PERFORMING ACCURATE SPEAKER RECOGNITION BY USE OF SVM AND CEPSTRAL FEATURES

Zülfikar ASLAN[1,*], **Mehmet AKIN**[2]

[1] Technical Sciences Vocational School, Gaziantep University, Gaziantep, Turkey
[2] Electrical-Electronics Engineering, Faculty of Engineering, Dicle University, Diyarbakır, Turkey
*Corresponding Author: zulfikaraslan@gantep.edu.tr

ABSTRACT: The task of performing speaker recognition over voice recordings is an active research area in the relevant literature in which many applications has been proposed so far. In this study, speaker recognition is performed over cepstral features extracted from raw voice recordings. Some of the most prominent cepstral feature selection methods, namely, LPC, LPCC, MFCC, PLP and RASTA-PLP are utilized and their contribution to the performance of the applied method is investigated. Obtained features are handled by SVM classification algorithm to finalize the speaker recognition task. As a result, it is observed that cepstral feature selection methods such as LPCC and MFCC combined with SVM classification result in around 97% accuracy.

Keywords: Speaker recognition, cepstral feature selection, SVM

1. INTRODUCTION

Performing automatic speaker recognition by use of voice recordings is an active research field in the relevant literature. Human voice, which is available in analog format in the real world, has to be digitized by a series of processes to be saved in the computer because the original format (analog) cannot be retained in the computer memory as it is [1]. This digital data then can be analyzed by use of several methods most of which rely on statistical and spectral synthesis of the data. Because automatic speaker recognition has gained a lot of attention and still preserves significance, several methods has been proposed in the literature [2]. Besides, such automatic systems have often been utilized in real-life practical applications. Forensic investigations, mobile banking, personal computer applications executed by voice commands are among several examples of such applications.

First and the most important step in automatic speaker recognition task is to extract useful features from the raw voice data. The digitized voice signal is observed to have stationary behaviour in the short-term time period while this situation may change in a longer period of time. Therefore, researchers in the relevant literature often found it more appropriate to analyze the signal in short periods of time intervals[3]. Especially cepstral analysis methods, such as LPCC [4] and MFCC [5], adopt this approach. In addition to these powerful feature extraction techniques, other methods, namely LPC [6], PLP and RASTA-PLP [7] have also found an important place in the relevant literature.

In this study, digitized voice signals are analyzed by use of different cepstral feature extraction methods. As a result of this analysis, obtained features are used to classify voice recordings.

Support Vector Machines (SVM), which is a state-of-art classifier and highly studied in the machine learning community, is used to classify the feature vectors. The aim of this study is to investigate the effect of different feature extraction methods over the performance of automatic speaker recognition task. Therefore, this study focuses on revealing which feature extraction method is more appropriate and useful to retain speaker related information that resides in the signal.

The paper is composed of two main sections: In the first one, the dataset and the utilized methods will be explained while in the second part, experimental results will be presented along with our remarks and conclusion about the results.

2. MATERIAL AND METHODS

2.1. Dataset

The dataset of the study is collected by having 10 people (6 male and 4 female) say 5 words each for 10 times and recording them on the computer. As a result, 50 files are recorded per person and that made a total of 500 files. During the recording process, each file is created in mono, 16 bit format by using a dynamic microphone. The frequency of the voice recordings is 11025 Hz. Each voice recording is edited to make all of them of the same length, namely a vector of 9000 size. Therefore, the digitized voice recordings are eventually transformed into a 500x9000 matrix (500 records each of which is of 9000 length) stored in the computer.

2.2. Methods

2.2.1. Linear Predictive Coding

Linear Predictive Coding (LPC), is a calculation method which linearly combines a few of the most recent samples. LPC [8],[9] has been a predominant technique in estimating the basic parameters of the speech. It is proven to be both accurate in estimating the parameters and effective in doing the calculations. The main idea behind the LPC algorithm is to approximate a speech sample by use of "predictive coefficients" that is calculated by using a linear combination of few previous samples. The error raised by the difference between an estimated sample and the real sample is reduced by use of an optimization process that improves the predictive coefficients. These coefficients form the basis for LPC algorithm [10]. Figure1 depicts the main steps involved in the process of LPC feature extraction.

