Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches

Ehab A. AlBadawy
University at Albany, SUNY
Albany, NY
ealbadawy@albany.edu

Yelin Kim
University at Albany, SUNY
Albany, NY
yelinkim@albany.edu

ABSTRACT

This paper presents a novel approach in continuous emotion prediction that characterizes dimensional emotion labels jointly with continuous and discretized representations. Continuous emotion labels can capture subtle emotion variations, but their inherent noise often has negative effects on model training. Recent approaches found a performance gain when converting the continuous labels into a discrete set (e.g., using *k*-means clustering), despite a label quantization error. To find the optimal trade-off between the continuous and discretized emotion representations, we investigate two joint modeling approaches: ensemble and end-to-end. The ensemble model combines the predictions from two models that are trained separately, one with discretized prediction and the other with continuous prediction. On the other hand, the end-to-end model is trained to simultaneously optimize both discretized and continuous prediction tasks in addition to the final combination between them. Our experimental results using the state-of-the-art deep BLSTM network on the RECOLA dataset demonstrate that (i) the joint representation outperforms both individual representation baselines and the state-of-the-art speech–based results on RECOLA, validating the assumption that combining continuous and discretized emotion representations yields better performance in emotion prediction; and (ii) the joint representation can help to accelerate convergence, particularly for valence prediction. Our work provides insights into joint discrete and continuous emotion representation and its efficacy for describing dynamically changing affective behavior in valence and activation prediction.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; Multimedia information system; • **Computing methodologies → Artificial intelligence**; Artificial intelligence;

KEYWORDS

Emotion recognition; Continuous emotion prediction; Joint representation; Bidirectional long- short-term memory

ACM Reference format:
Ehab A. AlBadawy and Yelin Kim. 2018. Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches. In *Proceedings of 2018 International Conference on MultimodalInteraction, Boulder, CO, USA, October 16–20, 2018 (ICMI '18),* 10 pages.
https://doi.org/10.1145/3242969.3242972

1 INTRODUCTION

Emotion recognition has gained a great interest in the multimodal interaction community [4, 22, 48]. Audio-visual expressive behavior includes salient information for understanding the overall tone and affective context of interaction [3, 26, 35]. Continuous emotion prediction, a process of estimating emotion in continuous time, uses dimensional representation of emotion (e.g., arousal and valence) [11, 31]. Continuous-time prediction allows us to better understand the dynamic behavior of affective interaction [28] and potentially inform more natural and timely responses in interactive systems [11, 36, 37]; however, the main challenge is that continuous emotion labels often contain inherent noise that results in negative consequences in training and prediction. In this paper, we explore jointly modeling continuous and discrete emotion labels and investigate the optimal trade-off between the two label representations.

Previous studies on continuous emotion prediction focused either on (i) regression approaches that directly used continuous emotion representation (e.g., arousal: 0.386) [5, 8, 16, 21, 24, 28, 30, 39, 41, 46] or on (ii) classification approaches that used discretized representation that quantized continuous labels into a discrete set of categories (e.g., arousal: "high" category) [27, 45, 47, 49, 51, 52]. The regression approaches with continuous emotion labels retain the full label information during training; however, inherent noise in emotion labels [12, 50] may present negative effects in training, such as decreased accuracy or increased model complexity [7]. In contrast, the classification approaches with discretized emotion labels attempted to reduce label noise using various quantization methods, such as Affinity Propagation–based clustering [52], binarization based on the mean (low and high classes) [27], label modeling over grid cells [49, 51], and *k*-means clustering [20]. These discretization and classification approaches have shown to be effective in emotion prediction at the expense of a label quantization error [8, 20]. The previous studies demonstrate the pros and cons of regression and classification approaches; however, open questions remain regarding how we can find the optimal trade-off between the two approaches, and whether we can jointly use continuous and discrete emotion representation.

This gap led us to explore two joint modeling methods for combining the regression and classification approaches: *ensemble* and *end-to-end* models. These models combine the two tasks at the prediction (decision) and representation levels, respectively. First, the ensemble model combines the predictions from two separate models that are optimized independently for regression and classification tasks. This model can make the optimal prediction for each

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3242972

task, and previous research has shown the efficacy of ensemble approaches in multiple domains, such as speech recognition [6], image recognition [13], and emotion recognition [29, 38]. Next, the proposed end-to-end model is trained to simultaneously optimize regression, classification, and the final combination between these two, in an end-to-end manner. The advantage of the end-to-end model is that it does not require any manually designed intermediary algorithms (e.g., averaging in ensemble models). For the end-to-end model, we further investigate how the classification and regression losses should be combined by introducing a new total loss function. We also explore the benefit of our proposed joint representation compared to a fully connected neural network that does not exploit this representation.

We use the benchmark Remote Collaborative and Affective Interactions (RECOLA) dataset [35] and Concordance Correlation Coefficient (CCC) performance to examine the proposed methods in the state-of-the-art context. We build strong baselines using deep bidirectional long short-term memory (D-BLSTM) for individual classification and regression tasks, similar to [20]. Our proposed ensemble and end-to-end models use D-BLSTM as a building block for joint representation. Since the previous state-of-the-art work that used a discretization method on RECOLA focused on speech emotion recognition [20], we first use speech to explore different joint modeling methods. We then investigate how our proposed joint modeling works in an audio-visual environment.

Our experimental results suggest that joint modeling of discrete and continuous emotion labels results in more accurate emotion prediction. Our joint modeling approaches also achieve the state-of-the-art performance on RECOLA when using speech. Our work can increase the understanding of the representation of multimodal interaction in continuous time. In summary, the main novelty of this work includes (i) the design and development of new joint modeling methods using D-BLSTMs, (ii) the new insight into a high-level joint representation of discrete and continuous emotion, and (iii) the investigation of the trade-off between classification and regression tasks for emotion prediction.

2 BACKGROUND: CONTINUOUS EMOTION PREDICTION

Continuous emotion prediction is a process of predicting high-level emotion in continuous time and modeling the dynamics of emotion fluctuation. For instance, Khorram et al. [18] proposed a combination of two Convolutional Neural Network (CNN) architectures to capture long-term temporal dependencies in speech emotion; the first architecture uses a stack of dilated convolutions that has been shown to improve image segmentation, speech synthesis, and automatic speech recognition, and the second architecture contains a downsampling/upsampling network that downsamples the input signal and upsamples the generated predictions. The second architecture helps to handle unstable prediction in the first architecture. The results showed that this network can not only improve the emotion prediction, but also generate smoothed predictions.

