

HMM INVERSION WITH FULL AND DIAGONAL COVARIANCE MATRICES FOR AUDIO-TO-VISUAL CONVERSION

Lucas D. Terissi and Juan C. Gómez

Laboratory for System Dynamics and Signal Processing, FCEIA, Universidad Nacional de Rosario
CIFASIS, CONICET, Riobamba 245bis, Rosario, Argentina

Keywords: Hidden Markov Models, Audio-Visual Speech Processing, Facial Animation.

Abstract: A speech driven MPEG-4 compliant facial animation system is proposed in this paper. The main feature of the system is the audio-to-visual conversion based on the inversion of an Audio-Visual Hidden Markov Model. The Hidden Markov Model Inversion algorithm is derived for the general case of considering full covariance matrices for the audio-visual observations. A performance comparison with the more common case of considering diagonal covariance matrices is carried out. Experimental results show that the use of full covariance matrices is preferable since it leads to an accurate estimation of the visual parameters, yielding the same performance as in the case of using diagonal covariance matrices, but with a less complex model.

1 INTRODUCTION

The widespread use of multimedia applications such as computer games, online virtual characters, video telephony, and other interactive human-machine interfaces, has made speech driven animation of virtual characters to play an increasingly important role. Several approaches have been proposed in the literature for speech driven facial animation. Among the different models used to generate the visual information (animation parameters) from speech data, the ones based on Hidden Markov Models (HMM) have proved to yield more realistic results.

In the approach proposed by Yamamoto *et al* (Yamamoto *et al.*, 1998), an HMM is learned from audio training data and each video training sequence is aligned with the audio sequence using Viterbi optimization (Viterbi, 1967). In the synthesis stage, the Viterbi alignment algorithm is used to select an optimal HMM state sequence for a novel audio and the visual output associated with each state in the audio sequence is retrieved. This technique provides video predictions of limited quality because the output of each state is the average of the Gaussian mixture components associated to that state, which is only indirectly related to the current audio vector by means of the Viterbi state. Rao and Chen (Rao *et al.*, 1998), (Chen, 2001) proposed a mixture based joint audiovisual HMM, least mean square estimation method is employed for the synthesis phase and the

visual output is made dependent not only on the current state, but also on the current audio input, resulting in improved performance. Brand (Brand, 1999) proposed a cross-modal HMM under the assumption that both acoustic and visual data can be modeled with the same structure. The training is conducted using video data. Once a video HMM is learned, the video output probabilities at each state are remapped onto the audio space using the M-step in Baum-Welch algorithm (Baum and Sell, 1968).

All the above methods rely on the Viterbi algorithm which is rather sensitive to noise in the audio input. To address this limitation, Choi *et al* (Choi *et al.*, 2001) have proposed a Hidden Markov Model Inversion (HMMI) method for audio-visual conversion. HMMI was originally introduced in (Moon and Hwang, 1995) in the context of robust speech recognition. In HMMI, the visual output is generated directly from the given audio input and the trained HMM by means of an expectation-maximization (EM) iteration, thus avoiding the use of the Viterbi sequence and improving the performance of the estimation (Fu *et al.*, 2005). Recently, Xie *et al* (Xie and Liu, 2007) proposed a coupled HMM approach and derived an expectation maximization (EM)-based A/V conversion algorithm for the CHMMs, which converts acoustic speech into decent facial animation parameters.

In this paper, a speech driven MPEG-4 compliant facial animation system is proposed. A joint audio-

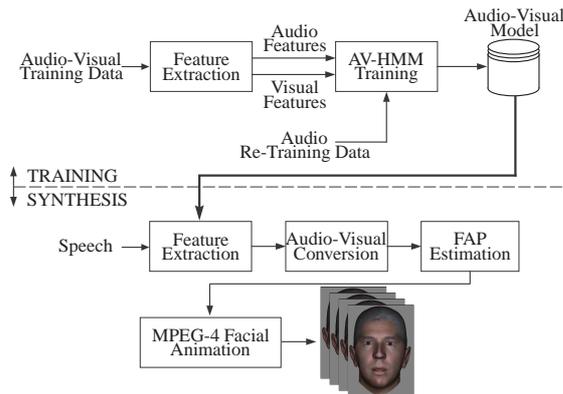


Figure 1: Schematic representation of the speech driven animation system.

