Mechanical properties of lip movements: How to characterize different speaking styles?

Shinji Maeda† and **Martine Toda**‡
† CNRS-URA820 and ENST-Dept. TSI, Paris, France
‡ Université Paris 3-ILPGA, France and ATR-HIS, Kyoto, Japan
E-mail: maeda@tsi.enst.fr and mtoda@his.atr.co.jp

ABSTRACT

Face deformations during speech in different styles are observed by measuring movements of 65 markers glued on the subject's face including the lips. A factor analysis decomposes lower lip movements generated by jaw gestures and those by intrinsic lip gestures. We then characterize mechanical properties of these movements in terms of global variables calculated over an utterance as the total movement time, distance traveled, average speed, and some variables related to muscular energy expenditure. Both rapid and clear speech require high muscular energy expenditure. The energy requirement in the former is a demand for power to maintain motion speed at a high level, whereas that in the latter is that for force to maintain a high level of accelerations along the utterance.

1. INTRODUCTION

In order to properly control an articulatory synthesizer, it is important to understand the structural organization underlying articulatory movements and, perhaps more basic, their properties imposed by the biomechanical characteristics of the muscles that act as the motor of the movements.

The purpose of this paper is, then, to examine such mechanistic properties of motions by analyzing kinematic data of the face deformations under different speaking styles as slow, normal, rapid, and hyper articulated speech. Our study, therefore, is in the same line as that reported by Perkell *et al.* [1] in the articulatory level and by Moon and Lindblom [2] in the acoustic level. Our study is different from [1] and [2] in two aspects, however. First, we use motion variables, which are global in the sense that their values are calculated over the whole length of an utterance. We assume here that a speaking style affects throughout the utterance, and thus global variables and their relations could show up their difference better than variables defined over a short segment, like a syllable.

Second, we concentrate here the analysis of lower lip motions, given that the lower lip is the most mobile part in the face during speech. We don't directly calculate global variables from observed lip motions however. Rather the observed motions are decomposed into those generated by the jaw gestures and others by the intrinsic lip gestures using a factor analysis [3]. We think that the characteristics of the muscles underlying a particular gesture manifest better on the decomposed motions than otherwise.

2. DATA AND PROCESSING

One American English male speaker (Speaker-A) and two French male and female speakers (respectively, Speaker-V and Speaker-P) read a corpus consisting of a sequence of nonsense VCV syllables and a short text in the corresponding language. These three speakers were instructed to read the VCV syllables with a clear or a hyper articulation and the text with three different speaking rates, slow, normal, and rapid. In the English VCV sequence, vowel (V) includes /i/, /a/, and /u/ and 24 consonants (C). These three vowels and plus the high front rounded vowel /y/ are combined with 20 consonants in the French VCV sequence. English text consists of 28 syllables and the French text 62 syllables.

Maeda et al. in [3] have reported, for Speaker-A, a detail of the data acquisition and of the statistical analysis to extract uncorrelated articulatory factors that efficiently describe the measured face data. The same method was used for the data from the two French speakers. Briefly, a Vicon Motion Capture machine with six infrared video cameras tracked 3D coordinates of 61 markers glued on the Speakers-A's face. For the two French speakers, eight cameras tracked 63 face markers. For all the three subjects, common 61 face markers were approximately placed at the same relative locations. The camera speed was always 120 frames/s.

Before the factor analysis, effects of head motion on the positions of markers are eliminated by a head alignment. Let Y be a matrix of head-aligned motion data of a speaker. For example, Y of Speaker-V consists of 171 interlaced 3 coordinates of 57 markers in column and 23164 frames of observation in row. Y is normalized to obtain its z score as

$$Z=(Y-m)/s, \qquad (1)$$

where m and s are, respectively, mean and standard deviation vector. Then, Z is assumed to be a weighted sum of factors as

$$Z=AX, \qquad (2)$$

where A is a matrix of weights that can be determined by a factor analysis and X is the frame-by-frame succession of factor values, i.e., factor scores.

