International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 23–35
DOI: https://doi.org/10.2991/ijcis.d.201019.002; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Emotion Recognition from Speech: An Unsupervised Learning Approach

Stefano Rovetta[1], Zied Mnasri[1,2,*], Francesco Masulli[1], Alberto Cabri[1]

[1]*DIBRIS, University of Genoa, Via Dodecaneso 35, Genova, 16146, Italy*
[2]*ENIT, University Tunis El Manar, BP37, Le Belvedere, Tunis, 1002, Tunisia*

ARTICLE INFO

Article History

Received 01 May 2020
Accepted 02 Oct 2020

Keywords

Emotion recognition
Speech signal
Feature extraction
K-means
Fuzzy clustering
Membership function

ABSTRACT

Speech processing is quickly shifting toward affective computing, that requires handling emotions and modeling expressive speech synthesis and recognition. The latter task has been so far achieved by supervised classifiers. This implies a prior labeling and data preprocessing, with a cost that increases with the size of the database, in addition to the risk of committing errors. A typical emotion recognition corpus therefore has a relatively limited number of instances. To avoid the cost of labeling, and at the same time to reduce the risk of overfitting due to lack of data, unsupervised learning seems a suitable alternative to recognize emotions from speech. The recent advances in clustering techniques make it possible to reach good performances, comparable to that obtained by classifiers, with much less preprocessing load and even with generalization guarantees. This paper presents a novel approach for emotion recognition from speech signal, based on some variants of fuzzy clustering, such as probabilistic, possibilistic and graded-possibilistic fuzzy c-means. Experiments indicate that this approach (a) is effective in recognition, with in-corpus performances comparable to other proposals in the literature but with the added value of complexity control and (b) allows an innovative way to analyze emotions conveyed by speech using possibilistic membership degrees.

1. INTRODUCTION

Nowadays applications are more and more interactive, which requires an optimal human–machine interaction. One of the most obvious ways to achieve this goal is spoken communication. Since a few decades, speech processing has registered considerable progress in different applications, such as speech recognition, synthesis and enhancement, source separation, etc.

Speech is a complex communication form that conveys information at several levels in addition to verbal content. One of these is emotion, a key component that may enable a much more effective interaction. Unfortunately, speech processing applications perform much better for acquiring the verbal content than for the recognition of expressive components. For instance, a benchmark comparison of emotion recognition based on deep learning architectures has shown that emotion recognition is providing accuracy rates under 80% for most expressive speech databases [1]. Furthermore, a recent evaluation of deep learning architectures for emotion recognition on a state-of-the-art expressive speech database, i.e., IEMOCAP [2], has not shown better results [3].

The literature offers approaches that can showcase a very high recognition performance. However, they are usually tested via cross-validation on the same limited-size corpora that are used for their training. It turns out [4] that *cross-corpus* evaluation is a much more difficult task, with some experimental results bordering random guessing. Due to the high number of parameters characterizing current deep learning models, this appears to be a clear indication that these methods are overfitting, i.e., learning the database rather than its information content. In other words, the good results reported by recent supervised methods on limited-size corpora are not reproducible in different contexts. While the ability of large machine learning models to work well in a regime of overfitting is the subject of current studies [5], this phenomenon cannot be relied upon in the absence of very extensive training sets.

Given the typical size of available corpora, the only alternative approach to avoid overfitting is *capacity control*, which consists in using machine learning methods whose ability to learn (and thus also overfit) depends on a controllable number of effective parameters. Classic theories developing this approach include Vapnik and Chervonenkis' statistical learning theory, the fat-shattering dimension and Rademacher complexity [6]. However, many methods that do not explicitly refer to these approaches can nevertheless be studied under the framework of complexity control, for instance those based on regularization or stochastic regularization (stochastic gradient descent, early stopping, dropout).

The work described in this paper has the goal to explore the task of emotion recognition from speech signal using a mainly unsupervised workflow. The use of this class of techniques can be justified in the light of capacity control theories, as will be briefly exposed

* *Corresponding author. Email: zied.mnasri@enit.utm.tn*

in the following. This gives it an advantage in the reliability of the attained experimental results with limited data.

The methodologies adopted include (a) combined techniques of features analysis, i.e., feature embedding by autoencoders and feature selection by analysis of variance (ANOVA) or mutual information (MI); (b) different clustering methods, such as crisp clustering using K-means, and fuzzy clustering using probabilistic, possibilistic and graded-possibilistic c-means; (c) a novel way to analyze emotion recognition using the sum-of-membership matrix, which is made possible thanks to the use of a possibilistic-type fuzzy clustering, as will be detailed hereafter.

The main contribution of this work consists in proposing a novel methodology for speech emotional content analysis based on clustering, using either crisp or fuzzy methods. The methodology uses unsupervised learning methods, such as autoencoders, to extract features. Up to our knowledge, this is the first work totally relying on unsupervised learning, both for feature extraction and speech clustering. This work is an extension of results presented at the 11th Conference of the European Society for Fuzzy Logic and Technology [7].

The rest of the paper is organized as follows: Section 2 presents the state-of-the-art of emotion recognition from speech, including databases, feature sets, emotion representation models, and the use of unsupervised learning. Section 3 describes the speech materials used in this work, including the expressive speech database chosen, the standard feature set employed and the psychological emotion model adopted. Section 4 details the methods employed for this study. Section 5 reports on the results and the interpretation of the experimental work; finally the conclusion (Section 6) presents some comments and perspectives.

2. RELATED WORK

2.1. Emotion Models

Emotion recognition from speech relies upon established psychological models. For instance, the Ekman model [9] states that there are six basic emotions, i.e., *neutral*, *anger*, *fear*, surprise, *joy* and *sadness*, that are recognized whatever the language, the culture or the means (speech, facial expressions, etc.). More detailed models of emotions rely on continuous dimensions rather than atomic "basic" emotions. Russel's circumplex model [10] suggests that emotions can be represented in a bi-dimensional space, where the x-axis represents valence and y-axis represents arousal (cf. Figure 1). Furthermore, Plutchik proposes a tri-dimensional model [11] which combines the basic and the bi-dimensional models. Thus, the outer emotions are a combination of the inner ones.

Classically, and like in speech recognition, emotion recognition was achieved using different methods, namely generative models such as hidden Markov models with Gaussian mixture models (HMM-GMM) [12], artificial neural networks (ANNs) [13,14], and support vector machines (SVMs) [15], yielding nearly the same accuracy [16]. Also, the combination of such models, either in series, in parallel or in a hierarchical way, has given better results than those obtained by single models [16]. Recently, deep learning tools like deep feedforward, recurrent or convolutional neural networks, have outperformed all the aforementioned models for emotion recognition [3,17].

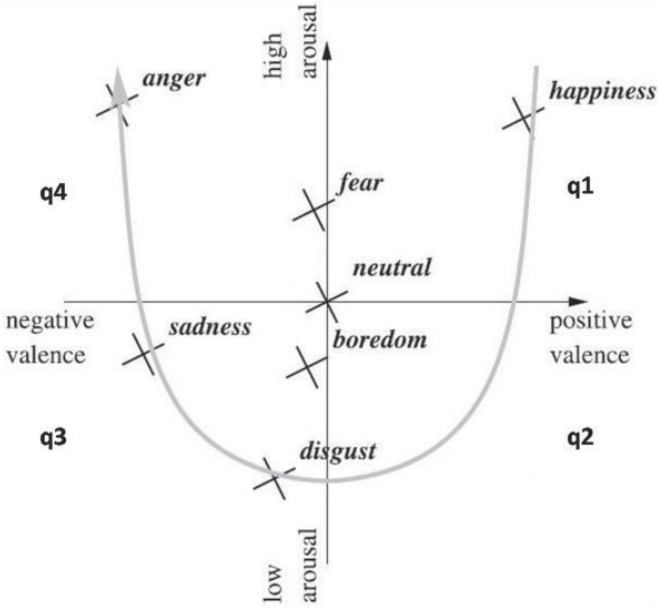

Figure 1 Valence/arousal model of emotion classes for the Emotion Database (EMO-DB) [20]. Figure adapted from [8].

