Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 12(1): 41-57, 2022
Journal of the Institute of Science and Technology, 12(1): 41-57, 2022
ISSN: 2146-0574, eISSN: 2536-4618
Bilgisayar Mühendisliği / *Computer Engineering* **DOI: 10.21597/jist.957879**
Araştırma Makalesi / *Research Article*
Geliş tarihi / Received: 17.09.2021 **Kabul tarihi / Accepted: 18.11.2021**

To Cite: Onder, M and Bayat, H.I., 2022. Classification vowel-consonant letters with deep neural networks in Turkish and text-voice synchronızatıon on a basis syllable sıze. Journal of The Institute of Science and Technology (JIST), 12(1): 41-57.

Classification Vowel-Consonant Letters with Deep Neural Networks in Turkish and Text-Voice Synchronization on a Basis Syllable Size

Mürsel ÖNDER, Halil İbrahim BAYAT

ABSTRACT: In the study, a syllable-scale synchronization study was carried out by considering the grammatical structure of Turkish to emphasize simultaneously the sound and the text. Therefore, it was aimed to classify the vowels and consonants in Turkish within the word. For this purpose, two different Artificial Neural Network (ANN) models were preferred for this classification, and also the Mel-Frequency Cepstrum Coefficients method was preferred for extracting features of voice data. It has been observed that ANNs give the best results with deep learning. Tests were made with different numbers of coefficients in feature extraction. In the first stage of this study, a certain number of recordings were taken from the vowels and consonants in Turkish. Then, their feature was extracted and prepared for the training of networks. The best network structure and parameters were selected as a result of training and test made with different parameters. In this training, networks were asked to distinguish vowels from consonants. Afterward, the vowel-consonant distinction was made among 10 predetermined vectors of words and phrases. Layer-recurrent Neural Network and Pattern Recognition Network achieved an average success of 97.43% and 98.04%, respectively, in deep learning training carried out through the Mathworks Matlab software. Because Pattern Recognition Network achieved 98.82% success in recognizing vowels and 97.27% in recognizing consonants, this network model was preferred in vowel-consonant classification. After the classification process, timing files were created by determining the transition times of the vowels in the word. In the last step, an interface was created on the C# .NET platform for the synchronization process, and a syllabic algorithm was developed in this interface to emphasize the syllable synchronization of the text. Thus, the desired high precision was achieved in the simultaneous highlighting of the words.

Keywords: Artificial Neural Networks, Deep Learning, Mel-Frequency Cepstrum Coefficients, Sound-Text Synchronization

[1]Mürsel ÖNDER (**Orcid ID:** 0000-0003-4475-3955), Halil İbrahim BAYAT (**Orcid ID:** 0000-0002-3014-7113) Gaziosmanpasa University, Department of Mechatronics Engineering, Tokat, Turkey

*Sorumlu Yazar/Corresponding Author: Halil İbrahim BAYAT, e-mail: hibrahimbyt@gmail.com

The subject of this article is taken from Halil İbrahim Bayat's master's thesis. Oral presentation was made by Halil İbrahim BAYAT in conference of Mas 14th International European Conference On Mathematics, Engineerıng, Natural&Medical Sciences.(2021)

INTRODUCTION

Thanks to the breakthroughs in digital technology, the rapidly developing visual media (Internet, TV, cinema) has an important place in human life. Undoubtedly, one of the most important features of this visual media is the subtitles of these images. Subtitles are important both for language differences and for people with hearing impairments to understand the visual in question. In this study, based on this motivation, a study was carried out to simultaneously highlight the voice and the text belonging to that voice data in the Turkish language. The path followed in this study is quite different and new from the studies done so far. However, some literature studies may be related to the subject of this study:

In the literature, studies on ANNs and voice recognition differ from each other in terms of both the different ANN models and the way the voice data process. Also, there are cases of being independent or dependent on the speaker and independent or dependent from the text. Studies that we think are closely related to our study are as follows: In 1990, Cosi and his colleagues implemented vowel classification for English. They created this study independently of the speaker (Cosi et all, 1990). In these studies, multi-layer ANN was used and they achieved 95% success. In 1994, a study was carried out by Parlaktuna et al. On recognizing vowels and consonants in Turkish within themselves. The study here is independent of the speaker. In the study, vowel-consonant recognition, vowel recognition, and consonant recognition are discussed in three groups. In recognizing vowels and consonants, 84.7%, 91.1%, respectively; an average of 80.1% in recognizing vowels within itself; 50.7% success was achieved in consonants (Parlaktuna et al, 1994). In his study conducted in 1997, Üstün achieved 97.5% success in recognizing vowels in Turkish by using multi-layered ANN (Ustun, 1997). The system is a speaker-dependent system. Yavuz and his friend conducted a study on recognizing vowels in Turkish with the Probability Neural Network in 2010 and achieved 95% success. His work is a speaker-independent system (Yavuz and Topuz, 2010).

In our study; First, the study of distinguishing Turkish vowels from consonants was carried out. The aim here is to find the times of the vowels in the word. Finding the time of the vowel is important as it will give us the time of the syllable in that word. Here, while developing this method, the linguistic structure of Turkish has been taken into consideration. Artificial Neural Networks (ANN) are preferred for this classification. Two different ANN structures have been tested in the study. Appropriate ANN and its parameters were obtained as a result of the tests performed. Before starting the training, Mel-Frequency Cepstrum (MFC) method was applied to the characteristics of 10 different word groups together with vowels and consonants. In the next step, the hyphenation algorithm was developed to show that the word groups are emphasized simultaneously in syllable size. At the last stage, the study was concluded by emphasizing the sound and the text belonging to this sound in the developed interface simultaneously in terms of syllables. There is currently no study conducted following such a path (Bayat, 2020).

