Multimedia Mapping using Continuous State Space Models

Tue Lehn-Schiøler
The Technical University of Denmark
Informatics and Mathematical Modelling
Email: tls@imm.dtu.dk

***Abstract*— In this paper a system that transforms speech waveforms to animated faces are proposed. The system relies on a state space model to perform the mapping. To create a photo realistic image an Active Appearance Model is used. The main contribution of the paper is to compare a Kalman filter and a Hidden Markov Model approach to the mapping. It is shown that even though the HMM can get a higher test likelihood the Kalman filter is easier to train and the animation quality is better for the Kalman filter.**

I. Introduction

The motivation for transforming a speech signal into lip movements is at least threefold. Firstly, the language synchronization of movies often leaves the actors mouth moving while there is silence or the other way around, this looks rather unnatural. If it was possible to manipulate the face of the actor to match the actual speech it would be much more pleasant to view synchronized movies (and a lot easier to make cartoons). Secondly, even with increasing bandwidth sending images via the cell phone is quite expensive, therefore, a system that allows single images to be sent and models the face in between would be useful. The technique will also make it possible for hearing impaired people to lip read over the phone. If the person in the other end does not have a camera on her phone, a model image can be used to display the facial movements. Thirdly, when producing agents on a computer (like Windows Office Mr. clips) it would make communication more plausible if the agent could interact with lip movements corresponding to the (automatically generated) speech.

The idea of extracting phonemes or similar high-level features from the speech signal before performing the mapping to the mouth position has been widely used in the lip-sync community. Goldenthal (1) suggested a system called ”Face Me!”. He extracts phonemes using Statistical Trajectory Modeling. Each phoneme is then associated with a mouth position (keyframe). In Mike Talk (2), phonemes are generated from text and then mapped onto keyframes, however, in this system trajectories linking all possible keyframes are calculated in advance thus making the video more seamless. In ”Video rewrite” (3) phonemes are again extracted from the speech, in this case using Hidden Markov Models. Each triphone (three consecutive phonemes) has a mouth sequence associated with it. The sequences are selected from training data, if the triphone does not have a matching mouth sequence in the training data, the closest available sequence is selected. Once the sequence of mouth movements has been determined, the mouth is mapped back to a background face of the speaker. Other authors have proposed methods based on modeling of phonemes by correlational HMM's (4) or neural networks (5).

Methods where speech is mapped directly to facial movement are not quite as popular as phoneme based methods. However, in 'Picture my voice' (6), a time dependent neural network, maps directly from 11×13 Mel Frequency Cepstral Coefficients (MFCC) as input to 37 facial control parameters. The training output is provided by a phoneme to animation mapping but the trained network does not make use of the phoneme representation. Also Brand (7) has proposed a method based on (entropic) HMM's where speech is mapped directly to images. In (8) Nakamura presents an overview of methods using HMM's, the first MAP-V converts speech into the most likely HMM state sequence and the uses a table lookup to convert into visual parameters. In an extended version MAP-EM the visual parameters are estimated using the EM algorithm. Methods that do not rely on phoneme extraction has the advantage that they can be trained to work on all languages, and that they are able to map non-speech sounds like yawning and laughing.

There are certain inherent difficulties in mapping from speech to mouth positions an analysis of these can be found in (9). The most profound is the confusion between visual and auditive information. The mouth position of sounds like /b/,/p/ and /m/ or /k/,/n/ and /g/ can not be distinguished even though the sounds can. Similarly the sounds of /m/ and /n/ or /b/ and /v/ are very similar even though the mouth position is completely different. This is perhaps best illustrated by the famous experiment by McGurk (10). Thus, when mapping from speech to facial movements, one cannot hope to get a perfect result simply because it is very difficult to distinguish whether a ”ba” or a ”ga” was spoken.

The rest of this paper is organized in three sections, section II focuses on feature extraction in sound and images, in section III the model are described. Finally experimental results are presented in section IV.

II. Feature extraction

Many different approaches has been taken for extraction of sound features. If the sound is generated directly from text phonemes can be extracted directly and there is no need to process the sound track (2). However, when a direct mapping is performed one can choose from a variety of features. A non-complete list of possibilities include Perceptual Linear

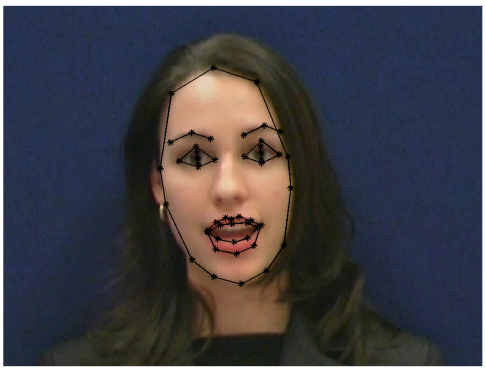


Fig. 1. Image with automatically extracted feature points. The facial feature points are selected from the MPEG-4 standard