2.2. Emotional Speech Databases

A variety of emotional speech databases were designed or recorded, covering more or less the emotion models described above, i.e., the Eckman [9], Russel [10] and Plutchik [11] models. An inventory of emotional speech databases [16] shows that the main differences between them lie in (a) the size, varying from a few tens of sentences to a few thousands [18]; (b) the number of speakers; (c) the type of speech, whether uttered by professional actors, or recorded from spontaneous conversation like telephone recordings; and (d) the number of emotions, which depends on the emotion model. A recent and updated comprehensive inventory [19] lists the main emotional speech databases. In particular, the EMO-DB database [20] has been widely used, since it covers all basic emotions in equal and sufficient proportions.

2.3. Standard Emotion-Recognition Features

Whatever the chosen model of emotions, a speech signal provides a representation of it that lies at a much lower level, i.e., a train of audio samples (see Figure 2). Therefore, special importance should be given to intermediate representations that make it possible to discriminate the types of emotional content; in other words, features have a crucial importance.

In machine learning there are two main approaches to obtaining good features: Using standard, expert-engineered feature sets that are known to be well correlated with the desired recognition task; or using optimization to learn features that will provide the best performance. In this work we will adopt both approaches, by encoding speech signals using standard feature sets at the lower level, but subsequently extracting higher-level features by means of unsupervised learning.

Generally in speech recognition the commonly used features can be divided into prosodic vs. acoustic (or spectral) ones. Prosodic

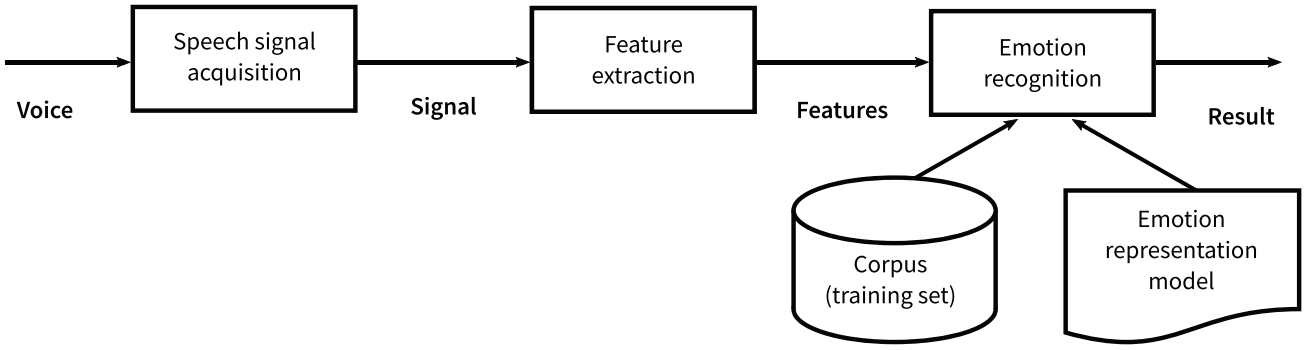

Figure 2 | General scheme of an emotion recognition system.

features include f_0, intensity and phone duration, whereas acoustic features are extracted from spectrum such as Mel-frequency cepstral coefficients (MFCCs), Linear spectral parameters (LSPs), and their first and second time-derivatives (Δ and Δ- Δ).

Another classification of features relies on the level of extraction, i.e., local vs. global features. Local features are extracted at each frame, like f_0, intensity, MFCC, etc., whereas global features are calculated using statistics all over the speech signal, like mean, variance, range, skewness, kurtosis, min and max values. This allows making the extracted features less dependent on the linguistic aspects of speech signal [16]. Interspeech'09 feature set [21], ComParE [22] and GeMAPS [23], which use mainly global features, are among the most used standard feature sets for emotion recognition.

2.4. Feature Learning Methods: Extraction and Selection

The ultimate goal of feature analysis is to optimize the input space, either by discarding the irrelevant (or redundant) features, i.e., feature selection, or by a nonlinear combination of features in order to obtain more discriminant ones, i.e., feature extraction. In particular, one interesting technique for feature extraction is feature embedding.

In the particular case of emotion recognition, feature analysis, either by extraction or by selection, has been widely used. For instance, feature extraction by sequential format floating search (SFFS) was used to choose the most relevant features for emotion recognition based on Bayesian classification [24]. The SFFS criterion was applied with the assumption that the features follow a multivariate Gaussian distribution. Then the variance of the correct classification rate of the Bayesian classifier during cross-validation is estimated.

Also, principal component analysis (PCA) was used in several emotion recognition-related works [25–28]. For emotion recognition, it has been noticed that the classification accuracy increases when the number of principal components is increased up to a certain order, after which the accuracy starts to decrease [16].

Albeit to a lesser degree, linear discriminant analysis (LDA) was also used in some works about emotion recognition. However, the results about the relevance of each group of features, i.e., pitch-related, energy-related and spectral features are not coherent enough [16]. This may be due to the difference of databases and feature sets used in each work.

To compare PCA and LDA for emotion recognition, both techniques were applied on BHUDES, a Chinese emotional speech corpus [29], before undertaking classification with ANN and SVM [30]. For both classifiers, the results have shown that using LDA for feature selection gives better recognition rates than using PCA, either for all classes or for every single emotion.

Furthermore, Eyben *et al.* evaluated the feature relevance in real-life conditions [31]. To fulfill that, an experience was set up by corrupting clean speech by different noise level, before extracting the standard ComParE feature set [32]. Then the Pearson correlation coefficients (CCs) of each feature with continuous target label was computed. The selected features were the subset of 400 features (among 6353 ones) having the best CC coefficients for arousal, valence and level of interest (LOI) tests. However, it has been shown in the tests that change in feature group relevance depends more on the individual tests than on the level of noise.

The feature extraction methods described so far, PCA and LDA, are linear mappings which make strong assumptions about the structure underlying the data. An alternative approach is nonlinear feature embedding.

The autoencoder is a neural network whose objective approximates the identity function. It is commonly used as an unsupervised learning technique, that aims to extract features from unlabeled data. To achieve this goal, the autoencoder optimizes the weights to minimize the mean square difference between the given input and the obtained output; then, the value of a hidden layer is used as an encoded representation of the input.

As shown in Figure 3, a simple autoencoder has only one hidden layer. It is therefore parameterized by weights ($w \in \mathbb{R}^{mxn}$, $\tilde{w} \in \mathbb{R}^{nxm}$) and biases ($b, \tilde{b} \in \mathbb{R}^m$), as follows:

$$\begin{cases} h = f(wx + b) \\ \tilde{x} = \tilde{f}(h\tilde{w} + \tilde{b}) \end{cases} \tag{1}$$

where $x = (x_1, x_2, \dots, x_m) \in \mathbb{R}^m$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_m) \in \mathbb{R}^m$ and $h = (h_1, h_2, \dots, h_n) \in \mathbb{R}^n$ are respectively the inputs, the outputs and the hidden layer code, and $f, \tilde{f}$ are nonlinear activation functions, such as the sigmoid function, $f(z) = \frac{1}{1+e^{-z}}$. [33].

It can be shown that the encoding obtained from a simple *linear* autoencoder, i.e., with $f(z) = \tilde{f}(z) = z$, spans the n principal components of the data space, recovering therefore the same embedding as PCA of order n. In this sense we may state that an autoencoder is a nonlinear generalization of PCA.

