A Glimpse into Emotion through the Lens of Computer Science

Mariana Rodríguez Castañeda
Universidad Nacional Atónoma de México
Facultad de Ingeniería Eléctrica - Procesamiento Digital de Señales
Mexico City, Mexico
mariana.rodriguez.c@comunidad.unam.mx

***Abstract*—In recent years, deep learning algorithms have introduced various approaches to emotion classification. In the audio domain, several representations can be utilized to train models. However, some of the primary issues with current architectures include high computational requirements, complexity, and the lack of interpretability of the data they entail. This work focuses on addressing these problems by concentrating the attention on pre-processing the database to emphasize features that effectively represent emotions. Resulting in inputs with reduced complexity and more significant features for the task. This approach leads to less noise in the data and a better understanding of emotions through their features.**

I. Introduction

Mental health has received significant attention in recent years, along with the recognition of emotions as a key factor in achieving a fulfilling life as a human being. In science, emotions are studied as chemical processes, but recognizing emotions in a person has been a challenging task, requiring a non-intrusive method to acquire information for analysis.

Images have been used to recognize emotions through expressions; the position of the lips or eyebrows can indicate disagreement or happiness. The results of image processing were then applied to video processing to detect emotions in real-time. However, using images for this task can be challenging due to variations in scene lighting, which can affect the recognition of facial expressions. Additionally, processing video information in real-time requires a significant amount of computational resources.

The issue of intrusion in emotion detection extends to the measurement of vital signs. The computational resources required for image and video processing in real-time applications have made to consider another approach for classifying emotions. Audio signals do not depend on lighting and do not require intrusive methods for information acquisition. Moreover, advancements in noise suppression and audio source location research, which have increased in recent years, positioned audio processing as a competitive method for classification tasks.

II. Databases

In emotion recognition, the impact of these advancements can be observed in the increasing number of datasets available for the task. Some of the most relevant datasets include RAVDESS, EMODB, IEMOCAP, CREMA-D, and for 3D audio, the SAVEE dataset. These datasets are being used in various approaches to the problem. In this document, the RAVDESS and CREMA-D datasets will be explored due to their similarities.

The CREMA-D dataset consists of 7,442 original clips from 91 actors. These clips were recorded by 48 male and 43 female actors, aged between 20 and 74, representing various races and ethnicities including African American, Asian, Caucasian, Hispanic, and Unspecified. Actors spoke a selection of 12 sentences, each presented with one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified)[1].

The RAVDESS dataset contains 7,356 files and features 24 professional actors (12 female, 12 male), each vocalizing two lexically-matched statements in a neutral North American accent. Speech includes expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, while song includes expressions of calm, happiness, sadness, anger, and fear. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only, Audio-Video, and Video-only[2].

From the CREMA-D database, 22 Caucasian actors and actresses were selected to ensure concordance between both datasets. The same was done with the intensity levels; each actor spoke a phrase at two intensity levels: medium and high. Therefore, each emotion comprises 44 audio clips. It is worth mentioning that in the case of the neutral emotion, there are no different intensity levels. This was resolved by recording two audio clips for each actor, ensuring the same number of audio clips as for the other emotions. The final dataset consists of 264 audio clips from Caucasian actors, representing six different emotions.

The RAVDESS database, in contrast to CREMA-D, has fewer actors but more data representations and emotions. Similar to CREMA-D, emotions in RAVDESS were chosen to align with those in CREMA-D. This database does not include multiple races, so 22 actors were selected, divided equally between 11 women and 11 men. With both intensity levels, the database contains two audio clips for each actor, resulting in 44 audio clips for each emotion. The neutral emotion does not have intensity levels, so two audio clips were selected to

Features List	
Continuous	Spectral
Energy	Spectral centroid
Energy entropy	Spectral spread
Zero crossing rate	Spectral entropy
Delta energy	Spectral flux
Delta energy entropy	Spectral roll off
Delta ZCR	MFCC 1-13
Frame energy	Chroma 1-12
Frame entropy	Chroma std
	Delta spectral centroid
	Delta spectral spread
	Delta spectral entropy
	Delta spectral flux
	Delta spectral roll off
	Delta mfcc 1-13
	Delta chroma 1-12
	Delta chroma std
	Spectral spread
	Formants

Tabla 1

Features List
Speech Production
GCI
OQ
VOQ
NAQ
NAQ
H1H2
H1H2
HRF
VHRF

Tabla 2

ensure each class has the same number of audio clips.

III. Features Extracted

For data analysis, feature extraction and construction must be performed to represent our data in different ways, as features provide different information about the audio. For the purpose of this document, various features were collected. Table 1 illustrates the different features that were either extracted or constructed from the continuous and spectral domains. Additionally, a set of speech production characteristics were added, including prosody, phonation, and articulation features.

The prosody features are based on the reconstruction of the glottal source from sustained vocals and continuous dialogue in the voiced frames [3]. Table 2 shows these features, where; OQ, stands for the average opening quotient for consecutive glottal cycles, which is the rate of opening phase duration or duration of the glottal cycle; VOQ represents the variability of the opening quotient; NAQ, denotes the average normalized amplitude quotient.

