LIP MOVEMENTS SYNTHESIS USING TIME DELAY NEURAL NETWORKS

Sergio Curinga, Fabio Lavagetto, Fabio Vignoli
D.I.S.T. - University of Genova
Via Opera Pia 13A, 16145 GENOVA
Tel +39 10 3532802 Fax: +39 10 3532948
E-mail: sergio@dist.dist.unige.it

ABSTRACT

A method exploiting the audio-visual correlation of speech in order to estimate the lip and mouth movements is presented. Its applications are in the field of aids and services for elderly people, in videotelephony, in cartoons and movie dubbing. Notice that lip movements synthesis does not imply speech recognition and that the mouth shape is not only specified by the phoneme currently uttered but it also depends on some past and future speech information. In order to take into account this temporal correlation, and considering the constraint of computational effectiveness, the Time Delay Neural Networks (TDNNs) seem to be the most appropriate analysis tool in comparison with methods like Markov Models, which are more resource consuming.

1 INTRODUCTION

In a model-based videophone the encoder performs a visual analysis and transmits the extracted parameters to the receiver, which then animates a facial model.

The measured visual parameters describing the shape of the person's lips are strictly correlated to the acoustic information of the speech. For example, if a /p/, /b/, or /m/ phoneme is being spoken, one could predict a closed mouth, while a wovel pronounciation suggests an open mouth. This correlation between the acoustic and visual modalities could be used to devise a reliable acoustic–to–visual conversion system able to improve video quality in model–based coding systems [4] or to animate a mouth model only on the basis of the acoustic information [1][2][3].

The basic problem encountered in estimating the mouth parameters, like height and width, is to overcome the coarticulation process for which the mouth shape depends not only on the phoneme currently uttered but also on some past and future information [5].

A neural network approach seems to be appropriate to manage both temporal correlation and computational effectiveness. The reproduction of the lip movements is performed by means of a synthetic mouth, a structure based on a wire-frame representation and animated by moving its vertices according to the parameter estimates. Each of these parameters is estimated independently by a TDNN, trained according to a supervised procedure, by means of audio data represented in the cepstrum space and visual articulatory parameters extracted from the speaker's mouth within the video sequence. Adaptive algorithms have been devised in order to overcome training problems concerning local minima convergence and silence learning.

2 THE TDNN APPROACH

The Time Delay Neural Networks are a special kind of multi-layer perceptron structures (MLP) that perform, through a suitable architecture, the extraction of temporal invariant features [9][10] . To do that, every unit in the hidden layers of the architecture performs feature extraction and evaluates not only the current input (time t) but also the inputs at time t-1, t-2,..., t-n (n units time delay). These delays are realized by a window of suitable size that slides over the input data, thus realizing a temporal-spatial conversion.

A procedure like the Least Mean Square (LMS) algorithm cannot be used for training multi-layer networks, due to the fact that the required responses of the hidden nodes are not known. A learning algorithm exists for multi-layer networks, known as BackPropagation, that is a generalized form of the LMS procedure. This algorithm is an iterative gradient method which tries to minimize the mean square error between the actual output and the desired response.

The error surface of such a network exhibits many local minima, in contrast to the single layer network where it has only one minimum (when a solution exists). While training the network with speech data, each internal unit becomes sensitive to a specific acoustic phenomenon independently from its position within the word.

The network is connected in a feedforward manner, that is, the input signals to a node come from nodes in a lower layer. Each node consists of a summation and a sigmoid logistic function. For every iteration of the minimization process the mean square error is reduced until the error approaches zero. In the backpropagation algorithm the changes of the connection weights of the

Layer	0	1	2	3
Units	12	8	3	1
Delay	-	2	4	6

Table 1: TDNN adopted topology: each delay corresponds to a 20 msec interval.

networks can take place in parallel. In our work each articulatory parameter is estimated independently by a TDNN whose topology is summarized in Table 1.