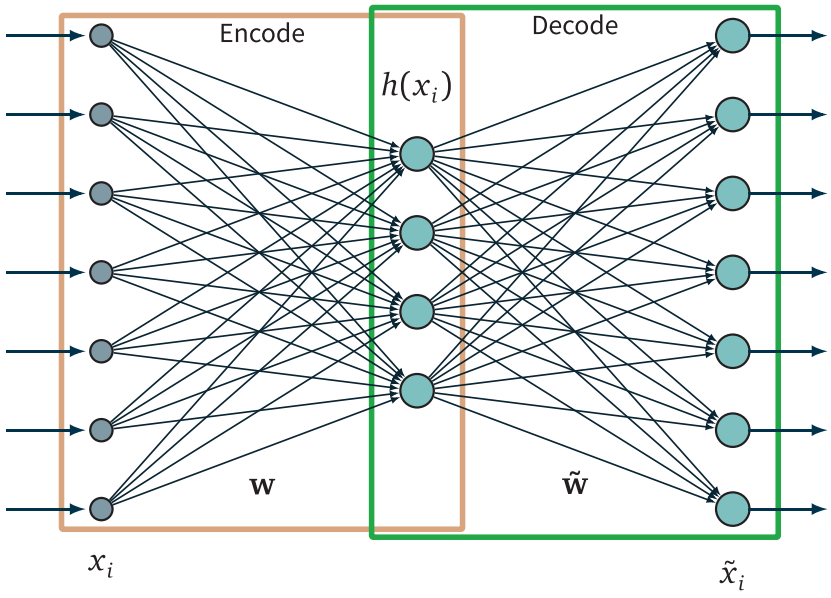

Figure 3 | Autoencoder architecture.

Deep autoencoders with several hidden layers are also possible, although this may imply an excessive overparameterization with increased risk of overfitting, or, correspondingly, the need for exponentially more data.

A simple autoencoder can be split into two parts: (a) the encoder, from the input layer to the middle layer and (b) the decoder, from the middle layer to the output layer. The encoded features are obtained at the output of the encoder layer. Hence, to reduce the dimension of the input space, the encoder layer should have a lower dimension than the input layer (cf. Figure 3). The encoder layer provides a useful transformation of input features, that allows, first discovering hidden structure in the input features, and second generating new features through the nonlinear transformation of the input features by the activation functions of the hidden layers.

The autoencoder has been used as a feature extraction method in several clustering-based works. For instance, Song *et al.* [34] trained an autoencoder with a new objective function, where the centroids are updated at the same time as the weights and biases of the neural networks. In the work of Xie *et al.* [35], clustering was based on deep autoencoders. This process starts by initializing the parameters of the clustering model with an autoencoder, and then the centroids and the autoencoder's parameters are optimized using Kullback–Leibler (KL) divergence to maximize the similarity between the distribution of the embedded features and the centroids. A comparison between using or not autoencoded features for spectral clustering was conducted on different data sets [36], such as documents (20-Newsgroup), biological data (DIP [37] and BioGrid [38]) and chemical data (WINE [39]), to reveal that the autoencoded features yield better results.

2.5. Clustering

Clustering techniques can be inventoried following several criteria, whether they are hierarchical, partition-based, density/neighborhood-based or model-based [8]. Obviously clustering was less used than classification for emotion recognition, since most works deal with labeled databases. However, in a big data context, unsupervised recognition methods can prove more useful, since labeling a huge quantity of expressive speech data would be a tedious and expensive task.

For instance, self-organizing maps (SOMs) were used by Szekely *et al.* to detect emotions in audiobooks [40], based on articulatory features. Also, hierarchical K-means were used by Eyben *et al.* to detect emotions in a corpus dedicated for expressive speech synthesis, using prosodic and acoustic features [31]. It should be noted that, since clusters do not necessarily correspond to classes, a "vector quantization" approach is usually used, whereby the number of clusters is overestimated, and subsequently the detected clusters are grouped into classes; methods based on this type of approach can even be competitive with entirely supervised approaches, with theoretical guarantees [41], and can be easily used in a semi-supervised context (partially labeled data).

Experiments have shown that in clustering the choice of features is more important for the final accuracy than in classification, where the generalization power of the classifier and the direct minimization of a loss function could mask the irrelevance or the aberrance of some features.

In this work, we are particularly interested in applying some of our recent results in fuzzy clustering [42] for emotion recognition. The soft/fuzzy clustering approach consists in using a real-valued membership function instead of a categorical or binary membership decision. Then an object belongs to all clusters, but with different membership degrees, having values between 0 and 1 [43].

The fuzzy clustering problem can be stated as follows: Given a set $X = x_1, ..., x_n$ of data objects, a set $\Omega = \omega_1, ..., \omega_c$ and a membership function $\mu(x, \omega)$, find $\omega \in \Omega$ such that $\forall\, x \in X$, $0 \leqslant \mu(x, \omega) \leqslant 1$.

Considering the membership function, fuzzy clustering methods can be categorized as either probabilistic or possibilistic. Then the pair (Ω, ω) is

- A possibilistic partition if $\mu(x,\omega) \in \mathbb{R}\ \forall x,\ \forall \omega$ such that $0 < \sum_{i=1}^{c} \mu(x,\omega_i) < c$
- A probabilistic partition if it is a possibilistic partition such that $\sum_{i=1}^{c} \mu(x,\omega_i) = 1$

The extremely popular *central* clustering approach consists in defining clusters ω and memberships $\mu(\cdot,\cdot)$ in terms of reference points in the data space, or centroids. The fuzzy clustering literature abounds with variations over the basic K-means algorithm. We now briefly review the specific approach followed in this work.

All fuzzy versions of K-means are in principle based on minimizing the following objective function:

$$\hat{E} = \sum_{j=1}^{c}\sum_{l=1}^{n} \mu_{jl} d_{jl} \tag{2}$$

where $\mu_{jl} = \mu(x_l, \omega_j)$ is the degree of membership of pattern x_l to cluster ω_j and $d_{jl} = \| x_l - y_j \|$ is the Euclidean distance between the pattern x_l and the cluster centroid y_j, so that this objective represents the average weighted distance of data points to centroids, also termed *distortion* in the vector quantization literature.

However, directly minimizing this objective yields a degenerate problem, whose solution is given by the (crisp) K-means method. For a real-valued (i.e., fuzzy) solution, the distortion objective is regularized in either of two ways, which were proved to be related by a common framework [44]: Either by introducing an exponent $m > 1$ in the weights, that become μ_{jl}^{m}, or by additive regularization terms. The first choice results in the fuzzy c-means algorithm [45]. The second choice has been the basis of a number of further methods and will be adopted in the following.

Optimization of the objective is customarily done via alternate minimization, which iteratively solves two problems assumed as independent: minimization with respect to centroid positions and minimization with respect to memberships.

In all cases, the cluster centroids y_j are not included in the regularization terms, therefore their locally optimal values only depend on the basic objective and are computed as follows:

$$y_j = \frac{\sum_{l=1}^{n} \mu_{jl} x_l}{\sum_{l=1}^{n} \mu_{jl}} \tag{3}$$

Regarding the membership function, in the cases of our interest it can be expressed as

$$\mu_{jl} = \frac{v_{jl}}{Z_l} \tag{4}$$

where v_{jl} depends on the specific regularization scheme adopted and Z_l, a "generalized partition function," realizes the choice of the particular clustering model. For a pretty wide family of objectives regularized with entropy-like terms [44,46–49], $v_{jl} = e^{-d_{jl}\beta_j}$ with $\beta_j > 0$ a free parameter related to cluster width.

Regarding the generalized partition function Z_l, given a transformation $f(\cdot)$ it can be written as $Z_l = f(\sum_{j=1}^{c} v_{jl})$: here the choice of $f(\cdot)$ defines the clustering paradigm, i.e.,

- $Z_l = (\sum_{j=1}^{c} v_{jl}^{\frac{1}{m-1}})^{m-1}$ in case of probabilistic clustering.
- $Z_l = 1$ in case of possibilistic clustering.
- $Z_l = (\sum_{j=1}^{c} v_{jl})^{\alpha}$ in case of graded-possibilistic clustering.

The *fuzziness parameter* $m \in \mathbb{R}$, $m > 1$ recovers Bezdek's [45] original formulation, but it is not really needed in this context.

3. SPEECH MATERIAL

To perform this work, an emotional speech database was selected from the available speech corpora. We chose EMO-DB [20] since it has been widely used and cited as a reference. Besides, a special attention was addressed to choosing the feature set, since several ones have been proposed in the literature.

