Comparison of Single-model and Multiple-model Prediction-based Audiovisual Fusion

Stavros Petridis[1], *Varun Rajgarhia*[1], *Maja Pantic*[1,2]

[1]Dept. Computing, Imperial College London
[2]EEMCS, University of Twente

stavros.petridis04@imperial.ac.uk, m.pantic@imperial.ac.uk

Abstract

Prediction-based fusion is a recently proposed audiovisual fusion approach which outperforms feature-level fusion on laughter-vs-speech discrimination. One set of predictive models is trained per class which learns the audio-to-visual and visual-to-audio feature mapping together with the time evolution of audio and visual features. Classification of a new input is performed via prediction. All the class predictors 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. In all the previous works, a single set of predictors was trained on the entire training set for each class. In this work, we investigate the use of multiple sets of predictors per class. The main idea is that since models are trained on clusters of data, they will be more specialised and they will produce lower prediction errors which can in turn enhance the classification performance. We experimented with subject-based clustering and clustering based on different types of laughter, voiced and unvoiced. Results are presented on laughter-vs-speech discrimination on a cross-database experiment using the AMI and MAHNOB databases. The use of multiple sets of models results in a significant performance increase with the latter clustering approach achieving the best performance. Overall, an increase of over 4% and 10% is observed for F1 speech and laughter, respectively, for both datasets.

Index Terms: Prediction-based fusion, Audiovisual fusion, Nonlinguistic Information Processing

1. Introduction

Audiovisual fusion has recently attracted a lot of attention and has been successfully applied to several problems like speech recognition [1, 2, 3], affect recognition [4] and laughter recognition [5, 6]. However, the optimal fusion type remains an open issue and largely depends on the problem. The most common types of audiovisual fusion are feature-level fusion, where the audio and visual features are concatenated and fed to a classifier, and decision-level fusion, where each modality is modeled independently and the decisions of the classifiers are combined. A new type of audiovisual fusion has been recently presented, prediction-based fusion [7], [8], [9], [10] which consistently outperforms feature-level fusion for laughter-vs-speech classification and non-linguistic vocalisation classification.

Prediction-based fusion is based on the idea that the relationship between audio and visual features is different in each class. This is achieved by explicitly modelling the spatial relationship between audio and visual features using predictive models which learn the audio-to-visual and visual-to-audio feature mapping for each class. Similarly, we model the temporal evolution of the audio and visual feature using predictors which learn the relationship between past and future values for audio and visual features for each class. It is expected that during testing the models which correspond to the actual class will produce a better prediction than all the other models since they have learned the audiovisual relationship for that class. Classification is performed by combining all the prediction errors per class and selecting the model that produces the lowest error. 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 models prediction.

In all the previous studies [7, 8, 9, 10] a single set of predictors was trained on the entire dataset for each class, i.e., laughter or speech. One drawback of this approach is that if the examples vary a lot within each class then the performance may degrade since a single set of predictors will try to model the high class variability. In this study we aim to solve this problem by training multiple sets of predictors for each class. Some sort of time series clustering should be first performed in each class in order to create homogeneous clusters of laughter and speech. Then, predictors for each cluster can be trained. The main idea is that the predictors trained per cluster will be more specialised and therefore will more accurately model the audiovisual feature relationship. As a consequence, it is expected that they will produce lower prediction errors which in turn will lead to better classificaton performance.

Time series clustering is a challenging task especially when the time series have different lengths [11]. In this work, we take advantage of the natural clustering embedded in the data and we create clusters based on subjects and on different laughter types. In the former case, predictive models are trained per class for each subject separately. In the latter case, laughter is divided into voiced and unvoiced according to [12], and separate models are trained for each type using examples from all subjects.

In this study we aim to discriminate laughter and speech, since both events are audiovisual in nature. We perform cross-database experiments where the SAL dataset is used for training and validation and the AMI and MAHNOB datasets are used for testing. The use of multiple sets of models is beneficial outperforming the use of a single set of predictors for both clustering approaches. In particular, an increase of over 4% in the F1 measure for speech is observed for both datasets and both clustering approaches. Similarly, an increase of over 10% is observed for F1 laughter.

(a) (b) (c) (d)

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

(a) (b) (c) (d))

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

2. Databases

For the purpose of this study we used three datasets corresponding to 3 different scenarios as explained below. Examples for each dataset are shown in Fig. 1, 2 and 3.

AMI: In the AMI Meeting Corpus [13] 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) 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: In the SAL dataset [14] 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, out of which 10 are used for training and 5 for validation. 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 frames per sec) 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.