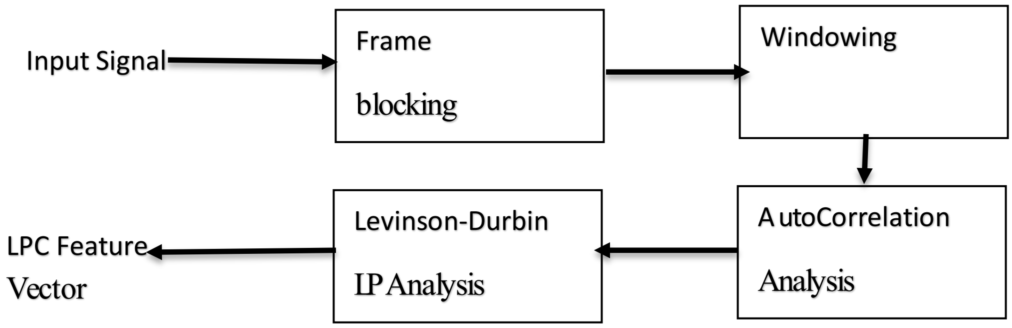

Figure 1. LPC flow diagram

2.2.2. Linear Prediction Cepstral Coefficient

Linear Prediction Cepstral Coefficient (LPCC) briefly takes features obtained after LPC and uses them to perform cepstral analysis. LPCC is also used to estimate the spectrum of the signal [4]. Like LPC, it too uses a linear combination of previous samples to approximate a sample. The coefficients required by LPCC is calculated during the optimization process to minimize the squared error rate. Pre-emphasis of the voice signal is the first step in the process of linearizing speech spectrum. This step, increases high frequency samples. The next step is to frame the signal and to multiply it with the windowing function in order to reduce the spectral leakage that may be available in the speech frame. LPC is an all-pole resonance modeling and it can be used to obtain automatic regression coefficients. In the final step, the cepstrum is calculated through cepstral analysis. The cepstral coefficients can be calculated by using a series of recursive functions applied over LPC [11]. The flow diagram of the LPCC is shown in Figure2.

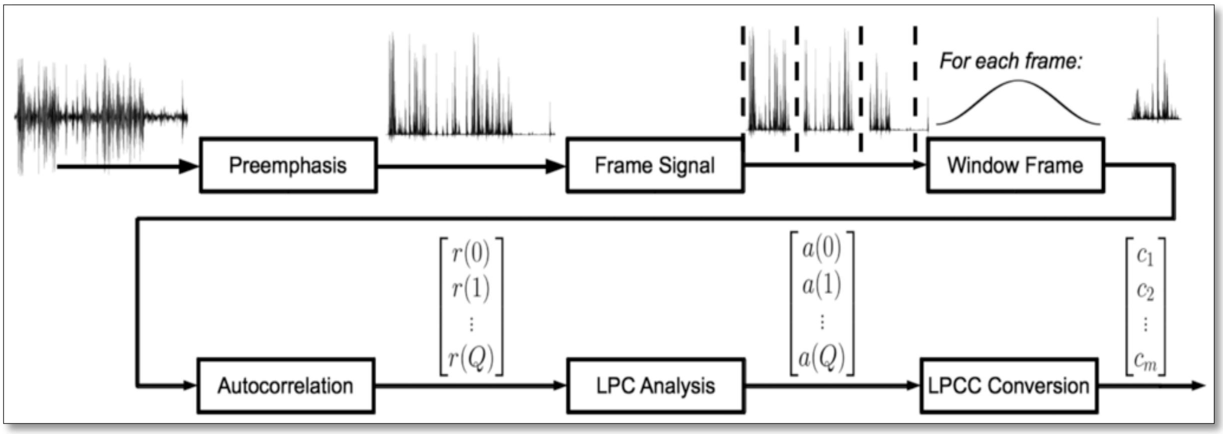
Preemphasis
Frame Signal
For each frame:
Window Frame
Autocorrelation
LPC Analysis
LPCC Conversion
$r(0)$, $r(1)$, $\vdots$, $r(Q)$
$a(0)$, $a(1)$, $\vdots$, $a(Q)$
c_1, c_2, $\vdots$, c_m

