

# Affective computing using speech and eye gaze: a review and bimodal system proposal for continuous affect prediction<sup>☆</sup>

Jonny O' Dwyer\*, Niall Murray, Ronan Flynn

*Department of Electronics and Informatics,*

*Athlone Institute of Technology,*

*Athlone, Ireland*

---

## Abstract

Speech has been a widely used modality in the field of affective computing. Recently however, there has been a growing interest in the use of multi-modal affective computing systems. These multi-modal systems incorporate both verbal and non-verbal features for affective computing tasks. Such multi-modal affective computing systems are advantageous for emotion assessment of individuals in audio-video communication environments such as teleconferencing, healthcare, and education. From a review of the literature, the use of eye gaze features extracted from video is a modality that has remained largely unexploited for continuous affect prediction. This work presents a review of the literature within the emotion classification and continuous affect prediction sub-fields of affective computing for both speech and eye gaze modalities. Additionally, continuous affect prediction experiments using speech and eye gaze modalities are presented. A baseline system is proposed using open source software, the performance of which is assessed on a publicly available audio-visual corpus. Further system performance is assessed in a cross-corpus and cross-lingual experiment. The experimental results suggest that eye gaze is an effective supportive modal-

---

<sup>☆</sup>Paper submission for review

\*Corresponding author

*Email addresses:* [j.odwyer@research.ait.ie](mailto:j.odwyer@research.ait.ie) (Jonny O' Dwyer),  
[nmurray@research.ait.ie](mailto:nmurray@research.ait.ie) (Niall Murray), [rflynn@ait.ie](mailto:rflynn@ait.ie) (Ronan Flynn)

ity for speech when used in a bimodal continuous affect prediction system. The addition of eye gaze to speech in a simple feature fusion framework yields a prediction improvement of 6.13% for valence and 1.62% for arousal.

*Keywords:* Affective computing, Speech, Eye gaze, Bimodal, Fusion

---

## 1. Introduction

Affective computing involves the computational analysis, recognition, prediction, and synthesis of emotion. This field brings together research ranging from artificial intelligence to social science (Poria et al., 2017). Within affective computing, emotion recognition involves the computational recognition of an emotion that has been expressed by a subject. The emotions to be recognized may be classed, for example, as happy, sad, or angry. Multi-dimensional classes such as high/medium/low arousal or valence are also possible. Continuous affect prediction, another form of emotion computation, is the task of predicting a continuous numerical value for emotion dimensions, examples of which include arousal and valence. Arousal is a measure of how calming or exciting an experience is and valence is a measure of how positive or negative an experience is (Kossaifi et al., 2017).

Affective computing tasks incorporating speech are now well developed. There are nearly thirty years of literature available on this topic (Fernandez and Picard, 2011). A recent review carried out by Feraru et al. (2015) showed that 66% of all international languages are represented by an affective speech dataset. These speech datasets may be focused on emotion classification (Burkhardt et al., 2005; Soleymani et al., 2012) or continuous affect prediction within emotion dimensions such as arousal and valence (McKeown et al., 2012; Ringeval et al., 2013b; Valstar et al., 2014). The emotions within the datasets may be acted, elicited, or natural as specified by the corpus data collection protocol. Within continuous affect prediction, speech has been found to perform well for arousal prediction, but less so for valence (Mencattini et al., 2017; Ringeval et al., 2015a). Research into speech as a modality for affective computing is fa-

cilitated by the availability of research tools such as openSMILE (Eyben et al., 2013), which is used to extract affective feature sets from speech. Examples of such feature sets include AVEC 2014 (Valstar et al., 2014), ComParE (Schuller et al., 2016), and GeMAPS (Eyben et al., 2016).

Increasingly, multi-modal approaches to emotion classification and continuous affect prediction have been employed by the affective computing research community (Nicolaou et al., 2011; Ranganathan et al., 2016; Ringeval et al., 2015a; Thushara and Veni, 2016). Datasets for the training and evaluation of multi-modal systems include YouTube (Morency et al., 2011), SEMAINE (McKeown et al., 2012), MAHNOB-HCI (Soleymani et al., 2012), RECOLA (Ringeval et al., 2013b), and AVEC 2014 (Valstar et al., 2014). However, multi-modal affect recognition that includes eye gaze features extracted from video has not received much attention from the research community, based on available literature. This is surprising given the cost effective, non-intrusive nature of extracting data for this modality and the amount of audio-visual corpora available today. The AVEC 2016 (Valstar et al., 2016) challenge did provide OpenFace (Baltruaitis et al., 2016) eye gaze approximation features from video, but they were only provided for the depression corpus and not for the affect prediction challenge. Affect prediction using eye gaze from video deserves further investigation as shown by O’Dwyer et al. (2017a), particularly for continuous prediction of valence.

This work has two key contributions. Firstly it presents a review of research within the emotion classification and continuous affect prediction sub-fields of affective computing for both speech and eye gaze modalities. Feature sets, classification and regression methodologies, and performance levels in the literature are reviewed. When carrying out the literature review, a gap in the use of eye gaze for continuous affect prediction was observed. Secondly this work presents unimodal and bimodal continuous affect prediction experiments using speech and eye gaze from audio-visual sequences. Based on these experiments, a bimodal speech and eye gaze affective computing system using the CURRENNT tool kit (Weninger, 2015) is proposed. The results obtained clearly show the

benefits of using eye gaze from video combined with speech in a bimodal continuous affect prediction system. Evaluations of the system include cross-corpus and cross-lingual experiments with promising results for arousal prediction.

This paper is structured as follows. Section 2 provides a review of affective computing using speech. In Section 3, multi-modal affective computing that includes speech information is reviewed. Section 4 presents a review of the observed gap in the literature, namely, affective computing using eye gaze. The proposal of a novel speech and eye gaze continuous affect prediction system, along with evaluation data, is presented in section 5. Section 6 presents the results and discussion of this work. Concluding remarks end this paper in Section 7.

## **2. Unimodal affective computing using speech**

This section presents a review of emotion recognition and continuous affect prediction using speech. Data and methods used are presented, along with the results achieved for affective computing systems that were based only on speech features. A summary of each section of this review is given in Tables 1 and 2.

### *2.1. Emotion recognition using speech*

Emotion recognition involves the computational recognition of an emotion that has been expressed by a subject. The emotions to be recognized are classed discretely, such as happy, sad, or angry. Within the reviewed literature, both intra-corpus and cross-corpus emotion recognition experiments are presented. Intra-corpus evaluation is where an individual corpus is split into known training and unknown test partitions, the partitions are then used for machine learning model training and testing respectively. Cross-corpus evaluation involves multiple different corpora in varying configurations, for example, a model training partition might be gathered from one corpus and an unseen test set from a different corpus. This section reviews different methods applied to speech emotion recognition.

Fayek et al. (2017) investigated the application of end-to-end deep learning for speech emotion recognition. The IEMOCAP (Busso et al., 2008) corpus was used for the experiments and feedforward neural network (FF), convolutional neural network (CNN), and long short-term memory recurrent neural network (LSTM-RNN) topologies were employed for evaluation and comparison. The results presented showed CNN performed best for the network architectures evaluated for the recognition of anger, happiness, sadness, and neutral emotion classes. CNN was also shown to outperform deep neural network (DNN) and extreme learning machine (ELM), support vector machine (SVM), and hierarchical binary decision tree approaches for frame-based emotion recognition. A 60.89% test set unweighted average recall (UAR) was achieved for frame-based emotion recognition using the CNN system.

Motamed et al. (2017) introduced an optimized brain emotional learning model (BEL) that merged an adaptive neuro-fuzzy inference system (ANFIS) and multilayer perceptron (MLP) model for speech emotion recognition. The ANFIS was intended to model the human amygdala and orbitofrontal cortex in order to make rules that were passed to the MLP network. Mel-frequency cepstral coefficients (MFCCs) (B. Davis and Mermelstein, 1980) from speech were used as input to the system for the recognition tasks, which were performed on the Berlin EMO-DB (Burkhardt et al., 2005). The proposed algorithm performed better on average over all emotions when compared to ANFIS, MLP, BEL, BEL based on learning automata (BELBA), K-nearest-neighbour (KNN), and SVM approaches (72.5% accuracy for anger, happiness, and sadness). KNN and SVM approaches used for comparison did outperform the proposed algorithm for both anger and happiness class recognition however.

C.K. et al. (2017) presented work on higher order spectral features and feature selection approaches for emotion and stress recognition from speech. The authors added 28 bispectral and 22 bicoherence features to the Interspeech 2010 speech feature set (Schuller et al., 2010a) and reported improved emotion recognition when compared to the Interspeech set alone. Additionally, the authors carried out biogeography-based optimisation (BBO), particle swarm

optimisation (PSO), and BBO PSO hybrid optimisation techniques for feature selection. The feature additions and feature selection techniques were assessed using speaker-dependent, speaker-independent, male-dependent, and female-dependent approaches on the EMO-DB (Burkhardt et al., 2005), SAVEE (Haq et al., 2008), and SUSAS (Hansen and Bou-Gazale, 1997) corpora on an intra-corpus basis. SVM and ELM algorithms were used for model generation. The best result achieved on the EMO-DB (Burkhardt et al., 2005) set resulted in improvements compared to previous work on that set for speaker independent experiments (93.25% recognition rate). Also, a 100% recognition rate was achieved using this method in a speaker-dependent experiment. These results were achieved using BBO optimisation and a SVM learning scheme.

Chakraborty et al. (2016) evaluated a knowledge-based framework for the recognition of angry, happy, neutral, and sad emotion classes from speech. The proposed framework adds linguistic and time-lapse information about the conversation to the speech features. To check the validity of the approach the authors provided this linguistic and time-lapse information to annotators of an emotional corpus. It was observed that agreement between the annotators, as measured by Fleiss' Kappa (L. Fleiss, 1971), improved during the annotation process when the linguistic and time-lapse information was provided. This preliminary experiment was carried out on the Interactive Voice Response Speech Enabled Enquiry System (IVR-SERES) (Bhat et al., 2013). The utterance datasets used for the final experiments included one acted dataset, EMO-DB (Burkhardt et al., 2005), and two spontaneous emotion datasets, IVR-SERES (Bhat et al., 2013) and Call Center (Kopparapu, 2014). Within the experimental framework, the authors multiplied each emotion by a weight vector depending on how long the utterance was. The authors also extracted words from speech using automatic speech recognition (ASR) and then assigned a prominence measure, for example +, -, or 0, to words associated with specific emotions. The prominence measure contains information about how much emotion is contained in a word. The speech features for the system were that of the Interspeech 2009 Challenge (Schuller et al., 2009). The performance of the system always

increased with the addition of the time-lapse and emotion prominence from linguistic content. SVM, ANN, and KNN machine learning schemes were used for the experiments. The best results obtained were 82.1% (IVR-SERES), 78.1% (Call Center), and 87.3% (Emo-DB) using a combination of all of the machine learning methods for classification.

Vlasenko et al. (2016) investigated cross-corpus arousal classification using the VAM (Grimm et al., 2008) dataset for training and the EMO-DB (Burkhardt et al., 2005) dataset for testing. During the experiments the authors trained models on data from the full VAM dataset, and two subsets of this dataset, VAM I, and VAM II. The VAM I and VAM II subsets contained very intense and intense emotions respectively. The authors describe the intensity in the VAM subsets as how well the emotions were conveyed. A hidden Markov model (HMM) approach was taken for low/high arousal classification. The authors concluded that there were big classification performance gaps for arousal models trained on spontaneous data with different emotional intensities. The entire VAM dataset for training provided the best test set performance UAR of 86%.

Song et al. (2016) investigated cross-corpus and cross-lingual emotion recognition using transfer learning combined with novel non-negative matrix factorization (NMF) on EMO-DB (Burkhardt et al., 2005), eNTERFACE (Martin et al., 2006), and FAU Aibo (Schuller et al., 2009) emotional speech corpora. EMO-DB and FAU Aibo corpora were the German datasets for the experiment and eNTERFACE represented English. Transfer graph regularized NMF (TGNMF) and transfer constrained NMF (TCNMF) additions to NMF were presented and evaluated for emotion recognition. The Interspeech 2010 feature set (Schuller et al., 2010a) was the affective speech feature set used for the experiments. In addition to the proposed approaches, other algorithms that were used for comparison included transfer component analysis (TCA), conventional NMF, graph-regularized NMF (GNMF), and constrained NMF (CNMF). The proposed TGNMF and TCNMF methods always outperformed the other algorithms used for cross-corpus emotion recognition, with TCNMF performing best. TCNMF emotion recognition rates from the experiments ranged from

36.81% (anger) to 74.81% (disgust) on EMO-DB for a model trained using the eNTERFACE dataset.

