MAPPING FROM SPEECH TO IMAGES USING CONTINUOUS STATE SPACE MODELS

Tue Lehn-Schiøler, Lars Kai Hansen & Jan Larsen

The Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, Bld. 321, Web: www.imm.dtu.dk
Email: (tls,lkh,jl)@imm.dtu.dk

ABSTRACT

In this paper a system that transforms speech waveforms to animated faces are proposed. The system relies on continuous state space models to perform the mapping, this makes it possible to ensure video with no sudden jumps and allows continuous control of the parameters in 'face space'.
The performance of the system is critically dependent on the number of hidden variables, with too few variables the model cannot represent data, and with too many overfitting is noticed.

Simulations are performed on recordings of 3-5 sec. video sequences with sentences from the Timit database. From a subjective point of view the model is able to construct an image sequence from an unknown noisy speech sequence even though the number of training examples are limited.

1. 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 work is supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778.

Lewis [1] provides an early overview paper about state of the art lip-sync in 1991. He concludes that using loudness to control the jaw is not a useful approach since sounds made with closed mouth can be just as loud as open mouth sounds. He also notes that the spectrum matching method used by MIT in the early 1980's has severe problems due to the formants independence of pitch. In this method the shape of the mouth is determined from the frequency content of the speech. The problem is illustrated by the fact that the mouth shape is the same when a sound e.g. an 'a' is spoken with a high or a deep voice. Final he mentions that it is possible to automatically generate speech from text and in this way gain control of what phoneme to visualize. In his view the speech synthesis in 1991 was not of sufficient quality to sound natural, and although progress has been made in the field automatic generated speech is still far from perfect. The suggestion in [1] is to extract phonemes using a Linear Prediction speech model and then map the phonemes to keyframes given by a lip reading chart.

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 [2] 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 [3], 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" [4] 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 [5] or neural networks [6].

Methods where speech is mapped directly to facial movement are not quite as popular as phoneme based methods. However, in 'Picture my voice' [7], a time dependent neural network, maps directly from 11×13 Mel Frequency Cepstral Coefficients (MFCC) as inputto 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 [8] has proposed a method based on (entropic) HMM's where speech is mapped directly to images. 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 or laughing.

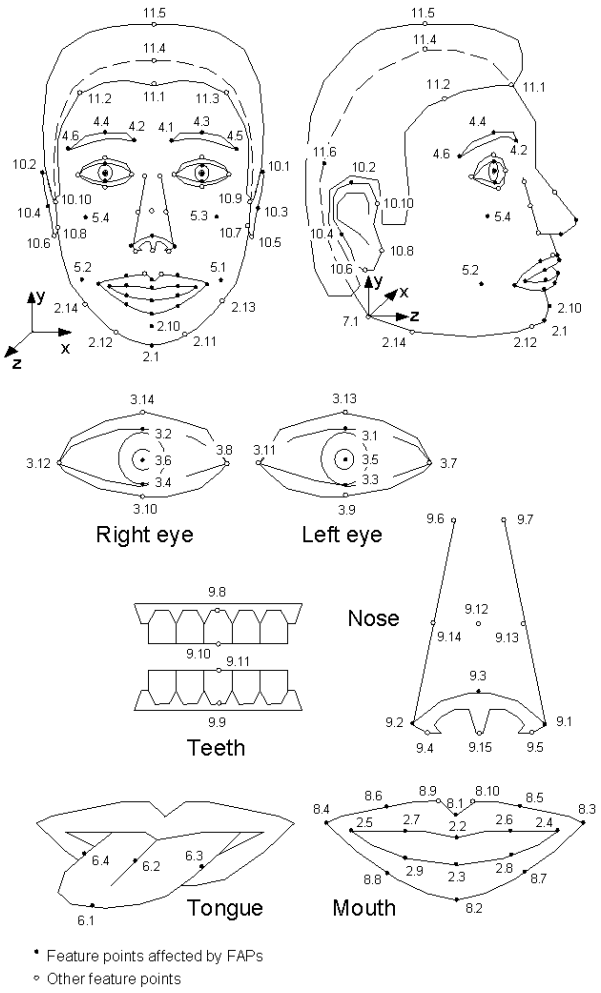
There are certain inherent difficulties in mapping from speech to mouth positions an analysis of these can be found in [9]. The most profound is the confusion between visual and auditive information. The mouth position of sounds like /b/,/p/ and /m/ or /k/,/n/ and /g/ can not be distinguished even though 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.

