An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction

Zhaocheng Huang
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
zhaocheng.huang@student.unsw.edu.au

Ting Dang
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
ting.dang@student.unsw.edu.au

Nicholas Cummins
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
n.p.cummins@unsw.edu.au

Brian Stasak
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
b.stasak@student.unsw.edu.au

Phu Le
School of Electrical Eng. and Tele.
The University of New South Wales
Sydney NSW 2052 Australia
phule@unsw.edu.au

Vidhyasaharan Sethu
School of Electrical Eng. and Tele.
The University of New South Wales
Sydney NSW 2052 Australia
v.sethu@unsw.edu.au

Julien Epps
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
j.epps@unsw.edu.au

ABSTRACT
Continuous emotion dimension prediction has increased in popularity over the last few years, as the shift away from discrete classification based tasks has introduced more realism in emotion modeling. However, many questions remain including how best to combine information from several modalities (e.g. audio, video, etc). As part of the AV+EC 2015 Challenge, we investigate annotation delay compensation and propose a range of multimodal systems based on an output-associative fusion framework. The performance of the proposed systems are significantly higher than the challenge baseline, with the strongest performing system yielding 66.7% and 53.9% relative increases in prediction accuracy over the AV+EC 2015 test set arousal and valence baselines respectively. Results also demonstrate the importance of annotation delay compensation for continuous emotion analysis. Of particular interest was the output–associative based fusion framework, which performed very well in a number of significantly different configurations, highlighting that incorporating both affective dimensional dependencies and temporal information is a promising research direction for predicting emotion dimensions.

Categories and Subject Descriptors
G.3 [**Mathematics of Computing**]: Probability and Statistics – *Correlation and regression analysis; Robust regression*
I.5.4 [**Computing Methodologies**]: Pattern Recognition – *Signal processing*; *Computer vision*; *Waveform analysis*

General Terms
Algorithms, Performance, Design, Human Factors, Verification.

AVEC'15, October 26 2015, Brisbane, Australia

DOI: http://dx.doi.org/10.1145/2808196.2811640

Keywords
Emotion Dimension Prediction, Support Vector Regression, Relevance Vector Machine, Output-Associative Fusion, Annotation Delay Compensation, Multimodal Fusion.

1. INTRODUCTION
Using behavioral signal processing techniques to model, analyze, detect or predict human emotions is an actively emerging area of research [1]. In recent years, there has been a shift away from extensive investigation into lab-based recognition of prototypical emotion categories (e.g. anger, fear, etc) towards continuous prediction of emotional dimensions (e.g. arousal and valence) in more naturalistic communication. Affective dimensions are considered a more descriptive representation of subtle and complex emotions and emotion-rated states [1, 2]. For continuous emotion prediction, a number of physiological and behavioral modalities have been investigated, such as: audio [3], video [4], body language [5] and EEG [6]. Also, a combination of the modalities can lead to further improvements [7, 8].

The *2015 Audio/Visual Emotion Challenge and Workshop* (AV+EC 2015) provides an opportunity for advancing continuous emotion prediction by combining information gained from audio, video and physiological data [9]. AV+EC 2015 requires participants to continuously predict arousal and valence by utilizing multimedia signal processing and machine learning techniques. The primary aim of the analysis provided herein is to outperform the challenge baseline benchmark, as well as to provide novel insights into continuous emotion analysis.

The investigations presented within this paper compare the performance of a range of multimodal prediction systems designed to capture relevant audio, video and physiological information. The experimental results demonstrate that significant gains in affect prediction performance can be found by compensating for annotator delays introduced when forming the ground truth labels and via the application of an output-associative regression framework for multimodal fusion.

2. RELATED WORK

Both *Support Vector Regression* (SVR) and *Long Short Term Memory Recurrent Neural Network* (LSTM-RNN) have been proven to be effective for predicting emotion dimensions [3, 10]. The baseline system in the AV+EC 2015 challenge adopted these two widely used methods [9], alongside standard feature sets extracted from the audio, video and physiological signals [11]. Both unimodal and multimodal systems were investigated using a linear combination of SVR and LSTM-RNN as decision level fusion. The decision level fusion, which employed features from all modalities, produced the best system performance [9]. However, a range of alternative approaches can be considered for achieving a more robust and effective multimodal prediction system. Approaches investigated in this paper include; compensation of annotation delay, alternative regression methods and investigations into fusion methods which model emotion dimension dependencies.

It has been realized, within the affective computing community, that delay problems are introduced during the manual annotation process used to establish the ground truth in many continuously annotated emotion corpora [2, 8, 12, 13]. When annotators make decisions based on vocal and visual signals, there is an inherent annotation delay between their perceptual observations and decision-making [2, 13]. Moreover, this delay varies between different raters and can range anywhere between 2 - 10 seconds [13]. This delay can significantly degrade system performance when predicting emotion due to unreliable modeling caused by asynchronous ratings. The compensation of annotation delay is commonly achieved by shifting the input features relative to the ground truth labels during system training and testing [8, 12, 13].

