Automatic Head Motion Prediction from Speech Data

Gregor Hofer and Hiroshi Shimodaira

Centre for Speech Technology Research
University of Edinburgh, UK
g.hofer@sms.ed.ac.uk, h.shimodaira@ed.ac.uk

Abstract

In this paper we present a novel approach to generate a sequence of head motion units given some speech. The modelling approach is based on the notion that head motion can be divided into a number of short homogeneous units that can be modelled individually. The system is based on Hidden Markov Models (HMM), which are trained on motion units and act as a sequence generator. They can be evaluated by an accuracy measure. A database of motion capture data was collected and manually annotated for head motion and is used to train the models. It was found that the model is good at distinguishing high activity regions from regions with less activity with accuracies around 75 percent. Furthermore the model is able to distinguish different head motion patterns based on speech features somewhat reliably, with accuracies reaching almost 70 percent.

Index Terms: audio-visual speech, multimodal, talking head

1. Introduction

The prevalence of computer animation in games and movies has prompted an interest in generating human-like behaviour in animated characters that is synchronised with speech. The conventional approach requires the collection of large motion capture databases where an actor performs all the different actions that are required for the animated character. Collecting such databases is expensive and they can only be employed in the exact scenarios they were collected for. To overcome the shortcomings of motion capture databases, methods to automatically generate human like behaviour have to be developed. In particular when animating faces speech needs to be used to drive the animation to ensure proper synchronisation between the facial movements and the audio.

So far speech has only been used extensively to drive lip animation, where phoneme sequences extracted from the speech signal are mapped to a viseme (visual counterparts of phonemes) sequence[1]. Speech has also been used to drive general facial animation [2]. Very few attempts exist to drive head motion from speech. Munhall et al.[3] suggested that head motion is important in speech perception and therefore could enhance our perception of animated characters. Notably Busso et al.[4] reported one of the first systems that used speech for head motion synthesis. The system of Busso et al. was based on a framewise relationship between speech and head motion which makes the modelling more straightforward but might fail to capture more long range relationships between speech and motion.

There have been several studies by Hadar et al.[5] investigating the relationship of head motion in the speech production process. It was found that the head moves almost constantly during speech and motoric functions are attributed to that movement. Although the researchers are mostly concerned with motoric functions of movements, it is realised that head motion is fundamentally influenced by two processes: Functional movement during conversations like nodding for agreement and motoric movement that is tied to the speech production. Following Hadar and colleagues this paper is only concerned with the latter and although it is not clear how to separate the two processes, it is hoped that by using statistics over enough data the former influence can be randomised as it bears no direct relationship with speech.

The link of head motion to the speech production process suggests that in order to drive head motion with speech data the temporal relationship between the two streams has to be taken into account. Since frame wise analysis of the data streams is not sufficient to model temporal relationships, the data has to be segmented into longer parts. In this paper, head motion is modelled by introducing a conceptual unit of motion that is based on manual labels that spawn over several frames. Since the link between the two streams of speech and head motion is not straightforward a modelling layer is introduced where speech and motion features are used together to train models that represent units of motion. It is hoped that the temporal relationship and the long range dependencies between the two streams can be better captured by this approach. Our approach is based on HMM's that act as a sequence generator and can be evaluated by an accuracy measure similar to word error rate used in speech recognition.

2. Data

2.1. Data collection and processing

An audiovisual database was recorded for this project from one actor with 7 markers on his face and body. He was asked to tell several fairy tales. A Qualisys Moiton Capture System was employed to capture the head and body motion of the actor at a sampling frequency of 500 Hz. His voice was recorded using a close talking microphone at a sampling frequency of 44 kHz. The motion capture and audio recording was synchronised automatically by the system. In total 25min of pure speech were recorded.

To calculate the head motion, the marker positions were used to estimate 2 rigid bodies, the head and the upper body. For each frame the rotation in Euler angles of the local coordinate system of each rigid body was calculated. The Euler angles of the upper body were subtracted from the Euler angles of the head to calculate the pure head motion angles. The minimum and maximum Euler angles for each dimension can be seen in Table 1. The acoustic signals were processed using the ESPS algorithms provided by the Snack toolkit that yield pitch and

