Prediction-Based Audiovisual Fusion for Classification of Non-Linguistic Vocalisations

Stavros Petridis, *Member, IEEE* and Maja Pantic, *Fellow, IEEE*

Abstract—Prediction plays a key role in recent computational models of the brain and it has been suggested that the brain constantly makes multisensory spatiotemporal predictions. Inspired by these findings we tackle the problem of audiovisual fusion from a new perspective based on prediction. We train predictive models which model the spatiotemporal relationship between audio and visual features by learning the audio-to-visual and visual-to-audio feature mapping for each class. Similarly, we train predictive models which model the time evolution of audio and visual features by learning the past-to-future feature mapping for each class. In classification, all the class-specific regression models produce a prediction of the expected audio/visual features and their prediction errors are combined for each class. The set of class-specific regressors which best describes the audiovisual feature relationship, i.e., results in the lowest prediction error, is chosen to label the input frame. We perform cross-database experiments, using the AMI, SAL, and MAHNOB databases, in order to classify laughter and speech and subject-independent experiments on the AVIC database in order to classify laughter, hesitation and consent. In virtually all cases prediction-based audiovisual fusion consistently outperforms the two most commonly used fusion approaches, decision-level and feature-level fusion.

Index Terms—Prediction-based fusion, audiovisual fusion, nonlinguistic vocalisation classification

◆

1 INTRODUCTION

Audiovisual fusion approaches have been successfully applied to various problems like speech recognition [1], [2], affect recognition [3], [4], laughter recognition [5], [6], biometric systems [7] and meeting analysis [8]. Their success lie in the redundant visual information not corrupted by auditory noise, and to a lesser degree to the complementary visual information, which is not present in the auditory channel. Although various works on audiovisual fusion have been recently presented, feature-level fusion (FF) and decision-level fusion (DF) remain the two most common types [7], [9].

In this work, we present a new audiovisual fusion approach, which is based on prediction, tackling the problem from a different perspective. The proposed approach has been inspired by recent computational models of the brain [10], [11]. The memory-prediction framework [11] was of particular interest to us since it emphasises the notion of multisensory spatiotemporal predictions. Based on the input from one sense, e.g., vision, the brain can make predictions about future events in the same sense, as well as current and future events in other senses, e.g., hearing. This means that based on what we see (hear) now we can predict what we expect to hear (see) now and see (hear) and hear (see) in the future.

Similar findings have been reported in psychology and cognitive science. It has been suggested in [12] that visual information has a predictive role in processing audio information. The audio signal and the mouth expression share common temporal properties [13] and this helps to reduce temporal uncertainty related to the onset of syllables. In other words, the mouth opening can be used to predict the acoustic envelope of the speech, which in turn reduces temporal uncertainty. This assumption has been experimentally tested [14], [15] and demonstrated to be valid. In [16], [17], it was shown that vision is used as a predictive signal and certain facial movements are better predictors of subsequently voiced speech than others.

Driven by those findings we propose a new audiovisual fusion approach based on prediction, which has received little attention so far. We explicitly model the spatiotemporal relationship between audio and visual features using regressors[1] which learn the audio-to-visual and visual-to-audio feature mapping for each class. This set of regressors learn to predict the audio features from the visual features and vice versa and constitute the cross-modal prediction fusion module as shown in Fig. 1. Similarly, we model the temporal evolution of the audio and visual features using regressors which learn the relationship between past and future values for the audio and visual features, respectively, for each class separately. These regressors learn to predict the current audio and visual features from their past values and constitute the intra-modal prediction fusion module as shown in Fig. 1.

In classification, each regressor produces a prediction error which is combined with the prediction errors of the other regressors from the same class in a hierarchical way as shown in Fig. 1. By selecting the model that produces the

---

- S. Petridis is with the Department of Computing, Imperial College London, London, United Kingdom. E-mail: stavros.petridis04@imperial.ac.uk.
- M. Pantic is with the Department of Computing, Imperial College London, London, United Kingdom, and the Department of Computer Science, University of Twente, Enschede, The Netherlands. E-mail: m.pantic@imperial.ac.uk.

Manuscript received 3 Feb. 2015; revised 19 May 2015; accepted 3 June 2015. Date of publication 16 June 2015; date of current version 7 Mar. 2016. Recommended for acceptance by B.W. Schuller. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2015.2446462

[1] The terms regressor, predictor and predictive model are used interchangeably in this paper.

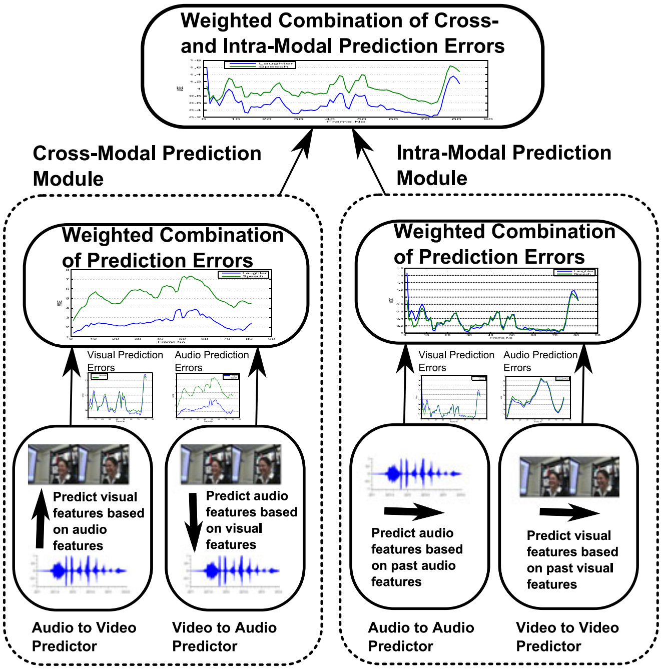

Weighted Combination of Cross- and Intra-Modal Prediction Errors
Cross-Modal Prediction Module
Intra-Modal Prediction Module
Weighted Combination of Prediction Errors
Weighted Combination of Prediction Errors
Visual Prediction Errors
Audio Prediction Errors
Visual Prediction Errors
Audio Prediction Errors
Predict visual features based on audio features
Predict audio features based on visual features
Predict audio features based on past audio features
Predict visual features based on past visual features
Audio to Video Predictor
Video to Audio Predictor
Audio to Audio Predictor
Video to Video Predictor

Fig. 1. Overview of the proposed prediction-based fusion. The first layer consists of the cross-modal prediction module, which models the relationships between audio and visual features, and the intra-modal prediction module, which models the temporal evolution of the audio and visual features separately. In the second layer the prediction errors of the two modules are combined. All four predictors are trained for each class separately. A sequence is classified based on the class-specific model which produces the lowest prediction error, i.e., best explains the audiovisual feature relationship. This example corresponds to a two-class problem and that is why there are two error curves in each module.

lowest prediction error, i.e., that best describes the audiovisual feature relationship, the presented input can be labelled accordingly. It is expected that the models corresponding to the actual class will produce a better prediction than all other models, since they have learnt the audiovisual relationship for that class. It does not matter if the absolute value of the prediction error is high or low, what really matters is the ranking of the errors. As long as the correct model produces the lowest error, the input example is correctly classified.

This study is an extension of our previous works [18], [19], [20], where we compared cross-modal prediction fusion with feature-level fusion for laughter-versus-speech discrimination. Here, we provide an extensive comparison of cross-modal prediction fusion, intra-modal prediction fusion, and their combination, with feature-level fusion, decision-level fusion, and their combination on two different problems laughter-versus-speech discrimination and classification of various nonlinguistic vocalisations.[2] We also compare the performance of various fusion approaches across different audio noise levels and provide some insight into the advantages of prediction-based fusion.

We have chosen nonlinguistic vocalisations as the target application since they are audiovisual in nature. Previous works have successfully used both decision-level and feature-level fusion to discriminate between laughter and speech [5], [6], [22], [23] and nonlinguistic vocalisations [24]. In all cases audiovisual fusion achieved higher classification performance over audio-only classifiers indicating that visual information is beneficial for such tasks.

We use the AMI, SAL, and MAHNOB databases to discriminate laughter from speech. We conduct cross-database experiments which pose a significant challenge due to the different recording conditions. Prediction-based fusion outperforms the standard fusion methods in virtually all cases. It results in a 4 and 5.4 percent increase in the mean F1 over all classes on the AMI and MAHNOB datasets, respectively. We also use the AVIC database in order to classify laughter, consent, hesitation and other human noises. In this case prediction-based fusion leads to an 8.3 percent increase in the mean F1 over all classes. We also repeat the same experiments while adding audio noise to the test sets. Again, prediction-based fusion outperforms decision-level and feature-level fusion in almost all noise levels.

Section 2 provides an overview of related works and Section 3 describes the proposed prediction-based fusion approach. In Sections 5 and 4 we present the databases and the audio and visual features, respectively. Section 6 describes the experimental setup and results are presented in Sections 7 and 8. Finally, Section 9 provides insight why the proposed fusion approach outperforms the standard fusion approaches and this is followed by Section 10 which concludes the paper.

2 RELATED WORK

2.1 Audiovisual Fusion

Multiple audiovisual fusion approaches have been proposed in the literature and have been applied to a variety of applications. In this section, we present the most popular fusion approaches: feature-level, classifier-level and decision-level fusion. A full review of existing audiovisual fusion approaches and applications can be found in [7], [9].