Dai et al. (2015) proposed a support vector regression-based (SVR-based) method for emotion recognition on vocal social media in terms of position-arousal-dominance (PAD) emotion dimension estimation followed by categorical emotion mapping. There were 25 proposed emotional features gathered from speech for the experiments, which included prosody, spectral, and sound quality features. The authors then carried out model training using 180 chats from historical data on WeChat <sup>1</sup>, a Chinese vocal social media platform. Following this the authors used the same WeChat group to create a test set for their model. The emotion recognition accuracy achieved was 82.43% on average across happy, sad, angry, surprised, afraid, and neutral emotion classes. The reported PAD emotion dimension estimation error on this test set was 13.76%. The authors then tested the developed model on a cross-corpus basis where a test set was gathered from QQ <sup>2</sup>, another Chinese vocal social media platform. The aim of the cross-corpus testing was to assess model generalisability. The results from the cross-corpus test increased the model error by an absolute increase of 11.24% above the initial WeChat error. The authors claimed that the personal features of the groups on social media had a significant impact on the estimation of PAD values from the model.

Modulation spectral features (MSF) were presented and their performance compared against short-time spectral features, MFCC and perceptual linear prediction (PLP), for the task of emotion recognition by Wu et al. (2011). The authors also added the MSFs to prosodic features to obtain improvement in emotion recognition rates. In addition, the authors carried out a continuous affect prediction experiment where the system is reported to have achieved performance comparable to human annotators. The Berlin EMO-DB (Burkhardt et al., 2005) and VAM (Grimm et al., 2008) datasets were used for the experi-

---

<sup>1</sup><https://web.wechat.com/>

<sup>2</sup><https://im.qq.com/index.shtml>



ments. SVM and SVR machine learning schemes were used for classification and regression respectively. The results presented for the Berlin dataset indicate that MSF can outperform MFCC and PLP features for emotion classification tasks. Also, during a complimentary feature experiment, PLP, MFCC, and MSF were added to prosodic features for emotion recognition. The authors found that, on average, MSF outperformed PLP or MFCC when added to prosodic features over all emotion classes for emotion recognition. The authors also carried out a cross-database evaluation where a model was trained using the VAM dataset and then tested using the Berlin dataset. The proposed MSF feature set was used along with prosodic features. The recognition task for this experiment included continuous valence, arousal, and dominance level assessment and discrimination between two classes: anger versus joy, anger versus sadness, and anger versus fear. The results ranged from 58.6% to 100% with anger versus sadness being the best performing discrimination pair.

Fernandez and Picard (2011) evaluated a model developed to infer affective categories using a hierarchical graphical model in the form of Dynamic Bayesian Networks (DBN) within their cross-corpus experiments. The hierarchical graphical model was intended to depict changes in prosodic-acoustic parameters over time (dynamic) and time-scales (hierarchical). The authors used 105 prosodic features gathered from speech, which included duration, intonational, loudness, and voice-quality features for model input. The authors gathered an acted corpora of emotions (afraid, angry, happy, neutral, and sad) which was used for model validation. CallHome (Canavan et al., 1997), and BT call center (Durstun et al., 2001) corpora were employed for the model testing experiments. The model testing corpora contained natural, spontaneous emotion that was intended to be a more realistic assessment of the developed model. The emotion classes for model testing included high-arousal and negative-valence, high-arousal and positive-valence, low-arousal and negative-valence, and low-arousal and positive-valence. A neutral classification region was employed for some experiments. The best performance evaluation of the model achieved a 70% emotion classification rate.

Cross-corpus emotion recognition using speech was investigated by Schuller et al. (2010b). The motivation for this research was based on the view that an overestimation of machine learning performance was presented in the literature due to single, intra-corpus training and testing being employed. The corpora for the experiments included Danish Emotional Speech (DES) (Engbert and Hansen, 2007), EMO-DB (Burkhardt et al., 2005), Speech under Simulated and Actual Stress (SUSAS) (Hansen and Bou-Gazale, 1997), eNTERFACE (Martin et al., 2006), Audio-visual Interest Corpus (AVIC) (Schuller et al., 2007), and smartKom (Steininger et al., 2002). The authors used SVM for the classification of up to six emotion classes. They also discriminated between positive or negative valence and high or low arousal within their experiments. Speaker, corpus, and speaker-corpus normalization techniques were used for the experiments with speaker normalization performing best. The highest performing test set result was achieved on the acted emotion dataset EMO-DB where 2-class emotion classification achieved a median 70% UAR. In general the results indicated that UAR performance dropped with a higher number of emotion classes. The authors concluded that cross corpus emotion recognition extends well only to acted data in clearly defined scenarios, for example controlled room acoustics, noise, and microphone-to-subject distance. The authors stated that corpus construction needs to improve to address the aforementioned issues and increase the generalisability of the emotion recognition models.

The review of speech emotion recognition showed that there are numerous corpora available for model research, validation, and testing (Haq et al., 2008; Grimm et al., 2008; Canavan et al., 1997; Durston et al., 2001; Engbert and Hansen, 2007; Burkhardt et al., 2005; Hansen and Bou-Gazale, 1997; Martin et al., 2006; Schuller et al., 2007; Steininger et al., 2002). Speech features used for input to emotion recognition models include MFCC (Motamed et al., 2017; Wu et al., 2011), prosody (Dai et al., 2015; Wu et al., 2011; Fernandez and Picard, 2011), and voice-quality (Fernandez and Picard, 2011; Dai et al., 2015) features. The use of a pre-compiled, publically available speech feature set was investigated as part of the work undertaken by C.K. et al. (2017), but this is not

as common as manual feature extraction according to the reviewed literature. The novel approach taken taken by Fayek et al. (2017) includes end-to-end deep learning that removes the need for speech feature extraction, lowering the human effort required for emotional speech model building. Novel speech features from the literature worthy of further investigation include modulation spectral features (Wu et al., 2011), and bispectral and bicoherence features (C.K. et al., 2017). While SVM machine learning approaches have been widely used for speech emotion recognition (C.K. et al., 2017; Chakraborty et al., 2016; Wu et al., 2011; Schuller et al., 2010b), there is a trend toward neural network modelling of emotional speech (Fayek et al., 2017; Motamed et al., 2017). A gap observed in the literature was that no author incorporated an explicit time-shift for ground-truth annotations at a frame level prior to model training and testing. The human annotation process adds a delay to the ground-truth annotations provided with corpora, and studies within continuous affect prediction are incorporating these delays into ground-truth as shifts back in time (Valstar et al., 2016; He et al., 2015). This is an area that merits further investigation for speech emotion recognition.

Table 1: Summary of methods for emotion recognition in speech

Author(s)	Learning Scheme	Comment	Dataset(s)	Emotion Classes	Performance
(Fayek et al., 2017)	CNN	Frame-level recognition	IEMOCAP	Anger, happiness, sadness, neutral	60.89% UAR
(Motamed et al., 2017)	NN	ANFIS addition to NN	EMO-DB	Anger, happiness, sadness	72.5%
(C.K. et al., 2017)	SVM	BBO feature optimisation	SAVEE	Anxiety, disgust, happiness, boredom, neutral, sadness, anger	93.25%
			SUSAS	Anxiety, disgust, fear neutral, sadness, surprise, happiness	62.38%
				Angry, lombard, loud, neutral	90.09%
(Chakraborty et al., 2016)	SVM+ANN+KNN	Word emotion prominence added	EMO-DB IVR-SERES Call Center	Anger, happy, sad, neutral	87.3% 82.1% 78.1%
(Vlasenko et al., 2016)	HMM	Cross-corpus	EMO-DB	Low vs high arousal	86% UAR
				Anxiety, disgust, happiness, boredom, neutral, sadness, anger	52.10%
(Song et al., 2016)	NMF, transfer learning	Cross-corpus, cross-lingual	FAU Aibo eNTERFACE	Neutral, non-neutral Happiness, anger, disgust, fear, sadness, surprise	47.25% 44.86%
(Dai et al., 2015)	SVR+emotion mapping	Cross-corpus	QQ	HAPPY	25% error
(Wu et al., 2011)	SVM, SVR	Modulation spectral features	EMO-DB	Anger vs joy Anger vs sadness Anger vs fear	63.6% 100% 83.7%
(Fernandez and Picard, 2011)	DBN	Hierarchical graphical model	Call Home Call Center	low/high arousal and positive/negative valence high arousal-negative valence (binary 1/0)	44.1% 70% detection rate, 30% false alarm
(Schuller et al., 2010b)	SVM	Cross-corpus, various normalization	EMO-DB DES eNTERFACE SMARTKOM	2 emotion classes 3 emotion classes 4 emotion classes Arousal and valence levels	70% UAR 50% UAR 39% UAR 50% UAR

## 2.2. Continuous affect prediction using speech

Continuous affect prediction in speech is the task of predicting continuous numerical values for emotion dimensions. Numerical values for each emotion dimension typically range between +/- 1.0 for each speech frame. Emotion dimensions include arousal, which is a measure of how calming or exciting an experience is, and valence, which is a measure of how positive or negative an experience is (Kossaifi et al., 2017). Some common emotions are shown plotted on the arousal-valence plane in Figure 1. Only intra-corpus experiments for continuous affect prediction have been found in the literature. The review of continuous affect prediction using speech is important for this work, as emotion dimensions, corpora, feature sets, and performance evaluations used inform the speech and eye gaze continuous affect prediction experiments presented in Section 5.

Mencattini et al. (2017) used single-speaker-regression-models (SSRMs) with the ComParE speech feature set (Schuller et al., 2016) for arousal and valence prediction on the RECOLA corpus (Ringeval et al., 2013b). The regression techniques used for the experiment included partial least squares (PLS) and SVR. The authors incorporated weighted averaging of annotator ground-truth, annotator reaction lag, and feature selection for positive or negative arousal and positive or negative valence, which they called quadrant-based temporal division (QBTD). A concordance correlation coefficient (CCC) fusion measure was used to decide which SSRMs to include in the cooperative regression model (CRM) for the final test set predictions. The results showed that the SVR performed better than the PLS regression method for SSRM. However, the PLS method performed best for the CRM, which was the more general predictor. The authors suggest that SVR is more prone to over-fitting and simpler regression techniques such as PLS may be more suitable for machine learning techniques such as boosting, which is a similar ensemble method to the CRM method presented by the authors.