ISBN 1-876346-48-5 © 2003 UAB

We use an arbitrary orthogonal factor analysis (AFA) followed up by a standard PCA [3]. A factor analysis determines factors so that they explain structures of correlations between data variables (in our case Z) with a constraint as maximal extractions of the variance in PCA. In AFA, we can (arbitrarily) specify for a factor to extract a particular correlation structure. We, therefore, determine the first factor (f1) so that it extracts the correlations between the z-coordinate of a marker on the chin and the three coordinates of all other face markers. Since this z-coordinate can be considered as a measure of the vertical jaw position, f1 represents effects of vertical close/open jaw motions upon the face including the lips. In this way, we determine the second (f2) and then third factor (f3) representing, respectively, effects of horizontal front/back jaw motions (along the y-axis) and those of horizontal left/right motions (along x-axis). The principal factors are determined from the residual after the extraction of these first three arbitrary factors. Table 1 summarizes variances explained by each of the first six factors determined for the three speakers.

	Arbitrary factor: Jaw			PCA		
Speaker	f1	f2	f3	f4	f5	f6
A	31	13	11	26	7	3
V	34	9	9	17	15	3
P	21	23	8	28	7	3

Table 1: Variances (%) explained by the first six factors of the three individual speakers.

A visual inspection of the effects of individual factors indicates that those of Speaker-A are related to a particular articulatory gesture (or maneuver) in a simple manner. As responses to the first three factors, the lower lip appears to passively follow the jaw motions, close/open (f1), front/back (f2), and left/right (f3). The upper lip, especially in the central region is, hardly affected by these jaw factors. The PCA derived factors can be interpreted as follows: f4 corresponds to a horizontal lip control, spreading and closing (with rounding) of the lip tube, f5 to a vertical control, opening (with protrusion) and retracting, and f6 a raising and lowering of the cheeks and of the upper lip. Factors of the other two speakers, especially of Speaker-P however, are less clean than those of Speaker-A in the sense that their factors, except f1, often represent a combination of articulatory maneuvers observed in Speaker-A. This is one of reasons for grouping the jaw gesture specified by f1, f2, and f3 together and the intrinsic lip gestures by f4, f5, and f6 in the decomposition of the lower lip movements.

Since the first six factors of all the speakers explain about 90% of the variance, the factors higher than f6 can be neglected. Let us denote the truncated factor-weight matrix as $A_{tr.}$ The observed positions of markers are approximately calculated by

$$Y \cong (A_{tr} X)s+m. \qquad (3)$$

This equation can be regarded as a face synthesizer having a face model A_{tr} and a model parameter set X. A_{tr} is calculated from the entire face data of each speaker. X can be calculated for a selected utterance corresponding to a text or a VCV sequence. We denote a selected face data as Ys and its normalized version as Zs, then the corresponding face parameter set, Xs, is calculated from the inverse of Eq. (2) as

$$Xs=ZsA^{-1} \qquad (4)$$

Applying Xs to Eq. (3), we re-synthesize Ys. In the jaw-lip decomposition therefore, to obtain the lip marker motions generated by to the jaw gesture, we set the values of the three lip factors equal to zero for all the frames in Xs. Contrarily, to compute those generated by the lip gesture, we let the values of the three jaw factors in Xs equal to zero. Figure 1 shows portions of decomposed vertical motions of lower lip generated by the jaw gesture, calculated for the same text uttered with the three different rates.

