Global Journal of Computer Sciences: Theory and Research

Volume 07, Issue 1, (2017) 48-57

www.gjcs.eu

Speech recognition system for Turkish language with hybrid method

Cigdem Bakir*, Computer Engineering Department, Yildiz Technical University, Davutpasa Campus, Istaanbul 34220, Turkey.

Suggested Citation:
Bakir, C. (2017). Speech recognition system for Turkish language with hybrid method. *Global Journal of Computer Sciences: Theory and Research. 7*(1), 48-57.

Received December 21, 2016; revised February 8, 2017; accepted April 16, 2017.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

Currently, technological developments are accompanied by a number of associated problems. Security takes the first place amongst such problems. In particular, biometric systems such as authentication constitute a significant fraction of the security problem. This is because sound recordings having connection with various crimes are required to be analysed for forensic purposes. Authentication systems necessitate transmission, design and classification of biometric data in a secure manner. The aim of this study is to actualise an automatic voice and speech recognition system using wavelet transform, taking Turkish sound forms and properties into consideration. Approximately 3740 Turkish voice samples of words and clauses of differing lengths were collected from 25 males and 25 females. The features of these voice samples were obtained using Mel-frequency cepstral coefficients (MFCCs), Mel-frequency discrete wavelet coefficients (MFDWCs) and linear prediction cepstral coefficient (LPCC). Feature vectors of the voice samples obtained were trained with k-means, artificial neural network (ANN) and hybrid model. The hybrid model was formed by combining with k-means clustering and ANN. In the first phase of this model, k-means performed subsets obtained with voice feature vectors. In the second phase, a set of training and tests were formed from these sub-clusters. Thus, for being trained more suitable data by clustering increased the accuracy. In the test phase, the owner of a given voice sample was identified by taking the trained voice samples into consideration. The results and performance of the algorithms used for classification are also demonstrated in a comparative manner.

Keywords: Speech recognition, hybrid model, k-means, artificial neural network (ANN).

* ADDRESS FOR CORRESPONDENCE: **Cigdem Bakir,** Computer Engineering Department,Yildiz Technical University, Istanbul 34220, Turkey. *E-mail address*: cigdem.bakr@gmail.com / Tel.: +0-212-383-5756

1. Introduction

Currently, security problems have begun to arise along with developments in technology. Studies have been accomplished, especially in order to prevent information belonging to some people from being transferred to other people in commercial transactions. Some of these studies are hand script recognition, signature recognition, face recognition, iris recognition and voice recognition [1].

Various studies have been carried out on voice and speaker recognition. Jie-Fu *et al.* [2] collected voice samples in Chinese from seven males and five females aged between 25 and 45. Attempts were made to identify the owner of the voice by trying to analyse these voice samples by means of their tone, vowels, consonants and syllables. Voice samples were separated into four frequency groups, and each frequency band was analysed. However, this study did not test very big data. In addition, the intended success was not exactly achieved because it was performed by taking the similarities with the English language into consideration.

Tokuda *et al.* [3] developed the English speech synthesis system using hidden Markov model. This system was developed for speaker recognition and specifies the structure by changing the voice feature. However, the characteristic feature of the synthesised voice in the study is pretty low.

Reynolds *et al.* [4] implemented the SuperSID project to enhance the performance of speaker recognition systems. The purpose of this project was to develop speaker recognition systems and employ the most suitable features in order to increase the accuracy. However, this study failed to fully achieve the acoustic characteristics of the voice and removal of noise.

Reynolds *et al.* [5] attempted to substantiate speaker identification and verification using Gaussian mixture model (GMM) method. Attempts were made to determine speaker verification in the system according to the probability distribution. Eleven different hypotheses were developed for this probability distribution. Data used in the study were extracted from telephonic conversations.

Danio *et al.* [6] developed a voice recognition application in English and Mandarin Chinese by using the end-to-end deep learning method. This system is the structure of recurrent neural network, which contains multiple repetitive layers over each other. This method aims to reduce the calculation costs in voice recognition systems and was done using noises that occur in different languages.

