Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network

Duc Le, Zakaria Aldeneh, Emily Mower Provost

University of Michigan, Ann Arbor, MI 48109, USA
Computer Science and Engineering
{ducle,aldeneh,emilykmp}@umich.edu

Abstract

Estimating continuous emotional states from speech as a function of time has traditionally been framed as a regression problem. In this paper, we present a novel approach that moves the problem into the classification domain by discretizing the training labels at different resolutions. We employ a multi-task deep bidirectional long-short term memory (BLSTM) recurrent neural network (RNN) trained with cost-sensitive Cross Entropy loss to model these labels jointly. We introduce an emotion decoding algorithm that incorporates long- and short-term temporal properties of the signal to produce more robust time series estimates. We show that our proposed approach achieves competitive audio-only performance on the RECOLA dataset, relative to previously published works as well as other strong regression baselines. This work provides a link between regression and classification, and contributes an alternative approach for continuous emotion recognition.

Index Terms: emotion recognition, multi-task learning, deep learning, emotion decoding, emotion language model

1. Introduction

Estimating a speaker's affective state over time has a wide range of real-world applications and is a problem of great interest in the emotion recognition community [1, 2]. The emotional state at each time step is typically encoded as a point in the latent dimension space, usually consisting of arousal, valence, and dominance [3]. The 2015 and 2016 Audio-Visual Emotion Challenge (AVEC) [4, 5], conducted on the Remote Collaborative and Affective Interactions (RECOLA) dataset [6], provided a common platform for evaluating machine learning methods for continuous emotion recognition. The objective of the challenge was to make temporal predictions of continuous arousal and valence values given audio, video, and physiological data. We focus exclusively on audio in this work.

Given the continuous nature of the labels, existing works have mostly tackled this task as a regression problem, training systems to output the exact arousal/valence values. Yet, a common method in dimensional emotion recognition is to quantize the latent dimension into different levels of intensity [7–12]. This helps simplify the modeling problem at the cost of reduced label resolution. We argue that a similar idea can be used to improve performance on RECOLA. Instead of predicting the exact values, the system can be trained to output a sequence of classes corresponding to coarse levels of arousal/valence. Labels with different coarseness might provide complementary information and can be modeled jointly with multi-task learning (MTL). This approach is promising given the relatively small amount of training data and the inherent label uncertainty.

Our approach consists of three primary steps. Firstly, we map the continuous label values onto a small set of discrete emotion classes using k-means. Secondly, we train a multi-task deep bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) with cost-sensitive Cross Entropy (CE) loss to jointly predict label sequences at different resolutions. Finally, we employ a decoding framework that incorporates an emotion "language model" to produce more robust time series estimates. We show that this approach gives competitive performance on RECOLA, outperforming the audio-only baseline [5] by a large margin with minimal feature engineering. The contribution of this work lies in the novel problem formulation, the introduction of cost-sensitive CE loss to exploit inter-class relationship, and the emotion decoding framework that can leverage known temporal patterns in the label sequences.

2. Related Work

2.1. AVEC Approaches

All participants in AVEC 2015 and 2016 utilized features from multiple modalities and, except for the baseline papers [4,5], reported audio results only on the development set. In this section, we review some of the approaches taken for continuous emotion recognition on RECOLA using audio features.

The baseline approaches involved extracting the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [13], a relatively small, manually engineered feature set of acoustic low-level descriptors (LLDs), trained with Support Vector Regression (SVR) [4, 5] and linear fusion with various types of neural networks [4]. The winner of AVEC 2015, He et al., utilized a comprehensive set of 4684 features including energy, spectral, and voicing-related features, followed by feature selection and BLSTM training [14]. The best performer of AVEC 2016, Brady et al., used SVR trained on sparse-coded higher-level representations of various types of audio features [15]. Povolny et al. trained a set of linear regressors on eGeMAPS augmented with deep bottleneck features from Deep Neural Network (DNN) acoustic models [16]. Trigeorgis et al. trained a convolutional RNN directly on raw waveform [17]. However, they used non-public portions of RECOLA and evaluated on a larger test set, making the results non-comparable.