In Gupta and Narayanan (2016), the correlations between depression and affect were studied. Self-reported depression scores were incorporated into a

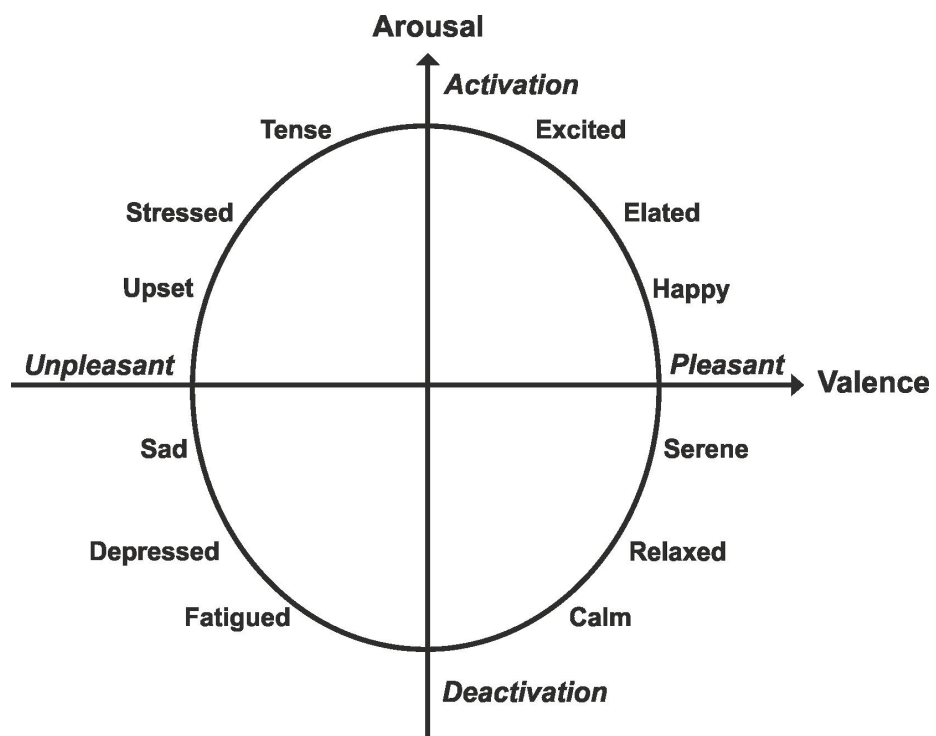


Figure 1: Arousal-Valence diagram from Abhang and Gawali, (2015)

continuous affect prediction system that used the AVEC 2014 (Valstar et al., 2014) database and speech features. Statistically significant correlations between affect dimensions and depression severity were found using a Student’s t-test at the 5% level. The authors based their subsequent investigation of an affect prediction system incorporating depression severity on these results. The final average results presented across arousal, valence, and dominance emotion dimensions were 0.33 correlation on the Freeform subset of AVEC 2014 and 0.52 on the Northwind subset. Including the depression severity score was found to improve the affect dimension correlations when compared with the authors’ baseline system, which did not include the depression scores.

A bag-of-audio-words (BoAW) approach to continuous affect prediction of arousal and valence on the RECOLA (Ringeval et al., 2013b) dataset was presented in Schmitt et al. (2016). The BoAW was created from MFCCs only and SVR was used for prediction from the codebook. The performance of BoAW was compared with that of functionals calculated from MFCCs and also against an end-to-end CNN approach to continuous affect prediction from the literature (Trigeorgis et al., 2016). The results showed the BoAW performed better than functionals on the test set. Additionally, the BoAW final arousal result in terms of CCC = 0.753, and valence = 0.465 was found to be better than a CNN approach from the literature (Trigeorgis et al., 2016).

Zhang et al. (2016) carried out arousal and valence prediction on the RECOLA (Ringeval et al., 2013b) corpus using SVR on speech after non-stationary additive and convolutional noise was added to speech. The experiment used the ChiME15 (Barker et al., 2015) database for additive noise and a convolution microphone input response (MIR) from a Google Nexus and room input response (RIR) for convolutional noises to be applied. The authors carried out feature enhancement on the noisy speech using LSTM-RNN and used this for comparison with the noisy speech input that they used as a baseline. SVR was used for affect prediction, temporal window sizes of 8 seconds, a temporal window step rate of 0.04 seconds, and a ground-truth delay of 4 seconds was applied to the speech input. The experiments included additive noise and smartphone noise

continuous affect prediction performance. The smartphone noise consisted of MIR, RIR, or varying levels of ChiME noise. In addition two feature enhancement methods were employed using LSTM-RNN: (1) matched, in which several feature enhancement models were trained with different noise conditions, and (2) mixed, in which one feature enhancement model was trained with different noise conditions. The results showed that matched feature enhancement performed better than mixed feature enhancement. The CCC values for the matched condition were 0.596 for arousal and 0.223 for valence when ChiME15 noise was included, and 0.684 for arousal and 0.162 for valence when smartphone noise was included. The feature enhancement methods always improved affect recognition when compared with the baseline results.

Asgari et al. (2014) used speech, prosody, and text features as inputs to SVR systems for continuous affect prediction experiments on their own recorded dataset of children posing as actors in a short story re-enactment. Speech performed best for arousal prediction with a product moment correlation score of 0.86 and text performed best for valence with a correlation of 0.57. The authors concluded that the arousal of speech can be measured reliably, but not valence. It was also suggested that better text features need to be developed for valence dimension prediction.

The review of continuous affect prediction clearly shows arousal and valence as the emotion dimensions of choice for two-dimensional affect prediction (Mencattini et al., 2017; Ringeval et al., 2013b; Asgari et al., 2014; Gupta and Narayanan, 2016; Schmitt et al., 2016; Zhang et al., 2016; Valstar et al., 2014). However, the AVEC 2014 corpus (Valstar et al., 2014) and the experiment presented by (Gupta and Narayanan, 2016) included dominance as a third dimension, along with arousal and valence. The affective corpora for the experiments included AVEC 2014 (Valstar et al., 2014) and RECOLA (Ringeval et al., 2013b). Pre-compiled speech feature sets for affect prediction from the literature included ComParE (Schuller et al., 2016) and the AVEC 2014 speech feature set (Valstar et al., 2014). The majority of works reviewed carried out performance evaluation using CCC as a metric (Mencattini et al., 2017; Schmitt et al., 2016;



Zhang et al., 2016) and this is justification for the use of CCC in this paper. No cross-corpus continuous affect prediction experiments could be found by the authors of this paper during the review. As such, continuous affect prediction needs cross-corpus experiments such as those undertaken in the speech emotion recognition community (Schuller et al., 2010b; Fernandez and Picard, 2011; Dai et al., 2015; Song et al., 2016; Vlasenko et al., 2016) to advance the field.

Table 2: Summary of continuous affect prediction methods for arousal (Ar), valence (Val), and dominance (Dom) prediction in speech

Author(s)	Learning Scheme	Comment	Dataset(s)	Performance
(Mencattini et al., 2017)	PLS	Cooperative regression model, QBTD	RECOLA	CCC = 0.7 (Ar), 0.2 (Val)
(Gupta and Narayanan, 2016)	SVR	Depression score inclusion	AVEC 2014 (Northwind)	COR = 0.52 (Avg. of Ar, Val, Dom)
(Schmitt et al., 2016)	SVR	BoAW	RECOLA	CCC = 0.753 (Ar), 0.465 (Val)
(Zhang et al., 2016)	SVR	Noise addition, feature enhancement	RECOLA	CCC = 0.684 (Ar), 0.223 (Val)
(Asgari et al., 2014)	SVR	Speech, prosody, text, ASR text	(Asgari et al., 2014)	COR = 0.86 (Ar), 0.57 (Val)

### 3. Multi-modal affective computing using speech

This section presents a review of multi-modal emotion recognition and continuous affect prediction using speech. Multi-modal affective computing is the study of more than one mode, or form of communication, during affective computing tasks. Of interest to this review are multi-modal systems that include speech. Again, the focus of this review within affective computing is emotion recognition, or recognizing an emotion displayed, and continuous affect prediction, which is the prediction of numerical values for emotion dimensions. The multi-modal systems reviewed include video, context, and physiological features. The physiological features in the literature include EEG, eye gaze, EDA, heart rate, ECG, body movement, posture, or facial gesture features. Summaries of each section of the review are given in Tables 3 and 4.

#### 3.1. Multi-modal emotion recognition using speech

Emotion recognition involves the computational recognition of emotion. The literature review of multi-modal emotion recognition using speech is intended to

highlight the features commonly used with speech in multi-modal systems, along with machine learning methods used and emotion recognition performances achieved.

Nguyen et al. (2017) carried out audio-visual emotion recognition using 3-dimensional CNNs (C3Ns) and deep-belief networks (DBNs) on the eNTERFACE corpus (Martin et al., 2006). Spatio-temporal features were gathered using C3Ns from both audio and visual modalities and these features were passed further as input to audio and video DBNs. The audio and video DBNs were later fused using score-level fusion to produce the final emotion class decision. Input feature vectors for the C3Ns included a log-power spectra feature matrix with a  $257 \times 72$  dimension for the audio C3N and a face region image extracted using a modified Viola-Jones algorithm from video frames for the video C3N. The modified Viola-Jones algorithm used by the authors employed a face cropping region the same as the previous frame if the face could not be detected. An emotion recognition rate of 82.83% was achieved on average across anger, disgust, fear, happiness, sadness, surprise emotion classes.

Poria et al. (2016) investigated multi-modal sentiment and emotion recognition from audio-visual and text sequences using temporal deep CNN. The IEMOCAP corpus (Busso et al., 2008) was used for the unimodal and multi-modal emotion recognition experiments presented. The temporal deep CNN proposed combined  $t$  and  $t + 1$  images, where  $t$  is an image index in time, into one image feature vector and additionally used an RNN to model spatial and temporal dependencies for the CNN. A total of 6,373 speech features were extracted using openSMILE (Eyben et al., 2013) from audio for the experiments, and a positive or negative sentiment feature was extracted from text using a neural network approach. For multi-modal fusion, the authors used multiple kernel learning (MKL), proposed by Subrahmanya and Shin (2010). The authors describe multiple kernel learning as a fusion method whereby features are placed in groups, and then each group has its own kernel for model learning. From the experiments, the video emotion classifier performed best when compared against audio and text unimodal emotion classifiers. The multi-modal

audio, text, and video classifier performed best overall when compared against the unimodal emotion recognition systems, however, the authors note that the multi-modal system performed only slightly better than the video classifier. The emotion recognition accuracies achieved for the multi-modal system were 79.2% (angry), 72.22% (happy), 75.63% (sad), and 80.63% (neutral).

Yang and Narayanan (2016) performed utterance-level multi-modal emotion recognition on the USC CreativeIT database (Metallinou et al., 2016) using a semi-supervised learning approach. Pitch, energy, and MFCC features were gathered from audio, while hand gesture features were gathered from video for emotion classification. The audio-visual feature vector was combined using a proposed canonical correlation analysis (CCA) where both audio and video feature vectors were transformed using CCA projection vectors from each input modality prior to feature concatenation of each modality data. The authors used a multi-modal codebook, inspired by bag of words (BoW) approaches from text analysis, which used k-means clustering to create a codebook of multi-modal words for emotion recognition. The classification experiments included 2- and 3-class arousal and valence level discrimination. The arousal and valence classification results were 70.9% for 2-class arousal and valence classification, and 53% for 3-class arousal and valence classification using a code book of size 35.

Xie et al. (2015) proposed a fusion method for multi-modal emotion recognition from audio-visual data. The experiments were carried out using prosody and MFCC features from audio, and Gabor filter and elastic body spline (EBS) facial features from video. The fusion method proposed incorporated both early and late fusion. For early fusion, separate audio and video modality feature transformation and fusion based on kernel entropy component analysis (KECA) (Jenssen, 2010) was carried out. Following the early feature fusion, separate audio and video HMMs were used to classify the emotion. The individual HMM classifications were fused by way of score-level fusion to produce the final emotion decision. The score-level fusion proposed by Xie et al. (2015) is based on maximum correntropy criterion (MCC) optimisation, where correntropy is

presented as “a generalized similarity measure which has the stability to variation or noise”. The MCC optimisation problem was solved using the algorithm presented by He et al. (2011). The emotion corpora used for the experiments included eNTERFACE (Martin et al., 2006) and RTL (Wang and Guan, 2008). Both of the corpora include anger, disgust, fear, sadness, surprise and happiness emotion ground-truth annotations. The results achieved over both corpora were 83% average recognition accuracy over all emotions on the two corpora. Additional experiments showed that the bimodal system outperformed unimodal audio or video systems and that among the video features EBS features outperformed Gabor filter features.

Soleymani et al. (2012) created an audio-visual affect database (MAHNOB-HCI) and investigated emotion recognition using speech, eye gaze, EEG, and physiological signals in their experiments. A Tobii X120<sup>3</sup> eye tracking device was used for eye gaze data capture from subjects. The authors divided arousal into classes of medium aroused, calm, and excited for the emotion recognition experiment and used SVM for classification. Valence was divided into classes of unpleasant, neutral valence, and pleasant. For multi-modal fusion, confidence measure summation fusion was used. The emotion recognition results showed that eye gaze performed best during unimodal affect recognition experiments and a combination of eye gaze and EEG proved best overall. It must be stated however, that the speech modality may not have been maximally utilised as the subjects were required to watch emotion provoking video only. The unimodal eye gaze emotion classification accuracies achieved were 63.5% for arousal and 68.8% for valence. The EEG and eye gaze results were 67.7% for arousal classification and 76.1% for valence classification. While speech may not have been maximally used by Soleymani et al. (2012), it is of interest for the work presented by this paper that eye gaze performed best during the unimodal affect recognition experiments performed by Soleymani et al. (2012).

