HUMAN-LIKE EMOTION RECOGNITION: MULTI-LABEL LEARNING FROM NOISY LABELED AUDIO-VISUAL EXPRESSIVE SPEECH

Yelin Kim

Electrical and Computer Engineering
University at Albany, SUNY, USA
yelinkim@albany.edu

Jeesun Kim

The MARCS Institute
Western Sydney University, Australia
j.kim@westernsydney.edu.au

ABSTRACT

To capture variation in categorical emotion recognition by human perceivers, we propose a multi-label learning and evaluation method that can employ the distribution of emotion labels generated by every human annotator. In contrast to the traditional accuracy-based performance measure for categorical emotion labels, our proposed learning and inference algorithms use cross entropy to directly compare human and machine emotion label distributions. Our audio-visual emotion recognition experiments demonstrate that emotion recognition can benefit from using a multi-label representation that fully uses both clear and ambiguous emotion data. Further, the results demonstrate that this emotion recognition system can (i) learn the distribution of human annotators directly; (ii) capture the human-like label noise in emotion perception; and (iii) identify infrequent or uncommon emotional expression (such as frustration) from inconsistently labeled emotion data, which were often ignored in previous emotion recognition systems.

Index Terms— Emotion recognition, prototypicality, label noise, multi-label learning, soft labeling, audio-visual emotion

1. INTRODUCTION

Our aim is to develop emotion recognition algorithms that go beyond one-hot label assignment (e.g., 'happy or 'valence: 2.5') to infer the emotion distribution produced by multiple human annotators. The main challenge in developing emotion recognition systems has been the subjectivity and ambiguity in ground truth labels for emotion. This paper presents a new 'human-like' emotion recognition system that represents, learns, and evaluates categorical emotion labels as multi-label distributions. This system is based on multi-label learning and inference algorithms that can directly learn the emotion distribution of multiple human annotators; in addition, we introduce new performance measures based on the consistency of the emotion label distribution between machine and human.

Traditional emotion recognition systems measure system performance using accuracy based on a comparison between the aggregation of annotator outcomes (as a ground truth label) and the estimated emotion label from the system. There are two basic approaches for emotion representation, dimensional and categorical approaches. To overcome label noise, these approaches usually either aggregate (dimensional approach) or take the majority vote (categorical approach). For instance, using the traditional one-hot labeling, Valstar et al. calculated the average dimensional emotion ratings from all raters [1]; Ringeval et al. employed a normalization technique to increase the inter-rater agreement, while preserving the original balancing of the dimensional ratings [2]; Shah et al. used a majority vote-based categorical emotion ground truth for multimodal emotion recognition [3].

Recent studies have attempted to combat label noise using soft labeling approaches [4–6]. For instance, Mower et al. presented a pioneering work that represented emotions with soft-labeling by using a set of binary emotion classifier outputs [4]. Lotfian and Busso presented an innovative probabilistic method for soft labeling of emotion in speech emotion recognition [5]. These methods have shown improved emotion recognition performance and provided more interpretable representation for ambiguous emotions compared to traditional systems. However, these previous soft labeling methods discarded inconsistently labeled data (data with no majority vote (NMV) from annotators). So, although effective in reducing annotator variations, these methods may not preserve all the variations of human annotators in ground truth and so may remove potentially useful information about emotionally expressive behavior.

In this work, we propose a soft-multi-labeling technique that is based on annotator distribution over emotion classes and as such use all the available data, even when the data indicates no agreement between human evaluators. Our multi-label approach can represent the distribution of emotion, even for NMV; so we can retain, learn, and infer the subtle difference within NMV data.

Using a multi-label categorical approach for emotion representation, we investigate how to capture and utilize the emotion label noise in emotion recognition systems to provide a within-category emotion dimension. We propose to use a feedforward neural network with the output layer activated using the full emotion distribution over multiple human annotators. Our network then learns the multi-label distribution directly. To shift the performance goal of emotion recognition to the learning of subjective annotator evaluation, we also propose to use cross entropy between the true and estimated emotion distribution, rather than using a one-hot-label based accuracy. Our proposed method uses more descriptive and richer emotion labels than traditional methods and can shed light on the relationship between audio-visual emotion expressions and emotion perception.

