Enhancing Acoustic-to-EPG Mapping with Lip Position Information

Asterios Toutios, Konstantinos Margaritis

Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
{toutios,kmarg}@uom.gr

Abstract

This paper investigates the hypothesis that cues involving the positioning of the lips may improve upon a system that performs a mapping from acoustic parameters to electropalatographic (EPG) information; that is, patterns of contact between the tongue and the hard palate. We adopt a multilayer perceptron as a relatively simple model for the acoustic-to-electropalatographic mapping and demonstrate that its performance is improved when parameters describing the positioning of the lips recorded by means of electromagnetic articulography (EMA) are added to the input of the model.

Index Terms: speech production, inversion, EPG, EMA

1. Introduction

Electropalatography (EPG) [1] is a relatively well-known technique that records patterns of contact between the tongue and the hard palate during continuous speech. It utilizes an artificial palate in which a number of electrodes are embedded (62 in the Reading EPG system, considered herein). When the tongue contacts any of the electrodes, a signal is conducted to an external processing unit and recorded, with a typical sampling rate of 100-200 Hz. For a given utterance, the sequence of EPG data consists of a stream of vectors with binary elements, representing contacts or non-contacts between the tongue and any of the electrodes. Figure 1 shows part of such a stream. EPG has been successfully used to study several phenomena in phonetic descriptive work, in studies of lingual coarticulation and in the diagnosis and treatment of various speech disorders [2].

For the acquisition of Electromagnetic Articulography (EMA) [3] data, sensor coils are attached to specific places on the speaker's articulators, such as the lips, teeth, tongue and velum. The speaker then wears a special helmet that produces an alternating magnetic field which records the positions of the coils at the endpoints of small fixed-time intervals. The outcomes are trajectories that illustrate the movement of the coils. Usually there are two trajectories for each coil, for the projection of its position on two axes on the midsaggital plane, an horizontal (x-axis) and a vertical one (y-axis).

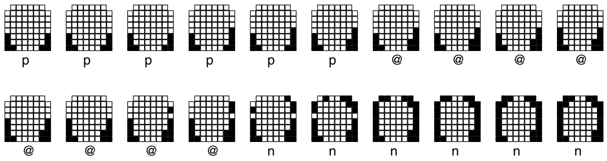

Figure 1: Part of typical EPG sequence. The shape of the figures (*EPG vectors* or *electropalatograms*) follows that of the palate, the alveolar part being in the top and the velar part in the bottom. Black squares indicate a contact between the tongue and the palate. Segment is from the utterance “The hallway o**pen**s into a huge chamber”. Speaker is fsew0 from the MOCHA database. Corresponding MOCHA labels are shown. Sampling rate is 200 Hz.

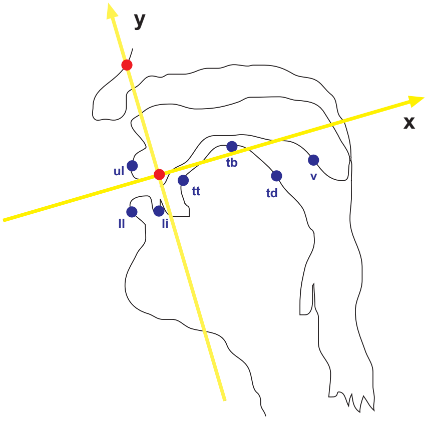

Figure 2: EMA sensor coils and axes in the MOCHA database. The coils on the bridge of the nose and the upper incisors are used only for normalization purposes.

The freely available MOCHA database [4] includes speech waveforms, sampled at 16 kHz, EMA data for the coils and axes presented in Figure 2 sampled at 500 Hz and EPG data sampled at 200 Hz, for three speakers (at the time of writing) reading a set of 460 British TIMIT sentences. The utterances are phonemic labeled, nevertheless the labels are the result of an automatic alignment process and considered prone to errors.

In several previous occasions (e.g. [5]), we have presented works towards building systems that estimate EPG sequences from the corresponding acoustic information. We view this task as special case of the acoustic-to-articulatory mapping problem, or speech inversion [6], which draws considerable attention in the speech community. We believe that a successful acoustic-to-EPG mapping could have all the same applications attributed to the more general acoustic-to-articulatory mapping in fields such as speech therapy, visualization, recognition, synthesis, coding or phonetics [7, 8].

