Comparison of HMM and TMDN Methods for Lip Synchronisation

Gregor Hofer and Korin Richmond

Centre for Speech Technology Research, Edinburgh University, United Kingdom
ghofer@inf.ed.ac.uk, korin@inf.ed.ac.uk

Abstract

This paper presents a comparison between a hidden Markov model (HMM) based method and a novel artificial neural network (ANN) based method for lip synchronisation. Both model types were trained on motion tracking data and a perceptual evaluation was carried out comparing the output of the models, both to each other and to the original tracked data. It was found that the ANN based method was judged significantly better than the HMM based method. Furthermore the original data was not judged significantly better than the output of the ANN method.

Index Terms: hidden Markov model, mixture density network, lip synchronisation, inversion mapping

1. Introduction

Talking computer animated characters are now commonplace in video games and films. Additionally, there has been an increasing amount of research in the field of interactive virtual agents. For all of these applications, the synchronisation of the character's face to their speech is essential to make the immersion or the interaction believable. Although performing mouth animation by hand gives the best results, as evident in the quality of animation in today's computer animated films, it is not always feasible because of cost or time constraints. Therefore, producing lip animation automatically is highly desirable. This paper addresses this problem, which may be summarised as mapping from speech to lip animation. In other words, given speech input we desire a method to output lip animation automatically.

Previous approaches have utilised dominance functions [1] which are linear combinations of trajectories selected according to a phonetic transcription of the speech signal. Although this method produces acceptable results, it is very difficult to tune to new speakers. More recent methods are based on machine learning techniques that automatically learn a mapping from speech to some representation of the animation. This can also be called acoustic-to-articulatory *inversion* because the configuration of the mouth is inferred from an acoustic signal. The benefit of using trainable models is that they can easily learn an inversion mapping for different speakers, as long as training data is available. The two main machine learning approaches used in this area are artificial neural networks (ANNs) [2] and statistical techniques like Hidden Markov Models (HMMs) [3, 4]. In theory, the HMM approach has certain advantages arising from its inherent construction and operation in terms of phones. The expected phone string for an utterance can easily be used in the HMM-based system (i.e. Viterbi alignment) as well as the input acoustic signal. This provides a significant extra source of information when generating lip movements phone-by-phone for a novel utterance. In addition, working at the phone level can make it straightforward to combine an HMM-based lip synchronisation system with speech synthesis, as the synthesiser can provide a phone string. Conversely, a possible advantage of ANN-based approaches is that they typically work on a frame-wise basis, mapping one input acoustic feature vector directly to one output lip configuration. Theoretically, this means an ANN-based approach can offer closer, more direct synchronisation with the acoustic signal than the HMM, in which the mapping is mediated through longer phone-size units. This paper aims to evaluate these advantages by presenting a direct experimental comparison between an HMM-based method and a novel ANN-based method.

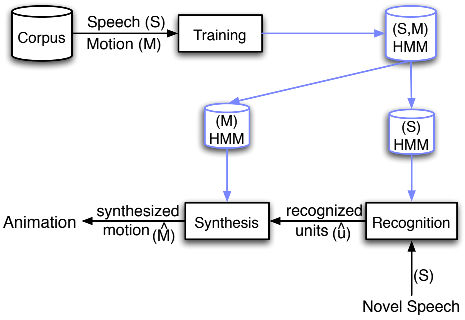


Figure 1: HMMs are trained on speech (S) and lip motion (M) data simultaneously. For synthesis the motion models that correspond to the predicted speech models are used to generate the output trajectories.

2. HMM-based Inversion Method

Multi-stream HMMs are trained on speech and lip motion data simultaneously. To synthesise lip motion, recognition is first performed using only the acoustic feature streams of the multi-stream HMMs, and a sequence of lip motion units is derived. These units yield the sequence of context-dependent models that are used for synthesising the lip motion trajectories. An overview of the training and synthesis process for the multi-stream HMMs is shown in Fig. 1.

