Prediction-Based Classification for Audiovisual Discrimination Between Laughter And Speech

Stavros Petridis
Department of Computing
Imperial College London
stavros.petridis04@ic.ac.uk

Maja Pantic
Dpt. of Computing / EEMCS
Imperial College London /
Univ. of Twente
m.pantic@imperial.ac.uk

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
jeffcohn@cs.cmu.edu

Abstract—**Recent evidence in neuroscience support the theory that prediction of spatial and temporal patterns in the brain plays a key role in human actions and perception. Inspired by these findings, a system that discriminates laughter from speech by modeling the spatial and temporal relationship between audio and visual features is presented. The underlying assumption is that this relationship is different between speech and laughter. Neural networks are trained which learn the audio-to-visual and visual-to-audio feature mapping together with the time evolution of audio and visual features for both classes. Classification of a new frame / sequence is performed via prediction. All the networks produce a prediction of the expected audio / visual features and their prediction errors are combined for each class. The model which best describes the audiovisual feature relationship, i.e., results in the lowest prediction error, provides its label to the input frame / sequence. Using 4 different datasets, the proposed system is compared to standard feature-level fusion on cross-database experiments. In almost all test cases, prediction-based classification outperforms feature-level fusion. Similar conclusion are drawn when adding artificial feature-level noise to the datasets.**

I. Introduction

Prediction is believed to be very important for human perception, thought and action. It has been proposed that the human brain continuously generates predictions that anticipate the future [1]. Previous studies have confirmed that humans use prediction to estimate the consequence of motor commands [2] and to segment continuous activity into discrete events [3]. Recent research in neuroscience links cognitive deficits with the breakdown of the prediction system [4], [5]. Prediction also plays an important role in recent models of the brain [6], [7] and particularly in [8] it is proposed that predictions about both temporal and spatial patterns are important.

Predictive models which predict ahead in time have been used for a long time in research areas like predictive control and audio processing and they have also been used for time series classification [9], [10]. Predictive models which make cross-modal predictions have been widely used in speech driven facial animation [11] and audiovisual speech enhancement [12]. However, there are just a few works using both temporal and spatial predictive models, and most of them use the hierarchical model proposed in [8]. In this work, we use both types of predictive models in order to discriminate audiovisual laughter and speech episodes.

Laughter is one of the most common and useful human social signals [13]. It helps humans to express their emotions and intentions in social interactions and provides useful feedback during interpersonal interactions. Laughter consists of an audio component, the laughter vocalization, and a visual component which involves facial activity around the mouth, the cheeks, and often the upper face. Recently, few audiovisual efforts have been reported aiming to discriminate laughter from speech combining audio and visual information or recognizing other non-linguistic vocalizations [14], [15], [16]. These works use either feature-level fusion with discriminative classifiers like neural networks (NNs) and Support Vector Machines (SVM), or generative classifiers like Hidden Markov Models. In the former case mostly spatial information about the features is used, whereas in the latter case mostly temporal information is used. Previous studies [17], [18] have shown that both approaches result in similar performance when presegmented episodes are used. The motivation for this work is to compare feature-level fusion, which is one of the most commonly used types of fusion, with a new type of fusion, presented in this paper, based on prediction.

There has been a lot of research in examining the relationship between acoustic and visual speech features [19], [20], [21]. Most of the studies are focused only on the audio-to-visual features mapping. On average visual features are predicted with a correlation of 0.7, when linear models are used [19], [21], although measures as high as 0.8 have been reported [20] and 0.85 when nonlinear models are used, like (NNs), [19]. Of course the correlation varies depending on the features and datasets used. To the best of our knowledge there is no work which performs such correlation analysis for laughter. However, it is reasonable to believe that the correlation between audio and visual features is different in speech and laughter. Similarly, it is reasonable to assume that the time evolution of audio and visual features is different between speech and laughter. There are a couple of works that have used these principles separately. Kumar et al. [22] model the time evolution of audiovisual features in speech using linear models, and use the model parameters for synchrony detection between the audio visual streams. In another work [23] speech and laughter were successfully discriminated by learning the frame-to-frame correlations between audio and visual features.

Driven by those results we propose a system that uses this difference in correlation and time evolution of audio and visual features to discriminate laughter and speech. This is achieved by explicitly modeling the spatial relationship between audio and visual features and the temporal relationship between past and future values of audio and visual features using predictive models. This approach is inspired by the memory-prediction framework [8]. The key idea is that an audio / visual input can make a prediction for an expected audio and visual input. Our implementation is much simpler and very different from the proposed framework, but it is based roughly on the same idea. The models make an audio and a visual prediction based on video, i.e., they predict what they expect to "hear" now and "see" in the future based on what they "see" now. A similar prediction is made based on audio, i.e., the models predict what they expect to "see" now and "hear" in the future based on what they "hear" now.