Dimitros and Kontropulos [7] conducted a voice recognition study to find the emotions of persons by their speech. The most frequently used acoustic feature of the voice was determined according to the emotional contexts. In this study, the number of the status of the emotions in the current sound recordings was determined with the number and variety of speakers. The voice vectors obtained with Mel-frequency cepstral coefficients (MFCCs) were coached by using classification techniques such as ANN, HMM, *k*-nearest neighbours and SVM.

Speech has an important place in communication, and for this reason the voice recognition study was carried out. In this study, simulation was also performed to solve the voice recognition problem related to security risk. However, certain difficulties came in the way while creating the voice database. The first and the greatest difficulty was that words were vocalised at different speeds and in different pronunciations by different persons. In addition, factors such as noise occurring in the environment and voice while recording the voice data, toning effect and syllable stress make the voice recognition process difficult [8].

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

2. 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 as *k*-means, artificial neural network (ANN) and hybrid method (*k*-means + ANN). 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 1, 3 and 5 feature vectors and the success of the methods was presented comparatively by training the feature vectors, obtained from speech signals.

2.1. Mel Frequency Cepstral Coefficients

MFCC is a feature extraction method that is used in sound processing. It is used to extract important information and features by dividing sound data into its subsets. The steps in the feature extraction technique of MFCC are shown in Figure 1 [9].

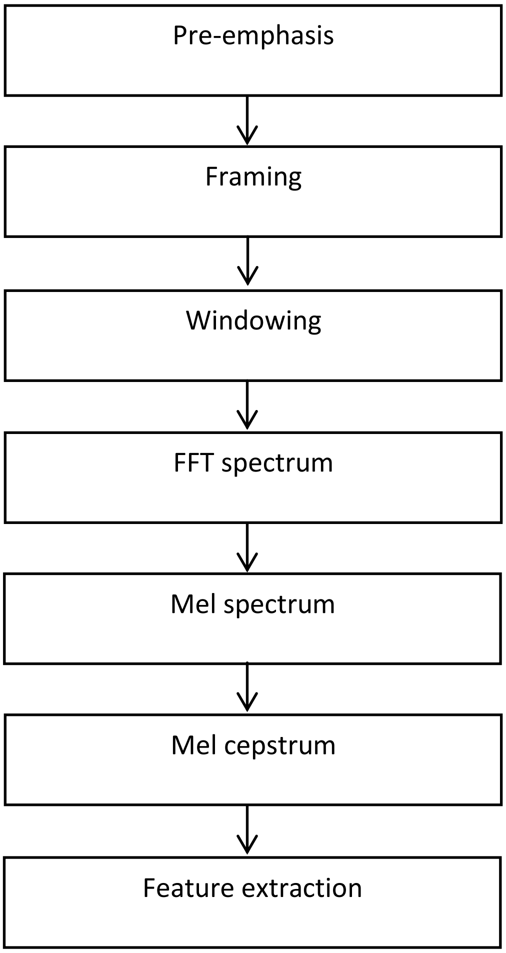


Figure 1. DFIM supplied by two PWM inverters

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

High-frequency sound signals are passed through a filter at this stage. In this way, the energy of the sound is increased at high frequency. The sound signals are analogs and are converted to digital by dividing into small frames between 20 and 40 ms in the framing stage; they are divided into *N* frames.

The sound signal is moved by sliding the sound signal at the windowing stage. In this way, the closest frequency lines and the frame that comes by windowing are combined. The window type, width and sliding amount are determined at this stage. Each *N* frame 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 that are formed while passing through a series of filters. The MEL spectrum reduces 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 the frequency space to time space. The MEL frequency cepstrum factors are obtained using discrete cosine transform in time space.

2.2. Mel-Frequency Discrete Wavelet Coefficients

The study in question was performed based on a unique database comprising Turkish voice samples collected from both men and women. These voice samples were separated into various feature vectors with MFDWC, and trained. MFDWC is a feature extraction method employed in speech processing. It is used to extract significant information and features by dividing voice data into subsets. The feature extraction steps of the MFDWC technique are shown in Figure 2 [10].