Figure 2. Linear predictive cepstral coefficient (LPCC) calculation.[12]

2.2.3. Mel Frequency Cepstral Coefficients

In voice signals, Mel-frequency Cepstrum (MFC), is a short-term power spectrum representation of the voice that is based on cosine transformation of a frequency into the log-power spectrum in the non-linear mel scale. Mel-frequency cepstral coefficients (MFCC) are coefficients that constitute the MFC. They are derived from a cepstral representation of the voice signal (i.e., non-linear “spectrum of a spectrum”). The difference between cepstrum and mel-frequency cepstrum is that in the latter the frequency bands are equally spaced on the mel scale which approximates human voice response better than cepstrum. MFCCs are often used in speech recognition systems such as the application in which numbers mentioned in a phone conversation is automatically detected. Speaker recognition systems, as well as recognition of human voice, are also among possible application areas of MFCCs [13]. The flow diagram of the MFCC is shown in Figure3.

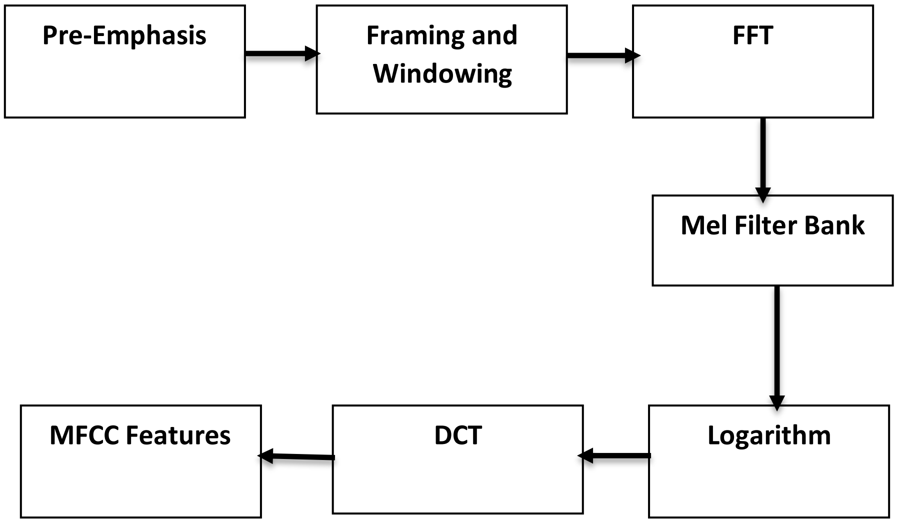

Figure 3. Block Diagram of MFCC[14]

2.2.4. Perceptual Linear Predictive

The Perceptual Linear Predictive (PLP) model was developed by Herman Sky in 1990. The original PLP model aimed at better defining human hearing psychophysics during the feature extraction phase. Similar to LPC, PLP is based on short-term spectrum of the voice signal. On the contrary to the use of pure linear predictive analysis of the voice signal, PLP modifies the short-term spectrum of the voice by using various psychophysics based transformations [15]. The flow diagram of the PLP is shown in Figure 4.