Towards this direction we train four NNs (two for each class), which learn the audio-to-visual and visual-to-audio feature mapping for speech and laughter and four NNs which learn the time evolution of audio and visual features. It is expected that laughter networks will produce a better prediction than speech networks when the input is laughter, since they have learned the audiovisual relationship and time evolution of the features for laughter, and vice versa. When new input comes, which is usually a window of past values, then its audio and visual features are fed to all 8 networks which produce a prediction error. The audio-to-video and video-to-audio mapping systems can be combined with the audio-to-audio and video-to-video systems from the same class in order to take advantage of the bidirectional and the past to future relationship between audio and visual features (see Section IV). Therefore a combined error is produced for each class. Selecting the model that produces the lowest error a new frame or a sequence, by summing the errors across all frames, can be labeled accordingly. In other words, a frame or a sequence is labeled based on the model which best describes the audiovisual feature relationship. It does not matter if the prediction is good or bad, just that it is better than the other network's prediction.

The proposed approach is compared to feature-level audiovisual fusion on cross-database experiments using two challenging datasets, AMI and DD and two easier datasets, SAL and AVLC. Both systems perform similarly when trained on the AMI dataset, however when trained either on the SAL or DD dataset the proposed system outperforms the feature-level fusion. This is an indication that the prediction system is able to learn a good model even when a less diverse and challenging dataset is used for training. In a second experiment artificial feature-level Gaussian noise was added to the datasets and similar conclusions were drawn confirming the benefits of combining spatial and temporal predictive models.

II. Databases

For the purpose of this study we used four datasets corresponding to 4 different scenarios. The first scenario involves social interactions between 4 subjects (AMI dataset), the second one involves interaction between a subject and an artificial agent (SAL dataset), the third one involves an interview between a therapist and a patient who meets the criteria for major depressive disorder (DD dataset) and the last one involves a subject watching funny video clips (AVLC dataset).

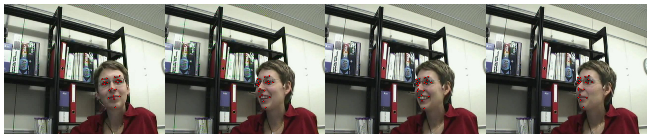
TABLE I: Description of the four datasets used in this study.

AMI			
Type	No Episodes / No Subjects	Total Duration (sec)	Mean / Std (sec)
Laughter	124 / 10	145.36	1.17 / 0.73
Speech	154 / 10	285.92	1.86 / 1.12
SAL			
Laughter	94 / 15	136.96	1.46 / 0.78
Speech	177 / 15	377.32	2.13 / 0.80
DD			
Laughter	217 / 36	135.10	0.62 / 0.41
Speech	327 / 36	469.40	1.44 / 0.82
AVLC			
Laughter	421 / 8	1601.32	3.80 / 6.43
Speech	0 / 0	0	0 / 0

AMI: In the AMI Meeting Corpus [24] people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face (720 x 576, 25 frames per second (fps)) and the related individual headset audio recordings (16kHz). Although there is a personal microphone for each subject there is background noise present from the other subjects. The camera is fixed and since people are involved in a discussion they tend to move their head a lot and they are rarely in frontal pose. The language used in the meetings is English, with speakers being mostly non-native speakers. For our experiments we used seven meetings (IB4001 to IB4011) and the relevant recordings of ten participants, 8 males and 2 females.

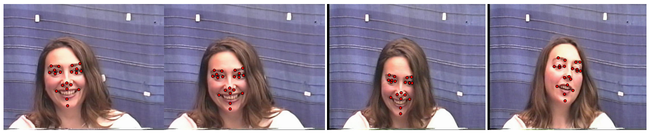
SAL: The Sensitive Artificial Listener (SAL) technique as described in [25] "focuses on conversation between a human and an agent that either is or appears to be a machine and it is designed to capture a broad spectrum of emotional states". The subjects interact with 4 different agents that have different personalities and the audiovisual response of the subjects while interacting is recorded. For our experiments we used 15 subjects, 8 males and 7 females. We used the close-up video recordings of the subjects face (720 x 576 for 12 subjects and 352 x 288 for 3 subjects, 25 fps) and the related audio recording (48kHz for 12 subjects and 44.1kHz for 3 subjects). Most of the time the subjects have frontal pose and head movements are small. The language used in the meetings is English, with all speakers being native.

