Predicting the Distribution of Emotion Perception: Capturing Inter-rater Variability

Biqiao Zhang
University of Michigan
Ann Arbor, Michigan, USA
didizbq@umich.edu

Georg Essl
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin, USA
essl@uwm.edu

Emily Mower Provost
University of Michigan
Ann Arbor, Michigan, USA
emilykmp@umich.edu

ABSTRACT

Emotion perception is person-dependent and variable. Dimensional characterizations of emotion can capture this variability by describing emotion in terms of its properties (e.g., valence, positive vs. negative, and activation, calm vs. excited). However, in many emotion recognition systems, this variability is often considered "noise" and is attenuated by averaging across raters. Yet, inter-rater variability provides information about the subtlety or clarity of an emotional expression and can be used to describe complex emotions. In this paper, we investigate methods that can effectively capture the variability across evaluators by predicting emotion perception as a discrete probability distribution in the valence-activation space. We propose: (1) a label processing method that can generate two-dimensional discrete probability distributions of emotion from a limited number of ordinal labels; (2) a new approach that predicts the generated probabilistic distributions using dynamic audio-visual features and Convolutional Neural Networks (CNNs). Our experimental results on the MSP-IMPROV corpus suggest that the proposed approach is more effective than the conventional Support Vector Regressions (SVRs) approach with utterance-level statistical features, and that feature-level fusion of the audio and video modalities outperforms decision-level fusion. The proposed CNN model predominantly improves the prediction accuracy for the valence dimension and brings a consistent performance improvement over data recorded from natural interactions. The results demonstrate the effectiveness of generating emotion distributions from limited number of labels and predicting the distribution using dynamic features and neural networks.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; **Neural networks**;

KEYWORDS

Automatic Emotion Recognition; Inter-rater Variability; Convolutional Neural Networks

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136792

ACM Reference Format:
Biqiao Zhang, Georg Essl, and Emily Mower Provost. 2017. Predicting the Distribution of Emotion Perception: Capturing Inter-rater Variability. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3136755.3136792

1 INTRODUCTION

Emotions are not perceived uniformly across individuals. In emotion recognition experiments, inter-rater variability is often mitigated by averaging the ratings of groups of evaluators, under the assumption that this amalgamation can remove perceptual "noise". However, inter-rater variability contains signal, in addition to noise. It provides information about the subtlety or clarity of the an emotional display. In this paper, we investigate methods that can effectively capture and predict the variation that is present in a population of evaluators.

We focus on dimensional descriptions of emotion, which characterize emotion in terms of continuous values. This characterization naturally captures variation in emotion perception [38], allowing us to retain rich information about the emotional content of a given expression. This information could provide high-level features for tasks such as the prediction of mental health (e.g., depression, autism), complex emotions, and cross-corpus / cross-language emotion recognition where categorical labels may have different meanings as a function of context [24].

Among the various definitions of the dimensions, valence (positive vs. negative) and activation (calm vs. excited) are the most commonly accepted [7, 23]. Compared to categorical emotion descriptions (e.g., anger, happiness, and sadness), these dimensions are less dependent on context or language [24] and therefore more commonly used in cross-corpus / cross-language emotion recognition [28, 29, 41, 43].

Previous work in music emotion recognition has explored methods for generating and predicting distributions of emotion perception [25, 26, 34, 35, 38, 42]. However, they often require a large number of real-valued evaluations, focus on a single modality (i.e., audio), and do not fully exploit short-term temporal information. In speech emotion recognition, there has been work incorporating inter-rater consistency into systems using categorical labels [31, 37]. However, most work using dimensional labels focuses on predicting either the mean evaluation across multiple evaluators [10, 13, 17, 19, 21, 36] or the classes caterogized from the mean evaluation [18, 28]. Works that seek to provide emotion variation as a usable and modelable signal for speech are still missing.

In this paper, we present a new approach that generates probabilisty distributions on the valence-activation space and captures

the variability of emotion perception from speech, using a limited number of ordinal evaluations. We demonstrate how these two-dimensional distributions can be predicted using frame-level audio-visual features. We then ask the following two research questions: (1) can we predict a probability distribution more accurately by modeling local temporal patterns; and (2) can we combine audio and video modalities to result in better performance compared to when a single modality is used?

We conduct experiments on the MSP-IMPROV dataset [5]. We upsample the evaluations for each utterance by repeatedly performing random subsampling and averaging. We use the resulting set of evaluations to calculate ground-truth probability distributions. We use convolutional neural networks (CNNs) to predict these distributions by leveraging regional temporal patterns for both unimodal and multimodal input. We compare the proposed CNN approach with support vector regression (SVR), the state-of-the-art approach [38, 42], to answer our first question, and compare the performance of different modalities and fusion methods to answer our second.

Our experimental results suggest that modeling local temporal patterns is beneficial with respect to both Total Variation and Jensen-Shannon Divergence compared to SVR with utterance-level statistical features. Combining audio and video modalities at the feature-level outperforms approaches that either use a single modality or combine the modalities at the decision-level. The proposed CNN model predominantly improves the prediction of valence. The novelty of this work includes: (1) a label processing method for generating two-dimensional probability distribution from scarce ordinal labels; (2) the first attempt to predict two-dimensional probability distributions of emotion perception for speech using a dynamic approach; (3) an exploration of the influence of modality on predicting the distribution of emotion perception.

2 RELATED WORKS

2.1 Emotion Recognition using Dimensional Labels

Dimensional descriptions of emotion have become increasingly common in emotion recognition research. The majority either predict the mean of a group of evaluations [10, 13, 17, 19, 21, 36] or the mean weighted by rater-reliability [12], or restructure the emotion recognition problem as classification along each dimension [18, 28, 29, 41, 43]. While it is common practice for emotion datasets to collect multiple evaluations [4, 8, 22], neither approach models the variability in emotion perception captured by these dimensional evaluations. This is compounded by the low inter-rater agreement common in these datasets [30].

