Predicting Tongue Positions from Acoustics and Facial Features

Asterios Toutios, Slim Ouni

Predicting Tongue Positions from Acoustics and Facial Features

Asterios Toutios, Slim Ouni

University Nancy 2 / LORIA, UMR 7503, BP 239, 54506 Vandœuvre-lès-Nancy, France
{toutiosa,slim}@loria.fr

Abstract

We test the hypothesis that adding information regarding the positions of electromagnetic articulograph (EMA) sensors on the lips and jaw can improve the results of a typical acoustic-to-EMA mapping system, based on support vector regression, that targets the tongue sensors. Our initial motivation is to use such a system in the context of adding a tongue animation to a talking head built on the basis of concatenating bimodal acoustic-visual units. For completeness, we also train a system that maps only jaw and lip information to tongue information.
Index Terms: acoustic-to-articulatory mapping, facial features, electromagnetic articulography

1. Introduction

We have been working on talking head synthesis by concatenating units composed of bimodal acoustic-visual information [1]. One of the future goals of our project is to enhance our facial animation with an animation of the tongue. We plan to drive this tongue animation using electromagnetic articulography (EMA) information.

Ideally, to achieve such a goal, we should concurrently record facial and EMA information. However, this is difficult in practice with our current data acquisition setup. In fact, for the facial data acquisition, we use a stereovision system with two cameras that capture the positions of a large number of markers painted on the face [2]. For EMA acquisition, we use the AG500 articulograph [3], where the speaker has the head positioned inside a plexiglas cube on which electromagnetic coils are mounted. The sides of this box occlude the field of view of the cameras.

In both facial and EMA acquisitions we concurrently record (and synchronize) the acoustic signal. We could obtain concurrent facial and *estimated* EMA data by building a system that maps acoustics to EMA information (using concurrently recorded acoustics and EMA) and then mapping the acoustic information recorded concurrently with facial data to EMA information. In fact, we have previously built systems [4, 5] that perform the mapping from acoustics to EMA with results that are on a par with published state-of-the-art solutions [6, 7, 8, 9] to the acoustic-to-articulatory inversion problem using the same dataset, namely the MOCHA database [10]. Nevertheless, we know that there are limits to the accuracy we can get from such a system: all studies using MOCHA seem to converge to a root mean squared error of around 1.5 mm to 1.6 mm, averaged over the 14 channels involved.

But with our setup we will have available not only acoustics but also visual information. In fact, we can manage that the positions of some of the markers in our stereovision acquisition are exactly equivalent to the positions of the lip and jaw sensors in our EMA acquisition. Our hypothesis is that adding visual information regarding the lips and jaw to acoustics can improve the accuracy of predicting the tongue sensors, since such an addition can make the mapping less ambiguous. This information will regard EMA sensors on the lips and jaw in the acquisition of training data, and then equivalent stereovision markers at testing. The present paper tests this hypothesis using EMA data from MOCHA, both for training and testing.

Adding visual information to an acoustic-to-EMA mapping setup using MOCHA was done in [11] with no significant improvement to the prediction of the tongue sensor positions. Nevertheless, in that case the visual information was extracted from video images in a complex way, that might be a source of inaccuracies. We believe that using as visual information simply the positions of the EMA sensors on the jaw and lips, may be more informative.

The authors of [12] studied the case where the tongue sensors are predicted using *only* information on the lip and jaw sensors, i.e. no acoustic information at all. As probably expected, the results were not good, suggesting that visual information alone is not enough to predict tongue position. Just for the sake of completeness, we replicate their experiments using our mapping method on MOCHA, as they used another dataset.

2. Data and Method

MOCHA includes electromagnetic articulography (EMA) information for the coils shown in Figure 1. The two coils at the bridge of the nose and the upper incisors are used for the normalization of the data from the rest. Seven coils, located at the lower incisors (li), upper lip (ul), lower lip (ll), tongue tip (tt), tongue blade (tb), tongue dorsum (td) and velum (v), offer useful location information, namely trajectories of the projections of their position on two axes on the midsagittal plane: one with direction from the front to the back of the head (x-axis) and one with direction from the bottom to the top of the head (y-axis). The information flows from individual coils on individual axes are referred to as *EMA channels*.