2.1.1 Feature-Level Fusion

The extracted audio and visual features are combined, usually by concatenation, and then fed to a classifier. Processing all features increases the dimensionality of the problem and makes the problem more complex since it requires a large amount of training data. One constraint of this approach is that once the classifier has been trained the relative weights of each stream cannot change as they are determined internally by the classifier. The main advantage of this type of fusion is that it takes into account the spatiotemporal relationship between the audio and visual features, i.e., the coevolution of the audiovisual features over time.

2.1.2 Classifier-Level Fusion (CF)

This type of fusion lies between feature-level and decision-level fusion. The audio and visual features are processed independently and fusion takes place in the classifier. Hence, this approach cannot be used with any classifier but only with specific types like hidden Markov models (HMMs) and Dynamic Bayesian Networks. Two of the most commonly used architectures in audiovisual speech recognition are the coupled HMMs [25] and multistream HMMs [1]. In the former case, two HMMs are used, one for the audio stream and

---

[2] According to Scherer [21] nonlinguistic vocalizations (or nonverbal vocalizations) are very brief, discrete, nonverbal expressions of affect in both face and voice, like laughter, sigh, hesitation etc.

one for the visual stream, which are coupled such that the next state in both streams depends on the current state of the audio and visual stream. In the latter case, two independent HMMs are used in parallel and their likelihoods are combined in pre-defined synchronisation points. Another variant is the multistream fused HMMs proposed in [26] for affect recognition. One HMM is created for the audio and visual streams and the hidden states are connected using the maximum mutual information criterion.

Since this fusion method is not generally applicable to any classifier but requires specific classifier architectures, we do not consider it further in this work. A review about other classifier-level fusion approaches can be found in [3], [27].

2.1.3 Decision Level Fusion

In this type of fusion the audio and video modalities are processed independently and they are combined at a higher level using various integration rules like the weighted sum. As a consequence the correlation between the audio and visual features is not taken into account. This fusion type does not increase the dimensionality, but comes at the expense of requiring multiple classifiers to be trained. It also allows for separate weighting of the different streams based on their reliability and the relative importance of the streams can be easily changed, by adjusting the weights, once the classifiers are trained.

2.2 Prediction-Based Approaches

In this section we review the most relevant works based on prediction. In order to be consistent with our approach we divide the works into those which predict ahead in time and those which make cross-modal predictions.

2.2.1 Prediction Ahead in Time

Predictive models which predict ahead in time have been mainly used for time series classification [28], [29] and to a lesser degree for time series segmentation [30]. The standard approach is that a predictive model per class is trained, usually either a feedforward or a recurrent neural network, which learns to predict the signal/feature values at the next time step. Classification is performed based on the model that produces the lowest prediction error similar to this work's approach.

The most common application for prediction-based approaches is gesture recognition. The prediction-error-classification approach [31], [32] has been proposed which builds predictive models based either on neuro-fuzzy predictors or continuous-time recurrent neural networks (NNs). These models learn to predict the acceleration values in the X, Y and Z axes in the next time step for eight different gestures. Classification is performed based on the model which produces the lowest prediction error over an entire gesture. A similar approach has been followed in [33]. A recurrent fuzzy network models the time evolution of the 2D coordinates on the image plane for each of the ten gestures considered and classification is performed based again on the lowest prediction error principle.

A variant of the above approach for object classification has been presented in [34] where only one recurrent neural network is used to model all time series. This work assumes that the time series are periodic and therefore they can be extended so they all have the same length. The use of a single model is possible because the context units, i.e., neurons that receive input from the feedback connections, are set to different initial values for each class. This means that instead of feeding a time series to all models, as described above, it is fed several times to the same model, where each time class-specific values for the context units are used.

In all the above approaches, the predictive models learn the evolution of raw signals, i.e., no features are extracted. The same prediction-based approach has been used in [35] where recurrent neural fuzzy networks model the time evolution of linear predictive coefficients extracted from the audio signal of birdsongs. In this case the task is birdsong classification, but classification is performed in exactly the same way.

Another application of the prediction-based approach has been feature extraction [36]. Feedforward neural networks are trained, using a window of past values, to perform one-step-ahead predictions for EEG time series. The mean squared errors (MSEs), over a window of the predicted values, of all the models are used as features for linear discriminant analysis.

It is obvious that in all the previous approaches prediction is performed in one stream only and audiovisual fusion has not been attempted. The main difference with our approach is that we create predictive models for the audio and visual streams and fusion occurs through the combination of their prediction errors.

2.2.2 Cross-Modal Prediction

Predictive models which make cross-modal predictions have been mainly used to examine the relationship between acoustic and visual speech features. Most of the studies are focused only on the audio-to-visual feature mapping. Linear predictors are commonly used to predict the visual features, usually facial points [37], [38] or distances between the facial points [39], [40], based on the audio features, usually line spectrum pairs [37], [38], [39], [40] or linear predictive coefficients [39], [40]. A correlation coefficient of about 0.70 between the predicted and actual visual features is reported in almost all studies, which increases to 0.85 when non-linear predictors are used like neural networks [37]. Similar conclusions have been drawn also for emotional speech where correlation values of over 0.80 have been reported when a linear predictor is used to estimate facial points based on mel frequency cepstral coefficients (MFCCs) [41]. It should be noted that all studies report results in a subject-dependent way, i.e., each predictor is tested on the same subject that has been trained on. As expected the correlation is weaker for subject-independent experiments [39].

There are also a few works which attempt to predict the audio features based on the visual features. The results are not as consistent in this case as in the audio-to-visual mapping. A correlation coefficient of 0.55 is reported in [40] whereas a correlation coefficient of 0.73 is reported in [38] using linear predictors.

Cross-modal prediction models have also been widely used in speech driven facial animation. In this case, the audio features, usually MFCCs [42], [43], are used as input to a non-linear predictor, usually a neural network [42], [43], [44]. The goal is to predict the visual features which in turn control the generated facial animations. The correlation coefficients reported vary significantly, from 0.64 [42], [45] to 0.96 [43], but this depends on the visual features used, which can simply be control parameters of the animated face [42].

Finally, to the best of our knowledge the only hybrid approach that combines intra-modal prediction with cross-modal prediction is presented in [46]. It is an interesting approach but it is used for synchrony detection in speech and not for audiovisual fusion and it makes no use of the prediction error. The time evolution of the audio features is modelled based on the assumption that the current audio features can be linearly predicted using past audio and visual features and the present visual features. It is expected that the visual features can be used to predict the audio features when they are synchronised, and therefore their correlation is higher, but not when they are asynchronous.

Based on the findings presented above it is obvious that there is a significant correlation between audio and visual features in speech. To the best of our knowledge there is no work which performs a similar correlation analysis for laughter and other nonlinguistic vocalisations. The closest work is [47] which attempts to produce facial animations based on the sound of laughing, crying, sneezing and yawning but without providing any correlation analysis. It is reasonable to assume that a correlation exists between audio and visual features in nonlinguistic vocalisations, although this may be weaker than in speech. Consequently, it makes sense to model audiovisual behaviour by models which predict the audio features from the visual features and vice versa. Yet, none of the prediction-based approaches have been used either for classification or fusion of audiovisual time series, as we propose in this work.

3 PREDICTION-BASED FUSION

The prediction-based fusion framework consists of two components as shown in Fig. 1. The first is the cross-modal prediction component, which combines the audio and visual features by modelling the spatiotemporal relationship between them. This component corresponds to feature-level fusion where the concatenation of audio and visual features is replaced by two predictors which learn the mapping between audio and visual features and vice versa for each class separately.

The second one is the intra-modal prediction component which models the temporal evolution of the audio and visual features separately. This component corresponds to decision-level fusion where each audio/visual stream is modelled by two predictors which learn the mapping between past and current audio or visual features for each class separately.

Finally, these two components are combined in a hierarchical fashion. In the first level, the two predictors of the cross-modal prediction component are combined in order to take into account the bidirectional relationship between audio and visual features. Similarly, the two predictors of the intra-modal prediction component are combined in order to merge the information about the temporal evolution of the audio and visual streams. In the second layer, the intra- and cross-modal prediction components are combined in order to benefit from both the audiovisual feature relationship and their temporal evolution. This corresponds to the combination of feature-level and decision-level fusion.

It is important to point out that all predictors are class-specific, since they learn the audiovisual features relationships for each class separately. This means that if there are C classes the number of predictors that should be trained is $4 \times C$. 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.

In the first set of predictors, which make predictions across modalities, the relationship between the audio (A^c) and visual (V^c) features of class c is modelled by 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:

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

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

In Eqs. (1) and (2), the size of the windows k^c_{AV} and k^c_{VA} depends on the mapping type and the modelled class. Note that the feature values at frame t are used as well in order to predict the feature values in the other modality at the same frame t.

In the second set of predictors, which make predictions within each modality, the relationship between past and future audio and visual features in each class c is modelled by 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:

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

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

In Eqs. (3) and (4), the size of the windows k^c_{AA} and k^c_{VV} depends on the mapping type and the modelled class. In this case the feature values at frame t are excluded since that is what we want to predict.

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 Eqs. (1) - (4) resulting in four prediction errors per frame for each class c. The prediction error measures we considered are the mean squared error, the mean absolute error (MAE) and the L^2 norm of the error (L^2-E). 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 Eqs. (5) to (8)

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

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

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

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

where Err is either the MSE or MAE or L^2-E. Then the two cross-modal prediction models (Eqs. (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).