Prediction or J-Rasta-PLP as in (7; 11), Harmonics of Discrete Fourier Transform as in (12), Linear Prediction Coefficients as in (13) or Mel Frequency Cepstral Coefficients (1; 5; 6; 8). In this work the sound is split into 25 blocks per second (the same as the image frame rate) and 13 MFCC features are extracted from each block. To extract features from the images an Active Appearance model (AAM) (14) is used. The use of this model for lipreading has previously been studied by Mathews et al. (15). AAM's are also useful for low bandwidth transmission of facial expressions (16). In this work an implementation by Mikkel B. Stegman (17) is used. For the extraction a suitable subset of images in the training set are selected and annotated with points according to the MPEG-4 facial animation standard. Using these annotations a 14-parameter model of the face is created. Thus, with 14 parameters it is possible to create a photo realistic image of any facial expression seen in the training set. Once the AAM is created the model is used to track the lip movements in the image sequences, at each point the 14 parameters are picked up. In Fig. 1 the result of the tracking is shown for a single representative image.

III. Model

In this work the mapping from sound to images is performed by two types of state space models, a HMM with a mixture of Gaussians observations and a Kalman filter. Both approaches uses the toolbox written by Kevin Murphy (http://www.ai.mit.edu/ murphyk/Software).

Normally, when using HMM's for speech to face-movement mapping a bank of HMM's are used. Each one is trained on a specific subset of data and when that model has the highest likelihood it is responsible for producing the image. In this work the entire sequence is considered at once and only a single state space model is trained. In case of the Kalman filter the model set up is as follows:

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{n}_k^x \quad (1)$$

$$\boldsymbol{s}_k = \boldsymbol{B}\boldsymbol{x}_k + \boldsymbol{n}_k^s \quad (2)$$

$$\boldsymbol{i}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{n}_k^i \quad (3)$$

In this setting $\boldsymbol{i}_k$ is the image features at time k, $\boldsymbol{s}_k$ is the sound features and $\boldsymbol{x}_k$ is a hidden variable without physical meaning. $\boldsymbol{x}$ can be thought of as some kind of brain activity controlling what is said. Each equation has i.i.d. Gaussian noise component $\boldsymbol{n}$ added to it.

During training both sound and image features are known, and the two observation equations can be collected in one.

$$\begin{pmatrix} \boldsymbol{s}_k \\ \boldsymbol{i}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{C} \end{pmatrix} \boldsymbol{x}_k + \begin{pmatrix} \boldsymbol{n}_k^s \\ \boldsymbol{n}_k^i \end{pmatrix} \quad (4)$$

By using the EM algorithm (18; 19) on the training data, all parameters $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Sigma}^x, \boldsymbol{\Sigma}^s, \boldsymbol{\Sigma}^i\}$ can be found. $\boldsymbol{\Sigma}$'s are the diagonal covariance matrices of the noise components.

When a new sound sequence arrives Kalman filtering (or smoothing) can be applied to equations (1,2) to obtain the hidden state $\boldsymbol{x}$. Given $\boldsymbol{x}$ the corresponding image features can be obtained by multiplication, $\boldsymbol{i}_k = \boldsymbol{C}\boldsymbol{x}_k$. If the intermediate smoothing variables are available the variance on $\boldsymbol{i}_k$ can also be calculated.

In case of the Hidden Markov Model the approach is similar, the transition probabilities, the emission probabilities for the sound and image features and the Gaussian mixture parameters are estimated during training. During testing the most probable state sequence can be found from the sound features and the image feature can be found using either the mean of the emitted Gaussian or by drawing a sample from it.

IV. Results

The data used is taken from the vidtimit database (20). The database contains recordings of large number of people each uttering ten different sentences while facing the camera. The sound recordings are degraded by fan-noise from the recording pc. In this work a single female speaker is selected, thus 10 different sentences are used, nine for training and one for testing.

To find the dimension of the hidden state ($\boldsymbol{x}$), the optimal parameters for both the KF and the HMM were found for varying dimensions. For each model the likelihood on training and test sequences were calculated, the result is shown in Fig. 2 and Fig. 3.

The test likelihood provides a statistical measure of the quality of the model and provides a way of comparing models. This allows comparison between the KF and the HMM approach. Unfortunately the likelihood is not necessarily a good measure of the quality of a model prediction. If the distributions in the model are broad, i.e. the model has high uncertainty, it can describe data well, but, it is not a good generative model.

