CLASSIFYING LAUGHTER AND SPEECH USING AUDIO-VISUAL FEATURE PREDICTION

Stavros Petridis, Ali Asghar

Imperial College London
Dept. of Computing, London, UK
{sp104;aa5908}@imperial.ac.uk

Maja Pantic

Imperial College London / Univ. of Twente
Dept. of Computing, UK / EEMCS, Netherlands
m.pantic@imperial.ac.uk

ABSTRACT

In this study, a system that discriminates laughter from speech by modelling the 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 features mapping for both classes. Classification of a new frame is performed via prediction. All the networks produce a prediction of the expected audio / visual features and the network with the best prediction, i.e., the model which best describes the audiovisual feature relationship, provides its label to the input frame. When trained on a simple dataset and tested on a hard dataset, the proposed approach outperforms audiovisual feature-level fusion, resulting in a 10.9% and 6.4% absolute increase in the F1 rate for laughter and classification rate, respectively. This indicates that classification based on prediction can produce a good model even when the available dataset is not challenging enough.

Index Terms— laughter-vs-speech discrimination, audio-visual speech / laughter feature relationship, prediction-based classification

1. INTRODUCTION

Recently, few efforts have been reported aiming to discriminate laughter from speech combining audio and visual information [1, 2, 3]. These works use either feature-level fusion, where audio and visual features are concatenated and then fed to a classifier, or decision-level fusion where the outputs of the audio- and video-only classifiers are fused. In the former case, the correlation between audio and visual features is taken into account by the classifier whereas in the latter case the correlation is lost. In this study, we present a system that discriminates laughter from speech by explicitly modelling the relationship between audio and visual features and based on the reasonable assumption that this relationship is different between speech and laughter.

There has been a lot of research in examining the relationship between acoustic and visual speech features [4, 5, 6]. 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 [4, 6], although measures as high as 0.8 have been reported [5] and 0.85 when nonlinear models are used, like neural networks (NNs), [4]. 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 are different in speech and laughter.

Driven by those results we would like to build a system that uses this difference in correlation to discriminate between laughter and speech. Towards this direction we train four NNs, which learn the audio-to-visual and visual-to-audio feature mappings for speech and laughter. It is expected that laughter networks will produce a better prediction than speech networks when the input is laughter, since they have learnt the audiovisual feature relationship for laughter, and vice versa. When a new frame comes then its audio and visual features are fed to all 4 networks, and the network which produces the best audio and visual feature prediction is the winner in the video-to-audio and audio-to-video case, respectively. The audio-to-video and video-to-audio mapping systems can be combined in order to take advantage of the bidirectional relationship between audio and visual features (see Section 4). The input frame is labelled based on the winner network. In other words, the frame is labelled based on the network / 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 present study is inspired by the memory-prediction framework [7]. The key idea is that an audio input can make a prediction for an expected visual input and vice versa. Our implementation is much simpler and very different from the proposed framework, but it is based roughly on the same idea. The networks make an audio prediction based on video, i.e., they predict what they expect to "hear" based on what they "see", and a video prediction based on audio, i.e., they predict what they expect to "see" based on what they "hear". Then the winner is the network with the best prediction. Depending on which class the winner network belongs to, laughter or speech, the input frame is labelled accordingly.

The proposed approach is compared to feature-level audiovisual fusion [1] on cross database experiments using one challenging dataset, AMI and one easy dataset, SAL. Both systems perform similarly when trained on AMI, however when trained on SAL the proposed system outperforms the feature-level fusion, leading to a 10.9% and 6.4% absolute increase in F1 rate for laughter and classification rate, respectively. 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.

