Predicting head motion from prosodic and linguistic features

Angelika Hönemann[1], *Diego Evin*[2,3], *Alejandro J. Hadad*[3], *Hansjörg Mixdorff*[1], *Sascha Fagel*[4]

[1]Beuth University Berlin, Berlin, Germany,
[2]INIGEM - Universidad de Buenos Aires- CONICET, Buenos Aires, Argentina
[3]Universidad Nacional de Entre Ríos, Facultad de Ingeniería, Oro Verde, Argentina
[4]zoobe message entertainment GmbH, Berlin, Germany,

{ahoenemann|mixdorff}@beuth-hochschule.de,diegoevin@gmail.com,
hadad@santafe-concet.gov.ar,fagel@zoobe.com

Abstract

This paper describes an approach to predict non-verbal cues from speech-related features. Our previous investigations of audiovisual speech showed that there are strong correlations between the two modalities. In this work we developed two models using different kinds of Recurrent Artificial Neural Networks: Elman and NARX, to predict parameters of activity for head motion using linguistic and prosodic inputs, and compared their performance. Prosodic inputs included F0 and intensity, while linguistic parameters included the former plus additional information such as the type of syllables, phrases, and different relations between them. Using speaker specific models for six subjects, performance measures in terms of root mean square error (RMSE) showed that there are significant differences between the models with respect to the input parameters, and that NARX network outperformed the Elman network on the prediction task.

Index Terms: predicting head motion, audiovisual speech, time-delayed NARX, Elman NN, linguistic vs. prosodic features

1. Introduction

Today the use of talking heads has become more common, they can be found for example in advertising, computer games or systems for learning languages. However, the goal of developing a realistic-looking talking head is still a major challenge. Different areas of research have focused on the synthesis of visual speech, and several approaches to synthesize gestures, head or facial motion can be found. Busso & Deng used Hidden Markov Models (HMM) to synthesize head motion in which each model represents an emotional state [1]. Cosker, et.al. used HMMs to estimate lip motion directly from the acoustic waveform [2]. Massaro, et al. trained an artificial neural network to map the cepstral coefficients of natural speech to the control parameters of an animated synthetic talking head [7]. As in that case, most of the methods are based on acoustic parameters such as the MFCC. According to what we know there exist only several studies which use linguistic parameters such as the type of syllable, phrase boundaries or the position of the syllable to predict head and facial gestures.

The ultimate goal of this work is to predict head motion during speech. In an earlier related study [9] authors analyzed the kinematic–acoustic relationship between head motion and fundamental frequency (F0) for two speakers, generating a regression model, used to animate a natural-looking talking head. Hofer, et al. proposed an HMM-based system to predict head motion from speech, and found a strong influence of F0 on the prediction model [3].

In previous work we analyzed audiovisual data to find significant parameters relating acoustic and linguistic information with head movements and facial gestures [5][6]. In the current study we apply some of the previous findings to develop a system predicting head motion from speech.

The paper is structured as follows: Section 2 gives an overview of previous work on the analysis of an audiovisual speech corpus and presents major findings. Section 3 describes the proposed prediction models. Discussion and conclusion are found in section 4.

2. Previous Work

In previous work we presented the development of an audiovisual speech corpus of spontaneous speech produced by seven native speakers of German [5], and a data analysis of its contents [6]. The analysis was performed on acoustic and visual data separately before we looked at the relationship between them. For the visual channel we recorded the video and the motion of 43 facial markers captured with a *Qualisys* motion capture system.

2.1. Acoustic Analysis

The acoustic data for each speaker was segmented at the syllable level and the accented syllable, as well as phrases, and phrase boundaries were marked. We calculated the duration means and standard deviations for all types of syllables and phrases. The syllables were subdivided into different classes depending on their position and degree of prominence. We also extracted the F0 and intensity contours from the speech signal and segmented the utterances into so-called accented and unaccented sequences. Accented sequences contain accented syllables and their left and right neighbors, unaccented sequences the remaining syllables. The minimum, maximum, and range of F0 and intensity were then computed within these sequences.

As expected the acoustic results showed that the syllable duration increases when speakers accent syllables. We also found that the proximity of an unaccented syllable to an accented syllable influences the duration of the former. It was also observed that the duration of a syllable is correlated with parameters such as the maximum, minimum, and range of F0 and intensity.

2.2. Visual Analysis

Initially we classified the visual data by identifying perceptible motion in the video. We then computed the mean, standard deviation, and the frequencies of the different motion classes. The analysis of the motion capture data was done using principal component analysis (PCA), in which we concentrated on head motion. We calculated the three most relevant PCs representing rigid head motion for each speaker, with an explained variance between 86% - 95%. Furthermore, correlations between the PCs and the translational and rotational movements were determined. We then calculated the maximum, minimum, and range of the PCs. As with the acoustic analysis, we used the accented and unaccented sequences as references, thereby imposing the same conditions and allowing for an alignment between the prosodic and visual features.