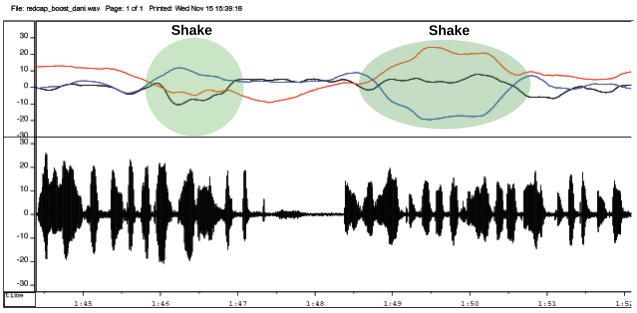

Figure 1: Example of two shakes as indicated by the marked regions

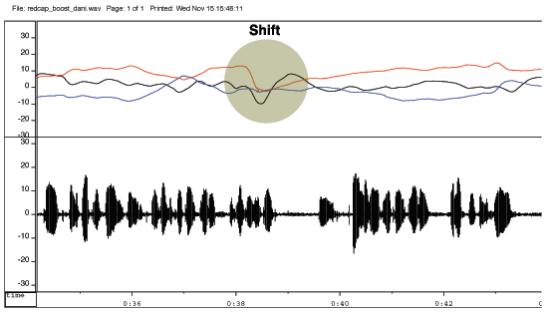

Figure 2: Example of a shift as indicated the marked region

RMS energy. The mean pitch for each utterance was subtracted from the voiced regions to normalise the pitch prediction. Furthermore the HTK toolkit was used to calculate the MFCC coefficients for the audio signal. The final speech feature vector had 42 dimensions and the motion feature vector had 9 dimensions.

Table 1: *Minimum and maximum Euler angles in the data.*

	pitch	yaw	roll
maximum	31	23	46
minimum	-21	-22	-29
SD.	8.71	7.71	9.4

2.2. Data labelling

To be able to better model the relationship between speech and head motion, the data was manually labelled to describe segments of distinct motion. The Euler angles were graphed and inspected visually. Four labels were applied:

- postural shift: the head shifts axis of movement
- shake and nod: lateral movement around one axis
- pause: no movement / rest position
- default: non-distinctive movement / slow movement

Labelling head motions is not straightforward as for example gait, where motions can be labelled as running, walking, dancing, etc. Head motion does not yield any clear distinctions between different movements, therefore we decided to use the most basic motions that can be seen in the data.

Figure 1 and Figure 2 show typical shake and shift motions in Euler angle representation. If the movement was not distinct a default label was applied. Table 2 shows the distribution of labels in the data and their average length.

Table 2: *Distribution of labels and their average length*

	shift	shake	pause	default
number	209	107	47	209
length(sec)	0.4	0.8	0.8	2.5

3. Correlation Analysis

Various researchers have suggested a close relationship between speech and motion, even frame-wise correlations were suggested [6]. Yehia et al. found correlations between F0 and Head motion within utterances but could not find any globally. We calculated frame wise correlations between our feature vector for all utterances and could not find any substantial correlations within utterances or globally. Figure 3 shows a matrix plot of the correlations between F0, RMS energy, and Euler angles, with their respective derivatives. The plot shows very clearly that there is no strong correlation among the different features contrary to some other findings.

To further investigate the absence of correlations, we used a more sophisticated technique to calculate correlations. Canonical Correlation Analysis (CCA) makes it possible to calculate correlations between vectors of different dimensions. We spilt the features into speech and motion features, resulting in a speech vector consisting of the first 12 MFCC coefficients, pitch, energy, and their respective first and second derivatives. To compare our analysis with the the analysis done by Busso et al.[4] we also performed CCA between the Euler angles (3D vector) and energy, pitch and their first and second derivatives (6D vector). The correlations found by our analysis were much lower than the correlations reported by Busso et al. Table 3 shows the correlation results between the features.

Table 3: *Frame-wise Canonical Correlation Analysis between Speech and Motion Features*

MFCC,E , F0	E, F0
0.08	0.07

The correlation analysis results indicate that it is not straightforward to model the relationship between speech and head motion as no apparent correlations between the 2 feature spaces exists. To model the speech and head motion the temporal properties of the two signals will have be taken into account.

4. Modelling Head Motion

The modelling approach is based the notion that head motion can be divided into a number of short homogeneous units that can each be modelled individually. For example all the data labelled as shift is modelled by one model and the data labelled as shake is modelled by another model. To model each of the units Hidden Markov Models (HMMs) are employed because they can model sequential data. When sequences are generated, the speech data is used to predict the motion label. For each input sequence of speech frames, a sequence of motion labels is produced that are chosen by the most likely sequence of models.

