ATLANTIS PRESS

Atlantis Studies in Uncertainty Modelling, volume 1

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Emotion recognition from speech signal using fuzzy clustering

Stefano Rovetta[a] and Zied Mnasri[a,b] and Francesco Masulli[a,c] and Alberto Cabri[a]
[a]DIBRIS, University of Genoa, Via Dodecaneso 35, 16146 Genoa, Italy
stefano.rovetta@unige.it
[b]Electrical engineering department, ENIT, University Tunis El Manar, BP 37, 1002 Tunis, Tunisia
zied.mnasri@enit.utm.tn
[c]Sbarro Inst. for Cancer Research and Molecular Medicine, Temple University, Philadelphia (PA, USA)
francesco.masulli@unige.it
alberto.cabri@dibris.unige.it

Abstract

Expressive speech modeling is a new trend in speech processing, including emotional speech synthesis and recognition. So far, emotion recognition from speech signal has been mainly achieved using supervised classifiers. However, clustering techniques seem well fitted to resolve such a problem, especially in huge databases, where speech labeling may be a hard and tedious task. This paper presents a novel approach for emotion recognition from speech signal, based on fuzzy clustering, including probabilistic, possibilistic and graded-possibilistic c-means. In comparison to crisp clustering, mainly using kmeans, fuzzy c-means look more fitted for this problem, and potentially offer an innovative way to analyze emotions conveyed by speech using membership degrees.

Keywords: Emotion recognition, speech signal, kmeans, fuzzy clustering, membership function.

1 Introduction

Nowadays applications are more and more interactive, which requires an optimal human-machine interaction. To fulfill this goal, speech seems as the most obvious and natural way. Though a considerable progress has been registered in many speech processing applications, such as automatic speech recognition, speaker verification, text-to-speech synthesis, etc., speech technology is still struggling to deal with natural aspects of speech, especially emotions. Actually it is still problematic to accurately detect emotion from speech, i.e. emotion recognition, or to generate an emotional speech, i.e. expressive speech synthesis.

Typically emotion recognition from speech is considered as a pattern recognition problem. Hence a variety of models, feature sets and databases have been developed and tested since many years to enhance recognition rates. Emotion recognition models could be split into classification vs. clustering ones. Emotion classifiers include all well-known supervised learning techniques, such as hidden Markov models using Gaussian mixture models (HMM-GMM), artificial neural networks (ANN) and support vector machines (SVM); whereas unsupervised techniques used for this purpose include K-means and SOM (self-organizing maps). Also, other models, mainly combining basic models were used [4]. In the same way, features could be classified into prosodic vs. acoustic ones. Prosodic features, like phone duration, fundamental frequency (F_0) and energy, are more related to perceptual/phonological aspects of speech, namely rhythm, intonation and loudness. On the other hand, acoustic features like MFCC (Mel-Frequency cepstral coefficients), LSP (Linear spectral pair) and others, are more linked to the physical/phonetic aspects of speech. However in both cases, using global statistics, like mean, variance, range, min and max may reduce the effect of the linguistic aspects. Furthermore, several emotional speech databases have been designed or collected, including different labeling schemes. Actually the number of emotion labels depends on the emotion model, containing either basic emotions like Ekman model [3] or detailed ones, like Russel [16] and Plutchik [14] models. Though speech and emotion recognition have been historically developed using classification tools, like HMM-GMM, ANN and SVM, clustering has been also successfully used for this goal. Hence, in [20] SOM were used to cluster expressive speech styles. Also, in [5] k-means were applied to detect emotions and voice styles. Though the results were satisfactory, classification techniques are still outperforming clustering. Besides, to the best of our knowledge, only crisp clustering has been used so far in emotion recognition from speech.

Therefore, in this paper we describe a study using fuzzy clustering for emotion recognition and analysis,

from speech signal. Actually, such a clustering method was preferred because the membership function used in fuzzy clustering allows a) detecting the emotion conveyed by speech and b) analyzing the "purity" of the detected emotion. This paper is organized as follows: section 1 presents the state of the art of emotion recognition from speech, section 2 describes the fuzzy clustering techniques used in this work, section 3 details the experimental process, and finally the results are discussed.

