured on spectrograms, indicated by closed circles and that calculated from EMG and jaw movement data, indicated by open circles. (a) Linear scaling of all EMG responses in the calculation of the articulatory parameters. (b) The same as (a), except some 1h (lip height) values are corrected (see Table 2). (c) The same as (b), except the EMG response of **GH** and **SG** are non-linearly scaled.

The comparison of the measured F1-F2 pattern and that derived from the EMG reveals two major discrepancies. The first is the noticeably low F1 frequencies for the low vowels, /æ, a, ʌ, and ɔ/. The excessively low F1 frequencies of these low vowels can be explained by the fact that **lh** determined from **OOs** alone results in an excessively small lip aperture. The second discrepancy occurs for the non-front vowels, /u, U, o, ʌ and æ/, whose F2 frequencies are systematically too high. Excessively high F2 values for these vowels can be explained by the fact that the EMG-derived tongue position is too fronted for these vowels. As a consequence of these two discrepancies, the EMG-derived F1-F2 scatter poorly compares with the spectrally measured one.

Experiment 2: Linear EMG scaling with corrected lip height

An excessively low F1 should be expected, since the lip shape is not fully specified because of the lack of consideration of the lip spreading muscles antagonistic to **OOs**. In order to demonstrate that this is indeed the case, we changed the **lh** value somewhat arbitrarily while respecting the order so that the degree of lip aperture monotonically increases from /u, U, ɔ, ʌ æ/, and to /a/. The recalculated F1-F2 scatter is shown in Figure 2b. Now most of the discrepancy remains with the F2 dimension. This is clearly seen by the straight lines connecting the measured and corresponding EMG vowels, which are now mostly horizontal and parallel with the F2 axis.

Experiment 3: non-linear EMG scaling with lip height correction

Unlike the excessively low F1 case, the predicted excessively high F2 values for those five low to high back vowels, /ʌ, ɔ, o, U, u/, deserve careful analysis. The F2 values are closely related to the tongue-body position, specified by the two parameters, **tp** and **ts**. An overly high F2 implies that the tongue is not sufficiently far back and/or the degree of constriction is not sufficient for those vowels.

One might suggest that correct tongue-body positions can be obtained by increasing the values of the EMG-to-position conversion coefficients, specifically c_2, which is associated with HG and c_5, which is associated with SG. This will not work, however. If the value of c_2 and/or c_5 were increased in order to enhance the tongue retracting function of HG and SG in the high/back vowels, it would result in a complete closure of the vocal tract for certain low/back vowels. The value of these coefficients, in particular, therefore, cannot be modified effectively for all vowels. It appears then that an effective remedy to the problem is to introduce a non-linear scaling of the agonists **HG** and **SG** by setting the value of α smaller than one. In this way, the force-displacement relations become non-linear in such a way that the displacement due to a small change in force at low level is greater than that at a higher level. Moreover, the scaled EMG never exceeds the value of the coefficients, since the maximum value of the EMG is normalized to one. Thus the non-linear scaling tends to result in a more rearward tongue position and a more bulging tongue shape in comparison with linear scaling for the same EMG activity level.

Let us try $\alpha = 0.5$, just to see the effects of the non-linear scaling of **HG** and **SG**. As in the case of the previous two experiments, the F1 and F2 values are calculated for the 11 target vowels and the results are shown in Figure 2c. The EMG specified vowels are closer to the corresponding acoustically measured vowels in comparison with the scatter plot shown in Figure 2b. Comparing Figure 2b and 2c, the effectiveness of the non-linearity can be seen. Probably, by optimizing the fixed coefficient values and finding more appropriate **lh** values for individual vowels, it would be possible to obtain a better match between the EMG-derived and measured F1-F2 patterns. We believe that this will be meaningless, however, since the aim of this computational experiment is to explain the observed relations among different vowels. In this sense, we are satisfied by the result here, although the EMG-predicted positions of the two vowels /I and æ / are still somewhat misplaced.

Non-linear scaling was employed to improve the F2 specification. It might be tempting to assume that the non-linear scaling reflects an increase in the stiffness of the muscle at higher degrees of contraction. The non-linearity, therefore, should be reasonable considering the relatively large excursions of these parameters. We doubt the validity of such an explanation, however.