For the processing of EMA data, first we subtracted a filtered version of the channel means across the dataset in the way and for the reasons explained in [13]. In our implementation we used a 15-point moving average window to filter the channel means. Then we low-pass filtered the data using a Hamming window at 20 Hz, and resampled from 500 Hz to 100 Hz. The authors of [8] showed that at least 99% of the energy of the EMA channels in MOCHA (and for the fsew0 speaker we use) lies well below 20 Hz (the only exception is the horizontal projection of the velum sensor with a slightly larger frequency bound, but this sensor is not important to our present work), so it is safe to say that low-pass filtering with this cutoff frequency does not lead to loss of important information. Regarding the acoustic speech signal, we extracted 12 MFCCs using HTK [14], with a window size of 25 ms and a shift of 10 ms. Both EMA and acoustic data were z-scored, and silent stretches

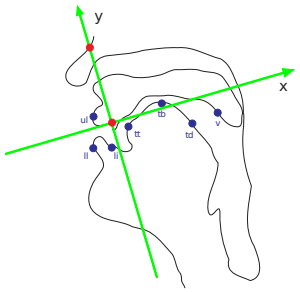

Figure 1: *Approximate positioning of sensors in the MOCHA database, and axes involved.*

at the beginning and end of the utterances were omitted.

We considered three different kinds of input information. For the *acoustic experiment*, the input information for each frame was its MFCC parameters. For the *facial experiment*, the input information for each frame was the values of the six EMA channels corresponding to sensors li, ul, and ll. For the *acoustic + facial experiment* the input information for each frame was the union of the previous two sets. In all cases, we constructed context input vectors spanning over 11 consecutive frames, centered around the output frame. The output information was the value of one of the EMA channels corresponding to sensors tt, tb, and td. Note that the support vector regression (SVR) algorithm that we used for the mapping considers a scalar output value, and not a vector.

We trained the ϵ-SVR algorithm with the gaussian kernel, using the LibSVM software [15]. The algorithm solves the following optimization problem:

$$\text{maximize}$$

$$-\varepsilon\sum_{i=1}^{n}(a_i^*+a_i)+\sum_{i=1}^{n}(a_i^*-a_i)y_i$$

$$-\frac{1}{2}\sum_{i,j=1}^{n}(a_i^*-a_i)(a_j^*-a_j)\exp(-\gamma\|\mathbf{x_i}-\mathbf{x_j}\|^2) \qquad (1)$$

$$\text{subject to}$$

$$0\le a_i,a_i^*\le C,\ i=1,\ldots,n \text{ and } \sum_{i=1}^{n}(a_i^*-a_i)=0,$$

over the n input vectors $\mathbf{x_i}$ and corresponding output scalars y_i, to provide with the mapping function

$$f(\mathbf{x})=\sum_{i=1}^{n}(a_i^*-a_i)\exp(-\gamma\|\mathbf{x}-\mathbf{x_i}\|^2)+b \qquad (2)$$

where b is calculated from the Karush-Kuhn-Tucker conditions for the problem [16]. The parameters C, ε, and γ are to be selected by the experimenter. Based on [17], we used

$$C=\max\left(|\bar{y}+3\sigma_y|,|\bar{y}-3\sigma_y|\right) \qquad (3)$$

where $\bar{y}$ and σ_y are the mean and the standard deviation of the output values of training data, and

$$\varepsilon=3\sigma_n\sqrt{\frac{\ln n}{n}} \qquad (4)$$

where n is the number of training examples, and σ_n is the median value of $\sqrt{(y-\bar{y})^2}$ across the training output data. Finally, based on [18], we used

$$\gamma=\frac{m^2}{\sum_{i,j=1}^{n}\|\mathbf{x_i}-\mathbf{x_j}\|^2}. \qquad (5)$$

The results of the mapping function of Eq. 2 were z-scored values of the EMA channels. After testing, we inverted z-scoring and added back the filtered version of the channel mean corresponding to the utterance in question. As a final post-processing step, we smoothed the resulting trajectories using a low pass-filter at 20 Hz, i.e. the same filter we used at pre-processing.

root mean squared error (mm)

channel	acoustic	acoustic + facial	facial
tt_x	2.34 (1.83)	2.15 (1.74)	3.52 (2.72)
tt_y	2.34 (1.98)	2.11 (1.54)	3.30 (2.39)
tb_x	2.19 (2.13)	1.98 (2.18)	3.16 (2.67)
tb_y	2.06 (2.23)	1.95 (2.23)	3.50 (4.23)
td_x	2.03 (2.08)	1.81 (1.97)	2.81 (2.51)
td_y	2.12 (1.91)	2.06 (1.91)	3.40 (3.48)
average	2.18 (2.03)	2.01 (1.93)	3.28 (3.00)

