Audiovisual Three-Level Fusion for Continuous Estimation of Russell's Emotion Circumplex

Enrique Sánchez-Lozano
Multimedia Technologies Group
AtlantTIC Research Center
University of Vigo
esanchez@gts.uvigo.es

Paula Lopez-Otero
Multimedia Technologies Group
AtlantTIC Research Center
University of Vigo
plopez@gts.uvigo.es

Laura Docio-Fernandez
Multimedia Technologies Group
AtlantTIC Research Center
University of Vigo
ldocio@gts.uvigo.es

Enrique Argones-Rúa
Multimodal Information Area
Gradiant
eargones@gradiant.org

José Luis Alba-Castro
Multimedia Technologies Group
AtlantTIC Research Center
University of Vigo
jalba@gts.uvigo.es

ABSTRACT

Predicting human emotions is catching the attention of many research areas, which demand accurate predictions in uncontrolled scenarios. Despite this attractiveness, designed systems for emotion detection are far off being as accurate as desired. Two of the typical measurements in human emotions are described in terms of the dimensions valence and arousal, which shape the Russell's circumplex where complex emotions lie. Thus, the Affect Recognition Sub-Challenge (ASC) of the third AudioVisual Emotion and Depression Challenge, AVEC'13, is focused on estimating these two dimensions. This paper presents a three-level fusion system combining single regression results from audio and visual features, in order to maximize the mean average correlation on both dimensions. Five sets of features are extracted (three for audio and two for video), and they are merged following an iterative process. Results show how this fusion outperforms the baseline method for the challenge database.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Audiovisual features, multimodal fusion, affective computing, eigen-models, support vector regression.

AVEC'13, October 21, 2013, Barcelona, Spain.

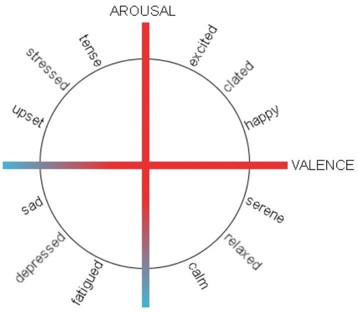

Figure 1: Russell's Emotion Circumplex

1. INTRODUCTION

Affective Computing [21] is an interdisciplinary subject whose goal is providing machines with the ability of recognizing, as well as synthesizing, people's emotions. In recent years, the interest on affective computing has been growing at the time as many applications were inquired for multiple markets: entertainment, education, marketing and health are the four major niches demanding affective solutions. Due to such increased interest, many products and applications have been developed in recent years, although research has not converged to an ultimate solution. Two of the most important cues on Affective Computing are the audiovisual signals. Beyond what is transmitted by our words, more than a 50% of our emotion state is transmitted by our expressive face and voice. Thus, many efforts are put towards improving emotion detection by audiovisual features.

Within emotion detection we can distinguish between two major trends: detecting appearance subtle changes and detecting complex states. The first one typically comprises detecting either an overall basic emotion or frame-by-frame muscular movements (Action Units). The second one comprises detecting complex states in a continuous space within which emotions lie. Both trends pave research lines of many leading groups. Great efforts have been made to measure the state of the art in both trends, but the lack of spontaneous