Our experimental results show that the proposed multi-label approach achieves higher accuracy than traditional one-hot labeling and provides human-like interpretation of automatic emotion recognition. The results also demonstrate the importance of emotionally ambiguous data in learning by showing that the use of NMV data in emotion recognition systems improves the overall performance. To the best of our knowledge, this is the first attempt to develop an emotion recognition system for five-class classification of anger, happiness, neutrality, sadness, and frustration, for IEMOCAP [7].

The key innovation of this proposed work is the inclusion of emotionally ambiguous data using new learning methods. Ambiguous, subtle expressions of emotion, which often obtain no majority

agreement from human annotators, are prevalent in the real world [8]. The use of such expressions in our learning or training will increase the size of the available data (e.g., 17.25% for the IEMOCAP benchmark dataset [7]) and make possible the application of big data approaches, such as deep learning, to emotion recognition.

2. BACKGROUND AND RELATED WORK

Audio-visual emotion recognition systems computationally classify emotion from expressive audio-visual behavior, such as speech, facial expression, and body gesture [8–12]. These systems mostly use the perceived emotion labels generated by multiple human annotators as ground truth emotion labels. However, a fundamental challenge in developing real-world emotion recognition systems is the noisiness in emotion labels due to subjectivity in emotion perception (class noise) and ambiguity in emotion expression, which we call 'emotion label noise'. This noise, need not relate to how accurate or reliable the data is, instead, it may reflect the ambiguity and subtlety of the emotion expressions themselves. Noise is not just an error in an emotion classification task; it may contain meaningful information that reflect ambiguity or mixed emotional phenomena, such as the mixture of happiness and sadness.

There are multiple factors that give rise to emotion label noise: for example, perceiver error, differential perceiver bias or subjectivity [13, 14], production variability within or across individuals [15], production of mixed emotion (multiple) expressions [16], and the domain – some emotions (e.g., disgust) simply are not clear-cut [17]. The level of noisiness in emotion labelling can be considered to indicate the degree of prototypicality of the expression. In other words, when an instance of an expression has low emotion label noise it can be considered to be prototypical, a result due to the combination of less variability in production, some robustness to perceiver' bias, and this instance being unambiguous (some emotions may have very few prototypical exemplars).

Several studies have investigated the use of non-prototypical data in emotion recognition systems. Mower et al. [18] studied a system's ability to interpret non-prototypical emotions, and proposed a new method to represent the confidence level of presence of certain emotion classes using outputs of binary classifiers [4]. Schuller et al. studied data selection methods to select emotionally salient training data [19]. They calculated prototypicality using the Euclidean distance of the class center of positive instances of SVM classifiers. The experimental results on eight emotion databases showed that the proposed method performs well for estimating arousal, but not for valence. Kim et al. [20] studied deep learning methods to learn complex interactions between audio and visual emotion expressions, and found that unsupervised feature learning that use deep neural networks is more effective for non-prototypical data than prototypical data. The key difference between our proposed method and those implemented in the above studies is that in using NMV as well as non-prototypical (and prototypical) data, the current system provides a method that employs the emotion distribution of multiple labels generated by human annotators.

Recent studies have employed deep neural networks [5, 6] and multi-task learning [21] to implement soft-labeling approaches. Lotfian and Busso estimated probabilistic distributions of emotion and considered the covariance matrix between emotion categories while training deep neural networks [5]. Fayek et al. compared an ensemble and soft-label method when modeling inter-rater variability, using speech and categorical emotions [6]. Han et al. used a multi-task learning approach, where two tasks are emotional states and the degree of uncertainty (measured using inter-rater agreement) [21]. These soft-labeling approaches show improvement in emotion recognition when using soft-label approaches, however, an open question remains concerning the use of inconsistently labeled emotion data or ambiguous emotion expressions, such as *frustration*. Our work differs from the above studies in that we develop a learning and inference algorithm that can utilize the data for which a 'gold standard' is difficult to find. Our work also differs in that we directly map machine and human confusions, rather than modeling uncertainty in emotion perception. We propose that this will lead to a more 'human-like' emotion recognition system that can produce similar responses to equivocal emotions as humans.