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

$$w^c_{\mathrm{AV}} + w^c_{\mathrm{VA}} = 1, \tag{10}$$

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).

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

$$w^c_{\mathrm{AA}} + w^c_{\mathrm{VV}} = 1, \tag{12}$$

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.

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

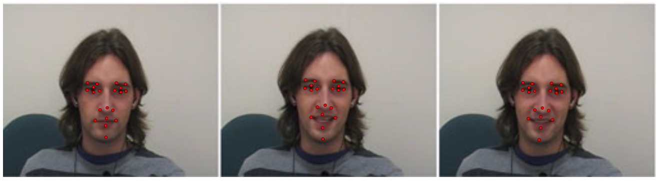
$$w^c_{\mathrm{CP}} + w^c_{\mathrm{IP}} = 1, \tag{14}$$

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. We have opted for combining the sub-systems in a hierarchical way since it allows for easier optimisation of the weights.

In Eqs. (9), (11), (13), the prediction errors are combined without being normalised first. It is expected that the errors will be in different scales since the predictors model different relationships. As a consequence, the weights indicate the relative importance of each predictor and act as scaling factors as well.

An alternative approach is to convert the prediction errors in the same scale by means of softmax normalisation. All errors in Eqs. (9), (11), (13) are normalised using the softmax function so their sum is equal to one. In this case, the weights simply indicate the relative importance of each predictor. In all the experiments conducted in this study, both softmax normalisation and no normalisation are considered.

In the final step, a label is assigned to the input sequence based on the C errors (Eq. (13)). This is done by selecting the label which corresponds to the lowest error. 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).



(a) 324 (b) 333 (c) 346

Fig. 2. Example of tracking a laughter episode from the MAHNOB database, Session S007-002, frames 324 to 346.

$$PredictedClass = \arg\min_{c=1\ldots C} e^c. \tag{15}$$

In case we wish to perform classification using either cross-modal prediction fusion or intra-modal prediction fusion only, this can be achieved by replacing the total prediction error e^c in Eq. (15) with either the cross-modal prediction error e^c_{CP} or the intra-modal prediction error e^c_{IP} from Eqs. (9) and (11), respectively.

4 FEATURES

4.1 Audio Features

Cepstral features, such as mel frequency cepstral coefficients, have been widely used in speech recognition [1], [2] and have also been successfully used for laughter detection [48] and laughter-vs-speech discrimination [5]. In addition, it has been shown that cepstral coefficients are more correlated to visual features than prosodic features [41]. Therefore we only use MFCCs for our experiments which were computed using the functions provided in [49].

The use of 13 MFCC coefficients is common in speech recognition, however, using 6 coefficients has been reported to lead to either the same or an improved performance in laughter detection [48] and language identification [50]. The same conclusion has been confirmed in one of our previous study where the performance of different number of coefficients was investigated through cross-validation in the AMI dataset [51]. Hence, we use 6 MFCCs which are computed every 10 ms over a 40 ms long frame, i.e., the frame rate is 100 fps. In addition, the ΔMFCCs are calculated as well since they capture local temporal characteristics. So in total, 12 audio features are computed.

4.2 Visual Features

To capture face movements in an input video, we track 20 facial points, as shown in Fig. 2. These points are the corners/ extremities of the eyebrows (2 points on each eyebrow), the eyes (4 points on each eye), the nose (3 points), the mouth (4 points), and the chin (1 point). To track these facial points we used the particle filtering tracking scheme proposed in [52], applied to tracking color-based templates centered around the facial points to be tracked. Hence, for each episode containing K video frames, we obtain a $K \times 40$ matrix which contains the x and y coordinates of the 20 points in each frame.

We wish to decouple rigid head movements from non-rigid movements, i.e., facial expressions, since we are mostly interested in the latter. To do so, we use a similar

TABLE 1
Description of the Four Datasets Used in This Study

Type	No. Episodes / No. Subjects	Total Duration (sec)	Mean / Std (sec)
AMI (25 fps, 720 × 576, 16 kHz)			
Laughter	124 / 10	145.36	1.17 / 0.73
Speech	154 / 10	285.92	1.86 / 1.12
SAL (25 fps, 720 × 576, 48 kHz)			
Laughter	94 / 15	136.96	1.46 / 0.78
Speech	177 / 15	377.32	2.13 / 0.80
MAHNOB (25 fps, 720 × 576, 48 kHz)			
Laughter	554 / 22	863.68	1.56 / 2.21
Speech	845 / 22	2430.92	2.88 / 2.28
AVIC (25 fps, 720 × 576, 44.1 kHz)			
Laughter	267 / 21	110.44	0.41 / 0.30
Hesitation	1136 / 21	356.96	0.31 / 0.16
Consent	308 / 18	80.88	0.26 / 0.11
Garbage	582 / 21	134.72	0.23 / 0.15

The frame rate in frames per second (fps), resolution and sample rate in kHz are shown next to each database.

approach to the one proposed by Gonzalez-Jimenez and Alba-Castro [53]. Using a point distribution model (PDM), by applying principal component (PCs) analysis to the matrix containing the point coordinates from the training frames, head movement can be decoupled from facial expression. 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. In this study we build a PDM based on the SAL training set, so our shape features are the projection of the 20 points to the three PCs which were found to correspond to facial expressions (PCs 5 to 7) [5]. Similarly, another PDM is built using the training set of AVIC using PCs 5 to 10, which correspond to facial expressions. These three and six visual features, are extracted at the video frame rate, i.e., 25 fps.

5 DATABASES

For the purpose of this study we use four databases corresponding to four different scenarios as described below. Details of the examples used in this study are given in Table 1.

Augmented multi-party interaction (AMI) corpus [54]. The AMI meeting corpus is a multi-modal database consisting of 100 hours of meeting recordings. In each meeting there are four participants which interact with each other. All meetings are held in English, although most of the subjects are non-native English speakers.

We use the same set of speech and laughter episodes used in [5] which can be found in [55]. Each participant is recorded by one camera positioned at a fixed location on the meeting table. Since subjects participate in a meeting they are rarely in a frontal pose. Audio for each participant is captured by a headset microphone and background noise is present from the other subjects.

Sensitive artificial listener (SAL) [56]. In this corpus subjects interact with four agents, which have different personalities and they are controlled by a human operator. The aim is to evoke emotionally coloured reactions from the users whose reactions are recorded by a camera and a microphone.

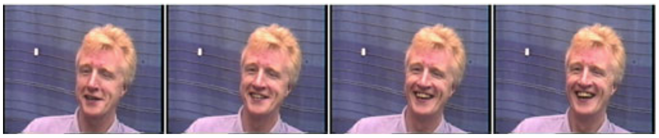

(a) 1449 (b) 1454 (c) 1464 (d) 1475

Fig. 3. Example of laughter from the SAL database (GHillSect3), frames 1449 to 1475.

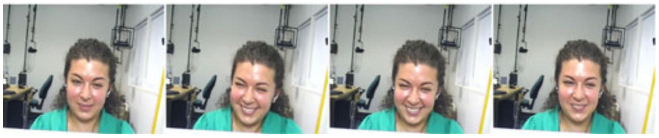

(a) 2104 (b) 2123 (c) 2163 (d) 2202

Fig. 4. Example of laughter from the MAHNOB database, Session S023-002, frames 2104 to 2202.

We use the same set of speech and laughter episodes used in [5] (see [55]) and most subjects are native English speakers. We use the close-up video recordings of the subjects and the related audio recording. Most of the time the subjects have frontal pose, head movements are small and audio noise is low. An example of a laughter episode is shown in Fig. 3.

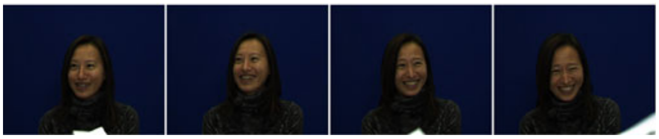
MAHNOB laughter database [57], [58]. In the MAHNOB laughter database funny video clips were shown to subjects and their reactions were recorded by two microphones, and a video camera. The subjects were also asked to speak about a topic of their choice in English and in their mother language.

We use the same set of speech and laughter examples as in [58] which can be found in [59], and most subjects are non-native English speakers. Each subject is recorded by a fixed camera and since subjects watch video clips they are mostly in frontal position and head movements are small except during intense laughter. The camera microphone audio is only considered since it is poses a more challenging problem as explained in [58]. An example laughter episode is shown in Fig. 4.

AudioVisual interest corpus (AVIC) [60]: The AVIC corpus is an audiovisual dataset containing scenario-based dyadic interactions. A subject is interacting with an experimenter who plays the role of a product presenter and leads the subject through a commercial presentation. The subjects role is to listen to the presentation and interact with the experimenter depending on his/her interest on the product.

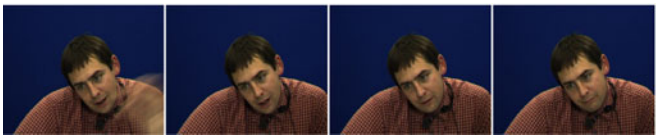
Annotations for laughter, hesitation, consent and other human noises, which are grouped into one class called garbage, are provided with the database and those are used in this study. Similarly to previous works [24], [60], [61] vocalisations that were very short (≤ 120 ms) were excluded.