Sample speech signal is shown in the 40–40,000 Hz range in the MFDWC feature extraction method. Speech signal is divided into frames after the pre-processing step. Hamming window was used in this study to smoothen the transition of speech samples between the frames. One Mel shows the frequency of voice tone. Mel-scale is scaled between the actual frequency of voice signal and the estimated voice frequency. For this reason, the total energy of every frame is calculated. The classification success in speaker identification was calculated on an individual basis for MFDWC-1, MFDWC-3 and MFDWC-5 vectors by training the feature vectors obtained from voice signals by means of ANN and hybrid method.

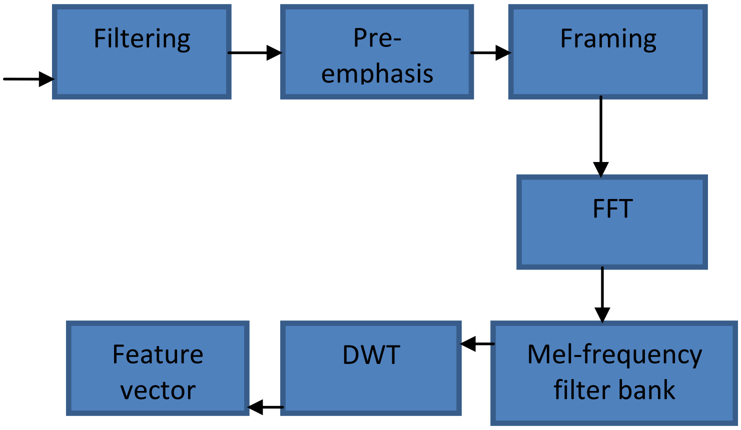


Figure 2. Feature extraction steps of MFDWC

2.3. Linear Prediction Cepstral Coefficient

LPCC is a commonly used technique to obtain the characteristic features from sound signals. In this technique, each sample of sound signal is based on the conversion of linear prediction coefficients obtained as a linear weighted total of the previous sound signals into cepstral coefficients. This is not a method preferred for sound signals exposed to various environmental effects or noise. LPCC utilises functions that model the sound path.

The LPCC method is obtained by converting LPC coefficients into cepstral coefficients through Fourier conversion. The preliminary process is completed by transmitting speech signals through high filter. Autocorrelation characterises the signal by determining the similarity of each sound signal with itself. This step is materialised in the frame of each signal. Signal is analysed by converting the autocorrelation values into LPC parameters using Levinson–Durbin recursion. In the final phase, LPCC parameters are obtained with cepstral analysis [11]. The steps in LPCC feature extraction are shown in Figure 3.

The LPCC method is calculated as in Equality 1 [12]. a_i LPC coefficients indicate the degree of **p** LPC coefficients.