3. DATA AND FEATURES

To evaluate our proposed approaches, we use an established audio-visual emotion dataset, IEMOCAP [7]. This dataset recorded ten speakers (five sessions of female-male pairs) during hypothetical emotional situations. The dataset includes audio, 3-D motion capture markers, and transcripts. We use both 3-D motion capture data from 55 markers on the faces and speech data that include pitch, energy Mel filter bank features, as in [4, 20]. For both facial and speech features, we compute 8 statistical functionals (mean, standard deviation, lower quantile, upper quantile, quantile range, and polynomial regression coefficients of degree three) at the utterance level. The resulting number of features are 1320 for face motion and 232 for speech. We calculate the global mean over the ten speakers for each feature dimension and normalize each speaker's audio-visual features using mean normalization as in [20, 22].

We use categorical labels, annotated by at least three human annotators for each utterance and assign the prototypicality labels as follows: prototypical (total agreement, 'Prot'), non-prototypical (majority agreement, 'Non-Prot'), and non-majority-vote (no agreement, 'NMV') labels. The emotion categories include anger (prot: 296, non-prot: 325), happiness (prot: 766, non-prot: 532), neutrality (prot: 127, non-prot: 327), frustration (prot: 372, non-prot: 626), surprise (prot: 1, non-prot: 30), fear (prot: 4, non-prot: 16), disgust (prot: 0, non-prot: 1), and other (prot: 0, non-prot: 2). In this paper, we used 1891 prototypical, 2338 non-prototypical, and 812 NMV utterances in total, retaining both data that have consistent (majority vote) and inconsistent (NMV) emotion labels.

4. METHOD

In this paper, we consider three hypotheses:
(H1) Learning from NMV data: An emotion recognition system trained with NMV data will outperform a system trained without NMV data, particularly improving test accuracy of NMV instances.
(H2) Multi-label approach for increased performance and 'human-like' interpretation: An emotion recognition system trained using multi-labeled data will achieve higher accuracy than a system trained with traditional one-hot-labeled data, particularly for non-prototypical data. Furthermore, the multi-label outputs enable us to build a more "human-like" emotion recognition system, that can provide an interpretable description of emotion distribution.
(H3) 5-class emotion classification: A multi-label approach will enable a more accurate classification of emotions that typically have greater noise, such as frustration; emotions that have been often discarded in previous systems [4, 22].

To test these hypotheses, we compare the emotion recognition performance of our proposed multi-label approach to that of a baseline (average cross-entropy error and accuracy). As a baseline we

use the traditional one-hot-label approach that assigns a single label to training and testing instances. In evaluating our method, we define expressions as prototypical (i.e., total agreement), non-prototypical (majority agreement), and NMV (no agreement). For both baseline and our proposed method, we analyzed the performance when the system was trained on (i) all data, (ii) prototype-only, (iii) non-prototype-only, and (iii) NMV-only.

We perform leave-one-subject-out cross validation across all of our experiments. We use paired t-tests for significance tests, and claim significance when $p < 0.001$.

4.1. Ground Truth for Human-Like Emotion

Our proposal is to use activation-based representation of emotion labels in neural networks to generate emotion soft labels. This approach is inspired by the pioneering work of Mower et al. [4], which presented SVM output-based vector representation of estimated emotion at the inference stage. We propose to go a step further and directly use the emotion distribution produced by human annotators in both training and inference. We use annotator distribution over different emotion classes. For instance, if six annotators label an utterance as "happy," two annotators label it "neutral," one annotator labels it "sad," and one annotator labels it "frustration" (out of the five classes of "angry," "happy," "neutral," "sad", and "frustration"), we then define and represent the ground truth of the utterance as a four-dimensional representation of $\{0, 0.6, 0.2, 0.1, 0.1\}$. To generalize this example to N classes of emotion, each unit of emotional data is represented as an N-dimensional vector, where each dimension of the vector is the fraction of annotators who chose that emotion class. The benefit of this representation is that we can further represent the NMV data's emotion labels, rather than discarding them as in current practices in traditional emotion recognition systems [20].

