ABSTRACT

There are two paradigms of emotion representation, categorical labeling 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. The proposed approach contains a regressor model which is trained to predict a vector of continuous values in dimensional representation for given speech audio. The output of this model can be interpreted as an emotional category using a mapping algorithm. We investigated the performances of a combination of three feature extractors, three neural network architectures, and three mapping algorithms on two different corpora. Our study shows the advantages and limitations of the classification via regression approach.

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

1. INTRODUCTION

The importance of extracting the paralinguistic information from speech has led the research community into Speech Emotion Recognition (SER). But the definition of emotions is ambiguous [1]. Consequently, there is no consensus on emotion representation and annotation. The two main emotional theories used in computer science are the followings: emotions can be described with categorical labels mostly based on Ekman representations [2] or emotional dimensions such as arousal (or activation), valence, dominance (AVD) [3] to precise the emotional state.

These two representations have merits and disadvantages. Usually, using categorical labels for describing emotional states would be more understandable for the public [4]. But it makes the representation of emotional states limited to certain categories, which may not cover all human emotions. On the other hand, using continuous value can precisely assign the emotional state to a point in dimensional space, which is less close to human language.

From a machine learning point of view, the advantages of dimensional representation are in favor compared with categorical representation [5]. In the following, the benefits of using continuous values for emotions in dimensional space are detailed. 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 value of 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 aggregation of 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 [6, 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 with disagreement of annotators 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 which provides a Distributed Representation can keep the intra and inter categories distance information. This continuous representation helps to break out the limitation of discrete labels.

It has been claimed by [10] that the perceived and expressed frequency of categorical emotions are not the same in the real life. In this case, an imbalanced classification problem would be faced [11]. As the dispersion of emotional labels in different corpus has been shown in [10], the impact of the non-homogeneous frequency of class would be less important when the dimensional representation is employed.

Last but not least, the dimensional representation has its application where a continuous value is needed more than only a category. In the task of continuous emotion recognition [12, 13], when a sequence of predictions over time is the goal, the dimensional approach helps to smooth the transition and take into account temporal dependencies. All these conveniences emphasize on the capacity of dimen-
sional representation of emotional states.

In this paper, the coherence of these two annotation types in two common used corpora, IEMOCAP [9] and MSP-Podcast [10], 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, which is theoretically and empirically supported in [5].

2. THE CHALLENGE OF EMOTION ANNOTATION

One of the main challenges of the emotional dataset creation and annotating human emotions is the annotator agreement. As an example, two common datasets (IEMOCAP and MSP-Podcast) are compared in the following.

Using hard labeling of categorical annotation, 19.4% of samples in MSP-Podcast corpus and 25% of samples in IEMOCAP could not get an agreed annotation from evaluators. The reliabilities of agreement in these two corpora are not high. The Fleiss’ Kappa of categorical annotations in the MSP-Podcast is only 0.23 and in the IEMOCAP is 0.48. A perceptual test in [14] showed that the human performance for emotion recognition of four main classes in IEMOCAP is only 69% overall accuracy.

The evaluators’ agreement on dimensional annotation is not high as well, except for the Arousal on the IEMOCAP. Table 1 shows the inter-evaluator reliability (Krippendorff’s alpha coefficient).

| Corpus        | Arousal | Valance | Dominance |
|---------------|---------|---------|-----------|
| IEMOCAP       | 0.68    | 0.30    | 0.27      |
| MSP-Podcast   | 0.27    | 0.33    | 0.21      |

To find a relation between two types of annotation, the density of samples with the same emotional label has been mapped in two main dimensions, arousal, and valence (See Figure 1). In order to have comparable experimental results and as it has been suggested by [15] [16], only the 4 main emotions (Neutral, Happy, Sad and Angry) from IEMOCAP and 5 main emotions (Neutral, Happy, Sad, Angry and Disgust) from MSP-Podcast have been used in the rest of this study.

Figure 1 confirms that the annotator in the different corpus has a different definition or perception of emotions. While the Angry class in the IEMOCAP has been evaluated with higher valence and lower arousal than Neutral, it has been assigned to a lower arousal and higher valence values than Neutral in the MSP-Podcast. This conflict indicates the ambiguity of emotions’ definition, which makes the emotion classification challenging in the cross corpus cases.

3. CLASSIFICATION VIA REgressor

In order to investigate the relation between these categorical and dimensional annotations, the main goal of this study is to evaluate the ability of classification, based on the continuous values in the dimensional space. Figure 2 demonstrates two pipelines for recognizing a class of emotion, a one-step classification on the left and classification via regression on the right. The main idea is that the categorical label of samples can be predicted based on the dimensional values, as long as the annotations are consistence. Some studies such as [5] [17] support the hypothesis. In [17], it has been observed that a model for the prediction of arousal and valence values can be useful to detect categorical emotions.
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 annotation and a mapping definition would be enough to have a prediction in two representations. This potential is the main focus of this study’s experiments.