Soleymani et al. [40] used a linear ridge regression algorithm to map features to each video clip. They used arousal, valence, dominance, and liking ratings. They used physiological responses of 32 participants, recorded when they were watching emotional music videos. They used a leave-one-out cross-validation strategy for each participant, and they achieved significant improvement in the continuous emotion prediction in comparison to random estimation.

Tzirakis et al. [43] used a CNN-Recurrent Neural Network (RNN) approach using both speech and visual data for continuous emotion prediction. They presented two different models for each of the audio and video modalities. The first model is called Visual Network, where they used a deep residual network (ResNet) with 50 layers [13]. They used pixel intensities from cropped faces of the subject's video as an input to the ResNet. The second model is called a Speech Network, where they used a two-layer CNN model on the raw audio signal. Both models are followed by an LSTM layer to handle the temporal variations in the data. The results showed that the proposed models perform significantly better than other models using the RECOLA database.

BLSTM models have shown to be effective in continuous emotion prediction, since they can effectively capture long-term temporal dependencies in data. For example, Metallinou et al. [28] proposed a hierarchical approach using BLSTM and Hidden Markov Model (HMM) classifiers to model emotion state for each utterance, and they demonstrated the efficacy of a hybrid HMM/BLSTM model. He et al. [14] used a D-BLSTM–based fusion architecture between different modalities (e.g., audio and video) for continuous emotion prediction. They smoothed the initial predictions from a unimodal D-BLSTM with Gaussian smoothing and input these predictions into a second layer of D-BLSTM for the final prediction. However, it is less explored how BLSTM models can be used for joint representation of continuous and discrete labels.

The winning submission to the Audio/Visual Emotion Challenge (AVEC) 2016 [44], Brady et al. [2], used a multimodal system for continuous emotion prediction using RECOLA. For audio data, they used sparse coding to learn higher-order representation of the extracted features. They used Mel-frequency cepstral coefficients (MFCC), shifted delta cepstral (SDC), and prosody features. They achieved a significant improvement in arousal compared to the baseline scores. For the visual data, they used a CNN with three convolution layers on the detected face in each frame. Additionally, they proposed a multi-sensor fusion of emotional states while maintaining the emotional states' variation over time. However, this study did not take into account the noise in the ground truth emotion labels [20].

To overcome that limitation, recent work by Le et al. [20] has demonstrated that using classification with discretized emotion labels approaches can help reducing the label noise in the continuous values. They proposed a novel clustering–based discretization method of continuous emotion labels for speech–based continuous emotion prediction. They trained a D-BLSTM model with multi-task learning to capture the discretized emotion labels over time. Further, they introduced a decoding framework of an HMM–based language model. The results showed that discretization–based classification outperforms a traditional regression model, and achieved the state-of-the-art performance on RECOLA dataset. This work motivated our investigation of joint modeling approaches; however, our work differs from Le et al. in that we focus on combining classification and regression tasks to simultaneously reduce the label noise and regress its prediction without any quantization error. We also provide insights into how our proposed models perform in a multimodal environment.

k	number of frames
4	8296, 15162, 22540, 21502
6	3839, 9827, 11283, 14816, 17487, 10248
8	1964, 5605, 7586, 7847, 10862, 12653, 14179, 6804
10	1173, 3343, 5446, 6346, 6673, 9006, 9854, 10792, 10495, 4372

Table 1: Number of frames in the training set (per each of the k clusters)

3 DATA AND FEATURES

In this work, we use the RECOLA dataset [35], a widely used recent benchmark dataset in continuous emotion prediction [4, 42]. Particularly, the dataset has been used in the AVEC [34, 44] and has resulted in numerous advancements in emotion recognition [2, 17, 18, 20, 23]. Hence, this dataset allows us to evaluate and compare our proposed methods in the state-of-the-art context, and makes it easier to reproduce our work.

RECOLA contains spontaneous and naturalistic interactions of 27 French-speaking subjects, each with five-minute recordings. All the subjects were recorded in dyads during a video conference while completing a task requiring collaboration. The corpus contains multimodal cues, including audio, video, electrocardiogram (ECG), and electrodermal activity (EDA). The dataset is divided into train, development and test sets, and nine different subjects present in each set. Ground truth emotion labels, arousal and valence, are obtained from six gender-balanced French-speaking annotators as continuous-time ratings (40 ms binned frames) using ANNEMO [35]. The continuous labels range from -1 to +1. In this paper, we use the Kaldi toolkit [32] to extract a 40-dimensional log Mel filter bank coefficient with 25 ms window size and 10 ms frame shift to be consistent with previous work [18, 20]. To have the same number of input and output frames, we concatenate every four consecutive frames of the input, resulting in a 160-dimensional feature vector for each frame [20]. We perform z-normalization per each session on both audio and visual features.

To explore joint emotion representation in a multimodal environment, we also use visual appearance and geometric features provided in AVEC 2015, which used the RECOLA dataset [34]. The use of the challenge baseline features allows us to directly compare our modeling with baselines and foster reproducibility of our work. The appearance features were calculated using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) for each frame. The dimensionality of the features was reduced to 84 after Principal Component Analysis–based feature selection. The geometric features were computed using the Euclidean distances between 49 facial landmarks [34], resulting in 316 features. The missing frames of these features were interpolated. In total, the dimensionality of visual features is 400.

4 PROPOSED METHODS

In this section, we first describe our baselines that use the state-of-the-art D-BLSTM architecture for classification (*clf*) and regression (*reg*) tasks (Section 4.1). We then describe our two joint modeling approaches: ensemble and end-to-end (Section 4.2).