The number of utterances that are labeled as surprise, fear and disgust is small because these emotions were less often produced in the IEMOCAP data which are not read speech (that has specific target emotion per sentence) but produced with hypothetical emotion-provoking situations. Previous studies used 4-class classification of anger, happy, neutral, sad emotion classes for performance measure, mainly due to this class imbalance. In our experiments, we include data for frustration, surprise, fear, and disgust.

4.2. Multi-Label Emotion Learning and Inference

Once we obtain the soft-label representation of emotion ground truth, we use a multi-label learning method that can use the multi-label similarity between the estimated emotion output and emotion ground truth. To that end, we modify ELM for soft-multi-labeling regression and classification experiments. ELM is computationally efficient in both training and testing, and has a tendency to reach a global optimum. We choose ELM because of its computational efficiency and the flexibility to use multi-label outputs in the output layer of the network. ELM is a neural network where the hidden layer is not required to be neuron alike [23]. The output activation function of ELM is formulated as: $f_l(\mathbf{x}) = \sum_{i=1}^{L} \boldsymbol{\beta}_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta}$, where L is the number of hidden nodes of the neural network. Also, we define the feature mapping as $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), ...h_L(\mathbf{x})]$, where $h_i(\mathbf{x})$ is a nonlinear piecewise continuous function that satisfies the ELM universal approximation capability theorems, defined in [24], such as sigmoid, gaussian, hard limit, or cosine functions. We define the output weight vector from the hidden nodes to the output nodes as $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ...\boldsymbol{\beta}_L]^T$. In this work, we use the same $L = 500$, a third of original feature dimensions, across all experiments to ensure the consistency in model complexity.

ELM has two main stages for training: (i) random initialization of hidden nodes and (ii) learning the weights between the hidden nodes and the output nodes. For the first stage, the initialization can be chosen as any mapping function $\mathbf{h}(\mathbf{x})$. In the current work, we use a sigmoid function to capture nonlinear relationships of inputs. For the second stage, ELM learns the weights by minimizing the training error. Here, we use cross entropy for emotion recognition tasks. ELM also learns the weights that minimize the norm of the output weights:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad ||\mathbf{H}\boldsymbol{\beta} - T||^2 \quad \text{and} \quad ||\boldsymbol{\beta}||,$$

where $\mathbf{H}$ is the hidden-layer output matrix for L hidden nodes.

4.3. Cross Entropy for Categorical Emotion Variations

To directly compare human and machine emotion label distributions, we use a cross-entropy based performance metric. Cross entropy has been widely used for measuring the performance of dimensional emotion recognition [25]. Our work differs from previous work in that we use cross entropy to investigate the effectiveness of soft-multi-labeling that captures the full categorical emotion distribution over human annotators. In this paper, we provide both cross entropy (for regression task) and accuracy (for classification task) as performance measure. In particular, the cross-entropy results can provide insight about the efficacy of our method compared to baseline and the interpretation of human-like emotion recognition. Cross entropy is calculated as follows: Cross Entropy $= -EMO_{\text{true}}. * \log(EMO_{\text{est}})$, where EMO_{true} and EMO_{est} are multi-dimensional vectors of true and estimated emotion distributions, respectively, and n is the number of utterances. '$.*$' denotes the element-wise multiplication. To compare our results using traditional accuracy measure, we assign a single emotion label to the multi-dimensional label based on the maximum emotion component as in previous work [4].

5. RESULTS AND DISCUSSION

To address our hypotheses **(H1)–(H3)**, we present the results of two sets of experiments: cross-entropy results ('CE') and 5-class emotion classification accuracy results ('Acc') in Table 1. The cross entropy results the best account of the difference between estimated and true emotion distribution; while the accuracy-based results demonstrate that the proposed approach allows the use of emotions that are often discarded in previous systems due to greater label noise (e.g., frustration). Each set of results are divided into five categories for training utterance types and four categories for test utterance types: general results ('All'), prototypical utterances ('Prot'), non-prototypical utterances ('Nonprot'), and NMV utterances ('NMV'). This allows us to investigate the respective efficacy of our proposed emotion recognition system when different prototypical types are considered in training and testing.

