Audio-to-Visual Conversion Via HMM Inversion for Speech-Driven Facial Animation

Lucas D. Terissi* and Juan Carlos Gómez

Laboratory for System Dynamics and Signal Processing
FCEIA, Universidad Nacional de Rosario
CIFASIS, CONICET
Riobamba 245bis, 2000, Rosario, Argentina
{lterissi,jcgomez}@fceia.unr.edu.ar

Abstract. In this paper, the inversion of a joint Audio-Visual Hidden Markov Model is proposed to estimate the visual information from speech data in a speech driven MPEG-4 compliant facial animation system. The inversion algorithm is derived for the general case of considering full covariance matrices for the audio-visual observations. The system performance is evaluated for the cases of full and diagonal covariance matrices. Experimental results show that full covariance matrices are preferable since similar, to the case of using diagonal matrices, performance can be achieved using a less complex model. The experiments are carried out using audio-visual databases compiled by the authors.

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

1 Introduction

Speech driven animation of virtual characters is playing an increasingly important role due to the widespread use of multimedia applications such as computer games, online virtual characters, video telephony, and other interactive human-machine interfaces. Among the different approaches proposed in the literature to model audio-visual data, the ones based on Hidden Markov Models (HMM) have proved to yield more realistic results when used in applications of speech driven facial animation.

Earlier approaches for speech-driven facial animation systems, such as the works in [1], [2], [3] and [4], resort to different HMM structures and require the use of Viterbi optimization algorithm [5] in the training or synthesis stages. This leads to video predictions of limited quality due to the high noise sensitivity of Viterbi algorithm. To address this limitation, Choi *et al* [6] have proposed a Hidden Markov Model Inversion (HMMI) method for audio-visual conversion. HMMI was originally introduced in [7] in the context of robust speech recognition. In HMMI, the visual output is generated directly from the given audio

* Corresponding author.

G. Zaverucha and A. Loureiro da Costa (Eds.): SBIA 2008, LNAI 5249, pp. 33–42, 2008.

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 [8]. Recently, Xie *et al* [9] proposed a coupled HMM approach and derived an expectation maximization (EM)-based A/V conversion algorithm for the CHMMs, which converts acoustic speech into reasonably good facial animation parameters.

In this paper, a speech driven MPEG-4 compliant facial animation system is proposed. A joint audio-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 a performance similar to that of using 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 an HMMI 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 Speech Driven Facial Animation 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. Retraining 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 the technique described in section 3. From these data, Facial Animation Parameters (FAPs) of the MPEG-4 [10] standard are computed to generate the facial animation.

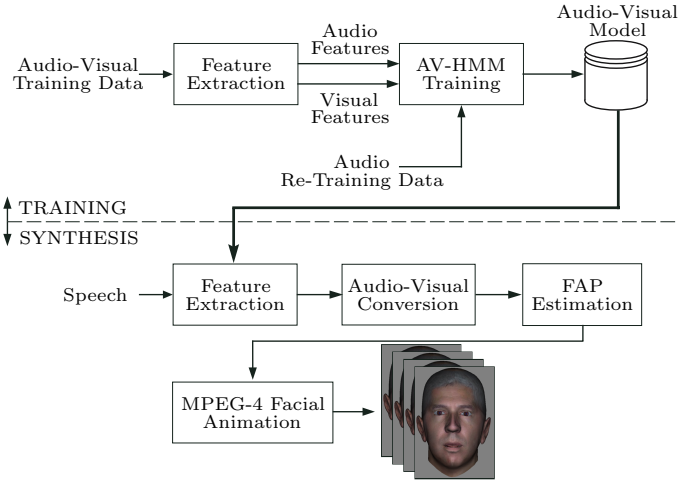

Fig. 1. Schematic representation of the speech driven animation system

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 $S = (s_1, s_2, \ldots, s_N)$, and audiovisual observation sequence $O_{av} = \{o_{av1}, \ldots, o_{avT}\}$. 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

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

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 audio-visual observation o_{avt} is partitioned as $o_{avt} \triangleq \left[o_{at}^T, o_{vt}^T\right]^T$, where o_{at} and o_{vt} are the audio and visual observation vectors, respectively.

A single ergodic (that is one in which transitions among all the states are allowed) 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 [9]. 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 [11]. 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 equation (1). To emphasize the mix composition of the AV-HMM, the mean and covariance parameters can be partitioned as

$$\mu_{jm} = \begin{bmatrix} \mu^a_{jm} \\ \mu^v_{jm} \end{bmatrix} \quad , \quad \Sigma_{jm} = \begin{bmatrix} \Sigma^a_{jm} & \Sigma^{av}_{jm} \\ \Sigma^{va}_{jm} & \Sigma^v_{jm} \end{bmatrix} \ , \tag{2}$$

where the superscript a and v denote the audio and visual parts, respectively. During the second training stage, only μ^a_{jm} and Σ^a_{jm} are re-estimated using 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 [7] in the context of robust speech recognition. Choi and co-authors [6] used this technique to estimate the visual features associated to audio features for the purposes of speech driven facial animation. Typically, it is assumed [7], [6], [9] 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 [11] that this optimization problem is equivalent to the maximization of the auxiliary function