The *Relevance Vector Machine* (RVM) is a relatively new approach to multi-dimensional regression which is gaining in popularity in continuous emotion prediction [14, 15]. RVM can be considered as a sparse Bayesian method analogous to the SVR [16]. A key advantage of RVM over SVR in the context of multi-modal learning is the RVM's *Heterogeneous Mapping* (HM) property, which allows *any arbitrary kernel function* to be used in conjunction with a RVM [17]. HM allows not only the mappings of contextual temporal information, but also a convenient multimodal fusion technique, which negates the need to train and heuristically combine multiple predictors [15].

Multimodal features are advantageous for continuous emotion prediction [1, 8, 9]. Fusing scores from different modalities is normally achieved through feature-level fusion or decision-level fusion [1, 8, 12]. Feature-level fusion is a concatenation of all features and decision-level fusion combines the outputs of models trained on different modalities to make a final prediction. Common systems for learning decision level fusion weights include either *Linear Regression* (LR) [9, 12], or SVR [18]. However, it is worth noting that fusing regressor outputs is difficult due to the multicollinearity problem, in which significant correlations between the prediction values being fused makes it hard to learn a reliable set of fusion weights [19].

Recently *Output-Associative* (OA) fusion techniques, which seek to utilize the correlations between arousal and valence values [1], have been investigated in order to achieve better prediction performance [8, 12, 15]. This framework learns the contextual dependencies that exist between predicted dimensional values, as well as the temporal dependencies that exist between an individual prediction output and all other outputs within a time frame around that prediction.

An extension of the OA framework is the OA-RVM technique presented in [15]. This technique takes the RVM's HM property to not only learn the contextual and temporal dependences that exist between prediction outputs but also the relationship between the predicted outputs and the input features space, and has been proven to be very effective for emotion tracking [15]. There are a range of extensions that can be trialed within the OA-RVM framework, including using other regression methods (e.g. SVR) as the initial modeling method and multimodal fusion.

Herein, the design and development of our systems for entry to the AV+EC 2015 challenge is presented. Our investigations focus on; compensation for annotation delay (Section 3.3), RVM (Section 3.4), a range of multimodal fusion techniques including; feature level fusion (Section 3.5), decision level fusion (Section 3.6) and the output-associative based fusion techniques; OA fusion (Section 3.7) and OA regression (Section 3.8). We present a range of system development experiments and results (Section 5), a series of additional investigations using a variety of audio features (Section 6) as well as our systems and results submitted to the AV+EC 2015 challenge (Section 7).

3. SYSTEM OVERVIEW

3.1 Introduction

Investigations in this research are centered on four different multimodal continuous emotion prediction systems. The aim of these investigations is to give insights into regression fusion. Each system uses different fusion strategies, introduced below. Furthermore, within these frameworks, the advantages offered by including delay compensation and Relevance Vector Machines, are also investigated.

3.2 Features and Modalities

The challenge provides a set of standard acoustic features to participants: the 102-dimensional *Extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [20]; two types of video descriptors (84-dimensional set of facial based appearance features and 316-dimensional set of facial based geometric features); two sets of physiological features (54-dimensional set of *electrocardiogram* (ECG) features and 60-dimensional set of *electro-dermal activity* (EDA) features). The reader is referred to [9] for a complete description of the challenge feature sets. Apart from the provided feature sets, audio features were also extracted using *VoiceSauce* [21] and *openSMILE* [22]. Experiments and results on these feature sets are outlined in Section 6.

Unless stated otherwise, all proposed systems used the provided audio, appearance based video, geometric based video and ECG features. During system design and development, the inclusion of EDA did not aid system performance. Furthermore the inclusion of EDA features when generating predictions of the challenge test set resulted in abnormally high valance predictions. Hence, EDA features are not included in any results presented herein.

3.3 Annotation Delay Compensation

Due to the delay caused by sensing and judgment between an annotator's perceptual observations and their decision-making, the numerical ratings associated with each dimension may not be reliable for system training. Similar to [12, 13], to compensate for annotation delay temporal shifts were applied for each file in the training set in order to realign the features with the ground truth. The frame shift was achieved by dropping first N ground truth scores and last N input feature frames

before regression training, and, unless stated is applied to all systems reported herein.

It should be noted that training a regressor with the shifted training files produces predictions which are shifted forward in time, by N frames, when compared to the ground-truth labels (Figure 1b). However a smoothing filter can be used to re-align the predicted outputs. Filtering has been shown to be effective for smoothing output predictions, helping to minimize adverse effects due to noisy predictions and offer rough estimations for undetected frames in facial features [23]. Filtering also introduces an output-delay proportional to the filter length; a FIR filter length of $2N + 1$ introduces a delay of N, where N is defined as per the previous paragraph. Hence, post-processing filters can also be used for resolving the synchronization issue in predicted outputs caused by the introduction of delay in the training phase (Figure 1c), and unless stated is applied to all systems reported herein.

In this paper, a smoothing filter is used not only to help remove high frequency noise present in predictions, but also to realign predictions generated by a system trained on frame shifted features. As this filter will be applied over longer timescales (2-4s), the commonly used mean filter [23], which applies equal weights to all samples, could be an unsuitable choice of filter. Therefore we apply a binomial filter, which is a Gaussian shaped filter that gives greater weight to predictions adjacent to the prediction and less weight to the predictions further away. Binomial filter coefficients are formed using a binomial expansion:

$$\left(\frac{1}{2},\frac{1}{2}\right)^{N} \quad (1)$$

The effect of both delay only and combined delay and smoothing on a set of predicted values can be seen in Figure 1.

