Speech Emotion Classification from Affective Dimensions: Limitation and Advantage

Meysam Shamsi

Speech Emotion Classification from Affective Dimensions: Limitation and Advantage

Meysam Shamsi
LIUM
Le Mans University
Le Mans, France
meysam.shamsi@univ-lemans.fr

***Abstract*—In affective computing, two main paradigms are used to represent emotion: categorical representation and dimensional description in continuous space. Therefore, the emotion recognition task can be treated as a classification or regression. The main aim of this study is to investigate the relation between these two representations and propose a classification pipeline that uses only dimensional annotation. Our approach contains a neural regressor which predicts a vector of arousal, valence and dominance values for a given speech segment. This vector can be interpreted as an emotional category using a mapping algorithm. We investigate the performances of a neural network architectures, and three mapping algorithms on two corpora. Our study shows the limitation and an advantage of the emotion classification via regression approach.**

***Index Terms*—Speech emotion recognition, Emotion representation, Classification, Regression**

I. Introduction

The importance of extracting the paralinguistic information from speech has led the research community into Speech Emotion Recognition (SER). One specificity of this field is that the definition of emotions is ambiguous [1], [2]. Consequently, there is no consensus on emotion representation and annotation. Two main theories of emotions are used in affective computing. Emotions can be described with categorical labels mostly based on Ekman representations [3] or emotional dimensions such as arousal (or activation), valence, dominance (AVD) [4].

These two representations have merits and disadvantages. On the one hand, the use of categorical labels for describing emotional states is usually more understandable for the public, as the words directly refer to common sense [5]. However, it makes the representation of emotional states limited to certain categories, which are usually prototypical and do not cover all diversity of human emotions. On the other hand, using affective dimensions can precisely assign an emotional state to a point in a continuous space. Moreover, the categorical labels can be an interpretation in dimensional space which can be personalized by a human perception.

In the following, we detail the advantages of dimensional representation favor compared with categorical representation from a machine learning point of view [6]. A supervised machine learning model typically uses ground truth annotation. But due to the complexity of human emotions, there is always a disagreement on the perceived emotions and then annotations. So usually the assigned values by annotators would be aggregated to generate one single annotation per input. One of the main differences between categorical representation, which makes emotion recognition a classification task, and the dimensional representation, which makes emotion recognition a regression task, is the conserved information after aggregating annotations. The most commonly used method for the aggregation is getting the majority vote of the annotator's opinion to have a hard label. Although, some studies such as [7], [8] followed a soft labeling approach to deal with the labeling complexity and ambiguity. For example, in the standard protocols of *IEMOCAP* dataset [9] and *MSP-Podcast* corpus [10], the samples where annotators disagree are discarded.

The most common approach for encoding emotional categories is one hot vector, which ignores the relation or distance between emotions. For example, anger can be very close to irritation, frustration, and rage, and they are usually perceived or expressed in similar situations. On the contrary, dimensional annotation that provides a distributed representation can keep the intra and inter categories distance information. This continuous representation helps to overcome the limitation of discrete labels.

In this paper, the coherence between these two annotation types is studied. Moreover, the capacity of classification models without using categorical annotations is investigated. This approach can show the advantages of dimensional annotation and representation. A similar conceptual theory, without any experimental reports, is discussed in [2].

II. Overview

To investigate the relation between the categorical and dimensional representation, we propose to study the capacity of classification based on affective dimensions. When the number of dimensions in most of the studies is three (AVD), the consistency of these dimensions to a set of categorical emotions is questioned. The idea is based on an assumption that the categorical labels of samples can be predicted based on the dimensional values, as long as these two representations are coherent and sufficient. Some studies such as [6], [11] support the hypothesis. In [11], it has been observed that a

This study has been realized under the project PULSAR supported by the Region of Pays de la Loire, France (grant agreement No 2022-09747).

model for the prediction of arousal and valence values can be useful to detect categorical emotions.

One of the main advantage of training on emotion recognition on dimensional space is the use of distributed representation feature, which contains between and within class distances, can inject additional information into the model. Moreover, a trained regression model on affective dimensions can be developed for a classification task as well. In this case, based on the definition of categorical labels in dimensional space, the output of the regression model can be mapped to emotional vocabularies. It means the parallel annotations, categorical and dimensional, of a dataset would not be necessary. Only dimensional annotations and a mapping definition would be enough.