3.1. Speech Database

EMO-DB [20] is a publicly available database of prepared emotional speech. Prepared speech corpora differ from spontaneous speech corpora, since they are elaborated by linguists to represent all the language phenomena in a balanced and normalized way. EMO-DB contains 10 German sentences (5 short and 5 long) uttered by 10 native-speaking professional actors (5 male and 5 female). Every sentence was uttered by every actor in 7 emotions (*neutral*, *anger*, *boredom*, *fear*, *disgust*, *joy* and *sadness*) once (or twice in a few cases). The sentences were recorded in an anechoic chamber, at 16 KHz sampling rate. The database was labeled including the emotion of each sentence, the syllabic segmentation and the stress level of each syllable. It is worth noting that EMO-DB has provided the highest emotion recognition rates using state-of-the-art classifiers, such as HMM-GMM and SVM [21].

3.2. Feature Set

Emotion recognition feature sets have been addressed extensive research, yielding a variety of proposed sets [19]. However most feature sets used a limited number of feature types, i.e., prosodic, spectral, voice-quality-related [19] and in a lesser proportion articulatory features [40]. Furthermore, the global features, i.e., statistics calculated all over the speech signal, were generally preferred to local features that are measured at each frame [50]. In particular, The Interspeech'09 emotion recognition challenge feature set was preferred to conduct this work, for two main reasons: (i) its preliminary results [21] and (ii) its compactness. Then features were extracted using the Opensmile toolkit [51]. Tables 1 and 2 show the complete set of features and the calculated statistics extracted for each, respectively, so that 384 features (16 descriptors + their 16 Δ-values) × 12 statistics) were extracted from each signal.

Table 1 Interspeech'09 emotion recognition challenge feature set.

Speech Parameter	Descriptors
Zero-crossing rate	ZCR, Δ-ZCR,
Root mean square energy	RMS energy, Δ-RMS energy,
fundamental frequency	F_0, Δ-F_0,
Harmonic-to-noise ratio	HNR, Δ-HNR,
12 Mel-Frequency cepstral coefficients	(MFCC (1–12)), Δ-MFCC(1–12)

3.3. Emotion Classes

Initially, classes consisted in single emotions, namely *neutral*, *anger*, *boredom*, *disgust*, *fear*, *joy* and *sadness*. However, we also employed a second way to label speech signals by using groups of emotions as classes instead of individual emotions. In fact, grouping emotions using the valence/arousal mapping was thought to increase clustering performance (cf. Table 3). Both sets of labels were evaluated during experiments.

4. METHODS

To conduct the experimental process, different steps were followed. First, the speech corpus was preprocessed to extract and select the most relevant features, second crisp and fuzzy clustering were performed following different strategies, third clusters were majority-labeled with class labels and finally the clustering and classification results were analyzed (cf. Figure 5).

4.1. Preprocessing

The experimental process starts with a preprocessing phase, in which the following steps are performed: (a) feature embedding, where the original descriptors (cf. Tables 1 and 2) were transformed by embedding with the autoencoder, so that a new set of features was extracted; (b) feature selection, where the final features were selected amongst the extracted ones, using either ANOVA or MI test.

4.1.1. Feature extraction

The first step in feature analysis consists in feature extraction. In this work, it was performed through feature embedding using a 3-hidden-layer deep autoencoder. The set of 384 features for each signal (cf. Table 1) is trained by the autoencoder to extract a smaller number of features at the encoder layer. Therefore, two schemes of preprocessing were tried out: (i) application of the autoencoder to the whole set of features and (ii) application of the autoencoder to each low-level descriptors (LLD) group, so that only one feature is extracted out of each 12-feature LLD. The autoencoder architectures used in (i) and (ii) are described in Table 4.

Table 2 Statistical parameters used for Interspeech'09 emotion recognition challenge feature set.

Features for Each Descriptor	Parameters
Global statistics	Mean, standard deviation, skewness, kurtosis
Minimum	Value, relative position, range,
Maximum	Value, relative position, range,
Linear regression coefficients	Offest, slope, Mean square error (MSE)

Table 3 Groups of emotions.

New Label	Grouped Labels	Common Characteristics
AJ	Anger and Joy	High absolute valence and arousal
NB	Neutral and Boredom	Low absolute valence and arousal
FD	Fear and Disgust	Low absolute valence and medium absolute arousal
S	Sadness	High absolute valence and medium absolute arousal

4.1.2. Feature selection

Once the features were extracted, through embedding, feature selection was set up. Although features seem to be mostly uncorrelated, a finer analysis was performed using ANOVA and MI to further reduce the cardinality of the set of extracted features. Furthermore, to keep a certain coherence between the selected features, two ANOVA strategies were adopted, the first evaluating individual features, and the second, denoted ANOVA group, evaluating groups of features, where each group contains the 12 statistics of each descriptor (cf. Tables 1 and 2).

4.2. Clustering

4.2.1. The graded possibilistic C means algorithm

As well as using the "crisp" K-means clustering method, we will follow a fuzzy approach by employing the graded-possibilistic c-means (GPCM) algorithm for unsupervised learning.

The graded-possibilistic paradigm [48] allows to switch continuously between the probabilistic and possibilistic paradigms by modulating the free parameter $\alpha \in [0,1]$, a *degree of probabilistic tendency*:

- $\alpha = 0 \quad \rightarrow$ fully possibilistic
- $\alpha = 1 \quad \rightarrow$ fully probabilistic
- $0 < \alpha < 1 \quad \rightarrow$ graded possibilistic

Also, it is worth noting that in case of probabilistic central clustering, $\sum_{j=1}^{c} \mu_{jl} = 1$, whereas this condition is not necessarily met in the possibilistic and graded-possibilistic paradigms [42].

To calculate the values for the cluster width parameters β_j, the following is suggested [42]:

$$\beta_j = -\frac{\ln(t)}{\min\limits_{h \neq j} ||y_h - y_j||^2} \tag{5}$$

where $t \in [0,1]$ is the threshold satisfying

$$\max_{h \neq j} \mu(y_h, y_j) \leqslant t,$$

i.e., it establishes the maximum accepted overlap between clusters. In particular $t = 1/2$ guarantees no overlap. This provides a minimum reference value for the width parameters, which can be subsequently increased if deemed necessary.

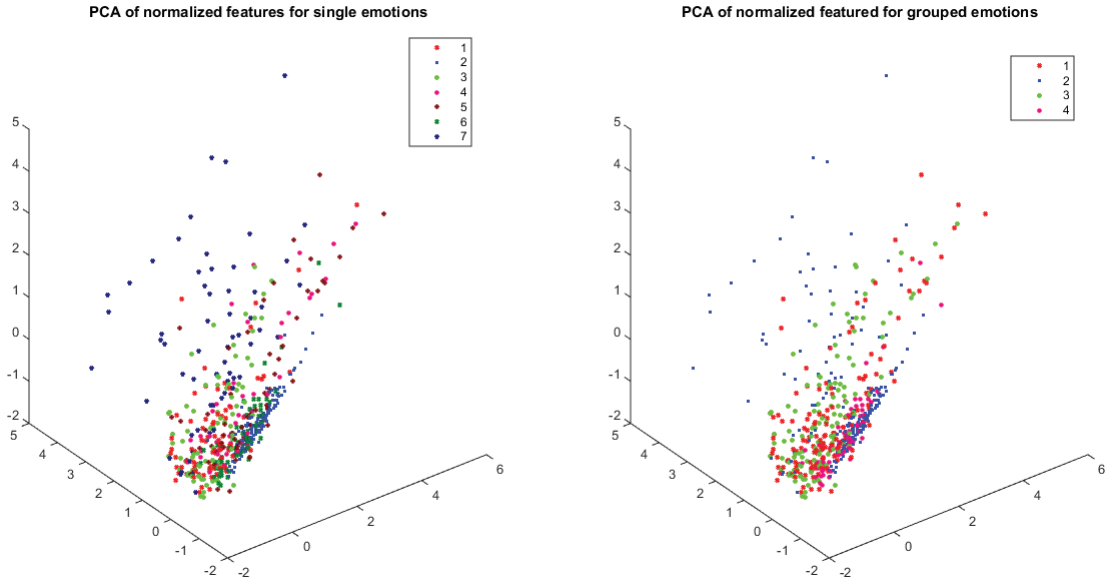

Figure 4 | Distribution of single and grouped emotion classes, visualized using principal component analysis (PCA) to project features in three dimensions: (a) for single emotions, three-dimensional projection suggests that the Interspeech'09 standard features are not discriminatory enough and (b) emotions were grouped the reduce feature scattering.