A video camera was used to record the subject's reaction, positioned in front of him/her and the audio signal was recorded by a lapel microphone. The audio noise is low, head movement is moderate and most of the time subjects have frontal pose. Examples of laughter and hesitation are shown in Figs. 5 and 6, respectively.



(a) 6532 (b) 6539 (c) 6555 (d) 6569

Fig. 5. Example of laughter from the AVIC corpus, Subject VP4, frames 6532 to 6569.



(a) 15476 (b) 15479 (c) 15489 (d) 15497

Fig. 6. Example of hesitation from the AVIC corpus, Subject VP8, frames 15476 to 15497.

6 EXPERIMENTAL SETUP

6.1 Decision-Level and Feature-Level Fusion

In this section we briefly present the two most common fusion types, decision-level and feature-level fusion, which are compared with the prediction-based fusion approach.

Decision-level fusion. In this type of fusion, one classifier is trained with the audio features and one with the visual features. In both cases, a window of past audio, k_A, or visual features, k_V, is fed to the classifiers which produce for each class, $c = 1...C$, a score per frame, s_A^c or s_V^c. The class scores of the audio and visual systems are combined as shown in Eq. (16) subject to constraint Eq. (17).

$$s_{\mathrm{DF}}^c = w_A \times s_A^c + w_V \times s_V^c \tag{16}$$

$$w_A + w_V = 1, \tag{17}$$

where w_A and w_V are the weights of the audio and visual classifiers, respectively.

Feature-level fusion. In this type of fusion, the audio and visual features are first concatenated and then a single classifier is trained. A window of past audiovisual features, k_{FF}, is fed to the classifier which produces a score for each class c and each frame, s_{FF}^c.

Feature-level + decision-level fusion. As shown in Fig. 1 the cross-modal prediction and intra-modal prediction components are combined hierarchically in order to merge the different types of information they encode. In the same way, decision-level and feature-level fusion can also be combined hierarchically. This is achieved by combining their outputs as shown in Eq. (18) subject to constraint Eq. (19).

$$s_{\mathrm{FF+DF}}^c = w_{FF} \times s_{\mathrm{FF}}^c + w_{\mathrm{DF}} \times s_{\mathrm{DF}}^c \tag{18}$$

$$w_{\mathrm{FF}} + w_{\mathrm{DF}} = 1, \tag{19}$$

where w_{FF} and w_{DF} are the weights for the feature-level and decision-level fusion systems.

In all the above cases, the total score per class over a sequence can be computed by summing the scores across all frames. Finally, a sequence is labelled based on the class with the highest score.

6.2 Preprocessing

As mentioned in Section 4 the audio and visual features are extracted at different frame rates. Therefore they need to be synchronised. This is achieved by upsampling the visual features, to match the frame rate of the audio features (100 fps), by linear interpolation similarly to [2]. In addition, since the recording conditions are different in each database, the features need to be normalised in order to remove (to some extent) the recording and subject variability. In order to do this, we follow the common approach in cross-database experiments [62], [63] where all audio and visual features are z-normalised per subject, to a zero mean and unity standard deviation.

Finally, the datasets are imbalanced as shown in Table 1 and this can significantly degrade the performance of discriminative classifiers [64]. Therefore, the training set is balanced by random downsampling when feature-level or decision-level fusion is used. In prediction-based fusion this is not required since each predictor is trained with examples from one class only.

6.3 Training

Feedforward neural networks with one hidden layer, using sigmoid activation functions, are used as classifiers in feature-level or decision-level fusion and as predictors in prediction-based fusion. In the former case the output layer consists of sigmoid activation functions and in the latter case of linear activation functions. Each network is trained using the resilient backpropagation algorithm [65] with an epoch limit of 1,000 and early stopping to avoid overfitting.

The NNs weights are initialised randomly and this can lead to slightly different performance each time a network is trained. This in combination with the random downsampling approach for balancing used in feature-level and decision-level fusion may lead to variable performance. In order to account for that randomness 30 networks are trained for each experiment and the mean performance and standard deviation are reported.

6.4 Parameter Optimisation

Prediction-based fusion. The first step is the optimisation of the number of hidden neurons in NNs and the window lengths from Eqs. (1) to (4). The number of hidden neurons varies between 5 and 60 neurons. The window lengths range is from 0 ms to 120 ms, which is the length of the shortest vocalisation, in steps of 10 ms. The combination of window length and number of hidden neurons that leads to the lowest prediction error (either MSE or MAE or $L^2 - E$) 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-modal prediction module, w_{AV}^c and w_{VA}^c, and intra-modal prediction module, w_{VV}^c and w_{AA}^c, 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_{CP}^c, and intra-modal prediction, w_{IP}^c, 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 Eqs. (13) and (15).

Decision-level fusion. The only parameters that need to be optimised are the number of hidden neurons and the window lengths, k_A and k_V, of the audio and visual classifiers, respectively. This is done in the same way as for the prediction-based fusion. The only difference is that the optimal combination is the one which maximises the mean F1 measure over all classes on the validation set. Weights, w_A and w_V from Eq. (16), are also optimised in the same way as the first layer weights in prediction-based fusion.

Feature-Level fusion. In this case, the only parameters to be optimised are the number of hidden neurons and the window length, k_{FF}. This is done in the same way as described above for the audio and video classifiers.

Feature-level + Decision-level Fusion: The only parameters that need optimisation are the weights w_{FF} and w_{DF} from Eq. (18). This is performed in the same way as for the second layer weights in prediction-based fusion.

6.5 Performance Measures

The performance measures used are the F1 measure, the classification rate (CR) and the unweighted average recall (UAR), which is simply the average of all recall rates per class. The use of F1 measure and UAR provide a more objective view of the performance over the CR which can be affected by imbalanced datasets.

In order to test the statistical significance of the results we use the randomization test [66] which has been shown to perform similarly to the commonly used T-test, when the normality assumption is met, but outperforms it when it is not met. In the following sections, whenever two methods are compared in terms of performance, a randomisation test is run. So unless mentioned otherwise, whenever one method performs better, this difference is statistically significant.

7 EXPERIMENTAL STUDIES

In order to compare the performance of prediction-based fusion with feature-level and decision-level fusion two sets of experiments are conducted. In all approaches, exactly the same audio/visual features are used, and the same classification protocol is followed. The only difference is how classification is performed, either via prediction or by using the standard feature-level and/or decision-level fusion.

7.1 Laughter-versus-Speech Discrimination

In this experiment, we use the SAL, AMI and MAHNOB databases in order to discriminate laughter from speech. The first 10 subjects of the SAL dataset are used for training, the last five subjects of SAL are used as a validation set and the AMI and MAHNOB datasets are used for testing. This cross-database setup presents a more challenging task since each database has different characteristics and the trained models and the optimal parameters found on one database will most likely be sub-optimal when tested on different databases.

TABLE 2
Optimal Weights for Prediction-Based Fusion, Feature-Level Fusion and Decision-Level Fusion

Prediction-Based Fusion Softmax Normalisation		Prediction-Based Fusion No Normalisation	
$[w_{VA}^L, w_{AV}^L]$	[0.80 0.20]	$[w_{VA}^L, w_{AV}^L]$	[0.90 0.10]
$[w_{VV}^L, w_{AA}^L]$	[0 1]	$[w_{VV}^L, w_{AA}^L]$	[0.10 0.90]
$[w_{VA}^S, w_{AV}^S]$	[1 0]	$[w_{VA}^S, w_{AV}^S]$	[0.85 0.15]
$[w_{VV}^S, w_{AA}^S]$	[0 1]	$[w_{VV}^S, w_{AA}^S]$	[0 1]
$[w_{CP}^L, w_{IP}^L]$	[0.55 0.45]	$[w_{CP}^L, w_{IP}^L]$	[0.30 0.70]
$[w_{CP}^S, w_{IP}^S]$	[0.65 0.35]	$[w_{CP}^S, w_{IP}^S]$	[0.30 0.70]
Decision-Level Fusion		Decision-Level Fusion + Feature-Level Fusion	
$[w_A, w_V]$	[0.90 0.10]	$[w_{FF}, w_{DF}]$	[0.05 0.95]

All parameters were optimised on five subjects from the SAL dataset.

The optimal weights are shown in Table 2. It is obvious that the video-to-audio and audio-to-audio prediction systems are heavily favoured for both classes and for both normalisation schemes with weights varying from 0.80 to 1. In case of decision-level fusion the audio classifier is heavily favoured.

Regarding the second layer weights, the cross-modal prediction module is favoured for both classes when softmax normalisation of the errors is used. In other words, the cross-modal prediction module is more important when the errors are in the same scale. On the other hand, the intra-modal prediction weights are higher when no error normalisation is applied. As explained in section 3 the weights in this case encode both relative importance and scaling information. Hence, this means that the cross-modal prediction errors are higher than the intra-modal prediction errors and a smaller weight is needed in order to convert them to the same scale as the intra-modal prediction errors.