visual Hidden Markov Model (AV-HMM) is trained using audio-visual data and then Hidden Markov Model inversion is used to estimate the animation parameters from speech data. The feature vector corresponding to the visual information during the training is obtained via Independent Component Analysis (ICA). Previous approaches based on HMMs consider diagonal covariance matrices for the audio-visual observation, invoking reasons of computational complexity. In this paper, the use of full covariance matrices is investigated. Simulation results show that the use of full covariance matrices leads to an accurate estimation of the visual parameters, yielding the same performance as with diagonal covariance matrices, but with a less complex model and without affecting significantly the computational load.

The rest of the paper is organized as follows. An overview of the speech driven facial animation system is presented in section 2. The AV-HMM is introduced in section 3, where a Hidden Markov Model Inversion algorithm for the general case of considering full covariance matrices for the audio-visual observations is also derived. In section 4, the proposed algorithm for feature extraction is described. The MPEG-4 compliant facial animation technique is presented in section 5. Experimental results and some concluding remarks are included in sections 6 and 7, respectively.

2 SYSTEM OVERVIEW

A block diagram of the proposed speech driven animation system is depicted in Fig. 1. An audiovisual database is used to estimate the parameters of a joint AV-HMM. This database consists of videos of a talking person with reference marks in the region around the mouth, see Fig. 2(a).

In a first training stage, feature parameters of the audiovisual data are extracted. The audio part of the feature vector consists of mel-cepstral coefficients, while the visual part are the coefficients in a ICA representation of the above mentioned set of reference marks. In a second training stage, the audio part of the AV-HMM is re-trained using audio data from a speech-only database. Re-training only the audio part of the model allows to obtain a more robust model against inter-speaker variability, avoiding the need to record videos of speakers with the reference marks on their faces.

For the speech driven animation, speech data is used to estimate the visual features by inversion of the AV-HMM using a technique described in section 3. From these data, Facial Animation Parameters (FAPs) of the MPEG-4 (ISO/IEC IS 14496-2, Visual, 1999) standard are computed to generate the facial animation.

3 AUDIO VISUAL MODEL

In this paper, a joint AV-HMM is used to represent the correlation between the speech and facial movements. The AV-HMM, denoted as λ_{av} , is characterized by three probability measures, namely, the state transition probability distribution matrix (A), the observation symbol probability distribution (B) and the initial state distribution (π), and a set of N states $S = (s_1, s_2, \dots, s_N)$, and audiovisual observation sequence $O_{av} = \{o_{av1}, \dots, o_{avT}\}$. In addition, the observation symbol probability distribution at state j and time t , $b_j(o_{avt})$, is considered a continuous distribution which is represented by a mixture of M Gaussian distributions

$$b_j(o_{avt}) = \sum_{m=1}^M c_{jm} \mathcal{N}(o_{at}, o_{vt}, \mu_{jm}, \Sigma_{jm}) \quad (1)$$

where c_{jm} is the mixture coefficient for the m -th mixture at state j and $\mathcal{N}(o_{at}, o_{vt}, \mu_{jm}, \Sigma_{jm})$ is a Gaussian density with mean μ_{jm} and covariance Σ_{jm} . The audiovisual observation o_{avt} is partitioned as $o_{avt} \triangleq [o_{at}^T, o_{vt}^T]^T$, where o_{at} and o_{vt} are the audio and visual observation vectors, respectively.

A single ergodic HMM is proposed to represent the audiovisual data. An alternative to an ergodic model, would be a set of left-to-right HMMs representing the different phonemes (with associated visemes) of the particular language. These models have been used in the context of speech modeling by several authors, see for instance (Xie and Liu, 2007). An ergodic model provides a more compact representation of the audiovisual data, without the need of

phoneme segmentation, which is required when left-to-right models are used. In addition, this has the advantage of making the system adaptable to any language.

