Mehran University Research Journal of Engineering & Technology
Vol. 37, No. 4, 453-462 October 2018
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1804.01

Application of ABM to Spectral Features for Emotion Recognition

SEMIYE DEMIRCAN*, AND, HUMAR KAHRAMANLI*

RECEIVED ON 11.04.2018 ACCEPTED ON 17.08.2018

ABSTRACT

ER (Emotion Recognition) from speech signals has been among the attractive subjects lately. As known feature extraction and feature selection are most important process steps in ER from speech signals. The aim of present study is to select the most relevant spectral feature subset. The proposed method is based on feature selection with optimization algorithm among the features obtained from speech signals. Firstly, MFCC (Mel-Frequency Cepstrum Coefficients) were extracted from the EmoDB. Several statistical values as maximum, minimum, mean, standard deviation, skewness, kurtosis and median were obtained from MFCC. The next process of study was feature selection which was performed in two stages: In the first stage ABM (Agent-Based Modelling) that is hardly applied to this area was applied to actual features. In the second stageOpt-aiNET optimization algorithm was applied in order to choose the agent group giving the best classification success. The last process of the study is classification. ANN (Artificial Neural Network) and 10 cross-validations were used for classification and evaluation. A narrow comprehension with three emotions was performed in the application. As a result, it was seen that the classification accuracy was rising after applying proposed method. The method was shown promising performance with spectral features.

Key Words: Agent-Based Modelling, Emotion recognition, Feature Extraction, Artificial Neural Networks, Optimization.

1. INTRODUCTION

From the perspective of human and computer interaction, it can be noticed that ER has gained importance recently. ER can be made via facial expressions, written text, and biomedical signals or speech signals [1-2]. Especially in the situation when face-to-face communication is not made, determining the emotion state of a person is made from speech data.

Nowadays, ER from speech data is successfully applied in many fields. For instance, Parket. al. [3] determined that a customer's speech was negative or non-negative by carrying out ER from speech on the application they developed in service robots [3]. An in-car board system, applications for safe drive and aircraft applications are among other application fields [4]. By using acoustic

Authors E-Mail: (semiye@selcuk.edu.tr, hkahramanli@selcuk.edu.tr)
* Department of Computer Engineering, Faculty of Engineering, Selcuk University/Konya Technical University, Konya, Turkey.

This is an open access article published by Mehran University Research Journal of Engineering and Technology, Jamshoro under the CC by 4.0 International License.

features of a speech, an application for determining differences among patients who are depressive and have suicidality also exists [5].

Feature extraction and selection are one of the most important steps in ER as in other signals. The feature selection methods can be examined in two ways as spectral and prosodic features. MFCC, and LPC (Linear Predictive Coefficients) can be considered as spectral features. Basic frequency (F0) value is one of the main prosodic features.

One of the vector features that have been used recently is LPC. Linear predictive analysis has importance in characterizing spectral features of a speech sign in time environment [6-7]. MFCC can be calculated without the necessity of LPC because while linear predictive analysis is modelling speech path, MFCC features models human ear [7]. As a result of this, MFCC method produces quite successful results in ER applications [7-8].

Milton et. al. [6] have taken Pitch, duration, energy and MFCC, LPC, features of AR (Autoregressive) parameters, which include gain and reflection coefficients to recognize the emotion from the speech. Single classifier or a combination of classifiers was applied to recognize emotions from the input features. Seven emotions (Anger, Boredom, Disgust, Fear, Happiness, Sadness and Neutral) were taken to recognize.

Lee et. al. [9] have introduced a hierarchical computational structure to recognize emotions. Structure maps were proposed as an input speech utterance into one of the multiple emotion classes through subsequent layers of binary classifications. The classification framework was evaluated on two different emotional databases using acoustic features, the AIBO database and the USC IEMOCAP database.

Multi-agent systems were widely utilized for machine learning systems. Montano et. al. [10] used multi-agent system for learning to identify an appropriate agent to answer free-text queries and keyword searches for defense contracting. Navarro et. al. [11] simulated an expert multi agent system that can compose harmony following specific rules. Taha et. al. [12] developed a novel agent-based design for Arabic speech recognition. The Arabic speech recognition was defined as a Multi-Agent-System where each agent has a specific goal and deals with that goal only.

