Speech Emotion Classification and Recognition with different methods for Turkish Language

C.Bakir, M.Yuzkat

Abstract— **In several application, emotion recognition from the speech signal has been research topic since many years. To determine the emotions from the speech signal, many systems have been developed. To solve the speaker emotion recognition problem, hybrid model is proposed to classify five speech emotions, including anger, sadness, fear, happiness and neutral. The aim this study of was to actualize automatic voice and speech emotion recognition system using hybrid model taking Turkish sound forms and properties into consideration. Approximately 3000 Turkish voice samples of words and clauses with differing lengths have been collected from 25 males and 25 females. In this study, an authentic and unique Turkish database has been used. Features of these voice samples have been obtained using Mel Frequency Cepstral Coefficients (MFCC) and Mel Frequency Discrete Wavelet Coefficients (MFDWC). Moreover, spectral features of these voice samples have been obtained using Support Vector Machine (SVM). Feature vectors of the voice samples obtained have been trained with such methods as Gauss Mixture Model(GMM), Artifical Neural Network (ANN), Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and hybrid model(GMM with combined SVM). This hybrid model has been carried out by combining with SVM and GMM. In first stage of this model, with SVM has been performed subsets obtained vector of spectral features. In the second phase, a set of training and tests have been formed from these spectral features. In the test phase, owner of a given voice sample has been identified taking the trained voice samples into consideration. Results and performances of the algorithms employed in the study for classification have been also demonstrated in a comparative manner.**

Index Terms—**MFCC, MFDWC, emotion, HMM, hybrid model.**

I. INTRODUCTION

Today, with the development of technology, security problems have also come to light. Biometric systems, such as authentication in particular, constitute an important part of the security issue. For this reason, it is necessary to determine the forensic soundings of the voice recordings subject to various crimes and the emotions of the people in these voice recordings at that moment. Especially in commercial transactions, some studies have been carried out to prevent the transfer of information belonging to persons to other persons. Handwriting recognition, signature recognition, face recognition, iris recognition, voice recognition constitute several of these studies [8].

Despite the fact that speech recognition technology has a very long history, attempts to extract emotion from human voice are still new and attract great attention. Obtaining the necessary data for extracting emotion constitutes an important problem. Because, there are so many kinds of emotions and it is very difficult to determine these emotions.

Various studies have been performed to determine the emotion of voice and speaker. Shami et al. have performed the study of emotion recognition from speech data with k-nearest neighbors, (kNN), Support Vector Machines (SVM) ve Ada-Boosted decision tree machine learning techniques on four different databases. In this study, the success of feature extraction techniques, AIBO and segmentation based approach, SBA on different databases and different classification techniques are compared. Both feature extractions gave different results on different databases [1].

Chen et al. have developed a three-level model to distinguish six emotions as independent of the speaker. Various features were selected from 288 individuals by using the Fisher ratio for each level of emotion. In order to measure the success of the proposed system, Principal Component Analysis (PCA) dimension reduction method was used and for classification, ANN and SVM were used. The results obtained have been presented comparatively. However, since the frequency of speech changes abruptly in some emotions, more study is required to be performed in this respect [2].

He et al. proposed two different methods of feature extraction for emotion classification from speech data. In the first method, they applied the EMD (Empirical Model Decomposition) method which calculates the average entropy of speech data. In the second method, however, they studied with a method that calculates the average spectral energy in the lower bands of the speech spectrogram. He et al. calculated the success of these two methods by using GMM and kNN classification algorithms on two different databases. They also compared the success of these two methods by using the MFCC feature extraction method [3].

Polzehl et al. conducted emotion recognition studies by using acoustic characteristics of speech data of children. They

C. BAKIR, is with Department of Computer Engineering University of Igdir University, Istanbul, Turkey, (e-mail: cigdem.bakir@igdir.edu.tr).

M. YUZKAT, is with Department of Computer Engineering University of Mus Alparslan Technical University, Istanbul, Turkey, (e-mail: m.yuzkat@alparslan.edu.tr).

Manuscript received August 22, 2017; accepted Nov 16, 2017.
DOI: 10.17694/bajece.419557

tried to distinguish angry feelings and feelings which are not angry. In this study, frame-based cepstral properties reduced in size were classified by acoustic properties ANN and SVM. Furthermore, in the study, feature selection was made with the Data Acquisition Ratio [4].