3.1 AV-HMM Training

The training of the AV-HMM consists of two stages, each one using a different database. In the first training stage, an audiovisual database consisting of a set of videos of a single talking person with reference marks drawn on the region around the mouth, is used to estimate the parameters of an ergodic AV-HMM, resorting to the standard Baum-Welch algorithm (Baum and Sell, 1968). Details on the composition of the audiovisual feature vector are given in Section 4, where procedures to take into account audio-visual synchronization and co-articulation are also described. In the second training stage, a speech-only database consisting of audio recordings from a set of talking persons is employed to re-train the audio part of the AV-HMM, leading to a speaker independent model. The re-training is carried out using an only audio HMM (hereafter denoted as A-HMM), with the same structure, which is constructed from the AV-HMM. The A-HMM has the same transition probability and initial state probability matrices obtained in the first stage, while the corresponding observation symbol probability distribution is re-estimated from the speech-only database. The observation symbol probability distribution is parameterized by μ_{jm} , Σ_{jm} and c_{jm} , see eq. (1). To emphasize the mix composition of the AV-HMM, the mean and covariance parameters can be partitioned as

$$\mu_{jm} = \begin{bmatrix} \mu_{jm}^a \\ \mu_{jm}^v \end{bmatrix} \quad \Sigma_{jm} = \begin{bmatrix} \Sigma_{jm}^a & \Sigma_{jm}^{av} \\ \Sigma_{jm}^{va} & \Sigma_{jm}^v \end{bmatrix} \quad (2)$$

where the superscript a and v denote the audio and visual parts, respectively. During the second training stage, only μ_{jm}^a and Σ_{jm}^a are re-estimated using the speech-only data. Finally, the re-estimated parameters are fed back into the AV-HMM.

3.2 Audio-to-Visual Conversion

Hidden Markov Model Inversion (HMMI) was originally proposed in (Moon and Hwang, 1995) in the context of robust speech recognition. Choi and co-authors (Choi et al., 2001) used this technique to estimate the visual features associated to audio features for the purposes of speech driven facial animation. Typically, it is assumed (Moon and Hwang, 1995), (Choi et al., 2001), (Xie and Liu, 2007) a diagonal structure for the covariance matrices of the Gaussian

mixtures, invoking reasons of computational complexity. This assumption is relaxed in this paper allowing for full covariance matrices. This leads to more general expressions for the visual feature estimates.

The idea of HMMI for audio-to-visual conversion is to estimate the visual features based on the trained AV-HMM, in such a way that the probability that the whole audiovisual observation has been generated by the model is maximized. It has been proved (Baum and Sell, 1968) that this optimization problem is equivalent to the maximization of the auxiliary function

$$\begin{aligned} Q(\lambda_{av}; \lambda_{av}, O_a, O_v, O'_v) &\triangleq \\ &\triangleq \sum_{j=1}^N \sum_{m=1}^M P(O_a, O_v, j, m | \lambda_{av}) \log P(O_a, O'_v, j, m | \lambda_{av}) \\ &= \sum_{j=1}^N \sum_{m=1}^M P(O_a, O_v, j, m | \lambda_{av}) \left[\log \pi_{j_0} + \sum_{t=1}^T \log a_{j_{t-1}j_t} \right. \\ &\quad \left. + \sum_{t=1}^T \log \mathcal{N}(o_{at}, o'_{vt}, \mu_{j_t m_t}, \Sigma_{j_t m_t}) + \sum_{t=1}^T \log c_{j_t m_t} \right] \quad (3) \end{aligned}$$

that is

$$O'_v = \arg \max_{O'_v} \{Q(\lambda_{av}; \lambda_{av}, O_a, O_v, O'_v)\} \quad (4)$$

where O_a , O_v and O'_v denote the matrices containing the audio, visual and estimated visual sequences from $t = 1, \dots, T$, respectively, π_{j_0} denotes the initial probability for state j and $a_{j_{t-1}j_t}$ denotes the state transition probability from state j_{t-1} to state j_t .