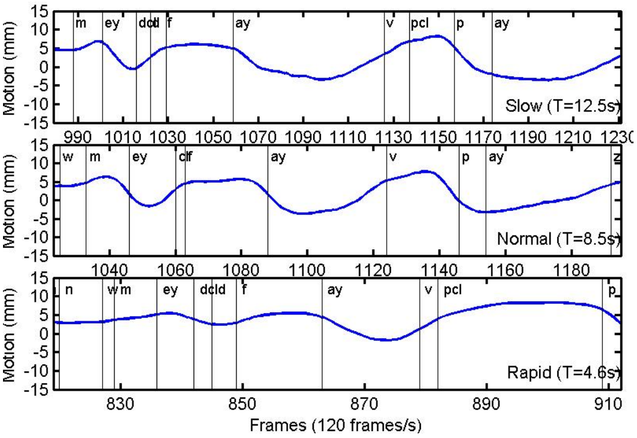

Figure 1. Vertical motion of the lower lip generated by the jaw gesture in portions, "…made five pies…", of slow, normal, and rapid utterance. The time scale of slow and of rapid utterance was, respectively, compressed and expanded linearly so that their duration matches with that of the normal utterance.

3. MOTION AND ENERGETIC VARIABLES

In order to characterize the lower lip motion in different speaking styles, the following basic global variables are considered: Utterance duration (T); Motion time (MT), T minus total pause duration; Total distance traveled during MT (D); Utterance average speed (Vav), =D/MT. Moreover, the range of variations due to different speaking rates is calculated for these variables. This range is defined as the maximum/minimum difference divided by the maximum and is expressed in %.

In addition to these basic variables, we include those related to the muscular energy expenditure. The calculation of such variables requires a formula relating a variable that can be determined from observed motions to energy expended by muscles that generate the motion. According to Hogan [4], input power requirement for muscles is proportional to the square of muscle force output. The mass involved in the motion generated by the muscles is unknown, but if we can assume the mass being constant, then force must be proportional to the acceleration, u, that is the second-order derivative of the observed lip motions. Since energy is the time integral of power, the muscular energy expenditure E in discrete formulation becomes

$$E = k\Delta T \sum_{n=1}^{N-2} u(n)^2 \text{ (joules)}, \qquad (5)$$

where ΔT denotes the inverse of video frame rate (120 frames/s), n the index of frames, and N the number of frames in an utterance. Although the value of the constant k is unknown, still Eq. (5) can give us the energy expenditure in an arbitrary energy scale. In order to compare utterances of the text and that of the VCV sequence for a given speaker E must be normalized as E/MT or E/D. Note that E/MT has, the physical dimension of power and E/D that of force.

4. RESULTS

Derived values of the range shown in Table 2 indicate that Speaker-A varied T (and also MT) much more than the two other speakers in a function of speaking rates. A considerably higher range of T relative to MT of Speaker-P suggests her heavy use of pauses to slow down utterance speed. For example, pauses occupy 8.4s of the slow utterance having the duration of 21.2s.

	Speaker-A		Speaker-V		Speaker-P	
	T (s)	MT (s)	T(s)	MT(s)	T (s)	MT (s)
rapid	4.6	4.4	10.2	7.1	11.9	8.2
normal	8.5	8.3	12.7	9.0	14.4	9.4
slow	12.5	12.3	15.1	10.8	21.2	12.8
range	63 %	64 %	33 %	34 %	44 %	35 %
VCV	25.1	19.8	96.1	28.8	121.5	37.3

Table 2: Measured utterance time T and motion time MT of a text read with three different rates and of the sequence of nonsense syllables, VCV.

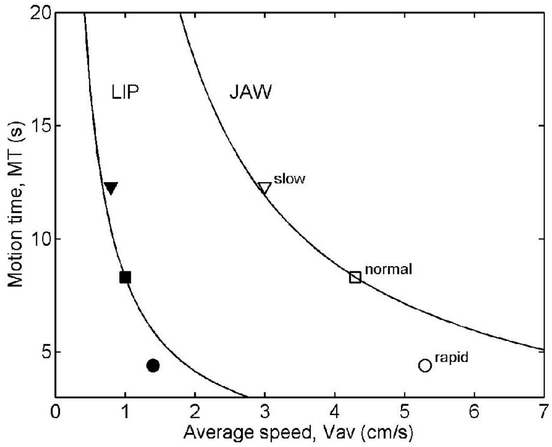

Figure 2: MT in function of Vav for Speaker-A. Each of two curves indicates the prediction by the inverse law.