Nwe et al. have worked to distinguish the six emotions on the speech data. They classified feelings with HMM. In this study, a database containing 90 different emotions taken from two speakers was used. However, this work was also carried out depending on the speaker [5].

Bhaskar et al. have proposed a hybrid approach to classify feelings in speech and text. In the study they made, both the textual and speech features were combined. For classification, they used the multi-class SVM method. However, only 11 features have been used in the study. More features need to be selected to achieve the desired performance [6].

Lee et al. Carried out and emotion study by using the Recurrent Neural Network (RNN) algorithm. In this study, they applied the Bidirectional long short-termmemory (BLST) algorithm to determine the time-varying emotions. In this way, the changes that occurred in the emotions, that is, unspecified emotions whose tag changed were tried to be determined [7].

Feature extraction, classification techniques used of the study performed, experimental study of results and conclusion were given respectively in the section 2, section 3, section 4 and section 5.

II. FEATURE EXTRACTION METHODS

The study has been realised on a unique database, which was formed from Turkish sound samples taken from both men and women. These sound samples are trained by getting dispersed to various feature vectors with MFDWC, MFCC and LPCC. In the second stage, the feature vectors of the recorded sound signals are trained with classification algorithms such artificial neural network (ANN), Dynamic Time Warping (DTW), Hidden Markov Model (HMM) and Gauss Mixture Model (GMM). The speech for recognition is decided by looking at sound signals in the test and training data after the system is trained. Furthermore, the classification success in recognising the gender of speaker was calculated separately for 5 feature vectors and the success of the methods was presented comparatively by training the feature vectors, obtained from speech signals.

A. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a feature extraction method, that is used in sound processing. It is used to extract important information and features by dividing the sound data to its subsets. The steps of feature extraction technique of MFCC is indicated in Figure 1[18].

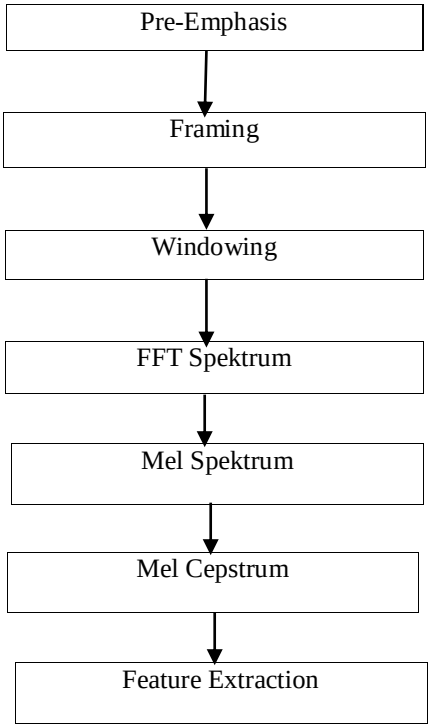

Fig.1. Feature extraction steps of MFCC

Two filters are used in MFCC feature extraction method. The first filter has a linear distribution of frequency values under 1000 Hz and the other has a logarithmic distribution of frequency over 1000 Hz. Pre-emphasis stage is the first stage in obtaining MFFC feature vector.

The sound signals, which have high frequency, are passed through a filter at this stage. This way, the energy of the sound is increased at high frequency. The sound signals are analog. The sound signals are converted from analog to digital by getting divided into small frames between 20 and 40 ms during the framing stage and it is divided into N frames. The sound signal is moved by sliding the sound signal at the windowing stage. This way, the closest frequency lines and the frame, which will come by windowing, that is used are combined. The window type, width and sliding amount are determined at this stage. Each of N frames is transmitted from the time space to the frequency space with Fast Fourier Transformer (FFT). The spectral features of sound signals are shown in frequency space.MEL spectrum is obtained by calculating the total weight of these spectral features. This MEL spectrum is formed from triangle waves and are formed by getting passed through a series of filters. MEL spectrum reduces the noise by lowering two neighbour frequencies. The logarithm of signal is taken at the stage of MEL spectrum and the signal is transmitted back again from frequency space to the time space. MEL frequency cepstrum factors are obtained by using DCT (Discrete Cosine Transform) in time space.