DD: The Depression Dataset (DD) consists of interviews between a therapist and a subject who suffers from depression [26]. During the interview subjects produce several non-linguistic vocalizations and laughter is one of them. For our experiments we used 36 subjects, 11 males and 25 females. In this study, the camera positioned approximately 15 degrees to the right was used, which recorded the patient's face and shoulders (640 x 480, 30 fps). There was no personal microphone for the patients, so the audio recorded by the



(a) Frame 2 (b) Frame 82 (c) Frame 162 (d) Frame 176 (Last Frame)

Fig. 1: Example of a laughter episode, from the AMI dataset, with illustrated facial point tracking results.



(a) Frame 1 (b) Frame 15 (c) Frame 30 (d) Frame 46 (Last Frame)

Fig. 2: Example of a laughter episode, from the SAL dataset, with illustrated facial point tracking results.

camera microphone was used (48kHz) and as a result the audio signal is noisy. The language used in the meetings is English, with all speakers being native. Non-frontal pose and moderate head motion were common.

AVLC: The AudioVisual Laughter Cycle (AVLC) database [27] consists of 24 subjects which were recorded while watching video clips for 10 minutes. The goal of this experiment was to elicit laugh from the participants. A webcam was used to record the subject's face (640 x 480, 25 fps) and a headset microphone (16kHz) was used for each subject. Head movements are small since the subjects watch a video, and audio noise is low since there are no other people in the recording room. In this study, we used 8 subjects (5, 6, 7, 11, 13, 14, 16, 18), 3 females and 5 males, since the webcam is positioned closer to them resulting in higher resolution of the face.

All laughter and speech episodes used in this study were pre-segmented based on audio. This means that the start and end point of a laughter episode is defined for the audio signal and then the corresponding video frames are extracted. For the AMI, DD and AVLC datasets laughter episodes were selected based on the annotations provided with the AMI Corpus and the DD and AVLC annotation files, respectively. After examining these episodes, we only kept those that do not co-occur with speech, do not contain profile views of the face (i.e. all facial components are still visible), and satisfy the criterion as suggested in [28]: "Laughter is defined as being any perceptibly audible expression that an ordinary person would characterize as laughter if heard under everyday circumstances". For the SAL dataset we manually annotated laughter episodes according to these rules. Similarly, speech segments for the AMI and DD datasets were determined by the annotations provided with the AMI Corpus and DD annotation files, respectively, and they were manually annotated and segmented for the SAL dataset. Speech annotation and segmentation was not performed for the AVLC dataset, since it does not contain speech. Speech segments were selected such that do not contain long pauses between two consecutive words. Fig. 1 and 2 show laughter episodes from the AMI and SAL datasets, respectively. Details of the four datasets are given in Table I.

III. Features

Audio Features: Cepstral features, such as Mel Frequency Cepstral Coefficients (MFCCs), have been widely used in speech recognition and have also been successfully used for laughter detection [29]. In addition, it has been shown that cepstral coefficients are more correlated to visual features than prosodic features [20]. Therefore we only use MFCCs for our experiments. Although it is common to use 12 MFCCs for speech recognition we only use the first 6 MFCCs, given the findings in [29], where 6 and 12 MFCCs resulted in the same performance for laughter detection. These 6 audio features are computed every 10ms over a window of 40ms, i.e. the frame rate is 100 fps.

Visual Features: Both appearance and shape features have been used in previous works on audiovisual speech and emotion recognition [30], [31], [32], [33]. In this work, we use shape features since they are less sensitive to registration errors than appearance features. Spontaneous data are used so registration errors are expected to be high as a result of the large head movements and non-frontal pose. In addition, there is evidence that shape features are more correlated with audio. Kumar et al [22], using a similar time evolution model with the one presented in section IV, found that shape features outperformed appearance features when combined with audio in a synchrony detection application. Therefore we use shape features by tracking 20 facial points using the Patras-Pantic particle filtering tracking scheme [34]. These points are the corners of the eyebrows (2 points), the eyes (4 points), the nose (3 points), the mouth (4 points) and the chin (1 point) as shown in Fig. 1 and 2. For each video segment containing K frames, we obtain a set of K vectors containing 2D coordinates of the 20 points. Using a Point Distribution Model (PDM), by applying principal component analysis to the matrix of these K vectors, head movement can be decoupled from facial expression. Using the approach proposed in [35], the facial expression movements are encoded by the projection of the tracking points coordinates to the N principal components (PCs) of the PDM which correspond to facial expressions. In this study we build a PDM based on AMI, so our shape features are the projection of the 20 points to the 4 PCs which were found to correspond to facial expressions (PCs 7 to 10). These 4 visual features, called shape parameters, are extracted at the video frame rate, i.e., 25 fps for the AMI, SAL, AVLC datasets and 30 fps for the DD dataset. The same PDM, built on AMI, is used in order to compute the shape parameters in all datasets. Further details of the feature extraction procedure can be found in [35].