2.3. Audiovisual Results

We noted that the syllable duration has a strong correlation with the maximum, minimum, and range of the PCs. The syllable duration therefore seems to be an indicator of perceptual prominence. The results of the acoustic analysis show that the syllable duration is dependent on the type of syllable. An accented syllable exhibits longer duration than an unaccented syllable. However, the position of a syllable is also important. Approximately 22% of all the motion begins during the unaccented syllable before an accented syllable, and approximately 29% of their movements end during the unaccented syllable following an accented syllable. These findings are very important and indicate that examining accented syllables only is insufficient.

Using the ranges of F0 and intensity of the accented and unaccented sequences we were able to analyze the connection between these features and the visual cues within these sequences. Our results indicate that F0 is strongly correlated with the speaker's degree of activity. It turns out that the ranges of the PCs for each speaker were significantly higher on accented than on unaccented sequences.

We found strong correlations between the prosodic features of F0 and intensity and head movements. Therefore we can assume that there exists some kind of alignment between the two modalities which we aim to exploit in the development of the prediction model discussed in the remainder of this article.

3. Model

Due to the differences in the motion data as well as the use of prosody between speakers we developed specific prediction models for each one of them. Because of the dynamic nature of the acoustic data we used recurrent neural networks (NN) to take into account the temporal context during the prediction task.

Furthermore, we aimed to examine the benefits of including linguistic parameters in the prediction model, given that the cost of hand-labeling the linguistic structure is very high.

We therefore evaluated the predictions of head motion using two sets of inputs. The input of the first model includes both, segmentation-based linguistic and raw prosodic information, whereas the second model just contains frame-wise raw prosodic features. Then we compared the RMSE of the outputs and interpreted those errors to infer the relative influence of the input parameters on the prediction accuracy. Table 1 summarizes the parameters used as input for the artificial neural networks.

The dataset of each speaker was split into three subsets: training (70%of data), validation (15% of data) and test. On average these are about 2490 training samples for training, 534 samples for validation, and 534 samples for testing for each speaker. The activity of the speaker movements was defined by the differences of the principal components calculated: ΔPC1, ΔPC2 and ΔPC3. These are the parameters to be predicted. The frame rate was set to 60Hz.

Table 1: *Input parameters of the network.*

raw prosodic input	
F0	normalized F0
INT	normalized intensity
linguistic/segmentation-based input	
SD	normalized syllable duration
SP	position of the syllable within a phrase 4 = one word phrases e.g. breaks 3 = last syllable 2 = syllables between first and last 1 = first syllable
BND	boundary of the phrase, only the last syllable is marked with a value 4 = intonation phrase 3 = intermediate phrase 2 = breaks/hesitations 0 = if the syllable is not the last one
PH	type of the phrase 4 =continuous phrase 3 = declarative phrase 2 = breaks/hesitation 1 = interrogative phrase
ACC	type of syllable 1 = accented syllable 0.5= unaccented before or after unaccented syllable 0 = unaccented syllable

In the preprocessing step we used min-max normalization on the input data. As mentioned, we used root mean square error (RMSE) to evaluate the network's performance [equation 1]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(o_{ij}-t_{ij})^2}{N}} \qquad (1)$$

Where $\boldsymbol{o}$ are the predicted outputs, and $\boldsymbol{t}$ the target values, N is the number of samples and M the number of components in the output and target vectors.

3.1. Elman NN

The first approach used Elman NN which were trained with linguistic and prosodic input parameters, generating two models: the Elman linguistic neural network (ELNN), and the Elman prosodic neural network (EPNN).ELNN uses seven input nodes, six hidden nodes in a single hidden layer, three output nodes and six context nodes belonging to each hidden node to map the previous output, as shown in Figure 1.

The EPNN architecture differs from ELNN just in the number of inputs, in this case two instead of six. To train the network we used the standard backpropagation algorithm with a learning rate of 0.3.