B. *Mel-frequency Discrete Wavelet Cofficients (MFDWC)*

The study in question has been performed, based on a unique database comprising Turkish voice samples collected from men and women. These voice samples were separated into various feature vectors with MFDWC, and trained. MFDWC is a feature extraction method employed in the speech processing. It is used to extract significant information and features by dividing voice data into subsets. Feature extraction steps of MFDWC technique is shown in the Figure 2 [19].

Sample speech signal is shown between the 40-40000 Hz range in the MFDWC feature extraction method. Speech signal is divided into frames after the pre-processing step. Hamming window has been used in this study in order to smoothen the transition of speech samples between the frames. One Mel shows the frequency of voice tone. Mel-scale is scaled between actual frequency of voice signal and estimated voice frequency. For this reason total energy of every frame is calculated. Classification success in speaker identification has been calculated on an individual basis for MFDWC-5 vectors by training the feature vectors obtained from voice signals by means of different methods.

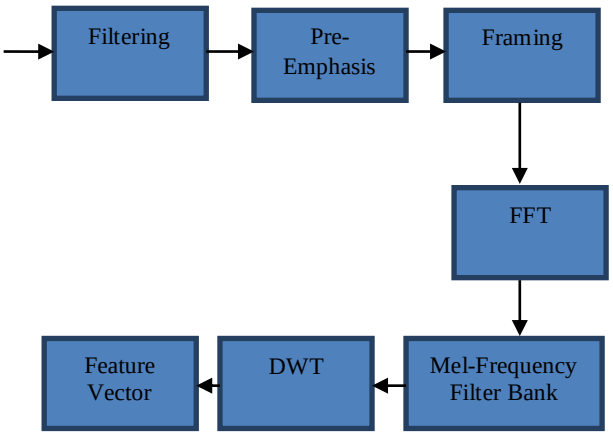

Fig.2. Feature extraction steps of MFDWC

III. METHOD

In Figure 3, the steps of the study are given. In this study, the sound samples, taken from 25 males and 25 females in different age ranges, have been separated to their feature vectors with SVM. The education and test samples have been formed from these voice feature vectors. These train and test samples have been coached according to ANN, DTW, HMM and recommended hybrid model and speech emotion recognition transaction has been realized automatically. Furthermore, the results, obtained with ANN, DTW, HMM and hybrid method, have been given comparatively.

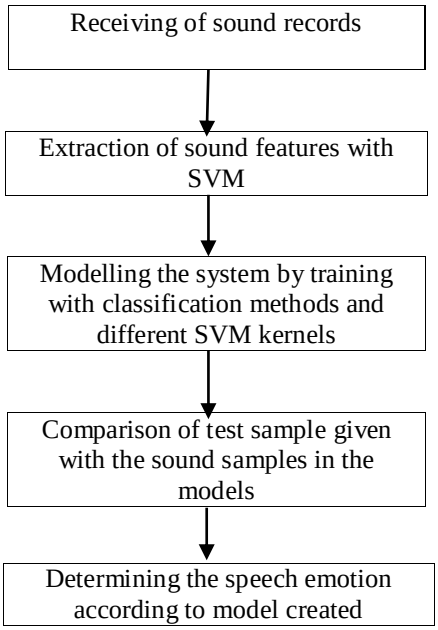

Fig.3. Study steps

A. *Artificial Neural Network*

ANNs have a very wide fields of application up to automotive, banking, defense industry, electronics, entertainment, finance, insurance, manufacture, oil and gas, robotics, telecommunication and transportation industry.

Artificial neural networks are information systems which mirror human brain function, and classify the data through learning. They have been developed, being based on a principle of human brain functioning. In other words; ANNs have been developed with a logic similar to the biological neural networks, and are data processing structures connected to each other with weights.

ANNs comprise of input layer, output layer and hidden layers. Data is received into neural networks through input layer. And it is transferred to outside through output layer. Layers between input and output layers constitute hidden layers.

Neurons in the feed-forward neural networks are connected just in the forward direction [11]. Each layer of neural network contains the connection of next layer and these connections are not in the backward direction. In a sense, there is a hierarchical structure between neurons, and the neurons located in one layer can only communicate data to the next layer. Structure of a feed-forward ANN is shown in the Figure 4.

Backward propagation network shows how to train a neuron [12]. Trainer is a sort of learning. Network is maintained both with the sample inputs and expected outputs when the trainer method is employed. Expected outputs are compared with actual outputs for the networks the inputs of which are given. Error is calculated in case the expected outputs are used, and weights of

various layers are adjusted in the backward direction from output layer to input layer. In other words, it is given for both input data and output data. Network updates its coefficients in order to obtain the expected output.