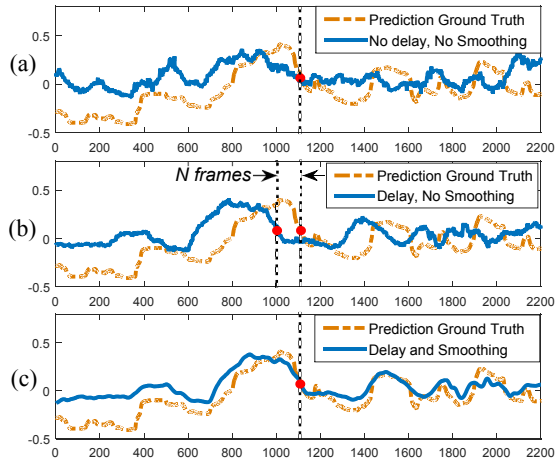

Figure 1. *Effect of annotation delay compensation on a set of predicted arousal ratings.* ***(a)*** *Predictions without delay and smoothing are noisy and not well matched with the ground truth labels.* ***(b)*** *Applying temporal shifts to the training data improves system performance but results in predictions which are advanced in time compared to their ground truth.* ***(c)*** *Applying a binomial filter to these predictions not only smooths the output but resolves the synchronization issue.*

3.4 Relevance Vector Machine

Relevance Vector Machines (RVM) are gaining popularity in continuous emotion prediction and have been shown to offer a wide range of advantages over SVR [15-17, 24]. The goal of RVM training is to learn the regression function:

$$\boldsymbol{y}(\boldsymbol{x}_*, \boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{\phi}(\boldsymbol{x}_*) + \boldsymbol{\epsilon} \quad (2)$$

in which $\boldsymbol{x}_*$ represents a multi-dimensional feature vector, $\boldsymbol{\omega} = [\omega_1, \cdots \omega_P]^T$ is an estimated set of sparse regression parameters, $\boldsymbol{\phi} = [\boldsymbol{\phi}_1(\boldsymbol{x}_*), \cdots \boldsymbol{\phi}_P(\boldsymbol{x}_*)]^T$ is a set of (potentially non-linear) transforms performed on $\boldsymbol{x}_*$ and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T$ is the training noise terms. In the Bayesian approach used in RVM's all noise terms are assumed to be distributed Gaussian $\sim N(0, \sigma^2)$.

RVMs learn a sparse representation of $\boldsymbol{\omega}$ where the majority of the ω_m are zero. This is achieved by giving $\boldsymbol{\omega}$ a zero-mean Gaussian prior which encourages sparsity by declaring smaller weights as more probable [16]:

$$\boldsymbol{p}(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{i=1}^{P} N(0, \alpha_i^{-1}) \quad (3)$$

$\boldsymbol{\alpha}$ is the inverse variance hyperparameter and is analogous to regularization terms in SVR or ridge regression.

RVM presents the learnt regression model as the most relevant set of extracted feature dimensions, meaning the technique explicitly performs both dimensionality reduction and feature selection without the need for a held-out validation data subset. In the context of the AV+EC 2015 challenge, this is a desirable quality as it helps to minimize the chances of overfitting during system development.

The training phase of the RVM Regression Model searches for $\boldsymbol{\alpha}_{MP}, \sigma^2_{MP}$, the *most probable* (MP) values of $\boldsymbol{\alpha}$ and σ^2 using iterative Bayesian inference procedure. In the testing phase, $\boldsymbol{\alpha}_{MP}, \sigma^2_{MP}$, are used to both make a prediction and to estimate a level of uncertainty associated with that prediction.

3.5 Feature Level Fusion

The first fusion system investigates the advantages of feature level concatenation. In this system different modalities are combined in the front end and used to train a SVR or RVM. During testing, the same set of features were combined before generating predictions (Figure 2).

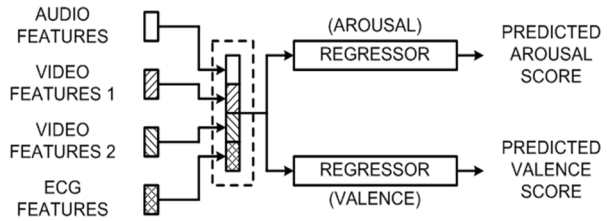

Figure 2: *Block diagram showing the feature level fusion strategy used to combine information from different modalities for the task of continuous emotion prediction*

3.6 Decision Level Fusion

Decision level fusion techniques have been shown to improve the performance in many continuous emotion prediction systems [1, 8, 12]. In this method, separate models are trained and tested on for each modality and a further regressor is trained to produce the overall prediction from the set of predictions from various modalities (Figure 3).

When performing this kind of fusion, multicollinearity problems, caused by the high correlations between prediction scores of different modalities, may arise [19]. Therefore we use either a RVM or a *Regularized Linear Regression* (RLR) [25] system to learn fusion weights.