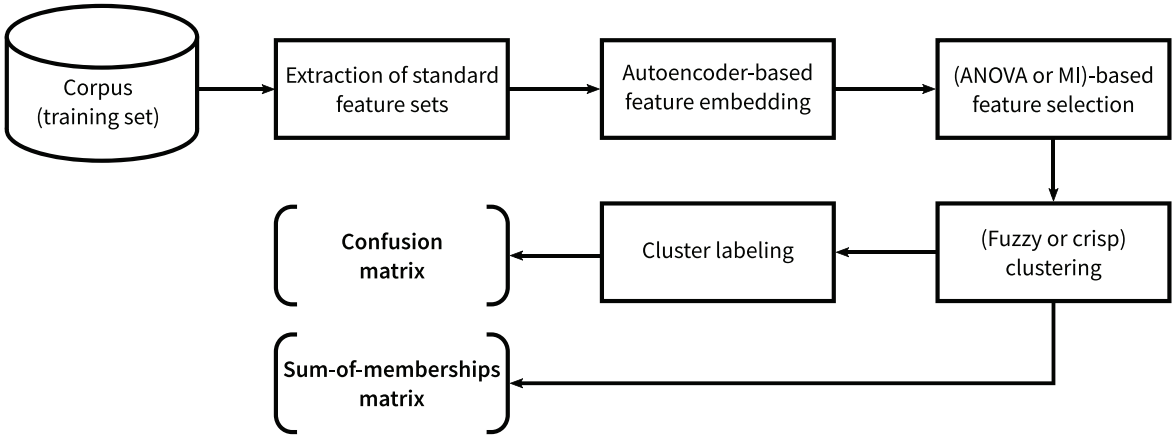

Figure 5 | Experimental process as applied in this work: Preprocessing includes hand-crafted feature computation, feature embedding and selection; clustering is achieved using crisp (K-means) and/or fuzzy methods, cluster labeling is applied to recuperate emotion classes from clusters. The evaluation outputs are the confusion and the sum-of-membership matrices.

4.2.2. Use of possibilistic membership for classification analysis

A distinctive advantage of the possibilistic paradigm, as compared to "probabilistic" fuzzy clustering, is represented by the ability to evaluate the typicality of patterns with respect to the learned clusters. Since memberships are not constrained to have a constant sum, for a given data instance it is possible to analyze the sum-of-memberships to each cluster and evaluate how well it fits the clusters distribution.

It is also possible to analyze the cumulative sum-of-memberships by classes, to gain insight on the "perceived" internal structure of emotions, to check for instance if two emotions are consistently considered similar. This particular analysis will be presented for the experimental results in Section 5.

Table 4 | Autoencoder architectures (layers and number of nodes).

Input Ffeatures	Input Layer	Hidden Layer 1	Code Layer	Hidden Layer 3	Output Layer
All features	384	500	32	500	384
LLD features	12	100	1	100	12

From the technical standpoint, the *graded* approach used in this work partially constrains the sum-of-membership; one advantage of this is ruling out degenerate solutions, and another one is allowing for an easier convergence of the optimization. However, the cumulative sum-of-membership matrix will have entries that depend on the degree of probabilistic tendency α. For this reason, this model parameter has been kept constant throughout the work (see below).

4.2.3. Parameter setting

In Rovetta *et al.* [42] it was observed that α should be close to 1 due to its exponential role. Therefore it was suggested to work with α on a logarithmic scale, by setting $\alpha = log_2(a+1)^2$ where $a \in [0.5, 1]$ to have $\alpha \in [0.9, 1]$. This makes parameter tuning easier.

4.3. Capacity Control in Unsupervised Learning

Now we provide a brief justification of the choice of supervised learning for limited-size data sets. The learning capacity of a central clustering model followed by supervised labeling was studied in a previous work [41]. In particular, the Vapnik–Chervonenkis dimension of this class of learning methods was established by Theorem 1 from the reference.

Theorem 1. *[41] The Vapnik–Chervonenkis dimension of a central clustering model with c centroids, used for classification by majority labeling, is c.*

The significance of this theorem, whose proof can be found in the referenced paper, lies in the fact that the only free parameter that influences the learning capacity is the number of centroids, so neither the dimensionality (number of features) nor the size (number of observations) of the data influence the Vapnik–Chervonenikis dimension. This capacity measure is computed in a worst-case scenario, so it is an overestimation of the actual learning capacity that can be expected in real cases. If we can upper-bound the Vapnik–Chervonenkis dimension, we can be confident that other, more realistic indexes like the fat-shattering dimension will not exceed it.

5. EXPERIMENTAL WORK

5.1. Experimental Protocol

Experiments were carried out following a protocol where the model parameters were varied, one at a time:

- The labels set, i.e., single emotions or groups of emotions (cf. Table 3).
- The number of clusters, increasing from the number of classes, to 3 times.
- The feature selection method, i.e., ANOVA or MI.
- The number of selected features, decreasing from all features, i.e., no feature selection, to 25% of features.
- β for the fuzzy c-means models, by increasing the parameter t from 0.1 to 0.5.
- α for the graded-possibilistic c-means model, by increasing the parameter a from 0.9 to 1.

These combinations yielded a high number of experiments, therefore only those providing the most relevant results are presented (cf. Table 5). In addition, at every execution of the fuzzy clustering algorithms, K-means was performed under the same conditions, i.e., number of classes, number of clusters, feature selection method and the number of selected features, and using a fixed number of replicates, equal to 10.

5.2. Results

Figure 4 shows a representation of clusters, projected in 3D using PCA as a visualization tool, for the original clusters, whereas Figure 6 shows the 2D distribution for predicted clusters using different methods, i.e., K-means, probabilistic, possibilistic and graded-possibilistic c-means.

5.2.1. Confusion matrix

The performances of both crisp and fuzzy clustering, followed by supervised labeling, were analyzed through the scores calculated from the confusion matrix, i.e., overall accuracy, precision, recall and F1-score (cf. (6a) to (6d)).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6a}$$

$$Precision = \frac{TP}{TP + FP}, \tag{6b}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6c}$$

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}, \tag{6d}$$

where TP, FP, TN and FN are respectively the numbers of true positives, false positives, true negatives and false negatives.

5.2.2. Sum-of-membership matrix

This proposed matrix (cf. Table 6) shows, for every recognized emotion class, the sum of the memberships to original classes. The coefficients of this matrix correspond to the sum-of-membership to an original class, calculated for all samples. This method is proposed in order to measure how many emotion classes were correctly recognized, or in other terms how much the features of every single emotion are shared by the other ones. The matrix can be analyzed like a confusion matrix, where the lines correspond to the recognized emotions and the columns to the original ones.

5.2.3. Initial distribution of classes

It looks since the beginning that in spite of using a standard feature set [21], the distribution of classes looks too dispersed (cf. Figure 4).

5.2.4. Fuzzy clustering performance

However, thanks to feature embedding and then feature selection, and to a good choice of the possibilistic and graded-possibilistic models parameters, i.e., α and β, the fuzzy clustering methods provide as good rate as the crisp clustering (cf. Figure 6). Besides, using

Table 5 Best recognition rates (in bold character) for different parameters combinations.