2.1. Optimal motion

Theoretically, the above procedures can be justified as follows: A motion sequence $\boldsymbol{O}_L = (\boldsymbol{o}_{L_1}, \boldsymbol{o}_{L_2}, \ldots, \boldsymbol{o}_{L_T})$ is generated from a given speech vector sequence $\boldsymbol{O}_S = (\boldsymbol{o}_{S_1}, \boldsymbol{o}_{S_2}, \ldots, \boldsymbol{o}_{S_T})$ with a length of T frames by solving the following optimisation:

$$\boldsymbol{O}_L^* = \underset{\boldsymbol{O}_L}{\operatorname{argmax}}\, P(\boldsymbol{O}_L|\boldsymbol{O}_S) \tag{1}$$

By incorporating the motion-unit sequence $\boldsymbol{u}_L = (u_{L_1}, ..., u_{L_e})$, which represents the lip movements corresponding to the given speech sequence, it can be approximated

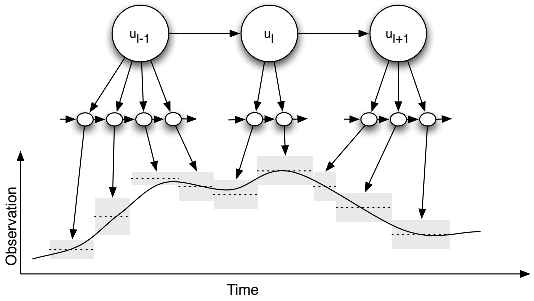


Figure 2: A sample trajectory generated from trajectory HMMs.

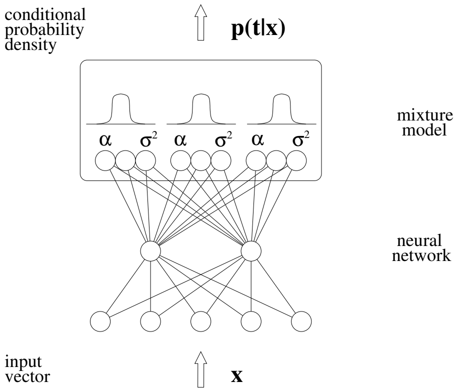


Figure 3: The mixture density network we use combines a multilayer perceptron and Gaussian mixture model.

by

$$\boldsymbol{O}_L^* = \operatorname*{argmax}_{\boldsymbol{O}_L} P(\boldsymbol{O}_L|\boldsymbol{O}_S) \tag{2}$$

$$= \operatorname*{argmax}_{\boldsymbol{O}_L} \sum_{\boldsymbol{u}_L} P(\boldsymbol{O}_L|\boldsymbol{u}_L, \boldsymbol{O}_S)P(\boldsymbol{O}_S|\boldsymbol{u}_L)P(\boldsymbol{u}_L) \tag{3}$$

$$\simeq \operatorname*{argmax}_{\boldsymbol{O}_L} P(\boldsymbol{O}_L|\boldsymbol{u}_L^*) \tag{4}$$

where

$$\boldsymbol{u}_L^* = \operatorname*{argmax}_{\boldsymbol{u}_L} P(\boldsymbol{O}_S|\boldsymbol{u}_L)P(\boldsymbol{u}_L) \tag{5}$$

Practically, we recognise the lip motion units $\boldsymbol{u}_L$ from the given speech data $\boldsymbol{O}_S$ using the Viterbi algorithm and then generate a lip motion sequence from HMMs corresponding to the recognised units. For the probability $P(\boldsymbol{u}_L)$, we use back-off bi-gram models estimated from the training database.

Trajectories are synthesised from the predicted units using the maximum likelihood parameter generation algorithm (MLPG) as described in [3]. When single Gaussian distributions are used as the emission probabilities, we can easily solve this problem in a closed form in a maximum likelihood sense [5]. A sample of the trajectory generated from HMMs is shown in Fig. 2, in which we can see that the generated trajectory (solid line) becomes smooth. When mixtures of Gaussian distributions are used as the emission probabilities, the trajectory is optimised via the EM algorithm to select the optimal Gaussian distributions.

3. MDN-based Inversion Method

In previous work, we have worked extensively on applying ANNs to the inversion mapping, and have demonstrated the superiority of one type of model in particular, the mixture density network, over other more common variants such as the multilayer perceptron (MLP) [6, 7]. Hence, we continue to use this model, and do not evaluate alternatives such as the MLP here.

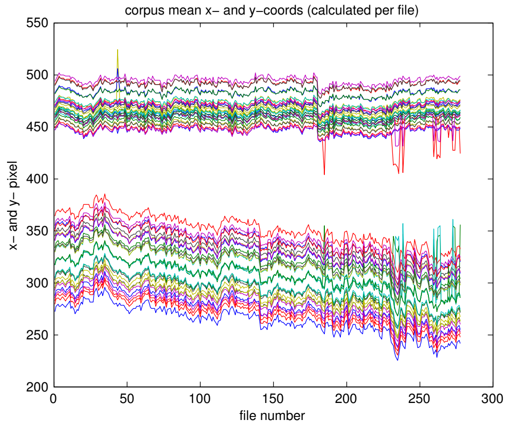


Figure 4: Average x- and y-coordinates (bottom and top group respectively) for tracked points calculated on a file-by-file basis.