2. DATABASES

AMI: We used the AMI Meeting Corpus [8] where people show a huge variety of spontaneous expressions. We only used the close-up video recordings of the subject's face and the related individual headset audio recordings. For our experiments we used seven meetings (IB4001 to IB4011) and the relevant recordings of ten participants. Laughter segments were selected based on the annotations provided with the AMI Corpus and those that co-occur with speech or laughter is not clearly audible were discarded. Speech segments were also determined by the annotations provided with the AMI Corpus. In total, we used 124 audio-visual laughter segments (149.1 sec), and 154 audio-visual speech segments (290.2 sec).

SAL: The Sensitive Artificial Listerner (SAL) technique as described in [9] "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 in total. We used the close-up video recordings of the subjects face and the related audio recording. In total, we used 94 audio-visual laughter segments (139.7 sec) and 177 audio-visual speech segments (382.8 sec).

3. FEATURES

Audio Features: Cepstral features, such as MFCCs, have been widely used in speech recognition and have also been successfully used for laughter detection [10]. In addition, it has been shown that cepstral coefficients are more correlated to visual features than prosodic features [5]. Only the first 6 MFCCs are used, given the findings in [10], and they are computed every 10ms over a window of 40ms, i.e. the frame rate is 100 frames per second (fps).

Visual Features: Changes in facial expression are captured by tracking 20 facial points. 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) [1]. 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 prinicpal component analysis to the matrix of these K vectors, head movement can be decoupled from facial expression. Using the approach proposed in [11], the facial expression movements are encoded by the projection of the tracking points coordinates to the N principal components of the PDM which correspond to facial expressions. N depends on the diversity of the dataset, so for AMI which contains large head movements and a lot of different facial expressions $N = 4$, whereas for SAL which is less diverse $N = 3$. These 4 or 3 visual features are extracted at the video frame rate, i.e., 25 fps. Further details of the feature extraction procedure can be found in [1, 2].

4. METHODOLOGY

For both speech and laughter we train two NNs, one that learns the audio-to-visual features mapping and one that learns the visual-to-audio features mapping. Training is done per frame and the goal is to minimize the error between the actual and the predicted visual/audio features. In other words, the first / second network takes as inputs the audio / visual features per frame and predicts the corresponding visual / audio features at the same frame. Therefore the relationship between the audio (A^L, A^S) and visual (V^L, V^S) features in speech and laughter is modelled by (NN^L_{AV}), (NN^L_{VA}) for laughter and (NN^S_{AV}), (NN^S_{VA}) for speech.

$$NN^L_{AV} : f^L_{AV}(A^L) = \hat{V}^L \approx V^L \quad (1)$$

$$NN^L_{VA} : f^L_{VA}(V^L) = \hat{A}^L \approx A^L \quad (2)$$

$$NN^S_{AV} : f^S_{AV}(A^S) = \hat{V}^S \approx V^S \quad (3)$$

$$NN^S_{VA} : f^S_{VA}(V^S) = \hat{A}^S \approx A^S \quad (4)$$

Once training is complete and the mapping functions (f^L, f^S) are learned then classification is performed based on the network that produces the lowest prediction error. When a new frame is available the audio and visual features are computed and then they are fed to all networks from eq. 1 - 4, and 4 errors are produced, eq. 5 - 8. The error metric used is the mean squared error (MSE). Then we can also combine the errors in order to generate a new error which takes into account the bidirectional relationship of audio and visual features as shown in eq. 9 and 10, where w is a weighting factor.

$$e^L_{AV} = MSE(\hat{V}^L, V^L) \quad (5)$$

$$e^L_{VA} = MSE(\hat{A}^L, A^L) \quad (6)$$

$$e^S_{AV} = MSE(\hat{V}^S, V^S) \quad (7)$$

$$e^S_{VA} = MSE(\hat{A}^S, A^S) \quad (8)$$

$$e^L = w \times e^L_{AV} + (1 - w) \times e^L_{VA} \quad (9)$$

$$e^S = w \times e^S_{AV} + (1 - w) \times e^S_{VA} \quad (10)$$

For the audio-to-video system a frame is labelled as laughter or speech depending on which network produced the best