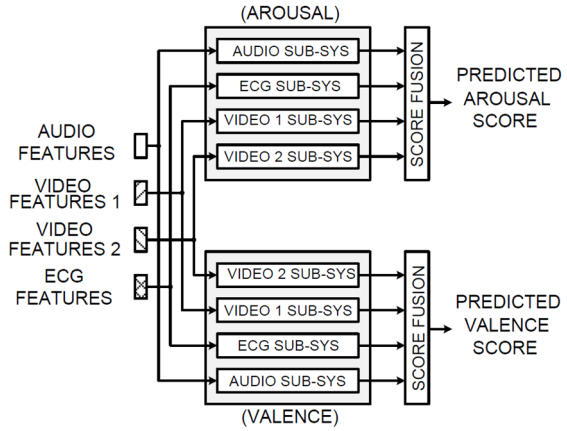

Figure 3: *Block diagram showing the decision level fusion strategy used to combine information from different modalities for the task of continuous emotion prediction*

3.7 Output-Associative Fusion

Output-Associative (OA) fusion techniques, which take into account the contextual and temporal dependencies that exist within and between predicted arousal and valence values when performing fusion, are gaining popularity in continuous emotion prediction [8, 12, 15]. OA-fusion is an extension of decision-level fusion and is achieved by learning fusion weights on an OA-matrix. The OA-matrix is formed by output associative vectors from a set of initial predictions, taken from each dimension and modality (Figure 4).

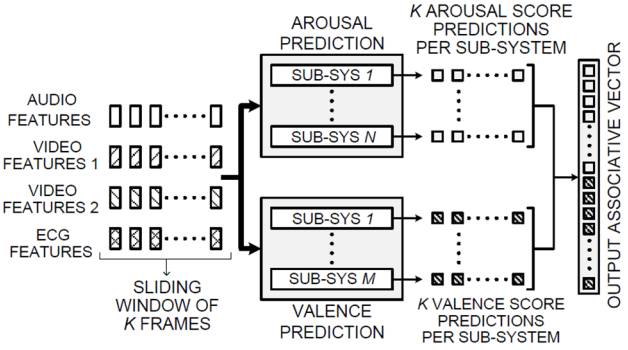

Figure 4. *Block diagram showing the formation of an* $(M + N) \times K$ *output associative vector from a set of multimodal predictions for use in an OA fusion or OA regression system*

Given a set of arousal predictions $\tilde{\boldsymbol{y}}_a = [\tilde{y}_a^1, \cdots, \tilde{y}_a^P]^T$, and valence predictions $\tilde{\boldsymbol{y}}_v = [\tilde{y}_v^1, \cdots, \tilde{y}_v^P]^T$, the k-th set of values are the set of initial arousal and valence predictions, learnt from one modality (Figure 4). The OA-matrix associated with the M-th modalities Y_{OA}^M can be formed as below:

$$\boldsymbol{Y}_{OA}^M = \begin{bmatrix} [\tilde{\boldsymbol{y}}_a^{1+t}]^T, [\tilde{\boldsymbol{y}}_v^{1+t}]^T \\ [\tilde{\boldsymbol{y}}_a^{2+t}]^T, [\tilde{\boldsymbol{y}}_v^{2+t}]^T \\ \vdots \qquad\qquad \vdots \\ [\tilde{\boldsymbol{y}}_a^{P+t}]^T, [\tilde{\boldsymbol{y}}_v^{P+t}]^T \end{bmatrix} \tag{4}$$

where $\tilde{\boldsymbol{y}}_a^{i+t}$ and $\tilde{\boldsymbol{y}}_v^{i+t}$ are a set of the temporal continuous prediction values taken from the range $[i-t, \cdots, i+t]$. A complete OA-matrix can then be formed by combining the OA matrices from all modalities:

$$\boldsymbol{Y}_{OA} = [\boldsymbol{Y}_{OA}^1, \boldsymbol{Y}_{OA}^2, \boldsymbol{Y}_{OA}^3, \boldsymbol{Y}_{OA}^4] \tag{5}$$

Fusion between modalities is simply a case of training a regressor with the OA-matrix (Figure 5). Again, to minimize multicollinearity effects, either a RVM or RLR was used to learn the fusion weights.

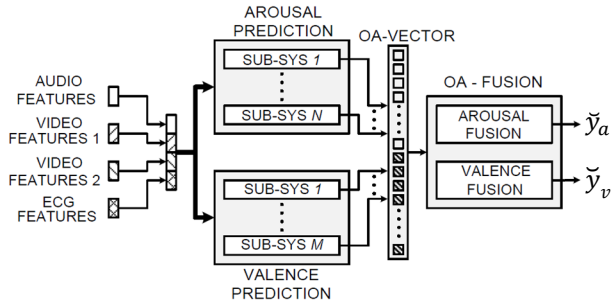

Figure 5: *Block diagram showing the output-associative fusion strategy used to combine information from different modalities for the task of continuous emotion prediction. Note* $\breve{\boldsymbol{y}}_a$ *and* $\breve{\boldsymbol{y}}_v$ *represent the predicted arousal and valence scores respectively*

3.8 Output-Associative Regression

The final fusion strategy is a combined feature-level fusion, decision-level fusion and OA fusion scheme, herein referred to as *output-associative regression* (OA-Regress.). This system is an extension of the OA fusion, in which the OA matrix is concatenated with the input feature space to learn the fusion weights (Figure 6). Fusion of modalities using this system will be performed using the Output-associative Relevance Vector Machine (OA-RVM, [15]).