2. FEATURE EXTRACTION

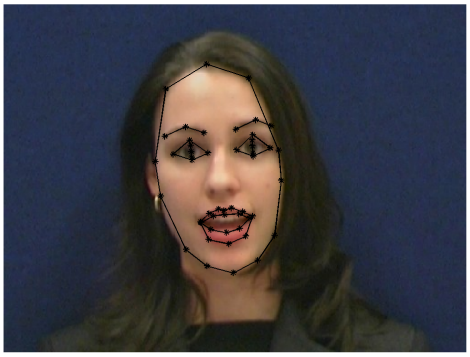
Many different approaches have been taken for extraction of sound features. If the sound is generated directly from text [3], phonemes can be extracted directly and there is no need to process the sound track. However, when a direct mapping is performed one can choose from a variety of features. A non-complete list of possibilities include Perceptual Linear Prediction or J-Rasta-PLP as in [11, 8], Harmonics of Discrete Fourier Transform as in [12], Linear Prediction Coefficients as in [1] or Mel Frequency Cepstral Coefficients [2, 7, 6]. 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) [13] is used. The use of this model for lipreading has previously been studied by Mathews et al. [14]. In this work the implementation by Mikkel B. Stegman [15] is used. For the extraction a suitable subset of images in the training set is selected and annotated with points according to the MPEG-4 facial animation standard (Fig. 1(a)). 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,

[1]Image from www.research.att.com/projects/AnimatedHead



(a) Facial feature points [1].



(b) Image with automatically extracted feature points

Fig. 1. The facial feature points used are selected from the MPEG-4 standard, points from main groups 2,3,4,8,10 and 11 are used.

at each point the 14 parameters are picked up. In Fig. 1(b) the result of the tracking is shown for a single representative image.

3. MODEL

Unlike most other approaches the mapping in this work is performed by a continuous state space model and not a Hidden Markov Model or a Neural Network. The reasoning behind this choice is that it should be possible to change the parameters controlling the face continuously (unlike in HMM) and yet make certain that all transitions happen smoothly (unlike NN's). Currently an experimental comparison of the performance of HMM's and the continuous state space models is investigated.

In this work the system is assumed to be linear and Gaussian and hence the Kalman Filter can be used [16]. This assumption is most likely not correct and other models like particle filtering and Markov Chain Monte Carlo are considered. However, as it will be shown below, even with the simplification the model produces useful results.

The model is set up as follows:

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

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

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

In this setting $\boldsymbol{z}_k$ is the image features at time k, $\boldsymbol{y}_k$ is the sound features and $\boldsymbol{x}_k$ is a hidden variable without physical meaning, but it 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.

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

By using the EM algorithm [17, 18] on the training data, all parameters $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Sigma}^x, \boldsymbol{\Sigma}^y, \boldsymbol{\Sigma}^z\}$ 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, $\boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k$. If the intermediate smoothing variables are available the variance on $\boldsymbol{y}_k$ can also be calculated.

4. RESULTS

The data used is taken from the vidtimit database [19]. 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.

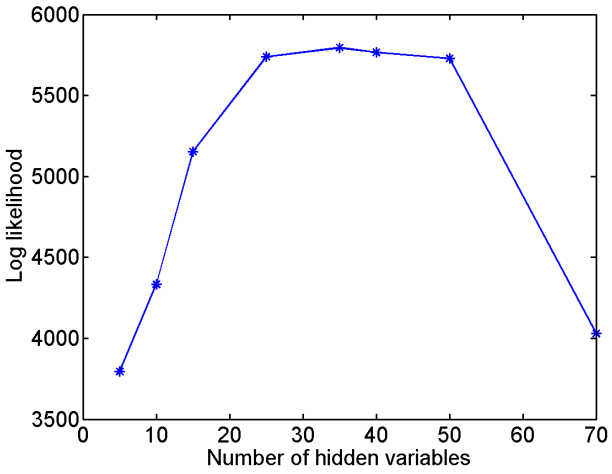
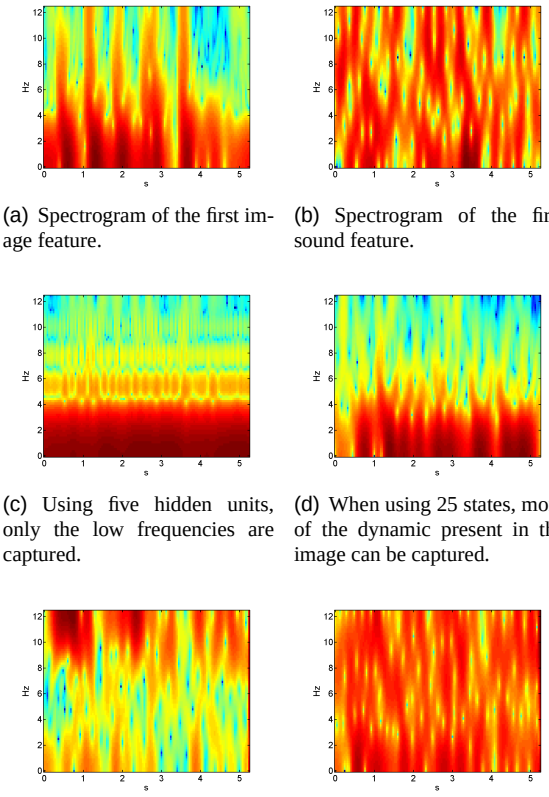


Fig. 2. The likelihood evaluated on the test data. With few hidden variables (dimesions in $\boldsymbol{x}$ space) the model is not rich enough. With too many parameters overfitting is experienced. An optimum is found in the range 25-40 hidden variables.