Pearson correlation

channel	acoustic	acoustic + facial	facial
tt_x	0.813 (0.794)	0.846 (0.811)	0.543 (0.573)
tt_y	0.861 (0.879)	0.888 (0.933)	0.710 (0.811)
tb_x	0.806 (0.795)	0.845 (0.791)	0.573 (0.622)
tb_y	0.857 (0.895)	0.873 (0.914)	0.546 (0.402)
td_x	0.791 (0.757)	0.839 (0.770)	0.587 (0.562)
td_y	0.796 (0.837)	0.809 (0.846)	0.420 (0.323)
average	0.821 (0.826)	0.850 (0.844)	0.562 (0.549)

Table 1: *Outside the parentheses are cumulative results for the 460 utterances spoken by fsew0, after cross-validation experiments (see text for the explanation of the inputs used). Inside the parentheses are results for an example of the single utterance "It's not easy to create illuminating examples" which is also illustrated in Figs. 2 and 3. Subscripts at sensor names denote the projection on the axes shown in Fig. 1.*

3. Results

We experimented using the data of speaker fsew0 from MOCHA, i.e. a female with a Southern English accent. We performed cross-validation experiments over the 460 numbered utterances available, using five partitions. For evaluation we used the two metrics typical in works on acoustic-to-EMA mappings: root mean squared error

$$E_{RMS}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i'-y_i)^2} \qquad (6)$$

where m is the number of examples in the test set, and y, y' are real and estimated values; and Pearson correlation

$$r=\frac{\sum_{i=1}^{m}\left(y_i'-\overline{y'}\right)(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{m}\left(y_i'-\overline{y'}\right)^2\sum_{i=1}^{m}(y_i-\overline{y})^2}} \qquad (7)$$

where overlines denote mean values over the test set.

These results are summarized in Table 1. The numbers outside parentheses refer to the whole dataset of 460 utterances.

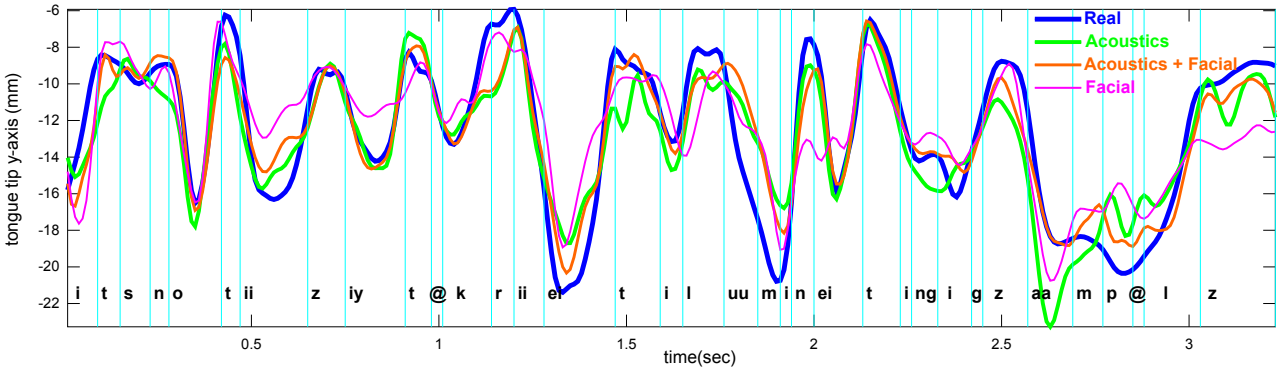

Figure 2: *Real and estimated trajectories describing the projection of tongue tip sensor position on axis y, as shown in Fig. 1, for the utterance "It's not easy to create illuminating examples", spoken by speaker fsew0 of MOCHA. Vertical lines mark the boundaries between phones. Vertical axis denotes millimeters; horizontal axis denotes time in seconds. RMS errors and Pearson correlations between the estimated trajectories shown and the real trajectory are given in parentheses in Table 1. The MOCHA labeling convention is used for the names of the phonemes.*

The numbers between parentheses refer to the utterance "It's not easy to create illuminating examples" which was chosen randomly from the dataset, for illustration purposes. For this utterance, the real and estimated trajectories for the projection of the tongue tip sensor on the y-axis of Fig. 1 are shown in Fig. 2. For the same utterance, Fig. 3 shows snapshots of an animation of the results.