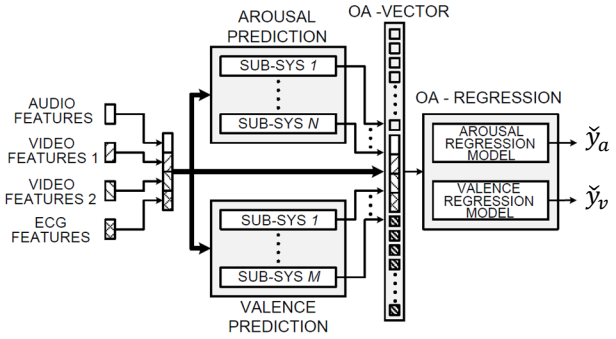

Figure 6: *Block diagram showing the output-associative regression strategy used to combine information from different modalities for the task of continuous emotion prediction. Note* $\breve{\boldsymbol{y}}_a$ *and* $\breve{\boldsymbol{y}}_v$ *represent the predicted arousal and valence scores respectively.*

The OA-RVM technique extends this contextual and temporal mapping performed in OA fusion to also incorporating the relationship between the input features space when updating the prediction values:

$$\breve{\boldsymbol{y}}_a^i = (\boldsymbol{\omega}_a)^T \boldsymbol{\phi}(\boldsymbol{x}_*) + (\boldsymbol{\varphi}_a)^T (\tilde{\boldsymbol{y}}_t^a) + (\boldsymbol{\psi}_a)^T (\tilde{\boldsymbol{y}}_t^v) + \boldsymbol{\epsilon} \tag{6}$$

$$\breve{\boldsymbol{y}}_v^i = (\boldsymbol{\omega}_v)^T \boldsymbol{\phi}(\boldsymbol{x}_*) + (\boldsymbol{\varphi}_v)^T (\tilde{\boldsymbol{y}}_t^a) + (\boldsymbol{\psi}_a)^T (\tilde{\boldsymbol{y}}_t^v) + \boldsymbol{\epsilon} \tag{7}$$

where $\tilde{\boldsymbol{y}}_t^a$ and $\tilde{\boldsymbol{y}}_t^v$ are the temporal independently-learnt set of arousal and valence prediction values, continuous on the range $[i-t, \cdots, i+t]$.

OA-RVM therefore uses the past, current and future prediction context associated with input feature frames, as well as the input features, to update a prediction result. Prediction using the non-causal relationship has been shown to be superior

to RVM and SVR when performing continuous emotion prediction [15]. The work presented within this paper aims to reinforce the usefulness of the OA-RVM framework and furthermore explore this paradigm in terms of a multimodal fusion technique.

4. EXPIREMENTAL CONDITIONS

4.1 Database

The corpus used in the AV+EC 2015 challenge is part of the *Remote Collaboration and Affective Interaction* (RECOLA) database [11]. In this database, data was collected during spontaneous dyadic interactions where multimodal signals were recorded, including audio, video, ECG and EDA. There are 27 subjects in total, which are evenly divided into training, development (devel.) and test partitions (9 speakers for each). The total recording comprises 5 minutes for each subject; all recordings have been continuously rated for arousal and valence in 40ms, resulting in 7501 pairs of affective score per file [9].

4.2 Performance Metric

The performance measure adopted is the *concordance correlation coefficient* (CCC), which combines the *Pearson correlation coefficient* (ρ) and the *mean square error* (MSE).

$$CCC = \frac{2Cov(\boldsymbol{x}, \boldsymbol{y})}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (8)$$

The baseline CCC's have been provided by the challenge organizers for both development set and test set (Table 1).

Table 1. *Challenge Baselines CCC's of the AV+EC 2015 development and testing partitions, reproduced from [9]*

	Arousal	Valence
Devel. Set	0.476	0.461
Test. Set	0.444	0.382

4.3 Key Experimental Settings

In this research, SVR and RVM were initially used as the regression methods. Before training, all training features were normalized to [0, 1] and the normalization coefficients were used to normalize the testing data. For SVR, 1 out of each 20 frames of training data were selected for training, for reasons of computational efficiency with negligible performance drop. A linear kernel was used and C was set to 0.005 and 0.05 for arousal and valence respectively, based on optimizing the performance with delay and smoothing in the development set in the range of $[10^{-4},1]$. The number of RVM training iterations was set to 30 for arousal and 40 for valence, based on the best performance on the development data. When performing OA fusion and OA-Regress K, OA window size (Figure 4), was set to 121 for both arousal and valence and the number of RVM training iterations was tailored for each system and ranged between 10-100.

5. SYSTEM DEVELOPMENT

5.1 Feature Level Fusion Results

The first system proposed was to concatenate all modalities into a single feature matrix to train a model. The back-end regression methods were SVR and RVM. As can be seen in Table 2, the RVM back-end outperformed the SVR, especially for valence. Notably, all non-delay compensated, feature level fusion systems outperformed single modality systems in our initial development phase, but none were able to match the challenge development set baseline.