Unlike these works, which tackled the problem as a regression task, our work approaches the problem from a classification perspective. Further, our system utilizes simple log Mel filterbank coefficients instead of manually engineered features.

2.2. Discretization of Continuous Emotion Labels

Discretization is a commonly employed method to simplify the modeling problem when dealing with continuous dimensional emotion values. Wöllmer et al. tackled arousal/valence prediction as a 4- or 7-class classification task [7, 9]. Meng and Bianchi-Berthouze tackled 2-class prediction for arousal,

valence, expectation, and power by modeling each class as a Hidden Markov Model (HMM) state [11]. Similar approaches are also common in the music emotion recognition literature. Zhang et al. used Affinity Propagation (AP) to cluster the 2D arousal/valence space, and manually mapped each cluster to a fixed arousal/valence value [8]. Yang et al. and Zhang et al. discretized the arousal/valence space into a G×G grid and modeled each label as a distribution over grid cells [10, 12].

A common theme among these works is that the discretization scheme is consistent with the evaluation metric. By contrast, both the ground-truth labels and the evaluation metric in this work are fully continuous, making it more difficult for discretization-based methods to achieve similar performance.

2.3. BLSTM and Speech Emotion Recognition

In this section, we give a quick overview of BLSTM and its applications in emotion recognition. A standard LSTM-RNN layer receives an input vector sequence $x = (x_1, \ldots, x_T)$ and produces a hidden vector sequence $h = (h_1, \ldots, h_T)$:

$$(h_t, c_t) = \mathcal{H}(x_t, h_{t-1}, c_{t-1}) \tag{1}$$

where h_t and c_t are the hidden and cell activation vectors at time step t, and $\mathcal{H}$ is the LSTM activation function [18]. BLSTM-RNN is an extension to this architecture, which adds a parallel LSTM-RNN layer that processes the input sequence backward:

$$(\overrightarrow{h}_t, \overrightarrow{c}_t) = \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}, \overrightarrow{c}_{t-1}) \tag{2}$$

$$(\overleftarrow{h}_t, \overleftarrow{c}_t) = \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}, \overleftarrow{c}_{t+1}) \tag{3}$$

The output of a BLSTM-RNN layer is the concatenated hidden vector $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$. Multiple BLSTM-RNN layers can be stacked on top of each other to create a deep BLSTM-RNN architecture. Finally, an output layer can be added:

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{4}$$

where $W_{\overrightarrow{h}y}$ and $W_{\overleftarrow{h}y}$ are the hidden-output weight matrices and b_y is the bias vector. For classification tasks, softmax normalization is applied to the output vector y_t.

BLSTM-RNN's primary advantage is its ability to model long-range temporal dependencies [18]. This architecture has achieved state-of-the-art performance on various automatic speech recognition benchmarks (e.g., [19–21]). It has also been used successfully for speech emotion recognition [7, 9, 14, 17].

3. Data and Feature Extraction

All experiments in this work are carried out on the RECOLA dataset [6]. RECOLA contains 27 five-minute recordings (i.e., utterances) partitioned into a training, development, and test set with nine utterances each. Ground-truth continuous annotations for arousal and valence were computed every 40ms from six gender-balanced annotators. We use the gold-standard ground-truth labels as they were defined in the AVEC 2016 affect recognition sub-challenge [5]. The official evaluation metric for this task is the Concordance Correlation Coefficient (CCC):

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{5}$$

where x and y are the predicted and ground-truth time series concatenated over all test utterances, s_{xy} is the covariance, s_x^2 and s_y^2 are the variances, and $\bar{x}$ and $\bar{y}$ are the time series means. In addition to CCC, Root Mean Squared Error (RMSE) is a commonly reported and analyzed metric.