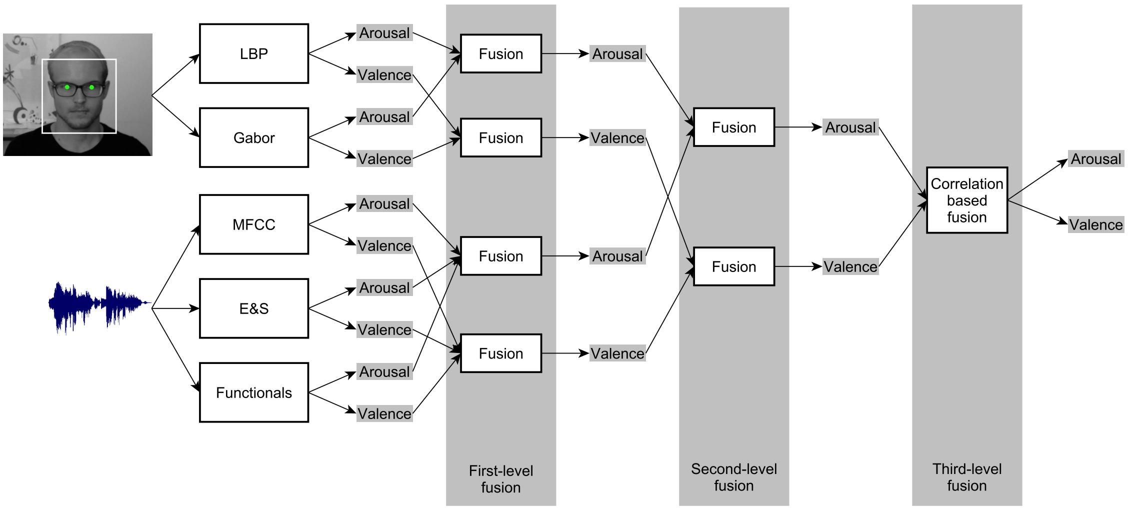

Figure 2: Overview of the emotion identification system

databases, jointly with the fact that each targeted research needs a specific database, make the task formidably complex. One of the best options to measure the state of the art is working in a common framework, like the one proposed in international challenges. This was revealed on the success of the first Facial Expression Recognition and Analysis challenge (FERA'11) [36], of the different challenges for extracting paralinguistic information from speech [27, 28, 30, 29, 31], and of the first and second AudioVisual Emotion Challenge (AVEC'11,'12) [33, 32]. This year, the challenge is also focused on depression detection, highlighting the fact that health is an important target for affective computing. The choice of valence (positivity or negativity) and arousal (activity) as dimensions to be estimated in AVEC'13 [37] remarks that both explain almost all complex states, as the Russell's circumplex model suggests [23]. This model (shown in Figure 1), shows how many emotional states can be described in terms of valence and arousal.

Most of the existing works on the continuous audiovisual emotion recognition task present systems whose key points address three aspects: the feature extraction for the audio and video signals, and the prediction and fusion strategies used for obtaining the continuous values representing the emotional state.

For the visual part, many typical appearance features have been used, based on their previous successful performance on other face analysis tasks, such as face alignment, face recognition or age and gender recognition. Classical approaches include Local Phase Quantisation (LPQ) [37], Gabor [16], Local Binary Patterns (LBP) [35], and Local Gabor Binary Pattern (LGBP) [34]. Recently, extensions from Three Orthogonal Planes (LPQ-TOP [12], LGBP-TOP [2]), have also been successfully used. Also, many geometric features have been proposed, including those based in Active Appearance Models [34] or Constrained Local Models [19]. Regarding the learning stage of emotion recognition using visual cues, the most extended methods used are SVM for classification tasks [16], and ϵ-SVR for regression tasks. What differs among different works is the type of kernel, as well as the number of kernels used [34]. Apart from ϵ-SVR, Kernel Regression was also explored with good results [19]. From all these techniques, it is a hard task selecting which combination is the most promising one. It remains unclear whether some features are better than others for the emotion recognition task, but on specific environments (same database, regressor, ...). The same occurs in the learning stage, since many regression techniques, such as, e.g., Gradient Boosting, have not been explored yet for the emotion recognition task.

In the audio part, the extracted features are generally based on measurements related to the prosodic and spectral characteristics of the audio signal. For example, the Mel-frequency cepstrum coefficients (MFCC), the energy, the zero-crossing rate, the speech rate and the pitch have been used successfully in affect recognition [41]. However, a few years ago a successful method for using a supra-segmental modeling of the audio characteristics emerged. This is based on a set of statistical functionals extracted from the above features (called acoustic Low-Level Descriptors (LLD)) [32, 33]. Approaches combining acoustic features and spoken words, as well as approaches using linguistic features to improve spontaneous emotion recognition, have also been proposed [25]. It is worth noting that deciding the optimal audio feature set is still an open research problem.

As in the visual part, different machine learning algorithms can be used for the learning stage. Several methods are based on context-dependent frameworks. For example, in [17] a system based on Hidden Markov Models was proposed. An advanced technique based on context modelling using Long Short-Term Memory Neural Networks (LSTM-NN) was investigated [40]. These systems provide the advantage of encoding dynamics within the learning algorithm. Other approaches use a static predictor as, for instance, the well-known Support Vector Machine [26, 5].

Several fusion strategies may be used for merging the visual and audio information. Feature-level fusion (also called early fusion) can be performed by merging different sets of features from each modality into one cumulative structure and feeding it to a single classifier. Another solution is decision-level fusion (or late fusion); each feature set feeds

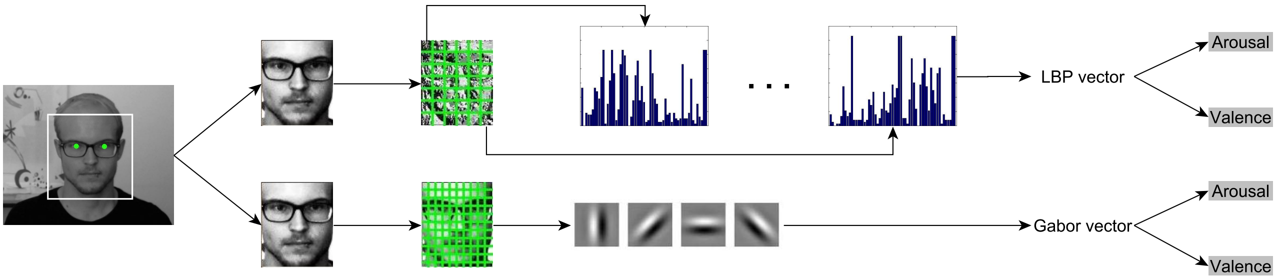

Figure 3: Overview of the video-based methods

one classifier and all the classifier outputs are fused to provide the final output. An output-associative fusion framework is presented in [18] and compared to feature-level and decison-level fusion approaches.