IV. Methodology

For each of the two classes, speech and laughter, we train two NNs which model the relationship between audio and visual features and two which model the relationship between past and future values of audio and visual features, respectively. In other words, the first network learns the audio-to-visual feature mapping, the second network learns the visual-to-audio feature mapping and the last two networks learn the mapping between past and future values for the audio and visual features, respectively. The input for each network is a window of past values and the output is the predicted audio or visual feature value.

In the first set of networks, which make predictions across modalities, the relationship between the audio (A^L, A^S) and visual (V^L, V^S) features in speech and laughter is modeled by (NN^L_{AV}), (NN^L_{VA}) for laughter and (NN^S_{AV}), (NN^S_{VA}) for speech. In other words, the first / second network takes as inputs the concatenated audio / visual features of the input window and predicts the corresponding visual / audio features at the same time t (eq. 1 - 4).

$$NN^L_{AV} : f^L_{A\to V}(A^L[t-k_1, t]) = \hat{V}^L_{A\to V}[t] \approx V^L[t] \tag{1}$$

$$NN^L_{VA} : f^L_{V\to A}(V^L[t-k_2, t]) = \hat{A}^L_{V\to A}[t] \approx A^L[t] \tag{2}$$

$$NN^S_{AV} : f^S_{A\to V}(A^S[t-k_3, t]) = \hat{V}^S_{A\to V}[t] \approx V^S[t] \tag{3}$$

$$NN^S_{VA} : f^S_{V\to A}(V^S[t-k_4, t]) = \hat{A}^S_{V\to A}[t] \approx A^S[t] \tag{4}$$

As shown in eq. 1, 2, 3, 4, different windows k_1, k_2, k_3, k_4 are used for each predictor. Also note that the feature values at time t are used as well in order to predict the feature values in the other modality at the same time t.

In the second set of networks, which make predictions within each modality, the relationship between past and future audio and visual features in speech and laughter is modeled by (NN^L_{AA}), (NN^L_{VV}) for laughter and (NN^S_{AA}), (NN^S_{VV}) for speech. In other words, the first / second network takes as inputs the concatenated past audio / visual features and predicts the corresponding audio / visual features at the same time t.

$$NN^L_{AA} : f^L_{AA}(A^L[t-k_5, t-1]) = \hat{A}^L_{A\to A}[t] \approx A^L[t] \tag{5}$$

$$NN^L_{VV} : f^L_{VV}(V^L[t-k_6, t-1]) = \hat{V}^L_{V\to V}[t] \approx V^L[t] \tag{6}$$

$$NN^S_{AA} : f^S_{AA}(A^S[t-k_7, t-1]) = \hat{A}^S_{A\to A}[t] \approx A^S[t] \tag{7}$$

$$NN^S_{VV} : f^S_{VV}(V^S[t-k_8, t-1]) = \hat{V}^S_{V\to V}[t] \approx V^S[t] \tag{8}$$

As shown in eq. 5, 6, 7, 8, different windows k_5, k_6, k_7, k_8 are used for each predictor. In this case the feature values at time t are excluded since that is exactly what we want to predict.

Once training is complete and the mapping functions (f^L, f^S) are learned then the networks can be used for classification. When a new sequence is available the audio and visual features are computed, which are fed to all networks from eq. 1 - 8 resulting in 8 errors per frame. The error measure used is the mean squared error (MSE). The total error for each predictor is computed by summing the errors across all frames, N, resulting in 8 errors per sequence. The errors for the 4 laughter predictors are computed using eq. 9 and 10. Similarly, the errors for the 4 speech predictors are computed by replacing the superscript L with S.

$$e^L_{A\,or\,V\to V} = \sum_{i=1}^{N} MSE(\hat{V}^L_{A\,or\,V\to V}[i], V^L[i]) \tag{9}$$

$$e^L_{A\,or\,V\to A} = \sum_{i=1}^{N} MSE(\hat{A}^L_{A\,or\,V\to A}[i], A^L[i]) \tag{10}$$

Then we can combine the errors in order to generate a new error which takes into account both the bidirectional and past to future relationships of audio and visual features as shown in eq. 11 and 12 subject to constraints eq. 13 and eq. 14.