4.1 Classification (*clf*) and Regression (*reg*)

We build strong baselines for classification and regression using D-BLSTM for *clf* and *reg* models that have similar architecture to those proposed in Le et al. [20], although we use these models as baselines. D-BLSTM is a stack of BLSTM layers; each layer is a standard LSTM cell [15] that processes the input sequence in both the forward and the backward direction. An LSTM cell has an internal cell state (c_τ) at time τ that is computed based on the current input (x_τ) and the previous cell state ($c_{\tau-1}$). The combination of input (i_τ) and forget (f_τ) gates determines how much the previous cell state $c_{\tau-1}$ and the current input x_τ contribute to the current cell state c_τ. The activation function for both forget f_τ in input i_τ gates is a sigmoid function (σ) that outputs values between 0 and 1. Specifically, the current cell state c_τ is computed as follows:

$$i_\tau = \sigma(W_{xi}x_\tau + W_{hi}h_{\tau-1} + W_{ci}c_{\tau-1} + b_i) \tag{1}$$

$$f_\tau = \sigma(W_{xf}x_\tau + W_{hf}h_{\tau-1} + W_{cf}c_{\tau-1} + b_f) \tag{2}$$

$$\tilde{c}_\tau = tanh(W_{xc}x_\tau + W_{hc}h_{\tau-1} + b_c) \tag{3}$$

$$c_\tau = f_\tau * c_{\tau-1} + i_\tau * \tilde{c}_\tau \tag{4}$$

The current cell output for time τ is computed using the current cell state c_τ:

$$o_\tau = \sigma(W_{xo}x_\tau + W_{ho}h_{\tau-1} + W_{co}c_{\tau-1} + b_o) \tag{5}$$

$$h_\tau = o_\tau * tanh(c_\tau) \tag{6}$$

where o_τ is the output gate that determines the current cell state c_τ contribution to the current output h_τ. We can also rewrite h_τ and c_τ as follows:

$$(h_\tau, c_\tau) = \mathcal{F}(x_\tau, h_{\tau-1}, c_{\tau-1})$$

where $\mathcal{F}$ is the LSTM activation function. For BLSTM, h_τ will be a composite of the forward and backward directions $h_\tau = [\overrightarrow{h_\tau}; \overleftarrow{h_\tau}]$ and it is defined as follows:

$$(\overrightarrow{h}_\tau, \overrightarrow{c}_\tau) = \overrightarrow{\mathcal{F}}(x_\tau, \overrightarrow{h}_{\tau-1}, \overrightarrow{c}_{\tau-1}) \tag{7}$$

$$(\overleftarrow{h}_\tau, \overleftarrow{c}_\tau) = \overleftarrow{\mathcal{F}}(x_\tau, \overleftarrow{h}_{\tau-1}, \overleftarrow{c}_{\tau-1}) \tag{8}$$

For the classification (*clf*) model, we use a D-BLSTM network with discretized labels for emotion prediction. As in Le et al. [20], we use k-means clustering to divide the output range into discretized regions. The discretization of the continuous labels helps to reduce the noise in label values and makes it easier to train. Choosing the number of the clusters for the k-means model is crucial, since it affects both how precise the conversion will be from the discretized signal to the original one and how difficult it is to train the networks [20]. In order to achieve the balance between both the precision and difficulty, Le et al. [20] proposed a suite of four k-means models with $k = 4, 6, 8$, and 10. Table 1 shows the number of frames in each cluster for the training set of RECOLA.

Le et al. introduced a cost-sensitive cross-entropy (CCE) loss [20] for the *clf* model. In this paper, we propose a new cost function C_{norm}, and the loss function is defined as follows, similar to the CCE:

$$\ell_{clf} = \sum_{t=1}^{4} \frac{1}{F} \sum_{f=1}^{F} C_{norm}(y_{tf}, \hat{y}_{tf}) \sum_{l=1}^{L_t} y_{tf}^{(l)}. \log \hat{y}_{tf}^{(l)} \tag{9}$$

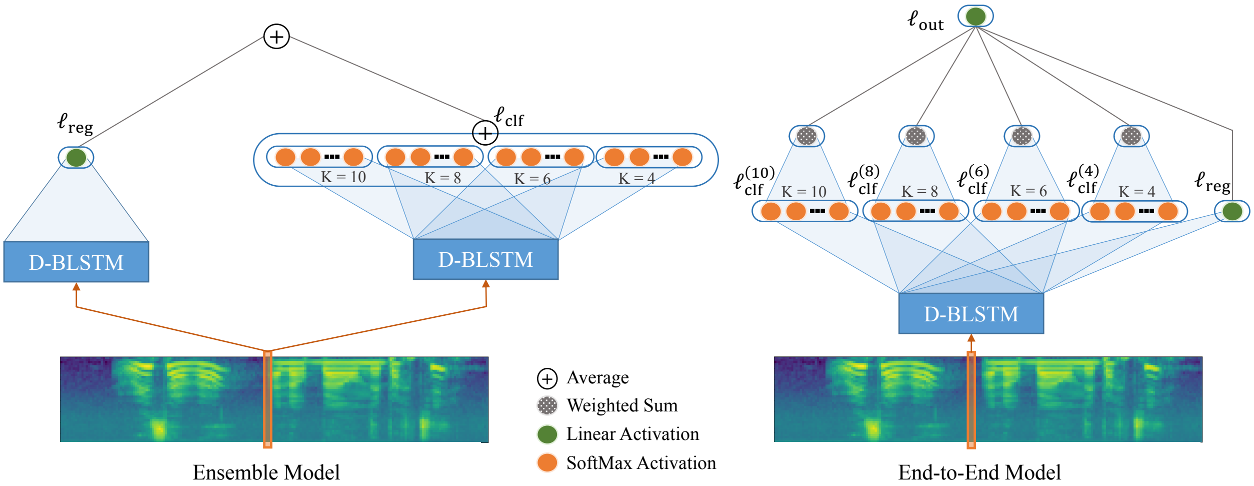

Figure 1: Our proposed joint modeling approaches using D-BLSTM architecture for ensemble (left) and end-to-end (right) models. The ensemble model combines the classification and regression predictions at the decision level, whereas the end-to-end model combines the classification (ℓ_{clf}), regression (ℓ_{reg}), and final combination of the two tasks (ℓ_{out}) at the representation level (using the total loss function, ℓ_{total}, in Eq. 12). The figure is best shown in color.

where F is the number of frames and L_t is the number of labels in the task t. The variables $y_{tf}^{(l)}$ and $\hat{y}_{tf}^{(l)}$ are the one-hot encoded ground truth and the predicted probability, respectively, at label index l for frame f and target function t.

