A Study about the Statistical Parameters Used in the Emotion Recognition

Marius ZBANCIOC 1,2, Monica FERARU 1
"Institute of Computer Science" Romanian Academy of Iasi
str. T. Codrescu nr.2, RO-700481 Iasi
Technical University “Gheorghe Asachi” of Iasi
monica.feraru@iit.academiaromana-iasi.ro

I. INTRODUCTION

Many researchers want to develop and to improve automatic classification techniques. In the literature there are many studies regarding the vocal parameters in order to recognize the emotional states [1]. One of these parameters is fundamental frequency (F0), and even if there are lots of automatically extracting methods for this prosodic information, we can not say yet that it was found a complete algorithm for this problem [2]. From this parameter can be constructed features vectors which include mean of F0, max and min of F0, variance of F0, intensity distributions, and patterns for F0 rising segments [3]. New methods or combinations of methods are welcome to improve the current results.

There is a need now to bring the knowledge about the emotions recognition in the voice, in Romanian language for the existing development for English, German, etc.[3-5]. This need is felt in the linguistic domain, as well as science and computer technology, psychology and medicine. A few studies have been made on the Romanian language [11, 12]. Most studies were performed on happiness, anger, sadness, fear, disgust, panic, anxiety states. The recognition rates for a few emotions as 4-6 emotions “are considered successful around 70% and pure chance guess” when there is 16 emotions studied [10]. The focus on this study is on 3 emotions (happiness, sadness, and fury states) and neutral tone. The sentences are from the SRoL database and they contain all the vowels from the Romanian language. The speakers are feminine and masculine persons with ages between 20-30 years, without pathological manifestation and without professional voices.

Among the difficulties encountered in the studies regarding the emotion recognition/classification are the speaker variability, the emotional speech databases which most of them are private and can not be accessed and others. The emotion recognition can be made using different methods and algorithms. In [6], the researchers “proposed a new entropy-based measure which makes possible a comparison between human labelers and machine classifiers”. They obtained a recognition rate around 60% for 4 emotions and with 5 persons which validated the corpus. They said that first the emotions must be classified by humans. It is good to know also that the more similar emotions we have, the results will be more confusing.

K-Nearest Neighbor classifier (kNN) and Fisher’s linear classifier [7] were chosen as they are commonly used in the classification [4].

Frank Dellaert and his colleagues explored some standard pattern recognition techniques and compared their performance. In their study they used maximum likelihood Bayes classifier (MLB), Kernel regression (KR) and K-nearest neighbors (KNN). The best results are obtained using a K-nearest neighbor’s classifier [8].

In our study we used the algorithm Fuzzy C-Means (FCM) in order to make a classification of the vocal signal according with the emotional states for Romanian language. In addition to previous research, we have included in the set of features the values of jitter and shimmer.

After [9], “the jitter is a measure of period-to-period fluctuations in fundamental frequency and number of voiced frames in the utterance and the shimmer is a measure of the period-to-period variability of the amplitude value”. After [10] “the jitter was obtained by counting the number of changes in sign of the pitch derivative in a window and the shimmer was obtained by counting the number of changes in sign of the intensity derivative in a window”. 

II. EMOTION DATABASE VALIDATION

The emotional database is a part of the SRoL annotated corpus and contains at this moment 396 sound files (199 with male voices and 197 with female voices) pronounced for neutral tone and three simulated emotions states: joy, sadness and fury. It contains the recordings of 7 phrases pronounced by 25 speakers, 11 of them female and 14 male.

The pronounced sentences are: “Vine mama / Mother is coming”, “Aseară/ yesterday evening”, “Cine a făcut așta? / Who done that?”, “Ai venit iar la mine/ You came back to me”, “Omul meu il lucră / My man done it”, “Îți vei câștiga locul dorit / You will get the desired place”, “Oricum îți poți câștiga locul dorit / Anyway, you can get the desired place”.

Index Terms — emotional database, FCM, recurrent coefficient, statistical parameters.