With the aim of predicting the Russell's Circumplex dimensions, we present an audiovisual system combining several simple audio and video features. Each feature vector is projected onto each emotion dimension using classical regression methods. The strength of the proposed method resides on how spectral audio features are computed and how each source is combined through an iterative algorithm. Moreover, based on an existing correlation among both dimensions, we present an algorithm for their mixing and reestimation.

The system is presented in response to the ASC subchallenge of the AVEC'13, which makes use of the DEPRESSION database, consisting of 150 videos equally divided into training, development, and test sets. This database has been continuously annotated by a team of 23 naive raters, and each video was annotated by only a single rater. Although each rater made annotations for each dimension separately, the fact that each video was tagged only by one rater suggests a possible correlation among dimensions, which is successfully explored in this work. As stated above, we have extracted three kinds of features for audio sequences, and two kinds of features for video sequences. Then, we have explored a three-level fusion system in order to maximize the average correlation. The first level fusion treats audio and video separately. The second one combines results from audio and video. Finally, the third one combines audiovisual results from both dimensions. The whole strategy is depicted in Figure 2. The rest of the paper is organized as follows: Sections 2 and 3 review the feature extraction method for video and audio, respectively; Sections 4 and 5 present the audio and video learning systems, respectively; Section 6 presents the audiovisual fusion strategy; Section 7 presents the experiments and results obtained for both the development and test sets; and Section 8 presents the conclusions.

2. VIDEO FEATURES

Faces were extracted using the bounding boxes given in the database. After the face was localised, the eyes were located using ASEF filters [3], and the image was rotated and scaled so that the eyes lie in the same horizontal line, and the interocular distance is set to 45 pixels (which will be the half of the width of the cropped face). After aligning and scaling, face is cropped to a single size of 105×90. Then, we have extracted two sets of features: LBP [20] and Gabor [8], which have been widely used in face analysis systems, such as face recognition [1, 9] or gender recognition [6]. LBP histograms were extracted using 6×7 regions of 15×15 pixels, and concatenated in a single vector. Each histogram consists of 59 bins, and is first normalised, truncated to 0.2, and renormalised again [39]. For Gabor, we used a 10×10 uniform grid. At each point of the grid, we have obtained the module of the output of 40 complex Gabor filters, using 5 frequencies and 8 orientations. Thus, we have a vector consisting of 2748 dimensions for LBP, and 4000 for Gabor. An overview of the feature extraction system can be seen in Figure 3. In order to deal with the high dimension of the feature space, we have applied PCA on a random subset of the training database, retaining 85% of the energy. The LBP feature vector was then reduced to ~ 500 dimensions, and the Gabor vector was reduced to ~ 300. This feature extraction system follows the one presented in [6] for gender recognition. Considering the similar results obtained for the studied sizes in that work, we have decided, for the sake of simplicity, to choose the smallest one. The same philosophy was taken into account for selecting the kind of features among existing ones. Studying other kinds of features is out of the scope of this paper, even though further work shall address it.

3. AUDIO FEATURES

As speech signals are not stationary, it is common in speech processing to divide them into frames of a few miliseconds (typically 5-100 ms) that can be considered to be approximately stationary [22]. From these frames many different types of temporal, spectral, energy and perceptual descriptors can be extracted from these frames. An issue that must be considered for the selection of efficient audio-related features for emotion characterization and recognition is the time extent used for feature extraction. Thus, it is possible to distinguish between low-level descriptors or instantaneous descriptors, which are computed for each time frame, and functionals or global descriptors, which are computed for the whole audio signal or an audio segment covering several frames. There is controversy in the literature about which of the above descriptors are more suitable for emotion recognition [7]; thus, the use of both of them is proposed in this work. Specifically, three different audio feature sets are extracted from the audio signals. Two of them are low-level descriptors (LLD) related to the temporal, energy and spectral characteristics of the audio signal, and the third

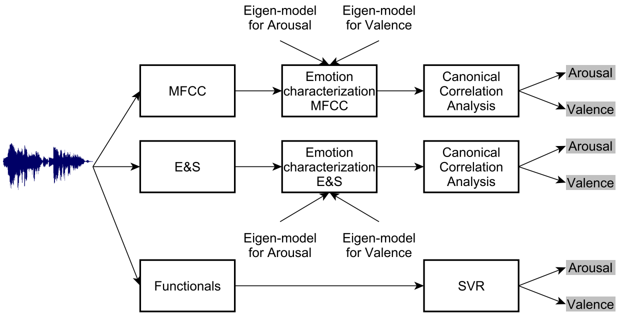

Figure 4: Overview of the audio-based methods

set consists of functionals extracted from a set of prosodic and spectral low-level descriptors.