estimate, i.e., the lowest prediction error, eq. 5, 7. The same prinicpal applies for the video-to-audio and the combined systems. In other words, a frame is assigned based on the following three rules:

$$\textbf{A}\rightarrow\textbf{V:}\ IF\ e^S_{AV} \geq e^L_{AV}\ THEN\ \textbf{L}\ ELSE\ \textbf{S} \quad (11)$$

$$\textbf{V}\rightarrow\textbf{A:}\ IF\ e^S_{VA} \geq e^L_{VA}\ THEN\ \textbf{L}\ ELSE\ \textbf{S} \quad (12)$$

$$\textbf{A}\rightarrow\textbf{V}+\textbf{V}\rightarrow\textbf{A:}\ IF\ e^S \geq e^L\ THEN\ \textbf{L}\ ELSE\ \textbf{S} \quad (13)$$

5. EXPERIMENTAL STUDIES

In order to assess the performance of the method presented in section 4, cross database experiments between AMI and SAL were performed. AMI is a challenging dataset since the subjects rarely have a frontal view and there are large head movements. On the other hand, SAL is an easy dataset since subjects almost always look straight at the camera and there are only small head movements. In the first experiment, a system is trained on AMI and tested on SAL and in the second one it is trained on SAL and tested on AMI.

As mentioned in section 3, 4 and 3 visual features are used when training on AMI and SAL, respectively. In both cases 6 audio features (MFCCs) are used. Before training, the audio and visual features are synchronised by upsampling the visual features, to match the frame rate of the audio features, by linear interpolation. All the audio and visual features are z-normalized per subject, to a zero mean and unity standard deviation. Subject normalisation helps removing subject and recording variability.

Following the approach of section 4, 4 NNs are trained, eq. 1 - 4, and the frames of each sequence are labelled using rules 11 - 13. Then the majority of the frame labels is assigned to the sequence. The NNs used in this study have one hidden layer with 15 neurons, using sigmoid activation funtions, and they are trained for 100 epochs. For comparison we also report the results of an audiovisual feature-level fusion approach based on NNs [1, 12]. 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 binary, labelling each frame as either speech or laughter. Again, the majority of the frame labels is assigned to the sequence. It has been shown that this approach can outperform Coupled Hidden Markov Models for discriminating laughter from speech [12].

Since NNs, which are initialised 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 the same audio / visual features are used, and the same classification protocol is followed. The only difference is how

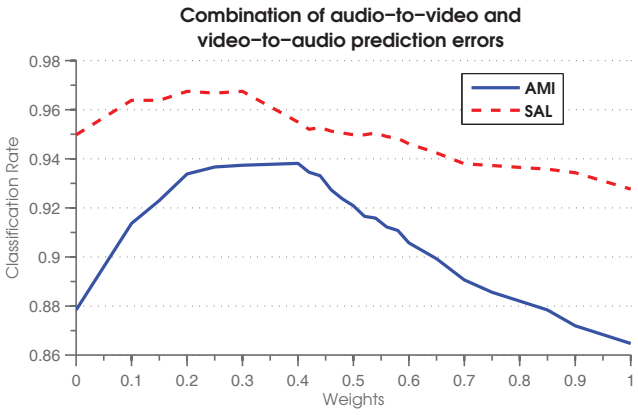

Fig. 1. Classification rate for laughter-vs-speech discrimination plotted for different weights used to combine the errors from eq. 9 and 10. Classification is performed using eq. 13. The solid / dashed line shows the classification rate for AMI / SAL. Both results were computed using a subject-independent cross validation within each dataset, i.e. cross validation is performed on AMI and SAL separately using examples only from one dataset at a time.

classification is performed, in the first approach via prediction and in the second case using the standard feature-level fusion.

