4th International Conference on Advanced Technologies
For Signal and Image Processing – ATSIP' 2018
March 21-24, 2018 – Sousse, Tunisia

ON THE USE OF PITCH-BASED FEATURES FOR FEAR EMOTION DETECTION FROM SPEECH

Safa Chebbi[1] and Sofia Ben Jebara[1]

***Abstract*—In this paper, we present a study that aims to evaluate the effect of pitch-related features on fear emotion detection from speech signal. In this context, several features have been tested. Only relevant ones are selected thanks to ANOVA tests. Next, they were decorrelated using principal component analysis. To select fear, emotion classification based on machine learning methods is used to extract fear from other emotions. Many classification tools are used and compared. We considered two types of emotion classification which highlights the fear emotion state, a simple classification as well as an hierarchical one. Results show that selected pitch-based features have a relatively great power in fear recognition. In fact, the highest accuracy rate reaches 78.7% using k-nearest neighbors algorithm.**

***Index Terms*—speech; emotion reognition; fear; pitch; classification**

I. Introduction

Emotion recognition represents a complex research field with a large background in the communities of life and social sciences. In this context, a growing interest in building human computer interfaces for automatic emotion recognition can be observed. Various bio-signals are hence used in emotion recognition such as electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), Electro dermal response (EDR), blood volume pulse ...

Moreover, people express emotion through some behavioral modalities such as facial expression, body postures and gestures and voice. As speech is one of the most fundamental and natural communication mean of human beings, it usually comes first to mind when thinking about possible sources to exploit for emotion detection. Speech emotion recognition is particularly useful for man-machine interaction applications in which the emotional state of the speaker plays an important role. For example, in tutoring systems which detect the learner's state and adjust the presentation style according to the detected mood [1], it is also useful in call center systems for detecting consumer's states and enhancing the service quality [2]. It can also be employed even in medicine as a diagnostic tool for detecting the emotional state of the patient during consultations [3] or even for emotion recognition of autistic individuals [4]. Speech emotion recognition has also been used in mobile communication which measures the degree of affection from the person's voice on the mobile phone [5].

Among many kinds of emotion (joy, anger, fear, disgust, happiness...), we addressed in this study fear emotion recognition through speech modality. It's useful for applications where physical and physiological aggression against human beings degrade their life. We related for example child maltreatment by a bully at school or even by a parent-in-law or caregiver at home. It can also be useful for detecting criminals during investigations, capturing smugglers in airports or ports or even detecting abnormal situations via surveillance cameras. In such cases, we need rapid intervention in order to protect victims and limit damage.

Many features have been explored in speech emotion recognition [6][7][8][9][10][11][12][13]. They are categorized into 4 classes: i) prosodic features [14,15,16] modeling the accent, the rhythm, the intonation and the melody of voice, ii) voice quality features [17,18] whose variation indicates physiological changes caused by emotion, iii) Spectral features [19] describing vocal folds and vocal track behavior and iv) perceptual features [20] related to the way of perceiving emotion through speech.

In this study, we targeted the recognition of fear using prosodic features specifically pitch-related ones. In fact, according to physiology, a study of the emotion production mechanism [21] has shown that fear sensation stimulates the sympathetic nervous system. Indeed, it generates a greater blood pressure, an increased heart rate, higher sub-glottal pressure, dryness of the mouth, and occasional muscle tremor. As a result vocal cords are impacted so that comes the idea of studying features related to vocal cords to detect fear.

In this study, we present a detailed research that evaluates the effect of pitch-based features on fear recognition. As a first step, a large amount of features derived from pitch are extracted. Then most relevant ones are selected thanks to ANOVA statistical tests. In order to validate the usefulness of features, four classification tools are used (decision tree, k-nearest neighbors, support vector machine and subspace discriminant analysis). Overall features are used for classification and then their dimensionality is reduced. Our approach consists on decorrelating features and eliminating the ones with less relevance using principal component analysis. Two types of emotion classification were adopted: simple and hierarchical one and then a comparison between them was carried out in order to give off the most appropriate classification method.