2 Emotion recognition from speech

2.1 Emotion classes

Emotion recognition from speech relies on the established psychological models. For instance, Ekman model [3] states that there is a set of six basic emotions, i.e. neutral, anger, fear, surprise, joy, sadness, that are recognized whatever the language, the culture or the means (speech, facial expressions, etc.). More extended models of emotions rely on dimensionality. Hence, Russel's circumplex model [16] 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).

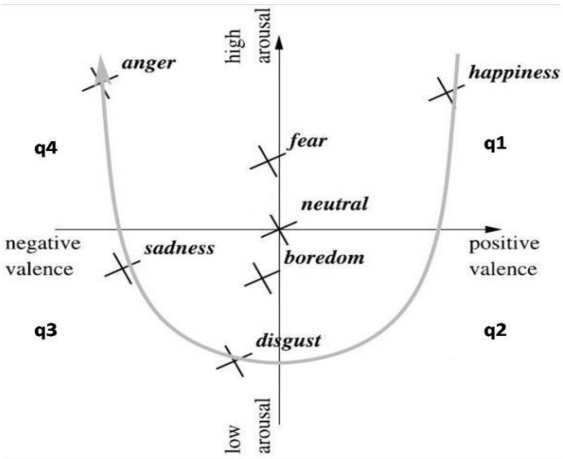

Figure 1: Valence/arousal model of EMO-DB emotion classes [21]

Furthermore, Plutchik proposes a tridimensional model [14] which combines the basic and the bi-dimensional models. Thus, the outer emotions are a combination of the inner ones.

2.2 Emotion recognition from speech

Historically, emotion recognition has inherited from speech recognition techniques. Actually, both tasks rely on speech signal analysis to extract a set of features that is fed into a classification/clustering model (cf. Figure 2).

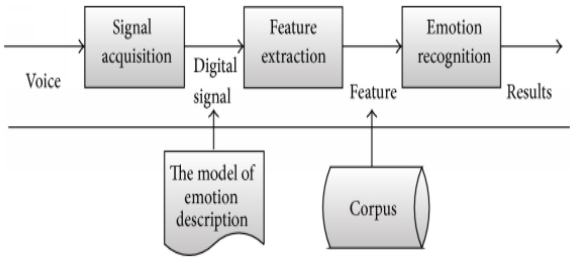

Figure 2: General scheme of an emotion recognition system [10]

2.2.1 Emotion recognition models

Classification models The same set of models already used for speech recognition have been applied to emotion recognition. These models include supervised learning techniques like HMM-GMM [13], ANN [12] and [9], SVM [17] and more recently DNN, LSTM and CNN [11]. In [4], it is reported that these models have been tested with approximately the same accuracy.

Another approach consists in using these classifiers in a combined scheme, either hierarchically, in series or in parallel. All these combined models were tested, giving nearly similar accuracy and outperforming single models, as reported by [4].

Clustering models Though clustering was less used than classification for emotion recognition, some works have proved that it could be successfully used to detect emotions. For instance, in [20] SOM were applied to detect emotions from audiobooks, based on articulatory features; whereas in [5], hierarchical k-means were used to detect emotions in a corpus for expressive speech synthesis, relying on a standard set of prosodic and acoustic features. Note that, in order to detect inherent clusters, in the unsupervised case the choice of features is more important than in the supervised case.

2.2.2 Features

The latter point, i.e. feature selection, is of a crucial importance. Actually feature selection depends on the problem. Generally in speech recognition, features can be divided into prosodic vs. acoustic (or spectral) ones. Prosodic features include f_0, intensity and phone duration, whereas acoustic features are extracted from spectrum such as MFCC (Mel-frequency cepstral coefficients), LSP (Linear spectral parameters), 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 those 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 [4]. Interspeech'09 feature set [18], GeMAPS [6] and Opensmile [7] are amongst the most used standard feature sets for emotion recognition, which use mainly global features. Still, feature selection using standard techniques, like ANOVA (Analysis of variance) and MI (mutual information) could be performed to choose the most contributory features.

