A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers

Matthew S. Craig
Institute of Biomaterials and Biomedical Engineering and Oral Dynamics Laboratory, University of Toronto, 160-500 University Avenue, Toronto, Ontario M5G 1V7, Canada

Pascal van Lieshout
Department of Speech Language Pathology, Oral Dynamics Laboratory, Institute of Biomaterials and Biomedical Engineering, and Department of Psychology, University of Toronto and Toronto Rehabilitation Institute, Toronto, Canada

Willy Wong
Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

(Received 11 September 2007; revised 13 August 2008; accepted 14 August 2008)

The relationship between acoustic and visual speech is important for understanding speech perception, but it also forms the basis behind a type of facial animator, which can predict facial motion during speech given an acoustic input. This relationship was examined by revisiting a linear transformation model of audio-visual speech production. A mathematical model is constructed whereby the visual aspect of speech is reproduced from the acoustic signal via a linear transformation. Unlike previous studies in this area, this paper will address specific aspects of the model as related to the effects of window size for acoustic framing and the critical size of the training set. On average, facial motion is predicted with a correlation of 0.70 to the recorded motion, when the model is trained and then tested on the same subject. This is comparable to previous studies using either similar or different model approaches. Using a model trained on other subjects and then applying it to a new subject resulted in a prediction correlation of 0.65. Furthermore, acoustic windows of 100 ms and a data set of approximately 40 sentences are required for maximum predictability. The results are interpreted in terms of the underlying assumptions of the model.


PACS number(s): 43.70.Jt, 43.70.Bk, 43.72.Ct, 43.70.Gr [AL]

Pages: 3183–3190

I. INTRODUCTION

Speech is both an acoustic and visual means of communication. People with extensive hearing loss can understand speech to some degree by lip reading (Dodd, 1987), while even normal hearing individuals lip-read constantly while communicating (Summerfield, 1992). The perceptual congruency between acoustics and motion can lead to “talking heads:” acoustic-to-visual models that predict and animate facial motion from acoustic input.

If synthetic animations of a face can be derived from acoustics alone, this would allow the recreation of the talking face with a significantly lower bit-rate when compared to transmitting video images. These talking heads can further be useful in the development of assistive devices for people with communication disorders, like the hearing impaired or for patients with limitations in facial expression, for example as found in those suffering from Parkinson’s disease. The rehabilitation of facial movement during speech, along with correcting problems with articulation for voicing, is essential to improve the quality of life of such patients. Using animation as a biofeedback device, patients might regain some functional facial motion during speech.

Current approaches involve the use of hidden Markov models (HMMs) or time delay neural networks (TDNNs) to predict motion from acoustics, as they are well-suited to model the nonlinear and time-asynchronous relationship between acoustic and visual speech. These models are designed to capture the effects of coarticulation, which would imply that the position of the face during speech cannot be predicted solely from the immediate acoustic parameters, but will also depend on values which precede and follow the current time instant. Accordingly, models will use as much as nine frames of acoustic information (i.e., the acoustic information in a time-window of nine frames of video information) to predict the facial position at a single frame (Fu et al., 2005; Savran et al., 2006). In the case of HMMs, which rely on artificial segmentation of the data, acoustic parameters are converted into phonemic units (i.e., a specific sound pattern for a given language) and the visual data into visemes (i.e., the basic facial configurations of visual speech). Animations are constructed by mapping each phoneme to its corresponding viseme. The motion between adjacent visemes is then interpolated for a smooth animation.

TDNNs search for a nonlinear transformation function that maps acoustics directly to facial motion, but also use acoustic information from adjacent frames to model coarticulation. The result is dependent upon the topology of the net-
work, the number of layers and nodes, and the initial estimates of node weights. Consequently, these models can vary widely between similar studies from as few as ten hidden units (Yehia et al., 2001) to as many as 600 (Massaro et al., 1999). Values between 25 and 100 are commonly used (Agelfors et al., 1999; Hong et al., 2002; Zelezny et al., 2006). Generally, phonemically richer speech data sets require more nodes, thereby exacerbating the “black-box” nature of TDNNs, and requiring considerable computational resources and time in the training phase (Hong et al., 2002).