This paper is organized as follows: In the first part, we will describe the whole set of pitch-related features that have

*This work was not supported by any organization

[1] COSIM Lab. Higher School of Communications of Tunis, Carthage University, Route de Raoued Km 3.5, Cité El Ghazala, Ariana, 2083, Tunisia safa.chebbi@supcom.tn , sofia.benjebara@supcom.rnu.tn

978-1-5386-5239-8/18/$31.00 ©2018 IEEE

been tested in our study. In the second part, we will identify the emotional database used during our study and the adopted emotion classes as well. The third part will be devoted to indicate the effect of pitch-based features on discriminating between different emotion classes thanks to statistical test results. Finally, we will describe the followed approach in order to identify the highest accuracy rate and present the classification results using machine learning algorithms.

II. FEATURE SET CALCULATION

Pitch is a feature related to vocal folds vibration. It's defined as the number of periods of vocal folds vibration per second. In this paper, we are interested in pitch-based features for fear emotion recognition since some modifications in speech during fear are due to vocal folds vibration (irregular voice, tremor, oscillation ...). The pitch was calculated frame by frame based on the robust algorithm for pitch tracking ”RAPT”[22]. Fig.1 shows an example of temporal evolution of speech signal for one utterance and its corresponding pitch graph.

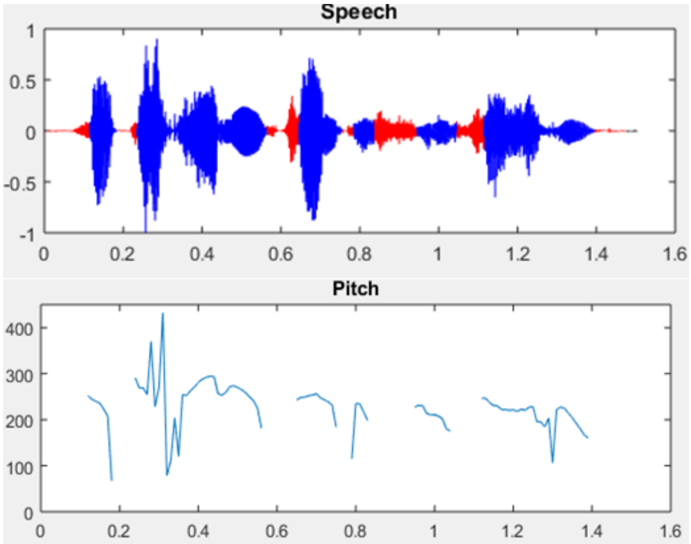

Fig. 1. Speech signal and pitch

Among a very wide range of features developed in the literature (hundreds) which were summarized by family in the introduction, the objective of this section is to select the most relevant ones leading to an accurate recognition of fear emotion. To attend such purpose, we have considered features with a global vision. It means that each feature is extracted from the whole sequence. Selected Pitch-related features are categorized as follows:

Usual measures: mean, maximum, minimum, variance, median, normalized standard deviation.

Features related to speech voicing: Number of voiced (resp. unvoiced) frames in the whole sequence divided by the total number of frames in the sequence, ratio of voiced frames on unvoiced frames, pitch of the first voiced frame, pitch of the last voiced frame, pitch of the second voiced frame, pitch of the before last voiced frame, pitch of voiced frame in the middle position.

Pitch contour derivative: mean of pitch's derivative, mean of the absolute value of pitch's derivative, variance of pitch's derivative, variance of the absolute value of pitch's derivative, maximum of pitch's derivative, maximum of the absolute value of pitch's derivative, maximum of pitch's second derivative, mean of pitch's second derivative.

And others: platitude expressed as the ratio of pitch's mean on its maximum, vehemence represented by the ratio of pitch's mean on its minimum, ratio of number of peaks on total of frames, maximum (resp. minimum) position calculated with regard to the total number of frames .

III. EMOTIONAL SPEECH CORPUS AND SELECTED EMOTION CLASSES

A. Emotional Speech Corpus

In our study, we used The EMO Database [23] which is a German emotional speech corpus containing more than 500 utterances. Ten actors, 5 males and 5 females, were asked to simulate 7 emotion states: fear, neutral, anger, boredom, disgust, joy and sadness. The total number of the text sentences to be read is 10 German utterances, 5 short and 5 longer sentences, which could be used in everyday communication and are interpretable in all applied emotions. Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz.

B. Selected Emotion Classes

Apart from our targeted emotion, we must consider other emotion types in order to study the fear behavior and then ensure its distinction from other emotional states. In this context, we opted to choose 2 types of emotion grouping for fear detection.