There has been work in music emotion recognition and cross-domain (song and speech) emotion recognition on predicting probability distributions on the valence-activation space. There are two popular approaches: parametric (e.g., bivariate Gaussian and GMMs) [25, 34, 35] or non-parametric (discrete grid representation) [26, 38, 42]. In general, both approaches rely upon a large number of real-valued annotations.

Schmidt et al. first proposed to model emotion perception from music as a probability distribution [25]. They assumed that the individual evaluations could be represented by a bivariate Gaussian. They formulated the task as a prediction of the Gaussian parameter associated with each short clip using several regression methods. They found that support vector regression (SVR) produced the best single-feature performance. However, the underlying assumption that the evaluations are guaranteed to follow a Gaussian distribution may not be valid, as noted in [35, 38]. Wang et al. proposed a generative model that learns two Gaussian mixture models (GMMs), one from acoustic features and the other from emotion labels [34, 35]. They predicted the emotional content of music as a probability distribution over the affective GMM components and summarized the prediction as a single Gaussian. However, while this approach used frame-level features directly, the utterance-level predictions were calculated by averaging the frame-level labels over the entirety of the utterance. The interactions across consecutive frames were not considered.

Another work of Schmidt et al. represented emotion perception as a probability heatmap [26]. The evaluations were discretized into equally spaced grids. No assumption of the distribution of labels was made. They predicted the heatmaps over 1-second periods using Conditional Random Fields (CRF). The acoustic features were averaged over the 1-second window to reduce the frame-rate to that of the labels. Therefore, while CRF is context-dependent, there was information loss in the feature downsampling process. Yang et al. generated a smooth probability density function from individual evaluations using Kernel Density Estimation (KDE) and then discretized the space [38]. The benefit is that the distribution is not biased by the position of the binning grids. They predicted the probability in each grid separately using SVR with utterance-level statistic features. Our previous work used a similar approach for predicting emotion perception across song and speech [42]. The difference is that we performed evaluator-dependent z-normalization to smooth the ordinal labels. This was valid because the evaluators were presented with relatively balanced data. However, both works used a static approach.

We note that these works often rely on a large number of real-valued labels, which are not available in most popular emotion datasets. Besides, they mostly focus on using the audio modality. In addition, either feature downsampling or ignoring interactions across frames will result in information loss. The short-time temporal information is not fully exploited. Therefore, we propose the following improvements: (1) developing a method to make use of a limited number of ordinal labels, (2) capturing emotionally salient temporal patterns using dynamic modeling, and (3) investigating the impact of modality. We posit that the combination of these three approaches will lead to a system with better performance.

2.2 CNN for Modeling Temporal Patterns in Emotion Recognition

CNNs have been used in affective computing to learn emotionally salient features from audio [14, 16, 32] and video [11, 39]. Recent works have explored the efficacy of CNNs for modeling temporal patterns. Mao et al. extracted emotion-salient features for speech emotion recognition using CNNs [16]. They first used a sparse auto-encoder to learn filters at different scales from unlabeled speech signals and convolved the spectrogram segments with learned filters to form a series of feature maps. They performed mean-pooling and

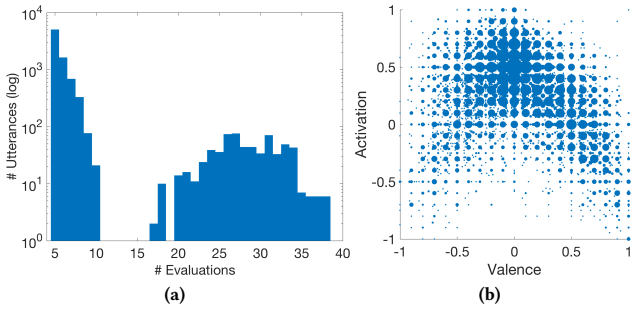

Figure 1: Dataset details about MSP-Improv: (a) number of evaluations per utterance (in log scale); (b) average valence-activation per utterance (size of dot proportional to the number of utterances).

stacked the feature maps into a feature vector. After that, they used the feature vector as the input to a fully-connected layer to learn emotion-salient features before feeding the learned features into a Support Vector Machine (SVM) classifier. Aldeneh et al. used a CNN with a convolutional layer, a global max-pooling layer, and several dense layers to identify emotionally salient local patterns and classify emotion from temporal low-level acoustic features [1]. They obtained comparable results to the state of the art utterance-level statistic features and SVMs. Khorram et al. used dilated CNN and downsampling-upsampling CNNs for predicting time-continuous valence and activation labels [15]. Their methods outperformed BLSTMs and were 46 times faster. These works support that CNNs can be used to model temporal patterns.

3 DATA

We experiment on MSP-Improv, an audio-visual dyadic emotion corpus [5]. We choose MSP-Improv because: (1) the available modalities and size of the dataset allow us to train multimodal models and (2) the crowdsourcing evaluation method and the number of evaluations capture variations in emotion perception.

MSP-Improv consists of six sessions, each including interactions between a male and a female actor. This results in 12 speakers in total. The emotional expressions of the speakers were elicited through carefully designed scenarios that include improvisations and target sentences with specific lexical content. Because of the recording paradigm of MSP-Improv, there are four types of recordings in the database: (1) the target sentences read by the actors; (2) the target sentences from the improvised scenes collected using emotionally evocative scenarios; (3) the speaker turns in the improvised scenes; (4) the natural spontaneous interactions during breaks between improvisations. The database includes over nine hours of data, segmented into 8,438 *utterances* (i.e., speaker turns or target sentences). The numbers of utterances corresponding to the four types of recordings are 620, 652, 4,381, and 2,785, respectively.