Our proposed norm cost function C_{norm} is defined as follows:

$$C_{norm}(y_{tf}, \hat{y}_{tf}) = 1 + \left\| \sum_{l=1}^{L_t} K_{L_t}^{(l)} (y_{tf}^{(l)} - \hat{y}_{tf}^{(l)}) \right\|_2 \tag{10}$$

where $K_{L_t}^{(l)}$ is the centroid of the label l for the k-means model that has L_t labels. For example, if K_{L_t} can be $[-0.31, -0.12, 0.05, 0.22]$ for $L_t = 4$, $\hat{y}_{tf}$ can be $[0.1, 0.6, 0.2, .1]$, and y_{tf} can be $[0, 0, 1, 0]$. The cost function C_{norm} takes into consideration the spatial relation between the labels, as it helps to have more stable training [20] where its value is 1 for normal cross-entropy calculation. The new norm cost function C_{norm} takes the l_2 norm distance between the actual and the predicted centroids. This allows us to take the full k-dimensional distribution of the predicted labels for each time step and compute the final predicted value. However, Le et al. [20] used an argmax–based cost function C that calculates the final predicted value based on the highest probability for each task, which may ignore the full label distribution. For example, if we use the norm cost, the two different softmax predictions of $[0, 0.3, 0.6, 0]$ and $[0, 0, 0.6, 0.3]$ will result in different final predicted values, $0.3 \times -0.12 + 0.6 \times 0.05 = -0.006$ and $0.6 \times 0.05 + 0.3 \times 0.22 = 0.096$, respectively. However, if we use the argmax–based cost C, both of the predictions will have the same final predicted value of $1 \times 0.05 = 0.05$, and the subtle difference between the two predictions will be ignored. Hence, the norm cost C_{norm} is more sensitive to inherent label distribution than the argmax–based cost C. We empirically found that the use of full distribution (using C_{norm}) can achieve higher emotion prediction CCC than the use of only the maximum component (using C).

For the regression (*reg*) model, we use a D-BLSTM network trained with CCC loss function, $CCC_{loss} = 1 - CCC$, where CCC is defined by

$$CCC = \frac{2 * cov(y, \hat{y})}{var(y) + var(\hat{y}) + \left(\mathbb{E}[y] - \mathbb{E}[\hat{y}]\right)^2} \tag{11}$$

where $cov(y, \hat{y})$ is the covariance matrix and $var(y)$ is the variance. $\mathbb{E}[y]$ and $\mathbb{E}[\hat{y}]$ are the expected values of ground truth labels $y = [y_1, y_2, y_3, ..., y_n]$ and predicted labels $\hat{y} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_n]$ respectively, where n is the total number of frames that are evaluated. Previous studies have shown that using CCC as the loss function enhances the continuous emotion predictions compared to using the Root-Mean-Square Error (RMSE) loss [20, 33, 42]. The CCC loss has the advantage over RMSE that it takes into consideration the overall shape of the time series.

4.2 Joint Representation: Ensemble and End-to-End

As shown in Fig. 1, we explore two different methods for joint modeling of *clf* and *reg*: ensemble and end-to-end. These approaches differ from previous methods that formulated continuous emotion prediction as either an individual classification or a regression task. We hypothesize that joint modeling of the two tasks will find the optimal trade-off between the easiness and precision of the training and improve the overall prediction power of the model. We use the D-BLSTM model as a building block for both classification and regression tasks.

The ensemble model combines two D-BLSTM models: one trained for the classification task (*clf*) and another trained for the regression task (*reg*). The *clf* model has four target tasks for each of the four k-means models ($k = 4, 6, 8,$ and 10) (Fig. 1, left). We use a softmax activation function for each target task of the *clf* model. To compute the predicted centroid value for each task, we take the highest probability in the softmax layer outputs. Using the example in Section 4.1, where $K_{L_t} = [-0.31, -0.12, 0.05, 0.22]$ for $L_t = 4$, and $\hat{y}_{tf} = [0.1, 0.6, 0.2, .1]$, the highest probability in $\hat{y}_{tf}$ is 0.6 for the second centroid; hence the predicted centroid for this case will be −0.12. The final predicted value will be the average of the four predicted centroids from each task. The *reg* model is trained with continuous emotion labels (Section 4.1). The *clf* and *reg* models are trained separately, and they do not share any weights. For the ensemble model, the final predictions are generated by averaging both models' predictions. We use the CCE loss ℓ_{clf} (Eq. 9) for classification and the CCC loss (Section 4.1) for regression.

The end-to-end model is a single D-BLSTM model that is trained end-to-end for a joint classification and regression task (Fig. 1, right). We then add a final node to combine the predicted values for the five nodes (four classification nodes and one regression node) with learnable weights. We propose to train the end-to-end model using the total loss ℓ_{total}, which jointly minimizes classification (ℓ_{clf}), regression (ℓ_{reg}) and the total (ℓ_{out}) losses. As shown in Fig. 1, the end-to-end model combines three tasks —each task minimizes one of the three losses— in a hierarchical manner. The classification (ℓ_{clf}) and regression (ℓ_{reg}) takes are positioned at the same layer; and the combination (ℓ_{out}) task is positioned at the next layer to automatically learn the optimal weights between ℓ_{clf} and ℓ_{reg}. Similar to the *clf* model (Section 4.1), the classification task (ℓ_{clf}) includes a suite of four k-means clustering models ($k = 4, 6, 8$, and 10). These models are combined to predict discrete emotion labels based on a softmax function. We apply the weighted sum operation to compute the predicted value for each classification task, where we multiply each probability by its corresponding centroid value for each task. On the other hand, the regression task (ℓ_{reg}) uses a linear activation function to predict continuous emotion labels. We formalize the total loss ℓ_{total} as follows:

$$\ell_{total} = \alpha_1 \ell_{clf} + \alpha_2 \ell_{reg} + \alpha_3 \ell_{out} \tag{12}$$

$$\alpha_3 = 2 - \frac{1}{2}(\alpha_1 + \alpha_2) \tag{13}$$

In Eq. 12, α_1 and α_2 are the hyper-parameters that take values between 0 and 1, and they control the contributions of ℓ_{clf} and ℓ_{reg} losses to the total loss ℓ_{total}. Eq. 13 for α_3 is designed so that we can enforce a loss value for the final node that is always equal to or greater than α_1 and α_2 and take a value between 1 and 2. For example, when $\alpha_1 = \alpha_2 = 1$, this means both tasks have the same weight of contribution to the final loss and α_3 will be 1 as well. Similarly, $\alpha_1 = \alpha_2 = 0$ means that ℓ_{clf} and ℓ_{reg} losses are completely ignored and the model will try to come up with its own representation for these tasks. The variable α_3 will be 2 in this case.