$$Q(\lambda_{av};\lambda_{av},O_a,O_v,O'_v) \triangleq \sum_{j=1}^{N}\sum_{m=1}^{M} P(O_a,O_v,j,m \mid \lambda_{av}) \ \log \ P(O_a,O'_v,j,m \mid \lambda_{av})$$

$$= \sum_{j=1}^{N}\sum_{m=1}^{M} P(O_a,O_v,j,m \mid \lambda_{av}) \ \left[\log \ \pi_{j_0} + \sum_{t=1}^{T} \log \ a_{j_{t-1}j_t} + \right.$$

$$\left. + \sum_{t=1}^{T} \log \ \mathcal{N}(o_{at},o'_{vt},\mu_{j_tm_t},\Sigma_{j_tm_t}) + \sum_{t=1}^{T} \log \ c_{j_tm_t}\right] \ , \tag{3}$$

that is

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

where O_a, O_v and O'_v denote the matrices containing the audio, visual and estimated visual sequences from $t = 1,\ldots,T$, 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

$$\frac{\partial Q(\lambda_{av};\lambda_{av},O_a,O_v,O'_v)}{\partial o'_{vt}} = \sum_{j=1}^{N}\sum_{m=1}^{M} P(O_a,O_v,j,m \mid \lambda_{av}) \times$$

$$\times \frac{\partial}{\partial o'_{vt}} \left[\sum_{t=1}^{T} \log \ \mathcal{N}(o_{at},o'_{vt},\mu_{j_tm_t},\Sigma_{j_tm_t})\right] = 0 \ . \tag{5}$$

Considering that

$$\log \ \mathcal{N}(o_{at},o'_{vt},\mu_{j_tm_t},\Sigma_{j_tm_t}) = \log \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_{j_tm_t}|}} -$$

$$- \frac{1}{2} \begin{bmatrix} o_{at} - \mu^a_{j_tm_t} \\ o_{vt} - \mu^v_{j_tm_t} \end{bmatrix}^T \begin{bmatrix} \Phi^a_{j_tm_t} & \Phi^{av}_{j_tm_t} \\ \Phi^{v_ta_t}_{j_tm_t} & \Phi^v_{j_tm_t} \end{bmatrix} \begin{bmatrix} o_{at} - \mu^a_{j_tm_t} \\ o_{vt} - \mu^v_{j_tm_t} \end{bmatrix} \ , \tag{6}$$

where d is the dimension of o_{avt} and

$$\Sigma^{-1}_{j_tm_t} = \begin{bmatrix} \Phi^a_{j_tm_t} & \Phi^{av}_{j_tm_t} \\ \Phi^{va}_{j_tm_t} & \Phi^v_{j_tm_t} \end{bmatrix},$$

the estimated visual observation becomes

$$o'_{vt} = \left[\sum_{j=1}^{N}\sum_{m=1}^{M} P(o_a,o_v,j,m \mid \lambda_{av})\Phi^v_{jm}\right]^{-1} \times$$

$$\times \sum_{j=1}^{N}\sum_{m=1}^{M} P(o_a,o_v,j,m \mid \lambda_{av}) \left[\Phi^v_{jm}\mu^v_{jm} - \Phi^{va}_{jm}(o_{at} - \mu^a_{jm})\right] \ . \tag{7}$$

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

$$o'_{vt} = \left[\sum_{j=1}^{N}\sum_{m=1}^{M} P(o_a, o_v, j, m \mid \lambda_{av})\Phi^v_{jm}\right]^{-1} \times \sum_{j=1}^{N}\sum_{m=1}^{M} P(o_a, o_v, j, m \mid \lambda_{av})\Phi^v_{jm}\mu^v_{jm} \ , \tag{8}$$

which is equivalent to the equation derived in [6].

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 $o_{at} = \left[a^T_{t-t_c}, \ldots, a^T_{t-1}, a^T_t, a^T_{t+1}, \ldots, a^T_{t+t_c}\right]^T$ 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 $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$ represent the training data collected from videos. Each vector $\mathbf{f}_t = [x^{(t)}_1, x^{(t)}_2, \ldots, x^{(t)}_P, y^{(t)}_1, y^{(t)}_2, \ldots, y^{(t)}_P]^T$ contains the coordinates $(x^{(t)}_p, y^{(t)}_p)$ of each mark $(p = 1, 2, \ldots, P)$ for the t-th frame, $t = 1, 2, \ldots, T$.

Let $\mathbf{f}_0$ be the neutral facial expression, mainly defined as the expression with all face muscles relaxed and the mouth closed [10]. 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

$$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_T] \ . \tag{9}$$

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 [12]. 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 [12] and the references therein. In this paper, the symmetric decorrelation based FastICA algorithm as implemented in [13] was employed.

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

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

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 [14].