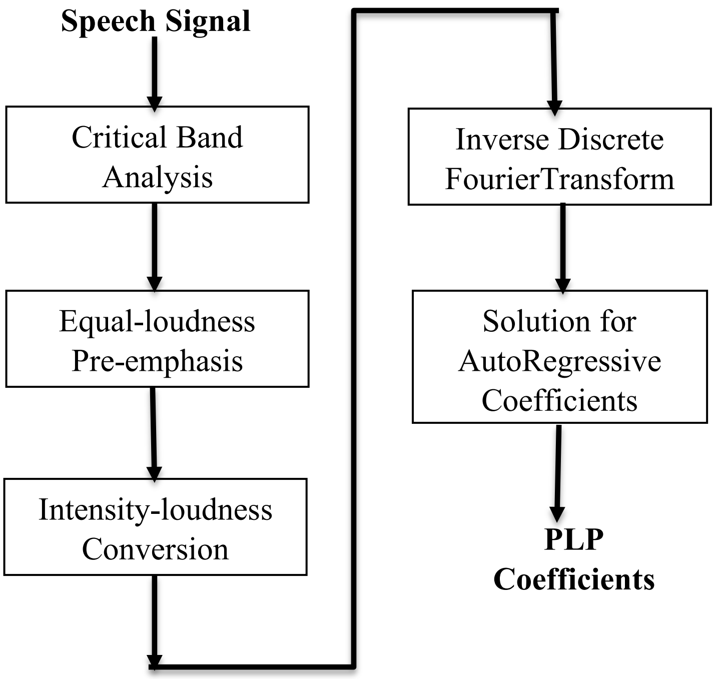
Speech Signal
Critical Band Analysis
Equal-loudness Pre-emphasis
Intensity-loudness Conversion
Inverse Discrete FourierTransform
Solution for AutoRegressive Coefficients
PLP Coefficients

Figure 4. Block diagram of Perceptual linear prediction[16]

2.2.5. Relative Spectra Perceptual linear Predictive

Relative Spectra Perceptual Linear Predictive (RASTA-PLP) is based on short-term spectrum of the speech. In order to improve the outcome of standard PLP, it uses different physiologic based transformations over short-term spectrum. RASTA-PLP is one of the most powerful and useful voice analysis techniques that is used to encode a low volume voice with low number of bits. It estimates the speech parameters highly accurately. Short-term spectral values are modified by the frequency response of the communication and that makes the technique vulnerable. The RASTA process provides feature extraction for speaker recognition, improvement and compression of the voice signal. Moreover, the RASTA filter can be used in log-spectral or cepstral analysis fields [17],[18].

2.2.6. Support Vector Machine (SVM)

Support Vector Machine (SVM), is a machine learning algorithm that can be used for classification and regression purposes [19]. It was developed by Vapnik [20] and relies mainly upon the Vapnik-Chervonenkis (VC) theory [21]. It aims to classify two classes of samples by providing the maximum margin between samples. Despite it can be defined as a linear classification algorithm, it can well solve non-linear classification problems by moving the sample space to a higher dimensional space through a technique called “kernel-trick”. SVM is often used in machine learning literature to solve several problems including all kinds of classification problems. Figure 5 shows how a SVM network operates.