Regarding overall results, we can see that the addition of facial cues to acoustics improves the performance on all six tongue channels, in terms of both metrics used. On average, the improvement in root mean squared error is 0.17 mm (7.8%), and the improvement in Pearson correlation is 0.29 units (3.5%). The performance of the system that uses only facial features as input is very poor, which verifies the findings of [12].

Regarding the single utterance, the presented results for both systems using acoustics (with or without facial features) are in general better than the cumulative results over the whole dataset, indicating that the specific utterance is a relatively good case among the 460. Nonetheless, the observed relative improvement after adding facial features is of similar importance compared to the whole (4.9% for root mean squared error, 2.8% for Pearson correlation, on average). Indeed, the trajectory estimated from the combination of acoustics and facial features shown in Fig. 2 is, in broad terms, a slightly better match to the real trajectory, than the trajectory estimated only from acoustics. But if we focus on the animation presented in Fig. 3 instead of just numbers (or a single trajectory), the improvement does not seem so important. The tongue contours based on the estimations from the combination of acoustics and facial features are closer to the real contours than their counterparts estimated just from acoustics, but only marginally so. Both sets of estimated contours present more or less the same problems in comparison to the real contours, for example the lack of making contact with the palate for velar consonants /k/ (2nd row, 6th column) or /g/ (5th row, 4th column), or the inverted overall tongue curvature for some instances of alveolar consonants like /t/ (in 1st row, 2nd column) and /n/ (in 1st row, 4th column and in 4th row, 4th column).

4. Concluding Remarks

In this paper, we presented experiments of adding facial information to a typical acoustic-to-EMA machine learning-based mapping setup. Overall numbers indicated that there is indeed a relative improvement of the results when doing so. However, when we animated the estimated tongue contours and compared them to the real ones, we found that these improvements were, at best, only marginal.

Our initial motivation was to use such a system in the context of adding a tongue animation to a talking head system. After our experiments, and, most notably, after observing animations such as the one shown in Fig. 3 we believe that, even *without* adding facial information, our acoustic-to-EMA mapping system is able to provide synthetic tongue trajectories useful for our purpose, i.e. a tongue animation that is intelligible to the interlocutor of the talking head. But on the other hand, we are especially doubtful about whether, even *after* adding facial information, the same system could be useful in the context of more sensitive applications, like providing articulatory feedback for speech training [19] or driving an articulatory model for the purposes of articulatory synthesis [20].

Several approaches have been tried on the acoustic-to-EMA mapping problem. When presented with such a work, aside from the main mapping method used, one should pay attention to the particular details of the implementation and presentation, such as the processing of the EMA trajectories, the speech signal parametrization, the size of context window used, or even the chosen subset of the dataset that is used for evaluation. But the single most important factor that can affect results is the corpus itself: recently Richmond applied the exact same methodology to MOCHA and to a newly acquired dataset and found a decrease of root mean squared error from 1.54 mm to 0.99 mm [21]. Perhaps in our case also the use of another corpus, which we plan to acquire, can further improve the results presented here.

5. Acknowledgment

This work was supported by the French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01).

6. References

[1] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger, "Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units," in *Interspeech*, Makuhari, Japan, 2010.

