ARTICULATORY CO-ORDINATION AND ITS NEUROBIOLOGICAL ASPECTS: AN ESSAY

Shinji MAEDA
Centre National de la Recherche Scientifique, URA 820
and Ecole Nationale Supérieure des Télécommunications, Département SIGNAL
(46, rue Barrault, 75634 Paris, France)

Kiyoshi HONDA
ATR Human Information Processing Research Laboratories
(2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan)

ABSTRACT

The aim of this talk is to describe a jaw-tongue articulatory compensation during unrounded vowels. Simulation experiments using an articulatory model indicated that an optimum compensation can be directly specified by linear relationships between these two articulators' positions. As if a speaker knows how to compensate the position of the two articulators, the linear relationships were very closed to those actually observed in the X-ray film data of the speaker. As implications of this finding, we propose a feedforward control model for compensatory articulation and then discuss on the organizations of motor control at different levels.

INTRODUCTION

In our previous factor analysis of X-ray film data showed that the tongue contours in the mid-sagittal plane can be described by four orthogonal components, *i.e.*, "articulators" [1]. Since the contribution of each component upon the tongue contours is specified by a single parameter, the observed contours are determined by the following four parameters; an extrinsic parameter, the jaw position (**jw**), and three intrinsic ones, tongue-body position (**tp**), tongue-body shape (**ts**), and tongue-tip position (**tt**). The statistics indicated that these four parameters explain more than 90% of the variance of the observed tongue contours. In this method, the set of parameter values can be calculated from an observed tongue shape. Thus, the variation of the tongue shapes along sentences is described by the corresponding frame-by-frame variations, *i.e.*, temporal patterns, of these articulatory parameters.

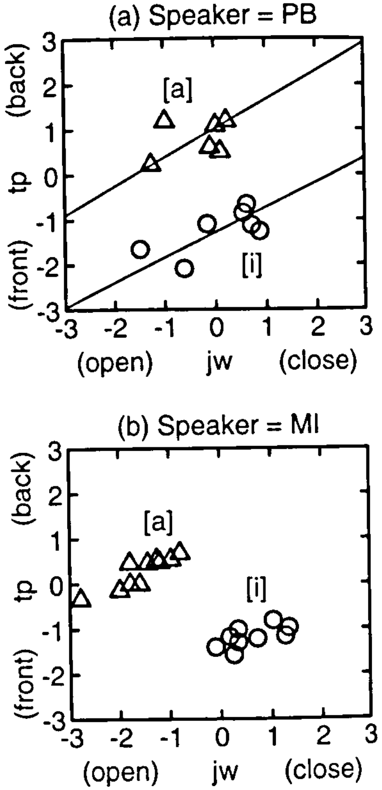

Figure 1. *Scattergrams of jaw (jw) vs. tongue-body positions (tp) at the "targets" for /i/ (circles) and /a/ (triangles) determined from the measured movements of the articulators in sentences read by PB in (a) and in nonsense words read by MI in (b).*

One of findings was that the temporal patterns exhibited a large contextual variability for the same vowel. In fact, this contextual variability seemed to be far greater than the inter-speaker variability for the same vowels. Figure 1a shows scattergrams on the **jw-tp** space concerning with two extreme unrounded vowels, /i/ and /a/, extracted from the different phonetic contexts in the 10 short French sentences uttered by a speaker (PB) [2]. Similar scattergrams are shown in Figure 1b where the data points were derived from X-ray film data for /pV1CV2/ nonsense words read by another speaker (MI). Those data points in the figures correspond to "targets" defined as turning points along the jaw-tongue trajectories during the vowels. The unit of both **jw** and **tp** is standard deviations. The value of these articulatory parameters rarely exceeds plus and minus three standard deviations. It is noticed that the points corresponding to each of the two vowels covers a large range. The dispersions are so large that jaw position or tongue position alone cannot distinguish these two extreme vowels, in particular for the speaker PB, exhibiting a significant degree of the context dependent variability in jaw and tongue positions for the same vowel. Nevertheless, the scattered points for the two vowels in both figures are well separated from each other.