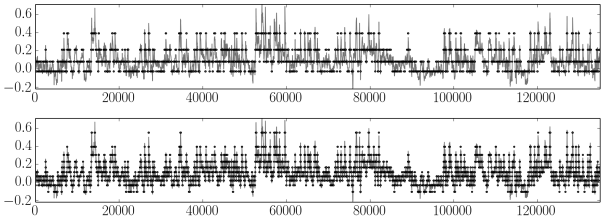

Figure 1: *Discretization of continuous valence values using 4 (top) and 10 (bottom) k-means clusters.*

Clusters	Arousal		Valence	
	CCC	RMSE	CCC	RMSE
4	0.953	0.057	0.934	0.045
6	0.978	0.040	0.970	0.031
8	0.987	0.031	0.983	0.024
10	0.991	0.025	0.989	0.019

Table 1: *K-means discretization reconstruction accuracy.*

We use the Kaldi toolkit [22] to extract 40-dimensional log Mel filterbank coefficients for each utterance, using a 25ms window and 10ms frame shift. We concatenate every four consecutive frames together to ensure that the features, which are now 160-dimensional and span 40ms, align with the labels. Finally, we perform per-utterance z-normalization on the features.

4. Methods

In this section, we will discuss three major components of our methods: (1) label discretization, (2) BLSTM-RNN training, and (3) emotion decoding. We will also describe two regression systems to compare with our proposed approach.

4.1. Label Discretization

The motivation behind label discretization is that raw dimensional emotion values are noisy, making it difficult to train regression models to predict the exact numbers. An alternative method is to divide these values into several discrete groups corresponding to different levels of arousal/valence. This approach can simplify the modeling process, converting the original regression problem into a discrete classification task. However, it may result in less accurate labels due to the loss of resolution.

There is no consensus in the literature as to what the best discretization technique is. One common method is to quantize the continuous values into equal-length segments on the arousal/valence axis. We found that this method led to extreme class imbalance, which is known to cause difficulties for classification. Instead, k-means can be used to partition arousal/valence values into several clusters, resulting in a more balanced class distribution. We can approximate the time series by replacing each label with the mean of the cluster to which it belongs. We will investigate other discretization techniques in future work.

We perform k-means clustering on all continuous labels of arousal/valence in the training and development set, with the number of clusters set to 4, 6, 8, and 10. Intuitively, few clusters are easier to model but do not approximate the original labels as accurately, whereas many clusters can reconstruct the labels accurately but are difficult to model due to data sparsity. Figure 1 shows the discretization results for valence using 4 and

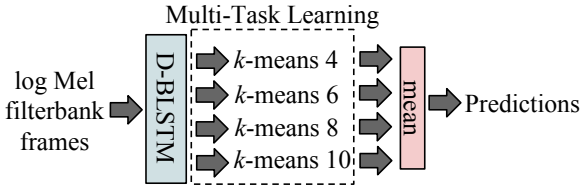

Figure 2: *Multi-task deep BLSTM-RNN architecture.*

10 clusters. As can be seen, this approach has a clipping effect on extreme continuous values. Table 1 lists the CCC and RMSE between the original and reconstructed time series (concatenation of all training and development utterances). High CCC and low RMSE can be achieved with as few as 4 clusters, demonstrating the potential of this discretization approach.

4.2. Multi-Task Deep BLSTM-RNN Classifier

After discretization, each continuous arousal/valence value is converted to four discrete class labels corresponding to the four k-means clustering configurations. We employ a multi-task BLSTM-RNN architecture (Figure 2) to model these four labels jointly, motivated by prior work demonstrating that MTL can improve performance if the component tasks are closely related [23–26]. Each task produces a sequence of class labels, which can be mapped to fixed arousal/valence values based on the cluster means. The final output of the network is the averaged arousal/valence values across all tasks. This helps increase the resolution and smoothness of the reconstructed labels, but will tend to push them toward more moderate values.