Table 3 shows the performance of the different approaches on the AMI and MAHNOB datasets. On the AMI dataset, the full prediction-based and cross-modal prediction fusion with no normalisation perform similarly and they are the best approaches for all performance measures. They achieve an absolute increase over the combination of decision- and feature-level fusion of up to 4.2% on the mean F1. It is worth pointing out that although intra-modal prediction fusion does not perform well, the full prediction-based system capitalises on the good performance of the cross-modal prediction system and the overall performance is not degraded. This happens because the correct class predictor produces a much lower prediction error than the wrong class predictor in the cross-modal prediction system and this difference is larger than the difference between the wrong class and the correct class prediction errors in the intra-modal prediction system.

On the MAHNOB dataset, the full prediction-based fusion approach independent of the normalisation scheme results in the best performance for all performance measures. It achieves an absolute increase over the combination of decision-level and feature-level fusion of up to 5.4% on the mean F1. In this case, both cross- and intra-modal prediction fusion approaches perform well so their combination results in improved performance.

TABLE 3
F1, UAR and CR for the Audio-Only Classifier (A), Video-Only Classifier (V), Feature-Level Fusion, D F, the Combination of DF and FF, Cross-Modal Prediction Fusion (C P), Intra-Modal Prediction Fusion (I P) and the Full Prediction-Based System with No Normalisation (P F - N) and Softmax Normalisation (P F - S)

Classification System	F1 Laughter	F1 Speech	F1 Mean	CR	UAR	F1 Laughter	F1 Speech	F1 Mean	CR	UAR
Test →			**AMI**					**MAHNOB**		
A	73.7 (3.4)	85.3 (1.4)	79.5 (2.4)	81.1 (2.0)	79.0 (2.2)	76.2 (3.3)	88.2 (1.1)	82.2 (2.2)	84.2 (1.7)	80.8 (2.2)
V	58.5 (5.2)	76.1 (1.0)	67.3 (2.8)	69.8 (1.7)	67.7 (2.2)	55.0 (5.6)	78.0 (1.0)	66.5 (3.0)	70.5 (1.8)	66.3 (2.9)
A + V (D F)	73.3 (3.2)	85.2 (1.3)	79.2 (2.3)	81.0 (1.9)	78.8 (2.1)	76.5 (3.4)	88.4 (1.2)	82.4 (2.3)	84.5 (1.8)	81.0 (2.3)
A + V (F F)	67.8 (2.8)	82.1 (1.1)	75.0 (1.9)	77.0 (1.6)	74.8 (1.7)	61.8 (2.5)	82.0 (0.7)	72.0 (1.5)	75.6 (1.1)	71.2 (1.4)
A + V (D F + F F)	73.5 (2.9)	85.4 (1.1)	79.5 (2.0)	81.2 (1.7)	79.0 (1.8)	76.5 (3.2)	88.4 (1.1)	82.5 (2.1)	84.5 (1.7)	81.0 (2.1)
A + V (C P - S)	76.6 (2.3)†	85.7 (1.0)†	81.2 (1.6)†	82.3 (1.4)	80.6 (1.5)†	81.7 (1.3)†	89.0 (0.6)	85.4 (1.0)†	86.3 (0.8)†	84.7 (1.1)†
A + V (C P - N)	**80.3 (2.5)**†	**87.0 (1.3)**†	**83.7 (1.9)**†	**84.3 (1.7)**†	**83.1 (1.9)**†	80.9 (2.4)†	88.3 (1.0)	84.6 (1.7)†	85.5 (1.5)	84.2 (1.9)†
A + V (I P - S)	62.3 (11.4)†	82.1 (3.2)†	72.2 (7.3)†	75.8 (5.3)†	73.0 (6.0)	73.7 (8.1)	87.6 (2.5)	80.7 (5.3)	83.2 (4.0)	79.5 (5.3)
A + V (I P - N)	68.2 (10.5)†	83.7 (3.1)†	76.0 (6.8)†	78.6 (5.1)†	76.1 (5.8)†	80.4 (7.3)†	89.8 (2.3)†	85.1 (4.8)†	86.6 (3.6)	84.1 (4.9)†
A + V (P F - S)	76.6 (1.9)†	86.2 (0.7)†	81.4 (1.3)†	82.6 (1.1)†	80.8 (1.2)†	**83.5 (1.2)**†	**90.4 (0.5)**†	**86.9 (0.8)**†	**87.8 (0.7)**†	**86.0 (1.0)**†
A + V (P F - N)	**79.4 (2.2)**†	**87.6 (1.0)**†	**83.5 (1.6)**†	**84.5 (1.4)**†	**82.9 (1.5)**†	**84.7 (2.2)**†	**91.1 (0.9)**†	**87.9 (1.6)**†	**88.7 (1.3)**†	**87.0 (1.7)**†

The AMI and MAHNOB datasets are used as test sets. The mean and (St. Dev.) over 30 experiments are presented. The two highest mean values in each column are given in bold. † denotes that the difference between the prediction-based approaches and D F + F F is statistically significant.

It is also obvious that none of the standard fusion methods outperforms the audio-only classification. This is due to the bad performance of the visual features which particulaty affects feature-level fusion.

We should also emphasise the different type of information encoded by the different fusion approaches. As shown in Table 2, prediction-based fusion is based on the one-way relationship between visual and audio features (video-to-audio predictor) and to a lesser degree on the temporal evolution of the audio features (audio-to-audio predictor). The combination of decision-level and feature-level fusion relies mostly on decision-level fusion which in turn is mainly based on the audio-only classifier, i.e., on the temporal evolution of the audio features. It is therefore obvious that prediction-based fusion offers a different representation of the audio and visual information and is also capable of extracting information which may not be easily accessible to standard fusion approaches.

It is also apparent that having no normalisation results in slightly better performance than softmax normalisation in most cases. Although softmax normalisation converts the errors in the same scale it distorts the difference between them achieving poorer performance results. On the other hand, having no normalisation does not introduce any distortion, and leads to slightly better performance, but it should be emphasised that the weights act as scaling factors as well and do not measure just the relative importance. For both types of normalisation the MAE led to the best performance on the validation set and that is the prediction error measure used in all experiments.

7.2 Non-linguistic Vocalization Classification

In this experiment, we use the AVIC database in order to classify 3 different non-linguistic vocalisations: laughter, hesitation, and consent, from a garbage class that contains other noises. Subjects 8 to 14 are used for training, subjects 15 to 21 are used for validation, and the first 7 subjects are used for testing.

The optimal weights for the non-linguistic vocalisation classification task are shown in Table 5. Similarly to laughter-vs-speech discrimination, the video-to-audio and the audio-to-audio predictors are clearly favoured for all classes and both types of normalisation. In case of decision-level fusion, the audio-only classifier is heavily favoured. However, the second layer weights follow a different pattern. The intra-modal prediction system is clearly favoured over the cross-modal prediction system independently of the normalisation used. This means that for this task the intra-modal prediction is more important. The intra-modal prediction weights are higher when no normalisation is applied, revealing also in this experiment that the cross-modal prediction errors tend to be higher than the intra-modal prediction errors.

Table 4 shows the performance of the different approaches on the AVIC dataset. On average, the intra-modal prediction module with softmax normalisation is the best performing approach achieving an absolute improvement of 8.3 percent on the mean F1 over the combination of decision-level and feature-level fusion. Similarly, all other prediction-based fusion approaches with the exception of cross-modal prediction fusion outperform all the standard fusion approaches.

In this experiment, cross-modal prediction fusion performs poorly because the audio and visual features are not highly correlated. For example, hesitation can be accompanied by either subtle facial expressions, like Fig. 6, or no facial expressions like Fig. 9. In other words, the facial expressions accompanying hesitation and consent are not as consistent as in the case of laughter or speech and as a consequence the audio and visual features are less correlated.

It is also clear from Table 4 that intra-modal prediction fusion and full prediction-based fusion perform similarly for most performance measures in case of no normalisation.

TABLE 4
F1 and UAR CR for the Audio-Only Classifier (A), Video-Only Classifier (V), Feature-Level Fusion, Decision-Level Fusion (D F), the Combination of DF and FF, Cross-Modal Prediction Fusion (C P), Intra-Modal Prediction Fusion (I P) and the Full Prediction-Based System with No Normalisation (P F - N) and Softmax Normalisation (P F - S)

Classification System	F1 Garbage	F1 Laughter	F1 Consent	F1 Hesitation	F1 Mean	CR	UAR
Test →				**AVIC**			
A	51.1 (3.8)	58.3 (2.6)	40.0 (5.2)	67.2 (2.8)	54.1 (2.2)	58.8 (2.4)	58.7 (2.4)
V	44.4 (4.1)	38.9 (2.6)	35.5 (3.4)	57.1 (3.7)	44.0 (2.0)	48.5 (2.6)	48.9 (2.5)
A + V (D F)	53.4 (3.9)	60.1 (2.4)	43.6 (5.5)	68.2 (2.8)	56.3 (2.2)	60.6 (2.4)	61.0 (2.3)
A + V (F F)	53.4 (2.5)	57.3 (2.4)	43.3 (2.8)	63.1 (3.1)	54.3 (1.8)	57.2 (2.3)	60.5 (1.6)
A + V (DF + FF)	54.3 (4.0)	60.5 (2.5)	44.8 (5.1)	68.4 (2.7)	57.0 (2.2)	61.1 (2.4)	61.8 (2.2)
A + V (C P - S)	38.9 (3.6)†	56.9 (2.2)†	37.3 (4.2)†	65.7 (1.8)†	49.7 (2.1)†	54.8 (1.9)†	53.6 (2.7)†
A + V (C P - N)	45.8 (3.1)†	54.3 (2.1)†	36.8 (5.4)†	67.0 (1.4)	51.0 (2.0)†	56.7 (1.6)†	54.3 (2.4)†
A + V (I P - S)	54.4 (3.4)	**77.1 (4.6)**†	**47.3 (7.8)**	**82.3 (2.6)**†	**65.3 (2.9)**†	**72.6 (3.0)**†	**64.9 (3.0)**†
A + V (I P - N)	50.2 (2.7)†	**72.8 (3.7)**†	46.1 (6.2)	**79.2 (1.9)**†	**62.1 (2.2)**†	**69.2 (2.2)**†	62.3 (2.4)
A + V (PF - S)	**56.9 (2.9)**	71.0 (2.6)†	44.0 (3.3)	75.9 (1.8)†	62.0 (1.6)†	67.7 (1.8)†	64.0 (1.9)†
A + V (PF - N)	**57.7 (2.2)**†	67.2 (2.5)†	**46.2 (4.2)**	74.9 (1.0)†	61.5 (1.6)†	67.0 (1.2)†	**64.2 (2.0)**†

Subjects 1 to 7 from the AVIC dataset are used as test set. The mean and (St. Dev.) over 30 experiments are presented. The two highest mean values in each column are given in bold. † denotes that the difference between the prediction-based approaches and D F + F F is statistically significant.