Since this area is young, there are many ways to perform and improve it. It is seen that MFCC and LPC are used predominantly in the results of the investigations, but it is observed that MFCC is more successful. One of the most important stages in emotion recognition is feature selection. Whereas there are many methods for this step, the desired success rate has not been achieved. The aim of this study is to develop a new feature selection method to increase the success rate.

For application stage data obtained from Berlin Emotion Database [13] for ER were used. Firstly, MFCC were obtained from the data. Since obtained features had too large dimensions to give them to classifiers and they were in different dimensions, new features were calculated by extracting some statistical features from these data. So as to determine the features which supplies to obtain the high accuracy level, opt-aiNET optimization algorithmon ABM were applied. ABM with opt-aiNET was applied in this study for the first time in literature. Obtained features were given to the classifier. As classifier, ANN was used. Classification results with ABM and without ABM were compared. It was observed that classification accuracy increases using ABM.

The remainder of this paper is structured as follows. In Section 2, information related to used methods was given. In Section 3, database was shortly explained. The experiments to assess the performance of proposed method were described in Section 4. In the last part, conclusion was given.

2. METHOD

ER from speech signals has been among the attractive subjects lately. As known the most important process steps are feature extraction and feature selection in ER.The process of ER can be made from facial expressions and speech signals. Schematic representation of ER process realized in this study is obtained via the steps depicted in Fig. 1.

Although directly ER from signal information seems theoretically possible, the dimension reducing process is needed by using features extracted from data through different methods because data dimension is very large. When it is considered from this point of view, it can be clearly said that the most important steps are feature extraction and selection in emotion detection.

Many methods exist for feature extraction in speech processing. The most known and used ones are MFCC, LPC, F0 value, Wavelet transform, AR Parameters.

2.1 Mel-Frequency Cepstrum Coefficients

MFCC comes first among spectral features that are used most widely to obtain feature from speech signals.

Input signal is divided into parts in a way that M is sample number and N is sample length (M<N). Whereas the first frame consists of N sample, the subsequent frame starts after M sample from the first frame and thus, the samples as N-M match up with [14].

Then, the process of windowing is carried out using the function of Hamming windowing. The function of Hamming windowing is indicated in Equation (1).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 1 \le n \le N \tag{1}$$

The windowed sign is passed through a FIR filter of first degree in pre-emphasis process. FFT (Fast Fourier Transform) is applied in order to transform speech part consisting of N samples from time domain into frequency domain. Mel scale is a scale explained with changing intervals in a way that it changes as linear up to 1 kHz and after 1 kHz, it changes as logarithmic. The logarithm of the sign obtained at the output of Mel filter is taken. Data at the logarithmic Mel scale in DCT (Discrete Cosine

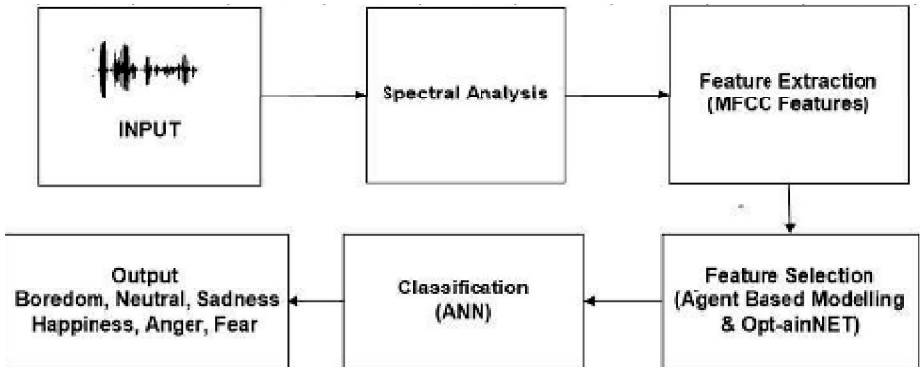

FIG. 1. SCHEMATIC REPRESENTATION OF ER PROCESS

Transform) that is the last step of MFCC extraction process is transformed into time environment again. As a result, data obtained are called as Mel Coefficients [14].