3.1 Spectral related features

Two different sets of low level descriptors are used, which consist of features that are widely used in speech emotion recognition. The first set is composed of cepstral-based features, and it includes 16 Mel-frequency cepstral coefficients (MFCC) plus their derivatives, a common choice for emotional speech analysis [24]. The second set is composed of temporal, energy and spectral based features. These feature sets are summarized in Table 1. All the features were extracted every 10 ms using a 40 ms Hamming window. It is not possible to identify an emotion with a context of only 40 ms, a bigger context is necessary. To obtain vectors that represent a bigger context, the audio is segmented using a sliding window of 3 s with 1 s of overlap. These segments are processed with an eigen-space technique in order to obtain a low-dimensional representation suited for arousal and valence estimation. Eigen-spaces techniques have been successfully used for speaker verification purposes, e.g. in [14] and [38]. In the eigen-space framework, statistical adaptation from a Universal Background Model (UBM) represented by a Gaussian Mixture Model is performed on a reduced subspace. Therefore, speakers are defined in a low-dimensional subspace where differences between them are easily found. The application of eigen-spaces techniques to continous emotion detection is not straightforward for two reasons: (1) the eigen-space must be able to represent the emotions taking out the speaker-dependent information; (2) speaker representation is discrete, whilst emotion representation is continuous, which means that emotional training data must be somehow discretized in order to apply the standard eigen-space techniques.

Table 1: Audio features.

MFCCs (48) (MFCC)
MFCC 1-16 plus delta and acceleration
Energy and spectral features (34) (E&S)
Loudness, zero crossing rate, energy in bands from 250–650Hz, 1 kHz–4 kHz, 25%, 50%, 75%, and 90% spectral roll-of points,spectral flux, centroid, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, flatness

Before addressing the emotion detection task, an overview of the eigen-space approach for speaker representation is presented. Given the UBM supermean vector:

$$\mathbf{m} = \left[\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_M'\right] \tag{1}$$

where $\boldsymbol{\mu}_i$ is the D-dimensional mean of the i^{th} UBM Gaussian mixture, M is the number of mixtures in the UBM, and D is the dimension of the acoustic features, the speaker u supermean vector is defined as:

$$\mathbf{m}_{\mathrm{u}} = \mathbf{m} + \mathbf{V}\mathbf{y}_{\mathrm{u}} \tag{2}$$

where $\mathbf{V}$ is a $DM \times R$ matrix characterizing the eigen-space, with $R \ll DM$, and $\mathbf{y}_{\mathrm{u}}$ is the low-dimensional speaker u characterization.

Unlike in the eigenspeaker characterization case described above, our aim is providing a compact representation of arousal and valence of speech. Hence, speech vectors are clustered into different emotional clusters, which we denote here as protoclasses, depending on the values of arousal and valence in the training set, instead of grouping speech vectors attending to the speaker. The LBG algorithm [15] is used for this task. The number of clusters chosen for this purpose was finally 8. All the vectors belonging to a given protoclass in the training set are here playing the same role as those belonging to a given speaker u in the eigenspeaker framework, thus the obtained matrix $\mathbf{V}$ is modelling the maximum variation directions of the UBM supermean regarding the values of arousal and valence. There are different techniques that can be used for obtaining both $\mathbf{V}$ and $\mathbf{y}_{\mathrm{u}}$. Taking into account that our main limitation is the low number of acoustic vectors used to infer the arousal and valence, the probabilistic method described in [13] was chosen for that purpose but, as stated above, u does not represent a speaker but a protoclass.

Anytime a test segment of emotional speech $\mathbf{s}_{\mathrm{e}}$ is processed, its posterior emotional characterization $\mathbf{y}_{\mathrm{e}}$ is obtained performing the adaptation described in [13]. This emotional characterization vector is later used as input for the arousal and valence regressors, as described in Sec. 5.1.

3.2 Functionals

The third acoustic feature set consists of feature vectors of dimension 2268 given to the participants of AVEC'13 challenge. These functionals were extracted using a sliding window of 3 s and an overlap of 1 s. As the different features range in different values, all the features were normalized to have zero mean and unit variance.

The dimension of these high dimensional feature vectors was reduced in order to consider only the most relevant ones using a correlation based feature subset selection (CFS) algorithm [10]. This widely used feature selection algorithm searches for the *best* subset of features, where *best* is heuristically defined taken two criteria into account: 1) the goodness of the individual features for predicting the class and 2) their correlation with the other features. Therefore, good subsets of features contain features that are highly correlated with the class but uncorrelated with each other. Thus, CFS directly handles correlated and irrelevant features. In order to avoid an exhaustive search through the 2268 different features, a best first search strategy was used: instead of removing the useless features, the algorithm starts with an empty set of features and the relevant features are added as they are chosen. CFS was applied to extract the most significant functionals for arousal and the most significant functionals for valence, resulting in 34 features for arousal and 47 features for valence.