The same is not true when softmax normalisation is used and the bad performance of the cross-modal prediction has a negative effect on the the full prediction-based system. The prediction error difference between the wrong classes predictors and the correct class predictor in the cross-modal prediction module is high enough so it cannot be offsetted by the difference between the correct and wrong classes predictors in the intra-modal prediction module. This happens possibly due to distortion of the prediction error differences when softmax normalisation is applied.

As shown in Table 5, prediction-based fusion is based on the temporal evolution of the audio features (audio-to-audio predictor) and to a much lesser degree on the one-way relationship between visual and audio features (video-to-audio predictor). The combination of decision-level and feature-level fusion relies mostly on decision-level fusion which in turn is mainly based on the audio-only classifier, i.e., on the temporal evolution of the audio features. Hence, in this experiment the better performance of prediction-based fusion is mostly due to the different representation of the audiovisual information.

TABLE 5
Optimal Weights, for Prediction-Based Fusion, Feature-Level Fusion and Decision-Level Fusion

Prediction-Based Fusion Softmax Normalisation		Prediction-Based Fusion No Normalisation	
$[w^G_{VA}, w^G_{AV}]$	[1 0]	$[w^G_{VA}, w^G_{AV}]$	[0.80 0.20]
$[w^G_{VV}, w^G_{AA}]$	[0.05 0.95]	$[w^G_{VV}, w^G_{AA}]$	[0 1]
$[w^L_{VA}, w^L_{AV}]$	[0.75 0.25]	$[w^L_{VA}, w^L_{AV}]$	[0.90 0.10]
$[w^L_{VV}, w^L_{AA}]$	[0 1]	$[w^L_{VV}, w^L_{AA}]$	[0 1]
$[w^C_{VA}, w^C_{AV}$	[0.85 0.15]	$[w^C_{VA}, w^C_{AV}]$	[0.90 0.10]
$[w^C_{VV}, w^C_{AA}]$	[0 1]	$[w^C_{VV}, w^C_{AA}]$	[0 1]
$[w^H_{VA}, w^H_{AV}]$	[0.65 0.35]	$[w^H_{VA}, w^H_{AV}]$	[1 0]
$[w^H_{VV}, w^H_{AA}]$	[0 1]	$[w^H_{VV}, w^H_{AA}]$	[0.15 0.85]
$[w^G_{CP}, w^G_{IP}]$	[0.20 0.80]	$[w^G_{CP}, w^G_{IP}]$	[0.05 0.95]
$[w^L_{CP}, w^L_{IP}]$	[0.15 0.85]	$[w^L_{CP}, w^L_{IP}]$	[0.05 0.95]
$[w^C_{CP}, w^C_{IP}]$	[0.25 0.75]	$[w^C_{CP}, w^C_{IP}]$	[0.05 0.95]
$[w^H_{CP}, w^H_{IP}]$	[0.35 0.65]	$[w^H_{CP}, w^H_{IP}]$	[0.05 0.95]
Decision-Level Fusion		Decision-Level Fusion Feature-Level Fusion	
$[w_A, w_V]$	[0.80 0.20]	$[w_{FF}, w_{DF}]$	[0.05, 0.95]

All parameters were optimised on subjects 15 to 21 from the AVIC dataset.

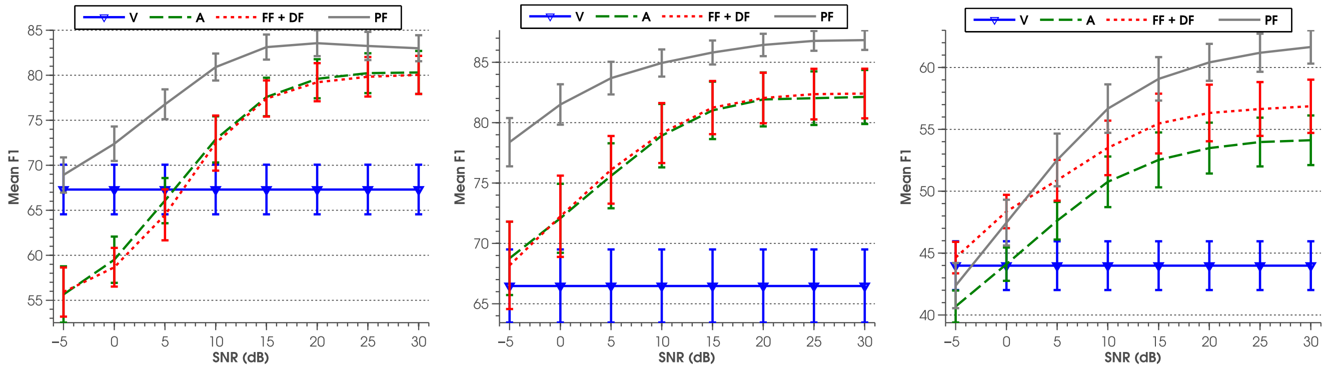
Finally, both types of normalisation perform similarly with the exception of intra-modal prediction fusion where softmax normalisation is superior. Overall, we see that softmax normalisation tends to distort the prediction error differences and this can have both positive and negative effects depending on the dataset, whereas no normalisation tends to be more stable. In case of softmax versus having no normalisation the $L^2 - E$ and MSE were found to be the best performing error measures, respectively, on the validation set and these are the prediction error measures used in all experiments.

8 EFFECT OF AUDIO NOISE

In order to investigate the robustness to audio noise of the audiovisual fusion approaches we run experiments under varying noise levels. The audio signal for each example is corrupted by additive babble noise from the NOISEX database [67] so as the SNR varies from -5 to 30 dB.

Results for the AMI, MAHNOB and AVIC datasets are shown in Figs. 7a, 7b and 7c, respectively. Overall, we see that prediction-based fusion is more robust to audio noise than the combination of decision- and feature-level fusion. The video-only classifier (blue solid line with triangle markers) is not affected by the addition of the audio noise and therefore its performance remains constant over all noise levels. On the other hand, as expected, the performance of the audio classifier (green dashed line) degrades as the audio noise increases.

The best performing approach over all noise levels on the AMI and MAHNOB datasets is the full prediction-based fusion (grey solid line). More specifically, its performance on the AMI dataset ranges from 82.5 percent (1.4) to 68.9 percent (2.0) and is the only approach which remains above the video-



(a) Test Set: AMI (b) Test Set: MAHNOB (c) Test Set: AVIC

Fig. 7. Mean F1 as a function of the babble noise added to the audio signal for different test sets. Prediction-based fusion is the best performing approach over all noise levels except for −5 dB and 0 dB on the AVIC dataset. V: Video-only classifier, A: Audio-only classifier, FF: Feature-level Fusion, DF: Decision-level Fusion, PF: Prediction-based Fusion with softmax normalisation, SNR: Signal-to-Noise Ratio.

only performance for all noise levels until −5 dB. On the MAHNOB dataset, it achieves a mean F1 of 86.8 percent (0.8) in 30 dB which decreases to 78.4 percent (2.0) in −5 dB. On both datasets the combination of decision- and feature-level fusion is almost identical to audio-only classification since it is mostly based on the audio-only classifier as shown in Table 2.

Similar conclusions can be drawn for the AVIC dataset, Fig. 7c. Between 10 and 30 dB prediction-based fusion is the best performing approach and the combination of decision- and feature-level fusion is the second best. The main difference with the other two datasets lies in the two noisiest levels. In 0 dB, both methods result in the same performance, whereas in −5 dB the combination of decision- and feature-level fusion performs slightly better, 44.6 percent (1.3), than prediction-based fusion, 42.4 percent (1.8).

9 DISCUSSION

The main advantage of the prediction-based fusion approach is that it does not explicitly rely on the actual values of the features as is the case for feature-level or decision-level fusion. The problem is converted into a competition between several models, e.g., a laughter and a speech model or a laughter, a hesitation, a consent and a garbage model. It does not matter if the prediction is good or bad, what matters is if the correct prediction model is closer to the actual values than the competitor models. This means that what matters is the relative position of the prediction errors and not their absolute values. Since the audiovisual feature relationship and their temporal evolution are different for each vocalisation, it is expected that the predictor which corresponds to the input vocalisation, i.e., was trained to model the audiovisual relationship for this vocalisation, will make a better prediction and hence the input example will be correctly classified.