MAHNOB: In the MAHNOB database [15], [16], laughter was elicited by showing funny videos to subjects. In total there are 22 subjects, 12 males and 10 females, and a large variety of laughter types is present. The camera is fixed therefore subjects are mostly in frontal pose. Two audio streams are available, one from the camera microphone and one from the lapel microphone. In this study, we only used the camera microphone since the audio signal is noisier and poses a more challenging problem.

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 [13] and MAHNOB [15] datasets laughter episodes were selected based on the annotations provided and for the SAL dataset we manually annotated laughter episodes. Details of the four datasets are given in Table 1.

Table 1: Description of the datasets.

Type	No. Episodes / No. Subjects	Total Duration (sec)	Mean / Std (sec)
AMI			
Laughter	124 / 10	145.4	1.17 / 0.7
Speech	154 / 10	285.9	1.86 / 1.1
SAL - Training			
Laughter	57 / 10	80.6	1.4 / 0.8
Speech	96 / 10	204.3	2.1 / 0.8
SAL - Validation			
Laughter	37 / 5	50.3	1.4 / 0.7
Speech	81 / 5	159.3	2.0 / 0.8
MAHNOB			
Laughter	554 / 22	863.7	1.56 / 2.2
Speech	845 / 22	2430.9	2.88 / 2.3

3. 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 [17]. In addition, it has been shown that cepstral coefficients are more correlated to visual features than prosodic features [18]. 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 [17], 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 frames per second (fps).

Visual Features: Changes in facial expressions 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) [19] as shown in Fig. 1, 2 and 3. 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 [20], 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 the SAL dataset, so our shape features are the projection of the 20 points to the 3 PCs which were found to correspond to facial expressions (PCs 5 to 7). These 3 visual features, called shape parameters, are extracted at the video frame rate, i.e., 25 fps. Further details of the feature extraction procedure can be found in [20].

4. Prediction-based Fusion

The prediction-based fusion framework consists of cross-modal and intra-modal predictors for each class c, where c is either laughter (L) or speech (S). The cross-modal predictors model the relationship between the audio (A^c) and visual (V^c) features

(a) (b) (c) (d)

Figure 3: Example of tracking a laughter episode from the MAHNOB database.

of class c with two regressors, $f^c_{A\to V}$ and $f^c_{V\to A}$, respectively. The first (second) predictor takes as input the audio (visual) features and predicts the corresponding visual (audio) features at the same frame t as shown in the following equations:

$$f^c_{A\to V}(A^c[t-k^c_{\mathrm{AV}},t]) = \hat{V}^c_{A\to V}[t] \approx V^c[t] \quad (1)$$

$$f^c_{V\to A}(V^c[t-k^c_{\mathrm{VA}},t]) = \hat{A}^c_{V\to A}[t] \approx A^c[t] \quad (2)$$

In eq. 1 and 2, the size of the windows k^c_{AV} and k^c_{VA} depends on the mapping type and the modelled class.

The intra-modal predictors model the relationship between past and future audio and visual features in each class c with two regressors $f^c_{A\to A}$ and $f^c_{V\to V}$. The first (second) predictor takes as input the past audio (visual) features and predicts the corresponding audio (visual) features at frame t as follows:

$$f^c_{A\to A}(A^c[t-k^c_{\mathrm{AA}},t-1]) = \hat{A}^c_{A\to A}[t] \approx A^c[t] \quad (3)$$

$$f^c_{V\to V}(V^c[t-k^c_{\mathrm{VV}},t-1]) = \hat{V}^c_{V\to V}[t] \approx V^c[t] \quad (4)$$

In eq. 3 and 4, the size of the windows k^c_{AA} and k^c_{VV} depends on the mapping type and the modelled class.

Once training is complete and the predictors f^c are learnt, they can be used for classification. When a new sequence is available, the audio and visual features are computed, which are fed to all predictors defined by eq. 1 - 4 resulting in 4 prediction errors per frame for each class c. The prediction error measure we use is the mean squared error (MSE). The total error for each predictor is computed by summing the errors across all frames, N, resulting in 4 prediction errors per sequence for each class. The errors for the 4 predictors of class c are computed using eq. 5 to 8.