The emotional content of MSP-Improv was evaluated using crowdsourcing (Amazon Mechanical Turk). A scheme was designed to ensure the reliability of the labels by stopping evaluators when their inter-rater agreement with known "gold-standard" evaluations dropped [5]. Each utterance was annotated by at least five evaluators using both dimensional and categorical rating paradigms. For the dimensional labels, the evaluators were required to access the valence, activation, dominance (dominant vs. submissive) and naturalness (acted vs. natural) of the utterances using a five-point Likert-scale. In this paper, we focus only on the dimensional labels of valence and activation. We rescale the evaluations to [-1, 1] from [1, 5].

We show the distribution of the number of evaluations per utterance in Figure 1a. The majority of the utterances were annotated by less than ten evaluators, yet a portion of the database has approximately 30 evaluations. Figure 1b illustrates the distribution of the mean evaluations of each utterance on the valence-activation space. The database is relatively balanced along the valence dimension but skewed towards positive for the activation dimension.

4 METHODOLOGY

4.1 Label Processing

Dimensional annotations are often collected using evaluations over m-point Likert-scales [4, 5, 42]. Previous work has approximated the distribution over evaluations using Kernel Density Estimation (KDE) either from original continuous labels [38] or after applying evaluator-dependent z-normalization [42]. KDE assigns a two-dimensional Gaussian "energy" to each evaluation. The probability density of any point in the valence-activation space can be calculated by summing over the "energy" emitted by all the evaluations. In this work, we adopt the same method because of its ability to generate smooth probability density distributions. However, there are a few challenges that we need to address first:

(1) The majority of the utterances have less than 10 evaluations, which may not be sufficient to conduct KDE.
(2) The dimensional labels are ordinal instead of continuous. Therefore, we cannot apply KDE directly.
(3) It is not guaranteed that each evaluator was given utterances with balanced emotional content. As a result, we cannot use evaluator-dependent z-normalization as in [10, 36, 42].

We argue that the mean of any subset of evaluations of each utterance can be considered a potential ground-truth label of that utterance, inspired by the fact that researchers often use the mean of evaluations as the ground truth, and that the number of evaluations varies within and across datasets. Therefore, for a given utterance, we randomly subsample from one to N evaluations, where N is the total number of evaluations for that utterance, and use the mean as a new annotation. We repeat the process 200 times for each utterance. We add random noise to each generated annotation to avoid the same value being repeated multiple times. The random noise follows a uniform distribution centered at zero, with the width and height corresponding to half of the standard deviation of the valence and activation for the given utterance, respectively. The generated annotations share similar statistical properties with the original evaluations. On a -1 to 1 scale, the mean absolute difference across all utterances between the mean of original labels and the mean of generated labels are 0.011 and 0.015 for valence and activation, respectively. The correlations between the per-utterance standard deviation of the original labels and generated labels across all utterances are 0.96 for valence and 0.95 for activation. We show an example of the original labels and the corresponding generated labels in Figure 2a-2b, respectively.

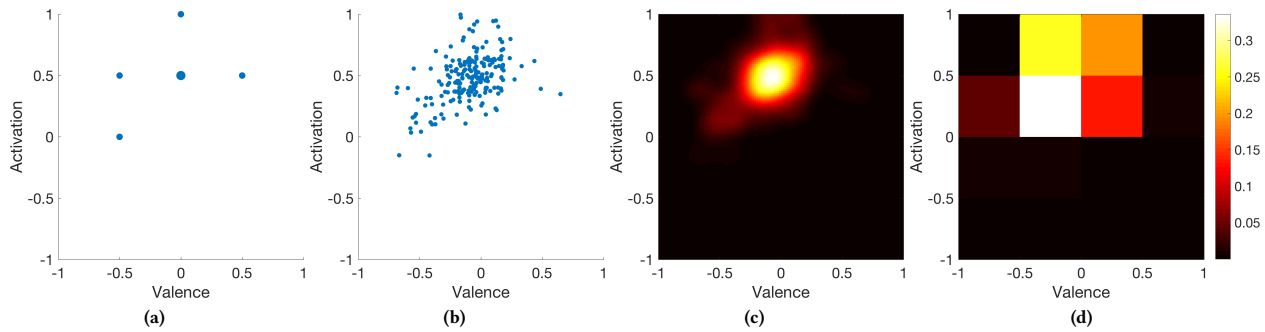

Figure 2: The process of generating the two-dimensional discrete probability distributions: (a) individual evaluations (size of dot proportional to the number of evaluations); (b) annotation cloud generated by averaging subsample of evaluations and adding random noise; (c) probability density distribution calculated by KDE; (4) discretized probability distribution at 4×4 resolution.

Table 1: Action Unit Features Used in Experiments.

AU	Description	AU	Description
1	Inner Brow Raiser	2	Outer Brow Raiser
4	Brow Lowerer	5	Upper Lid Raiser
6	Cheek Raiser	7	Lid Tightener
9	Nose Wrinkler	10	Upper Lip Raiser
12	Lip Corner Puller	14	Dimpler
15	Lip Corner Depressor	17	Chin Raiser
20	Lip Stretcher	23	Lip Tightener
25	Lips Part	26	Jaw Drop
45	Blink		

We then perform KDE using the approach from [3]. Since predicting a continuous function is both challenging and unnecessary, we transform the density function to a discrete probability distribution by creating equally spaced partitions along both valence and activation. Note that we create partitions from a smoothed density distribution instead of from individual labels directly, because the latter approach highly depends on the position of the partitions and can lead to biasing. For example, if the annotations are far apart and we want to predict at a higher resolution, we may end up with grids with high probability surrounding grids with zero probability. We use the mean of the density values within a grid to represent this grid. The values of all the grids are then normalized to sum to one. We show the density function from KDE and the discrete probability distribution in Figure 2c-2d, respectively.