H1H2 represents the average difference between the first two harmonics of the glottal flow signal, while variability H1H2 indicates the variability in the difference between the first two harmonics of the glottal flow signal. HRF stands for the average Harmonic Richness Factor, which is the ratio of the sum of the harmonics' amplitude and the amplitude of the fundamental frequency. GCI denotes the variability of the time between consecutive glottal closure instants, and VHRF represents the variability of HRF. These are some of the features from the speech production category, but a list of 103 features was constructed from the prosody category, 28 from the phonation, and 108 from the phonological features. The final list of features comprises 375 features.

IV. Implementation

Once all the features were extracted, the dimensionality of our space increased, and the problem with using all the features to feed a model is that it can be counterproductive. The features are divided into strong relevance, weak relevance, and otherwise. The otherwise features add noise to the data as they can be either too correlated with one or more of the other features, referred to as redundant features, or low correlated with the classes, referred to as irrelevant features.

The next step is selecting our features. The purpose of selecting the features is to reduce the dimensionality of our dataset in order to reduce computational resources. There are different types of feature selection algorithms, starting from basic methods like the feed-forward and back-forward methods. The feed-forward method aims to remove feature by feature from the starting set and feed them to the model to observe the accuracy reached with those features. Then, one can conclude which features provide more accuracy to the model and select them to try their models. On the other hand, the back-forward method starts with zero features, and by adding features, one can see whether the model's accuracy falls, remains the same, or increases with the selected feature.

There are three main categorization of feature selections:

1) Filter methods
2) Wrapper methods
3) Embedded methods

Filter methods, which are carried out independently as a preprocessing step before the use of any machine learning algorithm, tend to be faster as they rank variables with correlation coefficients [4]. Wrapper methods, such as the feed-forward and back-forward methods, fall into the category of feature selection that uses the model to verify the usefulness of the feature. This approach is often criticized because it seems to be a "brute force" method requiring massive amounts of computation [5]. Embedded methods combine filter methods and wrapper methods for feature selection.

A genetic programming approach is proposed. While there are different methods within the literature for wrapper methods, where the model is set as the fitness function of the algorithm, this work establishes a new perspective. It is suggested that feature selection should be done based on the specific task, as it is a matter of the domain of the data.

Therefore, to select features that better represent the task in a lower dimension without losing quality of the space analyzed, all methods that involve models for selection were discarded. This includes wrapper and embedded methods. It is crucial not to use models in this task because the goal is

to verify how well a feature represents emotions. One of the objectives of this work is to provide interpretability to the classification. By adding a model, the algorithm doesn't measure how well the feature characterizes the space of emotions, but rather how good the model is at distinguishing emotions with the given features, measuring model performance instead of features performance.

To this end, an exploration of our features has been done using strategies of dimensionality reduction. There are several dimensionality reduction algorithms that are commonly used; four were explored:

1) PCA
2) LDA
3) ISOmap
4) Umap

There are some advantages to using each one, but not all are appropriate for every task. For example, PCA and LDA are commonly used techniques, they are linear transformations that can be beneficial for tasks without high complexity. In the case of ISOmap and Umap, they are non-linear transformations. An advantage of Umap over ISOmap is that Umap not only retains the general structure like ISOmap but also the local structure. As our data is formed by variables like MFCCs of high-dimensional representation, it is of interest for our task that the local structure be preserved.

Due to the characteristics of Umap and after using it on the datasets described above, it was selected for the task. Umap was used as dimensional reduction, from which a fitness function was created. The formula involves the number of clusters, calculated by k-means with silhouette analysis, the standard deviation of clusters, the standard deviation of data in each cluster, and the mean distance between clusters. This is the fitness function for the genetic algorithm. The final results are positive, as a reduction of dimensionality with better clustering was obtained.

V. Conclusion

Nowadays, the use of feature selection has been left to the model, creating complex architectures, increasing computational resources, and losing interpretability of the data by selecting the best features for the model to classify, instead of focusing on the task and its domain. Much has been said about the lack of preprocessing of the data and the model perception of a black box. In this work has been proven the utilization of genetic programming is a helpful method in high dimensional datasets. The results were successful in eliminating irrelevant and redundant features and finding a space that better represents emotions based on their features and not on the model used to classify them. The fitness function created for the genetic algorithm performed well in evaluating the space for the selection of features and increase in clustering.

References

[1] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," in IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377-390, 1 Oct.-Dec. 2014, doi: 10.1109/TAFFC.2014.2336244. keywords: Crowdsourcing;Emotion recognition;Databases;Audio-visual systems;Emotional corpora;facial expression;multi-modal recognition;voice expression,

[2] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[3] UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

[4] JohnGeorge H. et al. Irrelevant features and the subset selection problem

[5] An Introduction to Variable and Feature Selection 2003 Journal of Machine Learning Research 3 (2003) 1157-1182