Due to space constraints, we can only give a high level introduction here. For a more explicit description, the reader is referred to [7]. At the heart of our inversion mapping model is the mixture density network (MDN). In the most general sense, the MDN combines a trainable regression function (typically a non-linear regressor such as an ANN) with a probability density function. In our work, we have been used an MLP and a Gaussian mixture model (GMM) for these purposes respectively, as illustrated in Fig. 3. The role of the MLP is to take an input vector in one domain ($\mathbf{x}$, acoustic features in this case) and map to the control parameters (priors, means and variances) of the pdf over the domain of the target parameters ($\mathbf{t}$, the lip parameters). In this way, the MDN offers a model of probability density over the target domain conditioned on the input domain, $p(\mathbf{t}|\mathbf{x})$. Training consists of updating the MLP weights to minimise the negative log likelihood of the target data.

To achieve a trajectory model, we can augment the target features with derived velocities and accelerations, and use the MDN to provide conditional pdfs over these. Hence, once trained, we can input the sequence of acoustic feature vectors for an utterance and get as output a sequence of pdfs over the static motion features and their delta and deltadeltas. We may then apply a maximum likelihood parameter generation algorithm (MLPG)[5] to this sequence of pdfs in order to obtain a single, most probable trajectory which optimises the constraints between the distributions of static features and their velocities and accelerations.

4. Comparison Experiment

4.1. Data selection and processing

For this evaluation we decided to use the data provided as part of the LIPS Challenge 2008 [8], since this matched our requirements and is likely to be familiar to other researchers as a standardised task. This dataset provides video of the face (50Hz frame rate) and audio for a single female subject reading 278 phonetically balanced sentences. The overall aim of data processing was to take the video and audio data from the LIPS Challenge and derive a small number of parameters to describe the movements of the subject's mouth in synchrony with the

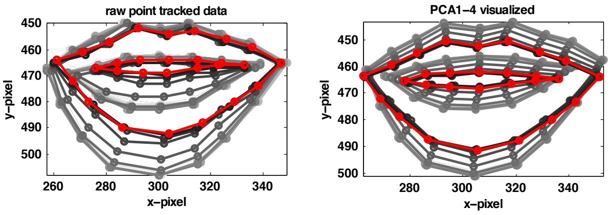
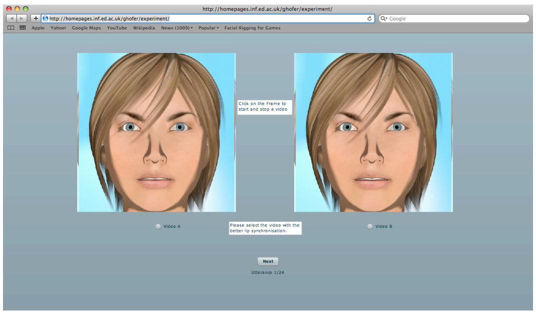


Figure 5: Comparison of raw point-tracking data (left) with visualisation of first 4 principle component weights (right) for the word "How". The red contours indicate the final frame, with frames going back in time represented with progressively lighter grayscale contours.

acoustic speech signal. As a first step, 28 points around the inner and outer contours of the lips were tracked in each video frame. This was done using linear predictors, as detailed in [3]. The next step was to correct for head movement which is present throughout the recorded utterances. To achieve this, we used principle components analysis (PCA). Through visualising the eigenvectors, it was observed that the first principle component weight corresponded entirely to horizontal movement of the mouth (i.e. head movement in the x-direction). It was therefore trivial to use this component weight to remove the global x-offset from the tracked points in each frame. Meanwhile, it was also observed that the second PCA weight largely corresponded to vertical displacement of the mouth, but with a small degree of associated mouth opening. To remove only the vertical displacement, we used the points at the mouth corners reconstructed from the second PCA weight for each frame to identify the displacement to remove from all points in that frame.

To verify data integrity, we calculated the mean x- and y-coordinate for each tracked point in each utterance, as shown in Fig. 4. We identified 17 utterances with gross tracking errors, and discarded those files. We also noted local trends in the mean positions of the tracked points, resulting from minor changes in the subject's pose over time. We used the adaptive mean normalisation technique described in [9] to minimise these effects.