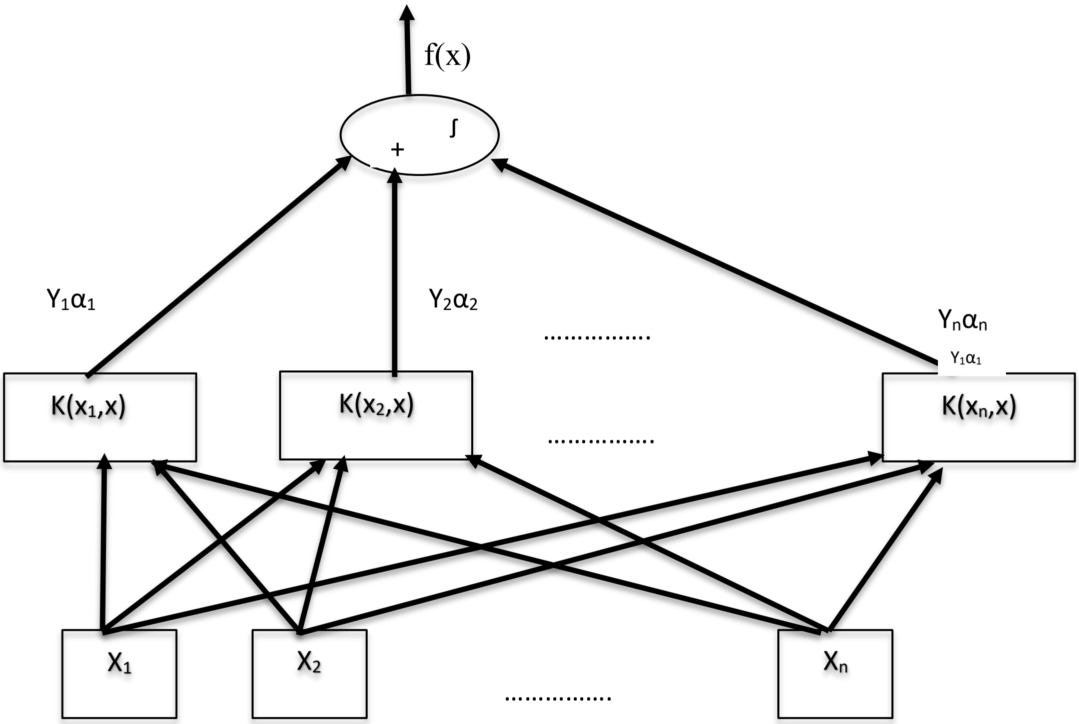

Figure 5. SVM network architecture [22].

When the network architecture given in Figure5 is investigated, it can clearly be seen that the main variables that affect how the system works are kernel functions and α Lagrange coefficients.

Table 1. Common kernel functions.

Polynomial	$K(a,b) = (1 + \sum_j a_j b_j)^d$
Radial Basis Functions	$K(a,b) = \exp\left(-(a-b)^2/2\sigma^2\right)$
Sigmoid	$K(a,b) = \tanh\left(ca^T + h\right)$

Kernel functions are used to transform the samples into a higher dimensional space while Lagrange coefficients are the weights of these calculations. Common kernel functions are shown in Table1. An output of a sample is obtained by summing over the multiplications of kernel function outputs by weights. The optimization process to find the weights aims at finding the hyperplane that separates two classes at the maximum margin. Although SVM is initially a binary classifier, i.e., designed to classify only two classes of samples, researchers also proposed ways to use SVM as a multi-class classifier [23].

EXPERIMENTAL RESULTS

As explained in Section 2.1, 500 voice recordings are processed and each of these files are transformed into a vector of size 9000. Further, each vector has an additional column that specifies which vector belongs to which speaker. In this study, this raw data is not used to detect speaker identity. Instead, this overly long vectors are processed by feature extractors to extract more useful features out of the vectors. This is mainly because the raw dataset does not directly reveal speaker specific information and in addition to that, processing vectors of that long reduces classification performance of SVM. Therefore, each vector is primarily processed by cepstral feature extraction methods before given to SVM.

The first cepstral feature extraction method used is LPC which requires us to choose an order parameter to analyze the signals. In accordance with the chosen order, number of features produced by the method changes. Table 1 presents best values for order parameters and respectively obtained number of features and classification accuracy.

Table 2. Classification accuracy and number of features obtained by varying order parameter of LPC

	Order=5	***Order=7***	***Order=9***	***Order=12***
1	0,950	0,969	0,980	0,980
2	0,970	0,980	0,960	0,970
3	0,959	0,970	0,960	0,970
4	0,990	0,990	0,970	0,940
5	0,950	0,950	0,970	0,970
Average	0,964	0,970	0,968	0,966

Number of Features	19	25	31	40

As can be seen in Table 2, the effect of order parameter over classification accuracy is slight and all results seem to be comparable to each other. With the increasing order value, however, number of features also increase which in turn results in more computation time.