$$e^c_{A\to V} = \sum_{i=1}^{N} MSE(\hat{V}^c_{A\to V}[i], V[i]) \quad (5)$$

$$e^c_{V\to A} = \sum_{i=1}^{N} MSE(\hat{A}^c_{V\to A}[i], A[i]) \quad (6)$$

$$e^c_{A\to A} = \sum_{i=1}^{N} MSE(\hat{A}^c_{A\to A}[i], A[i]) \quad (7)$$

$$e^c_{V\to V} = \sum_{i=1}^{N} MSE(\hat{V}^c_{V\to V}[i], V[i]) \quad (8)$$

Then the two cross-modal prediction models (eq. 5, 6) are combined in order to take into account the bidirectional relationship of audio and visual features as shown in eq. 9 subject to constraint in eq. 10.

$$e^c_{\mathrm{CP}} = w^c_{\mathrm{AV}} \times e^c_{A\to V} + w^c_{\mathrm{VA}} \times e^c_{V\to A} \quad (9)$$

$$w^c_{\mathrm{AV}} + w^c_{\mathrm{VA}} = 1 \quad (10)$$

where e^c_{CP} is the total cross-modal prediction error and w^c_{AV} and w^c_{VA} are the weights of the cross-modal prediction components.

Similarly, the two temporal evolution models (eq. 7, eq. 8) are combined in order to take into account past-to-future relationship between audio and visual features as shown in eq. 11 subject to constraint in eq. 12.

$$e^c_{\mathrm{IP}} = w^c_{\mathrm{AA}} \times e^c_{A\to A} + w^c_{\mathrm{VV}} \times e^c_{V\to V} \quad (11)$$

$$w^c_{\mathrm{AA}} + w^c_{\mathrm{VV}} = 1 \quad (12)$$

where e^c_{IP} is the total intra-modal prediction error and w^c_{AA} and w^c_{VV} are the weights of the intra-model prediction components.

Finally, the prediction errors of the two components are combined as shown in eq. 13, subject to constraint in eq. 14, in order to merge information from the two prediction-based models.

$$e^c = w^c_{\mathrm{CP}} \times e^c_{\mathrm{CP}} + w^c_{\mathrm{IP}} \times e^c_{\mathrm{IP}} \quad (13)$$

$$w^c_{\mathrm{CP}} + w^c_{\mathrm{IP}} = 1 \quad (14)$$

where e^c is the total prediction error and w^c_{CP} and w^c_{IP} are the weights for the cross-modal prediction and intra-model prediction fusion components, respectively.

It is important to point out that all predictors are class-specific, since they learn the audiovisual features relationships for laughter and speech separately. The key idea is that the class-specific predictors which correspond to the true class of a new input sequence will produce a better estimation of the audio/visual features than models corresponding to other classes, since they have been trained on the audiovisual features of the target class.

4.1. Single-Model Fusion

In single-model prediction-based fusion 4 predictors (eq. 1 - 4) are trained for each class, laughter and speech, using the entire training set and their prediction errors are combined. A label is assigned to the input sequence based on the two errors from eq. 13, one for laughter (e^L) and one for speech (e^S). In other words, the class-specific model that best explains the audiovisual feature relationship, i.e., leads to the lowest prediction error, labels the new sequence accordingly, as shown in eq. 15.

$$PredictedClass = \arg\min_{c=L,S} e^c \quad (15)$$

4.2. Multiple-Model Fusion

In multiple-model prediction-based fusion we build specialised predictive models trained on clusters of training data. In contrast to single-model prediction-based fusion, where a set of 4 predictors is trained on the entire training set for each class, in this case we train a set of predictors for each cluster within

each class. Since time series clustering is a challenging task especially when the time series have different lengths [11], we decided to use two types of natural clustering that already exist in the data.

Subject-based clustering: In this case we make the reasonable assumption that the characteristics of the audiovisual feature relationship is distinct in each subject. This means that we can train predictors to model laughter and speech for each subject separately. So if there are K subjects, there will be K models for laughter and K models for speech. Classification is performed based on the model which produces the lowest prediction error as follows:

$$PredictedClass = \underset{c=L_1,..,L_K,S_1,..,S_K}{\arg\min} e^c \qquad (16)$$

In other words, an input example is labelled as laughter if one of the K laughter models leads to the lowest prediction error, otherwise it is labelled as speech.

Laughter-type-based clustering: In this case, we cluster the time series based on the laughter characteristics. A widely accepted categorisation of laughter is between voiced and unvoiced laughter [12]. The SAL dataset contains annotations for voiced and unvoiced laughter, therefore two laughter predictors are trained using data from all subjects. Similarly, one speech predictor is trained using speech examples from all subjects. Classification is performed as follows:

$$PredictedClass = \underset{c=L_V,L_U,S}{\arg\min} e^c \qquad (17)$$

An input example is labelled as laughter if either the unvoiced laughter (L_U) or voiced laughter (L_V) predictor leads to the lowest prediction error, otherwise it is labelled as speech.