Abstract — The purpose of this research is to find a set of relevant parameters for the emotion recognition. In this study we used the recordings from the emotion database SRO, which is part of the project “Voiced Sounds of Romanian Language”. The database was validated by human listeners. The recognition accuracy of the correct expressed emotion (neutral tone, joy, fury and sadness) for the entire database was 63.97%. We used for the classification of input data the Recurrent Fuzzy C-Means (FCM) algorithm. We compared the cluster position with the statistical parameters extracted from vowels in order to establish the relevance of each parameter in the recognition of the emotions.

DEVELOPMENT AND APPLICATION SYSTEMS
International Conference on
Technical University “Gheorghe Asachi” of Iasi
Monica.feraru@iit.academiaromana-iasi.ro
Each phrase is pronounced for several times (on average about 4-5 times in the recording file). Thus we can estimate that the total number of sentences is about two thousands.

After the corpus validation made by three people we obtained two type of database: one (named DB100) in which all valuators have indicated the same emotional state and another (DB_validated) in which two of the three valuators (the majority of listeners) have confirmed the same emotion.

The valuators have the possibility to listen the wav. files how many time they want, the option to return to previous records and also the possibility to make the validation in sections. In fig. 1 is exemplify the interface between the valuators and the computer. When a human listener is not sure about the expressed emotion then he can use the "ambiguous" button.

![Image](image_url)

**Fig. 1** A screenshot of the emotion validation interface

The valuators have the possibility to listen the wav. files how many time they want, the option to return to previous records and also the possibility to make the validation in sections. In fig. 1 is exemplify the interface between the valuators and the computer. When a human listener is not sure about the expressed emotion then he can use the "ambiguous" button.

**TABLE I. CONFUSION MATRIX FOR FEMALE VOICES VALIDATION (IN PERCENTS)**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Neutral</th>
<th>Joy</th>
<th>Fury</th>
<th>Sadness</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>60.99%</td>
<td>2.13%</td>
<td>4.96%</td>
<td>30.50%</td>
<td>1.42%</td>
</tr>
<tr>
<td>Joy</td>
<td>3.85%</td>
<td>80.13%</td>
<td>9.62%</td>
<td>2.56%</td>
<td>3.85%</td>
</tr>
<tr>
<td>Fury</td>
<td>4.58%</td>
<td>10.46%</td>
<td>73.20%</td>
<td>8.50%</td>
<td>3.27%</td>
</tr>
<tr>
<td>Sadness</td>
<td>12.77%</td>
<td>2.84%</td>
<td>2.13%</td>
<td>78.72%</td>
<td>3.55%</td>
</tr>
</tbody>
</table>

**TABLE II. CONFUSION MATRIX FOR MALE VOICES VALIDATION**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Neutral</th>
<th>Joy</th>
<th>Fury</th>
<th>Sadness</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>59.03%</td>
<td>3.47%</td>
<td>6.25%</td>
<td>20.83%</td>
<td>10.42%</td>
</tr>
<tr>
<td>Joy</td>
<td>22.00%</td>
<td>54.67%</td>
<td>10.00%</td>
<td>3.33%</td>
<td>10.00%</td>
</tr>
<tr>
<td>Fury</td>
<td>24.53%</td>
<td>5.66%</td>
<td>61.01%</td>
<td>3.77%</td>
<td>5.03%</td>
</tr>
<tr>
<td>Sadness</td>
<td>39.58%</td>
<td>5.56%</td>
<td>4.17%</td>
<td>43.06%</td>
<td>7.64%</td>
</tr>
</tbody>
</table>