2.2 Opt-aiNET Algorithm

aiNET algorithm is a discrete immune network algorithm that is developed for clustering. Opt-aiNET algorithm is a little more developed state of it and its adaptation for optimization of problems [15]. In the study feature selection was made by applying opt-aiNET algorithm.

2.3 Agent-Based Modelling

Although agent does not have an exact definition, it can be defined as an object having features of target, action and state in a particular environment [16]. Moreover, it can be stated as a computer system that can act automatically in order to fulfil a particular aim.

An autonomous agent is defined as a system that receives (virtual or real) perception of the environment where it exists, creates situational awareness after that and by using this perception information in accordance with the aim, it determines its subsequent behavior and that realizes determined action in environment [17]. In a similar way, extracted feature groups are considered as agents and optimization process is considered as an environment. Action value of agent is determined according to classification success. If the classification success is maximum, state is being sent as "1" to the agent and the agent is used for ER, with the action value "1". Otherwise, it is decided that state will not be classified because state is "0" and at the same time, action value of the agent is "0".

3. MATERIAL

Many databases are applied in ER problems. It is possible to classify them in different ways according to their number of emotion included, with which language they are formed, as being public or private, their voicing by professionals or actors [18]. A list of the most important database according to their speaking language [19] is also given. The access of databases created by being vocalized by professionals is expensive because they are generally private.

EmoDB known as "Berlin Database of Emotional Speech" database [13] is a public and free database. Therefore, "Berlin DB" database was chosen in our study. This database was created in anechoic chamber in Technical University, Berlin. It was vocalized by 5 males and 5 females, 10 different actors in total. 10 different sentences were vocalized with 7 emotions. These emotions are: A (Anger), B (Boredom), D (Disgust), F (Fear), H (Happiness), S (Sadness) and N (Neutral). Vocalized sentences are sentences we frequently use in real life [13]. They consist of 535 audio segments that were sampled at 16 kHz. In this study, 520 segments from 535 segments were taken.

4. PROPOSED METHOD AND RESULTS

In this section, how the feature selection is done using the agent-based modeling and Opt-aiNET optimization algorithm is explained in detail.

The application was performed with narrow comprehension (3 emotions). The first group of emotions is BNS (Boredom, Neutral and Sadness) and the second group of emotions are HAF (Happiness, Anger and Fear).

Both performances firstly, statistical values were calculated from MCFF obtained by using MFCC methods on emotion data. In second stage of study, to select the features which increase the classification accuracy ABM was used.

As optimization algorithm for feature selection Opt-aiNET algorithm was used. Finally, selected features were used as inputs for ANN and emotion classification was done.

In this study, ANN classification ofWeka [20] (an open-source public available toolbox for automatic classification) was used. MLP (Multilayer Perceptron Algorithm) was preferred. 10-fold cross validation was applied in order to indicate reliability of the study.

In advance, 16 Mel-Coefficients of different lengths from each data were obtained. In the first step, data reduction was made by taking statistics of each coefficient. Extracted statistical values are indicated in Table 1.

As seen in Fig. 2, the segment dimension was 30896 at the first stage and then Mel coefficients were extracted, so the dimension was obtained as (388x16). In other words, 16 Mel coefficients were obtained. Finally, new fixed-dimensional feature vector of (7x16) by calculating statistical features for each Mel coefficient is given in Table 1. As a result, a new data set consisting of features of 112 (7x16) for each segment by transforming this matrix into row matrix was obtained.

TABLE 1. THE STATISTICAL FUNCTIONS USED FOR MFCC COEFFICIENTS

MFCC Features	Maximum Value of MFCC
	Minimum Value of MFCC
	Mean Value of MFCC
	Standard Deviation of MFCC
	Skewness Value of MFCC
	Kurtosis Value of MFCC
	Median Value of MFCC

In literature, it is seen that AHF and BNS are the most confused emotions. So we divided the emotions into two groups according to frequencies which are close to each other [8].

The first group of emotion consists of AHF. The steps of process for AHF are shown in Fig. 3. Fig. 3 feature group consisting of statistical values extracted from Mel coefficients was defined as Dataset1. After extracted feature is modelled with ABM, feature selection is performing with Opt-aiNET. The optioned data is named Dataset2. ANN is used for classification.