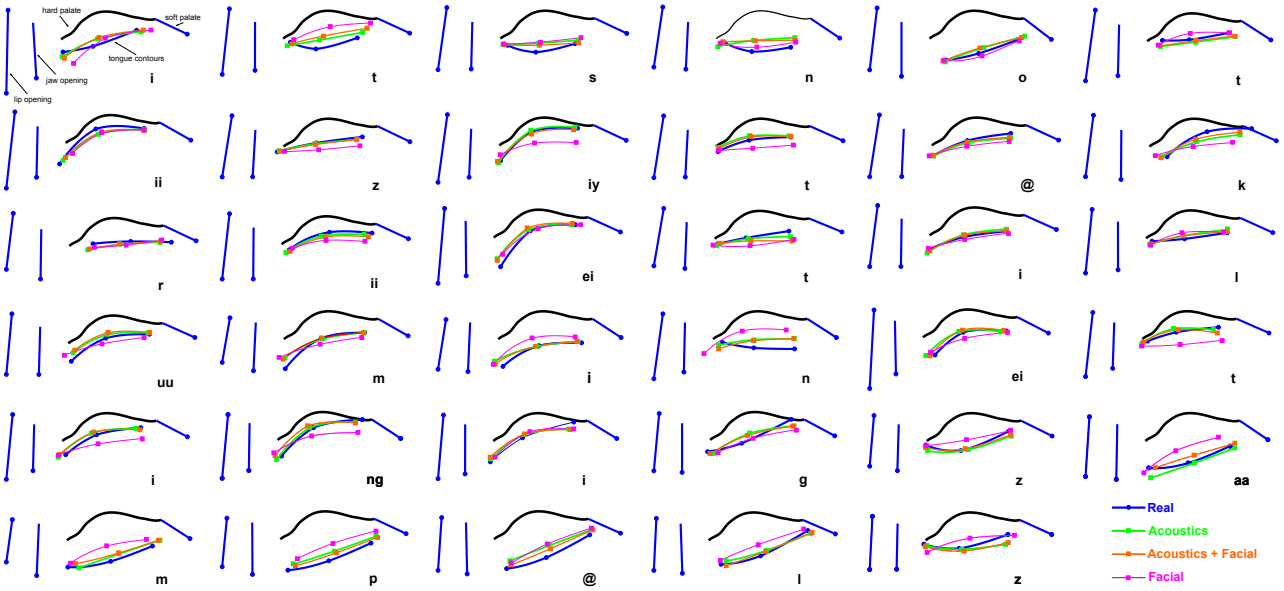

Figure 3: *Visualization of results for the phrase "It's not easy to create illuminating examples". Each subplot corresponds to approximately the middle of the phone duration, as indicated in Fig. 2. For the tongue sensors, we show the real positions of the sensors and the three estimation. In each case the three sensors are shown connected by a simple cubic spline. For the sensors on lip, jaw, and velum, we show only real positions. The leftmost straight line approximates the lip opening (connecting upper and lower lip sensor) and the second straight line jaw opening (connecting the constant position of upper incisor sensor to that of lower incisor sensor). The straight line connecting the rightmost point of the hard palate with the velum sensor position approximates very roughly the soft palate. The shape of the hard palate was approximated from the full fsew0 dataset. RMS errors and Pearson correlations between real and estimated positions for the full utterance are given in parentheses in Table 1. The MOCHA labeling convention is used for the names of the phonemes.*

[2] B. Wrobel-Dautcourt, M. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low-cost stereovision based system for acquisition of visible articulatory data," in *AVSP*, British Columbia, Canada, 2005.

[3] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillmann, "Extracting tongues from moving heads," in *5th Speech Production Seminar*, Kloster Seeon, Germany, 2000, pp. 313–316.

[4] A. Toutios and K. Margaritis, "Contribution to statistical acoustic-to-EMA mapping," in *EUSIPCO*, Lausanne, Switzerland, 2008.

[5] S. Demange and S. Ouni, "Acoustic-to-articulatory inversion using an episodic memory," in *ICASSP*. Prague, Czech Republic, 2011.

[6] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.

[7] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[8] P. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[9] G. Ananthakrishnan and O. Engwall, "Mapping between acoustic and articulatory gestures," *Speech Communication*, vol. 53, no. 4, pp. 567 – 589, 2011.

[10] A. Wrench and W. Hardcastle, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar on Speech Production*, Kloster Seeon, 2000, pp. 305–308.

[11] A. Katsamanis, G. Papandreou, and P. Maragos, "Face active appearance modeling and speech acoustic information to recover articulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 411–422, 2009.

[12] A. Ben Youssef, P. Badin, and G. Bailly, "Can tongue be recovered from face? The answer of data-driven statistical models," in *Interspeech*, Makuhari, Japan, 2010.

[13] K. Richmond, "Estimating articulatory parameters from the speech signal," Ph.D. dissertation, The Center for Speech Technology Research, Edinburgh, UK, 2002.

[14] S. Young, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005.

[15] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[16] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, December 2001.

[17] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, pp. 113–126, 2004.

[18] I. Tsang, J. Kwok, and J. Zurada, "Generalized core vector machines," *Neural Networks, IEEE Transactions on*, vol. 17, no. 5, pp. 1126–1140, 2006.

[19] J. Levitt and W. Katz, "The Effects of EMA-Based Augmented Visual Feedback on the English Speakers' Acquisition of the Japanese Flap: A Perceptual Study," in *Interspeech*, Makuhari, Japan, 2010.

[20] A. Toutios, S. Ouni, and Y. Laprie, "Estimating the Control parameters of an Articulatory Model from Electromagnetic Articulograph Data," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3245–3257, 2011.

[21] K. Richmond, "Preliminary inversion mapping results with a new EMA corpus," in *Interspeech*, Brighton, UK, 2009.