We use the same loss function ℓ_{clf} (Eq. 9) for classification, $RMSE_{loss}$ defined below for regression (ℓ_{reg}), and the CCC loss for the final node (ℓ_{out}) defined in Section 4.1. The loss functions of the regression and final nodes are chosen empirically. The $RMSE_{loss}$ is defined as follows:

$$RMSE_{loss} = \sqrt{\frac{1}{F}\sum_{f=1}^{F}(l_f - \hat{l}_f)^2} \tag{14}$$

5 EXPERIMENTAL SETTINGS

For all of our models, we perform two-stage training over each D-BLSTM model, similar to Le et al. [20]. The two-stage training first trains a model for 20 epochs using an Adam optimizer with a base learning rate of 0.002 and a batch size of one utterance (7,500 frames). For the second stage, we start with the best CCC value that we obtained from the first stage based on the development set. After each epoch, if the CCC value on the development set decreases for more than 0.01, we half the learning rate. If a better CCC value than the previous best is found, the learning rate is reset to 0.002. This continues until we reach 40 epochs or the learning rate drops below 0.00001.

The number of D-BLSTM layers is cross-validated using the development CCC ({5, 7} for arousal, {3, 5} for valence) with a size of 160 hidden units (80 for forward and 80 for backward paths) for speech–based experiments (Sections 6.1–6.3). For multimodal experiments (Section 6.4), we cross-validate the number of D-BLSTM layers over {5, 7, 9} and the number of hidden units over {160, 320} (80 for each path and 160 for each path, respectively) to increase the model complexity to adapt to the increased feature dimensionality. We cross-validate α_1 and α_2 (0, 0.25, 0.50, and 1) using the development CCC. Note that α_1 and α_2 in Eq. 12 cannot be directly trained from the data, since these weights are not part of model parameters. We implement all of our experiments using TensorFlow [1].[1] Our models were trained on Xeon CPU clusters, where the five-layer D-BLSTM model took 1.5 minutes per epoch for training and 21 seconds for validation.

We empirically found that initializing the final layer weights with Xavier initialization [10] and the biases with zero initialization is crucial to achieve stable training. To reduce the randomness of training, we run every experiment three times and use the average of three models' predictions as in Le et al. [20].

To be consistent with previous work using RECOLA [2, 20, 44], we use the CCC [19] (Eq. 11) as a performance metric of emotion prediction.

6 RESULTS AND DISCUSSION

In this section, we first explore our proposed methods with audio features for the comparison with the state-of-the-art discretization-based continuous emotion prediction on RECOLA [20] (Sections 6.1–6.3) and then carry out multimodal experiments (Section 6.4). In Section 6.1.1, we compare our baselines *clf* and *reg* against the previous state-of-the-art models. In Section 6.1.2, we compare our proposed joint representation approaches against both the strong baselines and previous studies. In Section 6.2, we specifically investigate the end-to-end model when the classification and regression losses are ignored ($\alpha_1 = \alpha_2 = 0$). This will help us to further explore whether it is beneficial to explicitly model ℓ_{clf} and ℓ_{reg} along side with ℓ_{out} or reduce the model complexity and model only ℓ_{out}. In Section 6.3, we investigate the classifier convergence for arousal and valence with *clf* and *clf+reg* (end-to-end) models to compare the convergence rate performance in the first stage of training. In Section 6.4, we compare the joint modeling and the baseline models in a multimodal (audio-visual) environment.

6.1 Performance Comparison

Table 2 shows CCC differences on the development (dev) and test (test) sets when using our proposed methods, baseline models, and previous state-of-the-art speech emotion recognition work. *clf* is

[1] https://gitlab.rit.albany.edu/kimlab/JointModeling

Model	**Arousal**		**Valence**	
	dev	test	dev	test
Valstar et al. [44]	0.796	0.648	0.455	0.375
Brady et al. [2]	0.846	-	0.450	-
Le et al. [20] (CLS-Raw)	0.858	0.682	0.563	0.448
Le et al. [20] (CLS-Decoded)	0.859	0.680	0.596	0.460
Khorram et al. [18]	0.867	0.684	0.592	0.502
clf [baseline]	0.860	0.686	0.558	0.527
reg [baseline]	0.863	0.686	0.595	0.544
clf+reg (ensemble)	**0.868**	0.693	0.601	**0.555**
clf+reg (end-to-end)	**0.868**	**0.697**	**0.623**	0.530

Table 2: CCC differences between previous state-of-the-art work, baselines (*clf* and *reg*), and our two proposed joint models for speech–based experiments. Dev: development set, test: test set.

Model			**Arousal**		**Valence**	
			dev	test	dev	test
	α_1	α_2				
clf+reg (end-to-end)	1	1	**0.868**	**0.697**	0.583	-
	0	0	0.854	-	0.608	-
	0.50	1	0.840	-	0.557	-
	0.25	1	0.864	-	0.591	-
	1	0.50	0.867	-	0.549	-
	1	0.25	0.858	-	**0.623**	0.530
reg (BLSTM-FC)			0.867	0.692	0.613	0.516
clf+reg (BLSTM-FC) (ensemble)			**0.869**	0.694	0.604	**0.538**

Table 3: CCC differences between the end-to-end model (with multiple α_1 and α_2) and BLSTM-FC models for speech–based experiments. Dev: development set, test: test set.

the classification model presented in Section 4.1, *reg* is the regression model presented in Section 4.1, *clf+reg* (ensemble) is the ensemble model presented in Section 4.2, and *clf+reg* (end-to-end) is the D-BLSTM model trained with classification and regression tasks end-to-end as presented in Section 4.2.

6.1.1 Individual Modeling Methods. Our proposed baseline models (*clf* and *reg*) achieve competitive results compared to previous state-of-the-art results, which indicates that the baselines we use to evaluate our work are strong (Table 2). Both *clf* and *reg* outperform the AVEC 2016 baseline model [44] by a large margin. For arousal, *clf* achieves 0.86 and 0.686 on the development and test sets respectively, and *reg* achieves 0.863 and 0.868. The Valstar et al. [44] baseline model achieved 0.796 on the development set and 0.648 on the test set. For valence, *clf* achieves 0.558 on the development set and 0.527 on the test set, and *reg* achieves 0.595 and 0.544, in comparison to 0.455 and 0.375 from Valstar et al. [44]. *clf* and *reg* also outperform the AVEC 2016 challenge winner, Brady et al. [2], which achieved 0.846 on arousal and 0.450 on valence for the development set.