Table 3. Classification performance of LPCC method depending on different cepstra numbers

	Number of cepstra		
	10	**5**	**3**
1	0,990	1	0,950
2	0,970	0,98	0,880
3	0,971	0,99	0,918
4	0,990	0,970	0,908
5	0,960	0,970	0,842
Average	0,976	0,982	0,899
Number of Features	31	16	10

In the next step, we utilized LPCC as the feature extractor method over raw vectors. This method also performs analysis of cepstral features and requires the number of cepstras given in advance. Like LPC, the best cepstral parameters (i.e., number of cepstral components) are experimented and respective results are compared (see Table 3). The classification accuracy based on LPCC based feature extraction is observed to significantly vary depending on the selected number of cepstras. If the number of cepstras are selected as 5 or 10, then the accuracy is around 96%-97% whereas if the number of cepstras is given as 3, then the accuracy drops to around 90%.

Table 4. Classification performance of MFCC method depending on different coefficient numbers

	Default Value	**In addition to 12 coefficients, including other coefficients**
1	0,960	0,980
2	0,940	0,970
3	0,980	0,980
4	0,930	0,980
5	0,970	0,980
Average	0,956	0,978
Number of Features	37	127

As the third option, MFCC is utilized to extract useful cepstral features from the voice signal. Initially, it is executed with the default parameters (12 coefficients). Moreover, in the second run, a new feature set is obtained by adding new features to the default features such as log-energy, 0. cepstral coefficient and delta-delta coefficients. Table 4 presents the results of these experiments.

Table 5. Classification accuracy with PLP and RASTA-PLP methods

	PLP	**RASTA-PLP**
1	0,600	0,496
2	0,620	0,518
3	0,578	0,490
4	0,589	0,486
5	0,680	0,663
Average	0,613	0,530
Number of features	81	81

Lastly, PLP and RASTA-PLP methods are utilized to analyze the voice signal and extract useful features from it to improve the classification accuracy. Table 5 presents the classification accuracy obtained by using these methods. As can be seen in Table 5, the features extract via PLP helped the classification algorithm output better results when compared with RASTA-PLP. Note that, both methods produce the same number of features from the dataset.

In this study, 5 different cepstral feature extraction methods are utilized and their outputs are given to SVM in order to measure their effect to classification accuracy. As a result, cepstral feature extractors (LPC, LPCC and MFCC) are observed to perform better than others (PLP and RASTA-PLP). The average classification performance with LPCC and MFCC is around 97% which is closely followed by the performance with LPC, with a result around 96%. Other methods, PLP and RASTA-PLP performed poorer than these methods. Therefore, the main conclusion of this paper is that cepstral features are observed to be better in retaining speaker specific features and characteristic than any other methods.

CONCLUSION

In this study, cepstral feature extraction methods are investigated as a requirement for a system that aims to perform automatic speaker recognition. The methods considered in this paper are LPC, LPCC, MFCC, PLP and RASTA-PLP. When results are analyzed, it is clear that the performance of LPC, LPCC and MFCC are highly promising as a feature extraction tool for the task of speaker recognition. Each of these methods relies upon some parameters to be tuned. The paper presents some of the most important parameters for each algorithm and which parameter results in best performance for each algorithm. For instance, LPC required us to choose the order of the analysis which affects the number of features produced by LPC but only slightly the performance of the algorithm. On the contrary to the order parameter of LPC, the number of cepstra parameter in LPCC requires careful selection of the parameter because it has more than a slight effect on the performance. Our study shows that the number of cepstras chosen appears to be directly proportional to the classification performance. That is, the more number of cepstras chosen, the more speaker specific characteristic retained by the feature selection algorithm. As for MFCC, the experiments show that its performance can be improved by adding some extra values to the feature set extracted by the algorithm. Lastly, any kind of parameter selection or addition does not help improve the performance of PLP and RASTA-PLP. Therefore, these algorithms are observed to be the weakest speaker related feature extraction algorithms when compared to the others. As a result, it is observed that LPCC and MFCC outperforms other methods in the task of automatic speaker recognition. This fact is empirically supported by the experiments conducted in the study. With these methods, the classification accuracy can be as high as 97% which appeared to be satisfactory for several application areas.