4. EMOTION PREDICTION WITH VIDEO FEATURES

We have trained an ϵ-SVR both for LBP and Gabor reduced features, using a linear kernel. For training, we randomly selected 750 images per video, and 10-videos-out cross-validation was performed on the training set. We have trained one regressor for each kind of features and dimension, resulting in four regressors, two per dimension. As a first idea, a simple linear combination should give the weight for each dimension. This linear regression could be carried out by rearranging the predictions of each ϵ-SVR into a two column matrix, and obtaining a regression vector through least squares against the ground-truth vector for the whole training set. However, beyond the prohibitive amount of training samples, this approach does not consider that each video should be itself a sample, since they have different time lengths, and the mean correlation between predictions and ground truth may significantly differ for each video. Thus, if we have several long videos that are wrongly predicted and few short videos that are fairly well predicted, the regression fusion vector will be unable to fit new samples correctly. On the opposite side, when few long videos are properly predicted, the regression vector will overfit these videos, suffering from lack of generalization. Thus, we propose to treat each video as a sample itself. Let d be the valence/arousal dimension. The regression vector for each video i is obtained as follows:

$$\mathbf{r}_d^i = \mathbf{d}_t^i(\mathbf{D}_p^i)^T \left((\mathbf{D}_p^i)(\mathbf{D}_p^i)^T\right)^{-1}, \tag{3}$$

where $\mathbf{D}_p^i$ is the two-column matrix obtained by concatenating the Gabor and LBP predictions for all the frames from video i, and $\mathbf{d}_t^i$ is the true dimension vector for the video i. Then, a mean vector could be a first attempt for having the fusion vector (called $\mathbf{f}_d$). However, the high variability on each $\mathbf{r}_d^i$ suggests that this is not a good choice. Let us consider an initial estimated $\mathbf{f}_d$, then,

$$\widetilde{\mathbf{d}}_p^i = \mathbf{D}_p^i \mathbf{f}_d, \tag{4}$$

is the vector prediction for each video, and

$$c_i = |corr(\mathbf{d}_t^i, \widetilde{\mathbf{d}}_p^i)|, \tag{5}$$

is the absolute value of Pearson's correlation coefficient for that video. The average of Pearson's correlation is then

$$c = \frac{1}{N}\sum_{i=1}^{N} c_i, \tag{6}$$

where N is the number of videos. Thus, what we propose is, instead of considering the mean regression vector, to equalize the contribution of each video by weighting each $\mathbf{r}_d^i$ with the term $(1 - c_i)$. This term will increase the contribution of regressor vectors of bad predicted videos. This equalization term was inspired on the way AdaBoost weighs training samples when updating the output regressor. Then, $\mathbf{f}_d$ is calculated as follows:

$$\mathbf{f}_d \leftarrow \mathbf{f}_d + \sum_{i=1}^{N} \mathbf{r}_d^i(1 - c_i). \tag{7}$$

After $\mathbf{f}_d$ is calculated, we can go back to Eqn. (4) and Eqn. (5), iteratively, until c is not improved. Algorithm 1 summarizes the proposed learning method.

Algorithm 1 Procedure for training the linear combination of regressors for dimension d.

```
Initialize $\mathbf{f}_d = \begin{pmatrix}1\\1\end{pmatrix}$
for each $video_i$ do
    Compute $\mathbf{r}_d^i = \mathbf{d}_t^i(\mathbf{D}_p^i)^T \left((\mathbf{D}_p^i)(\mathbf{D}_p^i)^T\right)^{-1}$
end for
repeat
    for each $video_i$ do
        $\widetilde{\mathbf{d}}_p^i = \mathbf{D}_p^i \mathbf{f}_d$
        $c_i = |corr(\mathbf{d}_t^i, \widetilde{\mathbf{d}}_p^i)|$
    end for
    $\mathbf{f}_d \leftarrow \mathbf{f}_d + \sum_i \mathbf{r}_d^i(1 - c_i)$
until $\sum_i (c_i - c_i^{prev}) \le 0$
```

When predicting new frames, a score is individually obtained for each kind of features, and then, the fusion is obtained as follows:

$$score = \mathbf{f}_d(1) * score_{LBP} + \mathbf{f}_d(2) * score_{Gabor} \tag{8}$$

5. EMOTION PREDICTION WITH AUDIO FEATURES

Two different approaches are used for performing arousal and valence identification. The first strategy is applied to the spectral based features and the second one is applied to the functionals.

5.1 Prediction with spectral based features

After obtaining the vectors corresponding to the emotion-characterized segments by applying the eigen-model strategy described in Sec. 3.1, four different sets of 50-dimensional vectors are obtained: one for the characterization of arousal with MFCC+Δ+$\Delta\Delta$ features, one for the characterization of valence with MFCC+Δ+$\Delta\Delta$ features, one for the characterization of arousal with energy and spectral features, and one for the characterization of valence with energy and spectral features. Estimated values of arousal and valence have to be extracted from each of these feature vectors; to do so, canonical correlation analysis is applied to each set of vectors. This technique finds basis vectors for two sets of variables (on the one hand, the features, and on the other hand, arousal or valence) such that the correlation between the projection of the variables onto these basis vectors are mutually maximized [11]. Given the feature vectors of the training data and their corresponding values for arousal (valence), two linear transformations, one for the feature vectors and another one for the arousal (valence) groundtruth labels, are learnt. These linear transformations are applied to the testing data, obtaining as a result a prediction for the arousal (valence) for each feature vector.