An illustration of this principle is shown in Figs. 8 and 9. Fig. 8f shows the output of decision-level and feature-level fusion system for a laughter episode from the MAHNOB database. The output is negative most of the time and the episode is incorrectly labelled as speech. Fig. 8g shows the MAE per frame for the laughter and speech predictors computed from Eq. (13). For almost all frames the laughter predictors give a better prediction than the speech predictors as expected, since they better model the audiovisual relationship for laughter. The total error over the entire episode is 33.7 and 37.7 for laughter and speech, respectively, and therefore the episode is correctly classified as laughter.

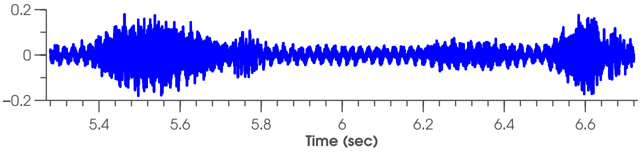
An example from the AVIC database is shown in Fig. 9. Fig. 9f shows the output of the combination of feature- and decision-level fusion approach for a hesitation episode. The garbage output (blue line) consistently produces the highest output so the episode in incorrectly classified as garbage. Fig. 9g shows the MSE per frame for all four prediction models. For almost all frames the hesitation model results in the lowest MSE. The total MSE error over the entire episode is 1.4, 2.5, 1.5, and 1.1, for the garbage, laughter, consent and hesitation predictors, respectively, which means that this episode is correctly classified as hesitation.

Fig. 10a shows the total score of a laughter episode, from the MAHNOB database, assigned by the combination of decision- and feature-level fusion for different noise levels. It can be seen that as the SNR becomes lower, the total score becomes lower as well. From 10 to 30 dB the total score is above zero and the episode is correctly classified as laughter. However, between −5 dB and 5 dB the score is negative, and the example is misclassified as speech. Fig. 10b shows the total laughter and speech prediction error of the same episode for the same noise levels. As the noise level increases, the prediction error of the correct model (laughter) becomes worse but lower than the error of the wrong model (speech) and hence the sequence is labelled correctly. Therefore, in this case, the episode is correctly classified for all noise levels. 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.

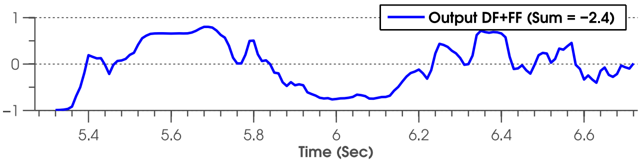
10 CONCLUSION

This paper has approached the problem of audiovisual fusion from a new perspective. Inspired by recent computational models of the brain, we have presented a new

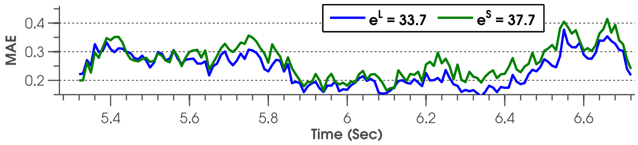
(a) Frame 133 (b) Frame 145 (c) Frame 157 (d) Frame 168



(e) Audio signal



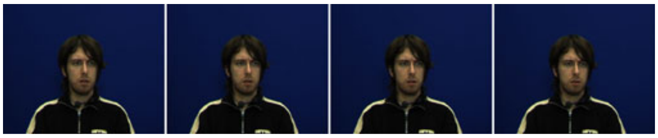
(f) Output of DF + FF. The caption shows the total score. A single output NN is used where positive/negative output correspond to laughter and speech, respectively. The example is misclassified as speech since the total score is negative.



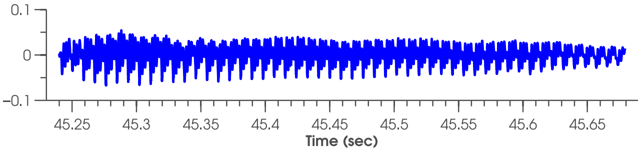
(g) MAE of the laughter and speech models. The caption shows the total MAE over the entire episode. The example is classified as laughter since this model leads to the lowest error.

Fig. 8. Example of laughter (L) from the MAHNOB database, Session S004-006, which is classified correctly by the full prediction-based approach but misclassified by the combination of decision-level (DF) and feature-level fusion as speech (S).

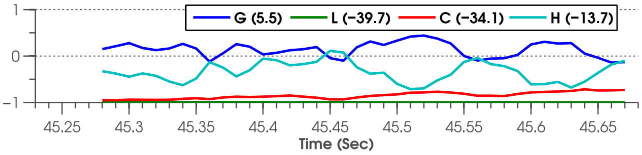
approach called prediction-based audiovisual fusion and compared its performance with two of the most commonly used fusion approaches, decision-level and feature-level. The main idea is that predictive models can be used to model the spatiotemporal relationship and the time evolution of the audio and visual features. The concatenation in feature-level fusion can be replaced by models which predict the visual features based on audio and vice versa for each class separately. Similarly, the audio and visual streams in decision-level fusion can be modelled by two one-step-ahead predictors. Fusion takes place by combining the prediction errors from all models in a hierarchical way. Classification occurs by labelling an input sequence based on the class-specific model that produces the lowest prediction error. When tested on classification of nonlinguistic vocalisations with and without added audio noise, prediction-based fusion outperforms the standard fusion methods in virtually all cases. A drawback of this approach is that if the time series 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. One line of research we are currently investigating in order to solve this problem is the creation of multiple sets of predictive models which are trained on different clusters of time series within each class. This has the potential to lead to more accurate predictions which can further enhance performance.



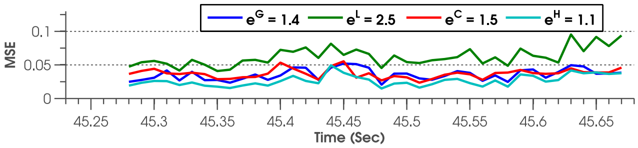
(a) Frame 1132 (b) Frame 1136 (c) Frame 1139 (d) Frame 1142



(e) Audio signal

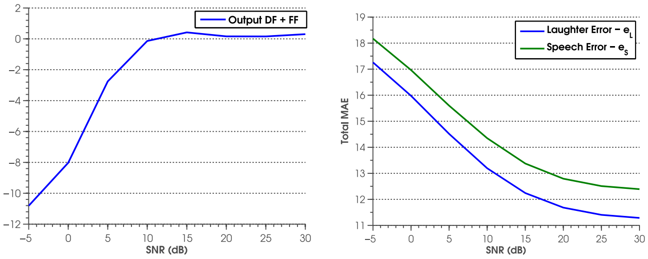


(f) Output of DF + FF. The caption shows the total score for each class. One NN with four outputs is used, where each output corresponds to one class. The example is misclassified as garbage since this output leads to the highest score.



(g) MSE of the four models. The caption shows the total MSE over the entire episode. The example is classified as hesitation since this model leads to the lowest error.

Fig. 9. Example of hesitation from the AVIC database, subject VP12 (VP12_part1), which is classified correctly by the full prediction-based approach but misclassified by the combination of decision-level (DF) and feature-level fusion as garbage. G: Garbage, L: Laughter, C: Consent, H: Hesitation.



(a) Output of the combination of DF and FF summed over the entire sequence for different audio noise levels. Between -5 dB and 10 dB the output becomes negative so the episode is wrongly labelled as speech.

(b) Total MAE over the entire sequence of the laughter and speech predictors for different audio noise levels. For lower SNRs the prediction error increases for both predictors so the example is classified correctly as laughter in all cases.

Fig. 10. Output of the combination of decision-level (DF) and feature-level fusion, and full prediction-based fusion with no normalisation, on a laughter episode from the MAHNOB database (session S014-001, start frame: 741, end frame: 754), as a function of the audio noise level.

ACKNOWLEDGMENTS

This work has been funded by the European Community Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work by

M. Pantic was further supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA).

REFERENCES

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

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

[3] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual and spontaneous expressions,” *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[4] H. Gunes and B. Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[5] S. Petridis and M. Pantic, “Audiovisual discrimination between speech and laughter: Why and when visual information might help,” *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 216–234, Apr. 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] P. S. Aleksic and A. K. Katsaggelos, “Audio-visual biometrics,” *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.

[8] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D’Errico, and M. Schröder, “Bridging the gap between social animal and unsocial machine: A survey of social signal processing,” *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 69–87, Jan.–Mar. 2012.

[9] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, “Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey,” *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[10] K. Friston, “The free-energy principle: A unified brain theory?” *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.

[11] J. Hawkins and S. Blakeslee, *On Intelligence*. New York, NY, USA: Owl Books, 2005.

[12] A. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lip Reading*, B. Dodd and R. Campbell, Eds. London, U.K: Erlbaum, 1987, pp. 3–51.

[13] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, “The natural statistics of audiovisual speech,” *PLoS Comput. Biol.*, vol. 5, no. 7, p. e1000436, 2009.

[14] K. W. Grant and P.-F. Seitz, “The use of visible speech cues for improving auditory detection of spoken sentences,” *J. Acoust. Soc. Amer.*, vol. 108, p. 1197, 2000.