For the mapping, three algorithms are proposed; Gaussian classifier (Gaussian), K-Nearest Neighbors (KNN) (optimized K=50) and Tow Layer Perceptrons, 5*5, (2LP). These models are constructed to predict the categorical label based on dimensional values. In order to have an upper bound of classification performance based on dimensional values in three dimensions, the result of these mapping algorithms on reference annotation of IEMOCAP and MSP-Podcast is evaluated. The results show that the valance is the most discriminative attribute (followed by arousal and dominance) to predict samples’ labels in both corpora. The combination of three attributes achieves the highest accuracy. Table 2 compares three dummy models with proposed mapping algorithms which predicts the class label based on the ground truth values of arousal, valance and dominance. The dummy models are Random labels, Prob. Random which generates labels randomly with respecting the probability of each class, Major Class which always generates the label of the most frequent class.

It is important to compare the performance of a Machine Learning model with these dummy models, especially in an imbalanced dataset. Based on the application, the evaluation metric can be different. When the task is emotion recognition in a real-life situation, the weighted performance can be more important. On the other hand, when the task is only to distinguish between different emotions, the unweighted performance can be applied. For this purpose, it is decided to report unweighted recall (UR) and weighted recall (WR). Needless to mention, contrary to classification models, the imbalanced data can cause less impact on regression models’ performance. As it is noted in Table 2, the weighted recall (WR) of selecting the only major class in the MSP-Podcast is 50.6%.

The 2LP archives the best performance for the prediction of the emotional labels based on three attributes in the IEMO CAP (with 4 classes). Although, it shows that a perfect regressor can map only 72.7% of samples from AVD space to the classes of emotion. The mapping algorithms on MSP-Podcast (with 5 classes) have different performances. It reveals the limitation of the proposed idea of using regression models for the classification task.

It should be mentioned that the objective of this study is not to outperform the classification task, but only to compare the performance of the two approaches. The main idea of this paper is to study the performance of a regression model on AVD space, which can be used for classification as well. The regression model can use all available information in a corpus (samples are not limited to certain categories) for training. Moreover, the class label can be an interpretation of the model’s output, which means it is not necessary to have a categorical annotation of samples. This interpretation or

---

**Fig. 1.** Density of the main emotion categories on normalized Arousal and Valence space. Non homogeneity of densities in IEMOCAP is due to the limited number of annotators and scoring steps.
mapping from dimensional space to categorical labels can be simply done by defining the emotional classes in continuous space in the posterior.

4. EXPERIMENT DESIGN

We propose to build a regression model to predict 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 models defines the categorical emotions in dimensional space. In our experiment, they are constructed using the training set of the corresponding corpus. The performance of classification via the regressor model (the right pipeline in the Figure 2) is compared to a classification model (the left pipeline in the Figure 2). The main aim is not to outperform the state-of-the-art system, but only to propose a different view toward emotion recognition and its potential. Although, the state-of-the-art methods such as data augmentation [16], reject option [18], use fine-tuning or more fancy feature extractors [19] [20] can be applied to improve the performances.

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’s weights). The regression models employ a linear layer as output and MSE as their loss function. The classification models are similarly designed, with some modifications. Their output layer is adapted to the number of classes, the output layer is modified to the softmax, and cross-entropy is used as their loss function.

In order to create the classifiers and regressors, a combination of three feature extractors and three different neural network architectures has been tested.

4.1. Feature extractors

By emerging of pretrained neural network models and their decent performance on different tasks, in particular for emotion recognition [16] [19], we propose to use pretrained wav2vec2 [21] and wavLM [22] models as the feature extractor.

The wav2vec2 [21] used self-supervised learning on raw audio to transform it into an embedding representation. We used the wav2vec 2.0 base model, pre-trained on Librispeech (960 hours of speech) without fine-tuning. Same as wav2vec2, the wavLM Base+ [22] extracts universal speech representations. But unlike wav2vec2, the wavLM model has been trained on massive unlabeled speech data (94k hours of speech). By using these two pre-trained models, the raw audios are encoded into a sequence of embeddings with a window length of 25ms and a stride of 20ms.

Moreover, the Mel spectrogram (MelSpg), which showed a decent performance in [23], has been used. In order to have the same length of feature sequence, the same configuration of the sliding window has been set for MelSpg.

These three feature extractors generate a vector for each frame of given audio. To treat all audio signals with variable lengths, in the same way, a padding/truncation method has been applied to have a fixed length (500 frames from the first 6.9 sec) of features. 128 features per frame have been extracted using MelSpg, while using wav2vec2 and wavLM would return 512 features per frame.