-A first classification with one level considering 3 groups: fear, neutral and other emotions. The ’Other emotions’ class includes the five remaining states (joy, anger, disgust, sadness and boredom). These classes repartition through the corpus is the following: 14% for fear, 14% for the neutral class and 72% for other emotions.

-A second hierarchical classification with 2 levels considering in the first level the 3 groups: positive emotions, negative emotions and neutral. Positive class includes joy's sequences. The second level divides the negative emotions group into 2 partitions: fear and other negative emotions. This latter contains anger, boredom, sadness and disgust emotional states. This second hierarchical classification presents a finer vision compared to the first one with only one level. The repartition of the first level classes through the corpus in terms of sequences is as follows: 73% for negative class, 14% for neutral class and 13% for positive emotions. For the second level, we have 14%

for fear and 59% for other negative emotions. An illustrative diagram explaining the two types of emotion classification is provided in the following figure (Fig2).

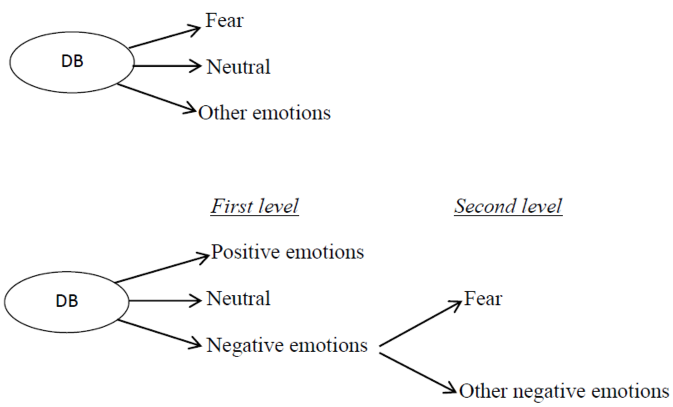

Fig. 2. Emotion classification types.

IV. RELEVANCE OF PITCH-BASED FEATURES

A. Pitch Effect for a Fixed Speaker

In order to illustrate the usefulness of pitch-based features to detect fear emotion, we considered one example of a speaker who pronounced different sentences in different emotion conditions. Pitch values are calculated frame by frame for each emotion and the mean value is extracted for each emotion sentence. As a result, we obtain a number of mean values as much as we have sequences simulated by that fixed speaker for each emotional class. Fig. 3 shows the repartition of pitch's mean for each emotionnel class. Each point of the graph represents one value of pitch's mean for a sequence pronounced by this speaker.
Fig.3 shows that pitch's mean differentiates clearly between different emotion classes. For simple classification, mean pitch values of fear class are scattered between 200Hz and 320Hz while they take a wider range area for other emotions class going from 100Hz to 300Hz. For the neutral class, mean values are dispersed on a more compacted area (compared to other classes) between 140Hz and 200Hz. For hierarchical classification, mean pitch values of positive emotions class are scattered between 160Hz and 320Hz while it takes a wider range area for negative emotions class going from 100Hz to 300Hz. For the neutral class, mean values are dispersed on a more compacted area compared to other classes between 140Hz and 200Hz. Dealing with the second level, mean pitch value reaches 320Hz as a higher threshold for fear emotion while it does not exceed 280Hz for other negative emotions class.

As a conclusion, we confirm that pitch distinguitches well the targeted emotion of fear from other states and can present a discrimination feature between them.

B. Boxplots

In statistical analysis, the boxplot is a useful tool for studying large sets of data. It provides a graphical comparison between different classes. In our case, it deals with the discrimination power of pitch's mean between different emotion classes. In Fig. 4, we drew the boxplots of pitch's mean for both simple and hierarchical classification. In the first case, we can see that it presents a clear difference between classes and especially between fear and neutral states. In the second case, boxplots revealed a great discrimination between positive, negative and neutral classes. The same thing is observed for the second classification level for fear and other negative emotions.