**TABLE III. NUMBER OF VOWEL OCCURRENCES FOR FEMALE VOICES (NUMBER OF FILES / NUMBER OF VOWELS)**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Neutral</th>
<th>Joy</th>
<th>Fury</th>
<th>Sadness</th>
<th>Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>4 / 25</td>
<td>10 / 97</td>
<td>8 / 62</td>
<td>8 / 76</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>5 / 32</td>
<td>10 / 104</td>
<td>3 / 27</td>
<td>7 / 22</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>6 / 34</td>
<td>10 / 191</td>
<td>8 / 132</td>
<td>9 / 141</td>
<td></td>
</tr>
<tr>
<td>a+</td>
<td>7 / 27</td>
<td>8 / 67</td>
<td>8 / 62</td>
<td>7 / 46</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>5 / 37</td>
<td>6 / 74</td>
<td>8 / 73</td>
<td>7 / 40</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>5 / 50</td>
<td>4 / 36</td>
<td>6 / 39</td>
<td>2 / 21</td>
<td></td>
</tr>
<tr>
<td>a-</td>
<td>5 / 29</td>
<td>4 / 44</td>
<td>6 / 53</td>
<td>2 / 29</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE IV. NUMBER OF VOWEL OCCURRENCES FOR MALE VOICES (NUMBER OF FILES / NUMBER OF VOWELS)**

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Neutral</th>
<th>Joy</th>
<th>Fury</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>6 / 23</td>
<td>4 / 19</td>
<td>4 / 41</td>
<td>3 / 10</td>
</tr>
<tr>
<td>i</td>
<td>7 / 26</td>
<td>7 / 55</td>
<td>4 / 44</td>
<td>5 / 25</td>
</tr>
<tr>
<td>a</td>
<td>7 / 40</td>
<td>3 / 10</td>
<td>7 / 78</td>
<td>5 / 19</td>
</tr>
<tr>
<td>a+</td>
<td>3 / 6</td>
<td>5 / 27</td>
<td>7 / 27</td>
<td>2 / 5</td>
</tr>
<tr>
<td>o</td>
<td>2 / 10</td>
<td>4 / 29</td>
<td>6 / 25</td>
<td>4 / 12</td>
</tr>
<tr>
<td>u</td>
<td>1 / 8</td>
<td>4 / 37</td>
<td>3 / 20</td>
<td>3 / 15</td>
</tr>
<tr>
<td>a-</td>
<td>1 / 12</td>
<td>7 / 40</td>
<td>3 / 23</td>
<td>3 / 21</td>
</tr>
</tbody>
</table>

In Table III and IV is specified the number of occurrences for every vowel depending on the expressed emotion for the DB100 validated database.

**III. FEATURE VECTORS AND STATISTICAL PARAMETERS**

In this study we used only the parameters extracted from the vowels of the Romanian language with Praat software. Each sound file (wav) was annotated manually generating a file with the TextGrid extension. In the statistical analysis was used only the “phoneme” annotation level, which specify the duration and the position of each vowel phoneme. Praat is a free software for acoustic analysis which can extract the pitch F0, the F1-F4 formants, the pulses and the intensity of the voiced signals.

The pitch is extracted using an analysis window of 40ms and a time step of 10ms in the frequency band [75, 500]Hz with the autocorrelation method, because compared with other methods is more robust, accurate and noise-resistant.

The formants F1-F4 are computed using a window length of 25ms and a time step of 6.25 ms using the LPC coefficients obtained with the Burg algorithm.

The statistical parameters are computed for each vowel (their boundaries were drawn in the annotation files). In the feature vectors are included the mean, the standard deviation and the median of F0 and of the formants F1-F4. The pitch is extracted using an analysis window of 40ms and a time step of 10ms in the frequency band [75, 500]Hz with the autocorrelation method, because compared with other methods is more robust, accurate and noise-resistant.

The formants F1-F4 are computed using a window length of 25ms and a time step of 6.25 ms using the LPC coefficients obtained with the Burg algorithm.

The statistical parameters are computed for each vowel (their boundaries were drawn in the annotation files). In the feature vectors are included the mean, the standard deviation and the median of F0 and of the formants F1-F4. The median is the central value of the sorted input vector.

In addition to previous research, we have included in the set of features the values of jitter and shimmer.