$$\begin{aligned} e^L =& w_1 \times e^L_{A\to V} + w_2 \times e^L_{V\to A} + \\ &+ w_3 \times e^L_{A\to A} + w_4 \times e^L_{V\to V} \end{aligned} \tag{11}$$

$$\begin{aligned} e^S =& w_5 \times e^S_{A\to V} + w_6 \times e^S_{V\to A} + \\ &+ w_7 \times e^S_{A\to A} + w_8 \times e^S_{V\to V} \end{aligned} \tag{12}$$

$$w_1 + w_2 + w_3 + w_4 = 1 \tag{13}$$

$$w_5 + w_6 + w_7 + w_8 = 1 \tag{14}$$

A sequence is labeled as laughter or speech depending on which model produced the best estimate, i.e., the lowest prediction error, eq. 11 and 12. In other words, a sequence is labeled based on the following rule:

$$IF\ \ e^S > e^L\ \ THEN\ \ \mathbf{L}\ \ ELSE\ \ \mathbf{S} \tag{15}$$

V. Experimental Studies

A. Experimental Setup

In order to assess the performance of the method presented in section IV, cross-database experiments between the AMI, SAL, DD and AVLC datasets were performed. As discussed in section II, AMI is a challenging dataset since the subjects rarely have a frontal view, there are large head movements and the audio recording is noisy. DD is also a challenging dataset containing noisy audio recordings and moderate head movements. On the other hand, SAL and AVLC are easy datasets since subjects almost always look straight at the camera, there are relatively small head movements and audio noise is low.

In all experiments the AMI dataset was used as a validation set in order to optimize the parameters. In the first experiment, a system is trained on the SAL dataset and tested on the DD and AVLC datasets and in the second experiment a system is trained on the DD dataset and tested on the AMI and AVLC datasets. The AVLC does not contain speech so it cannot be used for training. Results on the AMI dataset are also presented but should only be considered as validation and not test results.

Preprocessing: As mentioned in section III, 4 visual features and 6 audio features are used. Before training, the audio and

TABLE II: F1 and classification rates (CR) for the feature-level fusion (FF) system and the prediction based system on cross database experiments. The AMI dataset is used as a validation set to optimize the parameters. The AVLC dataset contains only speech so only the CR is reported.

Classification System	F1 Laughter	F1 Speech	CR	F1 Laughter	F1 Speech	CR	CR
Train SAL →	**Test AMI**			**Test DD**			**Test AVLC**
A + V (FF)	71.4	84.8	80.1	70.8	86.1	81.2	80.1
A + V Pred.	88.2	91.2	89.9	78.6	86.3	83.4	93.3
Train DD →	**Test AMI**			**Test SAL**			**Test AVLC**
A + V (FF)	68.7	83.7	78.6	87.1	94.3	92.1	55.1
A + V Pred.	83.0	87.1	85.3	89.0	93.6	91.9	90.3

TABLE III: Optimal window lengths, k_1 to k_8, in msec and weights w_1 to w_8, for prediction-based classification and feature-level fusion (FF). The AMI dataset was used as a validation set.

	Prediction-Based System				FF
Training Set	[k_1,k_2,k_3,k_4]	[k_5,k_6,k_7,k_8]	[w_1,w_2,w_3,w_4]	[w_5,w_6,w_7,w_8]	FF Window
SAL	[90 30 10 10]	[90 10 50 20]	[0.45 0.30 0.00 0.25]	[0.45 0.50 0.00 0.05]	90
DD	[90 60 30 10]	[90 20 10 10]	[0.00 0.15 0.15 0.70]	[0.05 0.15 0.20 0.60]	80

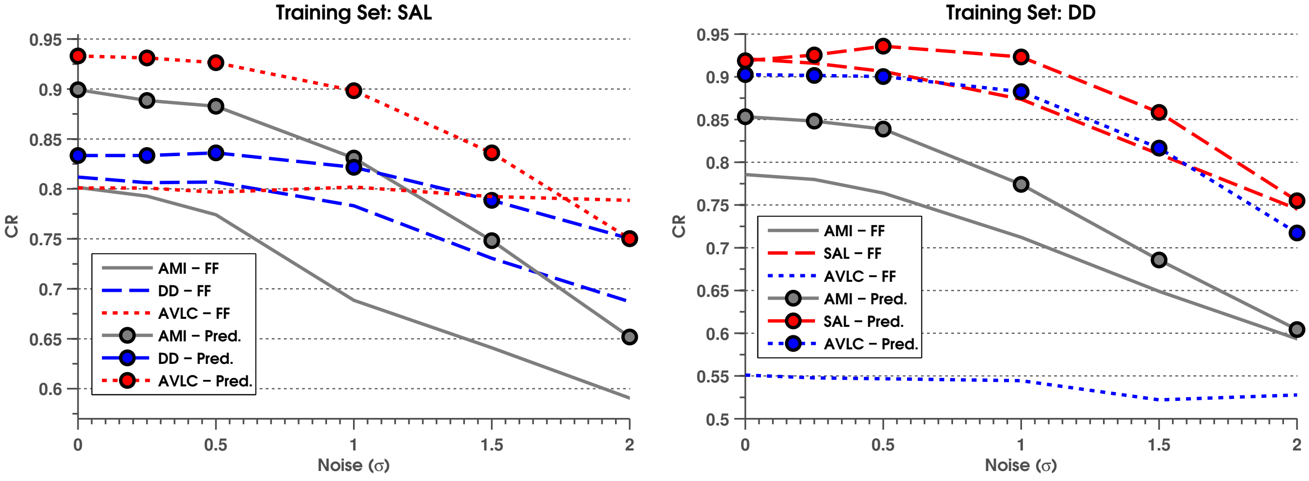
visual features are synchronized by upsampling the visual features, to match the frame rate of the audio features (100fps), by linear interpolation. All the audio and visual features are z-normalized per subject, to a zero mean and unity standard deviation. Since there is no speech in the AVLC dataset, the mean and standard deviation values of the entire AMI dataset were used for normalization.