[15] J.-L. Schwartz, F. Berthommier, and C. Savariaux, “Seeing to hear better: Evidence for early audio-visual interactions in speech identification,” *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.

[16] L. H. Arnal, V. Wyart, and A.-L. Giraud, “Transitions in neural oscillations reflect prediction errors generated in audiovisual speech,” *Nat. Neurosci.*, vol. 14, no. 6, pp. 797–801, 2011.

[17] C. Chandrasekaran and A. A. Ghazanfar, “When what you see is not what you hear,” *Nat. Neurosci.*, vol. 14, no. 6, p. 675, 2011.

[18] S. Petridis, A. Asghar, and M. Pantic, “Classifying laughter and speech using audio-visual feature prediction,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 5254–5257.

[19] S. Petridis, M. Pantic, and J. F. Cohn, “Prediction-based classification for audiovisual discrimination between laughter and speech,” in *Proc. IEEE Int. Automatic Face Gesture Recognit. Workshops*, 2011, pp. 619–626.

[20] S. Petridis, S. Bilakhia, and M. Pantic, “Comparison of prediction-based fusion and feature-level fusion across different learning models,” in *Proc. 20th Int. Conf. ACM Multimedia*, Nara, Japan, Nov. 2012, pp. 813–816.

[21] K. Scherer, “Affect bursts,” in *Emotions: Essays on Emotion Theory*, S. van Goozen, N. van Poll, and J. Sergeant, Eds. East Sussex, U.K.: Psychology Press, 1994, pp. 161–193.

[22] S. Petridis and M. Pantic, “Audiovisual discrimination between laughter and speech,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 5117–5120.

[23] S. Petridis and M. Pantic, “Audiovisual laughter detection based on temporal features,” in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2008, pp. 37–44.

[24] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5844–5847.

[25] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, “A coupled HMM for audio-visual speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2002, vol. 2, pp. 2013–2016.

[26] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang, “Audio–visual affective expression recognition through multistream fused HMM,” *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570–577, Jun. 2008.

[27] A. V. Nefian, L. H. Liang, X. B. Pi, X. X. Liu, and K. Murphy, “Dynamic Bayesian networks for audio-visual speech recognition,” *Eur. J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1274–1288, 2002.

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

[29] V. Petridis and A. Kehagias, *Predictive Modular Neural Networks: Applications to Time Series*. New York, NY, USA: Springer, 1998.

[30] A. Kehagias and V. Petridis, “Time-series segmentation using predictive modular neural networks,” *Neural Comput.*, vol. 9, no. 8, pp. 1691–1709, 1997.

[31] G. Bailador and S. Guadarrama, “Robust gesture recognition using a prediction-error-classification approach,” in *Proc. IEEE Intern. Conf. Fuzzy Syst.*, 2007, pp. 1–7.

[32] G. Bailador, D. Roggen, G. Tröster, and G. Triviño, “Real time gesture recognition using continuous time recurrent neural networks,” in *Proc. ICST 2nd Intern. Conf. Body Area Netw.*, 2007, pp. 15:1–15:8.

[33] C.-F. Juang and K.-C. Ku, “A recurrent fuzzy network for fuzzy temporal sequence processing and gesture recognition,” *IEEE Trans. Syst., Man, Cybern., Part B: Cybern.*, vol. 35, no. 4, pp. 646–658, Aug. 2005.

[34] L. Gupta, M. McAvoy, and J. Phegley, “Classification of temporal sequences via prediction using the simple recurrent neural network,” *Pattern Recognit.*, vol. 33, no. 10, pp. 1759–1770, 2000.

[35] C.-F. Juang and T.-M. Chen, “Birdsong recognition using prediction-based recurrent neural fuzzy networks,” *Neurocomputing*, vol. 71, no. 1, pp. 121–130, 2007.

[36] D. Coyle, G. Prasad, and T. McGinnity, “A time-series prediction approach for feature extraction in a brain-computer interface,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 4, pp. 461–467, Dec. 2005.

[37] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, “Linking facial animation, head motion and speech acoustics,” *J. Phonetics*, vol. 30, no. 3, pp. 555–568, 2002.

[38] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, “Quantitative association of vocal-tract and facial behavior,” *Speech Commun.*, vol. 26, no. 1, pp. 23–43, 1998.

[39] 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. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 3183–3190, 2008.

[40] J. P. Barker and F. Berthommier, “Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models,” in *Proc. Int. Conf. Auditory-Vis. Speech Process.*, 1999, p. 19.

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

[42] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodgriguez, “Picture my voice: Audio to visual speech synthesis using artificial neural networks,” in *Proc. Int. Conf. AV Speech Process.*, 1999, pp. 133–138.

[43] P. Hong, Z. Wen, and T. S. Huang, “Real-time speech-driven face animation with expressions using neural networks,” *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 916–927, Jul. 2002.

[44] A. Savran, L. M. Arslan, and L. Akarun, “Speaker-independent 3d face synthesis driven by speech and text,” *Signal Process.*, vol. 86, no. 10, pp. 2932–2951, 2006.

[45] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez, J. Castillo, and I. Rudomin, “Speech-driven facial animation with realistic dynamics,” *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 33–42, Feb. 2005.

[46] 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 *Proc. IEEE CVPR Workshops*, 2009, pp. 53–59.

[47] D. Cosker and J. Edge, “Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations,” in *Proc. Comput. Animation Social Agents*, 2009, pp. 21–24.

[48] L. Kennedy and D. Ellis, “Laughter detection in meetings,” in *Proc. NIST Meeting Recognition Workshop*, 2004, pp. 118–121.

[49] D. P. W. Ellis. (2005). PLP and RASTA (and MFCC, and inversion) in Matlab [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[50] B. Yin, E. Ambikairajah, and F. Chen, “Combining cepstral and prosodic features in language identification,” in *Proc. Int. Conf. Pattern Recognit.*, 2006, vol. 4, pp. 254–257.

[51] S. Petridis, “Audiovisual discrimination between laughter and speech,” Ph.D. dissertation, Imperial College London, London, England, 2011.

[52] I. Patras and M. Pantic, “Particle filtering with factorized likelihoods for tracking facial features,” in *Proc. Int. Conf. Automatic Face Gesture Recognit.*, 2004, pp. 97–104.

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

[54] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, “The AMI meeting corpus: A pre-announcement,” in *Proc. 2nd Int. Conf. Mach. Learn. Multimodal Interaction*, 2006, pp. 28–39.

[55] [Online]. Available: http://www.doc.ic.ac.uk/~maja/AMI-SAL-Annotations.xls, 2010.

[56] 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 *Proc. Workshop Corpora Res. Emotion Affect*, 2008, pp. 1–4.

[57] [Online]. Available: http://mahnob-db.eu/laughter/, 2012.

[58] S. Petridis, B. Martinez, and M. Pantic, “The mahnob laughter database,” *Image Vis. Comput. J.*, vol. 31, no. 2, pp. 186–202, 2013.

[59] [Online]. Available: http://mahnob-db.eu/laughter/media/uploads/annotations.xls

[60] B. Schuller, R. Mueller, F. Eyben, J. Gast, B. Hoernler, M. Woellmer, G. Rigoll, A. Hoethker, and H. Konosu, “Being bored? Recognising natural interest by extensive audiovisual integration for real-life application,” *Image Vis. Comput.*, vol. 27, no. 12, pp. 1760–1774, 2009.

[61] F. Eyben, S. Petridis, B. Schuller, and M. Pantic, “Audiovisual vocal outburst classification in noisy acoustic conditions,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 5097–5100.

[62] B. Schuller, B. Vlasenko, F. Eyben, M. Woellmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: Variances and strategies,” *IEEE Trans. Affective Comput.*, vol. 1, no. 2,, pp. 119–131, Jul.–Dec. 2010.

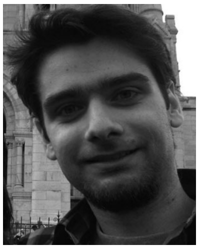
[63] C. Busso, S. Lee, and S. Narayanan, “Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 582–596, May 2009.

[64] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002.

[65] M. Riedmiller and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, vol. 1, pp. 586–591.

[66] M. Smucker, J. Allan, and B. Carterette, “A comparison of statistical significance tests for information retrieval evaluation,” in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2007, pp. 623–632.

[67] A. Varga and H. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.



Stavros Petridis received the BSc degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2004, the MSc degree in advanced computing, and the PhD degree from Imperial College London in 2005 and 2012, respectively. He is a research associate at the Department of Computing, Imperial College London. He has been a research intern in the Image Processing Group, University College London and the Field Robotics Centre, Robotics Institute, Carnegie Mellon University and a visiting researcher at the Affect Analysis Group, University of Piitsburgh. His research interests lie in the areas of pattern recognition and machine learning and their application to multimodal recognition of human non-verbal behaviour and non-linguistic vocalisations. He is currently working on deep learning approaches for audiovisual fusion. He is a member of the IEEE.

Maja Pantic is a professor in affective and behavioral computing in the Department of Computing, Imperial College London, United Kingdom, and in the Department of Computer Science, University of Twente, The Netherlands. She currently serves as the editor in chief of *Image and Vision Computing Journal* and as an associate editor for both the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *IEEE Transactions on Affective Computing*. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She is a fellow of the IEEE.