The processes of ABM and optimization were applied in order to observe the effect of feature selection upon

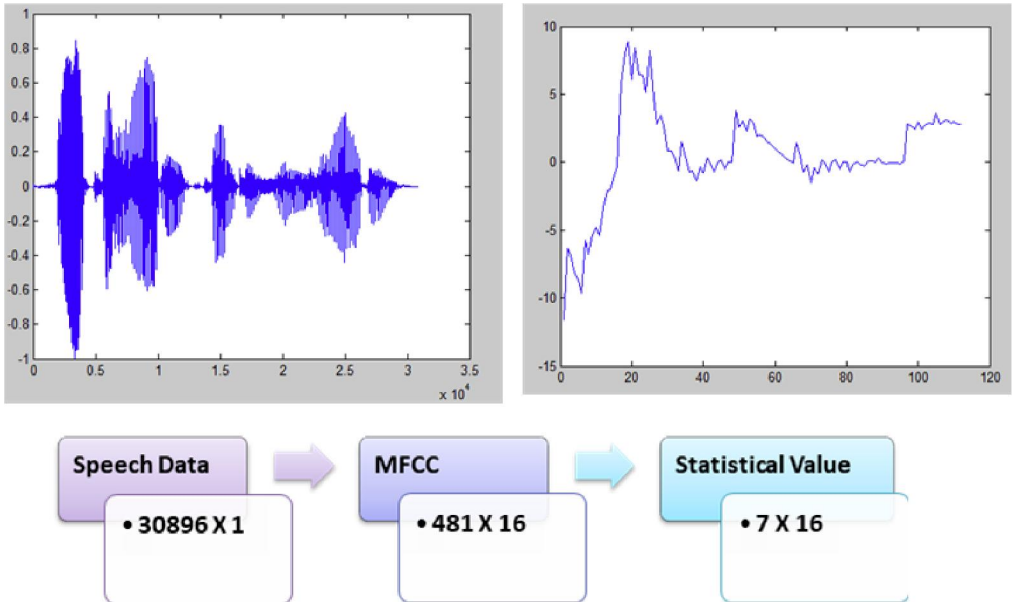

FIG. 2. AN EXAMPLE OF EXTRACTED FEATURE WITH MFCC USING A RANDOMLY SELECTED SEGMENT

classification success. Modeling process is indicated in Fig. 4. Fig. 4 each group of statistical values obtained from Mel coefficients was considered as an agent. Maximum value belonging to each group of coefficient was defined as Agent1, minimum value as Agent2, average value as Agent3, Standard Deviation value as Agent4, Median value as Agent5, Skewness value as Agent6 and Kurtosis value as Agent7. All the agents were given to Opt-aiNET algorithm. According toOpt-aiNET algorithm result action value was determined and it was used as a state value in the classification.

In the optimization step, the population is created first. Because the individual in the population represents the group of agents, it is foreseen that the length is as much as the number of agents. For this reason, the length of the individual in this study has been seven. Individuals consist of 0s and 1s. If the value of the agent is 1, the feature group represented by this agent will be included in the classification, 0 will not be included.