From the review of multi-modal emotion recognition using speech, it is clear

---

<sup>3</sup><https://www.tobii.com/>

that video is a popular modality for fusion with speech (Nguyen et al., 2017; Poria et al., 2016; Yang and Narayanan, 2016; Xie et al., 2015; Soleymani et al., 2012). In addition, it is encouraging to see temporal consideration taken into account in the recent CNN approaches taken in the literature, by way of spatio-temporal feature gathering (Nguyen et al., 2017), and temporal CNN and RNN combination (Poria et al., 2016).

The CCA fusion proposed by Yang and Narayanan (2016) is interesting for bimodal affective computing research due to its simplicity. This fusion approach was shown to exploit the correlations between input modalities and further investigation of this technique is required using different machine learning techniques. A direct comparison of this work, and the fusion work undertaken by Xie et al. (2015), against other approaches taken in the literature would be beneficial to advance multi-modal fusion research for affective computing.

Although not directly comparable, the average emotion recognition accuracy achieved by Xie et al. (2015) of 83% is higher than that of Nguyen et al. (2017) (82.83%). While Xie et al. (2015) used more complex fusion methods, for example, both feature fusion by KECA and score-level fusion by way of MMC optimisation, compared with Nguyen et al. (2017) who simply used score-level fusion, Nguyen et al. (2017) does use a more complex machine learning configuration of two C3Ns and two DBNs compared with two HMMs used by Xie et al. (2015). However, the result achieved by Xie et al. (2015) is reported on both eINTERFACE (Martin et al., 2006) and RTL (Wang and Guan, 2008) corpora, compared with only eINTERFACE for Nguyen et al. (2017). Reporting of the result achieved by Xie et al. (2015) on the eINTERFACE corpus only would yield a directly comparable result to that of Nguyen et al. (2017). This could offer interesting discussion on the complexity of fusion and learning methods, and their respective performance for multi-modal emotion recognition.

Table 3: Summary of multi-modal emotion recognition methods

Author(s)	Learning Scheme	Comment	Dataset(s)	Emotion Classes	Performance
(Nguyen et al., 2017)	DBN	C3N spatio-temporal features used	eNTERFACE	Anger, disgust, fear, happiness, sadness, surprise	82.83%
(Poria et al., 2016)	Temporal CNN/RNN	MKL fusion	IEMOCAP	Angry Neutral	79.2% 80.63%
(Yang and Narayanan, 2016)	BoW (k-means)	CCA fusion	USC CreativeIT	Arousal and valence levels	70.9%
(Xie et al., 2015)	HMM	KECA early fusion, MCC late fusion	eNTERFACE, RTL	Anger, disgust, fear, happiness, sadness, surprise	83%
(Soleymani et al., 2012)	SVM	Cross-corpus	MAHNOB-HCI	Arousal level Valence level	67.7% 76.1%

### *3.2. Multi-modal continuous affect prediction using speech*

Continuous affect prediction is the task of predicting continuous numerical values for emotion dimensions. There is an increasing interest by the affective computing community in multi-modal, continuous affect prediction as is evident by research challenges such as AVEC 2014 (Valstar et al., 2014), AV+EC 2015 (Ringeval et al., 2015b), AVEC 2016 (Valstar et al., 2016). Additionally, there is now a large number of affective corpora for multi-modal research including AFEW-VA (Kossaifi et al., 2017), AVEC 2014 (Valstar et al., 2014), RECOLA (Ringeval et al., 2013b), and SEMAINE (McKeown et al., 2012). Multi-modal continuous affective computing research using speech includes additional modalities in the overall system. Text, context, EEG, electro-cardiogram (ECG), electro-dermal activity (EDA), eye gaze, facial expression, gesture, posture, and video have all been used with speech for affect prediction. The variety provided for input data, advances in computer vision tools such as OpenFace (Baltruaitis et al., 2016), and the fact that this is such an active area of research greatly motivates the work presented here in this paper.

Stratou and Morency (2017) presented a framework designed to support multi-modal affective computing. The framework integrates open-source software, licensed software, and hardware support (camera, microphone, depth sensor) to achieve a unified framework for research and real-time applications for affective computing. Multi-modal input for the framework can include eye gaze, head pose, skeleton, speech, prosody, dialogue, and context. The framework incorporates context-based assessment at a local level by event creation and logging, such as when a question is asked, and at a global level by scenario definitions, for example, a job interview. The authors present a use case demonstrating the benefit of context-based features for automatic psychological distress analysis within a healthcare interview application. Subjects, which included US army veterans and general population, interacted with a virtual agent, SimSensei, and a behavioural report was produced after the user had finished their conversation with the virtual agent. A distress level prediction correlation of 0.7448 was achieved for a system that did not use context-based

features. A distress level prediction correlation of 0.8822 was achieved for an automatic system that included context-based features, demonstrating the benefit of features calculated using context information.

Brady et al. (2016) were continuous affect prediction challenge winners for their work in the AVEC 2016 challenge (Valstar et al., 2016). The authors used prosody, MFCC, and shifted delta cepstrum features from audio, in addition to the baseline audio features provided with the challenge, as input to SVR for arousal and valence prediction from audio. From the video modality, convolutional neural network (CNN) features were used as input to a recurrent neural network for arousal and valence regression. For prediction from the physiological channel, the authors opted to use the baseline features provided for ECG, skin conductance rate (SCR), and skin conductance level (SCL). LSTM regression features were trained from the baseline HRHRV and EDA features. The authors used Kalman filter fusion of audio, video, and physiological modalities for their multi-modal submission on the test set. The multi-modal fusion included model approximation, for example arousal or valence prediction, in addition to sensor channel measurements in the measurement matrix. The multi-modal results achieved on the test set were CCC values of 0.770 for arousal and 0.687 for valence.

Work presented in He et al. (2015) illustrates the deep BLSTM-RNN approach that these authors have taken in winning the AV+EC 2015 challenge (Ringeval et al., 2015b). The challenge was performed on the RECOLA dataset. Baseline features included with the dataset included speech, appearance and geometric video features, and physiological measures, which were comprised of ECG and EDA measurements. The authors added functionals of low-level descriptors extracted from speech using the YAFFE toolbox (Mathieu et al., 2010) to the speech baseline features and local phase quantization from three orthogonal planes (LPQ-TOP) from video to the video baseline features. The machine learning scheme used for prediction for each modality was BLSTM-RNN. Modality predictions were then fused by first applying Gaussian smoothing after which the prediction was passed to a final BLSTM-RNN for the final predictions. The



CCC values achieved on the RECOLA test set were 0.747 for arousal and 0.609 for valence prediction.

A recent continuous affect prediction experiment undertaken by Ringeval et al. (2015a) aimed to predict arousal and valence on the RECOLA (Ringeval et al., 2013b) corpus using neural networks. The modalities employed for this work included speech, video, ECG, and EDA. The machine learning schemes that were used included feed-forward neural networks, LSTM-RNN, and BLSTM-RNN. The CURRENNT tool kit (Weninger, 2015) was used for network creation and training. BLSTM-RNN was the best performing network from the experiments. The authors experimented with various temporal window sizes for feature calculation, their findings suggesting that valence required a longer time window for optimum system performance. The CCC performance achieved was 0.804 for arousal and 0.528 for valence prediction.

Work presented by Kächele et al. (2014) demonstrated their approach to the AVEC 2014 continuous affect prediction and depression classification challenges (Valstar et al., 2014). For the continuous affect prediction challenge, the authors investigated task dependent pattern templates for the prediction of each dimension. In addition to the challenge-provided features, the authors included application dependent meta knowledge features that included subject ID, gender, subject movement, subject age, estimated subject socio-economic status, and pixel statistics, among others. SVR and eigenvalue decomposition (EVD) with and without subject clustering approaches were used for the dimensional prediction from the audio-visual data. The subject clustering into 3 groups was carried out using a Wards distance measure based on the challenges depressive state features 1 - 29. The best Pearson correlations achieved were 0.6330 for arousal, and 0.5697 for dominance using an EVD and SVR approach with subject clustering. The best valence prediction result, a 0.5869 Pearson correlation was achieved using SVR with subject clustering.

Nicolaou et al. (2011) carried out continuous arousal and valence prediction experiments on the SEMAINE (McKeown et al., 2012) corpus using speech, facial features, and shoulder gesture modality inputs. The SEMAINE (McKe-

own et al., 2012) corpus is a spontaneous emotion dataset in which audio-visual recordings of a subject interacting with a particular emotional character (happy, sad) have been recorded. The experiments included a comparison of machine learning algorithms and multi-modal fusion techniques. The authors also presented an output-associative fusion framework designed to exploit correlations and covariances between arousal and valence for model development. Experimental results showed BLSTM-RNN outperforming SVR during unimodal experiments. Additionally, speech was the best performing modality for arousal but the worst performing modality for valence within the unimodal investigation. For the final fusion experiments, output-associative fusion outperformed model and feature-level fusion of features. The highest reported correlation scores of 0.796 for arousal and 0.643 for valence were achieved using features from all modalities combined as input to a BLSTM-RNN utilising the proposed output-associative fusion framework.

Multi-modal continuous affect prediction using speech is now both a well developed field, and an active area of research. Affect prediction performances achieved by Nicolaou et al. (2011), Ringeval et al. (2015a), He et al. (2015), and Brady et al. (2016) have achieved good results, given the challenging, non-acted emotion corpora used (Schuller et al., 2010b). BLSTM-RNN is widely used (Ringeval et al., 2015a; He et al., 2015; Nicolaou et al., 2011) for continuous affect prediction. In particular, Nicolaou et al. (2011) showed BLSTM-RNN outperforming SVR in his experiments and his findings are considered in the approach taken in the work presented in this paper. The work undertaken by Kächele et al. (2014) and Stratou and Morency (2017) showed the importance of including context information during affect prediction. The result achieved by Kächele et al. (2014) is particularly impressive, as the Northwind subset of AVEC 2014 (Valstar et al., 2014) does not explicitly elicit an emotional response; this is in contrast to other common emotion corpora such as SEMAINE (McKeown et al., 2012), RECOLA (Ringeval et al., 2013b), and the AVEC 2014 (Valstar et al., 2014) Freeform subset, which have explicit emotion elicitation protocols. Recent work on affective computing software tools by Stratou and

Morency (2017) shows that eye gaze is an emerging modality for affective computing. Unfortunately, from the review undertaken, there were no cross-corpus or cross-lingual continuous affect prediction experiments available. This gap in the literature presents a new research opportunity for continuous affect prediction.

Table 4: Summary of multi-modal continuous affect prediction methods for arousal (Ar), valence (Val), and dominance (Dom) prediction

Author(s)	Learning Scheme	Comment	Dataset(s)	Performance
(Stratou and Morency, 2017)	Linear regression	Distress prediction, context features	Stratou and Morency (2017)	COR = 0.8822
(Brady et al., 2016)	SVR, RNN	SDC features, Kalman filter fusion	RECOLA	CCC = 0.770 (Ar), 0.687 (Val)
(He et al., 2015)	BLSTM-RNN	Gaussian smoothed model fusion	RECOLA	CCC = 0.747 (Ar), 0.609 (Val)
(Ringeval et al., 2015a)	BLSTM-RNN	Fusion, varying temporal windows	RECOLA	CCC = 0.804 (Ar), 0.528 (Val)
(Kächele et al., 2014)	SVR	Application meta-knowledge	AVEC 2014	COR = 0.5965 (Avg. Ar, Val, Dom)
(Nicolaou et al., 2011)	BLSTM-RNN	Output-associative fusion	RECOLA	CCC = 0.796 (Ar), 0.643 (Val)

#### 4. Affective computing using eye gaze

From the review of the literature on emotion recognition and continuous affect prediction using speech, a gap has been observed concerning the use of speech combined with eye gaze in multi-modal affective computing systems. A further review is presented here for affective computing with eye gaze as an input modality. This section concludes with suggestions for combining speech and eye gaze as input modalities for affective computing. The eyes are a rich source of sociological (Itier and Batty, 2009), neuropsychological (Lappi, 2016), and arousal-level (Partala and Surakka, 2003) information. Eye-based features applied to affective computing include: eye blink (Soleymani et al., 2012; Lanatà et al., 2011), eye gaze (O’Dwyer et al., 2017a; Soleymani et al., 2012; Lanatà et al., 2011), visual focus of attention (Zhao et al., 2011), pupillometry (Soleymani et al., 2012), and pupil size variation (Lanatà et al., 2011). Other measurements, including eye saccade (saccadometry), are also possible to gather from the eyes for affective computing research.