Table 2 lists the values of scaled EMG activity levels and the resultant articulatory parameter values and calculated formant frequencies for all 11 vowels. It is seen that **tp** and **ts** have mostly positive values, indicating a rearward tongue position and a bulging tongue shape. The mean values of these parameters are, respectively, 1.26 and 0.8. The articulatory parameters are measures of the deviation from the neutral tongue shape. The values averaged over the 11 vowels may be assumed to be close to zero. The EMG-derived tongue body positions and shapes, therefore, are highly "skewed" toward a positive value. For comparison, we calculated the average values of these two parameters in the

Table 2. Values of scaled EMG activity levels and those of the corresponding articulatory parameters (in standard deviation) and formant frequencies (in Hz). The corrected lip height parameter values are listed at lh*.

vowel	F1	F2	jp	tp	HG	GGp	GH	ts	SG	GGa	lh*	lh
i	383	1841	-0.3	-0.8	2.0	-1.9	-0.9	-0.4	1.2	-1.6	---	-0.5
I	496	1306	-0.3	1.4	2.2	-0.2	-0.7	0.3	1.2	-0.9	---	-0.4
e	504	1442	-0.9	0.7	2.0	-0.3	-1.0	-0.4	1.0	-1.4	---	-0.4
ε	529	1327	-1.0	1.2	2.3	-0.1	-1.1	0.3	1.4	-1.1	---	-0.3
æ	642	1439	-2.0	0.9	2.5	-0.1	-1.5	-0.4	1.4	-1.8	0.5	-0.3
a	624	1232	-1.3	2.2	2.9	-0.1	-0.6	1.2	1.8	-0.6	0.8	-0.3
ɔ	464	1018	-0.5	2.0	2.6	-0.1	-0.5	2.2	2.7	-0.5	-0.2	-0.8
o	456	1089	-0.7	1.4	1.9	-0.2	-0.4	1.9	2.5	-0.5	---	-0.5
U	416	1056	0.5	1.8	2.3	-0.1	-0.3	1.3	1.7	-0.4	-0.7	-0.9
u	352	1112	0.7	1.1	2.2	-0.9	-0.2	1.7	2.4	-0.7	---	-0.8
ʌ	553	1173	-0.5	2.0	2.6	-0.1	-0.5	1.0	1.6	-0.5	0.0	-0.4

case of linearly scaled EMG activity patterns. These were, respectively, 0.8 and 0.16, corresponding to a configuration closer to neutral. The non-linear scaling turned out to be an effective means to skew the position farther back. Why is it necessary to skew the position to such a high degree, in order to predict correct F2 frequencies? We think that the need for the skewed position comes from the fact that the model vocal-tract geometry differs considerably from that of the EMG subject, the shape of the exterior vocal-tract walls from the palate to the pharynx, in particular, being different. The EMG subject most likely has a more advanced back pharyngeal wall in comparison with the model. Thus the skewed positions are understood to be a way to compensate for the difference in the overall vocal-tract dimensions between the EMG subject and the X-ray data subject on which the articulatory model is based.

It may be noted in Table 2 that the tongue position of the vowels /u, U, and o/ predicted from EMG is more fronted than the low back vowels /ɔ, ʌ, a, æ/. This relatively advanced tongue position is a direct consequence of relatively weak HG activity. The model-calculated profiles show that the constriction of these vowels occurs in the palatovelar region. This is not a contradiction, however. It is known from X-ray studies that the vowels, such as /u and o/ can be produced from two different vocal-tract configurations: one with the oral constriction in the palatovelar region and the other in the velopharyngeal region [e.g., Boë et al., 1992]. Apparently, our EMG subject articulated those vowels using the fronted constriction point.

5. Discussion

Surprisingly, such simple conversion from EMG to the value of the positional articulatory parameters is able to specify the vowel formant patterns rather satisfactorily. These results seem to support the following three plausible hypotheses. First, although the tongue muscular system is anatomically complex, it is organized into a small number of functional blocks for speech production. This was demonstrated by the fact that the tongue position for vowels can be specified using only two independent variables, the antagonistically combined EMG activities, **HG - GGp** and **SG - GGa** [Kusakawa et al., 1993]. One of the key elements in this description is that the genioglossus muscles are functionally split into two parts, the posterior fibers (**GGp**) and the anterior fibers (**GGa**). Second, the two essential model parameters, **tp** and **ts**, describe the effects of those two sets of antagonistically combined muscles on the tongue shapes. Third, the morphology of the vocal tract and the kinematic functions of the tongue muscles allow to produce an inherently "stable" and distinctive vowel acoustic pattern, regardless of detailed differences in the vocal-tract geometry.