Next, we performed PCA again, and visualised the new eigenvectors, both in isolation and in various combinations, to empirically identify a suitable small set of PCA weights with which to capture mouth shapes. Our aim was to find a combination that resulted in plausible and smoothly varying reconstructed mouth movements, with as little as possible of the noise present in the raw point-tracking data. For example, the first PCA weight was found to represent mouth opening and closing, while the second corresponded to lip spreading and puckering, and so on. In fact, we found using just the first four PCA weights resulted in reconstructed mouth shapes which were relatively noise free, yet subjectively plausible and intelligible for lip-reading. Fig. 5 demonstrates the effects of preprocessing. The raw data is far noisier and less consistent than the movements represented by the 4 PCA weights. Finally, these PCA weight trajectories were lowpass filtered using a 2nd order Butterworth filter with cutoff at 12.5Hz, upsampled to a frame rate of 200Hz, and normalised by subtracting their mean and dividing by 4 times their standard deviation. Meanwhile, the acoustic signal was parameterised with 25 mel cepstrum coefficients at a framerate of 200Hz to match the 4 PCA weights. A test and validation set were selected, each containing 26 utterances drawn evenly from throughout the corpus, while 209 utterances were used for training.


Click on the Frame to start and stop a video

Please select the video with the better lip synchronisation.

Utterance 1/24

Figure 6: Presentation of stimuli in the perceptual evaluation.

ANN HMM	ANN ORG	HMM ORG
17 %	50 %	8 %

Table 1: How often participants changed their mind for each condition when viewing the stimuli in reversed order.

4.2. Evaluation

We trained a range of MDN networks, with all combinations of hidden layer sizes in the set of [80, 100, 200] units, GMMs with [1, 2, 4] mixture components and acoustic input context window sizes of [1, 5, 10] frames. The scaled conjugate gradients algorithm was used for optimisation, using the validation set with an early stopping criterion to avoid overfitting. Comparison of validation set results showed that the MDN with 200 hidden units, an input context window of 10 frames and just a single Gaussian mixture component performed best. The fact that a single Gaussian performed marginally better than 2 or 4 was surprising, and conflicts with previous work on the inversion mapping using this model [7]. This may be due to the type of data used here, which may not be as consistent as the data provided by electromagnetic articulography that we have used before. This point needs further investigation. The configuration for the HMM was the same as the optimal system described in [3], with 5 states and 4 mixture components per state. Both model types were trained with exactly same features with the same training and test sets.

To generate animation the first two PCA components were mapped to deformation parameters of our talking head corresponding to lip opening and pucker. The presented lip synchronisation methods were evaluated perceptually by conducting a pairwise comparison experiment. A website was created to present the stimuli to the participants. An example page is shown in Fig. 6. For each stimulus the participant was presented with two videos that could be watched as many times as desired, although always in full. Participants were then requested to identify the video with better lip synchronisation. Four utterances were each synthesised in three different versions: using the ANN output (*ANN*); the HMM output (*HMM*); and the original tracking data (*ORG*). Three conditions were tested, consisting of the comparison between any two versions, where all utterances in each condition were presented in both orders. Thus participants were shown 24 stimuli in total. There were 17 participants: 11 male, 6 female; 8 native and 9 non-native speakers.

4.3. Results

Fig. 7 shows the raw scores for each condition averaged across all participants, as there was no significant difference between native and non-native speakers. From these results, we observe