#### 4.1. Emotion recognition using eye gaze

Unimodal emotion recognition using eye gaze was investigated by Aracena et al. (2015) using an EyeLink 1000<sup>4</sup> eye tracking headset. Eye gaze data were gathered from the eye tracker with emotion classification carried out using neural networks. The best performing method from this experiment included a decision tree with neural networks. The decision tree neural network correctly recognized emotion classes as either positive, negative or neutral with an average accuracy of 53.6% for 4 male subjects on a subject-independent basis. Subject-dependent scores for the decision tree neural network ranged from 62.3% to 78.4% classification accuracy.

Zhao et al. (2011) cited psychological research suggesting a correlation of direct eye gaze with angry and happy emotions in human-to-human communication. The authors also claimed that sadness and fear are associated with averted gaze. A geometrical eye and nostril model was used to identify averted gaze and direct gaze in video input. This eye gaze data was added to facial expression to improve emotion classification in angry, sad, fear, and disgust recognition. Happy and surprise recognition was not improved when eye gaze was added to facial features for the classification. The authors mention that facial illumination of subjects was controlled in the experiment, which limits the applicability of the results to more natural environments.

Lanata et al. (2011) assessed whether eye gaze tracking and pupil size variation could provide useful cues to discriminate between emotional states induced in subjects viewing images at different arousal content levels. The emotional states were defined as neutral and high arousal. Subjects were provided images from the international affective picture system (IAPS) (J Lang et al., 2008) which was intended to evoke these responses. A new wearable and wireless eye gaze tracker called HATCAM was proposed. Recurrence quantification analysis (RQA) (Webber Jr and Zbilut, 2005), along with fixation time and pupil area detection features were used for classification using k-nearest-neighbours

---

<sup>4</sup><http://www.sr-research.com/eyelinkII.html>

(KNN). The experimental results showed that the proposed hardware and features could be used to discriminate between users experiencing different emotion stimuli. Emotion recognition rates of 90% for neutral images and approximately 80% for high arousal images were achieved.

#### *4.2. Continuous affect prediction using eye gaze*

O’Dwyer et al. (2017a) investigated unimodal continuous affect prediction in terms of arousal and valence using eye gaze. The Freeform subset of the AVEC 2014 dataset (Valstar et al., 2014) was used for the experiments along with the speech features provided with AVEC 2014 for baseline comparison. An SVR learning scheme was employed for both speech and eye gaze experiments and a new eye gaze feature set for continuous affect prediction was proposed. The results showed that eye gaze could perform well for valence prediction when compared to speech, but speech was the far better predictor of arousal. The Pearson’s correlation scores from the experiment were: eye gaze valence 0.3318, speech valence 0.3107, eye gaze arousal 0.3322, speech arousal 0.5225.

#### *4.3. Affective computing using speech and eye gaze*

From the review of the literature, it appears that continuous affect prediction using the combination of speech and eye gaze data from audio-visual sources is still largely unexplored. The inclusion of speech in multi-modal continuous affective prediction systems is common, and good results have been achieved in the literature (Ringeval et al., 2015a; Brady et al., 2016). Combining speech and eye gaze, or more generally, including eye gaze in affective computing systems is receiving greater interest from the research community (Stratou and Morency, 2017; O’Dwyer et al., 2017a; Aracena et al., 2015; Zhao et al., 2011; Lanatà et al., 2011). Recent projects such as OpenFace (Baltruaitis et al., 2016) and MultiSense (Stratou and Morency, 2017) now make eye gaze features from video easily accessible to multi-modal affect systems research. However, the literature has focused on emotion recognition using eye gaze, only one paper could be identified as investigating eye gaze for continuous affect prediction (O’Dwyer

et al., 2017a). Furthermore, the combination of speech and eye gaze in a bimodal continuous affect prediction system has not been explored.

The next section details a proposal for a novel speech and eye gaze continuous affect prediction system based on LSTM-RNNs. The proposed system contains optimised ground-truth annotation delays to account for human perception of audio-visual sequences. The system training and performance evaluation data are also presented and discussed.

## 5. A bimodal speech and eye gaze affect prediction system

In this section, a bimodal speech and eye gaze continuous affect prediction system based on BLSTM-RNN and LSTM-RNN is proposed. Training and performance evaluation data are discussed.

### 5.1. System framework

The BLSTM and LSTM ((B)LSTM) neural networks presented are trained on a single-task basis. That is, any network is only trained to predict arousal or valence. The fusion framework used for the system is feature fusion, also known as early feature fusion, which is the concatenation of speech and eye gaze features into one large feature vector for each training or testing example. Additionally, the system framework is designed to achieve the best possible advantage from the ground-truth annotations for arousal and valence provided with each corpus, which the annotators provided in response to observed audio-visual recordings of subjects. Therefore, the system employs a ground-truth time-shift, back in time, prior to neural network training or testing. The optimal ground-truth time-shift is selected for the bimodal system, for both arousal and valence affect prediction tasks. A block diagram of the system framework is given in Figure 2.

### 5.2. Experimental approach

Following from Ringeval et al. (2015a), single-task BLSTM networks are trained using the CURRENNT toolkit (Weninger, 2015) for arousal and valence

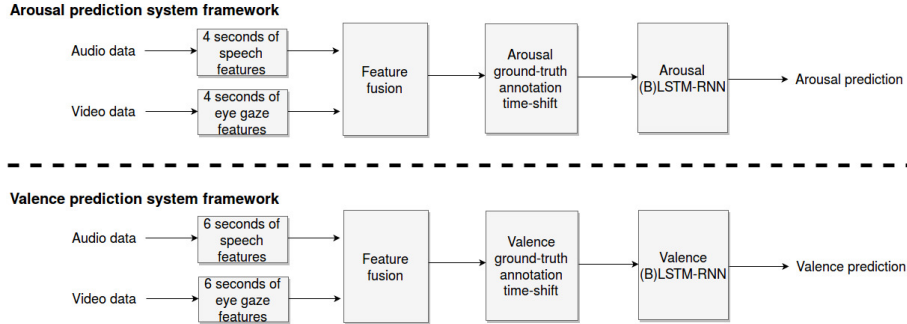


Figure 2: Proposed bimodal speech and eye gaze affect prediction system

prediction. There are two hidden layers, the first with 40 nodes and the second with 30 nodes. Additionally, unidirectional LSTM-RNN models consisting of two hidden layers with 80 and 60 nodes respectively are generated for performance comparison with BLSTM and system selection. Bimodal speech and eye gaze (B)LSTM models are first created using a number of different ground-truth time-shifts in order to select the optimal shift to be applied to arousal and valence annotations for training the networks. The ground-truth time-shift experiment is followed by unimodal (B)LSTM speech and eye gaze model creation to evaluate the contribution of eye gaze to speech in the proposed affect prediction system. Finally, cross-corpus and cross-lingual continuous affect prediction is carried out using the best performing models from the uni- and bimodal intra-corpus experiments. BLSTM-RNN has the advantage of both past and future temporal context during network training. The unidirectional LSTM recurrent neural network variant only has the benefit of past context, or memory, as it is trained. The LSTM addition to RNN was first presented by Hochreiter and Schmidhuber (1997) to avoid the vanishing gradient problem that RNNs suffer from. These neural networks contain what Hochreiter and Schmidhuber (1997) presented as memory units that allow nodes to store context information during network training. Today, LSTM-RNN is widely used for affective computing model generation, with good performance being achieved in recent works such as (Ringeval et al., 2015a), (He et al., 2015), (Valstar et al., 2016), and (Brady

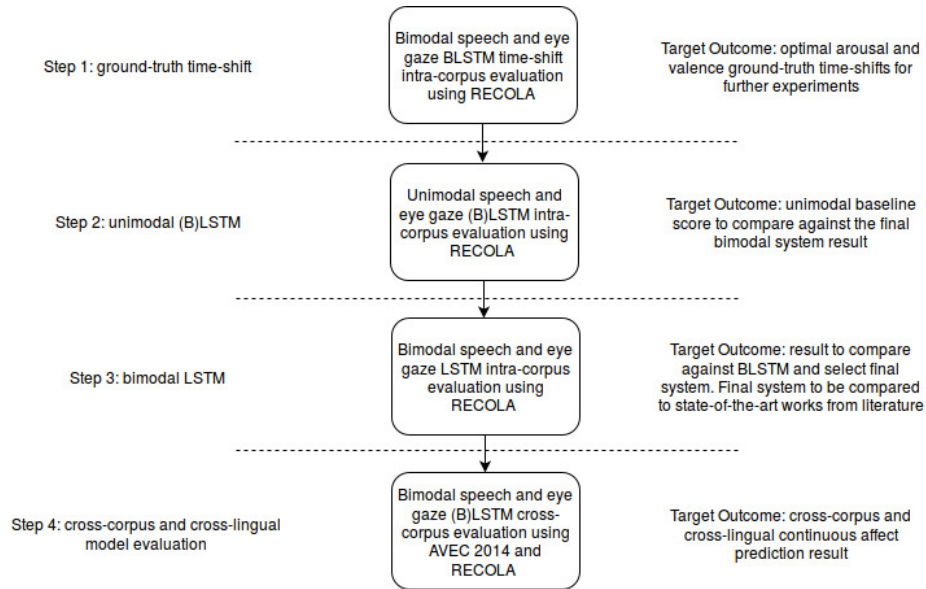


Figure 3: Summary of experimental steps and target outcomes in the design of the proposed speech and eye gaze continuous affect prediction system

et al., 2016). The experimental steps and their target outcomes are summarised in Figure 3.

### 5.3. Audio-visual corpora

The RECOLA corpus contains audio, visual, and physiological recordings of spontaneous affectively coloured dyad interactions in French. The subjects within each dyad were required to perform a task together requiring collaboration while being recorded. Human-provided numerical ground-truth annotations for arousal and valence are provided for each frame of each recording within the corpus for automatic prediction system training and performance evaluation.

The AVEC 2014 corpus (Valstar et al., 2014) is an audio-visual corpus recorded in German with ground-truth values provided for arousal and valence for each frame of each recording in the corpus. In particular, the Freeform partition of this corpus is used for the cross-corpus experiment. The Freeform partition of this dataset has subjects interacting with a computer to answer emo-



tionally provocative questions while audio-visual recordings were taken. The tasks, communication, emotion elicitation, and emotional responses for both AVEC and RECOLA corpora can be compared in Table 5.

Table 5: RECOLA and AVEC Freeform emotion corpora protocols

Corpus	Task	Communication	Emotion elicitation	Emotional response
AVEC 2014 (Valstar et al., 2014)	Responding to emotionally provocative questions	Human-computer	Explicitly provoked	Elicited
RECOLA (Ringeval et al., 2013b)	Collaborative dyadic problem solving	Human-human*	Interaction and task response	Spontaneous

\*Used remote computer interaction

#### 5.4. Affective feature extraction

Affective feature extraction is the process of calculating features from raw data provided from the input modalities of choice for a system. Features can include low-level descriptors, for example, energy or spectral parameters for speech (Eyben et al., 2016), and eye gaze distance (O’Dwyer et al., 2017a) for eye gaze. Statistics of the low-level descriptors are often calculated as part of the final feature set to be passed as input to machine learning models for classification or regression. The eGeMAPS (Eyben et al., 2016) speech feature set was used for affective speech feature input in this work. Speech features are extracted using openSMILE (Eyben et al., 2013) with a calculation window of 4 seconds worth of frames for arousal dimension feature calculation and 6 seconds for valence dimension feature calculation. The feature calculation window sizes are the same as those used for the audio information channel in (Valstar et al., 2016). The feature calculation windows were moved forward at a rate of 1 frame as was the case for (Valstar et al., 2016). The eGeMAPS (Eyben et al., 2016) feature set was used to provide baseline features in the AVEC 2016 Challenge (Valstar et al., 2016) and provides a total of 88 affective features from speech. For eye gaze affective features, the feature set presented by O’Dwyer et al. (2017a), containing a total of 31 features, is extracted. The affective eye gaze features from this set are listed in Table 6. Raw eye gaze data are gathered from video using OpenFace (Baltruaitis et al., 2016) and this is followed by

feature extraction from the raw data using the same time windows and window rates as for the speech features. In summary, for the bimodal affect prediction experiments a total of 119 features are extracted for model training, validation, and testing.