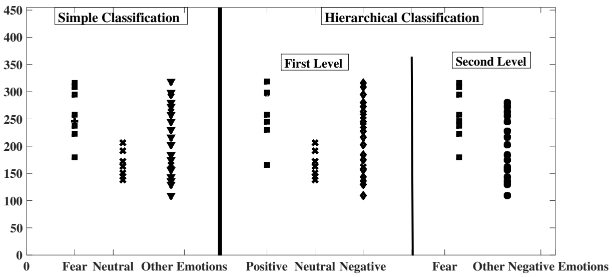

Fig. 3. Mean pitch values according to the considered classification

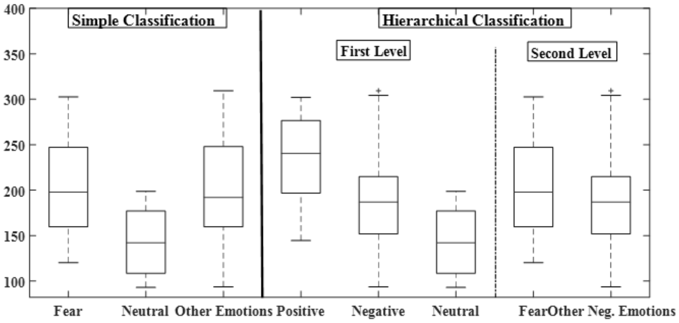

Fig. 4. Boxplots of mean pitch

C. Feature Set Preselection Based on Statistical Test

The aim of this subsection is to look for relevant features which permit a significant difference between the different groups of emotion. In other words, the feature which presents a significant difference between the different emotions is considered discriminant and is added to the selected list. In a first step, we collected the whole set of 30 features derived from pitch. All these features have a global vision as they are calculated towards the whole sequence. The final list of preselected features is obtained through the result of the statistical test ANOVA.
In statistics, ANalysis Of VAriance (ANOVA)[24] is a collection of statistical models used in order to analyze the differences between group means and their associated procedures (such as variation among and between groups). The normality was tested using the Shapiro Wilks test for all dependent variables. One-way ANOVA test was used to determine the effect of each feature on discriminating between different

emotion classes. Relative changes (%) in dependent variables are expressed with 95% confidence interval (95% CI). The test was used to separate emotion states with significance level equal to 5%. In other words, if pvalue is less than 0.05, it means that there is a significant difference between compared classes. If no, there is no significant difference between them. Tab. I lists all the tested features and gives the results of discrimination power between classes by giving p-values. Bold police is used to show p-values leading to discrimination ability.

TABLE I
DISCRIMINATION'S POWER OF PITCH-BASED FEATURES

Acoustic Features	**Simple Classification**	**Hierarchical Classification**	
		1st Level	*2nd Level*
Mean F0	**<0,001**	**<0,001**	**0.006**
Median F0	**<0,001**	**<0,001**	**0.005**
Variance F0	**<0,001**	**<0,001**	**0.015**
Normalised Standard Deviation	**<0,001**	**<0,001**	**0.001**
Min F0	0,05	0,06	0.09
Max F0	0,07	0,05	0.06
Platitude Pitch	**<0,001**	**<0,001**	**0.06**
Vehemence Pitch	**<0,001**	**0,001**	**<0,001**
Number of Voiced Frames /Number of Unvoiced Frames	0,19	0,08	0.43
Number of voiced Frames / Total Number Of Frames	0,19	0,17	0.43
Number of Unvoiced Frames / Total Number Of Frames	0,26	0,08	0.43
Number Of Peaks/Total Number Of Frames	**<0,001**	**0,001**	**0,002**
First voiced frame	**<0.001**	**0,01**	**0.015**
Last voiced frame	0,004	0,1	0.2
Second voiced frame	**0,003**	**0,004**	**0.05**
before last voiced frame	**0,002**	**0,05**	**0.004**
voiced frame in the middle position	**<0.001**	**<0.001**	**0.041**
position of the maximum	0,08	0,29	0.1
position of the minimum	0,9	0,8	0.93
mean of pitch's derivative	0,61	0,96	0.32
mean of the absolute value of pitch's derivative	**0,003**	**<0.001**	**0.05**
variance of pitch's derivative	0.12	0,9	0.68
variance of the absolute value of pitch's derivative	0,3	0,17	0.55
maximum of pitch's derivative	0,8	0,95	0.53
maximum of the absolute value of pitch's derivative	0,56	0,64	0.61
maximum of pitch's second derivative	0,76	0,86	0.004
mean of pitch's second derivative	0,024	0,7	0.99

V. CLASSIFICATION AND FEAR EMOTION DETECTION

A. Reduction of Dimensionnality of Feature's Vectors Using Principal Component Analysis (PCA)

Principal component analysis method [25] represents a statistical procedure that uses an orthogonal transformation in order to convert the set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the condition that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. In our case, we use principal component analysis in order to eliminate correlation between used features. We obtain as a result a reduced base of decorrelated vectors. In this way, we reduce the complexity induced by many factors such as: the base's size, dependency between features and redundancy of information.