The stop criterion was based on the number of iterations in order to allow an homogenous comparison of the performances provided by using different representations of the training data. This number figured out 15,000 in order to reach fairly good parameter estimates, judged by means of both subjective evaluations and cross-validation tests. [1]

3 ACOUSTIC - VISUAL ANALYSIS

Each of the articulatory parameters is estimated independently by a TDNN, trained according to a supervised procedure, by means of audio data represented in the cepstrum space and visual articulatory parameters extracted from the speaker's mouth within the video sequence.

3.1 Visual Analysis

The first problem to be addressed is parameter extraction from the video sequence. Each input image is analyzed, and a set of visual parameters is extracted. To accomplish this, some facial features have been tracked by a block matching technique, as shown in Fig 1.

In order to simplify the algorithm tests, a specific make-up was put on the subject's face to enhance the image contrast. The blocks to be tracked were manually defined on the first frame. For each of the ten image blocks (7X7 pixels), the analysis algorithms tracked and recorded the x, y coordinates of the center. Block tracking on both side and frontal views of the speaker's face was helpful in order to assess and correct the measures.

The following articulatory parameters have been then computed, starting from the coordinates of the tracked facial features, by means of geometrical relationships:

- W: distance between the corners of the mouth.
- H: distance between the external contours of the upper and lower lips.
- LM: distance between the chin and the nose.
- Lup: distance between the external contour of the upper lip and the nose.

It must be noticed that some unavoidable unaccuracies in block matching are reflected by the articulatory parameter measures, thus affecting the TDNNs training.

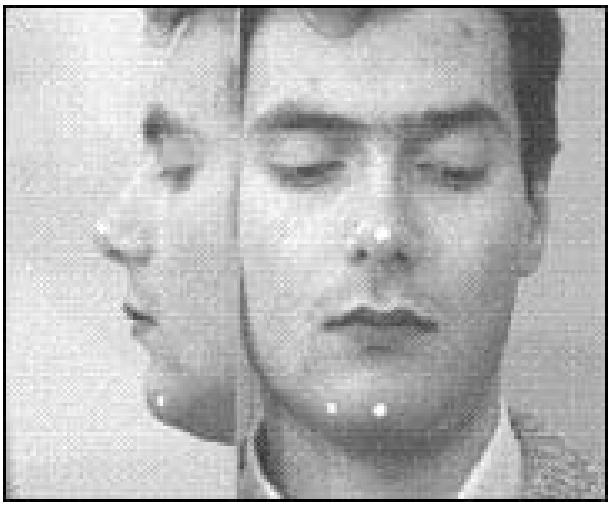

Figure 1: An image of the video sequence. The corners of the lips and the white markers denote the tracking points of the facial features.

Among the different parameters, H is the one which exhibits the largest dynamic range and therefore the one which is less sensitive to extraction errors. For this reason, the H parameter was chosen as a reference for assessing the net convergence.

3.2 Acoustic Mapping

Each acoustic frame (20 msec) extracted from the digitized audio (sampled at 8KHz) was converted to twelve LPC-derived cepstral coefficients. During the pauses within word utterances only noise is present in the audio signal and the corresponding cepstral coefficients have unpredictable values. Therefore, within these silence intervals, there is a sort of incoherence between the cepstral representation and the corresponding set of measured visual parameters. In other words, in the space of the cepstral coefficients, the cluster of the vectors corresponding to silence intervals is quite spread out. Due to the impossibility to associate the articulatory parameters that describe a closed mouth shape to a unique cepstral vector, the TDNNs cannot be trained to recognize the mouth closure between word utterances. If no countermeasures are taken, the so trained TDNNs would estimate unnatural mouth opening/closing in correspondence to inputs without speech content.

Furthermore, it is very important that the TDNNs distinguish between silence intervals associated to short pauses within a word and longer pauses between two consecutive word utterances. For example, due to the coarticulation phenomenon, the mouth is not closed at all during the "e" and "tt" utterances within the Italian word "spaghetti". In order to overcome this problem, a

[1] Cross-validation is a training specifically designed to determine the generalization capability of the net.