Table 6: Eye gaze feature set for continuous affect prediction from O’Dwyer et al. (2017a)

Data	Features
Eye gaze distance	eye gaze approach ratio, average eye gaze approach time in milliseconds
Eye gaze scan paths	average scan path length, standard deviation of scan path lengths
Vertical and horizontal eye gaze coordinates	average, inter quartile range 1-2, inter quartile range 2-3, standard deviation, skewness, power spectral densities at frequencies [0.011, 0.022, 0.033-0.044, 0.055-0.066, 0.077-0.133]Hz, average of standard deviation of coordinates in each fixation zone, standard deviation of standard deviation of coordinates in each fixation zone
Eye closure	average eye close frame count, standard deviation of eye close frame count, skewness of eye close frame count

### 5.5. Feature fusion

Feature fusion is the process of combining features gathered from different modalities with the intention of improving the performance of machine learning systems. Feature fusion can be performed early, late, or both, with regard to the machine learning process. Early feature fusion, in the form of feature fusion, is simply the row-wise concatenation of features from each modality into one large feature vector prior to model training and testing. An illustration of feature fusion is given in Figure 4. An example of late feature fusion is decision fusion. Decision fusion, shown in Figure 5, is where multiple models are trained, one for each modality, after which the model decisions that are made on a portion of the training data, or development set, are combined to make a further model for the final test set classification or regression. The simplest fusion approach taken in (Ringeval et al., 2015a), feature fusion, is employed for the work presented here. The feature fusion method was selected based on experimentation carried

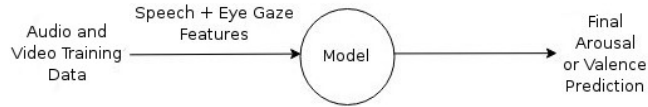


Figure 4: Speech and eye gaze feature fusion

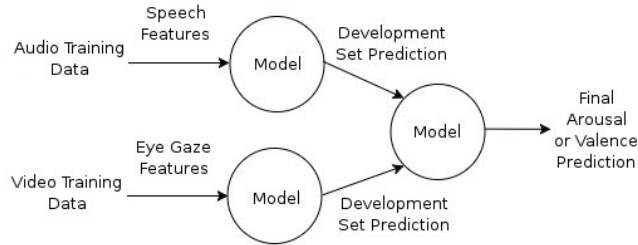


Figure 5: Speech and eye gaze decision fusion

out by O’Dwyer et al. (2017b), where feature fusion was shown to perform best for the continuous prediction of arousal.

### 5.6. Neural network training for affect prediction

The (B)LSTM-RNN neural network training methodology follows that of Ringeval et al. (2015a). As such, inputs and ground-truth targets are normalised to zero mean and unit variance prior to neural network training. In addition, Gaussian noise with a standard deviation 0.1 is added to all input features prior to training. The neural network training takes place for a maximum of 100 epochs, and training stops if no improvement of the performance as measured by sum of squared errors is observed on the validation set for more than 20 epochs. The network learning rate and random seed hyperparameters are selected based on the lowest validation set error achieved during experimentation.

One important difference between the proposed system training and that of Ringeval et al. (2015a) is the time-shift of ground-truth annotations prior to network training. To take into account the delay in human reaction times when producing ground-truth annotations for the corpora, a shift back in time is applied to the ground-truth prior to concatenation with the speech and eye gaze

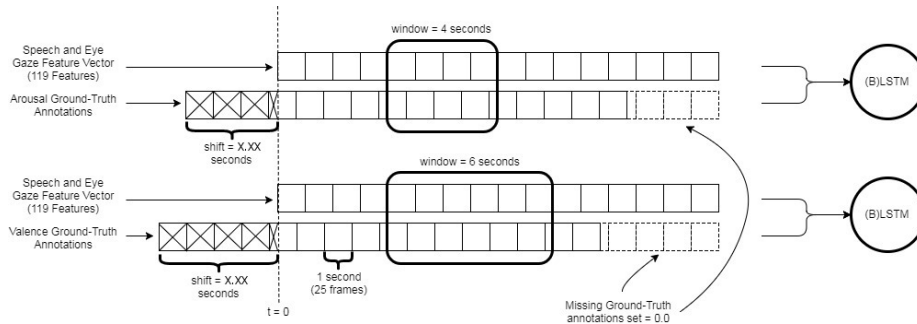


Figure 6: Arousal and valence ground-truth annotation time delay

features for network training and testing. Ringeval et al. (2015a) argued that LSTM-RNN can encode temporal context during network training, which allows this machine learning method to overcome the need for ground-truth time-shift to account for human annotation lag. However, based on results obtained by He et al. (2015), incorporating ground-truth delays prior to (B)LSTM network training, is used in the work presented in this paper. Therefore, the ground-truth annotation values for arousal and valence are shifted back in time for the start of each recording. At the end of each recording, missing annotation values are padded as 0.0 rated arousal and valence values. The annotation delay values for the experiments vary in range from  $\pm 1$  second around the audio modality ground-truth time-shifts proposed by He et al. (2015) for arousal and valence, respectively. The same delays are incorporated for both speech and eye gaze modalities given the early feature fusion framework being used. This process is illustrated in Figure 6.

### 5.7. Performance evaluation

Performance evaluation of the proposed system on the RECOLA corpus (Ringeval et al., 2013b) is carried out using CCC (Lin, 1989). CCC combines a penalty of mean-squared error with a Pearson correlation as in Equation (1), where  $x$  represents the machine predicted values,  $y$  represents the ground-truth values,  $\sigma$  is the covariance,  $\sigma^2$  is the variance, and  $\mu$  is the mean. The CCC gives a measure of agreement between the machine predicted values and the

ground-truth values.

$$CCC = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

In addition to intra-corpus evaluation for the bimodal system that is trained, validated and tested on the French language RECOLA corpus (Ringeval et al., 2013b), a cross-corpus and cross-lingual evaluation is carried out. The AVEC 2014 corpus (Valstar et al., 2014) is used for this evaluation. The Freeform partition of this German language corpus is used to assess the performance of the bimodal network model that has been trained exclusively using the French language RECOLA corpus (Ringeval et al., 2013b). An additional cross-corpus and cross-lingual evaluation of a system trained using AVEC 2014 and tested using RECOLA is also carried out.

## 6. Results and discussion

In this section results and discussion are presented for bimodal (B)LSTM ground-truth time-shift, intra-corpus unimodal (B)LSTM and bimodal (B)LSTM experiments. The proposed bimodal system is compared against other work from the literature. Finally, results for the proposed system during the cross-corpus and cross-lingual experiment are presented and discussed.

### 6.1. Ground-truth time-shift evaluations

For the ground-truth time-shift experiment, the ground-truth time shifts proposed by He et al. (2015) for the audio modality are used as a guide for the ground-truth time-shifts tested in this work. He et al. (2015) found a shift back in time of 59 frames for arousal and 78 frames for valence to be optimal when considering audio in a unimodal fashion on the RECOLA corpus (Ringeval et al., 2013b). For the experiments conducted here, bimodal (B)LSTM-RNNs were trained on the training partition of the RECOLA corpus (Ringeval et al., 2013b), and tested on the validation partition. Figure 7 shows the validation set

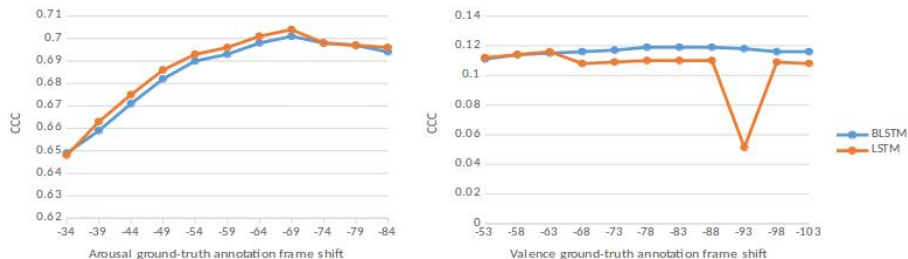


Figure 7: BLSTM-RNN arousal and valence prediction results using different ground-truth annotation time-shifts

performances of the arousal and valence (B)LSTM-RNNs with different time-shifts applied to the ground-truth annotations prior to training and testing.

Figure 7 clearly shows that altering the ground-truth annotation time-shift does have an effect on (B)LSTM-RNN performance. The highest performance achieved by both BLSTM and LSTM networks for arousal prediction had a 69 frame negative-shift applied to ground-truth annotations prior to training and testing. The highest performance achieved by BLSTM and LSTM networks for valence prediction had frame negative-shifts of 78 and 63 applied respectively. The RECOLA corpus (Ringeval et al., 2013b) was recorded at 25 frames per second, this results in a 2.76 seconds ground-truth time-shift for the arousal networks, and 3.12 (BLSTM) and 2.52 (LSTM) seconds ground-truth time-shift for the valence networks. These ground-truth time-shift values were used for the experiments described in the following sections.

## 6.2. Intra-corpora evaluations

Unimodal and bimodal LSTM and BLSTM network arousal and valence prediction performances achieved on the RECOLA corpus (Ringeval et al., 2013b) are shown in Table 7. The results for both LSTM and BLSTM systems clearly demonstrate the benefit of including eye gaze features with those of speech in a feature fusion framework for both arousal and valence prediction. The feature fusion BLSTM arousal prediction CCC of 0.754 is an increase in performance of 1.62% when compared to the highest performing unimodal speech system.

For valence, the feature fusion BLSTM CCC of 0.277 represents a 6.13% performance increase when compared to the highest performing unimodal speech system. LSTM provided better performance than BLSTM for arousal and valence prediction using speech alone. This suggests that past memory is more relevant for neural network training when using unimodal speech input. The highest performing networks for arousal and valence prediction overall, the BLSTM networks, show that future context in addition to past context is important for neural network training for continuous affect prediction in a bimodal system.

Table 7: RECOLA intra-corporis emotion dimension prediction CCC and arousal (Ar) and valence (Val) neural network hyperparameters for unimodal and bimodal experiments

Arousal	Valence	Learning Method	Modality	Random Seed	Learning Rate
0.732	0.261	LSTM	Speech	1787452436	$8 \times 10^{-5}$ (Ar)
				123456789	$5 \times 10^{-6}$ (Val)
0.232	0.02	LSTM	Eye gaze	1787452436	$7 \times 10^{-5}$ (Ar)
				123456789	$2 \times 10^{-6}$ (Val)
0.742	0.213	LSTM	Feature Fusion	1787452436	$9 \times 10^{-6}$ (Ar)
				1787452436	$3 \times 10^{-6}$ (Val)
0.735	0.255	BLSTM	Speech	1787452436	$1 \times 10^{-5}$ (Ar)
				123456789	$9 \times 10^{-7}$ (Val)
0.268	-0.001	BLSTM	Eye Gaze	1787452436	$3 \times 10^{-5}$ (Ar)
				1787452436	$4 \times 10^{-6}$ (Val)
<b>0.754</b>	<b>0.277</b>	BLSTM	Feature Fusion	1787452436	$1 \times 10^{-5}$ (Ar)
				1787452436	$1 \times 10^{-5}$ (Val)

### 6.3. Comparison with previous approaches from the literature

The (B)LSTM-based system proposed in this work is compared in Table 8 with previous approaches to intra-corporis evaluation on the RECOLA (Ringeval et al., 2013b) corpus. The central contribution of the work presented in this paper is to demonstrate the benefit that eye gaze can have when combined with speech in a continuous affect prediction system. However, a comparison against other works puts the achieved results in context within the literature. Therefore, the comparison includes other multimodal results from the literature that used the GeMAPS (Eyben et al., 2016) feature set for affect prediction on the RECOLA corpus (Ringeval et al., 2013b). Table 8 shows that the pro-

posed bimodal speech and eye gaze affect prediction system outperforms He et al. (2015) for arousal prediction. Compared to Brady et al. (2016) however, the proposed bimodal system presented here does not provide a performance increase for arousal prediction. Of note for this work when compared against Brady et al. (2016) is that a much simpler fusion approach and a smaller feature vector was used. Valence prediction performance for the proposed system does need improvement to be comparable to the literature. However, the central focus of this study is what the eye gaze modality can add to speech for continuous affect prediction, which was clearly demonstrated in Table 7.