Looking at the results in Fig. 2 and Fig. 3 it is seen that the likelihood of a HMM does not increase as expected with the model complexity. The KF on the other hand has a peak in the test likelihood around 40 hidden states. The test likelihood shows that the HMM is a better model than KF. However when examining the output feature vectors controlling the face movement (Fig. 4) it is seen that the output of the HMM is varying very fast and does not follow the true feature vector. The output from the KF on the other hand is smooth and closer to the desired. Visual inspection of the video sequence shows good results from the

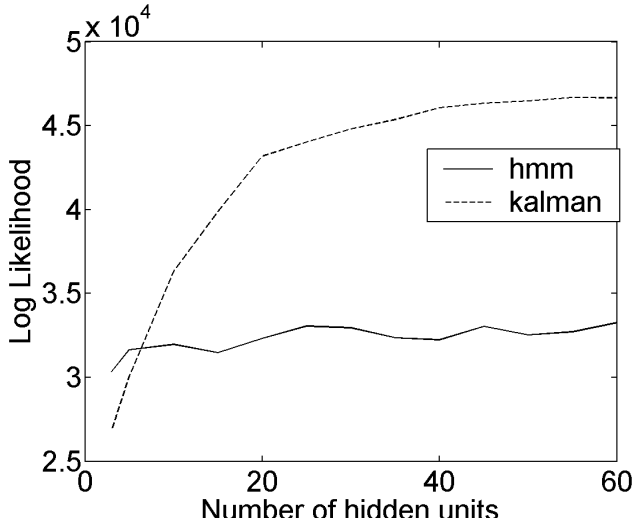


Fig. 2. The likelihood evaluated on the training data. The Kalman filter is able to utilize the extra dimension to improve the training result, whereas the HMM has almost the same performance no matter how many hidden states are used.

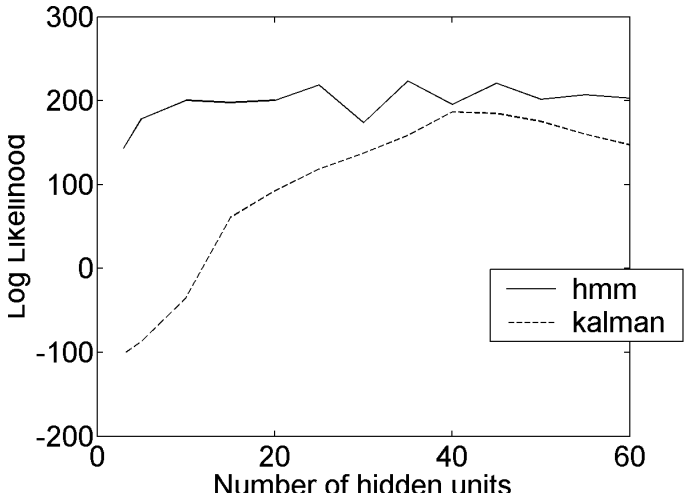


Fig. 3. The likelihood evaluated on the test data. Again the Kalman filter improves performance as more hidden dimensions are added and overfitting is seen for high number of hidden states. The HMM has the same performance independent of the number of hidden states.

KF but very jerky and unrealistic motion from the HMM. In Fig. 5 snapshots from the KF sequence are provided for visual inspection, the entire sequence is available at http://www.imm.dtu.dk/˜tls/code/facedemo.php, where other demos can also be found.

The failure of the likelihood to capture the quality of the final image sequence points to an interesting problem. No precise metric exist for evaluation of synthesized lip sequences. The distance between facial points in the true and the predicted image would be one way, another way would be to measure the distance between the predicted feature vector and the feature vector extracted from the true image. However, the ultimate evaluation of faces can be only provided by human interpretation. Unfortunately it is difficult to get an objective measure this way. One possibility would be to get a hearing impaired person to lipread the generated sequence, another to let people try to guess which sequence was real and which was computer generated. Unfortunately, such test are time and labor demanding. Further more these subjective test does not provide an error function that can be optimized directly.

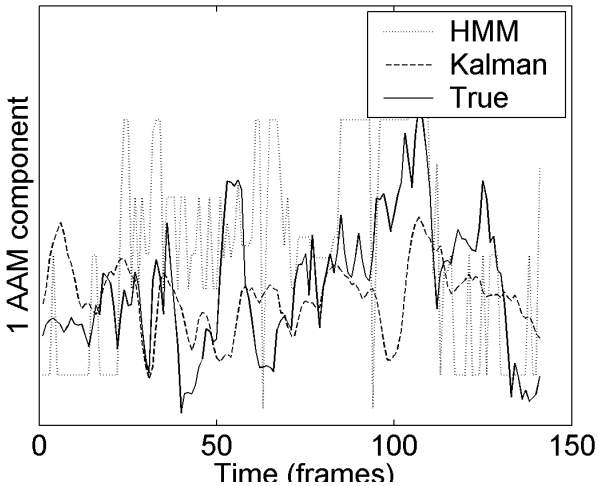


Fig. 4. Dynamics of the first AAM component true and predicted. The prediction from the Kalman model is smooth, but does not follow the curve completely. The HMM solution varies faster indicating that the uncertainty in the model is greater. On top of that the discrete nature of the model makes it jumps suddenly from frame to frame making the face movement look jerky.