Another popular method uses an audio-visual codebook (AVC) (Kakumanu, 2002; Savran et al., 2006). The salient acoustic and visual features of each audio-visual frame are extracted and stored as a pair in a codebook. Input acoustic features are compared to the acoustic features stored in the codebook until a “nearest match” is found. The visual partner of the nearest match is then used as the predicted facial motion for that frame. AVCs can also model coarticulation by using adjacent acoustic frames to identify the best nearest match (Kakumanu, 2002).

A much simpler model of audio-visual speech production involves the use of a multilayer regression (MLR) model (Yehia et al., 1998; Barker and Berthommier 1999; Jiang et al., 2000a, 2000b). MLRs relate acoustics to motion through a simple linear matrix operation. They perform frame-to-frame transformations from acoustic data to facial position. As a result, coarticulation cannot be explicitly modeled. MLRs do not solve for the transformation iteratively as do neural nets, nor do they require initial estimates. Although approximate, MLRs are thus substantially simpler than TDNNs, HMMs, and AVCs in terms of computational expense and objective physical interpretation.

Previous work has indicated that approximately 70% of facial motion can be recovered using a linear regression model, although these studies have been limited to small or restricted speech databases such as monosyllabic words (Jiang et al., 2002), bisyllabic words (Berthommier, 2003) or a small number of sentences (Yehia et al., 2001). Furthermore, these studies are based on a small number of subjects. The purpose of this study is to revisit the use of a linear regression model as a foundation for animation of human speech. Specifically, in extending previous work in this area, we use a significantly larger subject pool and a large set of phonetically varied stimuli to test the effect of the following model aspects on facial motion predictability:

1. Acoustic window size: typically about 10–16 ms of acoustic data is used to predict facial position at a given time. However, averaging a larger section of acoustic data may increase predictability by capturing adjacent context and smoothing facial motion trajectories.

2. Size of the data set: small training databases will create acoustic-to-motion transformations that are too narrowly defined and may not generalize well to new sentences. Our goal is to determine a minimum size of dataset for satisfying predictability.

3. Speaker independence: to see whether we can predict speech motions of subjects who are untrained in the system.

By examining these variables with a significantly larger subject pool, our goal is to present MLR as a conceptually and computationally simpler, yet comparably effective, basis for speech animation as compared to HMMs, TDNNs, and AVCs.

II. METHOD

A. Subjects

Sixteen subjects, eight adult males (M1–M8) and eight adult females (F1–F8), participated in the study. All subjects were over the age of 18 and spoke English as a first language, with the exception of subject M8, whose first language was Tamil but spoke English fluently. None of the subjects had a reported history of speech, language, or hearing problems and all gave informed consent before participating in the study.

B. Stimuli

Ninety sentences were used, 45 from the TIMIT sentences database and 45 from the HARVARD sentences database. The TIMIT sentences provide a rich mixture of phonemes (Brugnara, 2001) while the HARVARD sentences consist of phoneme proportions that are representative of their occurrence in the English language (Logan et al., 1989). Once intersentence pauses were removed from the data, the 90 sentences on average (depending on individual speech rate differences) contained 3 min and 20 s of continuous speech. To test how the size of a database would affect predictability, we ran predictions with sets containing different numbers of sentences, varying from a minimum of 2 to a maximum of 60 sentences (see Sec. III).