Then, our approach consists on bringing out for each iteration the classification performance using the first p principal components and give off as a result the optimal number which leads to the best classification rate. In other words, in the first iteration for example we do the classification using only the first component, then we test with the two first principal components. The process is re-iterated everytime by incresing the number of considered components until reaching the use of the whole set of components. The component's group giving the best classification performance is then identified. In this way, we avoid considering components which provide no additional information and leading on the contrary to worse classification results.

B. Identification of the Most Appropriate Classification Algorithm

Automatic classification of emotions is based on machine learning methods. These methods are built on a learning approach which is able to characterize an emotion class from a sufficient amount of data. Many classification techniques are developped in the litterature (see for example [26]).In this study, we performed the classification using different classifiers such as decision tree (DT) [27], Support Vector Machine (SVM)[28], K-nearest neighbors (KNN) [29] and Subspace discriminant algorithms [30]. The performance of classification was judged by the accuracy rate. It translates the percentage of well predicted emotion class among the total number of emotion samples.

For a fixed algorithm, the method accuracy calculation depends on the adopted emotion classification type . For the simple classification, we simply take the result as it is. For the hierarchical one, it's the product of the two accuracies obtained separately in each level as it is modeled by the following equation:

$$Accuracy\ for\ hier.\ classification = \frac{N_2}{N_1} \times \frac{N_1}{N_T} \qquad (1)$$

Where N1 represents the number of well predicted samples in the first level.

N2: designates the number of well predicted samples in the second level.

NT: is the total number of all emotion samples

Thus, the accuracy rate for the hierarchical classification would be calculated by multiplying the accuracy rate obtained independantly in the second level by the one obtained for the first level on whose depends.

The data was divided into training and validation data. 70% were used for training while 30% were used for testing. Fig.5 (resp. Fig. 6) presents classification results obtained by each classifier for simple (resp. hierarchical) emotion classification type. The x-axis shows the number of first PCA components (p = 1 .. 12) while the y-axis displays the accuracy rate obtained with each algorithm. The two figures leads to the following results:

-Decision Tree algorithm (DT): It turns out that the best accuracy rate has reached 72% using 6 first components for the simple classification type (fear/ neutral/other emotions), while it was 51.74% for the hierarchical classification obtained by the first component.

-K-nearest neighbors algorithm (KNN): It was found that the best accuracy rate has reached 78.7% using the 8 first components for the simple classification type while it didn't go over 57.1% for the hierarchical one using the 9 first components.

-SVM algorithm: It turns out that the best accuracy rate has reached 77.3% using the 6 first components for the simple classification type, while it was 56.44% for the hierarchical classification obtained by the three first components.

-Subspace Discriminant algorithm: It was found that the best accuracy rate has reached 72% using the 8 first components for the simple classification type while it didn't go over 59.32% for the hierarchical one using the 4 first components.

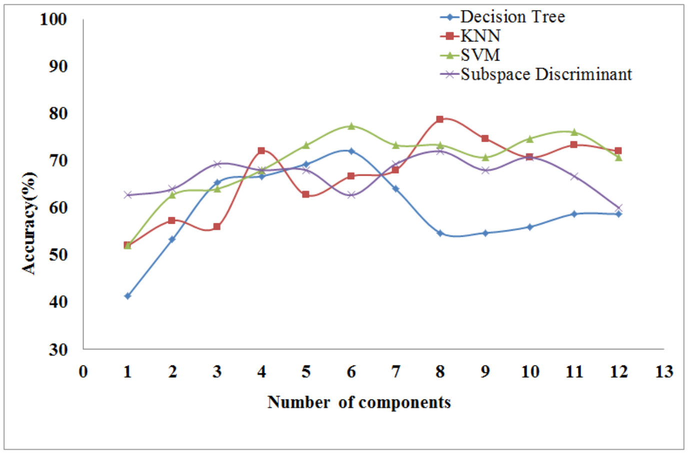

Fig. 5. Simple emotion type classification results

The study reveals that KNN provides the best accuracy rate of 78.7% for the simple classification type. For the hierarchical one, subspace discriminant classifier gives the highest accuracy rate with a percentage of 59.32%.