We train the network to minimize the total cost-sensitive CE (CCE) loss across all tasks. The target probability distributions are encoded as one-hot vectors (1 at the correct class index, 0 everywhere else), which help simplify the loss function:

$$\mathcal{L}_{CCE} = -\sum_{t=1}^{T}\sum_{f=1}^{F} \mathcal{C}(\boldsymbol{y_{tf}}, l_{tf}) \log \boldsymbol{y_{tf}}^{(l_{tf})} \tag{6}$$

where T is the number of tasks, F is the total number of feature frames, $\mathcal{C}$ is the cost function, $\boldsymbol{y_{tf}}$ is the model's output probability distribution for frame f and task t, l_{tf} is the correct class index for frame f and task t, and $\boldsymbol{y_{tf}}^{(l_{tf})}$ is the output probability of the correct class for frame f and task t. In regular CE loss, the cost function $\mathcal{C}$ is simply the constant 1. In our work, we leverage the fact that classes are spatially related as they correspond to different levels of arousal/valence. We use a cost function that places more weight on frames whose prediction is farther from the correct class label:

$$\mathcal{C}(\boldsymbol{y_{tf}}, l_{tf}) = 1 + \frac{|\arg\max_i \boldsymbol{y_{tf}}^{(i)} - l_{tf}|}{N_t} \tag{7}$$

where N_t is the number of classes in task t. We found that CCE led to more stable training compared to regular CE.

Training deep BLSTM-RNN on small datasets is challenging. We found that the following choices were crucial for this task. One, we train the network using the Adam optimizer [27] with a minibatch size of one utterance and full Backpropagation Through Time (BPTT). Two, we employ a 2-stage early stopping approach. In the first stage, the network is trained for an initial 20 epochs with a 0.002 learning rate, and the best model parameters (measured in CCC on the development set) are saved. In the second stage, the learning rate is halved after every epoch if the development CCC does not improve. This continues until the learning rate drops below 0.00001 or the total epochs exceed 40. This approach is especially important for valence, which takes longer to converge than arousal. Three, we scale the model output by the inverse of the class priors computed from training labels, followed by renormalization. This is a common technique for handling unbalanced class distributions.

We fix the BLSTM-RNN layer size at 160 (80 for forward, 80 for backward). The number of BLSTM-RNN layers (3, 4, or 5) and L2 regularization weight (0.0 or 0.00002) are cross-validated based on the development CCC. To help reduce training variation, each hyperparameter combination is run three times with different random seeds. The averaged prediction across these three runs is used for final testing and validation.

4.3. Emotion Decoding

The output of each task in our BLSTM-RNN is a sequence of probability distributions over classes. The end goal is to find a sequence of classes, which we can use to reconstruct the time series through the cluster means. The simplest way to achieve this is to select the class index with the highest probability for each frame. However, this approach does not take advantage of the fact that the label sequence is slow-moving; as a result, it may lead to erratic label transitions. Meng and Bianchi-Berthouze made a similar observation and tackled the decoding problem as finding the best path through a 2-state HMM [11].

We adopt this idea and model each class as a HMM state. The transition probabilities between states govern the short-term temporal property of the output sequence; they can be estimated from the training labels. We further extend this idea by incorporating an emotion "language model" to control the long-term temporal patterns of the label sequence. Each frame can be thought of as belonging to a larger emotion *region*. Hence, the frame labels define a sequence of regions, with each region spanning multiple frames. In this work, we define a region simply as a segment of frames belonging to the same class (i.e., HMM state). Let $\mathcal{B}(\boldsymbol{l})$ be a function that takes as input a sequence of frame labels $\boldsymbol{l}$ and collapses identical adjacent labels to produce a sequence of region labels. The emotion decoding problem[1] can then be formulated as:

$$\boldsymbol{l}^* = \arg\max_{\boldsymbol{l}} AM(\boldsymbol{l}) + \alpha LM(\mathcal{B}(\boldsymbol{l})) \tag{8}$$

where $\boldsymbol{l}$ is iterated over the set of all possible frame label sequences, AM and LM are the acoustic and language model scores, respectively, and α is the LM weight. For a frame label sequence $\boldsymbol{l} = (l_1, \ldots, l_F)$, the AM score can be computed as:

$$AM(\boldsymbol{l}) = \sum_{f=1}^{F} \log \boldsymbol{y_f}^{(l_f)} + \sum_{f=1}^{F-1} \log P(l_{f+1}|l_f) \tag{9}$$

where $\boldsymbol{y_f}^{(l_f)}$ is the deep BLSTM-RNN's output at frame f for label l_f, and $P(l_{f+1}|l_f)$ is the probability of transitioning into l_{f+1} given the previous label l_f. For LM, we employ an n-gram model computed from training and development data.