C. Audio processing

Speech was recorded using a WMS-PRO VHF condenser microphone with a sampling rate of 32 kHz. Audio time series were split into frames of 100 ms, with a 63.3 ms overlap. A shorter frame length of 10–50 ms is common in many speech coding systems (Kabal, 2003), including facial animators (Jiang et al., 2000a, 2000b; Yehia et al., 1998; Massaro et al., 1999). We found that facial motion predictability was improved with a longer frame in our time-invariant model. Frame sizes ranging from 20 to 150 ms were tested for the model described here (see Sec. III).

Each frame was multiplied by a Hamming window to reduce spectral distortions and the acoustic information was represented by an array of 65 parameters: The root mean squared energy, 16th order linear predictor coefficients (LPC), 16th order line spectral pairs (LSP), and the first derivatives of the previous two sets. LSPs are derived from LPCs (Jiang et al., 2000a) and are closely related to the formant values and bandwidths during speech (Fig. 1).

LPC and LSP coding is common in facial animators (Jiang et al., 2000a, 2000b; Yehia et al., 1998; Massaro et al., 1999). The use of time derivatives can be justified because they reflect the behavior of formant transitions which affect the perception of speech. Thus, the time derivative of LPC and LSP may also have a systematic effect on
characteristics of facial motion. Zelezny et al. (2006) used 52 parameters, including the first, second, and third derivatives of mel frequency cepstral coefficients, in their audio-visual speech synthesis. Time derivatives have also proven to be useful in speech recognition systems (Chou and Juang, 2003).

D. Video processing

Speech motions were recorded in three dimensions using two JVC DVL9800U digital video cameras placed approximately 1 m in front of the subjects’ faces, to the left and right. The cameras have a native shutter speed of 30 frames per second and a digital resolution of 720×480 pixels. Shutter speed was increased to 60 frames per second using interlaced scan mode, which effectively halves the vertical resolution.

Two 250 W ultraviolet black lights, fixed to the ceiling, were used to illuminate 13 glow-in-the-dark stickers, which are placed on the subjects’ faces at anatomical points of interest (see Fig. 2, left). Each sticker is 3.0+/-0.5 mm in diameter.

Four gestures (see Fig. 2, right) defined as the distance between two markers, were selected that represent facial motion in relation to relevant speech features: jaw, lip corners (LC), upper lip (UL), and upper lip / lower lip distance (UL/LL). Jaw and UL were measured in relation to a fixed point on the subjects’ foreheads. The UL gesture involves motion in the depth and vertical dimensions for protrusions and bilabial closures, the LC gesture covers mainly motion in the horizontal dimension, while UL/LL and jaw involve mainly vertical motions. Similar facial measures have been reported in the literature for linear regression studies as well as other visual speech synthesis models (Barker and Berthommier, 1999; Curinga et al., 1996; Jiang et al., 2000a, 2000b 2002; Kakumanu et al., 2006). The gestures were selected for clarity and simplicity, and represent anatomical motions that may be of interest during speech.

The recorded video from both cameras was processed using a specialized software suite, APAS, developed by Ariel Dynamics, Inc. (Trabuco Canyon, USA). APAS converts the locations of each marker in each video frame into a three-dimensional location, creating a time-series that can be read into Matlab®. This process is known as digitization. Figure 3 depicts the digitized three dimensional (3D) positioning of markers. For clarity, a polygon is added connecting the four points on the lips.

After digitization, gesture trajectories were smoothed using a Butterworth low-pass filter with a cut-off frequency of 16 Hz. Virtually all speech motions fall under 8 Hz (Craig et al., 2007).

E. System resolution

A detailed description of the system’s resolution and accuracy can be found elsewhere (Craig et al., 2007). The video cameras and digitization software are capable of resolving motions of amplitude as low as 0.7 mm with excellent accuracy. To be conservative, kinematic data from individual sentences with amplitudes less than 1 mm were discarded from this study. This was only the case for some instances of the UL gesture, which has an average amplitude of 1.2 mm for adult subjects during normal speech (Craig et al., 2007).