V. Conclusion

A speech to face mapping system relying on state space models is proposed. The system makes it possible to easily train a unique face model that can be used to transform speech into facial movements. The training set must contain all sounds and corresponding face gestures, but there are no language or phonetic requirements to what the model can handle.

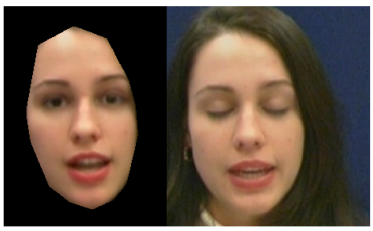
In this approach a single model is used for the entire sequence, making the problem one of system identification. For this task the Hidden Markov Model seem to be clearly inferior to the Kalman filter based on inspection of the output video. The likelihood of the HMM however shows that it is the better model. This confirms the suspicion that better error measures are needed for evaluation of lip-sync quality.

References

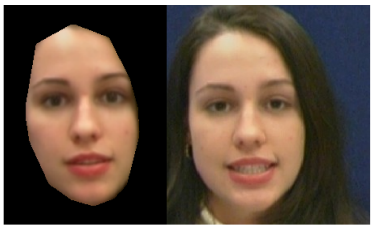
[1] W. Goldenthal, K. Waters, T. V. Jean-Manuel, and O. Glickman, “Driving synthetic mouth gestures: Phonetic recognition for faceme!” in *Proc. Eurospeech ’97*, Rhodes, Greece, 1997, pp. 1995–1998.

[2] T. Ezzat and T. Poggio, “Mike talk: a talking facial display based on morphing visemes,” *Proc. Computer Animation IEEE Computer Society*, pp. 96–102, 1998.

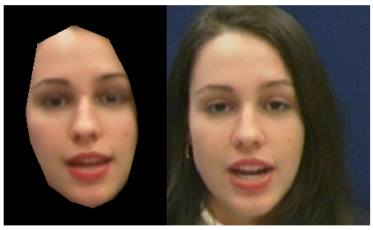
[3] C. Bregler, M. Covell, and M. Slaney, “Video rewrite: driving visual speech with audio,” in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1997, pp. 353–360.



(a)



(b)



(c)

Fig. 5. Characteristic images taken from the test sequence when using the Kalman filter. The predicted face is to the left and the true face to the right.

[4] J. J. Williams and A. K. Katsaggelos, "An hmm-based speech-to-video synthesizer," 1998.

[5] P. Hong, Z. Wen, and T. S. Huang, "Speech driven face animation," in *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, I. S. Pandzic and R. Forchheimer, Eds. Wiley, Europe, July 2002.

[6] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," *Proc. AVSP 99*, 1999.

[7] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 21–28.

[8] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, july 2002.

[9] L. F., "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. on Rehabilitation Engineering*, vol. 3, no. 1, 1995.

[10] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[11] S. Dupont and J. Luettin, "Audio-visual speech modelling for continuous speech recognition," *IEEE Transactions on Multimedia*, 2000.

[12] D. F. McAllister, R. D. Rodman, D. L. Bitzer, and A. S. Freeman, "Speaker independence in automated lip-sync for audio-video communication," *Comput. Netw. ISDN Syst.*, vol. 30, no. 20-21, pp. 1975–1980, 1998.

[13] J. P. Lewis, "Automated lip-sync: Background and techniques," *J. Visualization and Computer Animation*, vol. 2, 1991.

[14] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *Proc. European Conference on Computer Vision*, vol. 2, pp. 484–498, 1998.

[15] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198 –213, 2002.

[16] B. Theobald, S. Kruse, J. Bangham, and G. Cawley, "Towards a low bandwidth talking face using appearance models," *Image and Vision Computing*, vol. 21, no. 13-14, pp. 1117–1124, 2003.

[17] M. B. Stegmann, "Analysis and segmentation of face images using point annotations and linear subspace techniques," Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Tech. Rep., Aug. 2002, http://www.imm.dtu.dk/pubdb/p.php?922.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JRSSB*, vol. 39, pp. 1–38, 1977.

[19] Z. Ghahramani and G. Hinton, "Parameter estimation for linear dynamical systems," Tech. Rep., 1996, university of Toronto, CRG-TR-96-2.

[20] C. Sanderson and K. K. Paliwal, "Polynomial features for robust face authentication," *Proceedings of International Conference on Image Processing*, vol. 3, pp. 997–1000, 2002.

[21] T. Lehn-Schioler, L. K. Hansen, and J. Larsen, "Mapping from speech to images using continuous state space models," in *Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Apr. 2004.