4.2. Classifier/Regressor architectures

In order to predict the categorical emotions or the continuous values in AVD space, three different architecture has been designed (see Figure 3).

These downstream models are constructed with almost the same number of trainable parameters. The MLP model (see Figures 3 - a) constructed with 5 stacked fully connected layers. In order to generate a prediction, the last layer is a fully connected layer adapted corresponding to the number of classes in the classification task or the number of dimensions (three in AVD) in the regression task. While the MLP model is limited to the mean and variance of frames’ features, two other models can profit from the temporal information. The CNN model, inspired by [19], (see Figures 3 - b) is consecutive feed forward of 5 blocks of convolutional layer, batch normalization layer and max polling. The CNN-Trans model (see Figures 3 - c) is a parallel downstream architecture. On one side, it is 4 stacked convolutional layers. On the other side, the sequential information would be passed through an average pooling to reduce the dimensionality, afterward, it would be fed to two transformer encoder blocks with two heads of attention. Then the Transformer embedding will be concatenated to the CNN side. Finally, the result will pass through a fully connected layer to generate the prediction in the same way as the last layer of MLP.

4.3. Different datasets different challenges

Besides the challenges of different definitions and the consistency of two representation approach (mentioned in the section 3), the context of speech emotion recognition application plays a major role in the evaluation metrics. The context of application makes the designing of speech emotion corpora different. While IEMOCAP dataset [2] is recording of acted (some may call exaggerated) emotions with respecting the frequency of each class, the MSP-Podcast dataset [10] is extracted from recorded spontaneous speech without considering balancing the classes.

In this study, we use IEMOCAP and MSP-Podcast datasets to investigate our hypothesis on different contexts. The IEMOCAP dataset with 12h is smaller than the MSP-Podcast datasets with 27h of speech. In order to follow the previous studies [16] [19], four main emotions (Natural 30.9%, Happy
29.6%, Angry 19.9%, Sad 19.6%) in IEMOCAP has been selected which makes 5531 samples from all sessions. Contrary to this balanced distribution of emotions, for the MSP-Podcast datasets, like [15, 16], five main emotion classes (Natural 53.3%, Happy 29.3%, Angry 6.6%, Sad 5.4%, Disgust 5.3%) have been selected which contains 48754 samples.

The impact of imbalanced classes in the MSP-Podcast can be observed in the performance of dummy classifiers in table 2. The reported performance in this study is based on 5-fold cross-validation for the IEMOCAP dataset. The original partitioning of the MSP-Podcast dataset into train/dev/test is respected, and evaluations are based on the sum of test1 and test2 partitions [10].

5. RESULTS

The performance of regressor models (three feature types, three architectures) has been calculated based on the Concordance Correlation Coefficient (CCC) [24]. The output of the regressor model as AVD values has been mapped to categorical emotions using three algorithms mentioned in the section 3. The classification performance of these two approaches with different configurations on the IEMOCAP dataset and MSP-Podcast test set, are shown in tables 3 and 4. The results on two corpora show a low performance of classification if temporal information would not be taken into account. Using MelSpg in MLP architecture achieves a lower performance (almost equal to a dummy performance in MSP-Podcast) comparing with cases which temporal information have not been used in extracted features (wav2vec2 or wavLM) or in the model architectures (CNN or CNN-Trans).

In the IEMOCAP dataset (see table 3), the best classification results obtained by feeding the extracted embedding from wav2vec2 into MLP model with UR=63.16% and WR=61.81%. The same configuration achieved the best regression performance with the CCC of Arousal equal to 0.44, the CCC of Valence equal to 0.72, and the CCC of Dominance equal to 0.57. The best performance of classification via regression is obtained by applying Gaussian classifier as the mapping algorithm. Our experiment shows a lower performance of Gaussian Naive Bayes algorithm, which uses the frequency of classes in the training set as the prior probability. Comparing "classification" and "classification via regression" approaches shows that using dimensional annotation can successfully classify emotions with more than 80% proportional performance to using direct categorical annotation (UR=51.44% and WR=53.00%).

The best performances achieved using wav2vec2 or wavLM features and MLP or CNN-Trans models, which is decent compare to [16]. Same as IEMOCAP corpus, the best performance of regression and classification via regression obtained using wav2vec2 and MLP. We believe that there are several reasons for poor performance of models on the MSP-Podcast dataset, table 4. As it has been observed in the table 1 the annotator agreement is low in this corpus. Also, the emotions are not acted, which makes them more difficult to recognize. Moreover, most of the samples in this corpus are labeled as Natural, which makes classification more difficult in imbalanced problems (see poor UR in table 4).