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 [15]. 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 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 (320×240) pixels resolution. 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

(a) (b)

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

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

$$\epsilon = \frac{1}{TK} \sum_{k=1}^{K} \frac{1}{\sigma_{v_k}^2} \sum_{t=1}^{T} \left[o'_{vt_k} - o_{vt_k} \right]^2 , \tag{11}$$

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

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 states and mixtures could 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

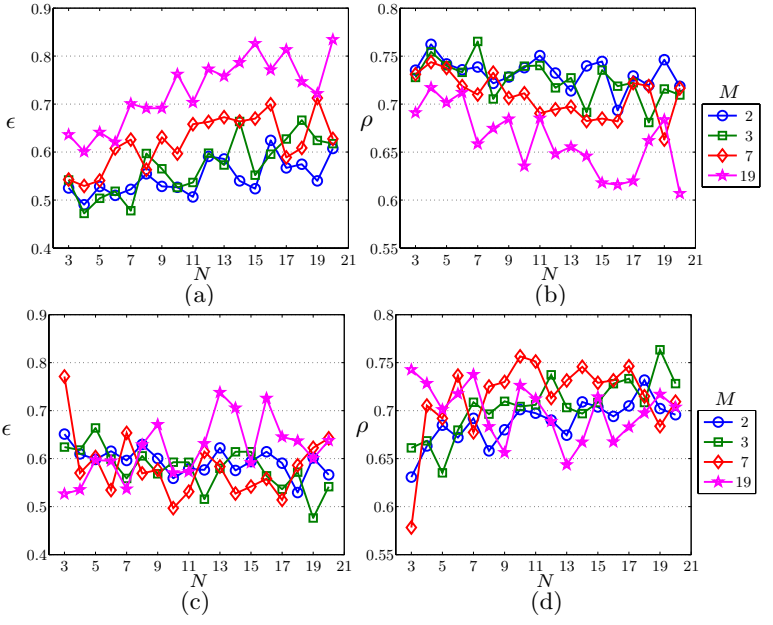

Fig. 3. *AMSE* (ϵ) and *ACC* (ρ) as a function of the number of states N and the number of mixtures M. Where (a) and (b) correspond to the case of full covariance matrices and, (c) and (d) correspond to the case of diagonal covariance matrices.

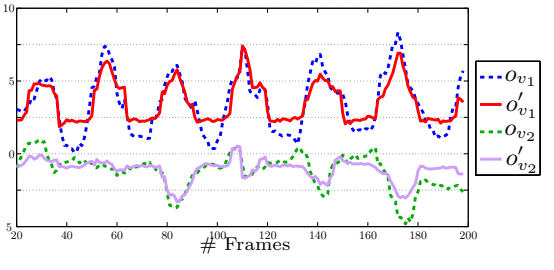

Fig. 4. True (dashed line) and estimated (solid line) visual observations

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 agrement 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 matrices 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 a performance similar to that of using diagonal covariance matrices, but with a less complex model and without affecting the computational load.

References

1. Yamamoto, E., Nakamura, S., Shikano, K.: Lip movement synthesis from speech based on Hidden Markov Models. Speech Communication 26(1-2), 105–115 (1998)
2. Rao, R., Chen, T., Mersereau, R.: Audio-to-visual conversion for multimedia communication. IEEE Trans. on Industrial Electronics 45(1), 15–22 (1998)
3. Chen, T.: Audiovisual speech processing. IEEE Signal Processing Magazine 18(1), 9–21 (2001)
4. Brand, M.: Voice puppetry. In: Proceedings of SIGGRAPH, Los Angeles, CA USA, pp. 21–28 (August 1999)
5. Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. on Information Theories 13, 260–269 (1967)
6. Choi, K., Luo, Y., Hwang, J.: 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 (2001)
7. Moon, S., Hwang, J.: Noisy speech recognition using robust inversion of Hidden Markov Models. In: Proceedings of IEEE International Conf. Acoust., Speech, Signal Processing, pp. 145–148 (1995)
8. Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P., Garcia, O.: Audio/visual mapping with cross-modal Hidden Markov Models. IEEE Trans. on Multimedia 7(2), 243–252 (2005)
9. Xie, L., Liu, Z.Q.: A coupled HMM approach to video-realistic speech animation. Pattern Recognition 40, 2325–2340 (2007)
10. ISO/IEC IS 14496-2, Visual (1999)
11. Baum, L.E., Sell, G.R.: Growth functions for transformations on manifolds. Pacific Journal of Mathematics 27(2), 211–227 (1968)
12. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, Inc., New York (2001)
13. Gävert, H., Hurri, J., Särelä, J., Hyvärinen, A.: FastICA package for MATLAB. Lab. of Computer and Information Science, Helsinki University of Technology
14. Terissi, L.D., Gómez, J.C.: Facial motion tracking and animation: An ICA-based approach. In: Proceedings of 15th European Signal Processing Conference, Poznań, Poland, September 3-7, pp. 292–296 (2007)
15. Ostermann, J.: Face Animation in MPEG-4. In: MPEG-4 Facial Animation - The Standard, Implementation and Applications, pp. 17–56. John Wiley & Sons, Chichester (2002)