Number of Classes	Number of Clusters	Feature Selection Method	Proportion of Selected Features (%)	t	α	K-means Rate (%)	GPCM Rate (%)
7	7	ANOVA group	50	0.1	0.9	51.9	**69.6**
7	14	ANOVA	25	0.1	0.9	56.6	55.9
7	21	ANOVA	25	0.1	0.9	**60.9**	63.0
4	4	ANOVA	75	0.1	0.9	62.4	61.3
4	8	ANOVA	50	0.1	0.9	69.9	**77.4**
4	12	ANOVA	50	0.1	0.9	**73.9**	75.1

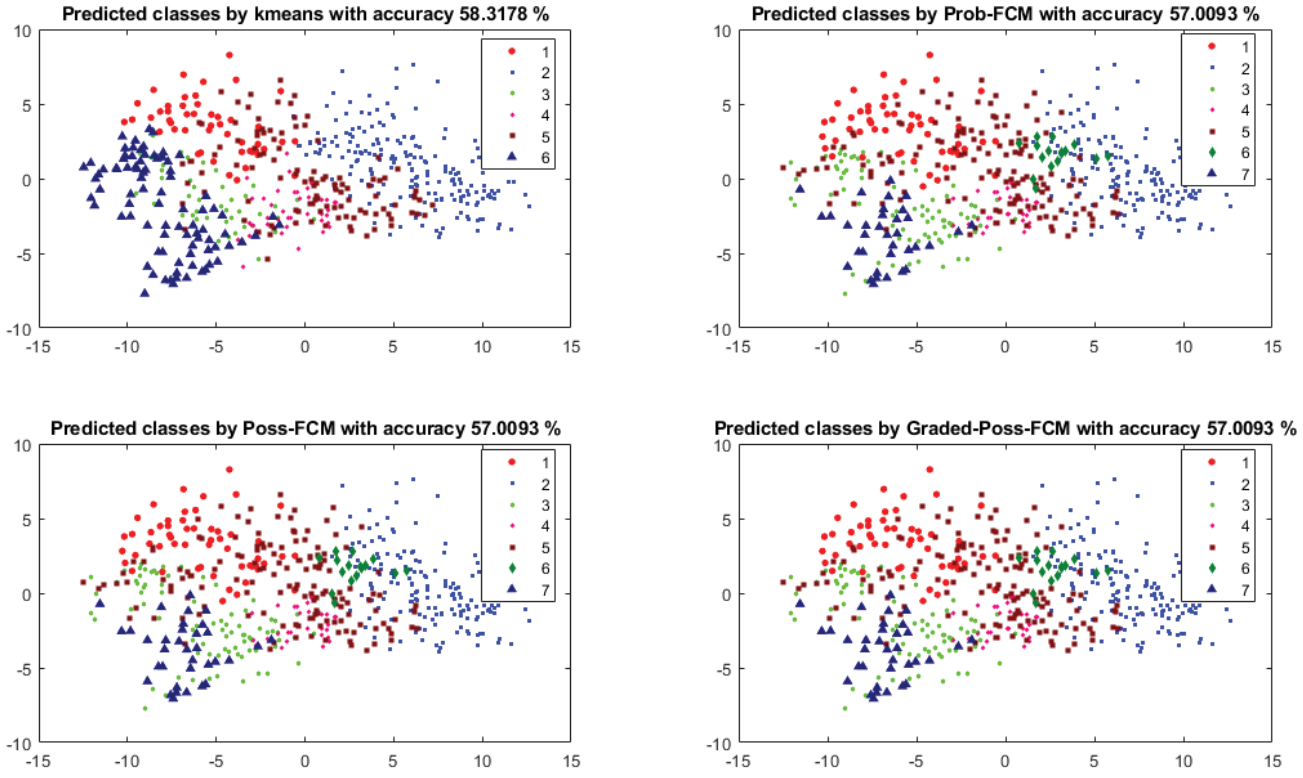

Figure 6 Clustering results for 7 classes using 21 clusters and 96 features selected by analysis of variance (ANOVA) method for K-means and different graded-possibilistic c-means (GPCM) methods (two-dimensional principal component analysis (PCA) projection): Though the two-dimensional projection does not show clearly disjoint clusters, K-means and GPCM methods seem able to separate some classes, e.g., Class 1 and Class 2.

Table 6 Sum-of-membership matrix calculated using 192 features selected by *MI*, GPCM with $t = 0.1$ and $\alpha = 0.9$. n.b The highest sum-of-membership matrix are in bold character.

Classes	Neutral	Anger	Boredom	Disgust	Fear	Joy	Sadness
Neutral	**35.1571**	5.0713	23.3220	5.1520	6.3261	2.7159	6.9717
Anger	5.2943	60.7539	3.9122	3.8892	9.9439	27.4724	0.1005
Boredom	6.0872	2.7400	**17.6296**	6.7532	4.8307	1.9224	17.0275
Disgust	**9.3078**	**6.3142**	**10.6536**	**12.7571**	5.9771	4.7399	0.5343
Fear	16.0078	**28.1121**	15.1840	11.4498	**35.0615**	12.5662	11.8798
Joy	0.5790	**18.8198**	0.1959	0.7331	2.8752	**18.7432**	0.0006
Sadness	6.5668	5.1887	10.1028	5.2657	3.9855	2.8400	**25.4855**

a number of clusters greater than that of classes helps increasing performance (cf. Table 5). However, the number of clusters should not be too big.

5.2.5. Single emotions vs. groups of emotions

Another result consists in increasing the clustering rate when emotions were grouped using the valence/arousal model. This could be explained by the fact that using more samples and less classes may increase the clustering rate, but it also tells about the relevance of grouping such emotions, despite some pairs contain opposite emotions (e.g., *anger* and *joy*). This last point may be useful in emotion analysis, using objective measures, such as the statistics used in the feature set.

5.3. Benchmarking

In order to compare the performance of the proposed approach, the results of other methods applied on the same expressive speech

database, i.e., EMO-DB [20], have been investigated. Table 8 shows the highest overall accuracy measured for such methods, combining supervised and unsupervised techniques for feature extraction and classification. It looks that the proposed approach, entirely based on unsupervised learning, both for feature extraction and classification, is not far away from the other methods that use supervised learning, either partly (for classification only) or entirely (in the whole process).

5.4. Interpretation and Discussion

In addition to emotion recognition, further results could be obtained from analyzing the results obtained by fuzzy clustering under the possibilistic framework. In the following we the confusion and the sum-of-membership matrices could provide as novelty regarding emotion analysis.

5.4.1. Analysis of the classification results

Though the obtained classification results may be lower than those that should have been provided by supervised learning, the obtained rates can be considered as satisfactory in the framework of unsupervised learning. In fact, cluster labeling has been utilized in this work only to check the performance. However, in real-world application of the proposed method, the data need not be entirely labeled, hence obtaining an overall accuracy of about 60% might be encouraging. A deeper analysis of the classification scores (cf. Table 7 and Figure 7) shows that: (a) crisp and fuzzy clustering do nearly the same for each class of emotions, e.g., *anger* and *sadness* are both well recognized, whereas *joy* is much less predicted by both techniques, though all classes have the same number of samples; (b) F1-score is higher than 50% for most emotion classes and for both methods, i.e., crisp and fuzzy clustering, which means that precision and recall are rather balanced (though it is less obvious for some emotions like *disgust* and *joy*). In fact the F1-score is a measure that reveals whether a high accuracy could hide unbalanced precision and recall.

Finally, though emotion classes are balanced in the EMO-DB database [20], the differences between classification results may be due to the choice of features. Though we opted for a standard feature set, that has been successfully used for Interspeech'09 emotion recognition challenge [21], it looks less efficient to detect all emotions equally. Another interpretation could be that some intense emotions, such as *joy*, may share a lot of their aspects like valence and arousal, and hence have similar features as other emotions that have similar levels of valence or arousal, such as *anger*.

As noted in Section 4, the sum-of-membership matrix provides a way to analyze these aspects.