$$s(n)=\sum_{n=1}^{p}a_i s(n-i) \tag{1}$$

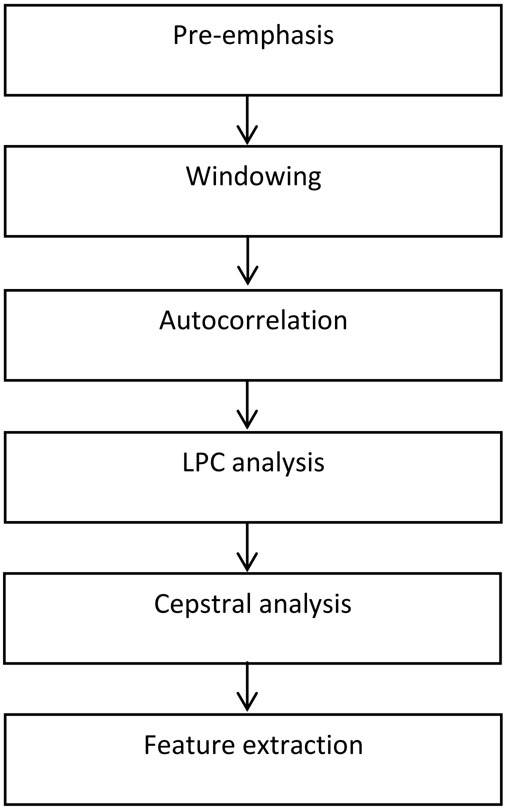


Figure 3. Feature extraction steps of LPCC

3. Feature Extraction Methods

In Figure 4, the steps of the study are given. In this study, sound samples taken from 25 males and 25 females in different age groups were separated from their feature vectors using MFCC, MFDWC and LPCCC. Education and test samples were formed from these voice feature vectors. These education and test samples were coached according to the ANN and recommended hybrid model, while the voice recognition transaction was realised automatically. The results obtained with the ANN and hybrid method are given comparatively.

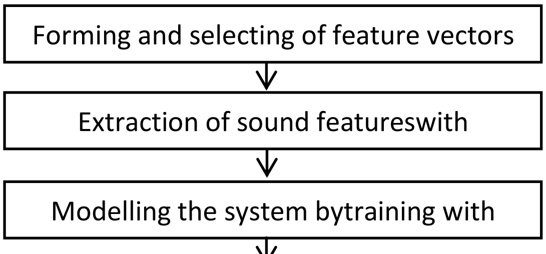

Forming and selecting of feature vectors
Extraction of sound featureswith
Modelling the system bytraining with

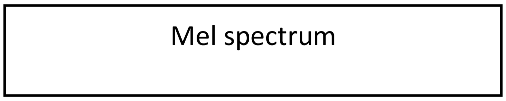
Figure 4. Study steps

3.1. K-Means Clustering

K-means is one of the learning methods that are non-educational and group objects based on their similarities. *K*-means clustering, which is a method based on being divided, calculates *N* pieces of objects according to their distance to the *k* cluster centre and include them in the cluster they are near to. The cluster centre is determined by randomly taking the average of one or more samples at the beginning and is recalculated at each iteration. At every turn, the similarities of all data are found according to the new cluster centres. These steps are iteratively repeated. The steps come to an end when the ratio of clustering error becomes a minimum.

The pseudo-code of *k*-means clustering method is as follows [13]:

```
Input:
D = {d1, d2,......,dn} //set of n data items.
k // Number of desired clusters
Output:
A set of k clusters.
Steps:
1. Arbitrarily choose k data items from D as initial centroids.
2. Repeat
Assign each item D to the cluster that has the closest centroid;
Calculate new mean for each cluster;
Until convergence criteria is met.
```



3.2. Artificial Neural Network

ANNs have very wide fields of application such as in the automotive, banking, defense, electronics, entertainment, finance, insurance, manufacture, oil and gas, robotics, telecommunication and transportation industries.

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

ANNs comprise an input layer, an output layer and hidden layers. Data are received into neural networks through the input layer, and are transferred outside through the output layer. The layers between the input and output layers constitute the hidden layers.

Neurons in the feed-forward neural networks are connected only in the forward direction [1]. Each layer of neural network contains the connection to the next layer; these connections are never 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. The structure of a feed-forward ANN is shown in Figure 5.

Backward propagation network shows how to train a neuron [14–16]. Trainer is a sort of learning. The 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 whose inputs are given. Error is calculated in case the expected outputs are used, and the weights of various layers are adjusted in the backward direction from the output layer to the input layer. In other words, it is given for both input data and output data. The 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. 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 proportional to their weights.

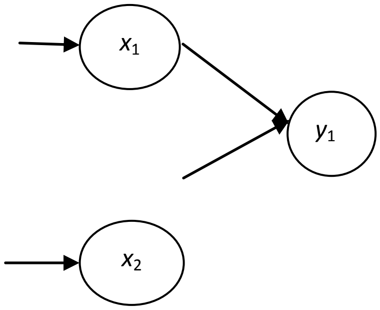


Figure 5. Feed-forward neural network

Layers are located one after another in a multilayer ANN. Outputs of neurons in a layer are given as their weights, to the input of the next layers, and these weights are used for calculation of outputs for the next layer. The weights of the hidden layer between the input and output layers are calculated [14].