5. Experimental Setup

Preprocessing: As mentioned in section 3, we used 3 visual features and 12 audio features in our experiments. Before training, the audio and visual features are synchronised 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 in order to remove subject and recording variabliy .

Parameter Optimization: Neural networks are used as regressors, hence the first step is the optimisation of the number of hidden neurons and the window lengths from eq. 1 to 4. We trained networks with only one hidden layer using resilient backpropagation. The number of hidden neurons varies between 5 and 60 neurons. The window lengths range is from $0\,\mathrm{ms}$ to $120\,\mathrm{ms}$, which is the length of the shortest vocalisation, in steps of $10\,\mathrm{ms}$. The combination of window length and number of hidden neurons that leads to the lowest prediction error over all sequences in the validation set is selected as the optimal one. It should be noted that the parameters of each network/predictor are optimised independently of the other networks.

The next step is the optimisation of the weights which is done hierarchically. In the first layer the weights of the cross-prediction module, w^c_{AV} and w^c_{VA}, and intra-prediction module, w^c_{VV} and w^c_{AA}, are optimised independently of each other. For each module a line search is performed between 0 to 1 in steps of 0.05 and classification is based either on eq. 9 or eq. 11. The weight combination in each module resulting in the best mean F1 measure over all classes on the validation set is selected as the optimal.

In the second layer, the weights that combine the cross-modal prediction, w^c_{CP}, and intra-modal prediction, w^c_{IP}, modules from eq. 13 are optimised. This is done in exactly the same way as in the first layer. The only difference is that the performance of the overall system is considered, i.e., classification is performed using either eq. 15 or 16 or 17 depending on the fusion type used.

We should also clarify that in the case of clustering based on the laughter type only one speech model is trained. However, this is paired with both the voiced and unvoiced laughter predictors in order to compute the weights from eq. 9, 11 and 17. This means that the same speech model is paired with the two laughter models but using a different set of weights in each case.

6. Results

In order to compare the performance of the methods presented in section 4, cross database experiments between SAL, AMI and MAHNOB were performed in order to discriminate laughter from speech. The first 10 subjects of the SAL dataset are used for training, the last 5 subjects of SAL are used as a validation set and the AMI and MAHNOB datasets are used for testing. The SAL dataset was used for training since it is the least diverse of all and therefore testing on the more diverse AMI and MAHNOB datasets is the most challenging scenario. For each experiment we use the recall and precision rates and the F1 measure as perfomance measures.

Results on the AMI and MAHNOB datasets are shown in Tables 2 and 3, respectively. Results of audio- and video-only classification using neural networks are also reported for completeness. Exactly the same patterns are observed for both datasets. The use of multiple predictive models trained per subject for each class significantly increases recall and decreases precision for laughter compared to a single set of models trained per class on the entire dataset. The opposite happens for speech, where recall decreases but precision significantly increases. Overall, the F1 measure for both laughter and speech significantly increases.

The same pattern is observed when predictive models are trained separately on voiced and unvoiced laughter. Recall for laughter and precision for speech increase even further whereas precision for laughter and recall for speech are reduced to a lesser degree. In other words what really happens when multiple models are trained on several clusters is that a significant number of additional laughter examples is correctly classified (that is why laughter recall and speech precision increases) whereas a small number of speech examples is misclassified (that is why laughter precision and speech recall decrease). As a consequence the F1 measures for both laughter and speech are further increased. In particular, an absolute increase of 12.6% and 4.9% is observed for F1 Laughter and Speech, respectively, on the AMI dataset. A similar increase of 10.7% and 4.3% for F1 Laughter and Speech, respectively, is observed on the MAHNOB dataset.

The significant increase in the overall performance confirms the assumption that training specialised sets of predictors on homogeneous clusters models more accurately the audiovisual feature relationship. The successful use of subject-based clustering reveals that there are differences in the audiovisual feature relationship in laughter and speech across subjects which cannot be modelled so accurately when a set of predictors

Table 2: Precision, Recall and F1 measure for laughter and speech when the single-model and multiple-model prediction-based fusion (PF) systems are tested on the AMI dataset. The performance of a classifier trained only on the audio or visual modality is reported for comparison purposes.