After BLSTM-RNN training, we sweep over different n-gram language model types (bigram or trigram) and α values $(0, 2, \ldots, 14)$. The best configuration in terms of development CCC is used for final testing and validation.

[1] In practice, this intractable problem is solved using beam search.

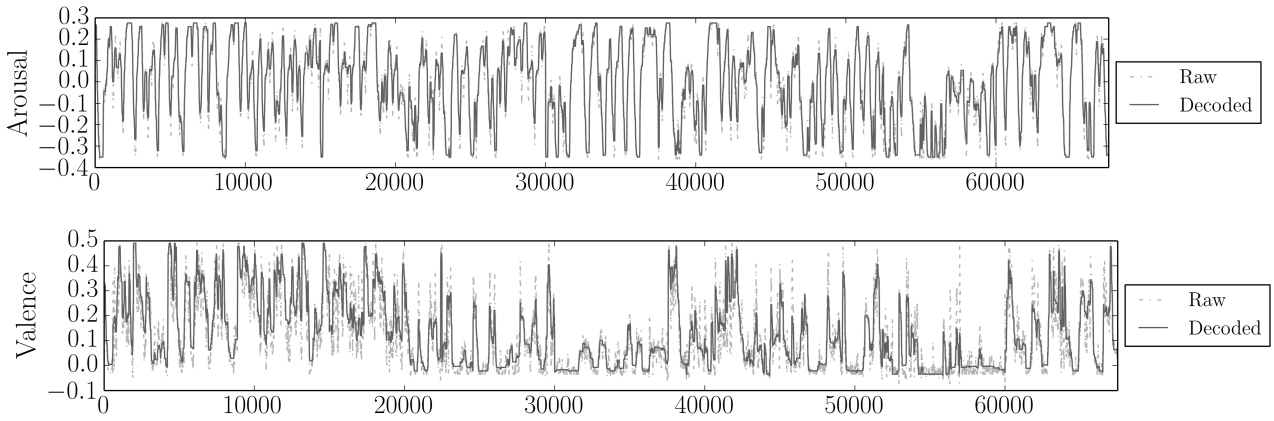

Figure 3: *Effect of emotion decoding on arousal (top) and valence (bottom) test prediction.*

	System	Arousal		Valence	
		Dev	*Test*	*Dev*	*Test*
AVEC	Valstar et al. [5]	.796	.648	.455	.375
	Povolny et al. [16][2]	.832	.682	.489	.349
	Brady et al. [15]	.846	–	.450	–
Proposed	REG-MSE	.835	.652	.268	.238
	REG-CCC	.855	.664	.518	.499
	CLS-Raw	.858	.682	.563	.448
	CLS-Decoded	.859	.680	.596	.460

Table 2: *Audio-only results on RECOLA in CCC.*

4.4. Deep BLSTM-RNN Regressor

In order to evaluate the effectiveness of our proposed approach, we compare it with regression-based approaches. The most directly comparable system is a deep BLSTM-RNN regressor trained with MSE loss, as both MSE and CCE are frame-level objectives. Previous works have reported that using CCC as the objective function consistently outperforms MSE for regression models [16, 17]. In addition to being more in line with the evaluation metric, CCC has the advantage over MSE and CCE in that it is an utterance-level objective that takes into account the overall shape of the time series. We will investigate adapting the CCC objective to the classification paradigm in future work.

We train two deep BLSTM-RNN regression models, one with MSE and one with CCC loss. The training and validation procedures are identical to those described in Section 4.2.