4.2 Feature Extraction

4.2.1 Acoustic Features. We use 40 log Mel-frequency filterbank features (MFBs) for the audio modality, as in [1]. We first trim the silence at the beginning and end of each utterance and then extract the MFBs from each frame with a window size of 25ms and a step size of 10ms using Kaldi [20].

4.2.2 Visual Features. We use the intensity of facial action units (AUs) for the video modality. The AUs, which are the contraction or relaxation of single or multiple facial muscles, stem from the Facial Action Coding System (FACS) proposed by Ekman et al. [9]. Using FACS, common facial expressions can be deconstructed into the specific Action Units (AU) that produced the expression. We choose to use AUs to represent the video modality because of the close relationship between facial expression and emotion.

We extract the intensity of 17 AUs (Table 1) using OpenFace [2], which provides a intra-class correlation coefficient of approximately 0.6 on the test set of the 2015 Facial Expression Recognition and Analysis challenge [33]. We use the "static" prediction model, which relies on a single frame to estimate the intensity of the AUs at each time step. This is because some videos have a limited dynamic range, thus using the dynamic model that attempts to perform pose calibration may be harmful [2].

4.3 Model

We ask two main research questions: (1) can we better predict the distribution of emotion perception by focusing on salient local regions and jointly optimizing the predictions across the grids; (2) can we understand the influence of modality?

We answer the first question by comparing a static regression approach from [38] and our approach that takes regional temporal patterns into account. We answer the second question by building four models for both approaches: two unimodal models (audio modality and video modality), a model combining the two modalities at decision-level by averaging (denoted as combined-late), and a model combining the two modalities at feature-level (denoted as combined-early).

Past research has found that ν-Support Vector Regression (SVR) with a Gaussian kernel is effective for predicting the discrete probability distribution of emotion [38, 42]. We use this approach with the same implementation (LibSVM [6]) as our baseline. SVR takes in static utterance-level features and predicts the probability of each grid separately. Because the regressors are optimized individually and the predictions are not bounded, we truncate negative values to zero and normalize the estimations over all grids to sum to one, as in [38, 42]. We concatenate the acoustic and visual features for the combined-early model.

We choose convolutional neural networks (CNNs) as our second approach. CNNs have been demonstrated to be effective in

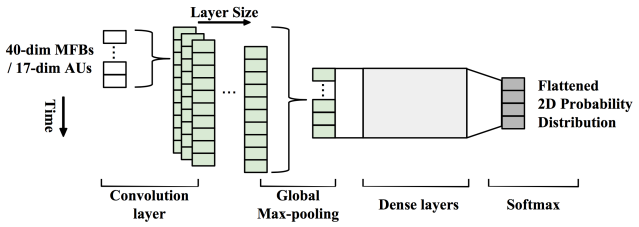

Figure 3: The structure of the unimodal CNN.

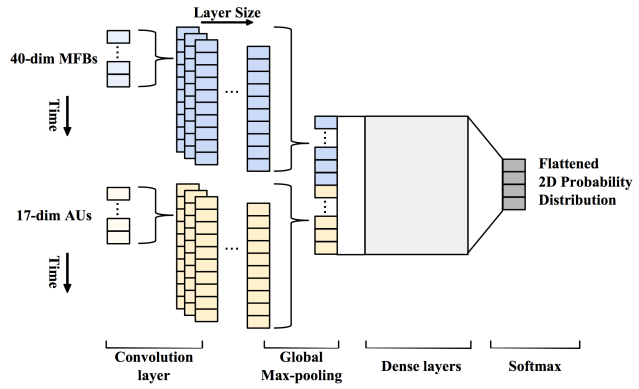

Figure 4: The structure of the multimodal CNN with the combined-early approach.

capturing regional saliency in speech emotion recognition [1]. We reframe the problem as classification using soft labels, rather than as regression. This leverages the fact that the output layer is usually a softmax, the output of which can be interpreted as the probability of each class.

We design our unimodal (i.e., audio or video) CNNs similar to [1], with an 1D-convolutional layer, a global max-pooling layer, several dense layers and a softmax layer (Figure 3). The 1D-convolutional layer takes in variable-length input and learns a sequence of feature representations by sliding N_F filters of length L_F through time. Each filter takes L_F consecutive frames and outputs an activation. By learning the filters, we are finding emotion-salient local temporal patterns. The global max-pooling layer identifies the highest activation of each filter over time and produces a feature vector of length N_F. This step allows us to focus on the most informative portion of an utterance and minimize the influence of padding and frames with invalid features. The interactions between the N_F features are further learned by applying several dense layers. Finally, we use a softmax layer to output the probability of emotion perception in each grid.

For the multimodal combined-early model, we build a separate 1D-convolutional layer and corresponding global max-pooling layer for audio and video, respectively. We concatenate the output of the two global max-pooling layers before feeding the features into dense layers. In this way, we allow for the difference in frame-rate of audio and video input, while still being able to explore the complex non-linear relationships between features across modalities. Our multimodal CNN is shown in Figure 4.

We use the Rectified Linear Unit (ReLU) [40] as the activation function for the convolutional and dense layers, and cross-entropy as the loss function. We apply L_2-regularization (0.0001) on the learned weights of the convolutional layers. The filter length of the convolutional layer (L_F), layer size (N_F), and the number of dense layers are treated as hyper-parameters.