In the following, the differences between conventional classification and proposed classification via regression will be presented.

1) Classification: A classification model would be trained to predict the categorical annotation of a given audio segment. In a neural classifier, the output layer would be limited to the number of targeted classes in the training set. They can only profit from the annotated samples in these categories for training as a supervised problem. Therefor, the capacity of these models are limited to the predefined target categories.

2) Classification via Regression: We propose to build a regression model which predicts a vector of values in the continuous space as the representation of the emotional state. The output of a trained regressor can be fed to a mapping model to transform into emotional labels. The training of mapping model defines the categorical emotions in the dimensional space. Using a similar architecture for the classifier and regressor provides the chance of comparing two approaches with the almost same capacity of learning (number of network weights).

For the mapping from dimensional to categorical representation, three algorithms are proposed; Gaussian classifier (*Gaussian*), K-Nearest Neighbors (*KNN*) (empirically optimized K=50) and Tow-Layer Perceptrons, 5*5, (*2LP*). These models are constructed to predict the categorical labels based on dimensional values. Using a classical machine learning model, do not require a lot of samples with parallel annotations.

III. Experiment

In this section, the data, the classification via regressor systems and results will be presented.

A. Data

To examine the idea of emotion classification from affective dimensions two common corpora, *IEMOCAP* [9] and *MSP-Podcast* [10], are employed which contain both annotation types. As it has been suggested by [12], only the 4 main emotions (*Neutral*, *Happy*, *Sad* and *Angry*) from *IEMOCAP* for the rest of this study. The same emotion categories have been selected from *MSP-Podcast*. The *IEMOCAP* dataset in this study is based on 5-fold cross-validation under a leave-one-session-out (LOSO) protocol. The original partitioning of the *MSP-Podcast* dataset version 1.8 [10] is respected, and evaluations are based on the test partition. In this study, the affective dimensions (AVD) in the two mentioned corpora are normalized to the range of -1 to 1.

In order to have an upper bound performance of classification based on three-dimensional values, the result of mentioned mapping algorithms on reference annotation (ground truth AVD) of *IEMOCAP* and *MSP-Podcast* is evaluated.

B. System

By emerging of pretrained neural network models and their decent performance on different tasks, particularly for emotion recognition [12], [13], we propose to use pretrained *wav2vec2* [14] model. The *wav2vec 2.0 base* model, pre-trained on Librispeech (960 hours of speech) and is fine-tuned in our training process. The *Wav2vec2* encoder is joined with a downstream head. The mean of the *Wav2vec2* encoder's output over time is passed through two layers of a Transformer encoder [15] with 2 attention heads, and then it is followed by ReLU activation function and linear layer. In terms of model's capacity, this model contains 99.9M trainable parameters.

There is small different between classifier and regressor network. The classification model employs a softmax layer as output and cross-entropy as its loss function. The regression model is similarly designed, with some modifications. Its output layer is adapted to the number of dimension (3 as A,V,D), the output layer is modified to the linear layer, and Concordance Correlation Coefficient (CCC) [16] is used as its loss function. Our CCC would be a value between 0 and +1 and optimized to be higher, so the loss value is defined as one minus the mean of three dimensions' CCC. Adam optimization algorithm with a learning rate of 2e-5 is used for classifier and regressor. The training process is continued to the maximum of 40 epochs, with an early stopping of 5 epochs when the loss of validation set would not improve.

The mapping models are trained using the same training samples as the classifier/regressor model. Finally, the classification via regressor model will be evaluated on the corresponding test set.

The regression model can use all available information in a corpus (samples are not limited to certain categories) for training. Moreover, the categorical labels can be an interpretation of the model's output, which means it is possible to have a various categorical label for a given audio according to a personalized perception. This interpretation or mapping from dimensional space to categorical labels can be simply done by defining the emotional classes (such as Gaussian mapping) in continuous space in the posterior.

C. Results

The results of classification model, the regression performances, and the ability classification using three affective dimensions (AVD) for two corpora is reported in the Table I. In this table, the first lines refer to the performance of classification via regression when the ground truth 3-dimensional values are used as its input of mapping. The second line compares the

TABLE I

CLASSIFICATION, REGRESSION AND CLASSIFICATION VIA REGRESSION RESULTS. THE FIRST LINE OF EACH CORPORA (*) IS THE RESULT OF CLASSIFICATION WITH AN IDEAL REGRESSION (USING GROUND TRUTH AVD VALUES FOR MAPPING). THE SECOND LINE IS THE RESULT OF CLASSIFIER/REGRESSOR AND CLASSIFICATION VIA REGRESSOR.