Recall Laughter	Precision Laughter	F1 Laughter	Recall Speech	Precision Speech	F1 Speech
Audio-only Classifier					
59.6	96.7	73.7	98.5	75.2	85.3
Video-only Classifier					
48.5	75.4	58.5	86.9	67.9	76.1
Single-model PF					
61.3	97.4	75.3	98.7	76.0	85.9
Multiple-model PF - Subject-based Clustering					
79.0	91.6	84.9	94.2	84.8	89.2
Multiple-model PF - Laughter-type-based Clustering					
80.7	94.3	87.0	96.1	86.1	90.8

Table 3: Precision, Recall and F1 measure for laughter and speech when the single-model and multiple-model prediction-based fusion (PF) systems are tested on the MAHNOB dataset. The performance of a classifier trained only on the audio or visual modality is reported for comparison purposes.

Recall Laughter	Precision Speech	F1 Laughter	Recall Speech	Precision Laughter	F1 Speech
Audio-only Classifier					
64.1	94.2	76.2	97.4	80.6	88.2
Video-only Classifier					
46.3	69.4	55.0	86.4	71.3	78.0
Single-model PF					
67.7	93.8	78.6	97.0	82.1	88.9
Multiple-model PF - Subject-based Clustering					
81.4	84.1	82.8	89.9	88.1	89.0
Multiple-model PF - Laughter-type-based Clustering					
87.7	91.0	89.3	94.3	92.1	93.2

is trained on all the subjects simultaneously. Similarly, the differences between voiced and unvoiced laughter are significant so the use of specialised sets of predictors is beneficial. In fact, the better performance of the second approach implies that the voiced and unvoiced clusters are probably more homogeneous that the clusters based on subjects. This is not unexpected, since all voiced laughs are quite different than unvoiced laughs even if expressed from different subjects. On the other hand, each subject produces both voiced and unvoiced laughs which are clustered together in the former approach and possibly leading to less homogeneous clusters. We should also take into account the fact that the sets of predictors are trained on fewer examples in the former case since there are more clusters, 20 compared to 3 in the latter case.

7. Conclusions

We have compared the standard prediction-based audiovisual fusion where only one set of predictors is trained per class on the entire training set to multiple-models prediction-based fusion where multiple models are trained on different clusters per class. We take advantage of the natural clustering embdedded in the data and we used two clustering approaches, subject-based clustering and clustering based on the laughter type, voiced or unvoiced. In the former case, a set of predictors for laughter and speech is trained per subject. In the latter case, 3 sets of predictors are trained, one for voiced laughter, one for unvoiced laughter and one for speech. Both approaches outperform the standard single-model prediction-based fusion. This confirms the hypothesis that when more specialised models are trained they can model the audiovisual feature relationship more accurately and in turn enhance the performance. An interesting direction for future work is the combination of the two natural clustering approaches, where voiced and unvoiced laughter clusters are defined within each subject. This approach has the potential to create even more homogeneous clusters and further enhance the performance. Finally, the use of time series clustering can reveal homogeneous clusters across subjects which can further improve the performance of prediction-based fusion.

8. Acknowledgment

This work has been funded by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). The work by S. Petridis is further supported by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA).

9. References

[1] 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.

[2] 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.

[3] I. Almajai and B. Milner, "Maximising audio-visual speech correlation," in *Audiovisual Speech Processing Conference*, 2007.

[4] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[5] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, April 2011.

[6] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm, "Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 1, pp. 4:1–4:31, Mar. 2012.

[7] S. Petridis, M. Pantic, and J. Cohn, "Prediction-based classification for audiovisual discrimination between laughter and speech," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, pp. 619–626.

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

[9] S. Petridis, S. Bilakhia, and M. Pantic, "Comparison of prediction-based fusion and feature-level fusion across different learning models," in *ACM Multimedia*, 2012, pp. 813–816.

[10] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing (accepted)*, 2015.

[11] T. W. Liao, "Clustering of time series data – A survey," *Pattern recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[12] J. Bachorowski and M. Owren, "Not All Laughs Are Alike: Voiced but Not Unvoiced Laughter Readily Elicits Positive Affect," *Psychological Science*, vol. 12, no. 3, pp. 252–257, 2001.

[13] 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.

[14] 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*, 2008, pp. 1–4.

[15] [online] http://mahnob-db.eu/laughter/.

[16] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image and Vision Computing Journal*, vol. 31, no. 2, pp. 186–202, 2013.

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

[18] 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.

[19] I. Patras and M. Pantic, "Particle filtering with factorized likelihoods for tracking facial features," in *FG*, 2004, pp. 97–104.

[20] 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.