To model the relationship between speech and head motion our specific model consists of two streams. Stream 1 was

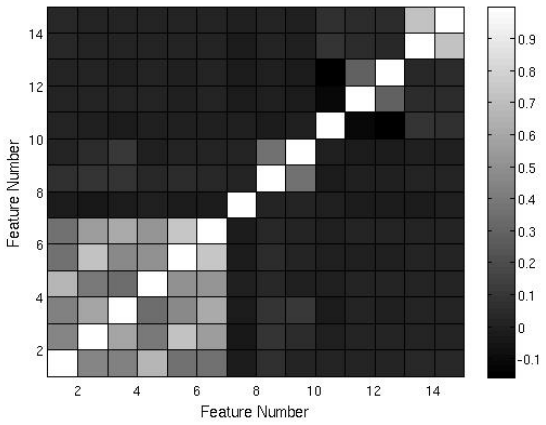

Figure 3: Correlations between Euler Angles and prosodic features. Feature 1-9 is the Euler Angels and its derivatives. Feature 10-12 is RMS energy and its first and second derivative. Feature 13-15 is the pitch and its derivatives.

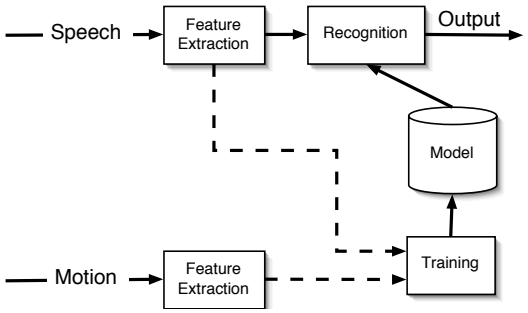

Figure 4: The models are trained on motion and speech features. Only speech features are used during recognition.

trained on the speech features and stream 2 was trained on the motion features. The transition probabilities for both models were shared and trained on the combined data. Figure 4 shows the training process and the recognition process. The standard left-to-right HMM used in speech recognition was chosen and one HMM is trained for each label from the labelled training data.

During recognition only the speech features were used to determine the sequence of motion labels. The motion stream was turned off and only the speech stream was used to recognise the most likely sequence of head motion labels. It is important to state that the transition probabilities were still the same and only the motion features were ignored. A pilot experiment compared the recognition accuracies of a model trained on both motion and speech and a model trained on speech. It was found that the model trained on both streams performed better (Acc=75) than the model trained only on speech (Acc=65). By training on both streams the transition probabilities can take both streams implicitly into account during recognition. This allowed us to produce a sequence of motions given some speech. Employing a framework that is similar to speech recognition allowed us to evaluate different aspects of the modelling process in a more principled way. A measure similar to word error rate was used to evaluate the models.

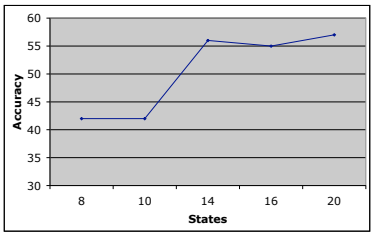

Figure 5: Results for different number of states per model. The number of mixtures in all models was 4.

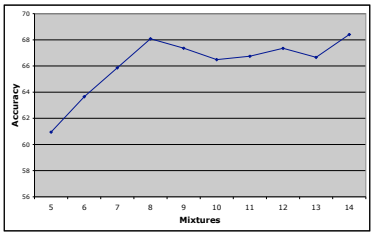

Figure 6: Results for different number of mixtures for the default model. The number of states in all models was 16.

5. Evaluation of Modelling

A number of experiments were conducted to determine the optimal modelling parameters. The accuracy was calculated like word error rate:

$$Acc = \frac{Correct - Insertions}{Total\ number\ of\ labels} \times 100$$

5.1. Manual labels

5.1.1. Model parameters

We conducted experiments with the models based on the manual labels to find the optimal model parameters. Since the motion and speech operate at different frame rates it was not clear what the optimal number of states should be. The actual test results were obtained by 7 fold cross validation. The results shown are the average results of this cross validation.

To determine the optimal model length the number of states was gradually increased. In the following experiments a reduced feature set without F0 and the motion delta features was used. From the results shown in Figure 5 it seems clear that around 16 states per model are required for adequate recognition performance. Furthermore the models had to be tuned to take the default class into account. We increased the discriminatory power by increasing the number of mixtures per state in the default model. The results shown in Figure 6 suggest that significantly more mixtures are needed for the default model than for the other models.