Table 2. *A comparison of CCC's of comparing feature level fusion in either a SVR or RVM back-end for the AV+EC 2015 development set.*

System	Arousal	Valence
SVR	0.276	0.145
RVM	0.340	0.361

5.1.1 Effect of Delay Compensation

Before training, a delay was introduced to the training data by frame dropping. The predictions generated by a system trained on shifted inputs features were then realigned using a smoothing filter. The pre-processing potentially increases reliability of the training labels if the introduced delay is equivalent to the annotation delay, whilst the post-processing produces a smoother and aligned set of predictions.

Initial investigations focused on searching for the best delay for arousal and valence, which were achieved using a number of delay values in pre-processing and post-processing. In the experiment, 16 delay values within 0 to 6 seconds with 0.4-second increments were used. However, unlike previous studies [13], rather than only using facial features, multimodal features were used to estimate the best delay values for arousal and valence. SVR was used as the regression method.

A substantial improvement in CCC was seen with the introduction of delay compensation, increasing from 0.276 to 0.691 for arousal and from 0.145 to around 0.5 for valence (Figure 7). Also it shows that valence rating responds more rapidly than arousal, which is consistent with previous studies, e.g. [13]. Based on these results, we selected a 4-second delay for arousal and a 2-second delay for valence as the optimal delay value. Unless specifically stated, these delay values were used in all subsequent systems.

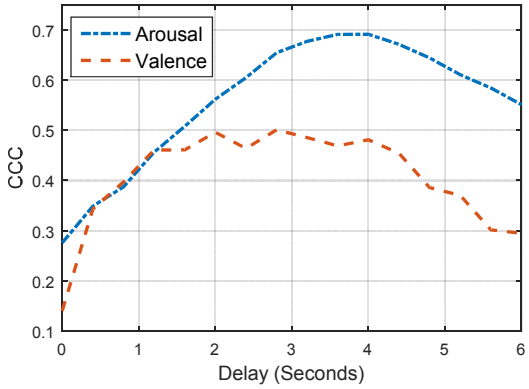

Figure 7. *Delay compensation using frame shift and smoothing. The best delay for arousal was 4s and the best delay value for valence was 2s***.**

Compensating for delay in our feature level fusion systems produced a substantial increase in system performance (Table 3). Both the SVR and RVM feature level fusion system easily outperformed the challenge development set baseline.

Table 3. *CCC's found for the AV+EC 2015 development set when combining delay compensation with either a SVR or RVM feature level fusion system.*

Front-End System	Arousal	Valence
SVR	0.691	0.496
RVM	0.683	0.508

5.2 Decision Level Fusion Results

In Section 5.1, all systems were trained using concatenated features from all modalities. The second proposed system, decision level fusion (Section 3.6), treated the modalities separately and fused each set of predictions. This fusion method was implemented using either a SVR or RVM as the feature level regression method, and either a RLR or RVM to learn the fusion weights.

The performance of the SVR decision level fusion systems is comparable with the SVR feature level fusion system (Table 4). However, there was a decrease in the performance of the decision level fusion RVM systems in comparison to the RVM feature level fusion systems (Table 4). This inconsistency could be due to the small amount of training data (4 instances) used to train the fusion regression systems.

Table 4. *CCC's found for the AV+EC 2015 development set found when performing decision level fusion using either a RVM or RLR to learn the fusion weights.*

Front-End Systems	Arousal		Valence	
	RLR	*RVM*	*RLR*	*RVM*
SVR	0.684	0.667	0.524	0.436
RVM	0.274	0.648	0.495	0.458

5.3 OA-Fusion Results

OA-fusion techniques seek to utilize the correlations between arousal and valence values [1]. Significant correlations can be seen when correlating the AV+EC 2015 gold standard arousal and valence scores in both the training, $r = 0.421$, $p < 0.001$, and development partitions $r = 0.556$, $p < 0.001$. Given the strength of these correlations, it can be expected that systems which utilize either OA fusion or OA regression will outperform the feature level and decision level fusion systems.

When performing OA fusion on the same four system configurations as in Section 5.2, there is an increase in system performance (Table 5). Again, all OA fusion systems tested outperformed the challenge development set baseline. These results indicate that consistent significant improvements can be obtained when OA fusion is used, in comparison to decision level fusion systems.

Table 5. *CCC's found for the AV+EC 2015 development set found when performing OA fusion using either a RVM or RLR to learn the fusion weights.*

Front-End Systems	Arousal		Valence	
	RLR	*RVM*	*RLR*	*RVM*
SVR	0.736	0.718	0.615	0.509
RVM	0.447	0.710	0.578	0.535

5.4 OA-Regression Results

In the fourth system, OA-fusion was extended to OA-regression. These systems combined feature-level fusion and OA fusion in order to further improve the performance (Section 3.8). Note that all OA-Regression systems used an OA-RVM framework [15], which was extended to fuse a set of predictions learnt from each modality.

OA-Regression provided a further increase in system performance when compared to the other fusion methods (Table 6). Interestingly, when compared to the OA-fusion the results appear very consistent across the different front-end system configurations. These results confirm the usefulness of OA-RVM for performing continuous emotion predictions [15].