When the first and second formant frequencies (F1 and F2, respectively) were calculated from those parameters, the corresponding F1-F2 scattergrams indicated a tighter pattern in comparison with the articulatory scatters [2]. We, therefore, recognize this large variability as a manifestation of the compensatory maneuver by the speakers. As the consequence of this, an unrounded vowel is not specified by a single set of parameter values, **jw** and **tp**, but rather by a compensation rule which determined (infinite) possible combinations of parameter values appropriate to produce that vowel. So, the question we ask is how we can describe this inter-articulator compensatory co-ordination.

We note in passing that for rounded vowels, such as /u/ and /o/, the compensatory maneuvers between the jaw and the intrinsic lip aperture (not tongue-body position as in the case of unrounded vowels) are predicted from simulation experiments with the articulatory model. We don't have sufficient number of rounded vowel tokens to validate the prediction, however. For this reason, we deal only with unrounded vowels in this paper. Moreover, we deal with only two parameters, **jw** and **tp**, since they are dominant in the determination of the tongue contours, especially for unrounded vowels (see the paper by Bouabana & Maeda in this congress.) Also we deal with only the speaker PB, since the complete articulatory model, which enables us to calculate acoustic characteristics, is available only for that speaker.

AN OPTIMUM JAW-TONGUE COMPENSATORY RULE

The target points associated with each vowel seem to exhibit a linear relation. Notice that the target points for each vowel in Figure 1a are scattered around the straight line which corresponds to the first principal axes. This means that it is possible to predict, although approximately, the tongue position for a vowel by a linear function of the actual jaw position or *vice versa*. The straight lines in Figure 1a were determined by the principal component analysis on a small number of points, for example only seven points for /i/ and six points for /a/ in the case of PB. These lines, therefore, are not necessarily "optimum" as compensatory rules. The term "optimum" means that a rule defining the jaw-tongue co-ordination results in the minimum of F1-F2 variability, when **tp** and **jw** are varied according to the rule. We thus carried out a simulation experiment, independently of the data, to determine optimum rules, *i.e.*, the linear relationships, for these two vowels and plus the posterior vowel /ɑ/.

Let us empirically assume that **tp** is linearly related to **jw** as follows:

$$\mathbf{tp} = a\,\mathbf{jw} + b, \qquad (1)$$

where "a" is a slope coefficient and "b" is the intercept. There is no mathematical reason to assume **jw** as a function of **tp**. It appears to us more reasonable to assume that tongue position is a function of jaw and not *vice versa*. Let us denote the intercept as **tp0**. Then, the linear equation becomes

$$\mathbf{tp} = \mathrm{a}\,\mathbf{jw} + \mathbf{tp0}. \qquad (2)$$

For given values of "a", a change in **tp0** (front/back tongue position) would result in a change in the phonetic value of vowel. It is easy to see, for example in Figure 1a, that varying **tp0** would vertically slide the straight line up and down, resulting in a change in the phonetic value of vowel, such as between /i/ and /a/.

It seems then reasonable to explore the variation of the acoustic variability index in function of the slope "a" for a fixed **tp0** value appropriate for each of the vowels /i/, /a/, and /ɑ/. Since the optimal rules should remain at the vicinity of the data derived principal axis, as already shown in Figure 1a, we used the following somewhat *ad hoc* search scheme to find out optimum rules. For a specified intercept **tp0**, we calculate variability index varying the value of "a" to determine its optimum value that results in the minimum index value. We then verify whether the determine value of "a" is also optimum along that line. To do so, we select anchor points, (**jw1**, **tp1**) along the determined line, such that **tp1** = a **jw1** + **tp0**, where **jw1** = -2, -1, 0, 1, and 2. The index values, therefore, are calculated varying the slope "a" for the five different straight lines defined as

$$\mathbf{tp} = \mathrm{a}\,\mathbf{jw} - (\mathrm{a}\,\mathbf{jw1} - \mathbf{tp1}). \qquad (3)$$

If the acoustic variability index determined with **jw1** = 0 is always smallest, we could conclude that the determine rule is indeed optimum and that the straight line is a reasonable specification of the compensatory rule.