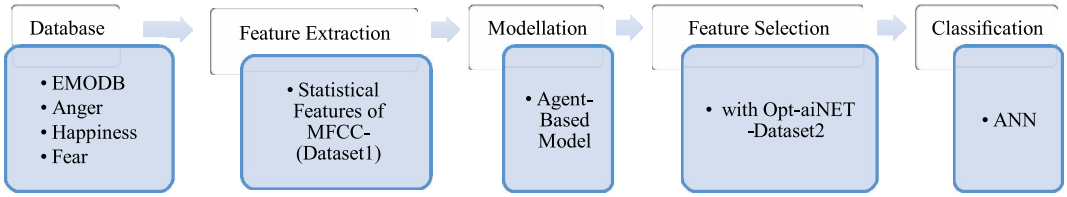

FIG. 3. THE STEPS OF PROCESS FOR DATASET1 AND DATASET2

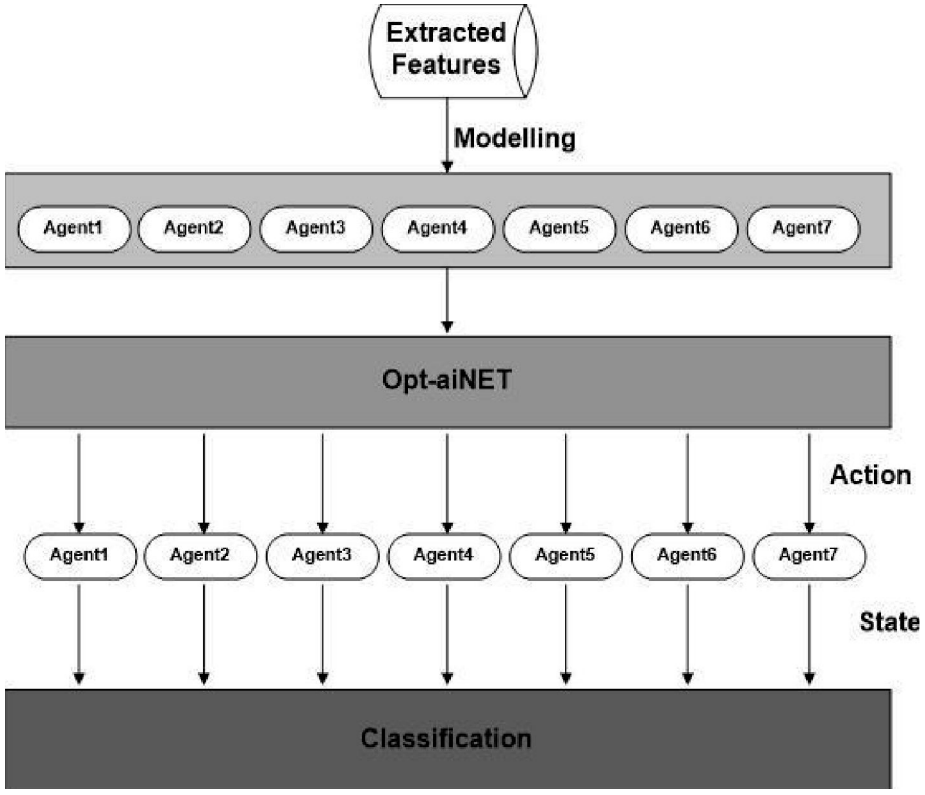

FIG. 4. AGENT AND OPTIMIZATION STRUCTURE

For example, let the individual is (0,1,0,1,0,0,1). This vector means that the 2nd, 4th, and 7th factors, namely the minimum value of MFCC, Standard deviation of MFCC, and the Median value of MFCC properties are used.

Classification success is used as fitness function. Since the goal is to choose agent (or agents) that will make the classification accuracy maximum, Opt-aiNET optimization algorithm was employed. The individual who has achieved the highest classification success as the result of the optimization has been selected for use in the future stages.

Agent groups chosen as a result of trials are modelled as feature groups with Agent3, Agent5, Agent6 and Agent7. All of these features were named as Dataset2. Selected features are shown in Table 2.

TABLE 2. SELECTED AGENTS FOR DATASET1

Agent	Dataset1	Dataset2
1	□	
2	□	
3	□	□
4	□	
5	□	□
6	□	□
7	□	□

Obtained results are shown in Table 3. Classification accuracy of 75.77% was obtained when Dataset1 was classified with ANN. As for the classification accuracy of Dataset2, it was obtained as 78.08%. Confusion matrix obtained as a result of classification is indicated in Table 4.

When Table 4 is examined it can be seen that 104 of Anger emotion in Dataset1 were accurately classified, while it was increased 106 of segments were for Dataset2. It means that the classification success of Anger emotion increases from 82.54-84.13%.It was observed that the number of defined segments in the emotion of Happiness was increased from 43-46. Before feature selection was not made with optimization in the emotion of Fear, 50 of them were classified as accurately. By applying the process in Fig. 5, 51 segments were accurately classified.

The second emotion groups BNS consist of Boredom, Neutral and Sadness. The steps of process carried out are indicated in Fig. 5. Fig. 5 after the statistical features extracted from Mel coefficients data was named as Dataset3. The data cluster obtained from optimization process that was applied after the modelling with ABM was named as Dataset4. Dataset3 and Dataset4 were classified with ANN.