5 EXPERIMENTAL SETTINGS

5.1 Feature Preparation

While the CNN models use frame-level features directly, the SVR models use a static approach with utterance-level features. We calculate 11 statistics over the frame-level MFBs and AUs and their first-order delta coefficient to generate the 880 and 374 utterance-level features for audio and video, respectively. The statistics include mean, standard deviation, max, position of the max frame, min, position of the min frame, range, interquartile range, mean absolute deviation, skewness, and kurtosis. For the video modality, the statistics calculation is applied only to frames with successfully extracted features (>98%). We also extract the state of the art Interspeech 2013 acoustic feature set [27] (6,373 utterance-level features, denoted as "IS13") to use in the SVR models for comparison.

We perform speaker-dependent z-normalization on all features before they are input into models. We normalize the features at the frame-level for CNN models and at the utterance-level for SVR models. Similarly, we exclude frames with unsuccessful AU extraction when z-normalizing the frame-level AUs. We replace these frames with zeros after normalization. We do not interpolate between frames that are successfully extracted because the unsuccessful extractions are usually a consecutive sequence of frames, and interpolation may introduce noise.

5.2 Performance Evaluation and Validation

We conduct experiments at two grid resolutions: 2×2 and 4×4. We use two metrics to evaluate the performance of the models: Total Variation (TV) and Jensen-Shannon divergence (JS). Both metrics can measure the difference between two discrete probability distributions. The metrics are calculated per utterance, and the lower TV and JS, the better the performance.

The value of Total Variation ranges in $[0, 1]$. Given two probability distribution X and Y over N states, The total variation between them is defined as

$$TV(X,Y) = \frac{1}{2}\sum_{i=1}^{N} |X_i - Y_i|, \tag{1}$$

Jensen-Shannon divergence is extended from the Kullback-Leibler divergence (denoted as KL). It is defined as

$$JS(X,Y) = \frac{1}{2}KL(X,M) + \frac{1}{2}KL(Y,M), \tag{2}$$

$$\text{where } M = \frac{1}{2}(X+Y) \text{ and } KL(X,Y) = \sum_{i=1}^{N} X_i \log \frac{X_i}{Y_i}.$$

We use JS instead of KL because: (1) JS is symmetric while KL is not, and (2) the value of JS ranges in [0, 1] when using log_2. Note that we replace zeros with 1e-8 when calculating the KL step in JS.

We train the models using a leave-one-speaker-out approach. At each round, a speaker is left out as the test set, while the other

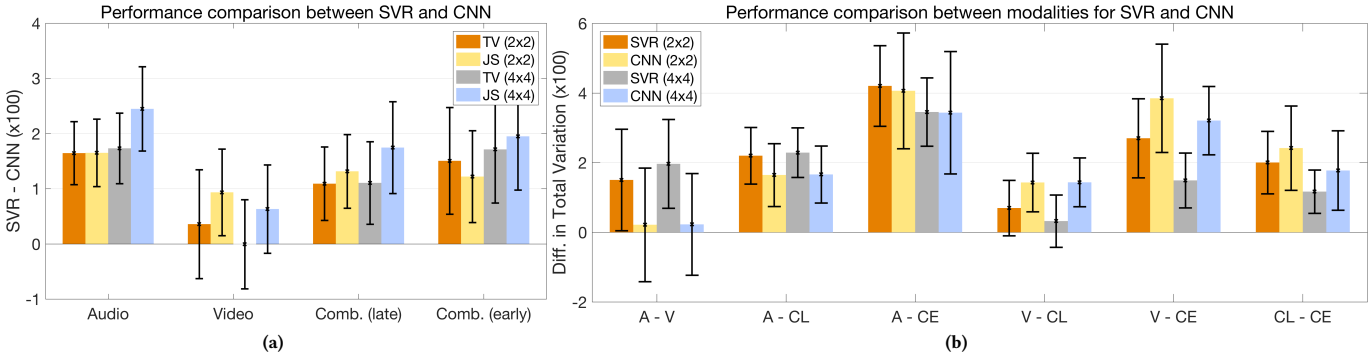

Figure 5: Performance difference at 2 × 2 and 4 × 4 resolutions, between: (1) SVR and CNN (in both total variation and JS-divergence), along with 95% confidence intervals of paired t-test; (2) different modalities for SVR and CNN (direction of subtraction shown as x-labels), along with 95% confidence intervals of Tukey's honest test.

Table 2: Range of Hyper-parameters in CNNs

Modality	Input	Filter length	Layer size	# Dense Layers
Audio	40	{8,16}	{128,256}	{1,2,3}
Video	17	{2,4}	{64,128,256}	{1,2,3}
Combined	40; 17	{8,16}; {2,4}	{64,128,256}	{1,2,3}

speaker in the same session is used as the validation set. The remaining ten speakers are used for training. We calculate the mean TV and JS for each test speaker and report the value averaged across all rounds as the performance of the models.

We use TV as the main validation metric because of its robustness to zeros. We select the hyper-parameters according to the validation TV. For the SVR models, the ranges are C (cost of error) $\in \{10^{-3}, 10^{-2}, ..., 10^{1}\}$, γ (kernel width) $\in \{10^{-5}, 10^{-4}, ..., 10^{-1}\}$ and ν (lower bound on the proportion of support vectors) $\in \{0.5, 0.6, 0.7, 0.8\}$. For the CNN models, the ranges of the filter length, layer size and the number of dense layers are shown in Table 2. Note that the number of filters in the convolutional layer and the size of the dense layers are kept the same. We use a training strategy of learning rate decay after N epochs. We randomly initialize the weights and start training with a learning rate of 0.001. We maintain the learning rate for 10 epochs, and select the one with the best validation TV to continue training. After that, we restore the previous weights and halve the learning rate when there is no improvement in validation TV after an epoch. We stop training when we reach the minimum learning rate or have five consecutive epochs with no improvement in validation performance, whichever comes first.