The results related to the remaining variables are summarized in Table 3, for only Speakers-A and V, since the results from Speaker-V and -P are relatively similar to each other. Let us discuss relationships between the basic variables, T, D, and Vav. The values of D indicate that its range of variations for Speaker-V (also –P) is much smaller than that of Vav. In Speaker-A, we see that these ranges don't differ very much. This is because the difference in D between normal and slow is rather small, whereas D of the rapid utterance is much shorter than D of normal and slow utterances, suggesting something special occurred in the motion of the rapid utterance. Figure 2 may clarify that.

(a) American-English Speaker-A

	D (cm)	Vav (cm/s)	E/MT	E/D
Motion generated by jaw gesture				
rapid	24.2	5.5	74.6	13.6
normal	36.8	4.4	38.1	8.6
slow	38.8	3.2	24.1	7.7
range	38 %	43 %	68 %	43 %
VCV	65.0	3.28	49.9	15.2
Motion generated by Intrinsic lip gesture				
rapid	6.8	1.5	11.0	7.2
normal	9.4	1.1	6.0	5.3
slow	10.8	.90	3.8	4.4
range	37 %	43 %	65 %	39 %
VCV	20.4	1.0	13.6	13.2

(b) French Speaker-V

Motion generated by jaw gesture				
rapid	31.3	4.4	71.4	16.2
normal	34.3	3.8	29.1	7.6
slow	36.5	3.4	19.3	5.7
range	14 %	23 %	73 %	65 %
VCV	67.2	2.3	10.1	4.3
Motion generated by Intrinsic lip gesture				
rapid	10.0	1.4	14.5	10.5
normal	10.4	1.2	8.2	7.1
slow	11.1	1.1	5.7	5.5
range	11 %	26 %	61 %	48 %
VCV	24.6	0.9	2.9	3.4

Table 3: Values of the global variables related to lower lip movements due to the jaw gesture (f1, f2, and f3) and to the intrinsic lip gesture (f4, f5, and f6).

If a speaker varied MT by changing instantaneous speed constantly proportional to Vav along an utterance and maintained the motion path invariant, and therefore D constant, then MT could be predicted by an inverse law, because MT=D/Vav. The two curves in Figure 2 illustrate the prediction by the inverse law. Both for motions generated by the jaw gesture (the data points are indicated by open markers) and for those by the intrinsic lip gesture (closed markers), we set D equal to the corresponding values derived from normal utterance. In the case of Speaker-V (not shown), both data points, for slow and rapid were located close to their prediction curves. It can be stated, then, that Speaker-V varies MT by just speeding up or slowing down the normal utterance, whereas Speaker-A

ISBN 1-876346-48-5 © 2003 UAB

also slows it down to produce the slow utterance but for the rapid utterance, he not only speeds it up but also shortens D to further reduce MT. In fact, the movement for the rapid utterance shown in Figure 1 seems to be much smoother than the others indicating a shorter D and thus an articulatory reduction.

Let us now examine the energetic variables with a particular question in mind: How can different speaking styles be characterized? Figure 3 illustrates E/D as a function of Vav for Speaker-A. The open markers indicate data points derived from the jaw induced lip motions and the closed markers from the lip motions generated by the intrinsic lip gesture. Each of two curves represents the predicted relationships derived by a square law. As seen in Eq. (5), E is proportional to the square of acceleration. If we assumed, as before, the motion velocity were scaled uniformly with Vav along an utterance, the acceleration undergoes also a proportional change in a function of Vav. Then, E and E/D (also E/T) must vary in proportional to the square of Vav. The two curves in Figure 3 are determined so that two curves intersect with the corresponding data points of the utterance with normal speaking-rate.