Parameter Optimization: This step is used to compute the optimal number of hidden neurons in NNs, the optimal number of windows k_1 to k_8 from eq. 1 to 8 and the optimal values for weights w_1 to w_8 from eq. 11 and 12. Ten different numbers of hidden neurons are considered equally spaced from 2 to 30. The window lengths considered are from 0 to 10, i.e. from 0ms to 100ms, for k_1 to k_4 and 1 to 10, i.e. from 10ms to 100ms, for k_5 to k_8. For each network (eq. 1 to 8) all possible combinations of window lengths and number of hidden neurons are tried. Training is performed on one of the three datasets, AMI, SAL, DD and the performance of each combination is evaluated on AMI. The performance measure used for the NN^L / NN^S is the distance between the predicted values of all the laughter / speech and speech / laughter frames. The motivation for this measure is that we want the laughter / speech networks to have a lower / higher prediction error when laughter frames are given as input and a higher / lower prediction error when the input is speech. Finally, the combination of windows and number of hidden neurons resulting in the maximum distance is selected as the optimal. The weight values considered vary from 0 to 1 in steps of 0.05. All possible combinations, subject to constraints 13 and 14, are tried and the one resulting in the best performance in terms of the F1 measure on the AMI dataset is selected as the optimal.

Training: Once the parameters for each network are optimized on AMI then training follows on the training dataset. The input for each network is a window of past values and the goal during training is to minimize the error between the actual and the predicted visual / audio features. Laughter networks are trained using only laughter examples and similarly speech networks are trained with speech examples only. The NNs used in this study have one hidden layer, using sigmoid activation functions and they are trained for up to 500 epochs.

Following the approach of section IV, 8 NNs are trained, eq. 1 - 8, and each sequence is labeled using rule 15. It has been shown that static classifiers like NNs or SVMs have similar performance with (Coupled) Hidden Markov Models when classifying presegmented episodes of non-linguistic vocalizations [17], [18]. Therefore for comparison we also report the results of an audiovisual feature-level fusion approach based on NNs, since we use NNs for prediction as well. This approach is based on concatenating the audio and visual at each frame, and then feeding them to a NN. The output of the network is continuous, between -1 and 1, corresponding to speech and laughter, and assigns a score to each frame. Again, the scores across all frames of a sequence are summed and if the final score is higher / lower than 0 then the sequence is labeled as laughter / speech. A window of past feature values is used as well and parameter optimization for selecting the optimal window length and the number of hidden neurons is performed as described above. The only difference is that there are no weights to optimize.

Since NNs, which are initialized randomly, are used for both approaches all experiments are repeated 5 times and the mean values for the performance measures are reported. The performance measures used in this study are the classification rate and the F1 rate. Therefore, in both approaches, exactly



(a) Training Set: SAL

(b) Training Set: DD

Fig. 3: Classification rate as a function of the noise added for the three different training sets. The added noise is Gaussian with zero mean and different values for standard deviation as shown in the y-axes and it is added to each subject separately. The solid, dashed and dotted lines correspond to the first, second and third test sets, respectively. The lines which correspond to the prediction systems are plotted with markers.

the same audio / visual features are used, and the same classification protocol is followed. The only difference is how classification is performed, in the first approach via prediction and in the second case using the standard feature-level fusion.

B. *Results*

The optimal window lengths and optimal weights are shown in Table III. As expected, the optimal window lengths and weights found are different in each dataset reflecting their different characteristics. It also interesting that the weights for speech and laughter are very similar in all cases. Finally, the windows for cross-modality prediction (k_1, k_2, k_5, k_6) tend to be longer than the windows used for within-modality prediction (k_3, k_4, k_7, k_8).