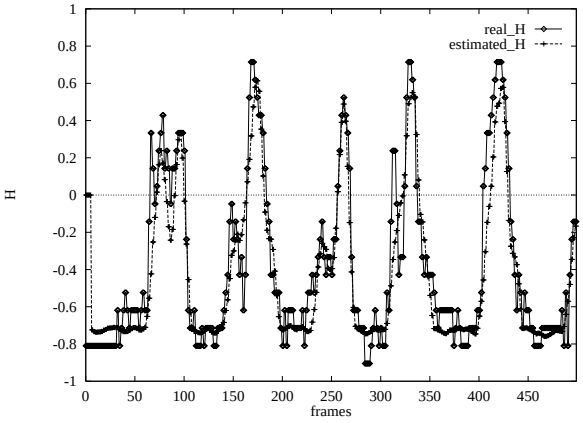

Figure 2: Comparison between the estimated and the actual trajectories of the H parameter. The estimation has been performed replacing cepstrum vectors during silence periods.

pre-defined cepstral configuration has been substituted in correspondence to quite long silence intervals (> 0.1 sec).

This strategy leads to fairly good improvements for the TDNNs estimates, as it can be noticed by comparing the plots in figure 2 related to the articulatory parameter H (results obtained after 15,000 iterations of the TDNNs trained with 40 words, ca. 100 sec).

The performances have been further improved by adaptively adding a random noise to the cepstral coefficients to reduce the likelihood of converging to local minima. Indeed, after adopting this procedure, the trained TDNNs produced better estimates of the articulatory parameters in case of telephone-line corrupted audio signal, the only drawback being a negligible accuracy loss in comparison with the previous trainings in case of noise-free signal.

4 VISUAL SYNTHESIS

The reproduction of the lip movements has been implemented by adopting wire-frame models that reproduce either the face of a synthetic actor [6], as shown in figure 3, or an actual person's face [7][8], as shown in figure 4. In case of a natural actor it is possible to accurately reproduce the pictorial information by texture mapping technique. The lips are animated by moving the vertices according to the parameter estimates.

The obtained performances show a sufficient subjective quality, which allows a quite realistic reproduction of the lip movements in video synthesis up to 25 frames/sec, even in case the audio signal is corrupted by telephone–line noise.

5 EXPERIMENTAL SET–UP

Our experimental data consisted of synchronized audio–visual data. The subject spoke 551 words separated by

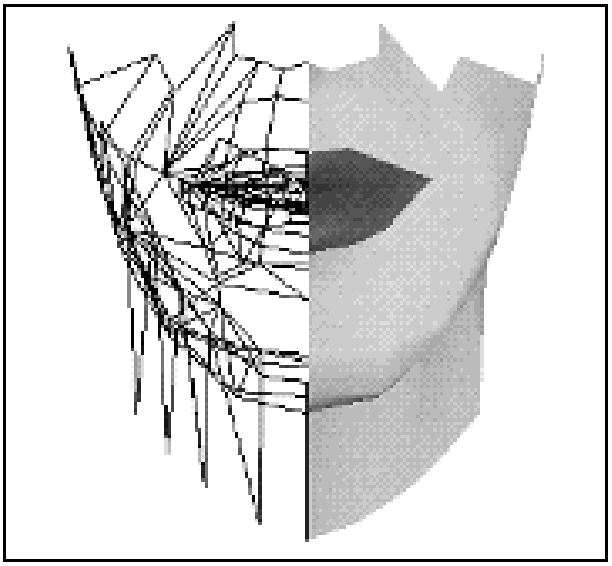

Figure 3: The adopted mouth model sketched both as a wire–frame (left) and as textured (right) patch.