2.2.3 Emotional speech databases

A variety of emotional speech databases were designed or recorded, covering more or less the emotion models described above, i.e. [3], [16], [14]. In [4], an inventory of emotional speech databases shows that the main differences between them lie in a) the size, varying from a few tens of sentences to a few thousands [8]; 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. In particular, EMO-DB database [2] has been widely used, since it covers all basic emotions in an equal and sufficient proportions.

3 Clustering methods

Clustering is the task of partitioning a set of data vectors or patterns into a fixed number of subsets, each defined by its centroid, so that each data sample is attributed to the subset which centroid is the closest. Clustering techniques could be inventoried following several criteria, whether they are hierarchical, partition-based, density/neighborhood-based or model-based [21]. However, partition-based grouping of clustering techniques looks to be the most adopted, since it allows discriminating crisp (hard) and fuzzy (soft) clustering. Furthermore, soft clustering techniques are also grouped into probabilistic vs. possibilistic ones.

3.1 Crisp clustering

Crisp clustering relies on the classical set theory, where an object either belongs or not to a cluster. In this case, clustering consists in partitioning objects into a fixed number of clusters.

Amongst the popular crisp clustering techniques, k-means is probably the most used one. The k-means algorithm proceeds as follows: Initially k centroids are randomly chosen, then repeat the following actions: a) assign each object to the group having the closest centroid and b) recalculate centroids from objects assigned to the groups, until centroids don't change or the maximum number of iterations is reached.

K-means main advantage is its scalability, since it is able to process a very large data set. Actually only centroids have to be stored in memory. Nevertheless, k-means presents three major drawbacks, first its inability to reach a global minimum, secondly the dependence of the solution on the initial values of the centroids, and at last the dependence of the solution on the order of the objects in the data set.

3.2 Fuzzy clustering

The soft/fuzzy clustering approach consists in using a membership function instead of discrete/binary membership decision. Then an object belongs to all clusters, but with different membership degrees, having values between 0 and 1 [1]. 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 $\forall x \in X$, $0 \leq \mu(x, \omega) \leq 1$.

Moreover, looking to the membership function, fuzzy clustering methods could be split into probabilistic and possibilistic ones. Then the pair (Ω, ω) is :

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

Among the most popular fuzzy clustering algorithms, c-means supports both crisp and fuzzy clustering. Actually fuzzy c-means is based on minimizing the following cost function:

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

where μ_{jl} is the degree of membership of pattern x_l to cluster ωj and d_{jl} is the euclidean distance between the pattern x_l and the cluster centroid y_j. Then the cluster centroid y_j is computed as follows:

$$y_j = \frac{\sum_{l=1}^{n} \mu_{jl} x_l}{\sum_{l=1}^{n} \mu_{jl}} \tag{2}$$

For c-means, the membership function is expressed by:

$$\mu_{jl} = \frac{v_{jl}}{Z_l} \tag{3}$$

In the particular case where $v_{jl} = e^{-d_{jl}\beta_j}$ and $\beta_j > 0$ (β_j is a cluster width parameter selected a priory), the generalized partition function Z_l is defined as $Z_l = f(\sum_{j=1}^{c} v_{jl})$. In this case, the function f defines the type of c-means, i.e.

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

where $\alpha \in [0, 1]$ is *the degree of probabilistic tendency*. Also, it should be noted that in case of probabilistic c-means, $\sum_{j=1}^{c} \mu_{jl} = 1$, whereas this condition is not necessarily met in possibilistic and graded-possibilistic c-means [15].

In [15], the following parametrizations are suggested:
$\alpha = 0$ for fully possibilistic c-means
$\alpha = 1$ for fully probabilistic c-means
$0 < \alpha < 1$ for graded-possibilistic c-means

Besides, the cluster width parameter β_j is suggested to be calculated as follows:

$$\beta_j = \frac{-\ln(t)}{min_{h \neq j} ||y_h - y_j||^2} \quad (4)$$

where $t \in [0, 1]$ is the threshold satisfying that $max_{h \neq j} \mu(y_h, y_j) \leq t$, i.e. minimal overlap condition. In particular $t = 1/2$ guarantees no overlap between clusters [15].

4 Speech material

To perform this work, an emotional speech database was selected from the available speech corpora. In particular, EMO-DB [2] has been widely used and cited as a reference emotion recognition database. Besides, choosing a feature set was addressed with a special attention, since several feature sets have been proposed in the literature.