6 RESULTS AND DISCUSSION

6.1 Performance Comparison

We present the performance of the SVR and CNN models at two resolutions for different modalities in Table 3. In addition to the baseline, we provide chance performance of: (1) a uniform distribution across grids (denoted as Uniform), and (2) the mean distribution of the training set (denoted as MTrain). To answer the two research questions, we compare the performance between: (1) the SVR and CNN models when controlling for modality, and (2) the different modalities and combinations when controlling for the model.

Table 3: The performance of SVR and CNN models at two resolutions for each modality and combination. The best performance for each metric-resolution combination is bolded. The chance performances are also provided.

Modality	Model	Features	2×2		4×4	
			TV	JS	TV	JS
Chance	Uniform	-	.531	.303	.680	.481
	Mtrain	-	.475	.260	.596	.391
Audio	SVR	IS13	.390	.204	.528	.329
	SVR	MFB	.399	.213	.536	.340
	CNN	MFB	.383	.196	.519	.316
Video	SVR	AU	.384	.200	.516	.318
	CNN	AU	.381	.191	.516	.312
Combined -late	SVR	IS13+AU	.373	.185	.510	.307
	SVR	MFB+AU	.377	.189	.513	.311
	CNN	MFB+AU	.366	.176	.502	.293
Combined -early	SVR	IS13+AU	.362	.181	.507	.304
	SVR	MFB+AU	.357	.178	.501	.300
	CNN	MFB+AU	**.342**	**.166**	**.484**	**.281**

We show the performance difference between the SVR models with utterance-level MFB and/or AU features and CNN models with the corresponding frame-level features in Figure 5a, along with the 95% confidence interval of paired t-test. We see significant performance improvement when using CNN for both resolution and evaluation metrics (in the order of TV (2×2), JS (2×2), TV (4×4), and JS (4×4)), for audio (p = 5.6e-5, 9.7e-5, 9.4e-5, and 2.1e-5, respectively), combined-late (p = 0.0041, 0.0012, 0.0077, and 7.4e-4, respectively), and combined-early (p = 0.0057, 0.0081, 0.0026, and 0.0010, respectively). We also compare our CNN models with SVR with the state of the art IS13 feature set for models using audio input. The performance improvement of CNN is significant for audio (p = 0.0011 and 0.013 for TV and JS, respectively), combined-late (p = 0.0081 for JS) and combined-early (p = 2.7e-4 and 0.0014 for TV and JS, respectively). This indicates that focusing on salient local

Table 4: TV of SVR and CNN models (using matching utterance-level and frame-level features) along valence and activation, calculated from the prediction at 2 × 2 and 4 × 4 resolutions. The best performance for each dimension-resolution combination is bolded. V: Valence; A: Activation.

Modality	Model	2×2		4×4	
		V	A	V	A
Chance	Uniform	.362	.356	.362	.356
	MTrain	.363	.271	.363	.271
Audio	SVR	.322	.182	.324	.189
	CNN	.300	.182	.298	.183
Video	SVR	.267	.216	.272	.216
	CNN	.258	.221	.257	.224
Combined-late	SVR	.285	.190	.291	.196
	CNN	.269	.191	.267	.193
Combined-early	SVR	.264	**.176**	.272	.184
	CNN	**.248**	**.176**	**.254**	**.173**

regions and jointly optimizing for all the grids together is beneficial for audio input and multimodal input with either decision-level or feature-level fusion. Audio input benefits the most from the CNN architecture. Of the two multimodal systems, the performance gain from feature-level fusion is higher than decision-level fusion. However, there is no significant difference between the performance of CNN and SVR for video input, except when using JS as metric at 2×2 resolution (p = 0.024). This may be because that while the close relationship between AUs and emotion ensured the relevance of the features, the small dimensionality of the input and the high-level nature of the AUs limit the learning ability of the CNN. In addition, the errors propagated from AU estimation may have larger influence on the dynamic and more complex CNN models.

We perform the repeated-measure ANOVA (denoted as RANOVA) to compare different modalities for SVR and CNN. As the compound symmetry assumption may not be satisfied, we evaluate significance of the influence of modality based on the p-value with Lower bound adjustment (p_{LB}). If p_{LB} is smaller than 0.05, we perform the Tukey's honest significant difference test (denoted as Tukey test) for pairwise comparison using the model statistics of RANOVA. For simplicity, we only compare TV because it is used as the validation metric. We find that the influence of modality is significant for both SVR ($F(3,33)$ = 48.4 and 47.9, p_{LB} = 2.4e-5 and 2.5e-5 for 2×2 and 4×4, respectively) and CNN ($F(3,33)$ = 35.1 and 31.5, p_{LB} = 1.0e-4 and 1.6e-4 for 2×2 and 4×4, respectively). We show the pairwise comparison with the 95% confidence interval of the Tukey test in Figure 5b. For both SVR and CNN, combined-early significantly outperforms both unimodal models and combined-late. This suggests that both models have the ability to learn the interaction between audio and video when we perform fusion at the feature level. While decision-level fusion also brings improvement, this improvement is not always significant (e.g., combined-late vs. video for SVR). For unimodal inputs, video performs significantly better than audio for SVR while the performance of audio and video modalities are comparable when using CNN. This again supports that the CNN architecture may not be ideal for the high-level AU features and that there may be a larger information loss in the video modality.