5. Results and Discussion

Table 2 summarizes the development and test CCC for our proposed approaches as well as other works targeting AVEC 2016. *CLS-Raw* and *CLS-Decoded* are our classification-based systems without and with emotion decoding, respectively. *REG-MSE* and *REG-CCC* are regression models trained with MSE and CCC loss. Compared with previously published results, our classification-based approaches outperform the baseline CCC in Valstar et al. [5] by a large margin for both arousal and valence, and perform favorably compared to Povolny et al. [16] and Brady et al. [15] (in terms of development CCC). *CLS-Raw* and *CLS-Decoded* clearly outperform *REG-MSE*. For *REG-CCC*, classification-based systems perform better on arousal but worse on valence. As previously discussed, the CCC loss, being an utterance-level objective and in line with the evaluation metric, has considerable advantage over MSE and CCE. Nevertheless, these results demonstrate the efficacy of our classification-based approaches. In future work, we will investigate objective functions that are more suitable for sequence classification.

It can be observed that classification-based approaches achieve high performance on the development set; however, the test improvement is comparatively lower. One of the challenges of this approach is that the cluster means obtained during training may not match the test labels perfectly. We expect this problem to lessen as the size of the training set increases.

Emotion decoding does not have a big impact on arousal, but improves valence performance. We plot the predicted time series for arousal and valence without and with emotion decoding to gain more insights into this result (Figure 3). In both cases, emotion decoding has a smoothing effect on the time series; however, the impact is much more pronounced in valence, which explains the larger CCC improvement. We hypothesize that, similar to in speech recognition, the language model is more effective when the acoustic model is less accurate, as is the case for valence. Compared to other time series smoothing techniques used on this task, such as Gaussian [14] and median [5, 15–17] filtering, our approach has the advantage of being completely data-driven. Future work will compare different smoothing methods in more detail.

6. Conclusion and Future Work

In this paper, we described a novel approach to the AVEC 2016 continuous emotion recognition task by using label discretization, multi-task deep BLSTM-RNN training, and a decoding framework that incorporates an emotion "language model." We showed that our proposed method achieved competitive performance compared to previously published works as well as other strong regression baselines. For future work, we will experiment with alternative discretization techniques and objective functions to further improve the results. We will also evaluate this approach on other datasets to demonstrate its efficacy.

7. Acknowledgements

This work was supported by IBM under the Sapphire project. We'd like to thank Dr. David Nahamoo and Dr. Lazaros Polymenakos, IBM Research, Yorktown Heights, for their support.

[2]Unpublished test results, courtesy of Povolny et al. [16].

8. References

[1] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, January 2010.

[2] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, Feb. 2013.

[3] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 12 2007.

[4] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "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*, 2015, pp. 3–8.

[5] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[6] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "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*. IEEE, 2013, pp. 1–8.

[7] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *INTERSPEECH*, 2008.

[8] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 1369–1372.

[9] M. Wöllmer, B. W. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[10] Y. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2184–2196, 2011.

[11] H. Meng and N. Bianchi-Berthouze, "Affective State Level Recognition in Naturalistic Facial and Vocal Expressions," *IEEE Trans Cybern*, vol. 44, no. 3, pp. 315–328, Mar 2014.

[12] B. Zhang, E. M. Provost, R. Swedberg, and G. Essl, "Predicting emotion perception across domains: A study of singing and speaking," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 1328–1334.

[13] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, pp. 14–27, 2016.

[14] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "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, 2015, pp. 73–80.

[15] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "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, 2016, pp. 97–104.

[16] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for avec 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 75–82.

[17] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "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*. IEEE, 2016, pp. 5200–5204.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 9, pp. 1735–1780, Nov. 1997.

[19] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, Vancouver, BC, Canada, 2013.

[20] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, Singapore, 2014.

[21] H. Sak, A. W. Senior, K. Rao, and F. Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," *CoRR*, vol. abs/1507.06947, 2015.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[23] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*, Vancouver, BC, Canada, 2013.

[24] D. Chen, B. Mak, C. C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *ICASSP*, Florence, Italy, 2014.

[25] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *ICASSP*, Brisbane, Australia, 2015.

[26] P. Bell and S. Renals, "Complementary tasks for context-dependent deep neural network acoustic models," in *INTERSPEECH*, Dresden, Germany, 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.