5.2 Prediction with functionals

Two ϵ-Support Vector Regressors (SVR) with a linear kernel were trained with the CFS-functionals extracted from the training set (one for arousal and one for valence). There is a training label for arousal (valence) every 1/30 s; as the

CFS-functionals were obtained at a rate of 2 seconds using overlapped segments of 3 seconds, the arousal (valence) value at the end of the 3-second window is assigned to the feature vectors in order to equal the rates of the feature vectors and the labels. SVR parameters were optimized on the training dataset using 5-fold cross-validation and validated on the development dataset. LibSVM library [4] was used for this task.

5.3 Fusion

Three different predictions for arousal and valence are obtained from the different audio features. A fusion of these predictions is performed by applying canonical correlation analysis once again, but this time the linear transformations are learnt from the predictions of arousal (valence) on the training data and the groundtruth labels. This fusion corresponds to the first level of the multilevel fusion strategy depicted in Fig. 2.

6. CORRELATION-BASED FUSION

Once the first-level fusions of the audio and video predictions are obtained, the next step consists of a fusion of these two inputs. We have applied Algorithm 1 again for the second-level fusion illustrated in Fig. 2. Instead of having the LBP and Gabor prediction vectors, now we have the audio and video predictions for each dimension. After the second level fusion, unique values for arousal and valence are given. Although ideally arousal and valence are independent dimensions, this actually does not occur on the challenge dataset. The fact that each video was labelled by a single rater suggests that there may be a correlation between both dimensions. The average of Pearson's correlations between both dimensions is 0.26 for the training set, and 0.25 for the development set, which implies that a combination of both dimensions may be considered in the last step of the prediction. Thus, we shall consider the information each dimension provides respect to the other. As stated before, a whole regression matrix is not an ideal choice. Previous works attempted to fuse both dimensions. In [18], an output-associative framework was presented, combining scores like the way presented in Fig. 2, by using a regression learning technique (Bidirectional Long Short-Term Memory Neural Networks, BLSTM-NN). In [19], a Kernel Regressor was learned for combining all the scores. However, this regression method was learned using a subset of the training set, which is not as accurate as considering the whole dataset. Thus, we propose to use a modified version of Algorithm 1, where a regression matrix ($\mathbf{F} \in \Re^{2\times 2}$) combining both dimensions is learned. Let $\mathbf{a}^i_{t,p}$, $\mathbf{v}^i_{t,p}$ be the ground-truth (t) and independently predicted (p) vectors for arousal (a) and valence (v), respectively, and $\mathbf{D}^i_{t,p} = [\mathbf{a}^i_{t,p}\, \mathbf{v}^i_{t,p}]$. Algorithm 2 summarizes the proposed method. Now, each sample is the regression matrix for each video:

$$\mathbf{R}^i = \mathbf{D}^i_t(\mathbf{D}^i_p)^T \left((\mathbf{D}^i_p)(\mathbf{D}^i_p)^T\right)^{-1}. \tag{9}$$

As in Algorithm 1, Pearson's correlations serve as weights for that samples. Now, we have one Pearson's correlation value for each dimension. These Pearson's correlations values are then used for weighting the corresponding row of $\mathbf{F}$ (the first row will be used for weighting arousal values, whereas the second one will weighs valence values):

$$\mathbf{F} \leftarrow \mathbf{F} + \sum_{i=1}^{N} \left((1-c^a_i)\mathbf{R}^i \circ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} + (1-c^v_i)\mathbf{R}^i \circ \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \right), \tag{10}$$

where $\circ$ denotes the Hadamard product. One mainly difference between our approach and that presented in [18] is that our approach attempts to maximize the average of Pearson's correlations, whereas SVR or a BLSTM-NN optimize other cost functions.

Algorithm 2 Procedure for training the fusion matrix.

```
Initialize $\mathbf{F} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$
for each $video_i$ do
    Compute $\mathbf{R}^i = \mathbf{D}^i_t(\mathbf{D}^i_p)^T \left((\mathbf{D}^i_p)(\mathbf{D}^i_p)^T\right)^{-1}$
end for
repeat
    for each $video_i$ do
        $\widetilde{\mathbf{D}}^i_p \leftarrow \mathbf{D}^i_p\mathbf{F}$
        $\mathbf{a}^i_p = \widetilde{\mathbf{D}}^i_p \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
        $\mathbf{v}^i_p = \widetilde{\mathbf{D}}^i_p \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
        $c^a_i = |corr(\mathbf{a}^i_t, \mathbf{a}^i_p)|$
        $c^v_i = |corr(\mathbf{v}^i_t, \mathbf{v}^i_p)|$
    end for
    Update $\mathbf{F}$ using Eqn. (10)
until $\sum_{i=1}^{N}(c^a_i + c^v_i - (c^a_i + c^v_i)^{prev}) \leq 0$
```