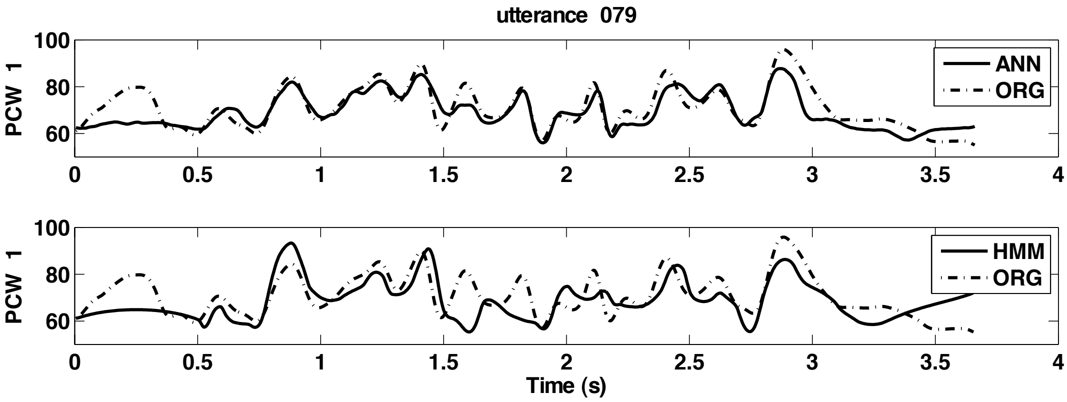


Figure 8: Comparison of trajectories for principle component weight 1 generated by HMM and ANN.

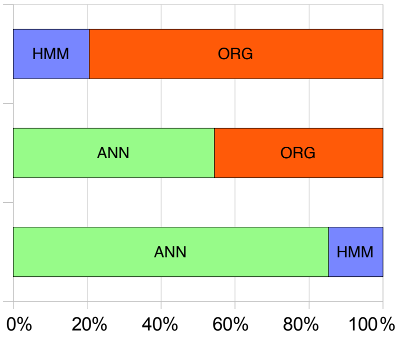


Figure 7: Preference scores for the 3 conditions.

that both the original lip motion and the lip motion generated by the ANN were preferred when compared to that generated by the HMM-based method. A two-tailed binomial test indicates both these preferences are statistically significant (p¡0.001). Meanwhile, there is a slight preference for the lip motion generated by the ANN compared to the original lip motion, but this difference is not statistically significant. Table 1 shows the consistency for each pairwise testing condition. We see that subjects were strongly consistent in their preference of ORG over HMM, and reasonably strongly consistent in their preference of ANN over HMM. However, the subjects were not so consistent in their selection when choosing between ANN and ORG. This indicates subjects had more difficulty in distinguishing between ANN and ORG, and supports the preference scores in Fig. 7.

Overall, these results support the view that the perceptual test subjects preferred the closer, frame-level synchronisation of lip movements offered by the ANN method. Finally, example output from the ANN and HMM for the 1st PCA weight throughout utterance `079` is shown in Fig. 8. Again, this supports the view that the ANN can offer closer synchronisation to the target movements (and thus the associated acoustic signal).

5. Conclusions

This paper has compared an ANN-based approach with an HMM-based approach for lip synchronisation. To the best of our knowledge our trajectory MDN method has not been used for lip synchronisation before, and based on our results here they seem very promising compared to HMMs. Whereas the original tracked data was clearly rated better than the HMM output, the MDN-based approach performed on a par with the original data. We interpret this indeed reflects a preference for the close, frame-level synchronisation offered by ANNs compared to the longer unit-based synchronisation of HMMs. In future work, we aim to apply our MDN-based method to a large set of high quality data.

6. Acknowledgements

Gregor Hofer is supported by EU FP7 SSPNET. Korin Richmond is supported by Engineering and Physical Sciences Research Council grant EP/E01609x/1.

7. References

[1] M. Cohen and D. Massaro, “Modeling coarticulation in synthetic visual speech,” *Models and Techniques in Computer Animation*, Jan 1993.

[2] J. Beskow and M. Nordenberg, “Data-driven synthesis of expressive visual speech using an MPEG-4 talking head,” *European Conference on Speech Communication and Technology*, Jan 2005.

[3] G. Hofer, J. Yamagishi, and H. Shimodaira, “Speech-driven lip motion generation with a trajectory HMM,” in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 2314–2317.

[4] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, “Text-to-visual speech synthesis based on parameter generation from HMM,” in *Proceedings of ICASSP 98*, 1998, pp. 3745–3748.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[6] K. Richmond, “Preliminary inversion mapping results with a new EMA corpus,” in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 2835–2838.

[7] ——, “Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion,” in *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007*, ser. Lecture Notes in Computer Science, M. Chetouani, A. Hussain, B. Gas, M. Milgram, and J.-L. Zarader, Eds., vol. 4885. Springer-Verlag Berlin Heidelberg, Dec. 2007, pp. 263–272.

[8] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, “Lips2008: Visual speech synthesis challenge,” in *Proc. Interspeech*, Brisbane, Australia, September 2008, pp. 2310–2313.

[9] K. Richmond, “Estimating articulatory parameters from the acoustic speech signal,” Ph.D. dissertation, The Centre for Speech Technology Research, Edinburgh University, 2002.