It was suggested to us by several sources, that the estimation of EPG sequences could benefit, if visual cues regarding lip activity were incorporated in such a system. In this paper we explore exactly this suggestion. Not having video feeds of the MOCHA speakers, we regard the available EMA lip data as an adequate supplement. The projections on the y-axis of the position of the EMA coils on the upper and lower lip offer information on vertical lip opening, while the projections on the x-axis are indicative of lip protrusion.

In the present paper we adopt a multilayer perceptron

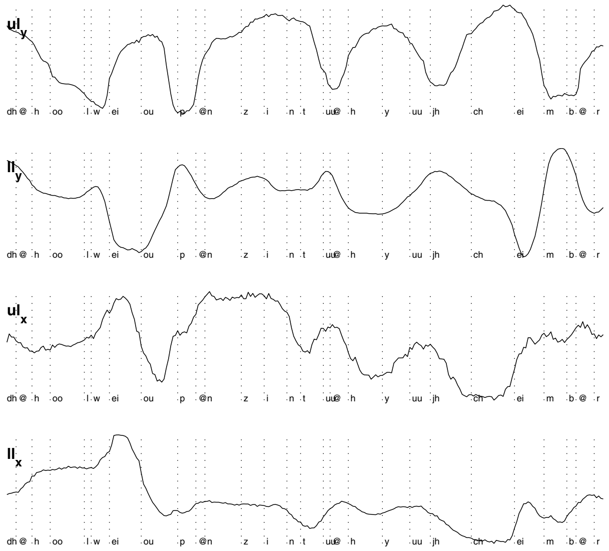

Figure 3: Processed EMA data for the utterance "The hallway opens into a huge chamber". Top to bottom: projection of the position of upper lip on y-axis; lower lip, y-axis; upper lip, x-axis; lower lip x-axis. MOCHA labels are shown.

(MLP) for the acoustic-to-EPG mapping, considering at first only acoustic parameters as inputs, and optimize the size of the input context window and number of hidden neurons so that the performance error on a test set is minimized. Having chosen the optimal size of acoustic context window we proceed by adding the lip EMA projections on the y-axis to the input, optimizing the context windows for these EMA parameters and again the number of hidden neurons. We then repeat the process adding the lip EMA projections on the x-axis to the input.

2. Data Processing

We utilize acoustic, EMA and EPG data from the fsew0 speaker of the MOCHA database, a female with a Southern English accent. These are processed as follows.

Based on the MOCHA label files, silent parts at the beginning and end of the utterances are omitted. 12-order MF-PLPs [9] plus log energies are extracted from the speech signal using 16ms windows with 10ms shifts and 40 filterbanks. The EMA data for the projections of the coils on the upper and lower lip on both axes are subjected to a series of processing steps similar to the ones presented in [7]. The procedure includes normalization, filtering and subsampling of the EMA data to 100 Hz. Resulting trajectories for a single utterance are shown in Figure 3. EPG data are subsampled from 200 Hz to 100 Hz.

Overall process results in 124,242 triplets, sampled at 100 Hz, of 13-dimensional acoustic vectors, 4-dimensional EMA vectors and 62-dimensional EPG vectors. From the 460 utterances, data from 92 (every 10th utterance beginning with the 2nd and every 10th beginning with the 6th, 24,388 triplets) constitute the test set and the rest the training set.

3. MLP Training and Results

We consider MLPs with one hidden layer of tanh neurons and an output layer of 62 logistic neurons, corresponding to the elements of the EPG vectors. EPG contacts are represented with

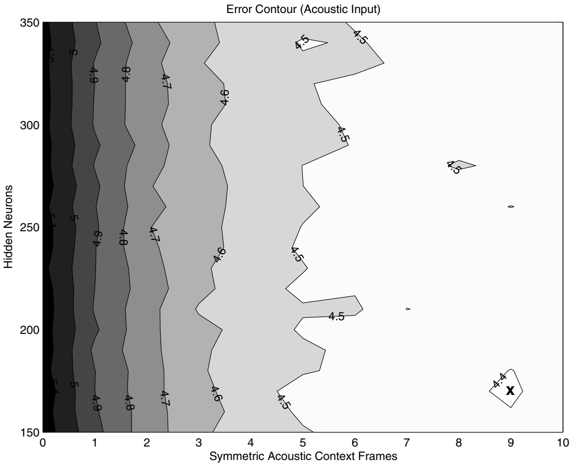

Figure 4: Contour plot of EPG estimation error versus number of symmetric frames and number of hidden neurons when only acoustic parameters are used at the networks' input. X marks the minimum.