REFERENCES

[1] "Principles of Data Acquisition and Conversion". Texas Instruments. April 2015. http://www.ti.com/lit/an/sbaa051a/sbaa051a.pdf (08.06.2017)

[2] Ambikairajah, E. (2007, December). Emerging features for speaker recognition. In Information, Communications & Signal Processing, 2007 6th International Conference on (pp. 1-7). IEEE.

[3] Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology, 3(12), 18006-18016.

[4] Makhoul, J. (1975). Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4), 561-580.

[5] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

[6] Yusnita, M. A., Paulraj, M. P., Yaacob, S., Fadzilah, M. N., & Shahriman, A. B. (2013). Acoustic analysis of formants across genders and ethnical accents in Malaysian English using ANOVA. Procedia Engineering, 64, 385-394.

[7] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America, 87(4), 1738-1752.

[8] Dumitru, C. O., & Gavat, I. (2006, June). A comparative study of feature extraction methods applied to continuous speech recognition in Romanian Language. In Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006 focused on (pp. 115-118). IEEE.

[9] O'Shaughnessy, D. (2003). Interacting with computers by voice: automatic speech recognition and synthesis. Proceedings of the IEEE, 91(9), 1272-1305.

[10] Maheswari, N. U., Kabilan, A. P., & Venkatesh, R. (2010). A hybrid model of neural network approach for speaker independent word recognition. International Journal of Computer Theory and Engineering, 2(6), 912.

[11] Dhonde, S. B., & Jagade, S. M. (2015). Feature extraction techniques in speaker recognition: A review. International Journal on Recent Technologies in Mechanical and Electrical Engineering (IJRMEE), 2(5), 104-106.

[12] Salomons, E. L., & Havinga, P. J. (2015). A survey on the feasibility of sound classification on wireless sensor nodes. Sensors, 15(4), 7462-7498.

[13] Chowdhury, M. H. (2014). Speech based gender identification using empirical mode decomposition (EMD) (Doctoral dissertation, BRAC University).

[14] Saksamudre, S. K., & Deshmukh, R. R. (2015). Comparative study of isolated word recognition system for Hindi language. International Journal of Engineering Research and Technology, 4(07).

[15] Kumar, J., Prabhakar, O. P., & Sahu, N. K. (2014). Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A Review. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2760-2269.

[16] Zouhir, Y., & Ouni, K. (2014). A bio-inspired feature extraction for robust speech recognition. SpringerPlus, 3(1), 651.

[17] Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. IEEE transactions on speech and audio processing, 2(4), 578-589.

[18] Kwon, O. W., Chan, K., & Lee, T. W. (2003). Speech feature analysis using variational Bayesian PCA. IEEE Signal Processing Letters, 10(5), 137-140.

[19] Soman, K.P., Loganathan, R. and Ajay, V. (2011). Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., 486 s.

[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[21] Li, S., Li, H., Li, M., Shyr, Y., Xie, L., & Li, Y. (2009). Improved prediction of lysine acetylation by support vector machines. Protein and peptide letters, 16(8), 977-983.

[22] Yildiz, M., Bergil, E., & Oral, C. (2017). Comparison of different classification methods for the preictal stage detection in EEG signals. Biomedical Research, 28(2).

[23] AYHAN, S., & ERDOĞMUŞ, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 9(1).