Now let us define the acoustic variability index. We use an averaged normalized variance of F1 and F2 frequencies as described as follows:

$$v = 100\sqrt{\frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_{1_{max}}^2} + \frac{\sigma_2^2}{\sigma_{2_{max}}^2}\right)} \quad (\%) \qquad (4)$$

where σ_i^2 (i = 1, or 2) is the variance of F1 or F2, respectively. The first and second formant frequencies are calculated using the articulatory model. The model specifies the vocal tract shapes with seven parameters including **jw** and **tp**. The value of **jw** is always varied from -3 to +3 with 1 standard deviation as the step size and the corresponding value of **tp** is determined by Eq. (3). Values of the remaining parameters, **ts**, **tt**, two lip parameters, and larynx position, are determined from the X-ray data frames. The area function and then formant frequencies were computed from the model-specified vocal-tract shape. The acoustic calculations take into account the effects of the non-rigidity of the tract walls and of sound radiation from the lips. The F1 (and F2) variance, finally, is computed from the seven different combinations of **jw** and **tp** associated with the straight-line rule. An example of such calculations is illustrated in Figure 2 for vowel [i].

The maximum possible variances of F1 and F2, $\sigma_{1_{max}}^2$ and $\sigma_{2_{max}}^2$ respectively, needed for the normalization are difficult to estimate. We, therefore, employed the values of half range of F1 and of F2 frequencies derived from the articulatory model. The half ranges, 300 Hz for F1 and 1000 Hz for F2, were determined from F1-F2 projections obtained by systematically varying the values of seven model parameters [3]. The result of the index calculations for the three vowels are shown in Table 1.

For the vowel /i/ shown at Table 1a, we choose the value of **tp0** as -1.3 at **jw0** = 0 based on the observation of the scattered points in Figure 1a. As described before, we calculated, first, variability index varying the slop from 0.3 to 0.8. The results are shown at the column, **jw1** = 0 and **tp1** = -1.3, framed vertically. The minimum variability occurs at slope ("a") equal to 0.6. Second, we selected four anchor points along the determined straight line rule, *i.e.*, **tp** = 0.6 **jw** - 1.3, which are listed at the top two rows. Note that variability indexes for a = 0.6, framed horizontally, have the identical value, since they are derived with the same straight line rule. It can be seen that acoustic variability is always minimum at "a" = 0.6 regardless of anchor points. Although the variability never becomes zero, in other words, the compensation isn't perfect, the linear relation appears adequate to specify the compensatory rule. It is noted, moreover, that the determined optimum value of slope, 0.6, favorably compares with the slope value calculated from the data points (shown in Figure 1a), 0.56. This good agreement suggests that the speaker knows an acoustically optimum way of co-ordinating the jaw and tongue-body movements in the production of this vowel in different phonetic context.

Table 1. *F1-F2 variability index (in %) for three vowels as a function of the slope coefficient "a" calculated at different anchor points (jw1, tp1) using Eq. 3.*

(a) Vowel /i/, where **tp1** = 0.6**jw1** - 1.3

jw1	-2	-1	0	1	2
tp1	-2.5	-1.9	-1.3	-0.7	-0.1
a					
0.3	127.5	96.6	23.1	19.3	16.9
0.4	95.4	15.7	14.4	12.8	11.9
0.5	7.5	6.9	6.8	6.7	6.4
0.6	4.6	4.6	4.6	4.6	4.6
0.7	9.8	10.7	10.4	10.6	11.4
0.8	15.8	16.7	17.9	23.5	100.8

(b) Vowel /a/, where **tp1** = 0.5**jw1**+ 0.9

jw1	-2	-1	0	1	2
tp1	-0.1	0.4	0.9	1.3	1.9
a					
0.3	16.2	14.7	14.1	14.0	12.6
0.4	11.4	11.3	10.9	10.9	10.6
0.5	9.4	9.4	9.4	9.4	9.4
0.6	9.8	9.9	9.6	9.5	9.5
0.7	12.4	12.1	11.8	11.3	11.2
0.8	15.9	15.3	15.0	14.4	14.3

(c) Vowel /ɑ/, where **tp1** = 0.4**jw1**+ 2.6

jw1	-2	-1	0	1	2
tp1	1.8	2.2	2.6	3.0	3.4
a					
0.2	12.2	11.9	11.4	10.9	10.3
0.3	10.0	10.0	9.7	9.8	9.8
0.4	9.5	9.5	9.5	9.5	9.5
0.5	11.6	11.0	10.9	10.5	10.3
0.6	16.1	14.4	11.3	12.4	11.5
0.7	24.9	20.4	16.5	14.8	13.8