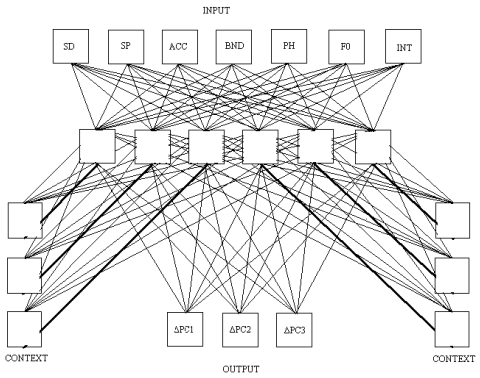

Figure 1: *ELNN architecture with seven input, six hidden, six context and three output units*

The speaker-dependent trained network differs with respect to the threshold, that is, the upper fire threshold of the hidden and context nodes was the maximum observed ΔPC of each speaker, and for each output node we set the maximum value of their observed ΔPCs. The maximum of the ΔPC was computed only for the ranges where we annotated head motion. Hence, the outputs of the neural networks are limited to these areas.
Table 2 shows the resulting RMSE for the ELNNs and Table 3 the resulting RMSE for EPNNs. The results indicate that the average performance of the EPNN (RMSE 0.0904) is much better than of the ELNN (RMSE 0.1848).

Table 2: *Test RMSE of the three output units and average for the ELNN model*

Speaker	ΔPC1	ΔPC2	ΔPC3	Mean
1	0.0423	0.0654	0.0504	0.0927
2	0.0716	0.1490	0.1348	0.2133
4	0.1400	0.0933	0.1093	0.2056
5	0.1087	0.1307	0.0775	0.1868
6	0.1444	0.1643	0.0791	0.2362
7	0.0851	0.1047	0.1107	0.1746

Table 3: *Test RMSE of the three output units and average for the EPNN model*

Speaker	ΔPC1	ΔPC2	ΔPC3	Mean
1	0.0428	0.0663	0.0488	0.0927
2	0.0478	0.0471	0.0827	0.1065
4	0.0392	0.0467	0.0607	0.0859
5	0.0451	0.0486	0.0547	0.0860
6	0.0393	0.0444	0.0422	0.0727
7	0.0613	0.0561	0.0530	0.0986

Figure 2 displays an example of prediction of ΔPC2 of one of our speakers. The solid line indicates the observed data, the dotted line the estimation of the test data with a trained ELNN network and the dashed line a prediction of a trained EPNN network. It shows that there are sections of varying prediction accuracy of both networks. Both networks show satisfactory performance e.g. between the time of 55.2 and 55.3 seconds, however, the prediction of the EPNN match more often with the observed data as can been seen between time 55.5 and 55.6 seconds.

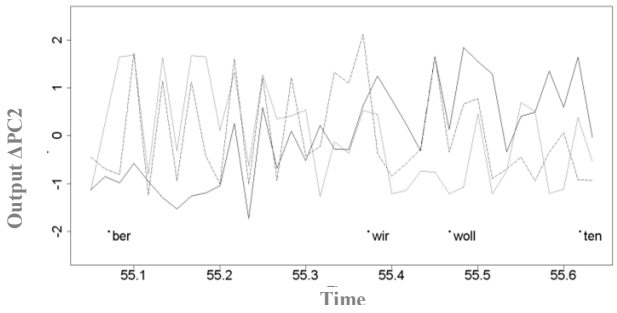

Figure 2: *Diagram with normalized ΔPC2 predicted output of the ELNN (dotted line) and EPNN (dashed line) networks with test data of one of our speaker. The solid line indicates the observed ΔPC2.*

3.2. NARX NN

We also trained speaker-dependent prediction models based on nonlinear autoregressive models with exogenous inputs (NARX) recurrent neural network[8].Unlike other recurrent neural network models, NARX architectures have limited feedback coming only from the output neuron rather than from hidden neurons. Figure 3 displays the architecture of a NARX ANN.

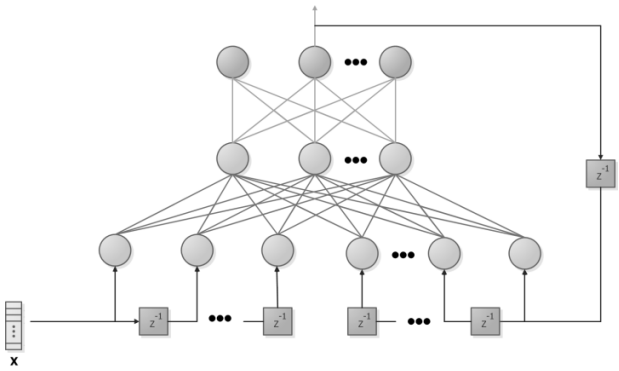

Figure 3: *Architecture of a NARX Neural Network.*

These networks have equal computing power than conventional recurrent networks, and in practice it has been reported that gradient-descent learning can be more effective in NARX networks than in other recurrent architectures with "hidden states" [4]. As for the Elman Networks, we constructed a model using linguistic information as inputs (NARXL) and another one just using raw prosodic features (NARXP). The best architecture on average contained six hidden neurons for NARXL and five hidden neurons for NARXP, with two delay units both for inputs and outputs. Tables 4 and 5 summarize the results for the models constructed using NARX ANN.As with Elman Networks, on average NARXP (RMSE 0.0523) outperforms NARXL (RMSE

0.0531), but in this case the difference between the two is smaller.