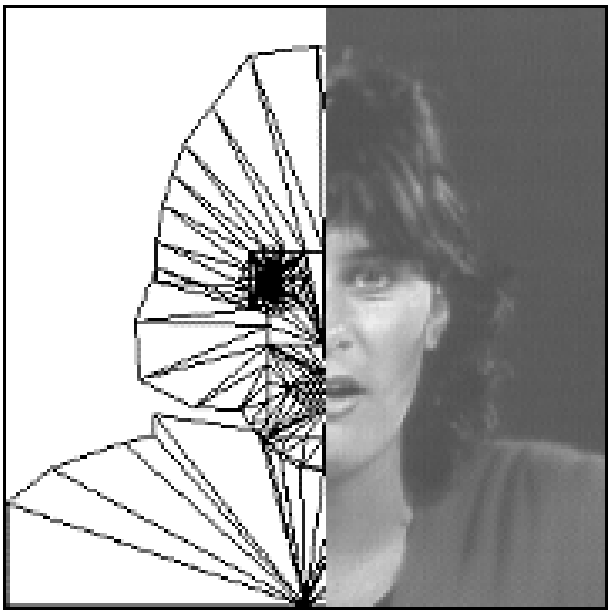

Figure 4: The adopted real actor model sketched both as a wire–frame (left) and as textured (right) patch.

a pause of about 1 sec ("word spotting"). The words were chosen to constitute a phonetically balanced data base, with roughly the same content of low and high frequencies. The audio was sampled at 48 KHz, represented with 16 bit/sample, and distorted by a synthetized telephone–line noise on one stereo channel.

The video was captured at a frame rate of 25 Hertz, with separated YUV components in CIF format with 24 bit/pixel precision. The side view of the person's face was obtained by a suitably oriented mirror.

A green lipstick has been used to enhance the contrast of the lips with respect of the surrounding skin and the teeth; some white markers have been also attached to the nose and the chin.

It was possible to accomplish this huge recording of A/V data thanks to a digital video recorder and massive storage supports. It must also be noticed that some recording options were far like 48 KHz sampling for the audio and chrominance components for the video. The TDNNs were trained on a Silicon Graphics Indigo XZ4000.

6 CONCLUSIONS

A study on the lip movements synthesis has been carried out. The reproduction of the lip movements is performed by means of a synthetic mouth based on a wire–frame model animated by moving the vertices according to the parameter estimates. Each of these parameters is estimated independently by a TDNN trained by means of speech data represented in the cepstrum space and visual articulatory parameters extracted from the synchrounous video sequence. The use of supervised TDNNs allows the analysis of synchronized audio and video data embedding a correct model of the temporal correlation.

Due to the unaffordable complexity, it has not been possible to carry out an exhaustive training of the system, so that performances are somehow low in the case of validation data which differ from training data. For this reason the system has been retrained by adding suitable noise to the input vectors in order to keep it far from undesired local minima configurations. The performances achieved through this new procedure have reached an acceptable level of quality. Adaptive algorithms have also been devised in order to overcome training problems concerning silence learning.

7 ACKNOWLEDGEMENTS

This work has been carried out as a part of the European TIDE project "SPLIT" and ACTS project "VIDAS".

References

[1] E.Casella, F. Lavagetto, R.Miani, "A Time–Delay Neural Network for Speech to Lip Movements Conversion", ICANN–94, Sorrento, May 26–29, 1994.

[2] E.Casella, F.Lavagetto, R.Miani, "A Neural Approach to Lips Movements Modeling", Proc. of EUSIPCO–94, Edinburgh, September 13–16, 1994.

[3] F.Lavagetto, "Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People", IEEE Trans. on Rehabilitation Engineering, Vol.3, n.1, 1995, pp. 90–102.

[4] Ram R. Rao, Tsuhan Chen, "Exploiting Audio-visual Correlation in Coding of Talking Head Sequences", International Picture Coding Symposium, Melbourne, Australia, March 13–15, 1996.

[5] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", 1989, Marcel Dekker, Inc.

[6] F.I. Parke, "Parameterized Models for Facial Animation", IEEE Computer Graphics Applications, Vol.2, n.9, November 1982, pp. 61–68.

[7] F. Lavagetto, S. Curinga, "Object-oriented Scene Modeling for Interpersonal Video Communication at Very Low Bitrate", Image Communication, Vol.6, n.5, October 1994.

[8] S. Curinga, "New techniques of video coding", Ph.D. Thesis, University of Genova, 1994 (in Italian).

[9] A. Waibel, "Neural Network Approaches for Speech Recognition", in Advances in Speech Signal processing, S. Furui and M.M. Sondhi, Eds., 1992.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. lang, "Phoneme Recognition Using Time Delay Neural Networks", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol.37, n.3, March 1989, pp. 328–339.