3.3. Hybrid Method

The steps of the hybrid method are shown in Figure 6. In this study, the voice feature vectors obtained with MFCC, MFDWC and LPCCC were decomposed with k-means. Random education and test samples were selected from these clusters. Education and test samples were coached with ANN and voice recognition transaction was realised. In this way, the performance and accuracy were increased by categorising and classifying more appropriate data.

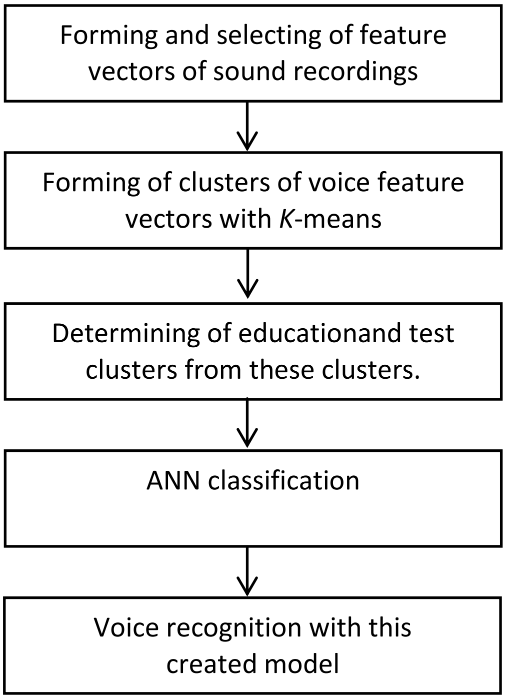

Forming and selecting of feature vectors of sound recordings
Forming of clusters of voice feature vectors with *K*-means
Determining of educationand test clusters from these clusters.
ANN classification
Voice recognition with this created model

Figure 6. Hybrid method of steps

4. Experimental Study and Results

In this study, an authentic and unique Turkish database was used. The names, surnames, age, sex and speeches of persons have been added to this database. The different numbers of feature vectors of sound components were extracted using MFDWC, MFCC and LPCC feature extraction method. The voice samples, which look alike, were combined in the same cluster with *k*-means. In the next stages, the sound samples were trained using the ANN and hybrid method. The features of the recorded sound samples are given in Table 1.

Table 1. Attributes of used databases

Age range	Number of speakers	
	Male	Female
18–25 range speakers	10	9
26–40 range speakers	7	11
41 and more speakers	8	5

Feature vectors of voice components with different quantities were extracted by means of MFDWC, MFCC and LPCC feature extraction method. Voice samples were tested by training them, using available feature vectors by means of ANN and hybrid method.

Table 2. Success in classification for MFCC

Feature extraction method	Classification method	
	ANN (%)	Hybrid method (%)
MFCC-1	75.81	81.24
MFCC-3	77.02	82.37
MFCC-5	82.86	87.78

The success rates of speech samples obtained utilising MFCC different feature vectors are given for ANN and the hybrid method in Table 2. Success in speaker identification increases as the number of words used increases in all the techniques employed. The hybrid method gave more successful results compared to ANN.

Table 3. Success in classification for MFDWC

Feature extraction method	Classification method	
	ANN (%)	Hybrid method (%)
MFDWC-1	70.14	75.27
MFDWC-3	72.67	79.31
MFDWC-5	73.68	81.37

The success rates of speech samples obtained utilising MFDWC feature vectors for ANN and hybrid method are tabulated in Table 3. Success in speaker identification increases as the number of words used increases in all the techniques employed. The hybrid method gave more successful results compared to ANN.

Table 4. Success in classification for LPCC

Feature extraction method	Classification method	
	ANN (%)	Hybrid method (%)
LPCC-1	71.06	76.27
LPCC-3	75.25	79.80
LPCC-5	80.71	85.35

The success rates of speech samples obtained utilising LPCC different feature vectors for ANN and the hybrid method are given in Table 4. Success in speaker identification increases as the number of words used increases in all the techniques employed. The hybrid method gave more successful results compared to ANN.