The calculations concerned with the vowel /a/, shown in Table 1b, indicate that variability index becomes minimum at the slope value equal to 0.5. This is always the case regardless of anchor points. However the optimum slope, 0.5, compares less favorably with that determined from the data (shown in Figure 1a), 0.63. This discrepancy might be due to, in part, the small number of the data points with a relatively large dispersion of the measured data points. In order to assess a general trend of the optimum slope in a function of **tp0**, *i.e.*, depending on the vowel identity, we calculated variabilites for the posterior vowel /ɑ/. The result is shown in Table 1c which indicates the optimum slope value of 0.4. From these calculations, the general trend appears to be a regular decreasing of the slope value from the front to back tongue-dorsum position.

Figure 2 shows the vocal-tract profiles and frontal lip shapes at the left and the corresponding transfer functions at the right for the vowel /i/. In this calculation, the value of jaw position was also varied from -3 to +3 with the step size of one standard deviation, as described before. The corresponding variations in the profile and vocal-tract transfer functions are indicated by the arrows. We used the jaw-tongue compensatory rule with the optimum slope that equals to 0.6. Since only the tongue-body position is compensated against a change in the jaw position, the dimensions of the lip tube, modelled by a uniform elliptic cylinder, vary significantly depending on the jaw position. Acoustically speaking, a decrease in the pharyngeal cavity volume is not important for F1 frequency, if that variation is compensated by the narrowing of the mouth channel (F1 of /i/ corresponds to the Helmholtz resonance.) Such an F1 compensation just occurs with "a" = 0.6. The F1 deviation due to the closing jaw is "corrected" automatically, but the pharyngeal cavity length is not. The non-corrected length change is the cause of the F2 drift toward lower frequencies, as seen in Figure 2.

When the vocal-tract transfers of the vowel /a/ were calculated with the optimum slope value of 0.5 and the intercept of 0.9, relatively large F2

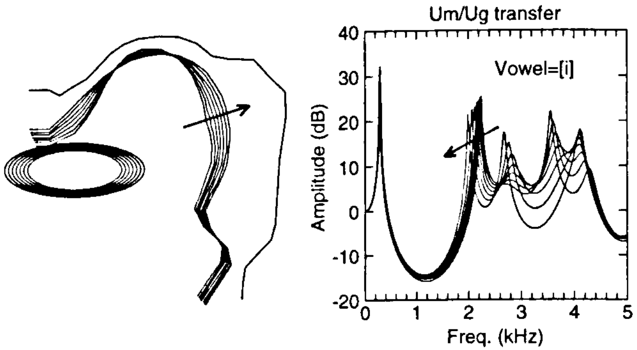

Figure 2. Calculated VT profiles and frontal lip shapes at the left, the corresponding VT transfer functions at the right for the vowel /i/: (a)The tongue-body position is varied according to the compensatory rule, tp = 0.6 jw - 1.3. The arrows indicate the direction of variations from jw = -3 (close) to +3 (open).

variation and, to a lesser extent, F1 variation occurs for this vowel. These variations are due to the lengthening of the front and back cavities combined with the narrowing of the lip aperture as the jaw position varies from low to high position. As consequence of this, the tongue-body compensation maneuver is less effective than in the case of the previous vowel /i/. This fact is manifested in the variability index value of 4.6% for /i/ and that of 9.4 % for /a/. The posterior vowel /ɑ/ exhibited the F1 and F2 variations similar to the vowel /a/. The variability index at the optimum condition (a = 0.4 and **tp0** = 2.6) was 9.5%.

Speech synthesis experiments confirmed, in particular for the vowel /i/, a stability of the perceptual phonetic value under the jaw-tongue compensation specified by the linear relation, even when jaw position was varied covering its maximum range. The vowel /a/ also exhibits the stability, but as expected, within a narrower range of the jaw position variation than for the /i/ vowel.