Jitter is the average variation of the F0, expressed as the difference between two consecutive pitch period T0=1/F0.

\[
\text{Jitter (phoneme)} = \frac{1}{N - 2} \sum_{k=3}^{N} |T_k - T_{k-1}| \\
= \frac{1}{N - 2} \sum_{k=3}^{N} (P_k - P_{k-1}) - (P_{k-1} - P_{k-2}) \\
= \frac{1}{N - 2} \sum_{k=3}^{N} [P_k - 2P_{k-1} + P_{k-2}]
\]

where \(N\) is the number of pulses corresponding to the duration of a phoneme/vowel. The pulses are used by Praat to compute the pitch time period and we stored these values in a file.

Shimmer represents the average variability of the peak-to-peak amplitude between two consecutive pitch periods.

\[
\text{Shimmer} = \frac{1}{N - 2} \sum_{k=4}^{N} \frac{20 \cdot \log_{10} A_k}{A_{k-1}} \\
= \frac{1}{N - 2} \sum_{k=4}^{N} \max(s(P_k : P_{k-1})) - \min(s(P_k : P_{k-1})) \\
= \frac{1}{N - 2} \sum_{k=4}^{N} \max(s(P_{k-1} : P_{k-2})) - \min(s(P_{k-1} : P_{k-2}))
\]

where \(N\) is the number of pulses \(P_k\) and \(s\) is the input signal.
IV. RECURRENT FUZZY K-MEANS ALGORITHM

The clustering algorithms divide the input sets based on their similarity. The most known are the hierarchical algorithms (agglomerative and divisive) and the partitioning algorithms. Among the partitioning algorithms, we mention EM (Expectation Maximization) based on the probabilistic models, QT (quality threshold), algorithms based on graph theory and K-means based on the calculation of Euclidean distance and squared distance error. From K-means algorithm was developed a series of other algorithms including the FGKA genetic algorithms (Fast Genetic K-means Algorithm) and fuzzy techniques such as Fuzzy C-means (FCM).

For the data clustering, we used an improved version of the FCM algorithm which gives a confidence coefficient to each input pattern associated with every center of a cluster/fuzzy partition. These coefficients are computed with the recurrent function in each new iteration of the algorithm. The FCMR algorithm (fuzzy c-means recurrent) eliminates one of the major disadvantage of the classical FCM algorithm, related to the fact that the learning is unsupervised and does not know the real number of clusters from the input data set. In this situation, the cluster centers found by FCM not converge to the real centers, which are compensated in the new algorithm FCMR by introducing the recurrent confidence coefficients.

In [13-15] Teodorescu proposed a model of the recurrent fuzzy systems with the justification that people tend to adjust their judgments based on the preliminary results. Thus, the knowledge (rules) that lead to achieve worst results should be associated with the lower coefficients compared with those that achieve better results. Because the recurrent fuzzy systems - Teodorescu model is a generalization of the existing classical systems (Mamdani, Sugeno, etc.), any classical fuzzy system (fuzzy algorithm) in any area of the analysis, can be transformed into a fuzzy model with recurrence to the confidence degree in meta-knowledge / rules [16].

The input data set \( \hat{x} = \{x_1, x_2, \ldots, x_n \} \) are vectors of \( n \) features, and the distance between two vectors is computed based on the Euclidean distance, but we can use any kind of distances, as Mahalanobis, Pearson, Manhattan, Chebyshev, Lee, Hamming, etc.. We note with \( N \) the size of the input data set \( X = \{\hat{x}_k\}, \ k = 1, N \). The FCM algorithm determines the distance from the each cluster center \( c_j \) to each feature vector \( \hat{x}_k \), and it associated a membership degree \( \mu_{jk} : R \rightarrow [0,1] \). We also note with \( C \) the number of clusters/partitions and with \( U \) the matrix (of size \( C \times N \)) which keeps all the membership grades.