F. Multilinear Regression

A MLR approach is used to map acoustic speech to facial position. The MLR defines a transformation, T, that multiplies each acoustic frame, A, to predict each of the four
gestures of motion of each visual frame, \( V \). There is one transformation function for each gesture. Specifically, for a data set of \( n \) frames

\[
T \cdot A = V,
\]

where

\[
A = \begin{pmatrix}
A_{1,1}, & \ldots, & A_{1,n} \\
A_{2,1}, & \ldots, & A_{2,n} \\
\vdots & & \vdots \\
A_{p,1}, & \ldots, & A_{p,n}
\end{pmatrix},
\]

\[
V = \begin{pmatrix}
V_{1,1}, & \ldots, & V_{1,n} \\
V_{2,1}, & \ldots, & V_{2,n} \\
\vdots & & \vdots \\
V_{m,1}, & \ldots, & V_{m,n}
\end{pmatrix}.
\]

\( A_{p,n} \) represents acoustic parameter \( p \) in frame \( n \) and \( V_{m,n} \) represents visual gesture \( m \) in frame \( n \). “\( p \)” ranges from 1 to 65 for the 65 acoustic parameters, “\( m \)” from 1 to 4 for the four visual gestures, and “\( n \)” is the total number frames in the training set which is approximately equal to 12 000 frames (3 min and 20 sec at 60 frames per second).

In multilinear regression, the optimal transformation, \( T \), is obtained by minimizing the squared error between the predicted motion, \( P = T \cdot A \) and recorded motion, \( V \) for all frames.

\[
SE = \sum_{k=1}^{n} (P_k - V_k)^2.
\]

\( T \) is found by setting the derivative of Eq. (3) to zero. The resulting solution is unique (Ferguson, 1984)

\[
T = [(A) \cdot (A)^T]^{-1}(A) \cdot (V)^T.
\]

G. Model evaluation

For a given data set of acoustic and visual data for \( S \) spoken sentences, the data is separated into training and testing sets. \( S-1 \) sentences comprise the training set [i.e., are used to define the transformation \( T \) from Eq. (4)]. The final sentence is used as the test set: \( T \) is multiplied by the acoustic parameters of each frame of the test repetition, yielding four predicted facial trajectories (time series of distances), one for each gesture. The predicted trajectories are compared to the recorded trajectories for that sentence and gesture by means of a Pearson correlation coefficient (CC) and normalized mean square error (NMSE). This procedure is repeated for all different combinations of training and testing data. CCs relate how well the signal’s trajectories match, with a value of 1 being a perfect directional match, while NMSEs relate how well the signal’s amplitudes match. We expect a negative correlation between CCs and NMSEs (i.e., a well predicted signal will have a high CC and low NMSE). However, for some predictions it may be the case that directionality is well predicted but scaling is inaccurate, leading to both a high CC and a high NMSE. Unlike CC there are no upper boundaries for NMSE. A zero value obviously means that there are no amplitude differences between the predicted and actual time series.

III. RESULTS

When training and testing was performed for the same subject’s data, the overall CC was 0.69 and overall NMSE was 0.58 across all gestures and subjects (Fig. 4).

In most subjects, jaw motion was predicted best across sentences CC=0.77 \( \pm \) 0.06, NMSE=0.42 \( \pm \) 0.07, while UL (upper lip) motion was predicted worst, CC=0.59 \( \pm \) 0.11, NMSE=0.76 \( \pm \) 0.13. A one-way ANOVA revealed that the differences in CC between all gestures was significant [\( F (3,15) =40.5, \ p<0.001 \)]. The distribution of CCs, averaged across gestures, is shown in Fig. 5. A sample of recorded and

<table>
<thead>
<tr>
<th>Gesture</th>
<th>CC (STD)</th>
<th>NMSE (STD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaw</td>
<td>0.77 (0.06)</td>
<td>0.42 (0.07)</td>
</tr>
<tr>
<td>UL/LL</td>
<td>0.72 (0.07)</td>
<td>0.30 (0.04)</td>
</tr>
<tr>
<td>LC</td>
<td>0.68 (0.10)</td>
<td>0.63 (0.12)</td>
</tr>
<tr>
<td>UL</td>
<td>0.59 (0.11)</td>
<td>0.76 (0.13)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.69 (0.06)</td>
<td>0.58 (0.07)</td>
</tr>
</tbody>
</table>

FIG. 4. Overall gesture CCs and NMSEs across sentences.