Once the independent scores are computed, they must be recomputed as follows:

$$\begin{aligned} arousal &\leftarrow \mathbf{F}(1,1) * arousal + \mathbf{F}(2,1) * valence \\ valence &\leftarrow \mathbf{F}(1,2) * arousal + \mathbf{F}(2,2) * valence \end{aligned} \tag{11}$$

7. EXPERIMENTAL RESULTS

7.1 Analysis of the eigen-space modelling approach

A novel use of eigen-space techniques for continuous emotion prediction was described in Sec. 3.1. The suitability of that strategy for continuous detection of arousal and valence is supported by results in Table 2, where it can be seen that the correlation between the groundtruth labels and the estimated labels obtained from the MFCCs increases to a great extent when using the eigen-space technique, while dramatically reducing the dimensionality of the feature vectors. Specifically, when no eigen-space technique is used, the feature vectors are equal to the supermean vectors $\mathbf{m}$ of dimension $M \cdot D = 256 \cdot 48 = 12288$, according to the

Table 2: Pearson's correlation averaged over the development set with and without eigen-space modelling when using 16 MFCCs plus their derivatives.

Eigen-space	Dimension	Arousal	Valence	Average
Yes	**50**	**0.1721**	**0.1404**	**0.1562**
No	12288	0.0903	0.0905	0.0904

Table 3: Pearson's correlation averaged over the development set of the Video Fusion strategies.

Method	Arousal	Valence	Average
LBP	0.1204	0.1453	0.1328
Gabor	**0.1334**	0.1383	0.1358
Fusion	0.1192	**0.1536**	**0.1364**

Table 4: Pearson's correlation averaged over the development set for the Audio Fusion strategies.

Method	Arousal	Valence	Average
MFCC	**0.1721**	**0.1404**	**0.1562**
S&E	0.1585	0.0990	0.1288
Functionals	0.1257	0.1268	0.1263
Expert-level fusion	0.1548	0.1327	0.1437
Feature-level fusion	0.1418	0.1214	0.1316

notation in Sec. 3.1, while the dimension is set to $R = 50$ when applying the eigen-space technique.

7.2 Analysis of fusion results

Table 3 shows the results obtained by the two video-based prediction strategies compared with their fusion, performed as described in Sec. 4. Table 5 also shows the results obtained when applying this type of fusion to the fused audio and video predictions (second-level fusion). These results show that the fusion strategy described in Algorithm 1 improves the average of arousal and valence with respect to the corresponding individual systems, even though the prediction for arousal (valence) is slightly worse than without the fusion, as can be seen in Table 3 (Table 5).

Table 4 shows the correlations achieved by the three different audio-based detection strategies. The eigen-space technique using MFCCs achieves the best results, followed by the eigen-space technique using spectral and energy features. Two different fusions were applied to the three systems: the first one (expert-level fusion), which was described in Sec. 5.3, consists of a fusion of experts, where each expert produces an estimated value of arousal or valence, and these three values are used for estimating a final value. The second strategy performs fusion at feature level, i.e. the functionals and the eigen-space characterized features are concatenated to create a unique vector, and then canonical correlation analysis is applied, obtaining a predicted value for arousal or valence. Table 4 shows that the expert-level fusion is more effective than the feature-level one, but none of them was able to outperform the eigen-space technique with MFCCs on the development dataset. Nevertheless, fusion results on the training dataset were superior than the individual systems, which may suggest an overfitting issue.

Table 5 shows an important quality of the third-level fusion strategy described in Algorithm 2. An improvement of the average value is obtained after the correlation-based fusion, although the prediction for valence works worse than without this fusion, as shown in Table 5. These results are the expected ones, as Algorithm 2 aims at improving the average of arousal and valence, not each dimension individually. Although the average can be increased by setting the final valence as the one predicted in the second-level of the proposed approach, there is no evidence about whether this improvement is going to be kept on the test set. Thus, these results show the effectiveness of the proposed method.

Table 5: Pearson's correlation averaged over the development set of the described systems.

Method	Arousal	Valence	Average
Video	0.1192	0.1536	0.1364
Audio	0.1548	0.1327	0.1437
Video+audio	0.1542	**0.1727**	0.1635
Correlation-based fusion	**0.1921**	0.1669	**0.1795**

Table 6: Pearson's correlation coefficient averaged over the test dataset.

Method	Arousal	Valence	Average
Our method	0.1318	**0.1352**	**0.1335**
Baseline video	**0.1340**	0.0760	0.1050
Baseline audio	0.0900	0.0890	0.0890

7.3 Results on the test dataset

Table 6 shows the results obtained on the test set of the AVEC'13 database, compared with those given in the baseline paper [37]. This table shows that the proposed strategy obtains a Pearson's correlation coefficient higher than the baseline systems. Also, comparing Tables 6 and 5, it can be seen that the results obtained on the development and test datasets are quite similar. Thus, the system demonstrates that it is able to generalize among different data.