Table 8: Emotion dimension prediction CCC of proposed bimodal system compared with approaches from the literature

Arousal	Valence	Feature Count	Fusion Method	Authors
0.747	0.609	270 (arousal), 259 (valence)	Gaussian smoothed model fusion	(He et al., 2015)
<b>0.770</b>	<b>0.687</b>	622	Kalman filter	(Brady et al., 2016)
0.754	0.277	119	Feature fusion	This work

#### 6.4. Cross-corpus and cross-lingual evaluations

During cross-corpus and cross-lingual experiments, the highest performing neural network models trained and tested using the French RECOLA corpus (Ringeval et al., 2013b) were tested using the German AVEC 2014 Freeform corpus (Valstar et al., 2014). As such, the feature fusion BLSTM-RNNs are selected for this task. Also, to complete the cross-comparisons, BLSTM arousal and valence networks were trained using AVEC 2014 Freeform (Valstar et al., 2014) and tested using RECOLA (Ringeval et al., 2013b). The ground-truth time-shift had to be altered to 84 frames for arousal and 96 frames for valence for the AVEC 2014 corpus training, validation and test sets for the experiment due to it being recorded at 30 frames per second. The cross-corpus test set results from the best performing neural networks from intra-corpus validation set experiments are shown in Table 9.

The cross-corpus and cross-lingual experiment undertaken provides a significant challenge for the neural network models developed. In contrast to Schuller



Table 9: Proposed BLSTM system cross-corpus and cross-lingual emotion dimension prediction CCC and arousal (Ar) and valence (Val) neural network hyperparameters

Arousal	Valence	Training corpus	Testing corpus	Random Seed	Learning Rate
0.263	0.097	RECOLA	AVEC Freeform	1787452436	$1 \times 10^{-5}$ (Ar)
				1787452436	$1 \times 10^{-5}$ (Val)
0.363	-0.081	AVEC Freeform	RECOLA	1787452436	$1 \times 10^{-5}$ (Ar)
				3926481610	$1 \times 10^{-5}$ (Val)

et al. (2010b), where Germanic languages only were used for cross-corpus emotion classification, this work uses Germanic (German) and Romance (French) languages for continuous affect prediction performance assessment. Additionally, the protocols used to compile the emotion corpora are also different, as presented in Table 5. The contribution of these results is that they are the first set of cross-corpus and cross-lingual results for continuous affect prediction from the literature reviewed. From the results in Table 9, clearly, arousal is more easily predicted across corpora and languages. Interestingly, networks trained using the German AVEC corpus (Valstar et al., 2014) and tested using the French RECOLA corpus (Ringeval et al., 2013b) performed 38.02% better on average during arousal prediction than networks trained using the French RECOLA corpus and tested using the German AVEC corpus. It is recognised by the affective computing research community that valence is much more difficult to predict compared to arousal. The difficulty for valence prediction here is compounded by the cross-corpus and cross-lingual prediction task.

## 7. Conclusions

From the review of the literature on affective computing using speech, it is clear that considering eye gaze with speech is relatively unexplored for affective computing, specifically for continuous affect prediction of emotion dimensions. This paper proposed a new bimodal affect prediction system using speech and eye gaze input to a BLSTM learning scheme. The proposed system design is simple, based on openly available software including OpenFace (Baltruaitis et al., 2016), openSMILE (Eyben et al., 2013), and CURRENNT (Weninger,

2015), and does not require the use of expensive or intrusive hardware.

From the experiments presented, it is clear that eye gaze features extracted from video improves arousal and valence prediction system performance when combined with speech features. The results from the proposed system are promising for arousal dimension prediction with the BLSTM variant outperforming LSTM for bimodal affect prediction. The best performing network for valence prediction needs further development for the proposed framework to produce results comparable to the literature for this emotion dimension. Altering temporal window sizes for feature calculation across different modalities is one option to improve valence dimension prediction, based on work by Ringeval et al. (2015a), and Valstar et al. (2016). More complex modality fusion and feature selection methods, and ground-truth reliability weighting (Grimm and Kroschel, 2005) are other potential avenues of investigation to improve system performance for this difficult to predict (Ringeval et al., 2013a) emotion dimension.

The novel cross-corpus results provided in this work serve as a baseline for researchers attempting cross-corpus, cross-lingual, and cross-protocol continuous affect prediction. The results achieved indicate that predicting arousal under these conditions is reasonable, but additional work is needed for the continuous affect prediction of valence during cross-corpus and cross-lingual experiments. From the results presented, it is hoped that the research community will consider eye gaze combined with speech in future multi-modal systems. Further evaluations including the proposed modalities within cross-corpus, cross-lingual, cross-task, and even cross-culture affective computing experiments are hoped for, to increase model generalisability and impact.

Future work will include ground-truth reliability weighting, a model fusion comparison against the proposed system, and transfer learning during cross-corpus experiments. These studies will allow investigation of the contribution of eye gaze when combined with speech for continuous affect prediction under further experimental conditions. Further cross-corpus continuous affect prediction studies on an intra-lingual basis and the addition of further non-intrusive

modalities to the proposed (B)LSTM-RNN based system are also planned.

### **Acknowledgement**

Funding: This work was supported by the Irish Research Council [grant number GOIPG/2016/1572].

### **References**

### **References**

Aracena, C., Basterrech, S., Snel, V., Velsquez, J.. Neural Networks for Emotion Recognition Based on Eye Tracking Data. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics. 2015. p. 2632–2637. doi:10.1109/SMC.2015.460.

Asgari, M., Kiss, G., Santen, J.v., Shafran, I., Song, X.. Automatic measurement of affective valence and arousal in speech. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. p. 965–969. doi:10.1109/ICASSP.2014.6853740.

B. Davis, S., Mermelstein, P.. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Processing* 28 1980;28:357 – 366.

Baltruaitis, T., Robinson, P., Morency, L.P.. OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). 2016. p. 1–10. doi:10.1109/WACV.2016.7477553.

Barker, J., Marxer, R., Vincent, E., Watanabe, S.. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). 2015. p. 504–511. doi:10.1109/ASRU.2015.7404837.

- Bhat, C., Mithun, B.S., Saxena, V., Kulkarni, V., Kopparapu, S.. Deploying usable speech enabled IVR systems for mass use. In: 2013 International Conference on Human Computer Interactions (ICHCI). 2013. p. 1–5. doi:10.1109/ICHCI-IEEE.2013.6887794.
- Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., Huang, T.S.. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: ACM; AVEC '16; 2016. p. 97–104. URL: <http://doi.acm.org/10.1145/2988257.2988264>. doi:10.1145/2988257.2988264.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.. A Database of German Emotional Speech. In: INTERSPEECH-2005. 2005. p. 1517–1520.
- Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 2008;42(4):335.
- Canavan, A., Graff, D., Zipperlen, G.. CALLHOME American English Speech. 1997. URL: <https://catalog.ldc.upenn.edu/l1dc97s42>.
- Chakraborty, R., Pandharipande, M., Kopparapu, S.K.. Knowledge-based Framework for Intelligent Emotion Recognition in Spontaneous Speech. *Procedia Computer Science* 2016;96:587–596. URL: <http://www.sciencedirect.com/science/article/pii/S187705091632049X>. doi:10.1016/j.procs.2016.08.239.
- C.K., Y., Hariharan, M., Ngadiran, R., Adom, A.H., Yaacob, S., Polat, K.. Hybrid BBO\_pso and higher order spectral features for emotion and stress recognition from natural speech. *Applied Soft Computing* 2017;56:217–232. URL: <http://www.sciencedirect.com/science/article/pii/S1568494617301369>. doi:10.1016/j.asoc.2017.03.013.

- Dai, W., Han, D., Dai, Y., Xu, D.. Emotion recognition and affective computing on vocal social media. *Information & Management* 2015;52(7):777–788. URL: <http://www.sciencedirect.com/science/article/pii/S037872061500018X>. doi:10.1016/j.im.2015.02.003.
- Durston, P.J., Farrell, M., Attwater, D., Allen, J., kwang Jeff Kuo, H., Afify, M., Fosler-lussier, E., hui Lee, C.. Oasis natural language call steering trial. In: In EUROSpeech-2001. 2001. p. 1323–1326.
- Engbert, L., Hansen, A.. Documentation of the Danish Emotional Speech Database DES. Technical; Aalborg Univ., Denmark; Center for PersonKommunikation; 2007.
- Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andr, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 2016;7(2):190–202. doi:10.1109/TAFFC.2015.2457417.
- Eyben, F., Weninger, F., Gross, F., Schuller, B.. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In: *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA: ACM; MM '13; 2013. p. 835–838. URL: <http://doi.acm.org/10.1145/2502081.2502224>. doi:10.1145/2502081.2502224.
- Fayek, H.M., Lech, M., Cavedon, L.. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* 2017;92:60–68. URL: <http://www.sciencedirect.com/science/article/pii/S089360801730059X>. doi:10.1016/j.neunet.2017.02.013.
- Feraru, S.M., Schuller, D., Schuller, B.. Cross-language acoustic emotion recognition: An overview and some tendencies. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2015. p. 125–131. doi:10.1109/ACII.2015.7344561.

- Fernandez, R., Picard, R.. Recognizing affect from speech prosody using hierarchical graphical models. *Speech Communication* 2011;53(9):1088–1103. URL: <http://www.sciencedirect.com/science/article/pii/S0167639311000689>. doi:10.1016/j.specom.2011.05.003.
- Grimm, M., Kroschel, K.. Evaluation of natural emotions using self assessment manikins. In: *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005. 2005. p. 381–385. doi:10.1109/ASRU.2005.1566530.
- Grimm, M., Kroschel, K., Narayanan, S.. The Vera am Mittag German audio-visual emotional speech database. In: *2008 IEEE International Conference on Multimedia and Expo*. 2008. p. 865–868. doi:10.1109/ICME.2008.4607572.
- Gupta, R., Narayanan, S.S.. Predicting Affective Dimensions Based on Self Assessed Depression Severity. In: *Interspeech 2016*. 2016. p. 1427–1431. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/0187.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0187.html). doi:10.21437/Interspeech.2016-187.
- Hansen, J., Bou-Gazale, B.. Getting Started with SUSAS: A Speech under Simulated and Actual Stress Database. In: *EUROSPEECH*. volume 4; 1997. p. 1743–1746.
- Haq, S., Jackson, P.J., Edge, J.. Audio-visual feature selection and reduction for emotion classification. In: *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP08)*, Tangalooma, Australia. 2008. p. 185–190.
- He, L., Jiang, D., Yang, L., Pei, E., Wu, P., Sahli, H.. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: ACM; AVEC '15; 2015. p. 73–80. URL: <http://doi.acm.org/10.1145/2808196.2811641>. doi:10.1145/2808196.2811641.
- He, R., Zheng, W.S., Hu, B.G.. Maximum Correntropy Criterion for Ro-

- bust Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2011;33(8):1561–1576. doi:10.1109/TPAMI.2010.220.
- Hochreiter, S., Schmidhuber, J.. Long Short-Term Memory. *Neural Computation* 1997;9(8):1735–1780. URL: <http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735>. doi:10.1162/neco.1997.9.8.1735.
- Itier, R.J., Batty, M.. Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews* 2009;33(6):843–863. URL: <http://www.sciencedirect.com/science/article/pii/S0149763409000207>. doi:10.1016/j.neubiorev.2009.02.004.
- J Lang, P., M Bradley, M., N Cuthbert, B.. International affective picture system (iaps): Affective ratings of pictures and instruction manual (rep. no. a-8) 2008;.
- Jenssen, R.. Kernel Entropy Component Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010;32(5):847–860. doi:10.1109/TPAMI.2009.100.
- Kächele, M., Schels, M., Schwenker, F.. Inferring Depression and Affect from Application Dependent Meta Knowledge. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: ACM; AVEC '14; 2014. p. 41–48. URL: <http://doi.acm.org/10.1145/2661806.2661813>. doi:10.1145/2661806.2661813.
- Kopparapu, S.K.. *Non-Linguistic Analysis of Call Center Conversations*. Springer Publishing Company, Incorporated, 2014.
- Kossaifi, J., Tzimiropoulos, G., Todorovic, S., Pantic, M.. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 2017;URL: <http://www.sciencedirect.com/science/article/pii/S0262885617300379>. doi:10.1016/j.imavis.2017.02.001.
- L. Fleiss, J.. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 1971;76:378. doi:10.1037/h0031619.