FIG. 5. Distribution of sentence mean CCs, averaged across subjects and gestures.
predicted trajectories for the four gestures is shown in Fig. 6, illustrating different combinations for CC and NMSE values.

Scatter plots in general revealed high negative correlations between CC and NMSE, indicating that in most cases good predictions yielded relatively high CCs and correspondingly low NMSEs. Figure 7 shows a typical example for subject M6. The mean correlations between CC and NMSEs across subjects are reported in Table I. It was seldom the case that both CC and NMSE were high indicating that there is no systematic scaling error for the predicted signal. This implies that either CC or NMSE alone is sufficient to report our results. For the sake of comparison with other studies, we will present our results mainly as CCs. Only in the case of UL predictions, and rarely for LC were there significant outliers to this trend for a few specific sentences. For example, for the sentence “Nothing is as offensive as innocence,” in which the UL has no functional use, predicted trajectories were scaled out of proportion often leading to extremely high NMSEs and moderate CCs, as shown in Fig. 8.

Figure 9 (top) depicts the sensitivity of predictability to acoustic frame size. Predictability was highest at a frame size of 100 ms, justifying the use of the 100 ms frame size for all subjects in these analyses. Figure 9 (bottom) shows the dependence of predictability on the size of the training database. At a training dataset of 37 sentences, predictability has reached 99% of its maximum value, but as shown in this figure, differences in predictability for data sets larger than 20 are very small.

Another purpose of the present study was to determine whether our system would be successful in predicting facial motion of a subject whose data were not used for training the transformation matrix. This can be tested in two ways within each gender: the first involves using the data from all subjects of one gender, excluding the test subject, to define a single transformation, which is then used on the test data. This method is called “multisubject transformation.” The second possibility is to create individual transformations for subjects and then average them together into one transformation, which is applied to the remaining subject’s data. This method is called “subject averaged transformation.” For the multisubject transformation, 301 sentences from seven subjects were combined into a 210-sentence training set. This training set was used to define a transformation, which was then tested on the 30 sentences of the remaining test subject. This process was repeated eight times for all combinations of training/testing subjects per gender.

![FIG. 6. Predicted (solid line) vs actual (dotted line) trajectories for the four different gestures (jaw, UL, LC, and UL/LL) for the sentence “The museum hires musicians every evening” from subject F3. These data illustrate different combinations for CC and NMSE values.](image)

![FIG. 7. Correlation between CC and NMSE for jaw predictions for subject M6.](image)

![FIG. 8. Overscaled predictions for UL from subject M5 for the sentence “Nothing is as offensive as innocence.” As in Fig. 6, solid lines depict predicted trajectories and dotted line depicts actual trajectories.](image)

**TABLE I. Correlation between CCs and NMSEs across subjects.**

<table>
<thead>
<tr>
<th>UL</th>
<th>LC</th>
<th>UL/LL</th>
<th>Jaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.62 (0.21)</td>
<td>−0.76 (0.12)</td>
<td>−0.92 (0.05)</td>
<td>−0.91 (0.05)</td>
</tr>
</tbody>
</table>
For the subject averaged transformation, the same 30 sentences from each subject were used to define transformations for each subject. Seven transformations were then linearly averaged into one transformation, which was tested on the 30 sentences of the remaining test subject. This process was repeated eight times for all combinations for training/testing subjects per gender.