The performance of classification via regression on these two corpora shows the potential and the limitation of this approach. Although following this approach can reduce the performance of classification, there are still several advantages. Using only dimensional annotation for training a regressor model can reduce the cost of preparing a dataset and using all available data. It also can prepare a model which can be used for the recognition of new emotion categories by only providing its definition in the AVD space. In order to show the potential of "classification via regression", we trained an MLP model using wav2vec2 features on samples from IEMOCAP with 4 emotions (Natural, Happy, Angry, Sad) as a regressor. Then we asked the pipeline to classify emotions with a
new test set containing Frustration emotion, which change the
problem from four to five classes. The model is able to rec-
ognize the Frustration class with a precision of 41.85% and
recall of 33.38%, although it has not seen any speech sample
of this class in the regressor’s training.

6. CONCLUSION

In this work, we investigated the relationship between dimen-
sional 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.

However, there are several benefits of following the “classi-
fication via regression” approach, our experiment on two
different corpora shows degradation of performance com-
pared to a classifier that profits from categorical labeled data.
We compared the performance of these two approaches by
employing three feature types, three architectures as regres-
sor/classifier, and three mapping algorithms for transforming
the continuous value in the AVD space to categorical labels.
While the main objective of this study was not to outperform
the state of the art, we showed the potential and limitations of
the “classification via regression” approach.

The benefits of the proposed approach are reducing the
cost of data preparation and breaking the limitation of the
predefined number of emotional categories. The performance
of the model can be improved by applying the state-of-the-
art techniques, such as fine-tuning the pre-trained wav2vec2
or wavLM feature extractors [19, 20]. Although the perfor-
manence of the best regressor model and mapping algorithm as
the classifier is decent, particularly in the IEMOCAP dataset.

The importance of dimensions in predicting categorical
labels (see Section 3) suggests modifying the loss function of
the regressor model to a weighted loss in the future works.

7. REFERENCES

[1] Roddy Cowie and Randolph R Cornelius, “Describ-
ing the emotional states that are expressed in speech,”
Speech communication, vol. 40, no. 1-2, pp. 5–32, 2003.

[2] Paul Ekman, “An argument for basic emotions,” Cogni-
tion & emotion, vol. 6, no. 3-4, pp. 169–200, 1992.

[3] James A Russell and Albert Mehrabian, “Evidence for
a three-factor theory of emotions,” Journal of research
in Personality, vol. 11, no. 3, pp. 273–294, 1977.
[4] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy 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.

[5] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso, “The ordinal nature of emotions: An emerging approach,” IEEE Transactions on Affective Computing, vol. 12, no. 1, pp. 16–35, 2018.

[6] Reza Lotfiand and Carlos Busso, “Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning,” Proc. Interspeech, pp. 951–955, 2018.

[7] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono, “Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4964–4968.

[8] 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 25th ACM international conference on Multimedia, 2017, pp. 890–897.

[9] 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. 335–359, 2008.

[10] R. Lotfiand 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] Takuya Fujioka, Takeshi Homma, and Kenji Nagamatsu, “Meta-learning for speech emotion recognition considering ambiguity of emotion labels,” in Proc. Interspeech, 2020, pp. 2332–2336.

[12] Hatice Gunes and Björn Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” Image and Vision Computing, vol. 31, no. 2, pp. 120–136, 2013.

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

[14] Vladimir Chernykh and Pavel Prikhodko, “Emotion recognition from speech with recurrent neural networks,” arXiv preprint arXiv:1701.08071, 2017.

[15] Reza Lotfi and Carlos Busso, “Curriculum learning for speech emotion recognition from crowdsourced labels,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 815–826, 2019.

[16] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, “x-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7169–7173.

[17] Vasudha Kowtha, Vikramjit Mitra, Chris Bartels, Erik Marchi, Sue Booker, William Caruso, Sachin Kajarekar, and Devang Naik, “Detecting emotion primitives from speech and their use in discerning categorical emotions,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7164–7168.

[18] Kusha Sridhar and Carlos Busso, “Speech emotion recognition with a reject option,” in Proc. Interspeech, 2019, pp. 3272–3276.

[19] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” in Proc. Interspeech, 2021, pp. 3400–3404.

[20] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba, “A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding,” arXiv preprint arXiv:2111.02735, 2021.

[21] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[22] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhou Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” arXiv preprint arXiv:2110.13900, 2021.

[23] Aharon Satt, Shai Rozenberg, and Ron Hoory, “Efficient emotion recognition from speech using deep learning on
spectrograms,” in *Proc. Interspeech*, 2017, pp. 1089–1093.

[24] Felix Weninger, Fabien Ringeval, Erik Marchi, and Björn W Schuller, “Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio,” in *IJCAI*, 2016, vol. 2016, pp. 2196–2202.