ANN is the most widely used method. In this algorithm, error in the output layer is calculated at the end of each iteration, so this error is transmitted to all neurons in the direction from output layer to input layer, and weights are readjusted according to the error margin. Such error margin is distributed to the previous neurons located before the said neuron in proportion to their weights.

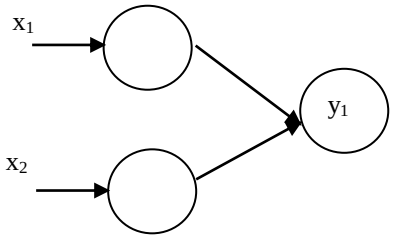

Fig.4. Feed-forward neural network

Layers are located one after another in a multilayer artificial neural network. Outputs of neurons in a layer will be given as their weights, to the input of next layers, and these weight are used in the calculation of outputs for the next layer. Weights of the hidden layer between input and output layers are calculated [12].

B. *Dynamic Time Warping*

Dynamic Time Warping (DTW) finds out to which speaker the voice signal given belongs, by calculating the similarity between the time-variant two speech signals. The most optimal time curve can be identified between two signals with this method.

$$Q= q_1, q_2, \dots, q_i, \dots, q_n \quad (1)$$

$$C= c_1, c_2, \dots, c_j, \dots, c_m \quad (2)$$

Q and C in the equation 1 and equation 2 demonstrate two distinct speech signals; n and m show the lengths of these speech signals [6]. In this case, the ratio of similarity between Q and C signals is calculated using Euclid length as in the equation 3 [14].

$$d(q_i, c_j)= (q_i, c_j)^2 \quad (3)$$

A matrix (i,j) is generated for Q and C. Accumulated distance matrix is calculated using this matrix.

$$D(i,j)=\min[D(i-1,j-1),D(i-1,j),D(i,j-1)]+d(i,j) \quad (4)$$

C. *Hidden Markov Model*

A lot of studies have been carried out with regard to the Hidden Markov Models (HMM) in many fields from past to today. HMM has been used in a wide manner in face recognition, speech recognition, voice recognition, hand script recognition, human body motion recognition, bioinformatics, estimation of gene, cryptanalysis, protein structure and sequence, DNA sequence and pattern recognition.

In Hidden Markov Model (HMM) the aim is to try to estimate future situations that will likely occur in cases when the existing situations are given as an input to the system. HMM is a stochastic process since it generates different output whenever it is operated. In addition, system in Markov models may move from its own state to another state according to the probability distribution, or remain in the same state. Probabilities occurred in the states are called as transition probabilities. States are not seen by the observer as distinct from HMM normal Markov model. However transition subject to the states may be observed. HMM speaker recognitions systems comprise of the following steps [13].

S= {$S_1, S_2, \dots, S_Q$ } shows current status of the speech signals generated where there are Q numbers of states.

- Initial state probabilities is determined in a discrete time, t. (π= { P_r, (S_i, |t=0, S_i, ε S})
- Transition probabilities are calculated according to the current states. a_{ij} = (P_r (S_j t in time t |S_j in time t-1), S_i ε S, S_j ε S))
- F, which is the number of features observed, is determined.
- Probability distribution of speech signal will be calculated in this way. (b_x= {b_j (x)= P_r (x(S_i), S_iε S, x ε F})
- HMM generated is demonstrated by λ =(a,b,π).

D. *Gauss Mixture Model*

Gauss Mixture Model is a statistical method based on the weight combination of the Gaussian distribution of one or more audio signals. The sum of the weighted combinations of Gaussian intensity is shown in equation 5 [10].

$$p(x|\lambda) = \sum_{i=1}^{M} p_i \, b_i(x) \quad (5)$$

x shows the feature vector, **D** shows the dimensional random vector b_i(x),i = 1,…,M shows the density components, and p_i shows the mixture weight. The parameters of this model are found by the ExpectationMaximization (EM) algorithm. All classes in the training data are expressed by

independent Gaussian density function. The most optimal density components to determine the mixture are found. Equation 6 is used to find the Gaussian model parameters that will maximize p(xIλ) [10].

$$p(xI\lambda) = \prod_{t=1}^{T} p(x_t I\lambda) \quad (6)$$

The GMM density function is shown in equation 7 [15].

$$p(x) = \sum_{i=1}^{N} w_i \; N\left(x; \mu_i \sum i\right) \quad (7)$$

N shows the Gauss density function, w_i , μ_i and $\sum i$ show weight, mean and covariance matrix of the Gaussian component **i**, respectively. The GMM super-vector consists of the sum of the averages of each Gaussian component [15].