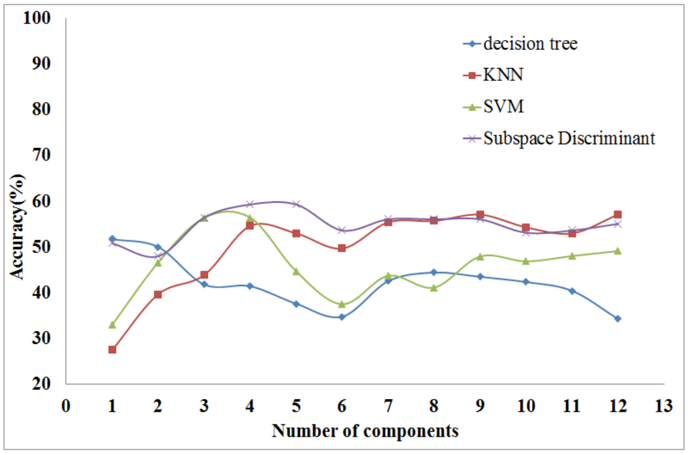

Fig. 6. Hierarchical emotion type classification results

C. The Effect of Grouping Emotions Hierarchism on Classification Performance

The aim of this subsection is to identify the effect of adopting hierarchism in emotion grouping on the classification performance. Tab. II compares the accuracy of the simple and the hierarchical classification using the different algorithms of classification. It turns out that simple classification is the best. In fact, hierarchical classification has a major drawback of accumulating errors of the levels.

Tab. III (resp. tab. IV) shows the confusion matrix for simple (resp. hierarchical) classification types. The rows and the columns represent respectively the true and the predicted emotion classes. For example, second row in tab. III says that 53.3% of utterances that were portrayed as fear were evaluated as fear, 20% as neutral and 26.7% as other emotions. We can also see that the most easily recognizable category is other emotions (93.6%). The results mentionned in these tables show a clear drop in the accuracy rates when passing from simple to hierarchical classification.

TABLE II
COMPARISON BETWEEN SIMPLE AND HIERARCHICAL CLASSIFICATION RESULTS

	Simple Classification	**Hierarchical Classification**
DT	72%	51.74%
KNN	78.7%	57.1%
SVM	77.3%	56.44%
Subspace Discriminant	72%	59.32%

the results mentioned in these tables show a clear drop in the accuracy rates when passing from simple level to hierarchical one. As a summarry, we conclude that the more than we have levels on adopted emotions grouping, the more we get a worse accuracy rate.

TABLE III
CONFUSION MATRIX FOR SIMPLE CLASSIFICATION

Emotion Class	Fear	Neutral	Other Emotions
Fear	53.3	20	26.7
Neutral	15.4	53.8	30.8
Other Emotions	6.4	0	93.6

TABLE IV
CONFUSION MATRIX FOR HIERARCHICAL CLASSIFICATION

Emotion Class	Positive emotions	Neutral	Fear	Other Negative Emotions
Positive Emotions	30	0	70	
Neutral	0	53.8	46.2	
Fear	5.8	7.7	53.3	46.7
Other Negative Emotions			5.6	94.4

VI. CONCLUSION

In this work, we presented a study that evaluated the effect of pitch-based features on fear recognition. It was shown that acoustic features related to pitch has a relatively great discrimination power between emotion states. Thus, the highest accuracy rate obtained in this study, reached 78.7% using k-nearest neighbors algrithm. We can conclude, as a result, that acoustic features related to vocal cords are relevant in emotion recognition, more precisely in fear detection. We have also proved that adopting hierarchism in emotion grouping considered classes affect negatively the classification performance. Studying fear recognition basing on the decomposition of sequences into segments will be the topic of further research.

REFERENCES

[1] N. Banda and P. Robinson, ”Multimodal affect recognition in intelligent tutoring systems,” Affective Computing and Intelligent Interaction, 200-207.2011.

[2] V. Petrushin, ”Emotion in speech: Recognition and application to call centers,”, Proceedings of Artificial Neural Networks in Engineering, 1999.

[3] E.Kramer, ”Judgment of personal characteristics and emotions from non verbal properties of speech,” Psychological Bulletin, 60(4), 408, 1963.