But where does such inherent stability come from? We speculate that the right-angled vocal-tract exterior walls surrounding the tongue body are the key to the formation of the stable vowel triangle (space) that is characterized by the three extreme vowels /i, a, and u/. It is important to note that the configurations for these extreme vowels are produced with the maximum activity levels of the participating muscles and they have special acoustic properties, in the sense their F1 and F2 frequencies correspond to distinctly different resonance modes [Fant, 1960]. For /i/, the narrow and long frontal (oral) cavity and the wide pharyngeal tract form a Helmholtz resonator. F1 corresponds to this Helmholtz resonance, which can be at a very low frequency, since the oral cavity can form a narrow tube of any degree by a proper articulatory maneuver. The F2 of /i/ is always highest among the vowels, because F2 is primarily determined by the half-wave length resonance of the front or the back cavity. Since the total length of the vocal tract is about 17 cm, the length of the front and back cavity would be about a half of that, F2 being roughly 2 kHz. For /a/, the vocal tract is characterized by the narrow back cavity connected to the wide front cavity. Both F1 and F2 are related to the quarter-wave length resonance of the front or back cavities, having about a half of the tract length. In this condition, F1 is most likely below 1 kHz and F2 above 1 kHz. For/u/, both F1 and F2 can be low, since both formants correspond to the coupled Helmholtz resonators in the two parts of the vocal tract.

The formant frequencies, especially F1 and F2, of these extreme vowels, therefore, are determined also by a small number of variables which characterize the area function of the vocal tract; the (equivalent) length of front or back cavity in the case of the quarter-wave and half-wave length resonances, and the volume and the length-area ratio of the "mouth" tube in the case of the Helmholtz resonance. It should be noticed, for example, that the frequency of the formant related to the quarter-wave or half-wave resonance is sensitive only to a variation in the (effective) length of the tube, but insensitive to a variation in cross-sectional areas. Therefore, speakers having a similar tract length should be able to produce similar vowel patterns. The original three-parameter model [Fant, 1960], which consist of only four uniform tubes to represent the area function, exploits these properties of the different resonance modes to cover all vowels, producing the famous "nomograms".

Indeed, the right-angled human vocal tract can quite readily form those three extreme configurations with a small number of control variables: Fronting the tongue body primarily by the activation of **GGp** muscle results in the /i/ configuration. Conversely, backing of the tongue body by the contraction of HG alone sets the tract to the /a/ configuration. The activation of SG creates a constriction in the middle of the vocal tract and, at the same time, forms the front and back cavities. The concomitant activation of OOs would result in the formation of the second constriction at the lips, thus completing the characteristic two enclosed cavities connected at the middle constriction of the /u/ configuration. Moreover, just in passing, a difference in the global tract shape could be compensated for, to a certain degree, by a systematic adjustment of the tongue position, which was demonstrated by the skewed tongue positions in our simulation experiment.

If we assume 1 kHz as being the middle (Mid) frequency, F1 and F2 of /a/ are characterized by Mid-Mid, that of /i/ by Low-High, and that of /u/ by Low-Low or Low-Mid, resulting in spectrally distinctive patterns. It is remarkable that these typical configurations can be derived from extremely simple tongue position control, involving only two pairs of antagonistic muscles, because of the morphology peculiar to the human vocal tract. Perhaps, this is why the large majority of world languages includes these three vowels [Maddieson, 1984; Lindblom, 1986]. Now it appears that the first computational articulatory model formulated by Coker and Fujimura [1966] has intuitively implemented this parsimonious tongue position control in a direct way.

Next, how are vowel positions determined when there are more than three vowels? Liljencrants and Lindblom [1972] have attempted to predict the vowel system on the basis of the notion of perceptual contrast, when the number of vowels is specified. Vallée [1989] has extended this work by adding the notion of "focal point", which is essentially the approaching of two formants. As characterized by Stevens' quantal nature [1989], the values of two approached formants are locally stable against a certain variation in the vocal-tract area function, and thus creating a certain stability against articulatory deviation. We agree that the perceptual contrast and its quantal nature are probably important factors in the formation of the vowel systems of languages. Such concepts are needed, since there is no particular articulatory reason to determined an intermediate vowel position. In principle, the tongue position can vary continuously from one extreme to another because the muscular action level can vary continuously. In fact, Lindblom and Sundberg [1971] have succeeded in describing the tongue shape of an arbitrary vowel through an interpolation of the three extreme tongue shapes.

6. Concluding remarks

This study has shown that the F1-F2 patterns for vowels can be recovered from tongue EMG and jaw movement data via the articulatory model. We speculate that the vowel space is formed in such a way that the vowel specification in terms of muscular activity pattern, vocal-tract configuration and formant pattern are related to each other in a simple manner. If the vowel specification in the human perceptual system can be assumed to be an analog of the F1-F2 pattern, it is speculated further that the brain optimally exploits the morphological and functional characteristics of the vocal-tract organs so that the acoustic pattern (perception) is connected to the production process in a relatively simple mapping, establishing an efficient communication system [Honda 1993]. Although such speculations, perhaps, are a matter of conjecture, we have, at least, successfully demonstrated how effectively the EMG-to-sound via the articulatory model can provide an objective means to evaluate the activity of the individual muscles in terms of the articulatory positioning and acoustic patterning, thus filling the gaps between the EMG activities and the acoustics.