the value of 0.7875 and non-contacts with 0.2125 (these are the values of maximum second derivative of the logistic function [10]). In the testing phase, the networks' output values are rounded to the nearest integer (0 for non-contacts or 1 for contacts). Estimation error is measured as the mean number of EPG electrodes for which the estimated values are wrong –that is, contacts instead of non-contacts and vice-versa– in each estimated EPG vector.

Considering first only the acoustic parameters as input to the networks, we experiment with various sizes of input context windows and various numbers of neurons in the hidden layer. Input vectors are constructed by concatenating acoustic parameters from the frames that are adjacent in time to the frame exactly corresponding to the EPG vector in question. We consider only symmetric context windows; that is, the input vectors are constructed by attaching the same number of adjacent frames before and after the frame in question. A value of 0 for the symmetric frames means that acoustic parameters from only one frame are used as the network input (13 parameters); a value of 1 means that acoustic parameters from 3 frames are used: the frame in question plus 1 previous frame plus 1 next frame (39 parameters); a value of 2 means a total of 5 frames and so on. For the number of neurons in the hidden layer we consider values from 150 up to 350 with an increment of 10.

All MLPs are trained for 50 epochs using the Scaled Conjugate Gradient optimization algorithm [11]. From the available training examples only one fifth (roughly the first out of five consecutive examples) is used for training, for speed of experiments and under the assumption of a certain degree of redundancy in the data.

Figure 4 is a contour plot of the EPG estimation error on the test set as function of number of symmetric frames and number of hidden neurons. Large context windows give in general better results, while the number of hidden neurons, at least for the values examined, does not influence much the performance. The minimum value of 4.385 is achieved with 170 neurons in the hidden layer and 9 symmetric acoustic frames; that is, the optimal input context window includes acoustic information from 19 frames in total, or 247 acoustic parameters, spanning over roughly 190 ms of speech.

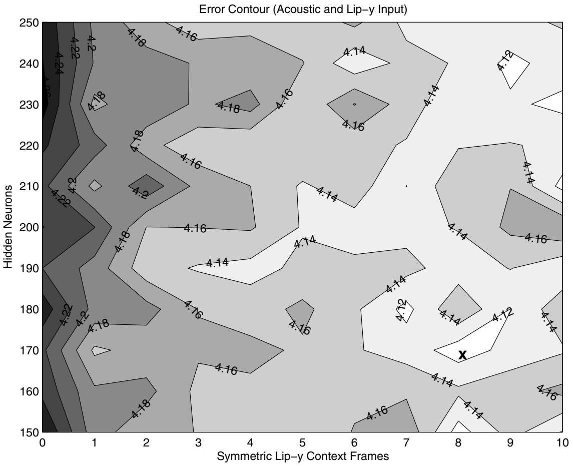

Figure 5: Contour plot of EPG estimation error versus number of EMA lip-y symmetric frames and number of hidden neurons when values of projections of upper and lower lip EMA coils on the y-axis are added to acoustic parameters at the networks' input. Size of the acoustic context window is fixed at a value of 9 symmetric frames. X marks the minimum.

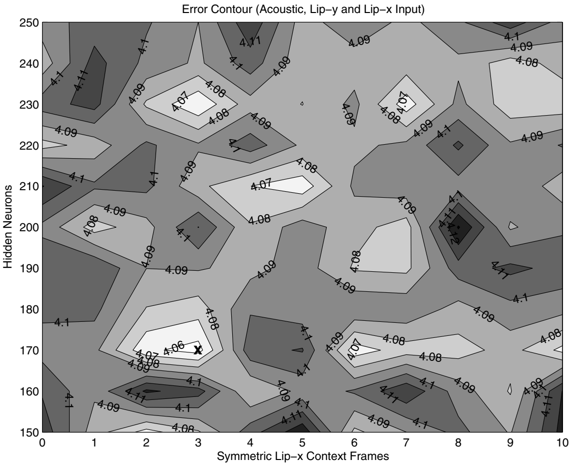

Figure 6: Contour plot of EPG estimation error versus number of EMA lip-x symmetric frames and number of hidden neurons when values of projections of upper and lower lip EMA coils on the y-axis are added to acoustic and lip-y parameters at the networks' input. Size of the acoustic context window is fixed at a value of 9 symmetric frames and size of the lip-y context windows at 8. X marks the minimum.