Table 6. *Comparison of CCC's for the AV+EC 2015 development set found using different OA-Regression systems using a OA-RVM set-up to learn the fusion weights*

Front-Ends System	Arousal	Valence
1-RVM (Feature Fusion)	0.743	0.600
4-SVR's	0.766	0.655
4-RVM's	0.742	0.608
4-SVR's + 4-RVM's	0.753	0.588

6. AUDIO-ONLY PREDICTION

Acoustic features played a crucial role in arousal predictions throughout all our experiments; results (not shown) found that systems trained with acoustic features only were comparable with those using multimodal features. This motivated further investigation into other audio features for arousal prediction. Links between acoustic features and arousal prediction are also well reported in the literature [1, 3]. In this section, two alternative sets of acoustic features were chosen for testing in a series of prediction systems.

The first was a set of 45 frame-based features was extracted using open-source software *VoiceSauce* [21], due to their effectiveness in previous paralinguistic research [26]. These features included: F0, F1-F3, formant amplitudes, harmonic amplitudes, cepstral peak prominence, and harmonic-to-noise ratios. Each feature output was then realigned by down sampling (1 out of 4 frames) to match the arousal ground truths. Note that all features were extracted using 20ms frames with a 10ms overlap. A feature-level fusion arousal-only RVM system was used, with the number of training iterations set to 100.

Frame-based audio features proved effective for arousal prediction (Table 7). The 13 *VoiceSauce features* (VSF) were able to match the challenge audio-only development set CCC of 0.287 [9]. By combining VSF with a set of *Shifted Delta Cepstrum* (SDC) features extracted, using *N-P-D-K* setting of 16-3-3-5, from a set of frame-level features including 4 MFCCs, 4 energy features, 4 Spectral Centroid Frequency and 4 Spectral Centroid Amplitude features, we find a *frame-level* audio system that outperforms the challenge development set baseline (Table 7). The inclusion of SDC's highlights the importance of including temporal contextual information.

Table 7. *CCC's found for the AV+EC 2015 development set using frame-based audio for arousal predictions using RVM*

Audio Features	Arousal Prediction
13 VSF	0.294
96 SDC	0.470
13 VSF + 96 SDC	0.494

Additionally, the eGeMAPS audio feature set was compared with the *Computational Paralinguistics Challenge 2013* (ComParE 2013) audio feature set. The ComParE feature set contains 65 Low Level Descriptor features and their first-order deviation [27], and previously performed strongly on predicting arousal on the RECOLA database [11]. Statistics of the features (e.g. mean, standard deviation, etc) were then extracted using the same window size (3s) and overlap rate (40ms) as the baseline features set. Both of these feature sets were tested separately and fused in an OA-RVM Regressor (Table 8). Not surprisingly, given previous results [11], ComParE features performed well when predicting arousal. Interestingly, fusion of both systems performed adequately on valence prediction, despite the fact that valence is typically better associated with video features [1, 7].

Table 8. *Comparison of audio feature sets on the AV+EC 2015 development set using OA-RVM regressor*

Audio Features	Arousal	Valence
eGeMAPS	0.689	0.317
ComParE 2013	0.791	0.272
ComParE 2013 + eGeMAPS	0.776	0.386

7. AV+EC 2015 CHALLENGE RESULTS

During system development significant gains in system performance were found by combining a*nnotation delay compensation* and *output-associative based fusion*. Therefore our five official entries to the AV+EC 2015 challenge combine both delay compensation and OA-based multimodal fusion.

The *first* system, *4-SVRs + OAfus*, is an OA-fusion system (Section 3.7). In this system the feature level predictions for each modality are performed using a SVR. The 8 sets of predictions are fused in an OAmatrix (Figure 4) and used to train separate RLR's per dimension.

The *second* system, *1-RVM + OAreg*, is an OA-Regress system (Section 3.8). The separate modalities are fused at the feature level and used to train an OA-RVM [15] per dimension.

The *third* system, *1-RVM + OAreg inc. ComParE*, is an extension of the second system. In this system the separate modalities are fused at the feature level along with the ComParE 2013 features and used to train an OA-RVM [15] per dimension.

The *fourth* system, *4-SVR's, 4-RVM's + OAreg*, is also an OA-Regress system (Section 3.8). In this system feature level predictions from each modality are performed using both a SVR and a RVM. The resulting 16 sets of predictions are fused in an OAmatrix (Figure 4) before being concatenated with the four sets of features and used to train an RVM per dimension.

The *fifth* system, *5/4-SVR's, 5/4-RVM's + OAreg*, is an extension of the fourth system. In this system arousal feature level predictions are learnt for each modality and the ComParE 2013 features using both a SVR (C= 0.05) and a RVM, resulting in 10 sets of predictions. Valence feature level predictions for each modality are learnt using both a SVR (C=0.01) and a RVM, resulting in 8 sets of predictions. The 18 predictions sets are fused in an OAmatrix (Figure 4) combined with the four challenge features sets and used to train an RVM per dimension.

All systems show a substantial increase in CCC over the challenge development set baseline (Table 9). Interestingly, the systems which used the challenge feature sets, provided consistent results despite being used in different configurations i.e. all three systems used different combinations of feature level and fusion level prediction methods. This consistency highlights the importance of OA modeling to our systems.

The highest development CCC's were produced with the *1-RVM + OAreg inc. ComParE* system. The stronger performances of both systems which include the ComParE 2013 is not surprising given the strong performance of the feature set in predicting arousal scores on the development set (Table 8).