5.4.2. Analysis of the sum-of-membership matrix

By inspecting the sum-of-membership matrix (Table 6), in line 2 the recognized emotion *anger* has the highest sum-of-membership to the same original emotion, which means that most samples recognized as belonging to the class *anger* have the highest membership value to the same class. This is also the case for classes *neutral* and *sadness*. However, for the other recognized emotions, like *boredom*, *disgust*, *fear* and *joy*, the sum-of-membership matrix shows that their sum-of-membership to some other classes are as high as for the original classes, e.g., the recognized class *boredom* shares nearly the same sum-of-membership to the original classes *boredom* and

Table 7 Confusion matrices calculated by K-means and GPCM, both with subsequent cluster labeling, using 192 features selected by *ANOVA* and with GPCM parameters $t = 0.1$ and $\alpha = 0.9$. n.b The highest confusion matrix values are in bold character.

		K-means							GPCM						
		Original Classes													
		Neutral	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral	Anger	Boredom	Disgust	Fear	Joy	Sadness
Pred. Classes	Neutral	**50**	0	18	5	14	6	1	**36**	0	5	3	10	6	2
	Anger	5	**115**	2	3	12	**46**	0	5	**112**	1	3	11	**47**	0
	Boredom	12	0	**42**	11	1	2	8	25	0	**48**	11	3	2	6
	Disgust	2	0	1	**21**	6	2	1	0	0	1	**17**	3	2	0
	Fear	7	5	4	6	**32**	7	0	9	8	8	12	**40**	6	4
	Joy	0	7	0	0	1	8	0	0	7	0	0	1	8	0
	Sadness	3	0	14	0	3	0	**52**	4	0	18	0	1	0	**50**

Table 8 Benchmark results of different methods combining supervised and unsupervised feature extraction and classification on EMO-DB expressive speech database (n.b. CNN: Convolutional neural networks, LSTM: Long short-term memory, BLSTM: Bidirectional long short-term memory, SVM: Support vector machines).

Learning	Input Features	Feature Extraction	Classification	Accuracy (%)
All supervised	Spectrogram images	CNN	BLSTM [52]	91.3
	Raw audio	CNN	LSTM [17]	88.9
Unsupervised feature extraction and supervised classification	Spectrogram images	K-means	SVM [53]	71.5
	Spectrogram images	Autoencoder	SVM [53]	67.4
All unsupervised	Interspeech'09 [21]	Autoencoder	K-means-VQ (Prop) [7]	60.9
	Interspeech'09 [21]	Autoencoder	GPCM-VQ (Prop) [7]	69.6

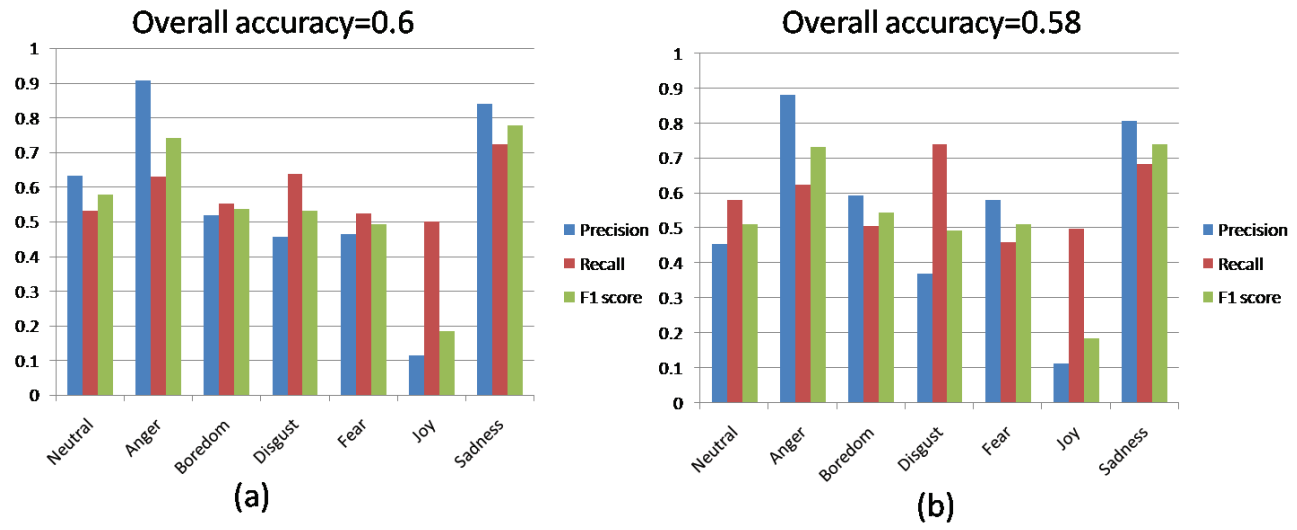

Figure 7 Scores of the confusion matrix for (a) K-means, (b) graded-possibilistic c-means (GPCM), for clustering followed by labeling, using 192 features selected by analysis of variance (ANOVA).

sadness, and recognized *joy* shares the same sum-of-membership with original *joy* and *anger*, etc.

In comparison to the scores calculated from the confusion matrix, cf. Figure 7, it could be easily noticed that highly misclassified emotions are the same which share most of their membership functions with other ones, such as *disgust* and *joy*. Hence, the sum-of-membership matrix could show the same tendency of the confusion matrix in case of fuzzy clustering. These findings could be interpreted as a novel way to approach emotion recognition/perception, e.g., when a recognized emotion class has most of its memberships in the same original class, e.g., for *anger*, this means that the model succeeds to identify most of the angry voices as belonging to the same class; however, when a recognized class has its sum-of-membership shared by more than one original emotion class, this could be interpreted, either as these classes share several features, e.g., for *joy* and *anger*, or the emotion itself is a mixture of more basic ones, e.g., *disgust* is a mixture of *anger*, *boredom* and *neutral*. Still, this interpretation could be deepened by human-listeners through subjective evaluation. At last, such an analysis shows the relevance of fuzzy clustering in (i) enhancing the emotion recognition from single signals, (ii) analyzing emotions, or at least to reveal some of their hidden characteristics thanks to the analysis of the sum-of-membership matrix.

6. CONCLUSION

In this paper, a novel approach for emotion recognition using fuzzy clustering is described. The main idea consists in clustering speech according to basic emotions, using (a) unsupervised learning for feature extraction, and more precisely feature embedding with autoencoders, (b) new advances in fuzzy clustering, such as possibilistic and graded-possibilistic c-means, in addition to probabilistic c-means, to recognize emotion from speech. Besides, the crisp approach was treated using K-means algorithm, for evaluation purposes. Several adjustments were also made to fine-tune the models, including feature embedding using autoencoders, feature selection using ANOVA and MI analysis, and finally varying the possibilistic models parameters. Also, using more clusters than classes helped increasing the recognition rates. In addition, choosing the optimal values of parameters had an impact on increasing the performance of possibilistic and graded-possibilistic c-means models.

The confusion matrix scores, i.e., accuracy, precision, recall and F1, confirm the efficiency of using fuzzy clustering as an alternative tool to supervised learning for emotion recognition. Either for single emotions or for groups of emotions, crisp and fuzzy clustering perform almost equally, yielding an overall accuracy of nearly 60% and a precision higher than 80% for some emotions such as *anger* and *sadness*, with equivalent recall. This may be quite useful as an alternative way for emotion recognition in large and especially unlabeled speech data sets.

In addition to the classical confusion matrix, utilized to show the classification performance, a novel representation based on the sum-of-membership matrix is presented. The analysis of such a matrix for fuzzy clustering shows a similar behavior than the confusion matrix, where highly recognized emotions tend to have a high membership to the same original emotion, whereas misclassified emotions tend to share their sum-of-membership with other emotions. This already allows differentiating between “strong” or “basic” emotions, which monopolize their membership values and “weak” or “mixed” emotions that tend to share their sum-of-memberships. This representation allows studying the dependence of each basic emotion to the other ones, and could be a helpful tool for emotion analysis, to understand how speech signal conveys emotions and how they are perceived.

ACKNOWLEDGMENTS

This work was supported by the research grant funded by “Fondi di Ricerca di Ateneo 2016” of the university of Genova. The authors declare that there is no conflict of interest regarding this work. Author 1 and author 2 have contributed equally to this work; author 3 and author 4 have contributed to the revision of the article.