We also found that using our proposed norm cost (C_{norm}) improves the classification task results on arousal (0.860 for the development set and 0.686 for the test set) compared to the original cost function computation (C) introduced by Le et al. [20]. Le et al. [20] reported 0.858 CCC value on the development set and 0.682 on the test set. For valence, *clf* with C_{norm} also outperforms the original classification model introduced by Le et al. [20] by a large margin (0.079) on the test set: *clf* with *norm cost* achieves 0.527 CCC value, which is higher the 0.448 from Le et al. [20].

6.1.2 Joint Modeling Methods. Joint modeling (ensemble and end-to-end) of discrete and continuous labels improves the overall CCC for both arousal and valence compared to our strong baselines (*clf* and *reg*) and the previous state-of-the-art models. As shown in Table 2, the end-to-end model shows the highest performance on arousal prediction among all models, achieving 0.868 for the development set and 0.697 for the test set. These results are higher than *clf* (0.860 and 0.686 on the development and test set, respectively) and *reg* (0.863 and 0.686) baselines. Our proposed joint modeling outperforms Valstar et al. [44] by a large margin (0.796 and 0.648 on the development set and test sets, respectively) and Brady et al. [2] (0.846 on the development set).

For valence prediction, the ensemble model performs better in terms of the test CCC, achieving 0.601 on the development set and 0.555 on the test set compared to 0.623 and 0.530 for the end-to-end model. These results are higher than *clf* (0.558 and 0.527 on the development and test set, respectively) and *reg* (0.595 and 0.544) baseline models. The end-to-end model showed a decrease in CCC (0.530) compared to *reg* (0.544) on the test set, which may be due

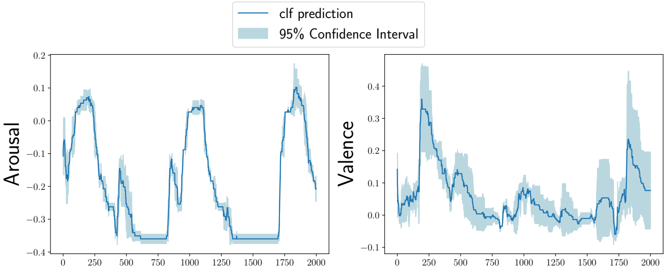

Figure 2: *clf* prediction and the 95% confidence interval for both arousal (top) and valence (bottom) on a given speaker from the development set while using the *clf* model.

to an inherent limitation of end-to-end approaches. For instance, previous research has shown that end-to-end models can be very inefficient when they are used for training neural networks that consist of multiple challenging tasks [9] (e.g., *clf* and *reg* tasks for valence prediction). Compared to previous state-of-the-art results, Valstar et al. [44] achieved 0.455 and 0.375 CCC value and Brady et al. [2] achieved 0.450 on the development set.

We also achieve the best performance compared to the best model proposed by Le et al. [20] (0.859 on the development set and 0.680 on the test set for arousal and 0.596 and 0.460 for valence). We found that both of our proposed joint models (ensemble and end-to-end) achieve new state-of-the-art results on arousal (0.868) on the development set where we use only the audio features. By comparison, Brady et al. [2] received 0.862 when they used audio, video, and physiological data. Our proposed models have the additional advantage that they achieve comparable accuracy without any post-processing step.

6.2 End-to-End vs. BLSTM-FC

In this section, we first explore the special case in the end-to-end model where both classification (ℓ_{clf}) and regression (ℓ_{reg}) tasks' contributions to the final loss are ignored ($\alpha_1 = \alpha_2 = 0$). This will help us to investigate the benefit of explicitly modeling ground truth labels for classification and regression tasks compared to replacing this mid-level layer (with ℓ_{clf} and ℓ_{reg}) with a fully connected (FC) layer without any explicit modeling. We call this model BLSTM-FC. Second, we investigate the different combinations of α_1 and α_2 that control the contribution of the classification and regression tasks for the end-to-end model. Third, we explore the ensemble model when BLSTM-FC is used as the regression model instead of our *reg* model.

First, we explore the special case in the end-to-end model when the *clf* and *reg* tasks are ignored. The end-to-end model is designed to explicitly model the discrete (ℓ_{clf}) and continuous (ℓ_{reg}) ground truths in its final hidden layer (Fig. 1). In addition, the final output layer computes the emotion prediction loss, ℓ_{out}. The total loss function is the sum of the three losses (Eq. 12), and this model uses supervised learning. However, when both α_1 and α_2 (in Eq. 12) are zero, ℓ_{clf} and ℓ_{reg} losses will not affect the total loss. Therefore, the representation of the hidden layer does not explicitly model the ground truths. Since softmax activation restricts the hidden layer outputs to have a probability distribution that sums up to 1, this can increase the model complexity. To address this complexity issue, we replace the mid-layer with a FC layer that has 16 units. The FC and the final node are calculated as follows:

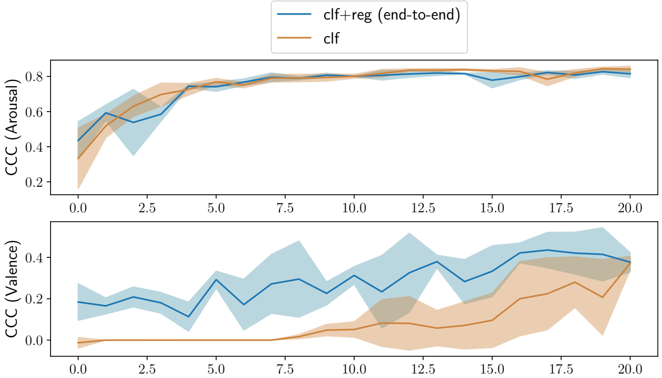

Figure 3: Learning curves for the convergence rate comparison between the *clf* and *clf+reg* (end-to-end) models, for arousal (top) and valence (bottom) prediction. The figure is best shown in color.

$$FC_\tau = tanh(W_{FC}h_\tau + b_{FC}) \tag{15}$$

$$\hat{l}_\tau = W_{out}FC_\tau + b_{out} \tag{16}$$

where FC_τ is the FC layer output at time τ, $tanh$ is the tanh activation function, and $\hat{l}_\tau$ is the predicted value for the time frame τ. We use the *CCC* loss function that is defined in Section 4.1. The comparison between the end-to-end model with $\alpha_1 = \alpha_2 = 0$ and the BLSTM-FC model will demonstrate whether it is more beneficial to keep the model architecture the same or to replace the layer before the final node with an FC layer, as in the BLSTM-FC model. As shown in Table 3, we found that it is better to use an FC layer instead of the softmax and linear activations. BLSTM-FC achieves 0.867 for arousal and 0.613 for valence on the development set, compared to 0.854 and 0.608 for arousal and valence respectively for *clf+reg* (end-to-end).