- Lanatà, A., Armato, A., Valenza, G., Scilingo, E.P.. Eye tracking and pupil size variation as response to affective stimuli: A preliminary study. In: 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops. 2011. p. 78–84. doi:10.4108/icst.pervasivehealth.2011.246056.
- Lappi, O.. Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference. *Neuroscience & Biobehavioral Reviews* 2016;69:49–68. URL: <http://www.sciencedirect.com/science/article/pii/S0149763415301317>. doi:10.1016/j.neubiorev.2016.06.006.
- Lin, L.I.K.. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989;45(1):255–268. URL: <http://www.jstor.org/stable/2532051>. doi:10.2307/2532051.
- Martin, O., Kotsia, I., Macq, B., Pitas, I.. The eNTERFACE'05 Audio-Visual Emotion Database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06). 2006. p. 8–8. doi:10.1109/ICDEW.2006.145.
- Mathieu, B., Essid, S., Fillon, T., Prado, J., Richard, G.. Yaafe, an easy to use and efficient audio feature extraction software. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. Utrecht, The Netherlands; 2010. p. 441–446. <http://ismir2010.ismir.net/proceedings/ismir2010-75.pdf>.
- McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing* 2012;3(1):5–17. doi:10.1109/T-AFFC.2011.20.
- Mencattini, A., Martinelli, E., Ringeval, F., Schuller, B., Natlae, C.D.. Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models. *IEEE Transactions on Affective Computing* 2017;PP(99):1–1. doi:10.1109/TAFFC.2016.2531664.



- Metallinou, A., Yang, Z., Lee, C.C., Busso, C., Carnicke, S., Narayanan, S.. The USC CreativeIT Database of Multimodal Dyadic Interactions: From Speech and Full Body Motion Capture to Continuous Emotional Annotations. *Lang Resour Eval* 2016;50(3):497–521. URL: <http://dx.doi.org/10.1007/s10579-015-9300-0>. doi:10.1007/s10579-015-9300-0.
- Morency, L.P., Mihalcea, R., Doshi, P.. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In: *Proceedings of the 13th International Conference on Multimodal Interfaces*. New York, NY, USA: ACM; ICMI '11; 2011. p. 169–176. URL: <http://doi.acm.org/10.1145/2070481.2070509>. doi:10.1145/2070481.2070509.
- Motamed, S., Setayeshi, S., Rabiee, A.. Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures* 2017;19:32–38. URL: <http://www.sciencedirect.com/science/article/pii/S2212683X16301219>. doi:10.1016/j.bica.2016.12.002.
- Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., Fookes, C.. Deep Spatio-Temporal Features for Multimodal Emotion Recognition. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017. p. 1215–1223. doi:10.1109/WACV.2017.140.
- Nicolaou, M.A., Gunes, H., Pantic, M.. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Transactions on Affective Computing* 2011;2(2):92–105. doi:10.1109/T-AFFC.2011.9.
- O'Dwyer, J., Flynn, R., Murray, N.. Continuous affect prediction using eye gaze. In: *2017 28th Irish Signals and Systems Conference (ISSC)*. 2017a. p. 1–6. doi:10.1109/ISSC.2017.7983611.
- O'Dwyer, J., Flynn, R., Murray, N.. Continuous affect prediction using eye gaze and speech. In: *2017 IEEE International Conference on Bioinformatics*

- and Biomedicine (BIBM). 2017b. p. 2001–2007. doi:10.1109/BIBM.2017.8217968.
- Partala, T., Surakka, V.. Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies* 2003;59(1):185–198. URL: <http://www.sciencedirect.com/science/article/pii/S107158190300017X>. doi:10.1016/S1071-5819(03)00017-X.
- Poria, S., Cambria, E., Bajpai, R., Hussain, A.. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 2017;37:98–125. URL: <http://www.sciencedirect.com/science/article/pii/S1566253517300738>. doi:10.1016/j.inffus.2017.02.003.
- Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). 2016. p. 439–448. doi:10.1109/ICDM.2016.0055.
- Ranganathan, H., Chakraborty, S., Panchanathan, S.. Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). 2016. p. 1–9. doi:10.1109/WACV.2016.7477679.
- Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.P., Ebrahimi, T., Lalanne, D., Schuller, B.. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 2015a;66:22–30. URL: <http://www.sciencedirect.com/science/article/pii/S0167865514003572>. doi:10.1016/j.patrec.2014.11.007.
- Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M.. 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*. New York, NY, USA: ACM; AVEC '15; 2015b. p. 3–8. URL: <http://doi.acm.org/10.1145/2808196.2811642>. doi:10.1145/2808196.2811642.

- Ringeval, F., Sonderegger, A., Noris, B., Billard, A., Sauer, J., Lalanne, D.. On the Influence of Emotional Feedback on Emotion Awareness and Gaze Behavior. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013a. p. 448–453. doi:10.1109/ACII.2013.80.
- Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013b. p. 1–8. doi:10.1109/FG.2013.6553805.
- Schmitt, M., Ringeval, F., Schuller, B.. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In: Interspeech 2016. 2016. p. 495–499. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/1124.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1124.html). doi:10.21437/Interspeech.2016-1124.
- Schuller, B., Meller, R., Hernler, B., Hethker, A., Konosu, H., Rigoll, G.. Audiovisual Recognition of Spontaneous Interest Within Conversations. In: Proceedings of the 9th International Conference on Multimodal Interfaces. New York, NY, USA: ACM; ICMI '07; 2007. p. 30–37. URL: <http://doi.acm.org/10.1145/1322192.1322201>. doi:10.1145/1322192.1322201.
- Schuller, B., Steidl, S., Batliner, A.. The interspeech 2009 emotion challenge. In: Tenth Annual Conference of the International Speech Communication Association. 2009. p. 312–315.
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.S.. The interspeech 2010 paralinguistic challenge. In: Eleventh Annual Conference of the International Speech Communication Association. 2010a. p. 2794–2797.
- Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., Evanini, K.. The

- INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity and Native Language. In: INTERSPEECH 2016. 2016. p. 2001–2005. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/0129.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0129.html). doi:10.21437/Interspeech.2016-129.
- Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing* 2010b;1(2):119–131. doi:10.1109/T-AFFC.2010.8.
- Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 2012;3(1):42–55. doi:10.1109/T-AFFC.2011.25.
- Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y., Liu, J., Yu, Y.. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication* 2016;83:34–41. URL: <http://www.sciencedirect.com/science/article/pii/S0167639316300966>. doi:10.1016/j.specom.2016.07.010.
- Steininger, S., Schiel, F., Dioubina, O., Raubold, S.. Development of User-State Conventions for the Multimodal Corpus in Smartkom. In: *The Proc. of the 3rd Int. Conf. On Language Resources and Evaluation, Workshop on Multimodal Resources and Multimodal Systems Evaluation*. 2002. p. 33–37.
- Stratou, G., Morency, L.P.. MultiSense #8212;Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. *IEEE Transactions on Affective Computing* 2017;8(2):190–203. doi:10.1109/TAFFC.2016.2614300.
- Subrahmanya, N., Shin, Y.C.. Sparse Multiple Kernel Learning for Signal Processing Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010;32(5):788–798. doi:10.1109/TPAMI.2009.98.

- Thushara, S., Veni, S.. A multimodal emotion recognition system from video. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT). 2016. p. 1–5. doi:10.1109/ICCPCT.2016.7530161.
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016. p. 5200–5204. doi:10.1109/ICASSP.2016.7472669.
- Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: ACM; AVEC '16; 2016. p. 3–10. URL: <http://doi.acm.org/10.1145/2988257.2988258>. doi:10.1145/2988257.2988258.
- Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.. AVEC 2014: 3d Dimensional Affect and Depression Recognition Challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. New York, NY, USA: ACM; AVEC '14; 2014. p. 3–10. URL: <http://doi.acm.org/10.1145/2661806.2661807>. doi:10.1145/2661806.2661807.
- Vlasenko, B., Schuller, B., Wendemuth, A.. Tendencies regarding the effect of emotional intensity in inter corpus phoneme-level speech emotion modelling. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). 2016. p. 1–6. doi:10.1109/MLSP.2016.7738859.
- Wang, Y., Guan, L.. Recognizing Human Emotional State From Audiovisual Signals\*. IEEE Transactions on Multimedia 2008;10(5):936–946. doi:10.1109/TMM.2008.927665.
- Webber Jr, C.L., Zbilut, J.P.. Recurrence quantification analysis of nonlinear

- dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences* 2005;:26–94.
- Weninger, F.. Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit. *Journal of Machine Learning Research* 2015;16:547–551. URL: <http://jmlr.org/papers/v16/weninger15a.html>.
- Wu, S., Falk, T.H., Chan, W.Y.. Automatic speech emotion recognition using modulation spectral features. *Speech Communication* 2011;53(5):768–785. URL: <http://www.sciencedirect.com/science/article/pii/S0167639310001470>. doi:10.1016/j.specom.2010.08.013.
- Xie, Z., Tie, Y., Guan, L.. A new audiovisual emotion recognition system using entropy-estimation-based multimodal information fusion. In: *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2015. p. 726–729. doi:10.1109/ISCAS.2015.7168736.
- Yang, Z., Narayanan, S.. Lightly-supervised utterance-level emotion identification using latent topic modeling of multimodal words. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016. p. 2767–2771. doi:10.1109/ICASSP.2016.7472181.
- Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., Schuller, B.. Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks. In: *Interspeech 2016*. 2016. p. 3593–3597. URL: [http://www.isca-speech.org/archive/Interspeech\\_2016/abstracts/0998.html](http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0998.html). doi:10.21437/Interspeech.2016-998.
- Zhao, Y., Wang, X., Petriu, E.M.. Facial expression analysis using eye gaze information. In: *2011 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSAS) Proceedings*. 2011. p. 1–4. doi:10.1109/CIMSAS.2011.6059936.



Jonny O' Dwyer received the BEng degree in Software Engineering from Athlone Institute of Technology. He is currently a postgraduate MSc research student at Athlone Institute of Technology. His research interests include affective computing and machine learning. He is a Government of Ireland Postgraduate Scholar.



Niall Murray is a faculty member of Engineering & Informatics in Athlone Institute of Technology. He received the BE degree in Electronic Engineering from National University of Ireland, Galway, the MEng degree from University of Limerick and the PhD degree from AIT. He is the founder of and the principal investigator in the truly Immersive and Interactive Multimedia Experiences research group. He is a Science Foundation Ireland funded investigator in the Confirm Centre for Smart manufacturing and an associate investigator on the Enterprise Ireland-funded COMAND technology gateway. His research interests include immersive and multi-sensory multimedia communication, and quality of experience.



Ronan Flynn received the BE degree in Electronic Engineering from University College Dublin, the MEng degree in Computer Systems from University of Limerick and the PhD degree from National University of Ireland, Galway. Dr. Flynn is currently a lecturer and researcher in the Faculty of Engineering & Informatics in Athlone Institute of Technology. He has been involved in the evaluation of H2020 proposals submitted to the European

Commision. He was on the management committee for European COST Action IC1206 "De-identification for privacy protection in multimedia content". His research interests include speech recognition, emotion recognition in speech and multi-modal affective computing.