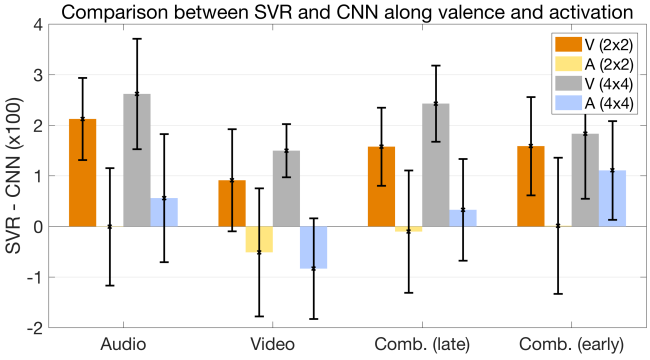

Figure 6: Performance difference (in TV) between SVR and CNN along valence and activation dimensions, calculated from predictions at 2 × 2 and 4 × 4 resolutions, with 95% confidence intervals of paired t-test. The best performance for each type-resolution combination is bolded.

6.2 Analysis along Valence and Activation

We assess the ability of the models to predict valence and activation. We marginalize by summing over the predictions along valence or activation to generate distributions of resolution 2×1 for negative and positive activation or 1×2 for negative and positive valence, and compare to the ground truth distribution processed in the same way. We show the resulting TV in Table 4, along with the chance of Uniform and MTrain for reference. We find that the audio modality is better at predicting activation, while the video modality is better at predicting valence. This is in line with previous findings [10, 21, 36]. We see multimodal improvement in the feature-level fusion setting (combined-early). In addition, when we compress the output to a two-state probability distribution along valence or activation, the predictions for 2×2 and 4×4 resolution have similar performance.

We illustrate the drop in TV of CNN compared to SVR along valence and activation dimensions in Figure 6, together with the 95% confidence interval of paired t-test. While most works using dynamic approaches witness higher performance improvement in activation [10, 21, 36], we surprisingly find that the performance gain of our CNN predominantly comes from valence, regardless of modality. More specifically, the difference between CNN and SVR is significant for valence for audio (p = 1.3e-4 and 2.6e-4 for 2×2 and 4×4, respectively), video (p = 6.1e-5 for 4×4), combined-late (p = 8.9e-4 and 2.0e-5 for 2×2 and 4×4, respectively), and combined-early (p = 0.0042 and 0.0093 for 2×2 and 4×4, respectively). The only significant result for activation comes from combined-early with 4×4 resolution (p = 0.030). These results indicate that identifying salient local patterns using CNN brings more benefit in predicting valence, compared to activation. This might be related to the observation in Section 3 that our data is more balanced along valence compared to activation.

6.3 Analysis of Different Types of Recordings

We investigate the performance difference between SVR and CNN for different types of recordings. As mentioned in Section 3, MSP-IMPROV consists of four types of recordings: read target sentences, target sentences from improvised scenes, other speaker turns from

Table 5: TV of SVR and CNN models (using matching utterance-level and frame-level features) for each type of recordings. T: Target sentences; I: Improvised turns; N: Natural interactions.

Modality	Model	2×2			4×4		
		T	I	N	T	I	N
Chance	Uniform	.530	.530	.529	.695	.676	.680
	MTrain	.489	.478	.456	.613	.595	.585
Audio	SVR	.410	.415	.366	.541	.549	.508
	CNN	.381 ∗	.411	.337 ∗	.533	.539 ∗	.481 ∗
Video	SVR	.372	.402	.358	.502	.535	.488
	CNN	.367	.405	.346	.519	.533	.486
Combined -late	SVR	.377	.395	.347	.509	.530	.483
	CNN	.357 ∗	.394	.326 ∗	.511	.522 ∗	.467
Combined -early	SVR	.342	.377	.328	.495	.519	.469
	CNN	**.315** ∗	**.370**	**.313**	**.479**	**.507**	**.451** ∗

improvised scenes, and natural interaction during the breaks. We combine the first two types in this analysis, because of the lack of the read target sentences for five of the speakers.

We present the TV of prediction for different types of recordings in Table 5, together with the chance performances. We find that in general, all the models are the best at predicting the emotion perceived from natural interactions, followed by target sentences. The emotion perceived from improvised scenes is the hardest to predict. This matches the classification accuracies of different types of recording using categorical labels reported in [5]. This might be because that the improvised scenes contain a wider range of emotion than the natural interactions since they were designed to enable the speakers to express a variety of emotions. The standard deviation of the mean evaluation of each utterance is higher in improvised scenes, suggesting a larger difference in the emotional content across utterances.

We mark the significant improvement (paired t-test, $p<0.05$) of CNN compared to SVR using "∗" in Table 5. We find that the performance gain of CNN is not consistent across different resolutions. For example, CNN significantly outperforms SVR for the target sentences for audio (p = 0.0028), combined-late (p = 0.011), and combined-early (p = 0.029) in the 2×2 case, but not in the 4×4 case. On the other hand, CNN significantly outperforms SVR for the improvised scenes for audio (p = 0.010) and combined-late (p = 0.011) in the 4×4 case, but not in the 2×2 case. The most consistent improvement is observed in natural interaction with audio and multimodal inputs. The performance difference between CNN and SVR is significant for audio at both resolutions (p = 0.0093 for 2×2, p = 0.018 for 4×4), combined-late at 2×2 (p = 0.011) and combined-early at 4×4 (p = 0.048), and is approaching significance for combined-early at 2×2 (p = 0.052).

7 CONCLUSION AND FUTURE WORKS

In this paper, we proposed a label processing method to generate two-dimensional discrete probability distributions on the valence-activation space from a limited number of ordinal labels. We showed that this method can preserve the mean evaluation of the original labels and that the correlation between the standard deviations of the original labels and up-sampled labels is high. Further, we explored the impact of modeling approaches (i.e., static SVR with individual optimization for each grid vs. dynamic CNN with joint optimization for all grids) and modalities on predicting the probability distribution of emotion perception. We hypothesized that using CNN models with a focus on salient local temporal patterns leads to a performance gain. In addition, combining audio and video modalities results in better performance compared to using each individual modality.