Fixing the number of symmetric frames in the acoustic context window at the value of 9, we repeat the process adding to the network's input values of the projections of the positions of the upper and lower lip EMA coils on the y-axis (let's call them lip-y projections). We now consider symmetric context windows applied to the EMA data; that is EMA information from more than one frame is added to the input. For the number of neurons in the hidden layer, values from 150 up to 250 with an increment of 10 are examined.

Figure 5 is the contour plot of the EPG estimation error as function of lip-y symmetric context frames and number of hidden neurons. Again, the performance of the network seems to benefit from large sizes of lip-y context windows, but now this depends on the number of hidden neurons. In total, the results are better than the ones achieved using only acoustic parameters. A minimal error of 4.110 is achieved with 170 neurons in the hidden layer and 8 symmetric lip-y frames.

The same process is repeated adding to the network's input the projections of the upper and lower lip EMA coils' projections on the x-axis (lip-x), with the number of acoustic symmetric frames fixed at 9 and the number of lip-y symmetric frames fixed at 8. This time, incremental values of lip-x symmetric context frames are considered. Figure 6 is the contour plot of the EPG estimation error as function of lip-x symmetric context frames and number of hidden neurons. A minimal error of 4.059 is achieved with 170 neurons in the hidden layer and 3 symmetric lip-x context frames. The shape of the contour plot does not allow for any concrete observations on the relationships between error, hidden neurons and size of context window. Nevertheless, overall results are better than those achieved using only acoustic and lip-y parameters, since the minimal error value in Figure 6 is roughly equal to the maximal error value in Figure 5.

In all, the best MLP out of those trained includes 170 hidden neurons and its input vectors consist of 295 parameters. Out of these $(9 \times 2 + 1) \times 13 = 247$ are acoustic parameters, $(3{\times}2{+}1){\times}2 = 34$ are lip-y parameters and $(3{\times}2{+}1){\times}2 = 14$ are lip-x parameters. Figure 7 shows a detailed example of estimation of an EPG sequence by this network.

4. Discussion

The findings of this paper do verify the hypothesis that information regarding the positioning of the lips may improve upon a system that estimates EPG patterns from the corresponding acoustic parameters. The minimal mean error achieved drops from 4.385, when only acoustic input is used, to 4.110, when the projections on a vertical axis of the positions of EMA coils placed on the upper and lower lips are added to the input, and to 4.059, with the further addition of the projections on the horizontal axis.

The use of large acoustic context windows in general speech inversion works may be justified as a means to overcome the one-to-many nature of the acoustic-to-articulatory mapping [12]. Our experiments using only acoustic input verify that large acoustic context windows are beneficial for the acoustic-to-EPG mapping since optimal results are achieved with as much as 9 symmetric context frames (19 frames in total), spanning over roughly 190 ms of speech. The same observation is also true for the lip position information: large context windows are beneficial for the projections on the vertical axis, since the optimal results are achieved using 8 symmetric context frames (17 frames in total, 170 ms of speech). For the projections on the horizontal axis, the observation holds in a lesser degree, since the optimum is achieved with only 3 symmetric frames (7 frames in total, 70 ms of speech). Additionally, it is perhaps interesting to notice that the optimal number of hidden neurons is the same in all three cases: 170.

The experiments presented in this paper could be refined in at least three ways. First, the networks were trained for a fixed number of epochs, without any attempt to prevent possible overfitting; assessing optimal number of training epochs via a separate validation set could possibly lead to smoother error contours, especially in Figures 5 and 6. Second, the use of MF-PLPs is not actually justified as the optimal acoustic