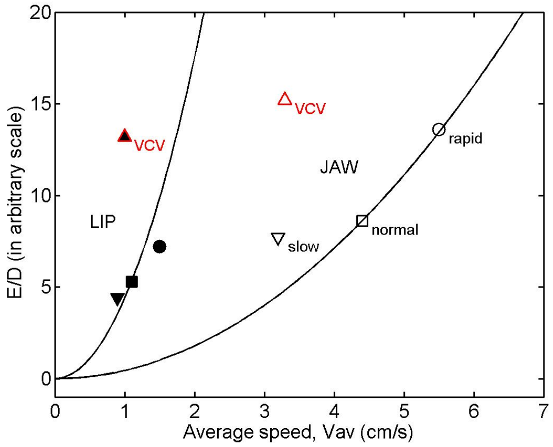

Figure 3: Muscular energy expenditure per distance (E/D) in a function of Vav for Speaker-A. Each of the two curves indicates the prediction by a square law.

It appears, first, that the data points from the utterances of the same text read in the three different rates follow the square law, except the jaw generated motions in the slow utterance, which seems to be hyper articulated to some degrees. The shortening of D in the rapid utterance could be a constraint imposed by the square law. Second, the two data points from the VCV utterance (indicated by the upward triangles) are located far above the corresponding prediction curves. It can be stated that hyper speech is characterized by high muscular energy expenditure per unit distance, E/D (force) for a given average speed.

Interestingly, the value of E/MT shown in Table 3 for the rapid utterance is greater than or close to that for the VCV utterance This is expected because E/MT=(E/D)Vav, which means E/MT increases proportionally with Vav. Motions in rapid speech, then, can be characterized by high muscular energy expenditure per unit motion time, E/MT (power).

As far as Speaker-V concerned, all the indications in the data presented in Table 3 are that the VCV sequence were uttered as if it was the text read with a slow speaking style without extra muscular efforts expected in a hyper mode of speech. An interesting implication to us is that the mechanical characteristics observed in the VCV utterance of Speaker-A is not due to the artificial VCV material but to the particular speaking style, i.e., a hyper speech.

5. CONCLUDING REMARKS

The analysis of lower-lip motions has indicated that both rapid speech and hyper speech require high muscular energy expenditure. However, power demand in virtue of a high average motion speed characterizes the former, whereas forces to maintain a high level of accelerations the latter. For any speaking style, a high average speed is expensive. The muscular energy expenditure, E, goes up with the square of Vav. Moreover, the inverse law, as shown in Figure 2, means that the effectiveness of increasing Vav to shorten MT quickly goes down at higher speeds. This might explain why two French speakers didn't vary much Vav and resorted to pause insertions to vary speaking rate. Moreover, it is tempting to speculate that muscles already reach to a biomechanical limit at a rate moderately higher than the normal speaking rate. The articulatory reduction observed in the rapid utterance of Speaker-A could be the manifestation of such a limit. Articulatory synthesis experiments may shed more lights on the mechanical nature of reduction and on its acoustic consequences.

ACKNOWLEDGMENT

This work was supported, in part, by the project SAALSA No. 7215/17/00 in the program PRIAMM of CNC, France.

REFERENCES

[1] J. Perkell, M. Zandipour, M. Matthies, and H. Lane, "Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modelling issues," *J. Acoust. Soc. Am.*, vol. 112, pp. 1627-1641, 2002.

[2] S.-J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," *J. Acoust. Soc. Am.*, vol. 96, pp. 40-50, 1994.

[3] S. Maeda, M. Toda, A. Carlen, and L. Meftahi, "Functional modelling of face movements during speech," *Proc. International Congress of Spoken Language Processing*, pp. 1529-1532, 2002.

[4] N. Hogan, "Adaptive control of mechanical impedance by coactivation of antagonist muscles," *IEEE Transaction on Automatic Control*, vol. **AC-29** (8), pp. 681-690, 1984.