4.1 Speech database

EMO-DB is a public available database of prepared emotional speech. Actually prepared speech corpora differ from spontaneous speech, 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 should be noted that EMO-DB has provided the highest emotion recognition rates using classical classifiers, such as HMM-GMM and SVM, as reported in [18].

4.2 Feature set

Since emotion recognition is a pattern recognition problem, different sets of features were proposed to solve such a problem. Actually most feature sets used the types of features described in section 2.2, i.e. prosodic, acoustic and in a lesser proportion articulatory features [20]. However, the global features, i.e. statistics calculated all over the speech signal, were generally preferred to local features, measured at each frame. Particularly, three features sets have widely been used for emotion recognition, i.e. Interspeech'09 feature set [18], ComParE [19] and GeMAPS [6]. though the three feature sets share some common descriptors, like f_0, loudness, MFCC, which are calculated using global statistics, the Interspeech'09 was preferred to conduct this work, for its preliminary results. Then features were extracted using Opensmile toolkit [7]. Table 1 and Table 2 show the complete set of features and the calculated statistics, so that 384 features ((16 descriptors + their 16 Δ-values) x 12 statistics) were extracted from each signal.

4.3 Classes

Initially, classes consisted in single emotions, namely neutral, anger, boredom, disgust, fear, joy and sadness. However, a second way to label speech signals was executed 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.

5 Experiments

The experimental process was achieved through the following steps: a) Preprocessing, where the final features were selected among the extracted ones, using either ANOVA or mutual information (MI) test; then fuzzy clustering parametrizations was set (β value for possibilistic and α value for graded possibilistic c-means). b) Fuzzy clustering, which was performed either on single emotions or on groups of emotions. c) Postprocessing, where clustering results were analyzed using principal component analysis (PCA) of features. Moreover, a finer analysis was achieved using the sum of memberships regarding each class.

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)

Table 1: Interspeech'09 emotion recognition challenge feature set [18]

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, MSE

Table 2: Statistical parameters used for Interspeech'09 emotion recognition challenge feature set [18]

5.1 Preprocessing

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

Parameter setting Fuzzy clustering parameters were set based on results reported in [15] where it was proven that $t = 1/2$ guarantees no overlap between clusters, and where it was also observed that α should be close to 1. Then it was suggested to set $\alpha = log_2(a+1)^2$ where $a \in [0.5, 1]$ to have $\alpha \in [0.9, 1]$.

5.2 Experimental protocol

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

- The label set, i.e. single emotions or groups pf emotions (cf. Table 3).
- The number of clusters, increasing from the number of classes, to 3 times.
- The feature selection method, i.e. ANOVA for individual features, ANOVA for a group of features (ANOVA-Group) 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 gave an high number of experiments, therefore only those giving the most relevant results are presented (cf. Table 4). 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.

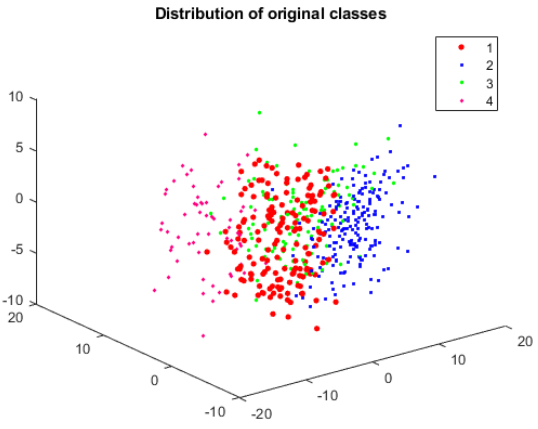

Figure 3: Distribution of original classes using PCA analysis of features (in case of 4 classes of grouped emotions)

5.3 Results

Figure 3 shows the 3-D distribution of original clusters using principal-component analysis (PCA) applied to

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

Table 3: Groups of emotions

# of classes	# of clusters	Feature selection method	Proportion of selected features	t	α	kmeans rate (%)	FCM 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

Table 4: Best recognition rates for different parameters combinations

features, whereas figure 4 shows the same PCA-based distribution for predicted clusters.