By running two subject-independent cross validation experiments, one using the AMI dataset only and one using the SAL dataset only, we find the optimal weight, w, which is used in eq. 9 and 10 to combine the prediction errors from the two systems. From Fig. 1 we see that the optimal weight for AMI is 0.4 and for SAL 0.3. This means that in both cases the system which predicts the audio feature values is weighted more. So for the cross database experiments the weight is set to 0.4 and 0.3 when training on AMI and SAL respectively.

Table 1 shows the performance for each system. For the first experiment (train AMI $\rightarrow$ test SAL) we see that feature-level fusion outperforms both the audio-to-video and video-to-audio prediction systems. However, the difference is marginal when compared to the combination of the two prediction systems (max absolute difference: 0.3%). For the second experiment (train SAL $\rightarrow$ test AMI) feature-level fusion is much inferior to the combination of the two prediction systems. The absolute difference is 10.9%, 4.3% and 6.4% for F1 Laughter, F1 Speech and classification rate, respectively. Even the visual-to-audio prediction system performs better than feature-level fusion. Only the performance of the audio-to-video system is comparable, although still the F1 obtained for laughter is much lower in feature-level fusion.

Overall, we see that when we train on a challenging dataset (AMI) then both feature-level fusion and the prediction system lead to similar performance. But when we train on a less challenging dataset (SAL) then the prediction system is able to generalize much better on an unseen difficult dataset than feature-level fusion. This remark can be of great

Table 1. F1 and classification rates (CR) for the feature-level fusion (FF) system and the prediction based system on cross database experiments

Classification System	F1 Laughter	F1 Speech	CR
Train AMI → Test SAL			
A + V (FF)	95.4	97.6	96.8
A→V Pred.	88.0	93.1	91.2
V→A Pred.	92.2	95.8	94.5
A→V + V→A Pred.	95.2	97.3	96.5
Train SAL → Test AMI			
A + V (FF)	65.4	81.4	75.8
A→V Pred.	70.0	80.3	76.2
V→A Pred.	72.2	83.0	78.9
A→V + V→A Pred.	76.3	85.7	82.2

practical importance since it implies that in order to train a good system we do not depend so much on the available dataset. However, further experiments are needed in order to verify this claim.

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 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.

The performance of the combination of the two prediction system is better than the individual system. This was expected since we take into account the bidirectional relationship between audio and visual features. It is also interesting that the audio-to-video prediction system is worse than the video-to-audio system. This was obvious also in the case of cross validation within the two datasets, since the weight for the former system is lower than then for the second one. This might imply that the video-to-audio relationship is more different between laughter and speech than the video-to-audio relationship. However, this is an issue that requires further investigation.

6. CONCLUSIONS

A new classification approach based on prediction was presented for the problem of audiovisual laughter-vs-speech discrimination. This approach outperforms feature-level fusion when both are trained on a simple dataset and tested on a hard dataset which indicates that classification based on prediction can produce a good model even when the available dataset is not challenging enough. Training and testing was performed frame-by-frame resulting in a memoryless system, in this proof-of-concept work. Therefore, in future work we aim to include memory in the system which is expected to further benefit the system's performance.

7. ACKNOWLEDGEMENTS

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).

8. REFERENCES

[1] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," in *Proc. ACM CIVR*, 2008, pp. 329–337.

[2] S. Petridis and M. Pantic, "Audiovisual laughter detection based on temporal features," in *Proc. ACM ICMI*, 2008, pp. 37–44.

[3] B. Reuderink, M. Poel, K. Truong, R. Poppe, M. Pantic, "Decision-level fusion for audio-visual laughter detection," *LNCS*, 2008, vol. 5237, pp. 137 - 148.

[4] 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.

[5] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 8, pp. 2331–2347, 2007.

[6] 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.

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

[8] 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.

[9] 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 *Workshop on Corpora for Research on Emotion and Affect*, 2008, pp. 1–4.

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

[11] D. G. Jimenez and J. L. A. 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.

[12] 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.