8. CONCLUSIONS

We have presented a complete audiovisual system for predicting both valence and arousal dimensions of human emotions. We have measured each contribution (audio and video) separately (level 1), as well as the merged contribution (level 2). In order to maximize the average between estimated arousal and estimated valence, we have presented an iterative algorithm for making the fusion considering the correlation between both dimensions (level 3). This algorithm penalizes one of the dimensions in increase of the average. Also, the presented fusion have proven to have a good generalization, since it ensures that all the training videos contribute equally, avoiding the problem of overfitting. These fusion algorithms can be generalized to other kind of features or dimensions. Further work should explore other measures for the updating step, as they were designed for maximizing the global correlation measurement, which is used for evaluating the AVEC'13 systems. Also, future work should consider including other features, as well as an optimization for the RMS Error.

A novel eigen-space based approach has been used for emotion characterization of speech. Experimental results showed that this technique is suitable for continuous prediction of arousal and valence. A more general approach using eigen-spaces, where features are not necessarily clustered into emotional protoclasses, will also be explored in future work.

9. ACKNOWLEDGMENTS

This work was partially supported by the Galician Government (Xunta de Galicia) under Grant 10TIC008CT, Research contract CN2011/019 (Modalidade: Grupos de Referencia Competitiva 2007), and the project CN 2012/260 "Consolidation of Research Units: AtlantTIC"; by the Spanish Ministry of Economy and Competitivity under the project TEC2012-38939-C03-01 and FPI grant BES-2010-033358; by the European Regional Development Fund; and by the EU FP7 - Capacities under grant agreement no. 285901 (LIFTGATE project). Finally, authors would like to thank Daniel González-Jiménez and Pablo Dago-Casas for their technical support and helpful comments.

10. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *Computer Vision-ECCV 2004*, pages 469–481. Springer, 2004.

[2] T. Almaev and M. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proc. Affective Computing and Interaction (to appear) (ACII'13)*, 2013.

[3] D. Bolme, B. Draper, and J. Beveridge. Average of synthetic exact filters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, 2009.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] A. C. Cruz, B. Bhanu, and S. Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *ACII (2)*, pages 341–350, 2011.

[6] P. Dago-Casas, D. González-Jiménez, L. L. Yu, and J. L. Alba-Castro. Single-and cross-database benchmarks for gender classification under unconstrained settings. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2152–2159. IEEE, 2011.

[7] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.*, 44(3):572–587, Mar. 2011.

[8] D. Gabor. Theory of communication. *Journal of Institute for Electrical Engnineering*, 53, 1946.

[9] D. González-Jiménez and J. L. Alba-Castro. Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. *Information Forensics and Security, IEEE Transactions on*, 2(3):413–429, 2007.

[10] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[11] D. R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-taylor. Canonical correlation analysis; an overview with application to learning methods. Technical report, 2007.

[12] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

[13] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3), 2005.

[14] R. Kuhn, P. Nguyen, J. C. Junqua, and L. Goldwasser. Eigenfaces and eigenvoices: dimensionality reduction for specialized pattern recognition. In *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pages 71–76, 1998.

[15] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(84), 1980.

[16] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[17] H. Meng and N. Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, ACII'11, pages 378–387, Berlin, Heidelberg, 2011. Springer-Verlag.

[18] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuos prediction of spontaneous affect from multiple cues and modalities in valence–arousal space. *IEEE Trans. Affect. Comput.*, 2(2):92–105, July 2011.

[19] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 501–508. ACM, 2012.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[21] R. Picard. Affective computing. *MIT Press*, 1997.

[22] L. Rabiner and B.-H. Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[23] J. Russel. A circumspect model of affect. *Journal of Psychology and Social Psychology*, 39(6), 1980.

[24] N. Sato and Y. Obuchi. Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848, 2007.

[25] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 485–492, New York, NY, USA, 2012. ACM.

[26] A. Sayedelahl, P. Fewzee, M. S. Kamel, and F. Karray. Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In *Proceedings of the 4th international conference on*

Affective computing and intelligent interaction - Volume Part II, ACII'11, pages 407–414, Berlin, Heidelberg, 2011. Springer-Verlag.

[27] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 312–315, 2009.

[28] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan. The interspeech 2010 paralinguistic challenge. In *In Proceedings of InterSpeech*, Makuhari, Japan, Sept. 2010.

[29] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss. The INTERSPEECH 2012 Speaker Trait Challenge. In ISCA, editor, *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, 2012.

[30] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The interspeech 2011 speaker state challenge. In *INTERSPEECH*, pages 3201–3204, 2011.

[31] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. 09/2013 2013.

[32] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

[33] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

[34] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):993–1005, 2012.

[35] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[36] M. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926, 2011.

[37] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013 - the continuous audio/visual emotion and depression recognition challange. In *Proceedings of the 3rd International Audio/Visual Emotion Challange and Workshop (AVEC'13)*, 2013.

[38] R. Westwood. *Speaker Adaptation Using Eigenvoices*. PhD thesis, Cambridge University, Cambridge, U.K., 1999.

[39] L. Wolf, T. Hassner, Y. Taigman, et al. Descriptor based methods in the wild. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[40] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vision Comput.*, 31(2):153–163, 2013.

[41] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions, 2009.