TABLE 3. THE RECOGNITION RATE AND CLASSIFICATION ACCURACY FOR DATASET1 AND DATASET2

Dataset	Recognition Rate			Classification Accuracy (%)
	Anger	Happiness	Fear	
1	82.54	61.43	78.12	75.77
2	84.43	65.71	79.69	78.08

TABLE 4. CONFUSION MATRIX

Dataset1				Dataset2			
	A	H	F		A	H	F
A	104	21	1	A	106	19	1
H	17	43	10	H	16	46	8
F	5	9	50	F	3	10	51

Selected agents are shown in Table 5. The classification successes obtained for Dataset3 and Dataset4 are given in Table 6. The classification success was obtained as 74.66% with Dataset3 and 80.09% with Dataset4. The highest classification success was obtained by selecting feature groups of Agent3 and Agent5 after optimization process. Confusion matrix obtained as a result of classification is indicated in Table 7.

When Table 7 was examined in terms of Boredom emotion it can be seen that 52 of segments were classified correctly in Dataset3, when it increases to 56 for Dataset4 and it means that the classification success of Boredom emotion was increased from 64.20-69.14%. The number of segments defined in Neutral emotion data was observed to increase from 56-66. In the emotion of Sadness, 57 of them were classified as accurate before selection with optimization. After implementation of process in Fig. 5, 55 segments were classified accurately.

The method suggested in the study was applied to the EMO dB using a 10-cross fold validation in a speaker independent manner, covering all emotions. The studies with these criteria are not often encountered in the literature. For these reasons, the obtained results were compared to the results obtained using MFCC features in the studies [8,21]. Albornoz et. al. [8] found the classification accuracy 71.48% for AHF and BNS emotion

TABLE 5. SELECTED AGENTS FOR DATASET3 AND DATASET4

Agent	Dataset5	Dataset6
1	☐	
2	☐	
3	☐	☐
4	☐	
5	☐	☐
6	☐	
7	☐	

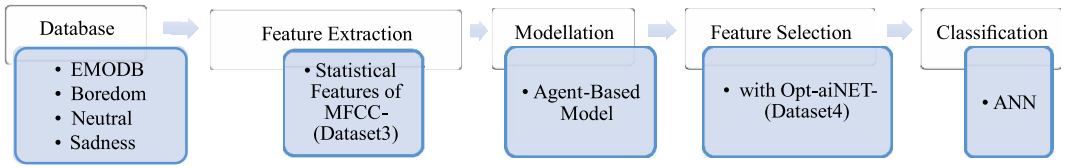

FIG. 5. THE STEPS OF PROCESS FOR DATASET3 AND DATASET4

TABLE 6. THE RECOGNITION RATE AND CLASSIFICATION ACCURACY FOR DATASET3 AND DATASET4

Dataset	Recognition Rate			Classification Accuracy (%)
	Boredom (%)	Neutral (%)	Sadness (%)	
5	64.20	70.89	93.44	74.66
6	69.14	83.54	90.16	80.00

TABLE 7. CONFUSION MATRIX

Dataset3				Dataset4			
	B	N	S		B	N	S
B	52	23	6	B	56	19	6
N	22	56	1	N	12	66	1
S	3	1	57	S	3	3	55

groups in their study. Fersini et. al. [21] also achieved classification accuracy to 64.78% using Berlin DB. It is seen that proposed method achieved better results for 3 group emotions than other studies.

4. CONCLUSION

Speech recognition, person recognition and ER come first among the studies carried out on speech signals. In this study, the process of ER from emotion data including speech signal was carried out. Mel coefficients were obtained from spectral features and so, statistical values were extracted. The obtained features were modelled as ABM and opt-aiNET optimization was carried out to select the agent group which provides better classification accuracy. The changes in classification success after optimization process using ABM were analyzed. It was observed that the success of 75.77% for the first emotion group (AHF) increased to 78.08%. It was analyzed the success of 74.67% for the second emotion group (BNS) increased to 80.00%. As a result, it was seen that the presented method gave successful results in feature selection. In our following study, we aim to apply this method on different features.