Classification (UAR,WAR)	Regression CCC (A, V, D)	Classification via Regression (UAR,WAR) 2LP	KNN	Gaussian
IEMOCAP				
Ground truth* (71.6, 69.5)	(1.00, 1.00, 1.00) (0.64, 0.73, 0.57)	(70.5, 71.1) (59.6, 58.8)	(68.2, 69.0) (57.9, 57.8)	(70.8, 71.0) (60.4, 58.6)
MSP Podcast				
Ground truth* (49.5, 67.7)	(1.00, 1.00, 1.00) (0.60, 0.41, 0.50)	(64.5, 75.5) (42.3, 65.6)	(62.5, 74.4) (42.7, 64.7)	(69.5, 73.0) (49.1, 63.6)

performance of the classifier model with the cascade pipeline of the regressor, followed by mapping algorithms.

The first line of the table I shows the limitation of classification via regression using the three affective dimensions. It shows that a perfect regressor can map only less than 72% (resp. 75%) of samples from AVD space to the four classes of emotion in the *IEMOCAP* (resp. MSP-Podcast). Comparing the performance of the "classification" and "classification via regression", the impact of regression error on final classification can be observed. It indicates that even by a state-of-the-art regressor, the performance of classification via regressor will degrade around 10% in the IEMOCAP (from **U**nweighted-**A**verage-**R**ecall:72%, **W**eighted-**A**verage-**R**ecall:69% to UAR:60%, WAR:59% with Gaussian mapping).

This degradation of classification performance is observed even though the regressor used more training samples. The regressor training samples are not limited to the four targeted categories. In an equivalent scenario, when the regressor is trained with only samples in four emotions on *IEMOCAP* which means using 5531 samples instead of 7532, the performance of classifier via regressor degrade drastically (from UAR=60%, WAR=59% to UAR=47%, WAR=43% for Gaussian mapping). Although, the use of other all samples in the *MSP-Podcast* does not change the performance of the classifier via regressor. One explanation may be the fact that the number of training samples in *IEMOCAP* is not enough, and samples from other classes can help to train a better regressor. Based on this observation, it can be proposed to follow classification via regression for cases with limited number of samples annotated with target categories. It means all collected (or recorded) samples can be used for the training no matter the label of annotators is, which does not waste annotators' efforts.

IV. ADVANTAGE AND LIMITATION OF CLASSIFICATION VIA REGRESSION APPROACH

In previous section, the limitation of classification via regression approach has been revealed by the degradation of its performance compare with classical classification model. This observation indicates that the three-dimensional representation of arousal, valence, dominance are not enough to categorize the emotion in even 4 basic categories. In this section, we analyze this limitation and name an advantage of classification via regression approach.

A. Classification in lower and higher dimensional space

An ablation study on different affective dimensions with mapping algorithm from ground truth annotation to categorical emotions is done. The results show that the valence is the most discriminative attribute (followed by arousal and dominance) to predict samples' labels in both corpora. Using only the valence dimension results UAR:61%, WAR:62% and using valence and arousal results UAR:69%, WAR:70%. Although the performance of mapping from these two dimension is slightly lower than using 3 dimensions, it shows that the dominance values are not necessary for the mapping.

A complementary experiment has confirmed that AVD values are not enough to distinguish 4 targeted categories. By following a probing approach, we trained a new mapping model using the output of *Trans* layer (before the linear layer) as the embedding vector to map from higher dimensional space, 768 values instead of AVD, to categorical labels. Although in this case, the predicted values by regressor are used as the input for training mapping model instead of the reference values, our experiment on the *IEMOCAP* shows an improvement of performance to UAR=65%, WAR=63% (compare to UAR=60%, WAR=59%). It can be concluded that by compressing the information to AVD values in regressor, we would lose at least some of the useful information for classification in 4 classes.