First of all, CE results address **(H1)** and **(H2)**. To address **(H1)**, we train our emotion recognition systems using prototypical-only ('P'), non-prototypical-only ('NP'), NMV-only ('NMV'), combined prototypical and non-prototypical ('P+NP'), and all ('All') utterances. We also compare the cross-entropy results of test utterances for proposed and baseline methods, to address **(H2)**.

Overall, our proposed method always significantly outperformed the baseline, supporting **(H2)**. The improved performance using our

Table 1: Our proposed cross-entropy results ('CE') and traditional unweighted accuracy results ('Acc'), averaged over ten test speakers of IEMOCAP. **Trained Data** rows represent what prototypcality type is used for training the system: Prot-only, Nonprot-only, NMV-only, Prot+Nonprot, and All (Prot+Nonprot+NMV) training utterance types. **Test Utterances** columns represent what test utterances are used to report the results: All, Prot, Nonprot, and NMV test utterances. "[*]" indicates the statistical significance levels ($p < 0.001$) between our proposed method and baseline.

Trained Data	Method	Test Utterances							
		All		Prot		Nonprot		NMV	
		CE	Acc	CE	Acc	CE	Acc	CE	Acc
Prot	Proposed	0.6779[*]	41.96	0.4996[*]	48.59	0.7823[*]	40.20	0.8012[*]	27.44
	Baseline	1.3514	42.54	0.9175	50.33	1.5430	40.80	1.8000	27.03
Nonprot	Proposed	0.4609[*]	40.22	0.3603[*]	44.59	0.5143[*]	37.20	0.5421[*]	32.58
	Baseline	1.0914	40.68	0.7960	46.37	1.2256	36.76	1.4015	32.84
NMV	Proposed	0.6637[*]	23.01	0.5925[*]	22.67	0.6995[*]	22.64	0.7275[*]	21.39
	Baseline	3.5192	N/A	3.5158	N/A	3.5193	N/A	3.5218	N/A
Prot+Nonprot	Proposed	0.5115[*]	44.58	0.3877[*]	51.37	0.5849[*]	42.80	0.5820[*]	28.99
	Baseline	1.0007	43.58	0.7190	51.93	1.1260	40.54	1.2983	27.93
All	Proposed	0.4837[*]	44.45	0.3723[*]	52.76	0.5531[*]	40.70	0.5478[*]	28.62
	Baseline	1.0362	N/A	0.7153	N/A	1.1934	N/A	1.3420	N/A

multi-label approach, across all training and testing environments, indicates the importance of our new representation and performance metric for emotion recognition. When comparing the systems trained with NMV (**Trained Data** are NMV or All) vs. without NMV (**Trained Data** are P, NP, or P+NP), the results significantly improve from baseline to proposed methods, particularly when NMV-only data are used for training (improvement from 3.5192 to 0.6637 when 'all' test utterances are used). The baseline methods work particularly poorly when only NMV utterances are used (all greater than 3.50). However, our proposed method achieves significantly higher cross-entropy results in the same experiment, demonstrating the importance of a multi-label approach, particularly for NMV utterances. Hence, we conclude that **(H1)** is supported.

Next, Acc results show the 5-class emotion classification accuracy after we transfer the multi-label outputs to discrete emotion classes. For our proposed method, we use the maximum component over the emotion dimensions of multi-label outputs for each utterance to assign a single label. The classification results address **(H3)**, and will provide insight into how our multi-label approach can be extended to categorical classification, and whether our new approach enables us to accurately classify emotions that are often discarded in previous systems due to greater label noise (e.g., frustration). As in Table 1, we train our systems using different prototypical types to address **(H1)**, and compare the accuracy results to address **(H2)**.