Second, by cross-validating α_1 and α_2 on the development set (Table 3), we found that $\alpha_1 = \alpha_2 = 1$ works the best for arousal (0.868 on the development set and 0.697 on the test set), and $\alpha_1 = 1$, $\alpha_2 = 0.25$ for valence (0.623 and 0.530). For this specific choice of α_1 and $\alpha_2 = 1$, *clf+reg* (end-to-end) performs slightly better than BLSTM-FC (0.867 on the development set and 0.692 on the test set for arousal, 0.613 and 0.516 for valence). For other combinations of α_1 and α_2, there is no consistent trend found. For instance, when α_1 is greater than α_2 and more weight is given to the classification loss than regression, there is no consistent performance. In the case where $\alpha_1 = 1$ and $\alpha_2 = 0.5$, we achieve higher CCC on arousal (0.867) compared to 0.840 when $\alpha_1 = 0.5$ and $\alpha_2 = 0.1$. However, for $\alpha_1 = 1$ and $\alpha_2 = 0.25$ we achieve lower CCC (0.858) compared to 0.864 when $\alpha_1 = 0.25$ and $\alpha_2 = 0.1$ for arousal.

Based on BLSTM-FC development set results, we found that it is better to perform the regression task for both arousal and valence with the additional FC layer before the final node. BLSTM-FC achieves a CCC value of 0.867 compared to 0.863 from the *reg* model for arousal, and 0.613 compared to 0.595 for valence. By replacing *reg* with BLSTM-FC in the *clf+reg* (ensemble) model, CCC improves from 0.868 to 0.869 for arousal and from 0.601 to 0.604 for valence, achieving the highest CCC for arousal.

6.3 Convergence Rate Performance Analysis

Figure 2 shows a visualization of a 95% confidence interval of the *clf* model. It demonstrates that the 95% confidence interval for valence has a large margin over three different runs. This is because valence highly depends on video features, and it is not easy to predict using only audio features [2]. For arousal, *clf* is more stable over the three runs, especially for the regions where there is no transition in the emotion state.

To investigate the large margin in confidence interval for valence, we further explore the learning curve of *clf* and *clf+reg* (end-to-end) models (Fig. 3). The learning curve analysis can provide insight into the convergence rate for *clf* and *clf+reg* models. The learning curves demonstrate that *clf+reg* (end-to-end) performs more efficiently in terms of valence convergence than the *clf* model. We found that for valence, the learning curve for the *clf+reg* (end-to-end) model has an accelerated convergence speed. For *clf+reg* (end-to-end), the model starts with CCC around 0.2 for the development set and increases over the first 20 epochs. However, for the *clf* model, the training starts with CCC around zero and does not increase until it reaches the seventh epoch. For arousal, on the other hand, both *clf* and *clf+reg* (end-to-end) models have similar learning curves for the first 20 epochs.

6.4 Joint Modeling in a Multimodal Environment

Model	Arousal		Valence	
	dev	test	dev	test
clf [baseline]	0.864	0.690	0.698	0.622
reg [baseline]	0.863	0.699	0.681	0.583
clf+reg (ensemble)	**0.869**	0.699	**0.705**	0.617

Table 4: CCC differences between baselines and proposed joint model (ensemble) using multimodal (audio-visual) features. Dev: development set, Test: test set.

Table 4 demonstrates how our best joint modeling method performs in multimodal (audio-visual) experiments. For the development set, the highest CCC is achieved when a joint model (ensemble) is used for both arousal and valence prediction. For arousal, *clf* and *reg* achieve 0.864 and 0.863, whereas the *clf+reg* (ensemble) achieves 0.869. For valence, *clf* and *reg* achieve 0.698 and 0.681, whereas the *clf+reg* (ensemble) achieves 0.705. These results are higher than the audio–based performance in Table 2, demonstrating the importance of using multimodal features in predicting emotion. The improvement compared to unimodal systems shows a larger margin for valence, where valence CCC improves from 0.555 to 0.705 (27.02% improvement). This finding is consistent with previous work [2] that demonstrated that visual features contribute more to valence prediction. Both arousal and valence results are higher than the RECOLA multimodal baseline [44]. These results demonstrate that the joint representation of discrete and continuous emotion helps the overall prediction compared to individual classification or regression models.

On the other hand, for the test set, the ensemble model achieves higher CCC than *clf* for arousal (0.699 versus 0.690, respectively); however, the CCC remains the same compared to *reg*. Also, the valence prediction achieves the highest CCC for *clf* (0.622), the second-highest for *clf+reg* (ensemble, 0.617), and finally *reg* (0.583). We assume that the different performance trends shown between the test and development sets may indicate the difficulty of valence prediction in the RECOLA dataset. Indeed, previous multimodal systems have shown a relatively higher performance for arousal than valence [2, 44].

7 CONCLUSIONS

In this paper, we propose joint modeling methods that combine discrete (classification, *clf*) and continuous (regression, *reg*) emotion representations. Recent studies have found improvement in continuous emotion prediction performance when noisy continuous labels are quantized and these quantized, discrete labels are used for training classification models. However, although effective, the previous quantization approach may introduce a quantization error when converting the continuous labels to discrete ones. To overcome this challenge and find the optimal trade-off between discretized and continuous emotion representations, we introduce two joint modeling methods, ensemble and end-to-end. The ensemble method combines the outputs from the *clf* and *reg* models at the prediction level. The end-to-end model simultaneously optimizes *clf*, *reg*, and the combined tasks of the two in an end-to-end manner. Our models are based on the state-of-the-art D-BLSTM architectures.