Results for the multisubject transformation and subject averaged transformation are given in Table II. Results using the same subject for testing and training are also given under “single subject.” Figure 10 shows the individual gesture differences.

The subject averaged model generalizes well across subjects of the same gender, with average CCs dropping from 0.71 to 0.66 for males and from 0.68 to 0.60 for females. The multisubject model did not perform as well, with drops in CC from 0.71 to 0.59 for males and 0.68 to 0.52 for females.

Differences in CC were largest for UL, particularly for female subjects, indicating that the UL gesture may be the least generalizable and this is probably also a reflection of the relatively poor UL prediction in females in general.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single subject</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>Multisubject</td>
<td>0.59</td>
<td>0.52</td>
</tr>
<tr>
<td>Subject averaged</td>
<td>0.66</td>
<td>0.60</td>
</tr>
</tbody>
</table>

FIG. 9. Top: Sensitivity of CCs to acoustic frame size, using 89-sentence training sets, averaged across subjects. Bottom: Sensitivity of CC to the number of sentences in the training data set using 100 ms acoustic frame, averaged across subjects.

FIG. 10. Single subject, multisubject, and subject averaged transformations, separated by gesture.

IV. DISCUSSION

Our results are consistent with the previously obtained results from smaller scale studies using regression modeling, that approximately 70% of facial motion is recoverable (Ye-hia et al., 1998; Barker and Berthommier 1999; Jiang et al., 2000a, 2000b). We have confirmed these results for a much larger subject pool and highly varied speech database.

The findings further show that acoustic windows of 100 ms and data sets of approximately 40 sentences are required to achieve maximum predictability. In addition, we showed that by using subject averaged transformations, the facial motions of new subjects (i.e., not included in the training phase) can be predicted with only a small drop in accuracy.

An important limitation with the MLR approach is that it assumes that a linear, time-invariant mapping exists between acoustic information and facial position. This is demonstrably untrue even for most simplistic speech. Schroeder (1967), modeling the vocal tract as a tube, showed that a unique area function could not be determined by the formants of the output acoustics. Computer simulations have shown that different vocal tract shapes can lead to almost
identical values of the first three formants (Atal et al., 1978). A recent paper describing a new mathematical approach called MIMICRI deals with this issue by applying continuity constraints in order to recover articulator motion from acoustics, “even if the mapping from articulatory positions to acoustics is not known, and possibly, even if the mapping is many-to-one” (Hogden et al., 2007, p. 379). This approach may hold some promise for future applications in the area of speech recognition and animation.

The situation for our model becomes more complicated for vocal-tract articulator position rather than facial articulator position because intra-oral articulator motion (e.g., tongue) is hardly reflected in the face (except when it directly correlates with jaw motion; see below). The articulations of /l/, /d/, and /n/ (as in /toe/, /doe/, and /know/) can look almost identical in the face but sound very different. Facial articulator position for a given sound can vary widely for a single phoneme across contexts. For instance, the tongue and jaw can act in a coordinated manner to achieve a single acoustic goal in a number of different ways, referred to as “coarticulation,” or “functionally-organized goal-oriented behavior” (Fletcher and Harrington, 1999). Moreover, the individual contribution of articulators to produce a phoneme can vary between repetitions of the same utterance, and may vary considerably between sessions for the same subject and context (Alfonso and Van Lieshout, 1997).

A number of methods have been used to capture the time-varying nature of articulation in facial animators using other models such as AVCs (Kakumanu, 2002), TDNNs (Massaro et al., 1999; Savran et al., 2006), and HMMs (Yamamoto et al., 1998; Xie and Liu, 2007). Despite the limitations of MLR, our results are comparable to other systems using non-linear, asynchronous models. Xie and Liu (2007) compared six different HMMs for audio-visual speech synthesis, the most effective model producing average correlations of 0.696. Zelezny et al. (2006) using a 39-input ANN achieved a highest correlation of 0.622 when predicting LC, UL/LL and UL. Fu et al. (2005) were able to achieve a higher average correlation of 0.83 for jaw, UL/LL and LC with a HMM that used five frames of video information (captured at 60 fps) to account for backwards and forwards coarticulation. Savran et al. (2006) using a time delay neural network with nine frames of context achieved correlations of 0.72. They were then able to further increase predictability by using text information as an additional input.