N shows the Gauss density function, w_i , μ_i and $\sum i$ show weight, mean and covariance matrix of the Gaussian component **i**, respectively. The GMM super-vector consists of the sum of the averages of each Gaussian component [15].

$$\mu = \begin{matrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{matrix} \quad (8)$$

Each emotion is trained by the spectral properties generated by the GMM super-vectors shown in equation 8.

E. Hybrid Model (Gauss Mixture Model with combined SVM)

SVM is a classification algorithm that determines the class of each training vector in high dimensional space. The SVM determines the classes that will determine the support vectors of the data and the output of the hyper plane and the system. At the moment of training, it determined the support vectors by linear, polynomial or sesamoid functions. In this study, linear and polynomial SVM kernels were used for GMM super-vectors.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^t \boldsymbol{x}_j \quad (9)$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^t \boldsymbol{x}_j + 1)^n \quad (10)$$

The stages of the hybrid model are shown in Figure 5.

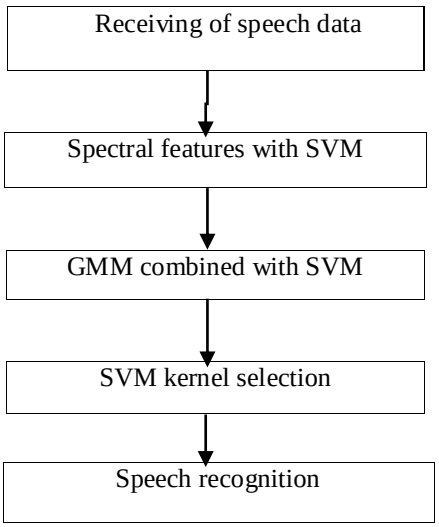

Fig.5. Hybrid model steps

IV. EXPERIMENTAL STUDY

A unique and genuine Turkish language database has been employed in this study. Names, family names, ages, speeches and genders of the persons were added to this database. In this database, the numbers of five senses in males and females as angry, fearful, sad, happy and neutral are shown in Table 1.

Voice samples have been tested by training them, using available feature vectors by means of ANN, HMM ,DTW, GMM and hybrid methods. Success rates of speech samples obtained utilizing are given for ANN, HMM ,DTW, GMM in the Table 2.

TABLE I
VOICE DATABASE

	Female	Male	Total
Anger	124	241	365
Fear	178	157	335
Sadness	274	256	530
Happiness	179	364	543
Neutral	572	634	1206

TABLE II
THE SUCCESS OF THE METHODS WITH SVM

	ANN	HMM	DTW	GMM
Male	74.62	75.71	69.97	71.60
Female	75.34	77.63	70.79	72.39

TABLE III
SUCCESS IN CLASSIFICATION FOR MFCC (5 FEATURE VECTORS)

	ANN	HMM	DTW	GMM
Male	72.64	77.25	67.21	70.25
Female	71.47	75.68	70.24	68.17

Success rates of speech samples obtained utilizing MFCC 5 feature vectors are given for ANN, HMM, DTW and GMM in the Table 3. HMM gave more successful results when compared to other techniques.

TABLE IV
SUCCESS IN CLASSIFICATION FOR MFDWC (5 FEATURE VECTORS)

	ANN	HMM	DTW	GMM
Male	71.37	75.09	65.38	69.36
Female	70.28	74.34	68.09	68.55

Success rates of speech samples obtained utilizing MFDWC 5 feature vectors are given for ANN, HMM, DTW and GMM in the Table 4. HMM gave more successful results when compared to other techniques.

TABLE V
HYBRID METHODS FOR DIFFERENT SVM KERNELS

	Linear kernel	Polynomial kernel
Male	76.78	80.67
Female	79.85	81.37

Success rates of speech samples obtained hybrid methods for different SVM kernels (linear, polynomial) in the Table 5. Hibrit Model gave more successful results when compared to all other techniques.