Initial distribution of classes It looks since the beginning that in spite of using a standard features set, as suggested by [18], the distribution of classes looks too scattered (cf. Figure 3).

Fuzzy clustering performance However, thanks to feature selection, and to a good choice of the possibilistic and graded-possibilistic models parameters, i.e. α and β, the fuzzy clustering methods outperform the crisp clustering one, i.e. kmeans (cf. Figure 4). Besides, using a number of clusters bigger than that of classes helps increasing performance. However, it should not be too big.

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 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.

5.4 Emotion analysis

In addition to emotion recognition, further analysis results could yield from applying fuzzy clustering. Actually, a matrix of the sum of memberships was calculated for each emotion, in the same way a confusion matrix is computed. However, instead of giving true positives, true negatives, etc., this matrix shows, for every detected emotion class, the sum of memberships to the original classes (cf. Table 5).

The analysis of the sum-of-membership matrix (cf. Table 5) shows for every recognized emotion class, the sum-of-memberships to original classes. Hence this could be interpreted as measuring the "purity" of an emotion, as expressed by speech signal features. For instance, in line 2, the recognized emotion *Anger* has a high sum-of-membership to the original class, i.e. *Anger*, whereas in line 4, for the recognized emotion *Disgust*, the sum-of-membership to classes *Neutral* and *Boredom* are as high as that to the original class, and the double of the sum of membership to class *Anger*. This result could be interpreted as follows: For the emotion *Anger* the selected features succeed to capture the original class, whereas for the emotion *Disgust*, either the selected features are not appropriate to detect such an emotion class, or the emotion *Disgust* is in fact a mixture of more basic ones, such as *Neutral*, and *Boredom* and in a lesser degree *Anger*. In all cases, such an analysis could be deepened further by subjective evaluation. Finally, the use of fuzzy clustering methods would be useful in corpus analysis and emotion labeling for expressive speech synthesis.

6 Discussion and conclusion

In his paper, a novel approach for emotion recognition using fuzzy clustering was described. The basic idea consisted in using 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

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**

Table 5: Sum-of-membership matrix calculated using 192 features selected by *MI*, FCM with $t = 0.1$ and $\alpha = 0.9$

was treated using k-means algorithm, for evaluation purposes. Several adjustments were also made to fine-tune the models, including feature selection, the use of a number of cluster higher than the number of classes and grouping classes which share common characteristics, and finally varying the possibilistic models parameters. Some of these modifications, mainly feature selection and using more clusters than classes, helped increasing the recognition rates. Also, choosing the optimal values of parameters had an impact on increasing the performance of possibilistic and graded-possibilistic c-means models. Moreover, and as an outlook, using fuzzy clustering, especially the analysis of the sum of the membership function could be an important tool for emotion analysis, since a single speech signal may convey more than one emotion.

Acknowledgement

This work was supported by the research grant funded by "Fondi di Ricerca di Ateneo 2016" of the university of Genova.

References

[1] R. Babuška, H. B. Verbruggen, An overview of fuzzy modeling for control, Control Engineering Practice 4 (11) (1996) 1593–1606.

[2] 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, 2005.

[3] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (3-4) (1992) 169–200.

[4] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.

[5] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. 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, 2012, pp. 4009–4012.

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE Transactions on Affective Computing 7 (2) (2016) 190–202.

[7] 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, 2010, pp. 1459–1462.

[8] J. H. 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, 1997.

[9] V. Hozjan, Z. Kačič, Context-independent multilingual emotion recognition from speech signals, International journal of speech technology 6 (3) (2003) 311–320.

[10] C. Huang, W. Gong, W. Fu, D. Feng, A research of speech emotion recognition based on deep belief network and svm, Mathematical Problems in Engineering 2014.

[11] 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, 2018.

[12] J. Nicholson, K. Takahashi, R. Nakatsu, Emotion recognition in speech using neural networks, Neural computing & applications 9 (4) (2000) 290–296.

[13] T. L. Nwe, S. W. Foo, L. C. De Silva, Speech emotion recognition using hidden markov models, Speech communication 41 (4) (2003) 603–623.