Acknowledgements

We thank Dr. Eric Vatikiotis-Bateson for his help and many valuable discussions throughout the investigation. We also thank Prof. Osamu Fujimura for his helpful criticisms and comments on a previous version of the manuscript. The EMG data in this paper were recorded and pre-analyzed at Haskins Laboratories in New Haven, Connecticut, US, and were supported by an NIH Grant DC-00121 to Haskins Laboratories (PI; Prof. Katherine S. Harris). An earlier version of this paper was presented at the Third Seminar on Speech Production: Models and Data, Saybrook, CT, USA, May 1993, and a final version at the Second Workshop (ESPRIT/BRA, EC), Lingual data and modeling in speech production, Institut d'Estudis Catalans, Barcelona, December 1993.

References

Baer, T.; Alfonso, P.J.; Honda, K.: Electromyography of the tongue muscles during vowels in /əpVp/ environment. Annual Bulletin of Research Institute of Logopedics and Phoniatrics 22: 7-19 (University of Tokyo, 1988).

Boë, L-J.; Perrier, P.; Bailly, G.: The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. Journal of Phonetics 22: 27-38 (1992).

Coker, C.; Fujimura, O.: A model for specification of the vocal tract area function. Journal of Acoustical Society of America 40, 1271 (1966).

Fant, G.: Acoustic theory of speech production. (Mouton, The Hague 1960).

Honda, K.: A statistical analysis of tongue muscle EMG and vowel formant frequencies. Journal of Acoustical Society of America 90, 4 (Part 2): 2310 (1991).

Honda, K.; Kusakawa; N. Kakita Y.: An EMG analysis of sequential control cycles of articulatory activity during /əpVp/ utterances. Journal of Phonetics 20: 53-63 (1992).

Honda, K.: Physiological background of speech production (tutorial). Journal of Acoustical Society of Japan 48 (1): 9-14, *in Japanese* , (1992).

Honda, K.: Modeling vocal tract organs based on MRI and EMG observations and its implication on brain function. Annual Bulletin of Research Institute of Logopedics and Phoniatrics 27: *in print* (University of Tokyo, 1993).

Johnson, K.; Ladefoged, P.; Lindau, M.: Individual differences in vowel production. Journal of Acoustical Society of America 94 (2): 701-714 (1993).

Kakita, Y.; Fujimura, O.: Mapping from muscular contraction patterns to formant patterns in vowel space. Transaction of the Committee on Speech Research, The Acoustical Society of Japan, S83-100, *in Japanese:* (March 31, 1984).

Kakita, Y.; Fujimura, O.; Honda K.: Computation of mapping from muscular contraction to formant patterns in vowel space. In Phonetic Linguistics, (V.A. Fromkin, editor): 133-144 (Academic Press Inc., 1985)

Kusakawa, N.; Honda, K.; Kakita, Y.: Construction of articulatory trajectories in the space of tongue muscle contraction force. ATR Technical Report, TR-A-0717, *in Japanese,* (1993).

Liljencrants, J.; Lindblom, B.: Numerical simulation of vowel quality system: the role of perceptual contrast. Language 48: 839-862 (1972).

Lindblom, B.; Lubker, J.; Gay, T.: Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. Journal of Phonetics 7: 147-161 (1979).

Lindblom, B.: Phonetic universal in vowel system. in Experimental phonology, (J. Ohara, editor): 13-44 (Academic Press, New York 1986).

Lindblom, B.; Sundberg, J.: Acoustical consequences of lip, tongue, jaw and larynx movements. Journal of Acoustical Society of America 50 (4, Part 2): 1166-1179 (1971).

Maddieson, I.: Patterns of sound. (Cambridge University Press, Cambridge 1984).

Maeda, S.: Un modèle articulatoire de la langue avec composantes linéaires. 10émes JEP, GALF: 152-164 (1979).

Maeda, S.: A digital simulation method of vocal-tract system. Speech Communication 1: 199-229 (1982).

Maeda, S.: Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. in Speech Production and Speech Modeling (W.J. Hardcastle; A. Marchal, editors): 131-149 (Kluwer Academic Publishers, 1990).

Stevens, K. N.: On the quantal nature of speech. Journal of Phonetics 17: 3-45 (1989).

Vallée, N.: Typologie des systèmes vocalique. Travail d'Etudes et de Recherche de Maîtrise (Université Stendhal, Grenoble, France 1989)