The results demonstrate that our proposed joint modeling approaches can predict continuous emotion labels more accurately than previous approaches, especially for valence prediction. Our results also indicate that joint modeling increases the emotion prediction performance compared to individual modeling baselines for both audio and audio-visual experiments. By further investigating the learning curves, we found that joint modeling has faster convergence in comparison to individual modeling for valence prediction. The results provide insights into the new joint representation of continuous and discrete emotion. Our findings can further the design and development of interactive systems that require the accurate description of dynamically changing affective phenomena.

In future work, we will further explore the effect of α_1 and α_2 in Eq. 12 and seek to understand in greater detail how the different combinations help to control for the contribution of both classification and regression tasks to the final loss. We will also study the effect of delayed annotations with RECOLA, which was studied by Mariooryad et al. [25]. They showed that a shifted delay in continuous-time annotations improves the final emotion prediction performance, achieving over 7% relative improvement in accuracy on the SEMAINE database [26]. In addition, we will explore the use of CNNs to extract more generic features for video and/or audio signals in the proposed D-BLSTM–based joint modeling networks. We will further investigate the case of jointly training arousal and valence as two related tasks in a multi-task learning approach. The two tasks will share parts of the model topology.

8 ACKNOWLEDGEMENTS

This research was supported by a Google Faculty Research Award and a University at Albany Faculty Research Award.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning.. In *OSDI*, Vol. 16. 265–283.

[2] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang. 2016. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 97–104.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[4] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 65–72.

[5] Ira Cohen, Ashutosh Garg, Thomas S Huang, et al. 2000. Emotion recognition from facial expressions using multilevel HMM. In *Neural information processing systems*, Vol. 2. Citeseer.

[6] Li Deng and John C Platt. 2014. Ensemble deep learning for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[7] Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2014), 845–869.

[8] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation Learning for Speech Emotion Recognition.. In *INTERSPEECH*. 3603–3607.

[9] Tobias Glasmachers. 2017. Limits of end-to-end learning. *arXiv preprint arXiv:1704.08305* (2017).

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

[11] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[12] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 890–897.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. 2015. 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*. ACM, 73–80.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 801–804.

[17] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* 65 (2017), 66–75.

[18] Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost. 2017. Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *arXiv preprint arXiv:1708.07050* (2017).

[19] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[20] Duc Le, Zakaria Aldeneh, and Emily Mower Provost. 2017. Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. *Interspeech, 2017 (to apear)* (2017).

[21] Gil Levi and Tal Hassner. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 503–510.

[22] Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo. 2014. iVectors for continuous emotion recognition. *Training* 45 (2014), 50.

[23] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 35–42.

[24] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. 2014. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia* 16, 8 (2014), 2203–2213.

[25] Soroosh Mariooryad and Carlos Busso. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing* 6, 2 (2015), 97–108.

[26] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.

[27] Hongying Meng and Nadia Bianchi-Berthouze. 2014. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics* 44, 3 (2014), 315–328.

[28] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. 2012. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing* 3, 2 (2012), 184–198.

[29] Donn Morrison, Ruili Wang, and Liyanage C De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech communication* 49, 2 (2007), 98–112.

[30] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech communication* 41, 4 (2003), 603–623.

[31] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[32] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[33] Filip Povolny, Pavel MatËĞejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. 2016. Multimodal emotion recognition for AVEC 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 75–82.

[34] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. 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*. ACM, 3–8.

[35] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013*. IEEE, 1–8.

[36] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. 2012. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing* 3, 2 (2012), 165–183.

[37] Marc Schröder, Sathish Pammi, Hatice Gunes, Maja Pantic, Michel F Valstar, Roddy Cowie, Gary McKeown, Dirk Heylen, Mark Ter Maat, Florian Eyben, et al. 2011. Come and have an emotional workout with sensitive artificial listeners!. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011*. IEEE, 646–646.

[38] Björn Schuller, Stephan Reiter, Ronald Muller, Marc Al-Hames, Manfred Lang, and Gerhard Rigoll. 2005. Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*. IEEE, 864–867.

[39] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *International Conference on Multimedia and Expo, 2003. ICME'03. Proceedings. 2003.*, Vol. 1. IEEE, I–401.

[40] Mohammad Soleymani, Sander Koelstra, Ioannis Patras, and Thierry Pun. 2011. Continuous emotion detection in response to music videos. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011*. IEEE, 803–808.

[41] Yang Sun, Louis Ten Bosch, and Lou Boves. 2010. Hybrid HMM/BLSTM-RNN for robust speech recognition. In *International Conference on Text, Speech and Dialogue*. Springer, 400–407.

[42] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016*. IEEE, 5200–5204.

[43] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[44] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

[45] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST*. Brisbane, Australia, 597–600.

[46] Martin Wöllmer, Angeliki Metallinou, Nassos Katsamanis, Björn Schuller, and Shrikanth Narayanan. 2012. Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.* IEEE, 4157–4160.

[47] Martin Wollmer, Björn Schuller, Florian Eyben, and Gerhard Rigoll. 2010. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing* 4, 5 (2010), 867–881.

[48] Xinzhou Xu, Jun Deng, Maryna Gavryukova, Zixing Zhang, Li Zhao, and Björn Schuller. 2016. Multiscale kernel locally penalised discriminant analysis exemplified by emotion recognition in speech. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction.* ACM, 233–237.

[49] Yi-Hsuan Yang and Homer H Chen. 2011. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 7 (2011), 2184–2196.

[50] Biqiao Zhang, Georg Essl, and Emily Mower Provost. 2017. Predicting the distribution of emotion perception: capturing inter-rater variability. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* ACM, 51–59.

[51] Biqiao Zhang, Emily Mower Provost, Robert Swedberg, and Georg Essl. 2015. Predicting Emotion Perception Across Domains: A Study of Singing and Speaking.. In *AAAI*. 1328–1335.

[52] Shiliang Zhang, Qi Tian, Shuqiang Jiang, Qingming Huang, and Wen Gao. 2008. Affective MTV analysis based on arousal and valence features. In *IEEE International Conference on Multimedia and Expo, 2008.* IEEE, 1369–1372.