Table II shows the performance for each system. In the first experiment, in which training is performed on the easiest dataset (SAL) we see that the prediction system leads to better performance than feature-level fusion in both test sets. The absolute difference in CR is 2.2% and 13.2% for the DD and AVLC datasets, respectively. In the second experiment (training on the DD dataset) the prediction system significantly outperforms FF on the AVLC dataset and achieves better performance for F1 laughter but slightly worse performance for F1 speech and CR on the SAL dataset.

Based on these results, we see that when we train on a relatively easy dataset (SAL) then the prediction system is able to generalize much better on an unseen difficult dataset than feature-level fusion. We also see that the prediction system is less sensitive to data normalization resulting in much better performance, up to 35.2% absolute difference in CR, on the AVLC dataset, which is an easy dataset with small head movements and low audio noise, than feature-level fusion. The latter is severely affected by the different way normalization is performed achieving low CR.

In order to have an indication about the robustness to noise of the prediction system, artificial feature-level Gaussian noise was added to the audio and visual features. In audiovisual speech recognition it is common to add Gaussian noise on the audio signal [31], [30]. In computer vision applications it is common to add noise or occlusions on the image. However, in an audiovisual setting there is no straightforward way to compare the amount of noise added to the audio signal with the amount of noise added to an image. Therefore, in this first approach we have added Gaussian noise per subject directly to the audio and visual features. This is an artificial scenario which allows us explicitly control the added noise to the features and therefore ensure that the same amount is added to both audio and visual features. Gaussian noise with zero mean and standard deviation between 0.25σ and 2σ is added to both the audio and visual features, where each time σ is the standard deviation of the feature that noise is added to.

Fig. 3a and 3b show the performance of both approaches in the presence of noise. Similar conclusions as above can be drawn. The prediction-based approach outperforms feature-level fusion for all noise levels, for almost all test cases that performs better in the noise-free case. The only exception is when training on the SAL dataset and testing on the AVLC for the maximum amount of noise, 2σ. It seems that this is the point that the combination of different normalization conditions for the AVLC and the added noise begins to affect the prediction-based approach more than feature-level fusion.

The main advantage of the prediction system is that it does not explicitly rely on the actual values of the features as in the case of feature-level fusion. The problem is converted in competition between two models, a laughter and a speech model. It does not matter if the prediction is good or bad, what

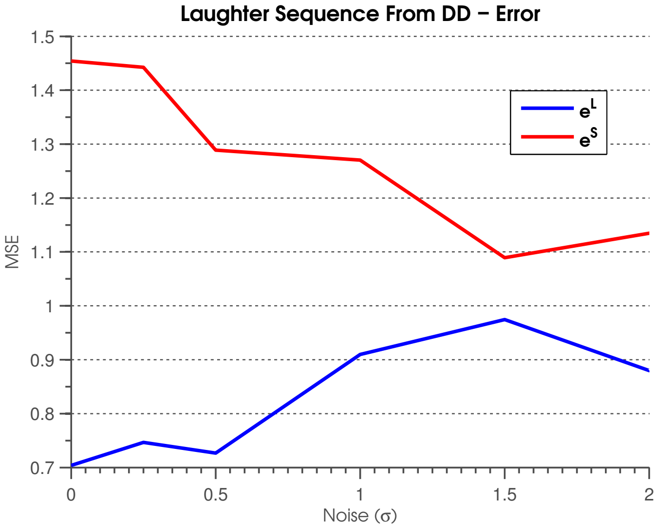


Fig. 4: Error of the laughter and speech models (eq. 11 and 12) as a function of the added noise. The predicion system was trained on the SAL dataset and tested on a laughter sequence from the DD dataset.

matters is if it is closer to the actual values than the competitor model. And since the audio-visual feature relationship is different in laughter than in speech, it is expected that the right model will be closer to the real feature values. An illustration of this principle is shown in Fig. 4, where the overall error of the laughter and speech networks for a laughter sequence from the SAL dataset is plotted. As the noise level increases the prediction error of the correct model (laughter) becomes worse but as long as it stays below the wrong model (speech) the sequence is labeled correctly. It does not matter if the absolute prediction error increases with the addition of noise, what matters is the relative position of the two errors.

VI. Conclusions

A new classification approach based on prediction was presented for the problem of audiovisual laughter-vs-speech discrimination. Neural networks were trained in order to model the time evolution and the correlation between audio and visual features for speech and laughter. The key idea is that classification is based on the model that best describes the spatial and temporal relationship between the audio and visual features. This approach outperforms feature-level fusion in most of the test cases that were considered. It also achieves good performance when a relatively simple dataset (SAL) is used for training, and a more challenging dataset (AMI or DD) is used for testing, which indicates that classification based on prediction can produce a good model even when the available dataset is not challenging enough. Initial results on artificial noisy conditions indicate that the system is also more robust to noise than feature-level fusion. However, more experiments with data noise rather than feature noise are needed. We are currently experimenting with other predictors, like Gaussian processes and support vector regression in order to investigate their performance and test the degree at which the prediction principle is dependent on the regressor used. Finally, since this work was inspired by [8] which states that use of memory is important for prediction, possible ways of explicitly integrating memory in the system are investigated.