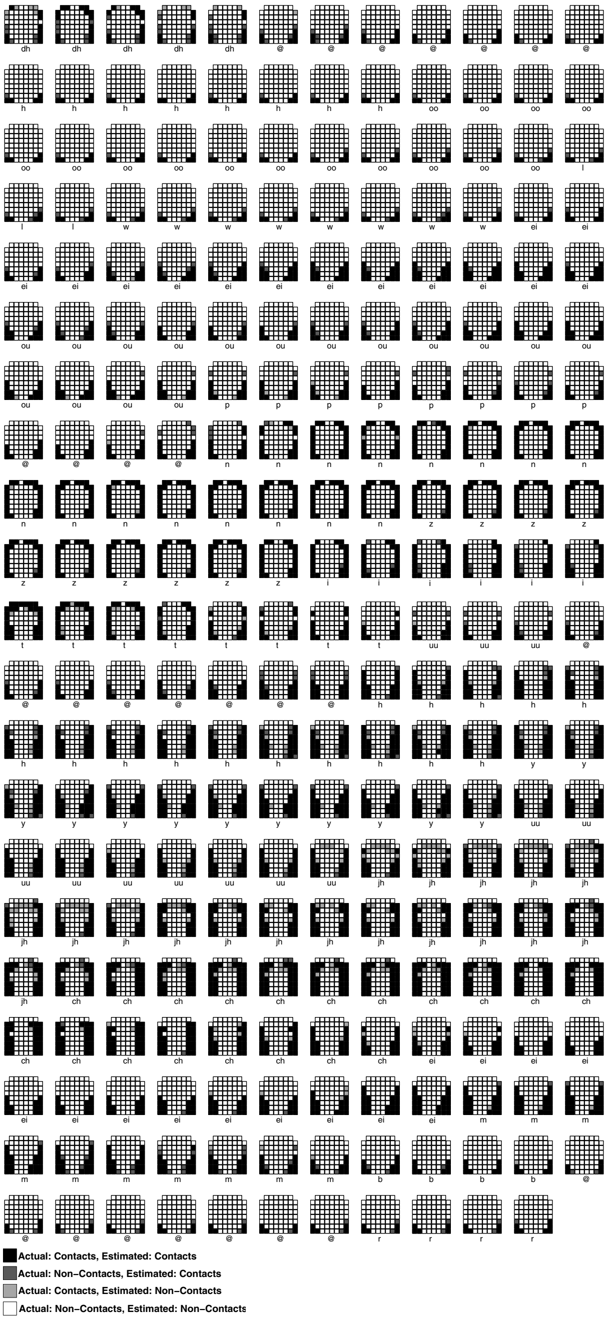

Figure 7: Detailed example of EPG estimation with the "best" MLP described in the paper. Fsew0 utters "The hallway opens into a huge chamber" (utterance number 246). Sampling rate is 100 Hz. MOCHA labels are shown.

parametrization for the task at hand; alternatives should be investigated. Third, only a subset of the available training data was actually used; using all data would probably lead to improvement of some degree.

We have already mentioned that we use EMA data as an adequate supplement to video feeds of the speakers talking. The projections of lip EMA coils on the vertical axes could be replaced by frontal views of the speaker's face, while the projections on the horizontal axis could equal lateral movies of the speaker. We believe that the extension of the method and results presented in this paper to an environment with such video feeds at hand is straightforward.

5. Acknowledgement

This work was supported by the Greek research program "HRAKLEITOS", which is co-funded by the European Social Fund (75%) and national resources (25%).

6. References

[1] F. Gibbon and K. Nicolaidis, "Palatography," in *Coarticulation in Speech Production: Theory, Data, and Techniques*, W. J. Hardcastle and N. Hewlett, Eds. Cambridge, England: Cambridge University Press, 1999, pp. 229–245.

[2] F. Gibbon, "Bibliography of electropalatographic studies in English (1957-2006)," Queen Margaret University College, Edinburgh, UK, Tech. Rep., 2006.

[3] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, pp. 26–35, 1987.

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

[5] A. Toutios and K. Margaritis, "Learning electropalatograms from acoustics," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. I–361–I–364.

[6] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal tract shapes from the speech signal," *IEEE Transactions on Speech and Signal Processing*, vol. 2, no. 1, pp. 133–150, 1994.

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

[8] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[9] P. C. Woodland, M. J. Gales, D. Pye, and S. J. Young, "Broadcast News Transcription Using HTK," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Munich, Germany, 1997, pp. 719–722.

[10] Y. Le Cun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks, Tricks of the Trade*, ser. Lecture Notes in Computer Science LNCS 1524. Springer Verlag, 1998.

[11] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[12] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004, pp. 1129–1132.