[4] R. P. Hobson, J. Ouston and A. LEE, ”Emotion recognition in autism: Coordinating faces and voices,” Psychological medicine, 1988.

[5] W. J. Yoon, Y. H. Cho and K. S. Park, ”A study of speech emotion recognition and its application to mobile services,” Ubiquitous Intelligence and Computing, 758-766, 2007.

[6] Cowie, Roddy and E.Douglas-Cowie, ”Automatic statistical analysis of the signal and prosodic signs of emotion in speech,” Proceedings,Fourth International Conference on Spoken Language,1996.ICSLP96.vol.3,1996,pp. 19891992.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz and J. G. Taylor, ”Emotion recognition in humancomputer interaction,” IEEE Signal Process.Mag. 18(2001)3280.

[8] R. Banse and K. Scherer, ”Acoustic profiles in vocal emotion expression,” J.Pers.Soc. Psychol. 70(3) (1996) 614636.

[9] I. Murray, J. Arnott, ”Towardas imulation of emotions in synthetic speech: A review of the literature on human vocal emotion,” J.Acoust.Soc.Am.93(2) (1993) 10971108.

[10] A. Oster and A. Risberg, ”The identification of the mood of a speaker by hearing impaire listeners,” Speech Transmission Lab.Quarterly Progress Status Report 4,Stockholm,1986,pp.7990.

[11] S. Beeke, R. Wilkinson and J.Maxim, ”Prosody as a compensatory strategy in the conversations of people with agrammatism,” Clin.Linguist.Phonetics23(2) (2009) 133155.

[12] M. Borchert and A. Dusterhoft, ”Emotions in speechexperiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments,” Proceedings of 2005 IEEE International Conference on Natural Language Process in gand Knowledge Engineering,IEEE NLP-KE05 2005,2005,pp.147151.

[13] J. Tao, Y. Kang and A. Li, ”Prosody conversion from neutral speech to emotional speech,” IEEETrans, Audio Speech Language Process.14(4)(2006)11451154.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz and J. Taylor, ”Emotion recognition in humancomputer interaction,” IEEE Signal Process. Mag. 18 (2001).

[15] C. Busso, S. Lee and S. Narayanan, ”Analysis of emotionally salient aspects of fundamental frequency for emotion detection,” IEEE Trans. Audio Speech Language Process. 17 (4) (2009).

[16] L. Bosch, ”Emotions, speech and the asr framework,” Speech Communication 40.1 (2003): 213-225.

[17] Scherer and R. Klaus, ”Vocal affect expression: a review and a model for future research,” Psychological bulletin, 1986, vol. 99, no 2, p. 143.

[18] J.R. Davitz, ”The Communication of Emotional Meaning,” McGraw-Hill, New York, 1964.

[19] T. Nwe, S. Foo, L. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (2003).

[20] S.Haque, , R.Togneri and A. Zaknich, ”A zero-crossing perceptual model for robust speech recognition,” In Inter-University Postgraduate Electrical Engineering Symposium, Curtin University, Perth, Western Australia, 27th September 2005.

[21] C. Williams and K. Stevens, ”Vocal correlates of emotional states,” Speech Evaluation in Psychiatry, Grune and Stratton, 1981.

[22] Talkin, David. A robust algorithm for pitch tracking (RAPT). Speech coding and synthesis, 1995.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier und B. Weiss, ”A Database of German Emotional Speech Proceedings,” Interspeech 2005, Lissabon, Portugal.

[24] R. Nuzzo, ”Scientific method: Statistical errors,” Nature, 506 (February), 150152, 2014.

[25] S. Lindsay, ”A Tutorial on Principal Component Analysis,” February 2002.

[26] DUDA and O.Richard, ”Patern classification and scence analysis,” Wiley, 1973.

[27] Quinlan, J. Ross, ”Induction of decision trees, Machine Learning,” 1986, vol. 1, issue 1, pp. 81 - 106.

[28] V. Vapnik, ”The Nature of Statistical Learning Theory,” Springer, 1995.

[29] Larose and T. Daniel, ”Knearest neighbor algorithm,” Discovering Knowledge in Data: An Introduction to Data Mining, 90-106.

[30] G. Baudat and F.Anouar, ”Generalized discriminant analysis using a kernel approach,” Neural computation, 12(10), 2385-2404, 2000