B. Using classification via regressor for new emotion

Although following classification via regression approach can degrade the performance of classification, there are still some advantages. Following classification via regressor approach can provide the chance of extending to new emotional categories. It is only needed to provide the position of the categorical label in the AVD space, without annotating new data. To show this potential, we use the proposed model trained on samples from 4 emotions (Natural, Happy, Angry, Sad) in *IEMOCAP* as a regressor. Then the Gaussian mapping trained on 5 class is employed. In this case, the *Frustration* category is added to test set, which change the problem from four to five classes. In total 1849 samples from *Frustration* category are evaluated in the 5 folds cross-validation. The model is able to recognize the *Frustration* class with a precision of 40.3% and recall of 33.2%, although the initial regressor has not seen any sample from this class during training.

V. DISCUSSION

The proposed approach does not necessarily apply to parallel annotation corpus, particularly by using Gaussian mapping (when only a mean and variance is enough to simulate the Gaussian distribution). The idea of using the distance to the centroid of each class of emotion as the mapping algorithm in a text context in [17] is very close to using Gaussian mapping in our approach. Although this study has shown the limitation of this approach with a state-of-the-art classifier/regressor, the benefits from dimensional annotation for classification are not new in the literature. For example, [18]–[20] suggested

employing both annotations in a multitask approach to improve classification results. Although these studies showed that dimensional information can be helpful for classification, our study has indicated the power of only dimensional information in the combined paradigm.

One perspective of classification via regressor is personalizing the regressor to the definition of emotional categories in the dimensional space with small number of samples per subject. While the perception of annotators can be predicted by a general classifier, which trained on samples with aggregation of several annotators label. Here, we propose to use a general regressor to predict the AVD values and implement a personalized mapping model for each annotator to adapt the classification to the emotion perception of each annotator.

VI. Conclusions

In this work, we investigated the relationship between dimensional representation and categorical labeling of emotions. We proposed to consider the speech emotion recognition task as a regression problem whose output can be interpreted as categorical emotions by using a mapping algorithm. We compared the performance of these two approaches by a state-of-the-art regressor/classifier, and three mapping algorithms for transforming the continuous value in the AVD space to categorical labels. Our experiment on two different corpora has shown degradation of performance compared to a traditional classifier that profits from categorical labeled data. It has been observed that the AVD dimensions are not necessary and enough for classification. Beside the limitation of the "classification via regression" approach, advantages of this approach have been presented, such as extending the classification to new categories and personalized emotion classification with limited number of parallel samples.

Ethical Impact Statement

While the proposed approach has the potential to enhance communication by enabling personalized emotion recognition, it is important to acknowledge the potential risks and negative implications associated. The ability to monitor and manipulate individuals through speech recording devices raises serious concerns about user privacy and consent. The author strongly emphasize that the misuse of this technology is not advocated and should be strictly prohibited. Additionally, this work highlights the limitations of current common parallel frameworks for emotion representation, emphasizing the need for caution and thorough interpretation in real-world applications where mapping affective dimensions to categorical representations can be prone to failure due to insufficient coherent.

Acknowledgment

My heartfelt gratitude to Dr. Marie Tahon, my esteemed colleague, for her invaluable contributions and unwavering support throughout this research study. This study has been realized under the project PULSAR supported by the Region of Pays de la Loire, France (grant agreement No 2022-09747).

References

[1] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[2] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation."

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[5] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 827–834.

[6] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.

[7] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," *Proc. Interspeech*, pp. 951–955, 2018.

[8] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4964–4968.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[11] V. Kowtha, V. Mitra, C. Bartels, E. Marchi, S. Booker, W. Caruso, S. Kajarekar, and D. Naik, "Detecting emotion primitives from speech and their use in discerning categorical emotions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7164–7168.

[12] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, oct 2020, pp. 38–45.

[16] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio." in *IJCAI*, vol. 2016, 2016, pp. 2196–2202.

[17] R. A. Calvo and S. Mac Kim, "Emotions in text: dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[18] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on affective computing*, vol. 8, no. 1, pp. 3–14, 2015.

[19] R. Cai, K. Guo, B. Xu, X. Yang, and Z. Zhang, "Meta multi-task learning for speech emotion recognition." in *Proc. Interspeech*, 2020, pp. 3336–3340.

[20] R. Sharma, H. Dhamyal, B. Raj, and R. Singh, "Unifying the discrete and continuous emotion labels for speech emotion recognition," *arXiv preprint arXiv:2210.16642*, 2022.