The solution to the optimization problem in (4) can be computed by equating to zero the derivative of Q with respect to o'_{vt} . Considering that the only term that depends on o'_{vt} is the one involving the Gaussians, this derivative can be written as

$$\begin{aligned} \frac{\partial Q(\lambda_{av}; \lambda_{av}, O_a, O_v, O'_v)}{\partial o'_{vt}} &= 0 \\ 0 &= \sum_{j=1}^N \sum_{m=1}^M P(O_a, O_v, j, m | \lambda_{av}) \times \\ &\quad \times \frac{\partial}{\partial o'_{vt}} \left[\sum_{t=1}^T \log \mathcal{N}(o_{at}, o'_{vt}, \mu_{j_t m_t}, \Sigma_{j_t m_t}) \right] \quad (5) \end{aligned}$$

Considering that

$$\begin{aligned} \log \mathcal{N}(o_{at}, o'_{vt}, \mu_{j_t m_t}, \Sigma_{j_t m_t}) &= \log \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma_{j_t m_t}|}} - \\ &\quad - \frac{1}{2} \begin{bmatrix} o_{at} - \mu_{j_t m_t}^a \\ o_{vt} - \mu_{j_t m_t}^v \end{bmatrix}^T \begin{bmatrix} \Phi_{j_t m_t}^a & \Phi_{j_t m_t}^{av} \\ \Phi_{j_t m_t}^{va} & \Phi_{j_t m_t}^v \end{bmatrix} \begin{bmatrix} o_{at} - \mu_{j_t m_t}^a \\ o_{vt} - \mu_{j_t m_t}^v \end{bmatrix} \quad (6) \end{aligned}$$

where d is the dimension of o_{avt} and

$$\Sigma_{j_t m_t}^{-1} = \begin{bmatrix} \Phi_{j_t m_t}^a & \Phi_{j_t m_t}^{av} \\ \Phi_{j_t m_t}^{va} & \Phi_{j_t m_t}^v \end{bmatrix},$$

the estimated visual observation becomes

$$o'_{vt} = \left[\sum_{j=1}^N \sum_{m=1}^M P(o_a, o_v, j, m | \lambda_{av}) \Phi_{jm}^v \right]^{-1} \times \\ \times \sum_{j=1}^N \sum_{m=1}^M P(o_a, o_v, j, m | \lambda_{av}) \left[\Phi_{jm}^v \mu_{jm}^v - \Phi_{jm}^{va} (o_{at} - \mu_{jm}^a) \right] \quad (7)$$

For the case of diagonal matrices, equation (7) reduces to

$$o'_{vt} = \left[\sum_{j=1}^N \sum_{m=1}^M P(o_a, o_v, j, m | \lambda_{av}) \Phi_{jm}^v \right]^{-1} \times \\ \times \sum_{j=1}^N \sum_{m=1}^M P(o_a, o_v, j, m | \lambda_{av}) \Phi_{jm}^v \mu_{jm}^v \quad (8)$$

which is equivalent to the equation derived in (Choi et al., 2001).

As is common in HMM training, the estimation algorithms (7) and (8) are implemented in a recursive way, initializing the visual observation randomly.

4 FEATURE EXTRACTION

The audio signal is partitioned in frames with the same rate as the video frame rate. A number of Mel-Cepstral Coefficients in each frame (a_t) are used in the audio part of the feature vector. To take into account the audiovisual co-articulation, several frames are used to form the audio feature vector $o_{at} = [a_{t-t_c}^T, \dots, a_{t-1}^T, a_t^T, a_{t+1}^T, \dots, a_{t+t_c}^T]^T$ corresponding to the visual feature vector o_{vt} .

For the visual part, the coefficients in an Independent Component representation of the coordinates of marks in the region around the mouth of the speaking person are used, see Fig. 2(a). Let $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T\}$ represent the training data collected from videos. Each vector $\mathbf{f}_t = [x_1^{(t)}, x_2^{(t)}, \dots, x_p^{(t)}, y_1^{(t)}, y_2^{(t)}, \dots, y_p^{(t)}]^T$ contains the coordinates $(x_p^{(t)}, y_p^{(t)})$ of each mark ($p = 1, 2, \dots, P$) for the t -th frame, $t = 1, 2, \dots, T$.