Our results show that the CNN models significantly outperform the SVR models when using the audio modality and combined audio and video modalities, supporting the effectiveness of modeling locally salient patterns and jointly predicting the distribution over all grids. CNN and SVR are comparable when the video modality is used. This indicates that the potential of CNN may not be fully explored when using a limited number of high-level AUs as inputs. In addition, the errors from AU estimation may have larger influence on the dynamic and more complex CNN models. We find that using both audio and video modalities is better than using either individually and that feature-level fusion is more beneficial than decision-level fusion. Analyses along different dimensions show that the audio modality is better at predicting activation while video modality is more advantageous at predicting valence, and that we can obtain improvement over the joined valence-activation space with feature-level fusion. This is in line with previous findings. In addition, we find that the performance gain brought by CNN mainly comes from the valence dimension. We see a consistent performance improvement over natural interactions when using CNN models, compared to SVR models.

A limitation of this work is that we combine features from the two modalities after global max-pooling. While this method allows us to overcome the difference in frame-rate, it disrupts the interaction between acoustic and visual features in real-time. Besides, using AU features and conducting global max-pooling may not be the best choice for dynamically modeling the video modality, as shown by our results. In the future, we plan to design models that combine audio and video inputs at an earlier stage, and explore the impact of using 2D or 3D facial landmarks or use raw video frames as input to the models. In addition, we will learn long-term interaction using methods such as recurrent neural networks with long short-term memory or dilated convolutional neural networks.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Michigan Institute for Data Science ("MIDAS"), by Toyota Research Institute ("TRI") and by National Science Foundation (NSF CAREER 1651740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the MIDAS, NSF, TRI, or any other Toyota entity.

REFERENCES

[1] Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *ICASSP*.
[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*. 1–10.

[3] Zdravko I Botev, Joseph F Grotowski, Dirk P Kroese, et al. 2010. Kernel density estimation via diffusion. *The Annals of Statistics* 38, 5 (2010), 2916–2957.

[4] 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.

[5] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed Abdel-Wahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing* 8, 1 (2016), 67–80.

[6] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27.

[7] Joel R Davitz. 1969. *The language of emotion.* Academic Press.

[8] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. 2007. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction* (2007), 488–500.

[9] Paul Ekman and Wallace V Friesen. 1977. Facial action coding system. (1977).

[10] Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces* 3, 1 (2010), 7–19.

[11] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 445–450.

[12] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication* 49, 10 (2007), 787–800.

[13] Hatice Gunes. 2010. Automatic, dimensional and continuous emotion recognition. (2010).

[14] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. 2014. Speech emotion recognition using CNN. In *ACM International Conference on Multimedia*. 801–804.

[15] 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. In *Interspeech*.

[16] 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.

[17] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[18] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. A multi-layer hybrid framework for dimensional emotion classification. In *ACM International Conference on Multimedia*. 933–936.

[19] Antonio Origlia, Francesco Cutugno, and Vincenzo Galatà. 2014. Continuous emotion recognition with phonetic syllables. *Speech Communication* 57 (2014), 155–169.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

[21] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (2015), 22–30.

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

[23] JA Russel. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (1980), 1161–78.

[24] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[25] Erik M Schmidt and Youngmoo E Kim. 2010. Prediction of Time-varying Musical Mood Distributions from Audio. In *ISMIR*. 465–470.

[26] Erik M Schmidt and Youngmoo E Kim. 2011. Modeling Musical Emotion Dynamics with Conditional Random Fields. In *ISMIR*. 777–782.

[27] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. (2013).

[28] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing* 1, 2 (2010), 119–131.

[29] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll. 2011. Using multiple databases for training in emotion recognition: To unite or to vote?. In *INTERSPEECH*. 1553–1556.

[30] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. 2014. Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces* 8, 1 (2014), 17–28.

[31] Stefan Steidl, Michael Levit, Anton Batliner, Elmar Noth, and Heinrich Niemann. 2005. Of all things the measure is manäutomatic classification of emotions and inter-labeler consistency. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings*, Vol. 1. I–317.

[32] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. 5200–5204.

[33] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. 2015. Fera 2015-second facial expression recognition and analysis challenge. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 6. 1–8.

[34] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. 2012. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *ACM International Conference on Multimedia*. 89–98.

[35] Ju-Chiang Wang, Yi-Hsuan Yang, Hsin-Min Wang, and Shyh-Kang Jeng. 2015. Modeling the affective content of music with a Gaussian mixture model. *IEEE Transactions on Affective Computing* 6, 1 (2015), 56–68.

[36] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn W Schuller, Cate Cox, Ellen Douglas-Cowie, Roddy Cowie, et al. 2008. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.. In *Interspeech*, Vol. 2008. 597–600.

[37] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. 2017. Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network. In *AAAI*.

[38] 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.

[39] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. 2015. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 451–458.

[40] Matthew D Zeiler, M Ranzato, Rajat Monga, Min Mao, Kun Yang, Quoc Viet Le, Patrick Nguyen, Alan Senior, Vincent Vanhoucke, Jeffrey Dean, et al. 2013. On rectified linear units for speech processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 3517–3521.

[41] Biqiao Zhang, Emily Mower Provost, and Georg Essl. 2017. Cross-corpus Acoustic Emotion Recognition with Multi-task Learning: Seeking Common Ground while Preserving Differences. *IEEE Transactions on Affective Computing* (2017).

[42] 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.

[43] Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. 523–528.