5.1.2. Features

The final model topology had 18 states per model and 4 mixtures per state in all models except the default model which had 8 mixtures. In addition the influence of F0 on the model was

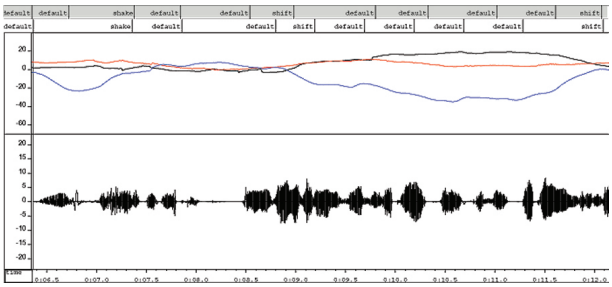

Figure 7: Predicted and actual labels for a short segment of speech. The upper labels are the predicted ones.

tested as F0 was attributed a great significance in the relationship between head motion and speech. Table 4 shows a comparison between the best models with F0 and without F0. Finally a model that just used F0, energy and their first and second derivative was constructed as well. The results are also shown in Table 4.

Table 4: *Results for models trained on different speech feature sets on 2 classes and 4 classes. The first and second derivative of each feature was also used.*

		MFCC + E	MFCC + E + F0	E + F0
4 class	Acc	68%	69%	50%
	SD.	2.97	6.32	11.88
	Max.	71%	75%	70%
	Min.	64%	61%	39%
2 class	Acc	74%	76%	73%
	SD.	2.71	4.34	3.24

The discriminatory power of the model between regions of high activity (shake and shift) and low activity (default and pause) was tested as well, termed 2 class. The results are shown in table 4 and suggest that the model can distinguish reasonably well between default/pause segments and other regions. Figure 7 shows a predicted label sequence in comparison to the actual labels. It shows quite well that the boundaries detected by the model are accurate most of the time.

5.2. LBG cluster labels

To build a baseline model LBG clustering was used to divide the 3D space of Euler angles into K clusters. This was done framewise at a framerate of 500Hz. To compare the results with the manual labels K=4 clusters were used. If the LBG algorithm clustered more than 5 consecutive frames with the same cluster index, these frames were treated as a sequence. The minimum length of a sequence was therefore 10ms. Each sequence was labelled with its corresponding the cluster index. The clusters indices were treated like labels and the training data, consisting of speech and motion was marked accordingly. One HMM was trained per cluster index on the marked sequences and recognition experiments were performed.

The model configuration for each label was 18 states with 4 mixtures per state. This configuration was determined experimentally. Table 5 shows how model a model trained on LBG labels compares to a model trained on the manual labels using the same feature set.

Table 5: *Comparison of recognition accuracy between a model trained on LBG labels and a model trained on manual labels.*

	LBG	Manual
Accuracy	52%	69%

6. Conclusion

A novel approach for predicting motion labels from speech data using long range dependencies modeled by Hidden Markov Models has been described in this paper. Long range dependencies are used because it has been show that frame wise correlations between speech and motion features are very difficult to find. Although previous work has shown these kind of correlations, the results could not be replicated in this paper. Of course we calculated the correlations on our own data, which is very different from the data used in other studies. The actor we recorded was speaking relatively freely and did not read predefined sentences which could account for the vastly different correlation results found. Our models were trained on data manually annotated for head motion and they were able to predict motion labels with accuracies reaching 70%. It was found that F0 helps in distinguishing different types of motion. When the model was tested for how well it could distinguish regions of high activity and regions of low activity, a model trained only on F0 and energy was able to perform almost on par with models trained on the full feature set. Furthermore it has been shown that our system outperformed a baseline based on LBG clustering labels.

To improve the model, longer range features need to be included in the future. Since the rate of change of speech and motion are very different, it will be challenging to come up with compelling features that improve the accuracy. Finally the labels will be used to synthesise head motion trajectories used in a talking head.

7. References

[1] Chang, Y. and Ezzat, T., Transferable Videorealistic Speech Animation, ACM Siggraph/Eurographics Symposium on Computer Animation, Los Angeles, CA, 2005.

[2] Graf, H.P., Cosatto, E., Strom, V., and Huang, F.C. Visual Prosody: Facial Movements Accompanying Speech. Conference on Automatic Face and Gesture Recognition. IEEE Computer Society, Washington, DC, 2002.

[3] Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., and Vatikiotis-Bateson, E. Visual Prosody and Speech Intelligibility. Head Movement Improves Auditory Speech Perception Psychological Science 15. 2004.

[4] Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. In Press IEEE Transactions on Audio, Speech and Language Processing, March 2007.

[5] Hadar, U., Steiner, T.J., Grand, E.C., and Rose, F.C. Head movement correlates of juncture and stress at sentence level. Language and Speech, 26. 1983.

[6] Yehia, H.C., Kuratate, T., and Vatikiotis-Bateson, E. Linking facial animation, head motion and speech acoustics. Journal of Phonetics, 30. 2002.