Let \mathbf{f}_0 be the neutral facial expression, mainly defined as the expression with all face muscles relaxed and the mouth closed (ISO/IEC IS 14496-2, Visual, 1999), the relative facial deformation (with respect to the neutral expression) at each frame can be computed as $\mathbf{d}_t = \mathbf{f}_t - \mathbf{f}_0$, and a deformation matrix can then be defined as

$$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_T] \quad (9)$$

The different facial expressions in the training data are represented by the columns of matrix \mathbf{D} . The idea is to represent any facial expression as the linear

combination of a reduced number of independent vectors. The dimensionality reduction can be performed by Principal Component Analysis (Hyvärinen et al., 2001). The PCA stage yields an uncorrelated set of vectors. It is desirable to have a statistically independent set of vector so that information contained in each vector will not provide information on any of the others. This is the main idea in ICA. Summarizing, ICA after PCA will be performed on the data matrix \mathbf{D} .

Several algorithms are available in the literature for ICA computation. The reader is referred to (Hyvärinen et al., 2001) and the references therein. In this paper, the symmetric decorrelation based Fast-ICA algorithm as implemented in (Gävert et al., 2005) was employed.

As a result of the ICA processing, any facial deformation can then be computed as

$$\mathbf{f}_t = \sum_{k=1}^K o_{vt_k} \mathbf{u}_k + \mathbf{f}_0 \quad (10)$$

where $\{\mathbf{u}_k\}_{k=1}^K$ are the independent components from \mathbf{D} and o_{vt_k} is the k -th component of the visual vector o_{vt} . The coefficients o_{vt_k} are computed in two stages. In the first stage, the mark locations are estimated using image processing techniques. In the second stage, the coefficients o_{vt_k} are computed in such a way that the facial expression is given by the linear combination of the ICs vectors that best match the mark estimation computed in the first stage. Details of this procedure can be found in (Terissi and Gómez, 2007).

5 FACIAL ANIMATION

As already mentioned, the facial animation technique proposed in this paper is MPEG-4 compliant. The MPEG-4 standard defines 64 Facial Animation Parameters and 84 Feature Points (FPs) on a face model in its neutral state (Ostermann, 2002). FAPs represent a complete set of basic facial actions such as head motion, and eye, cheeks and mouth control. FPs are used as reference points to perform the facial deformation.

Based on the estimated facial expression for each frame, the associated FAPs can be determined by computing the displacement of a set of marks from their corresponding position in the neutral facial expression. For instance, the marks encircled in red in Fig. 2(a) can be associated to FAP3 corresponding to jaw opening. Figure 2(b) shows the resulting expression after applying the estimated FAP3 to the neutral expression (several other FAPs, in addition to FAP3, have also been applied to produce the mouth opening



Figure 2: (a) Real person facial expression. Marks associated to FAP3 are encircled in red. (b) Synthesized facial expression.

and cheek movements). Similarly, several subsets of marks can be associated to the different FAPs.

6 EXPERIMENTAL RESULTS

For the audio-visual training, videos of a talking person with reference marks on the region around the person's mouth were recorded at a rate of 30 frames per seconds, with a resolution of (320×240) pixels. The audio was recorded at 11025Hz synchronized with the video. The videos consist of sequences of the Spanish utterances corresponding to the digits zero to nine in random order. For the re-training of the audio part of the AV-HMM, an only-audio database consisting of recordings of sequences of the utterances corresponding to the digits zero to nine by 25 speakers (balance proportion of males and females) was collected.