VOWEL SPECIFICATION AND MOVEMENT CONTROL: DISCUSSIONS

It is well known in the domain of motor control that there are two kinds of movements, fast and slow. The duration of fast movements is 200 ms or less. Articulatory movements, therefore, belong to fast movements. Since the latency in the neural transmission of sensory information can be more than 200 ms or longer, there is no way to adaptively control articulatory movement by means of sensory feedback. The present compensatory rule enables us to postulate a simple feedforward "compute-and-control" scheme without feedback. The term "compute" means here a simple arithmetics such as Eq.2 defined before. A feedforward scheme for tongue-body position control is conceptually illustrated in Figure 3. The input to the control "box" is assumed to be the value of invariant target or command associated with a vowel and the current jaw position as an "afferent" information. We don't think there exists an invariant jaw position for a given vowel. Jaw position is influenced by stress and suprasegmentals as well. Here we assume that the position is provided somehow. Recalling the compensatory rule Eq.2, the input to the tongue position would be the values of the intercept and the slope, *i.e.*, **tp0** and "a", for unrounded vowels. The actual tongue-body position is calculated with rule. If the slope value can be approximated by a single value for all unrounded vowels or predicted from the intercept itself, only the specification of **tp0** where its value depends on the vowel identity is necessary. It is noted that the outlined control scheme should be recognized as a feedforward or an open-loop control, since the sensory information comes from the jaw and not from the tongue itself.

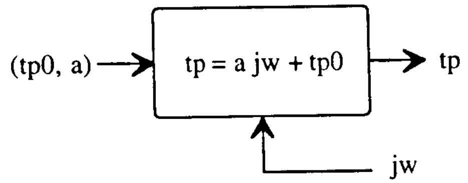

***Figure 3.** A conceptual model of feedforward control for tongue-dorsum position **tp**. The box can be thought of as a "motor program". The program receives two inputs; "command" (**tp0** and the slope 'a') and an "afferent" information about jaw position, **jw**.*

The compensatory rule was derived from position data and naturally it relates between two positions. An obvious question is whether or not the positional variables have neurobiological reality as a control parameter of the muscles. There seems to be no definitive answer yet in the literature about the neural coding of positions. It should be mentioned, however, that in the tongue system, the extrinsic muscles seem to be organized in a clean way for speech production: the genioglossus posterior and hyoglossus form an antagonistic pair controlling high-front to low-back orientation of the tongue movements. The styloglossus (SG) and genioglossus anterior (GGa) are also considered antagonistic; the SG pulls the tongue body toward the high-back direction, whereas the GGa compresses the tongue in the opposite direction [4]. A vowel, then, is produced by selecting and activating one of the paired muscles, which is termed as "muscle group selection (MGS)" [5 & 6]. This kind of control organization can correspond to an elementary motor program.

In the muscular system, moreover, the contractile force and position (displacement) are related to each other, as a gross approximation, by a proportional principle (Hooke's law) [7]. In fact, our simulation experiments have shown that the values of positional articulatory parameters, **tp** and **ts**, can be determined, by the linear law, from the EMG activity patterns of these paired muscles. The F1-F2 patterns calculated from EMG data formed a reasonable English vowel pattern [8]. It is then safe to state that the output of the arrow-and-box control model depicted in Figure 3 can be understood as a net force to be generated by the selected group of the muscles in order to produce an intended vowel. Whereas the input can be regarded as the corresponding invariant specification of tongue position in terms of force. It follows that the control model situates between the motor organisation at the low-level, *i.e.*, the level of MGS or of elementary motor programs, and that at the high-level, where mappings between invariant auditory and articulatory representations of speech occur [6]. Thus the proposed control model could be considered as a motor program operating at the intermediate level, interfacing the high and low level motor organizations.

Although many details must be worked out, such a control model has the following three attractive characteristics:

(1) The model can explain the mechanisms of the bite-block vowels in straight forward manner, say, "compute and control" instead of "simulate and control" proposed by Lindblom [9]. It is noted that a series of papers has been published concerning with the "motor equivalence" that explains the jaw and lip co-ordination to maintain relatively invariant lip aperture regardless of a relatively large individual variations of these two articulators ([10] & [11] to mention a few). Motor equivalence was postulated from observations such that when the jaw contributed a large displacement to the opening or closing of the oral cavity, the upper lip and lower lip contributed proportionally less and conversely. It assumes the existence of afferent pathways, from the jaw to the lips, that provide information to adjust lip position by means of an open-loop control. Neurological evidence of such

open-loop control mechanism is found in the vestibulo-ocular reflex (VOR) [12]. VOR contributes to automatically stabilize retinal image against fast head movements. Its reaction time can be as fast as 10 ms. It should be noticed that there is a distinctive difference between the motor equivalence and our proposed feedforward control. The motor equivalence automatically operates as reflex and its control is not intentional. It contributes to the "correction" against local perturbation. Our feedforward scheme is to control an intended articulation. Thus we think that our proposed feedforward control operates at a higher level, presumably, at the level of the motor programming in the motor cortex and in the cerebellum.