In general, the 5-class classification of anger, happy, neutral, sad, frustration achieves up to 47.17% in weighted accuracy, 2.26% higher than the highest accuracy of the baseline method (45.97%). Both proposed and baseline methods achieve much higher accuracy than a chance (20%). To the best of our knowledge, our 5-class classification is new for this benchmark IEMOCAP dataset. Our proposed multi-label approach enables us to accurately classify emotion labels with more noise, such as frustration, which have been often discarded in previous systems, supporting **(H3)**. Also, the proposed method trained using all utterances (UW 44.45%, W 46.38%) achieves significantly higher accuracy than the same method trained using only prototypical utterances (UW 41.95%, W 44.05%), with 2.50% ($p = 0.036$) and 2.33% ($p = 0.043$) for UW and W accuracy, respectively. This demonstrates that it is beneficial to use non-prototypical and NMV utterances in training, supporting **(H1)**.

Finally, the proposed method outperforms baseline when prototypical and non-prototypical utterances are used for training, as in the traditional emotion recognition benchmark. When all test utterances are evaluated, UW accuracies are slightly higher for the proposed method than baseline, 44.58% to 43.58% (1.00% increase, not significant). The proposed method also achieves slightly higher W accuracy than baseline, 47.17% to 45.97% (1.20% increase, not significant). The non-prototypical and NMV test utterances also achieve higher performance than baseline when our proposed method is used compared to baseline, whereas the prototypical test utterances achieve similar accuracy for the proposed method and baseline. This may indicate that our proposed multi-label method particularly helps learning from emotionally ambiguous data, i.e., non-prototypical and NMV utterances. Given the prevalence of such subtle and ambiguous emotion expressions in real-world applications, our method is promising. Hence, the results support **(H3)**.

6. CONCLUSION

In this work, we present a new representation, learning, and inference method that utilizes multi-label approach for emotion recognition. Unlike traditional emotion recognition systems that either use one-hot labeling method or discard inconsistently labeled data (NMV data), our proposed method can use the full data that include prototypical, non-prototypical, and NMV data. The key novelty of this work comes from our investigation of the hypotheses **(H1)**-**(H3)**.

This research points the way to unlocking the potential of big multimedia data with an approach can exploit full human annotator data resources, even those with no majority agreement, which are often rejected from current emotion recognition systems. We propose that using the full data set will improve the learning of universal emotion expression patterns across different users and emotion classes. We anticipate emotion recognition systems trained using this approach will be able to generate more accurate and robust emotion inference for a new user or emotion class.

7. REFERENCES

[1] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

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

[3] Mohit Shah, Chaitali Chakrabarti, and Andreas Spanias, "A multi-modal approach to emotion recognition using undirected topic models," in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 754–757.

[4] Emily Mower, Maja J Mataric, and Shrikanth Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[5] Reza Lotfian and Carlos Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classificatio," in *Affective Computing and Intelligent Interaction (ACII)*, October 2017, pp. 415–420.

[6] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 566–570.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[8] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "Paralinguistics in speech and language–state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[9] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, no. 5, pp. 1203, 2013.

[10] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[11] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 827–834.

[12] Andrea Kleinsmith and Nadia Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 15–33, 2013.

[13] Chee Seng Chong, Jeesun Kim, and Chris Davis, "Visual vs. auditory emotion information: how language and culture affect our bias towards the different modalities.," in *AVSP*, 2015, pp. 46–51.

[14] Simone Simonetti, Jeesun Kim, and Chris Davis, "Auditory, visual, and auditory-visual spoken emotion recognition in young and old adults," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015), 10-14 August 2015, Glasgow, Scotland, UK*, 2015.

[15] Chee Seng Chong, Jeesun Kim, and Chris Davis, "Exploring acoustic differences between Cantonese (tonal) and English (non-tonal) spoken expressions of emotions," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] Laurence Devillers, Laurence Vidrascu, and Lori Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[17] Chee Seng Chong, Jeesun Kim, and Chris Davis, "The sound of disgust: How facial expression may influence speech production.," in *INTERSPEECH*, 2016, pp. 37–41.

[18] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–8.

[19] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*. Citeseer, 2011.

[20] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3687–3691.

[21] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, "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, 2017, pp. 890–897.

[22] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, March 2010, pp. 2462–2465.

[23] Gao Huang, Guang Bin Huang, Shiji Song, and Keyou You, "Trends in extreme learning machines: A review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[24] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, vol. 42, no. 2, pp. 513–29, 2012.

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