REFERENCES

[1] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: a benchmark comparison of performances, in IEEE Workshop on Automatic Speech Recognition Understanding (ASRU 2009), IEEE, Merano, Italy, 2009, pp. 552–557.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, *et al.*, Iemocap: interactive emotional dyadic motion capture database, Lang. Res. Eval. 42 (2008), 335.

[3] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Netw. 92 (2017), 60–68.

[4] E. Avots, T. Sapiński, M. Bachmann, D. Kamińska, Audiovisual emotion recognition in wild, Mach. Vis. Appl. 30 (2019), 975–985.

[5] M. Belkin, S. Ma, S. Mandal, To understand deep learning we need to understand kernel learning, arXiv preprint arXiv:1802.01396, 2018.

[6] M. Anthony, P.L. Bartlett, Theoretical Foundations, Martin Anthony and Peter, Cambridge University Press, Cambridge, U.K., 1999. pp., 389, ISBN 0-521-57353X

[7] S. Rovetta, Z. Mnasri, F. Masulli, A. Cabri, Emotion recognition from speech signal using fuzzy clustering, in 2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019), Atlantis Press, Prague, September 9-13 2019.

[8] G. Ulutagay, E. Nasibov, Fuzzy and crisp clustering methods based on the neighborhood concept: a comprehensive review, J. Intell. Fuzzy Syst. 23 (2012), 271–281.

[9] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (1992), 169–200.

[10] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (1980), 1161.

[11] R. Plutchik, The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, Am. Sci. 89 (2001), 344–350.

[12] T.L. Nwe, S.W. Foo, L.C. De Silva, Speech emotion recognition using hidden markov models, Speech Commun. 41 (2003), 603–623.

[13] V. Hozjan, Z. Kačič, Context-independent multilingual emotion recognition from speech signals, Int. J. Speech Technol. 6 (2003), 311–320.

[14] J. Nicholson, K. Takahashi, R. Nakatsu, Emotion recognition in speech using neural networks, Neural Comput. Appl. 9 (2000), 290–296.

[15] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'04), IEEE, Montreal, Canada, 2004, vol. 1, pp. 1–577.

[16] M.E. Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (2011), 572–587.

[17] J. Kim, R. Saurous., Emotion recognition from human speech using temporal information and deep learning, in Annual Conference of the International Speech Communication Association (Interspeech 2018), Hyderabad, India, 2018.

[18] J.H.L. Hansen, S.E. Bou-Ghazale, Getting started with susas: a speech under simulated and actual stress database, in Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, 1997.

[19] M.B. Akçay, K. Oğuz, Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Commun. 116 (2020), 56–76.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005.

[21] B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, in Tenth Annual Conference of the International Speech Communication Association, Brighton, United Kingdom September 6-10, 2009.

[22] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. Narayanan, The interspeech 2010 paralinguistic challenge, in Proceeding INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 2794–2797.

[23] F. Eyben, K.R. Scherer, B.W. Schuller, J.Sundberg, E. André, C. Busso, *et al.*, The Geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2016), 190–202.

[24] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features, and methods, Speech Commun. 48 (2006), 1162–1181.

[25] C.M. Lee, S.S. Narayanan, R. Pieraccini, Classifying emotions in human-machine spoken dialogs, in Proceedings, IEEE International Conference on Multimedia and Expo, IEEE, Lausanne, Switzerland, 2002, vol. 1, pp. 737–740.

[26] D. Ververidis, C. Kotropoulos,I. Pitas, Automatic emotional speech classification, in 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Montreal, Canada, 2004, vol. 1, pp. 1–593.

[27] M. You, C. Chen, J. Bu, J. Liu, J. Tao, Emotion recognition from noisy speech, in 2006 IEEE International Conference on Multimedia and Expo, IEEE, Toronto, Canada, 2006, pp. 1653–1656.

[28] M. You, C. Chen, J. Bu, J. Liu, J. Tao, A hierarchical framework for speech emotion recognition, in 2006 IEEE International Symposiumon Industrial Electronics, IEEE, Montreal, Canada, 2006, vol. 1, pp. 515–519.

[29] X. Mao, L. Chen, L. Fu, Multi-level speech emotion recognition based on HMM and ANN, in 2009 WRI World Congress on Computer Science and Information Engineering, IEEE, Los Angeles, CA, USA, 2009, vol. 7, pp. 225–229.

[30] L. Chen, X. Mao, Y. Xue, L.L. Cheng, Speech emotion recognition: features and classification models, Digital Signal Process. 22 (2012), 1154–1160.

[31] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M.J.F. Gales, K. Knill, Unsupervised clustering of emotion and voice styles for expressive TTS, in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2012), IEEE, Kyoto, Japan, 2012, pp. 4009–4012.

[32] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, *et al.*, The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013.

[33] N. Andrew, Sparse autoencoder, 2011. https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf
[34] C. Song, F. Liu, Y. Huang, L. Wang, T. Tan, Auto-encoder based data clustering, in: J. Ruiz-Shulcloper, G. Sanniti di Baja (Eds.), Iberoamerican Congress on Pattern Recognition, Springer, Berlin, Heidelberg, Germany, 2013, pp. 117–124.
[35] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in International Conference on Machine Learning, 2016, pp. 478–487.
[36] F. Tian, B. Gao, Q. Cui, E. Chen, T.-Y. Liu, Learning deep representations for graph clustering, in Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, Québec City, Québec, Canada. 2014.
[37] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic Acids Res. 32 (2004), D449–D451.
[38] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M.S. Livstone, *et al.*, The biogrid interaction database: 2011 update, Nucleic Acids Res. 39 (2010), D698–D704.
[39] A. Asuncion, D. H.Newman, UCI Machine Learning Repository, 2007. University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml,
[40] E. Székely, J.P. Cabral, P. Cahill, J. Carson-Berndsen, Clustering expressive speech styles in audiobooks using glottal source parameters, in 12th Annual Conference of the International-Speech-Communication-Association, Florence, Italy August 27-31., 2011.
[41] S. Ridella, S. Rovetta, R. Zunino, K-winner machines for pattern classification, IEEE Trans. Neural Netw. 12 (2001), 371–385.
[42] R. Rovetta, F. Masulli, Soft clustering: why and how to, in The 12th International Workshop on Fuzzy Logic and Applications (WILF 2018), 2018.
[43] R. Babuška, H.B. Verbruggen, An overview of fuzzy modeling for control, Control Eng. Pract. 4 (1996), 1593–1606.
[44] M. Miyamoto, M. Mukaidono, Fuzzy C-Means as a regularization and maximum entropy approach, in Proceedings of the Seventh IFSA World Congress, Prague, Czech Republic, 1997, pp. 86–91.
[45] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[46] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (1993), 98–110.
[47] R. Krishnapuram, J.M. Keller, The possibilistic C -Means algorithm: insights and recommendations, IEEE Trans. Fuzzy Syst. 4 (1996), 385–393.
[48] F. Masulli, S. Rovetta, Soft transition from probabilistic to possibilistic fuzzy clustering, IEEE Trans. Fuzzy Syst. 14 (2006), 516–527.
[49] K. Rose, E. Gurewitz, G. Fox, A deterministic annealing approach to clustering, Pattern Recognit. Lett. 11 (1990), 589–594.
[50] M.-J. Caraty, C. Montacié, Detecting speech interruptions for automatic conflict detection, in: F. D'Errico, I. Poggi, A. Vinciarelli, L. Vincze (Eds.), Conflict and Multimodal Communication, Springer, Cham, Switzerland, 2015, pp. 377–401.
[51] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in Proceedings of the 18th ACM international conference on Multimedia, ACM, Firenze, Italy, 2010,
[52] L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, X. Li, Speech emotion recognition by combining amplitude and phase information using convolutional neural network, in INTERSPEECH, 2018, pp. 1611–1615.
[53] Z.-W. Huang, W.-T. Xue, Q.-R. Mao, Speech emotion recognition with unsupervised feature learning, Front. Inf. Technol. Electron. Eng. 16 (2015), 358–366.