ACKNOWLEDGEMENTS

The authors acknowledge the support of this study provided by Selcuk University Scientific Research Projects. The authors also thank Tubitak, for their support of this study.

REFERENCES

[1] Perikos, I., and Hatzilygeroudis, I., "Recognizing Emotions in Text Using Ensemble of Classifiers", Engineering Applications of Artificial Intelligence, Volume 51, pp. 191-201, 2016.

[2] Kim, J.-B., and Park, J.-S, "Multistage Data Selection-Based Unsupervised Speaker Adaptation for Personalized Speech Emotion Recognition", Engineering Applications of Artificial Intelligence, Volume 52, pp. 126-134, 2016.

[3] Park, J.S., Kim, J.H., and Oh, Y. H., "Feature Vector Classification based Speech Emotion Recognition for Service Robots", IEEE Transactions on Consumer Electronics, Volume 55, pp. 1590-1596, August, 2009.

[4] Schuller, B., Rigoll, G., and Lang, M., "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, pp. 577-580, 2004.

[5] France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., and Wilkes, D.M., "Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk", IEEE Transactions on Biomedical Engineering, Volume 47, pp. 829-837, July, 2000.

[6] Milton, A., and Tamil, S.S., "Class-Specific Multiple Classifiers Scheme to Recognize Emotions from Speech Signals", Computer Speech & Language, Volume 28, pp. 727-742, 2014.

[7] New, T.L., Foo, W., De, S., and Silva, L.C., "Speech Emotion Recognition Using Hidden Markov Models", Speech Communication, Volume 41, pp. 603-623, 2003

[8] Albornoz, E.M., Milone, D.H., and Rufiner, H.L., "Spoken Emotion Recognition Using Hierarchical Classifiers", Computer Speech and Language, Volume 25, pp. 556-570, July, 2011.

[9] Lee, C.C., Mower, E., Busso, C., Lee, S., and Narayanan, S., "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach", Speech Communication, Volume 53, pp. 1162-1171, November-December, 2011.

[10] Montano, B.R., Yoon, V., Drummey, K., and Liebowitz, J., "Agent Learning in the Multi-Agent Contracting System", Decision Support Systems, Volume 45, pp. 140-149, 2008.

[11] Navarro, M., Corchado, J.M., and Demazeau, Y., "MUSIC-MAS: Modeling a Harmonic Composition System with Virtual Organizations to Assist Novice Composers", Expert Systems with Applications, Volume 57, pp. 345-355, September 15, 2016.

[12] Taha, M., Helmy, T., and Alez, R.A., "Multi-Agent Based Arabic Speech Recognition", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, Marriott Fremont, CA, USA, 2007.

[13] Burkhardt, F., Paeschke, A., Rolfes, M., and Sendlmeier, W., "A Database of German Emotional Speech", Proceedings of Interspeech, 2005.

[14] Becchetti, C., and Ricotti, L.P., "Speech Recognition: Theory and C++ Implementation", Speech Recognition, pp. 10, 2004.

[15] de Castro, L.N., and Timmis, J., "An Artificial Immune Network for Multimodal Function Optimization", Proceedings of Congress on Evolutionary Computation, Volume 1-2, pp. 699-704, 2002.

[16] Stone, P., and Veloso, M., "Multi-Agent Systems a Survey from a Machine Learning Perspective", Autonomous Robotics, Volume 8, 2000.

[17] Weiss, G., "Multi-Agents Systems", 2nd Edition, MIT Press, 2013.

[18] El Ayadi, M., Kamel, S., and Karray, F., "Survey on Speech Emotion Recognition: Features, Classification M. Schemes, and Databases", Pattern Recognition, Volume 44, pp. 572-587, March, 2011.

[19] Ververidis, D., and Kotropoulos, C., "A State of the Art Review on Emotional Speech Databases", 1st Richmedia Conference, 2003.

[20] Hall, M., Frank, E., Holmes, G., Pfahringer, B., and Reutemann, PIH, "The WEKA Data Mining Software", ACM SIGKDD Explor, 2009.

[21] Fersini, E., Messina, E., and Archetti, F., "Emotional States in Judicial Courtrooms: An Experimental Investigation", Speech Communication, Volume 54, pp. 11-22, January, 2012.