The FCMR algorithm pseudocode is the following:

1. The matrix initialization with random numbers \( U^{(0)} \) and its normalization, setting the number of clusters \( 2 \leq C \leq N_1 \), the error convergence of the algorithm \( \varepsilon \) and the exponent \( m \) applied to the membership degrees. In addition, a matrix of the recurrent confidence coefficients is introduced: \( \hat{C}^{(0)}_{jk} = 1 \), \( t=0 \);

2. A set \( C \) of partitions’ centers \( \{\hat{c}_j^*\} \) is computed using a relationship with the recurrent confidence coefficients. The distances \( \{d_{jk}^*\} \) from each element of the \( X \) data set to each center are computed;

3. Based on the distances, the objective function \( J^* \) and the value of the membership grades for next step \( U^{(t+1)} \) are computed;

4. With the new values of the matrix \( U^{(t+1)} \) the recurrent coefficients for the \( t+1 \) time is updated;

5. If the differences in the objective function \( |J^{(t+1)} - J^{(t)}| \) are above the imposed convergence limit ( \( \varepsilon \) imposed error) and the maximum number of the iterations is not reached, then return to step 2.

By simulation, it can be concluded that the performance of the FCMR proposed algorithm may be improved if the confidence coefficients are not applied from the beginning but with a certain delay after a consistent movement of the cluster’s centers has been observed.

Figure 2 shows the situation when we have a real number \( M=3 \) clusters and \( C=2 \) specified clusters of the partitioning algorithm. The classical FCM algorithm places a center between two clusters while the FCMR algorithm achieve all three clusters’ centers and the FCMR algorithm applied with "delay" is more robust and frequently finds only two clusters. The simulations have been made by running the algorithms for 50 times.

Thus the FCMR algorithm partially solves another disadvantage of the classical K-means and FCM algorithms, which can converge only to a local, not global maximum and that the final values of the cluster centers depends on the selected values of the partitions / clusters’ centers.

V. RESULTS AND CONCLUSION

Combining the emotional states (‘Joy’, Sadness’, ‘Fury’ and ‘Neutral’) with the gender of the speakers (‘Female’ and ‘Male’) are obtained \( N_s = 8 \) situations notated with ‘JF’, ‘JM’, ‘SF’, ‘SM’, ‘FF’, ‘FM’, ‘NF’, ‘NM’. The structure DB where are stored the feature vectors has the dimension \([N_s \times N_v \times N_p]\). \( N_p=23 \) is the number of statistical parameters...
and \( N_v = 7 \) the number of the vowels (analyzed phonemes). We compared in the bi-dimensional space of characteristics F1 and F2 the cluster centers (in Fig.4) with the mean values of each vowel (in Fig.3). We have observed that formant space is properly partitioned; the FCMR is robust and places the partition centers in the same areas (Fig. 5). For the n-dimensional space of features \{F0, F1-F4, jitter, shimmer\} we computed the accuracy of classification of emotions, but results have been around a percentage of 41\%, for the database validated by human listeners with an accuracy recognition percent of 63.97\%.

From simulations it was observed that higher formants F3-F4 not help to differentiate emotions. By eliminating the superior formants from the features vectors were obtained the same results in classification.

The SRoL corpus contain a sets of software instruments for pitch extraction (using four methods: auto-correlation, cepstral, HPS - Harmonic Product Spectrum and AMDF - Average Magnitude Difference Function) and for the detection of formants using concatenated parts of “smoothed” spectrum. In this paper we used only features extracted from Praat files, but we intend to expand this research using our instruments for pitch detection.

For future research will take into account only the parameters coming from phonemes located in fixed positions, preferably in the center of sentences where the emotion is expressed better. Another approach which will be analyzed consists in extracting parameters not at phoneme level, but for an entire word or phrase.

ACKNOWLEDGMENTS

The authors thank all those who contributed with recordings and colleagues that contributed to the building of the website SRoL. The authors acknowledge the support of the Romanian Academy in the framework of the program “Cognitive systems”.

REFERENCES