(2) The model can handle the reflex reversal in speech production. Kelso *et al.* [13] reported that when jaw position had been unexpectedly perturbed during the vowel in /baeb/, the subject adjusted the lips, whereas during /baez/, the tongue was adjusted. This phenomenon can be explained by assuming that the selected motor program, such as depicted in Figure 3, activates the jaw-tongue sensory pathway, for example, in /baez/ but not in /baeb/.

(3) Since the final (and observable) tongue position is calculated as a function of the actual jaw position in the down stream using the explicit compensatory rule, the specification of tongue position for a given vowel in the motor program can be invariant as described above.

It is quite natural for one to raise a question about the neurobiological validity of such control model. To our knowledge there is no neurobiological evidence directly supporting the proposed control model. It is only possible to infer the model reality from the known functions of the human neurobiological system. The jaw-tongue and jaw-lip co-ordination including sensory pathways should functionally exist, at least, for vegetative functions such as food intake. Chewing food requires an intricate co-ordination between jaw and tongue movements. However, it is not clear whether or not the same co-ordinative mechanism is employed for speech. If the jaw-tongue co-ordination is acquired during learning speech, it might be the case that the basic (and innate) co-ordinative mechanisms including the selective sensory activation are specifically tuned for speech production, creating a new speech motor program. Learning the speech skill can be viewed as the process of such a "turning". It is known that the cerebellum plays a crucial role in learning of motor skill. It is suspected then that speech motor programs, at least part of them, are memorized in the cerebellum and they contribute to the formation of speech motor commands. A motor program is an abstract and functional concept, however, and consequently it could be wrong to consider that the speech motor programs and their functions are concentrated only in the cerebellum.

In summary, with the feedforward control of tongue movement, vowels can be specified by invariant muscular activity patterns (the input to the control system) which undergo lawful modifications according to the compensatory rule. Then a plausible scenario would be that in the stage of learning speech, the control process including feedback control play an important role in the speech production. But once the skill is acquired, the mode is transferred to the feedforward control discussed here. Or rather, when such change in the mode of control occurs, we recognize that a speaker has mastered how to speak.

ACKNOWLEDGEMENT

The first author acknowledges the support of the European project Esprit/BRA, N° 6975, SPEECH MAPS.

REFERENCES

[1] Maeda, S. (1990). 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.

[2] Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, **19**, 321-331.

[3] Boë, L-J. (1993). Speech Maps Interactive Plant "SMIPS", *Deliverable 6* (pp.3-20) of SPEECH MAPS (ESPRIT/BR N° 6975).

[4] Kusakawa. N., Honda, K. & Kakita, Y.(1993). Construction of articulatory trajectories in the space of tongue muscle contraction force. *ATR Technical Report*, TR-A-0171 (in Japanese).

[5] Honda, K., Kurita, T., Kakita, Y. & Maeda, S. (in press). Physiology of the lips and modeling of lip gestures. *Journal of Phonetics*.

[6] Honda, K. (in press). Organization of tongue articulation for vowels. *Journal of Phonetics*.

[7] Bouabana, S. & Maeda, S. (1994). Modélisation des mouvements articulatoires par la méthode de la LPC multi-impulsionnelle. presented at Troisième congrés français d'acoustique (Toulouse), in *Journal de Physique*, **4** (Colloque N° 5, Suppl. JP III, N° 5), 449-452.

[8] Maeda, S. & Honda, K. (1994). From EMG to formant patterns of vowels: the implication of vowel spaces. *Phonetica*, **51**, 17-29.

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

[10] Folkins, J.W. & Abbs, J.H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research*, **18**, 207-220.

[11] Gracco, V.L. & Abbs, J.H. (1988). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, **65**, 156-166.

[12] Ito, M. (1974). The control mechanisms of cerebellar control systems. In F. O. Schmitt & F.G. Warden (Eds.) *The Neuroscinces Third World Study Program*, MIT Press, 293-303.

[13] Kelso, J.A.S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C.A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812-832.