V. CONCLUSION

Speech recognition and speech emotion recognition plays an important role in our day due to security and many other reasons. Speech emotion recognition of systems have been developed, being based on an unique database obtained by utilizing Turkish language in this study. Classification success of the methods employed in the study have been calculated and results are demonstrated in a comparative manner. Hybrid Method provided more successful results compared to the other speech emotion methods when the results are taken into consideration. This hybrid model has been carried out by combining with SVM and GMM. In first stage of this model, with SVM has been performed subsets obtained vector of spectral features. Hybrid model yielded better results compared to other methods that are used in other literature. Moreover, success rates of speech samples obtained employing MFCC and MFDWC feature vector. Success rates of speech samples obtained employing 5 feature vector. MFCC gave more successful results compared to MFDWC.

REFERENCES

[1] Mohammed Shami, Wemen Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech", Speech Communication, 2007, 49(3), p.201-212.

[2] Lijiang Chen , Xia Mao, Yuli Xue , Lee Lung Cheng , "Speech emotion recognition: Features and classification models", Digital Signal Processing, 22(6), 2012, p.1154-1160.

[3] Ling He, Margaret Lech, Namunu C. Maddage, Nicholas B. Allen, "Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech", *Biomedical Signal Processing and Control*, 2011, 6(2), p.139-146.

[4] Tim Polzehl , Shiva Sundaram , Hamed Ketabdar , Michael Wagner and Florian Metze, "Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features", *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association*, 2009.

[5] Halicioglu, Tin Lay Nwe, Foo Say Wei and Liyanage C De Silva, "Speech Based Emotion Classification", TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, 2001.

[6] Jasmine Bhaskar, Sruthi Ka and Prema Nedungadi, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining", *Procedia Computer Science*, 2015, 46, p.635-643.

[7] Jinkyu Lee and Ivan Tashev. "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition", *Interspeech 2015*, 2015.

[8] S.Oh and C.Suen, "A class-modular feed forward neural network for handwriting recognition", Pattern Recognition, 2002, 35(1), p.229-244.

[9] Dimitros and Kontropulos, "Emotional speech recognition: Resources, features, and methods", Speech Communication, 2006, 48(9), p.1162-1181.

[10] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech Audio Proc., 1995, 3, p. 72–83.

[11] Seok, Oh and Ching, Suen, "A class-modular feed forward neural network for handwriting recognition", Pattern Recognition, 2002, 35(1), p.229-244.

[12] Lihang, Li, Dongqing, Chen and Sarang, Lakare etc, "Image segmentation approach to extract colon lümen through colonic material taggng and hidden markov random field model for virtual colonoskopy", Medical Imaging, 2002.

[13] Edmondo, Trentin and Marko, Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition", Elsevier Neurocomputing 37, p.91-126, 2001.

[14] Lindasalwa, Muda and Mumtaj, Began, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal Computing, 2010, 2(3), p.138-143, ISBN 2151-9617, 2010.

[15] Hao Hu, Ming-XingXu, and Wei Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition", *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007.

[16] Cigdem Bakir, "Automatic Speaker Gender Identification for the German Language", *Balkan Journal of Electrical&Computer Engineering*, 2015, 4(2), p.79-83, 2015.

[17] Cigdem Bakir, "Automatic voice and Speech Recognition System for the German Language", *1st International Conference on Engineering Technology and Applied Sciences*, 2016, p.131-134.

[18] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, vol.2, issue 3, p.138-143, ISSN 2151-9617, 2010.

[19] M., Fahid M. and M.A, "Robust Voice conversion systems using MFDWC", 2008 International Symposium on Telecommunications, p.778-781, 2008.

BIOGRAPHIES

CIGDEM BAKIR was born in İstanbul. She received the B.S. degrees in computer engineering from the University of Sakarya, in 2010 and the M.S. degree in computer engineering from Yildiz Technical University, İstanbul, in 2014. Since 2012, she was a Research Assistant with the Yildiz Technical University. She works a Research Assistant with the Igdir University. Her research interests include recommendation systems, information security, data mining, image processing and biomedical signal processing.

MECIT YUZKAT received the B.S. degrees in computer engineering from the University of Trakya and M.S. degrees in computer engineering from the University of Yildiz Technical University. He works a Research Assistant with the Mus Alparslan University. Her research interests include process mining algorithm, data mining, image processing and biomedical signal processing.