Experiments were performed with AV-HMM with full and diagonal covariance matrices, different number of states and mixtures in the ranges $[3, 20]$ and $[2, 19]$, respectively, and different values of the co-articulation parameter t_c in the range $[2, 5]$. In the experiments, the audio feature vector a_t is composed by the first eleven non-DC Mel-Cepstral coefficients, while the visual feature vector o_v is of dimension two ($K = 2$ in equation (10)). The performances of the different models were compared by computing the Average Mean Square Error (AMSE)(ϵ), and the Average Correlation Coefficient (ACC)(ρ) between the true and estimated visual parameters, defined as

$$\epsilon = \frac{1}{TK} \sum_{k=1}^K \frac{1}{\sigma_{v_k}^2} \sum_{t=1}^T [o'_{vt_k} - o_{vt_k}]^2 \quad (11)$$

$$\rho = \frac{1}{TK} \sum_{t=1}^T \sum_{k=1}^K \frac{(o_{vt_k} - \mu_{v_k})(o'_{vt_k} - \mu'_{v_k})}{\sigma_{v_k} \sigma'_{v_k}} \quad (12)$$

respectively, where μ_{v_k} and σ_{v_k} denote the mean and the variance of the true visual observation, respectively, and μ'_{v_k} and σ'_{v_k} denote the mean and variance of the estimated visual parameters, respectively.

For the quantification of the visual estimation accuracy, a separate audio-visual dataset, different from the training dataset, was employed. The following results correspond to a co-articulation parameter $t_c = 5$, which proves to be the optimal value in the given range. Fig. 3(a) and Fig. 3(b), show the AMSE and the ACC as a function of the number of states and the number of mixtures for an AV-HMM with full covariance matrix. In this case, equation (7) applies for the estimation of the visual observations o'_{vt} . As can be observed, as the number of states and the number of mixtures increase, the AMSE increases and the ACC decreases, indicating that the accuracy of the estimation deteriorates. This is probably due to the bias-variance tradeoff inherent to any estimation problem. The optimal values for the number of the states and mixtures would be for this case $N = 4$ and $M = 2$, respectively, corresponding to $\epsilon = 0.47$ and $\rho = 0.75$.

Fig. 3(c) and Fig. 3(d), show the AMSE and the ACC as a function of the number of states and the number of mixtures for an AV-HMM with diagonal covariance matrix. In this case, equation (8) applies for the estimation of the visual observations o'_{vt} . As can be observed, to obtain a similar accuracy a more complex model (larger number of states or mixtures) is required. For this case, the optimal values are $N = 19$ and $M = 3$, corresponding to $\epsilon = 0.47$ and $\rho = 0.76$.

The use of full covariance matrices affects the computational complexity during the training stage but, since this is carried out off-line, this does not represent a problem. During the synthesis stage (visual estimation through HMM inversion), and due to the low dimension of the visual feature vector ($K = 2$), the computational load is similar to the case of using diagonal covariance matrices for the same number of states and mixtures.

The above arguments allow one to conclude that the use of full covariance matrices is preferable from the point of view of both computational complexity and accuracy.

The true and estimated visual parameters for the case of full covariance matrices with $N = 4$ states and $M = 2$ mixtures (optimal values) are represented in Fig. 4, where a good agreement can be observed.

7 CONCLUSIONS

A speech driven MPEG-4 compliant facial animation system was introduced in this paper. A joint AV-HMM is proposed to represent the audio-visual data and an algorithm for HMM inversion was derived for the general case of considering full covariance matrix.