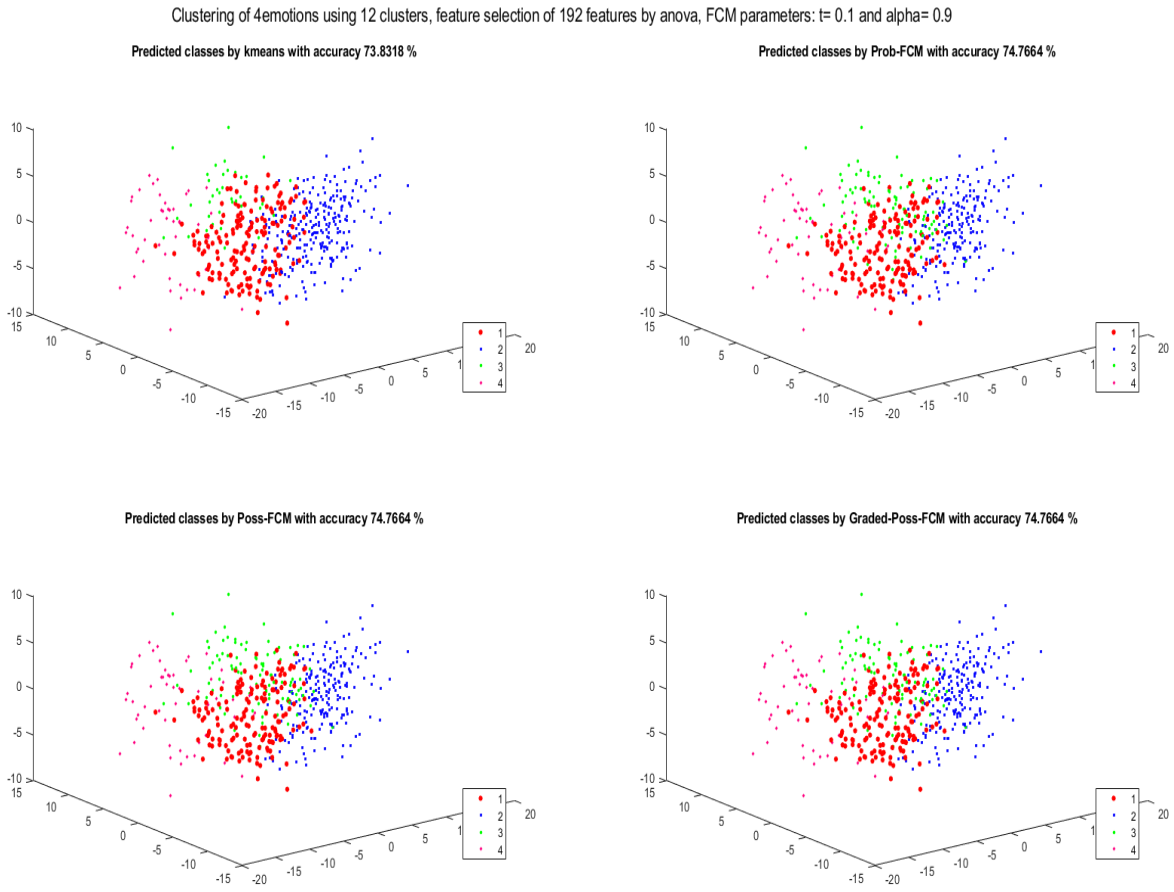

Figure 4: Emotion classification performance and distribution (in case of 4 classes of grouped emotions)

[14] 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, American scientist 89 (4) (2001) 344–350.

[15] S. Rovetta, F. Masulli, Soft clustering: why and how to, in: The 12th International Workshop on Fuzzy Logic and Applications (WILF 2018).

[16] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (6) (1980) 1161.

[17] 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, 2004. Proceedings.(ICASSP'04), Vol. 1, IEEE, 2004, pp. I–577.

[18] B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, in: Tenth Annual Conference of the International Speech Communication Association, 2009.

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

[20] 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 2011.

[21] G. Ulutagay, E. Nasibov, Fuzzy and crisp clustering methods based on the neighborhood concept: A comprehensive review, Journal of Intelligent & Fuzzy Systems 23 (6) (2012) 271–281.