The success rates of speech samples obtained employing MFDWC, MFCC and LPCC feature vectors. The success rates of speech samples obtained employing five feature vectors for all the feature extraction techniques.

A unique and genuine Turkish language database was employed in this study. Names, family names, age, speech and gender of the persons were added to this database. Feature vectors of voice components with different quantities were extracted by means of MFCC, MFDWC and LPCC feature extraction method. Voice samples were tested by training them, using the available feature vectors by means of ANN and hybrid method. In the testing phase, successful classification techniques were determined by the available testing example. It was also presented and compared by calculating the success of other methods used.

5. Conclusion

Voice recognition plays an important role currently, due to security and other reasons. Speech recognition of systems was developed in this study, based on a unique database obtained by utilising the Turkish language. The success of the methods employed in the study is calculated and the results are demonstrated comparatively. The hybrid method provided more successful results compared to ANN. The speech recognition system is more successful for MFCC-5 compared to the results obtained utilising all the other feature extraction techniques.

References

[1] O. Seok and S. Ching, “A class-modular feed forward neural network for handwriting recognition,” *Pattern Recognit.*, vol. 35, issue 1, pp. 229–244, 2002.

[2] J.-F. Quan *et al*., “Importance of tonal envelope cues in Chinese speech recognition,” *J. Acoust. Soc. Am.*, vol. 104, issue 1, pp. 505–510, 1998.

[3] T. Keiichi *et al.*, “An HMM-based speech synthesis system applied to English,” in: *Proceedings of 2002 IEEE SSW*, 2012, pp. 227–230.

[4] R. Douglas *et al.*, “The SuperSID project: exploiting high-level information for high-accuracy speaker recognition,” in: *Proceedings of ICASSP*, Hong Kong, 2003, pp. 784–787.

[5] R. Douglas *et al.*, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, pp. 19–41, 2000.

[6] D. Amodei and S. Ananthanarayanan, “Deep speech 2: end-to-end speech recognition in English and Mandarin,” in: *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 173–182.

[7] D. Ververids and C. Kotropoulos, “Emotional speech recognition: resources, features and methods,” *Speech Commun*., vol. 48, pp. 1162–1181, 2005.

[8] G. Wouter *et al.*, “Neural networks used for speech recognition,” *J. Autom. Contr.*, vol. 20, pp. 1–7, 2010.

[9] L. Muda *et al.*, “Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques,” *J. Comput.*, vol. 2, issue 3, pp. 138–143, 2010. ISSN 2151-9617.

[10] M. Fahid and M. A. Tinati, “Robust voice conversion systems using MFDWC,” in: *International Symposium on Telecommunications*, 2008, pp. 778–781.

[11] M. Zbancioc and M. Costin, “Using neural networks and LPCC to improve speech recognition,” in: *International Symposium on Signals, Circuits and Systems*, vol. 2, 2003, pp. 445–448.

[12] O. Eray, *Destek Vektör Makineleri ile Ses Tanıma Uygulaması*. Pamukkale Üniversitesi, 2008.

[13] K. A. Abdul Nazeer and M. P. Sebastian, “Improving the accuracy and efficiency of the k-means clustering algorithm,” in: *Proceedings of the World Congress on Engineering 2009 Vol IWCE 2009*, 2009.

[14] L. Lihang *et al.*, “Image segmentation approach to extract colon lümen through colonic material tagging and hidden Markov random field model for virtual colonoscopy,” in: *Proceedings of Medical Imaging 2002: Physiology and Function from Multidimensional Images*, 2002.

[15] C. Bakir, “Automatic voice and speech recognition system for the German language with deep learning methods,” *Int. J. Appl. Mathematics Electron. Comput.*, vol. 4, pp. 399–403, 2016.

[16] C. Bakir, “Automatic speaker gender identification for the German language,” *Balkan J. Electr. Comput. Eng.*, vol. 4, issue 2, pp. 79–83, 2016.