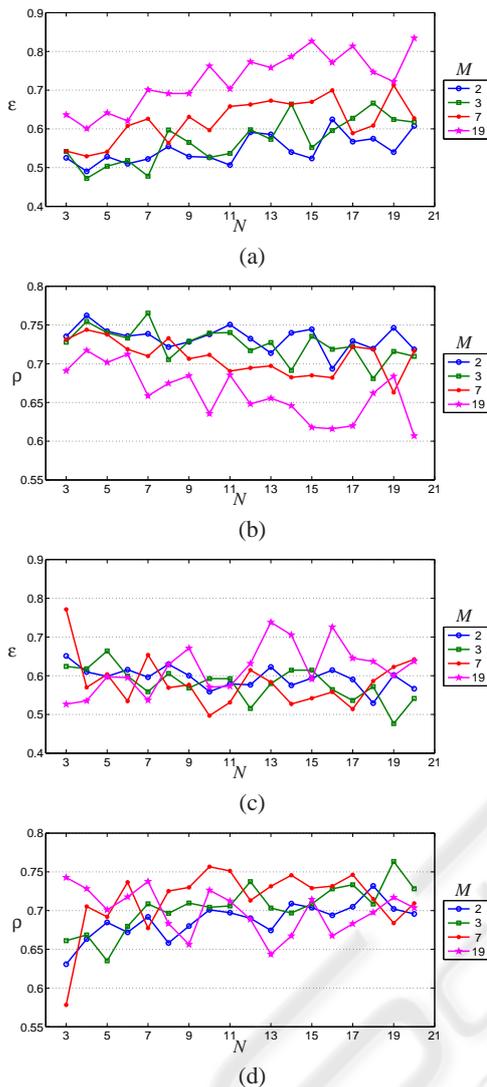


Figure 3: AMSE (ϵ) and ACC (ρ) as a function of the number of states N and the number of mixtures M . Where (a) and (b) correspond to the case of full covariance matrices and, (c) and (d) correspond to the case of diagonal covariance matrices.

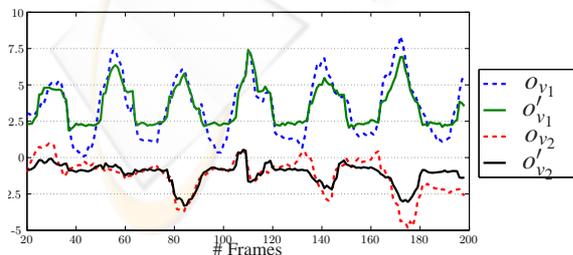


Figure 4: True (dashed line) and estimated (solid line) visual observations.

ces for the audio-visual observations. The influence on the visual estimation accuracy of the use of full

covariance matrices, as opposed to diagonal ones, was investigated. Simulation results show that the use of full covariance matrices leads to an accurate estimation of the visual parameters, yielding the same performance as with diagonal covariance matrices, but with a less complex model and without affecting significantly the computational load.

REFERENCES

- Baum, L. E. and Sell, G. R. (1968). Growth functions for transformations on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227.
- Brand, M. (1999). Voice puppetry. In *Proceedings of SIGGRAPH*, pages 21–28, Los Angeles, CA USA.
- Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, 18(1):9–21.
- Choi, K., Luo, Y., and Hwang, J. (2001). Hidden Markov Model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. *Journal of VLSI Signal Processing*, 29(1-2):51–61.
- Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P., and Garcia, O. (2005). Audio/visual mapping with cross-modal Hidden Markov Models. *IEEE Trans. on Multimedia*, 7(2):243–252.
- Gävert, H., Hurri, J., Särelä, J., and Hyvärinen, A.. FastICA package for MATLAB. Lab. of Computer and Information Science, Helsinki University of Technology.
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, Inc., New York.
- ISO/IEC IS 14496-2, Visual (1999).
- Moon, S. and Hwang, J. (1995). Noisy speech recognition using robust inversion of Hidden Markov Models. In *Proceedings of IEEE International Conf. Acoust., Speech, Signal Processing*, pages 145–148.
- Ostermann, J. (2002). *MPEG-4 Facial Animation - The Standard, Implementation and Applications*, chapter Face Animation in MPEG-4, pages 17–56. John Wiley & Sons.
- Rao, R., Chen, T., and Mersereau, R. (1998). Audio-to-visual conversion for multimedia communication. *IEEE Trans. on Industrial Electronics*, 45(1):15–22.
- Terissi, L. D. and Gómez, J. C. (2007). Facial motion tracking and animation: An ICA-based approach. In *Proceedings of 15th European Signal Processing Conference*, pages 292–296, Poznań, Poland.
- Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. on Information Theories*, 13:260–269.
- Xie, L. and Liu, Z.-Q. (2007). A coupled HMM approach to video-realistic speech animation. *Pattern Recognition*, 40:2325–2340.
- Yamamoto, E., Nakamura, S., and Shikano, K. (1998). Lip movement synthesis from speech based on Hidden Markov Models. *Speech Communication*, 26(1-2):105–115.