In our system, the use of long acoustic frames acted to smooth over local context and coarticulation effects, thereby improving predictability. Particularly interesting is the fact that the subject averaged transformations generalized well to new subjects (CC=0.66 for males, 0.60 for females) even though the number of subjects used to obtain the average transformation was relatively small (7), and predictability will likely improve with a larger pool. Savran et al. (2006) attempted similar subject-independent predictions by concatenating the audio-visual speech data of 14 subjects into an audio-visual codebook, and used a look-up algorithm to predict facial motion of a novel subject. However, the average correlation of 0.72 using a training database of 125 sentences was obtained using text information as an additional input.

The poorer results of the multisubject transformations are most likely due to the fact that all intersubject data, including large differences in orofacial dimensions and absolute values of formant frequencies, are being concatenated into the training set before the minimum squared error algorithm is performed. With the subject-averaged transformations, intersubject differences are being concatenated by an averaging process after each minimum squared error transformation is derived. This second process is less sensitive to variations in the total data set and will lead to better general predictions.² Savran et al. (2006) avoided this issue by associating acoustic information from multiple speakers to visual information from a single speaker in a codebook algorithm, using a specific weighting system. In their system, the acoustic information of a given speaker is independently compared to that of 14 other speakers before choosing the appropriate viseme.

Our subject averaged model reduces these differences by first creating individual transformations and then averaging them. This suggests that the important structural similarities across subjects for the mappings are related to the general acoustic-to-articulatory trends, e.g., that F1 is inversely related to tongue height while F2 is directly related to tongue advancement (Kent and Read, 1992). Both height and advancement of the tongue are partially controlled by the jaw and will thus be mirrored, to some extent, in the face (Hertrich and Ackermann, 2000; Stone and Vatikiotis-Bateson, 1995).

The speaker-independence of MLR modeling may make it useful for animation in the context of rehabilitation purposes. Berthommier (2003) has shown that animations based on linear regressions are realistic enough to improve speech perception, although that study was limited to one-syllable or two-syllable words with overt syllabic structures. We generated some preliminary crude point-animations and have found them to look surprisingly realistic despite mediocre correlation coefficients. This provides hope to its use in future multimedia/rehabilitative applications, although more objective testing with realistic animations is necessary.

V. CONCLUSIONS

We have shown that facial motion can be predicted from acoustics using a linear, time-invariant model with accuracies comparable to that of more computationally intensive models. Our results were generated by 16 subjects speaking 90 sentences each.

While HMMs, TDNNs, and AVCs use adjacent acoustic frames to model coarticulation, this is not possible using the technique of MLR. We compensated for this by using an enlarged acoustic window of 100 ms, which then improved predictability. To train our system effectively, approximately 40 sentences (less than 2 min of continuous speech) are required. Further, our model is capable of predicting the speech motions of subjects untrained in the system with only a small drop in predictability. This work demonstrates that MLR may be used as a general framework to create facial animations for multimedia or rehabilitative purposes.
ACKNOWLEDGMENTS

This research was supported by the National Sciences and Engineering Research Council (NSERC) and undertaken, in part, thanks to funding from the Canada Research Chairs Program. The authors wish to thank Aravind Namasivayam and Mark Noseworthy for their support and help in preparing the manuscript.

1As mentioned, predictability differences for datasets larger than 20 were very small and close to the 99% predictability level found for 37 sentences and 30 sentences were chosen as a practical number (all sentences were recorded in groups of 15) for the purpose of this evaluation.

2We thank an anonymous reviewer for highlighting this aspect.