To find the dimension of the hidden state ($\boldsymbol{x}$), the optimal parameters ($\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Sigma}\}$) where found for varying dimensions. For each model the likelihood on the test sequence was calculated, the result is shown in Fig. 2.

With few dimensions the model is not rich enough to capture the dynamics of the image sequence. This is illustrated by the spectrogram of a hidden variable which represent the dynamics of the hidden space, as shown in Fig. 3(c). It is noted that only low frequency components are present. As the hidden space gets larger it becomes possible to model more of the dynamics present in the image. The spectrogram of a representative hidden variable when using a 25 dimensional hidden space (Fig. 3(d)) has a structure very similar to what is found in one of the image features (Fig. 3(a)). When increasing the hidden units to 70, the model degrees of freedom becomes large and over fitting becomes possible. Fig. 3(e) and Fig. 3(f) show the spectrogram of two hidden variables and it is seen that the states specializes. In 3(e) high frequencies are dominant, and the other seemingly displays a structure, which resembles the dynamics of the sound features as seen in Fig. 3(b). This is not relevant due to the slower dynamics of the facial expressions. These specializations are furthermore specific to the training set and do not generalize according to Fig. 2. It should be noted that training a large model is difficult,



(a) Spectrogram of the first image feature.

(b) Spectrogram of the first sound feature.

(c) Using five hidden units, only the low frequencies are captured.

(d) When using 25 states, more of the dynamic present in the image can be captured.

(e) Using 70 units ...

(f) ... the states specializes.

Fig. 3. In the spectrograms of one of the predicted hidden states of on the test sequence, the effect of varying the size of the state space can be seen. Spectrograms of the first sound and image features are provided for comparison.

both in terms of computations and convergence. With this analysis in mind a model with 25 hidden units is selected.

The test likelihood provides a measure of the quality of the model in feature space and provides a way of comparing models. This also allows comparison between this model and a similar Hidden Markov Model approach. However, it does not measure the quality of the final image sequence. 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 and it has not been possible to perform them in this study.

In Fig. 4 snapshots from the 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.

5. CONCLUSION

A speech to face mapping system relying on continuous 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.

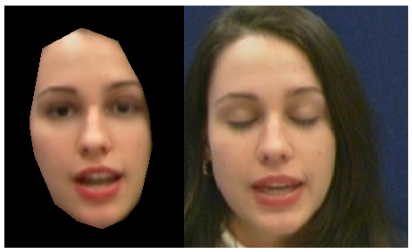
Surprisingly little attention has previously been paid to the training of state space models. In this paper it is shown that the Kalman filter is able overfit when the number of parameters are too large, similar effects are expected for the Hidden Markov Model.

All though preliminary, the results are promising. Future experiments will show how the Kalman model and other instances of continuous state space models compares to Hidden Markov Model type systems.

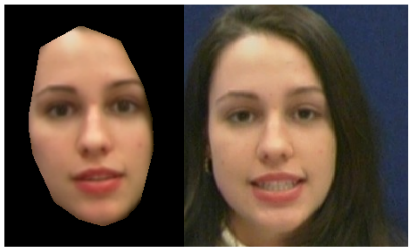
6. REFERENCES

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

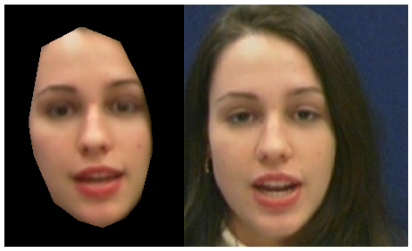
[2] William Goldenthal, Keith Waters, Thong Van Jean-Manuel, and Oren Glickman, "Driving synthetic mouth gestures: Phonetic recognition for faceme!,"



(a)



(b)



(c)

Fig. 4. Characteristic images taken from the test sequence. The predicted face is to the left and the true face to the right.

in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 1995–1998.

[3] 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.

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

[5] Jay J. Williams and Aggelos K. Katsaggelos, "An hmm-based speech-to-video synthesizer," 1998.

[6] Pengyu Hong, Zhen Wen, and Thomas S. Huang, "Speech driven face animation," in *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, Igor S. Pandzic and Robert Forchheimer, Eds. Wiley, Europe, July 2002.

[7] Dominic W. Massaro, Jonas Beskow, Michael M. Cohen, Christopher L. Fry, and Tony Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," *Proc. AVSP 99*, 1999.

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

[9] Lavagetto 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] David F. McAllister, Robert D. Rodman, Donald L. Bitzer, and Andrew 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] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *Proc. European Conference on Computer Vision*, vol. 2, pp. 484–498, 1998.

[14] I. Matthews, T.F. Cootes, J.A. 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.

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

[16] Rudolph Emil Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, 1960.

[17] 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.

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

[19] 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.