VII. Acknowledgments

The research presented in this paper has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB) and by the U.S. NIMH grant 51435.

References

[1] M. Bar, "The proactive brain: using analogies and associations to generate predictions," *Trends in Cognitive Sciences*, vol. 11, no. 7, pp. 280–289, 2007.

[2] D.M. Wolpert, Z. Ghahramani, and M.I. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880, 1995.

[3] J.M. Zacks and K.M. Swallow, "Event segmentation," *Current Directions in Psychological Science*, vol. 16, no. 2, pp. 80, 2007.

[4] J.M. Zacks and J.Q. Sargent, "Event perception: A theory and its application to clinical neuroscience," *Psychology of Learning and Motivation*, vol. 53, pp. 253–299, 2010.

[5] M.S. Kraus, R.S.E. Keefe, and R.K.R. Krishnan, "Memory-prediction errors and their consequences in schizophrenia," *Neuropsychology review*, vol. 19, no. 3, pp. 336–352, 2009.

[6] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[7] M. Haruno, D.M. Wolpert, and M. Kawato, "Mosaic model for sensorimotor learning and control," *Neural Computation*, vol. 13, no. 10, pp. 2201–2220, 2001.

[8] J. Hawkins and S. Blakeslee, *On intelligence*, Owl Books, 2005.

[9] A. Kehagias and V. Petridis, "Predictive modular neural networks for time series classification," *Neural Networks*, vol. 10, no. 1, pp. 31–49, 1997.

[10] D. Coyle, G. Prasad, and T.M. McGinnity, "A time-series prediction approach for feature extraction in a brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 4, pp. 461–467, 2005.

[11] R. Gutierrez-Osuna, PK Kakumanu, A. Esposito, ON Garcia, A. Bojorquez, JL Castillo, and I. Rudomin, "Speech-driven facial animation with realistic dynamics," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp. 33–42, 2005.

[12] Laurent Girin, Jean-Luc Schwartz, and Gang Feng, "Audio-visual enhancement of speech in noise," *The Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[13] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[14] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *Proc. ACM Int'l Conf. on Multimodal interfaces*, 2008, pp. 37–44.

[15] S. Scherer, F. Schwenker, N. Campbell, and G. Palm, "Multimodal laughter detection in natural discourses," *Human Centered Robot Systems*, pp. 111–120, 2009.

[16] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," *Lecture Notes in Computer Science*, vol. 5237, pp. 137 – 148, 2008.

[17] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. Dynamic Modelling of Human Nonverbal Behavior from Multiple Cues and Modalities," in *Proc. ACM ICMI*, 2009, pp. 23–30.

[18] B. Schuller, F. Eyben, and G. Rigoll, "Static and Dynamic Modelling for the Recognition of Non-verbal Vocalisations in Conversational Speech," *Lecture Notes in Computer Science*, vol. 5078, pp. 99–110, 2008.

[19] H.C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, no. 3, pp. 555–568, 2002.

[20] C. Busso and S. S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007, 1558-7916.

[21] M. S. Craig, P. Lieshout, and W. Wong, "A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers," *J. Acoustical Soc. America*, vol. 124, no. 5, pp. 3183–3190, 2008.

[22] K. Kumar, J. Navratil, E. Marcheret, V. Libal, G. Ramaswamy, and G. Potamianos, "Audio-visual speech synchronization detection using a bimodal linear prediction model," in *IEEE CVPR Workshops*, 2009, pp. 53 –59.

[23] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter and speech using audio-visual feature prediction," in *IEEE ICASSP*, 2010, pp. 5254–5257.

[24] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, "The AMI meeting corpus," in *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 2005, pp. 137–140.

[25] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation," in *Programme of the Workshop on Corpora for Research on Emotion and Affect*, pp. 1–4.

[26] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Intern. Conf. on ACII 2009*. IEEE, 2009, pp. 1–7.

[27] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle Database," *Proc. LREC*, 2010.

[28] J. A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 1581–1597, 2001.

[29] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST Meeting Recognition Workshop*, 2004.

[30] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[31] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[32] P.S. Aleksic and A.K. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2007.

[33] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *CVPR*, 2005, pp. 967–972.

[34] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *Int'l Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 97–104.

[35] D. Gonzalez-Jimenez and J. L. Alba-Castro, "Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry," *IEEE Trans. Inform. Forensics and Security*, vol. 2, no. 3, pp. 413–429, 2007.