Table 4: *Test RMSE of the three output units and average for the NARXL model*

Speaker	ΔPC1	ΔPC2	ΔPC3	Mean
1	0.0254	0.0325	0.0377	0.0559
2	0.0203	0.0253	0.0460	0.0562
4	0.0192	0.0355	0.0313	0.0510
5	0.0123	0.0373	0.0476	0.0618
6	0.0145	0.0214	0.0335	0.0423
7	0.0356	0.0307	0.0221	0.0520

Table 5: *Test RMSE of the three output units and average for the NARXP model*

Speaker	ΔPC1	ΔPC2	ΔPC3	Mean
1	0.0228	0.0288	0.0371	0.0522
2	0.0177	0.0216	0.0450	0.0530
4	0.0190	0.0347	0.0293	0.0492
5	0.0132	0.0387	0.0500	0.0646
6	0.0142	0.0212	0.0330	0.0417
7	0.0366	0.0311	0.0223	0.0530

4. Discussion and Conclusions

In this paper we evaluated two kinds of models to predict head motion from raw prosodic and segmentation-based features including linguistic information. The models were two types of recurrent artificial neural networks, Elman and NARX. Furthermore, we evaluated different input sets with each model, in one case we only used frame-wise F0 and intensity, and in the other case we added to these features hand-labeled linguistic features.

We found that NARX networks outperformed Elman's networks with the two types of input feature vectors.

In addition we could not find benefits in terms of prediction accuracy using hand-labeled linguistic data. The mean RMSE of the ELNN networks was approximately twice as great as the mean RMSE of the EPNN networks, where as there was almost no difference in mean RMSE between the NARXL and NARXP.

Given these results, there is obviously no reason to use linguistic features since their extraction requires a considerable amount of manual labeling. However, we argue that this is true when there is no silence during speech; however, during silent regions other types of information are required. In previous work we found that 5.4% of all head motion events occurred in pauses. With the prosodic inputs, the lack of F0 and intensity during pauses does not facilitate the prediction of motion. In those cases, some linguistic inputs like the type of the phrase boundary might be necessary to predict motion activity. One of the objectives of future work will be to solve the problem of motion prediction during silent intervals.

Furthermore we plan to develop models from larger datasets, as well as construct speaker-independent models. Finally, in the current study the outputs were given in terms of ΔPC, which represent a degree of activity in head motion. A further step will be the mapping from ΔPC to the type of annotated head motion.

5. Acknowledgements

The first author is funded by the European Social Fund (ESF) and supported by the Berlin Senate for Economics, Technology and Research.

6. References

[1] Busso, C., Deng, Z., Grimm, M. Ulrich Neumann, ShrikanthNarayanan, Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis, IEEE Transactions on Audio, Speech, and Language Processing, v.15 n.3, pp.1075-1086, March 2007

[2] Cosker, D. Marshall, P. Rosin, Y.A. Hicks, ``Speech Driven Facial Animation using a Hidden Markov Co-articulation Model", In Proc. of IEEE International Conference on Pattern Recognition (ICPR), Vol. 1, pp 128-131, 2004.

[3] Hofer, G., &Shimodaira, H.: Automatic Head Motion Prediction from Speech Data. In Proc. Interspeech 2007, Antwerp, Belgium, 2007.

[4] Horne, B., & Giles, C.: An experimental comparison of recurrent neural networks, in Advances in Neural Information Processing Systems 7, G. Tesauro, D. Touretzky, and T. Leen, Eds. Cambridge, MA: MIT Press, 1995, pp. 697-704.

[5] Hönemann, A., Mixdorff, H., Fagel, S. (forthcoming),"A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualization". In: Proceeding of Nordic Prosody XI. Peter Lang.

[6] Hönemann, A., Mixdorff, H., Fagel, S., "Alignment between Rigid Head Movements and Prosodic Landmarks". In: Tagungsband Elektronische Sprachsignalverarbeitung ESSV 2013. Bielefeld, Germany: pp. 181.188,Petra Wagner (Hrsg.)

[7] Massaro, D.W., Beskow, J., Cohen, M.M., Fry, C.L. and Rodriguez, T., "Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", in Audio-Visual Speech Processing, 1999.

[8] Menezes, J., Barreto, G.: Long-term time series prediction with the NARX network: An empirical evaluation. Neurocomputing 71 (2008) 3335–3343

[9] Yehia, H. C., Kuratate, T., &Vatikiotis-Bateson, E.: Linking facial animation, head motion and speech acoustics. Journal of Phonetics, 30(3), pp. 555-568.