Table 9. *Comparison of AV+EC 2015* ***development set*** *CCC's for systems chosen for challenge entry*

System	Arousal	Valence
Baseline [9]	0.476	0.461
4 SVRs + OAfus	0.736	0.615
1-RVM + OAreg	0.743	0.600
1-RVM + OAreg inc. ComParE	**0.845**	**0.642**
4-SVR's, 4-RVM's + OAreg	0.753	0.588
5/4-SVR's, 5/4-RVM's + OAreg	0.809	0.615

When performing prediction on the AV+EC 2015 test set, all systems were trained with both the training and development partitions, all other settings were set as previous experiment. All systems performed significantly better than the challenge baseline on the test set (Table 10).

The best arousal (0.740) and valence (0.588) test set CCC's were found using the *4-SVR's, 4-RVM's + OAreg* system. This result represents a 66.7% and a 53.9% relative improvement for arousal and valence predictions respectively over the AV+EC 2015 baseline. The strong and consistent performance of all our systems highlights the advantages afforded through exploiting the contextual and temporal dependencies that exist within and between the predicted values from the different modalities.

Table 10. *Comparison of AV+EC 2015* ***test set*** *CCC's generated using a range of systems which incorporate both annotation delay compensation and OA fusion framework*

System	Arousal	Valence
Baseline [9]	0.444	0.382
4 SVRs + OAfus	0.711	0.558
1-RVM + OAreg	0.739	0.535
1-RVM + OAreg inc. ComParE	0.733	0.569
4-SVR's, 4-RVM's + OAreg	**0.740**	**0.588**
5/4-SVR's, 5/4-RVM's + OAreg	0.719	0.569

8. CONCLUSIONS

Combining information from a range of different behavioral and physiological modalities has been shown to improve accuracy when performing continuous emotion prediction [1, 2, 7, 8]. Our systems submitted under AV+EC 2015 test conditions outperformed both the challenge development and testing baselines. We speculate that the stronger performance of our systems is due to both annotator delay compensation and multimodal output-associative based fusion.

The combination of pre-processing frame shifting and post-processing filtering for delay compensation provided significant improvements for both arousal and valence prediction. The optimal delay compensation, estimated using multimodal features, was found to be 4s for arousal and 2s for valence. Results show that an output-associative (OA) fusion framework, which exploits emotion dimension dependencies and temporal information of predictions, further improved our system performance. An OA regression system, which combined feature-level fusion, decision level fusion and OA fusion, gave the best performance on the AV+EC 2015 test set. Additional investigations into frame-based audio features for arousal-only prediction showed a strong correlation between audio modality and arousal, which is consistent with previous studies [3].

Future work will explore delay compensation further. It is possible, given evaluators make annotations based on only audio and video signals, performance increases can be found by searching for the best delay for each modality and treating them separately at the pre-processing and post-processing stages. We will continue our exploration of arousal prediction using frame-level audio features. We will also investigate whether delay compensation can be included within the OA matrix to further improve the usefulness of this fusion framework.

9. ACKNOWLEDGEMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

10. REFERENCES

[1] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing,* vol. 31, pp. 120-136, 2013.

[2] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: an overview," *International Journal of Synthetic Emotions,* vol. 3, pp. 1-17, 2012.

[3] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 1085-1088.

[4] H. Gunes, M. A. Nicolaou, and M. Pantic, "Continuous analysis of affect from voice and face," in *Computer Analysis of Human Behavior*, ed: Springer, 2011, pp. 255-291.

[5] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing,* vol. 31, pp. 137-152, 2013.

[6] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," in *IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 2662-2667.

[7] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 31-38.

[8] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing,* vol. 2, pp. 92-105, 2011.

[9] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne*, et al.*, "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[10] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie*, et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech*, 2008, pp. 597-600.

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proceedings of Face & Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[12] 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*, 2012, pp. 501-508.

[13] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing,* 2014.

[14] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Tenth IEEE International Symposium on Multimedia*, 2008, pp. 228-235.

[15] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing,* vol. 30, pp. 186-196, 2012.

[16] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," in *Advanced lectures on machine Learning*, ed: Springer, 2004, pp. 41-62.

[17] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Relevance Vector Machine for Depression Prediction," in *Proc. Interspeech*, Dresden, Germany, 2015.

[18] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi*, et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters,* 2014.

[19] C. J. Merz and M. J. Pazzani, "A principal components approach to combining regression estimates," *Machine Learning,* vol. 36, pp. 9-32, 1999.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. A. e, C. Busso*, et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing, to appear,* 2015.

[21] Y.-L. Shue, Patricia Keating, C. Vicenik, and K. Yu, "VoiceSauce: A Program for Voice Analysis," in *ICPhS*, 2011.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835-838.

[23] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 19-26.

[24] F. Wang, W. Verhelst, and H. Sahli, "Relevance vector machine based speech emotion recognition," in *Affective computing and intelligent interaction*, Memphis, US, 2011, pp. 111-120.

[25] C. M. Bishop, *Pattern recognition and machine learning*: Springer, 2006.

[26] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Voice source features for cognitive load classification," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5700-5703.

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval*, et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.