The main purpose of this study is to show that such a method can be used to emphasize or synthesize voice-text. Because it is thought that such a study will be beneficial for learning Turkish or for hearing impaired people to follow and comprehend the text of the voice (Yalçın, 2006).

MATERIALS AND METHODS

In our study, two different ANN models that can be classified as a classifier as dynamic and static were selected: In the ANN training study carried out with the Mathworks Matlab program; Pattern Recognition Network (*static*), one of which is a feed-forward ANN model (Figure 2), and the other with

Layer Recursive ANN model (*dynamic*), which is a different version of Elman Networks (Gupta and Homma, 2004), (Figure 1). MFC method was used to extract the properties of the audio data.

Artificial Neural Networks

The artificial neural network is a mesh recognition method inspired by the way the human brain processes information. To make a general definition, ANN: Large-scale interconnected networks of simple (usually adaptable) elements and their hierarchical organization aim to interact with real-life objects as the biological nervous system does (Kohonen, 1987; Haykin, 1999). In recent years, ANNs have been showing successful results in establishing relationships between these patterns without recognizing patterns. There are several studies on this subject in the literature

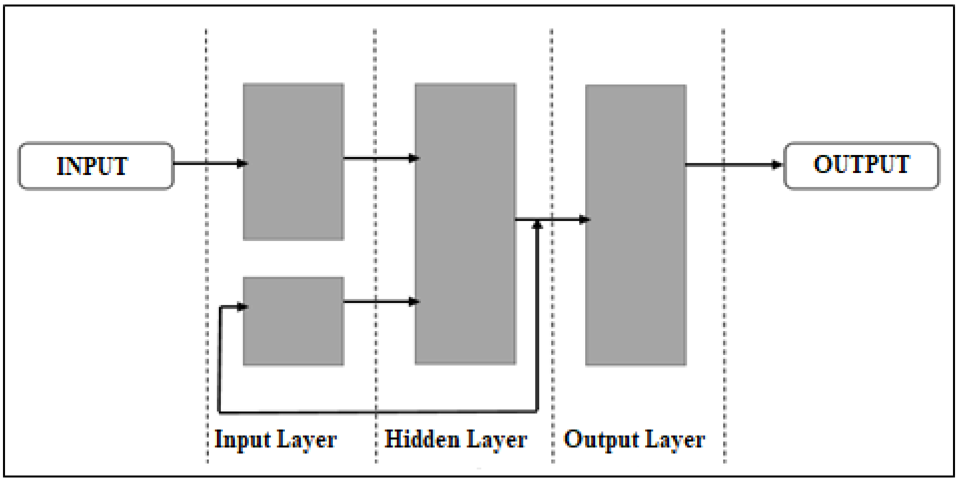

Figure 1. Schematic representation of Elman Networks (Bayat, 2020)

ANNs generally consist of the input layer from which input data is received, one hidden layer (the number may increase), and the output layer. Structures of ANNs; It may vary according to learning algorithms and data traffic between layers. For example, while there is a return from the hidden layer output to the input layer in Elman Networks in Figure 1 (Elman, 1990), this is not the case in the multi-layer feed-forward ANN model given in Figure 2. In this study, the performances of these two different ANN models were compared.

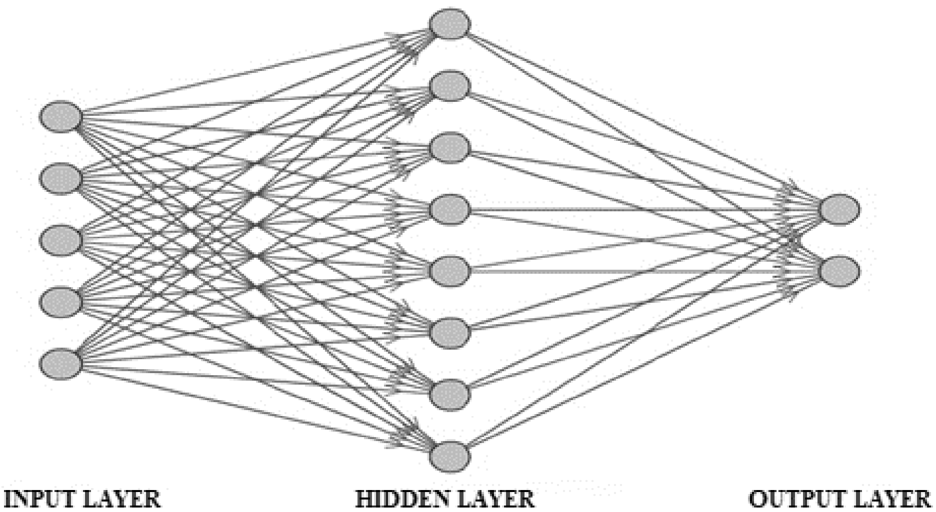

Figure 2. Schematic representation of Feed Forward ANN (Bayat, 2020)

Deep Neural Networks

Deep neural networks (DNN) are referred to in the literature as ANN consisting of two or more (usually more than two) hidden layers (Hinton et al, 2006; Bengio, 2012). ANN is depicted as a single hidden layer structure in general terms. However, in complex problems where training data is insufficient and inputs do not contain enough features, this single layer may not perform adequately. The purpose of using multiple layers is to find high-level abstract features from data with low-level features defined (Bengio, 2012). These highly abstract features help to distinguish independent distributions in training data (Bayat, 2020).

Shallow network structures, that is, ANN structures with one or two hidden layers, usually require a large number of neurons to represent their inputs well. As the number of neurons increases, the number of network parameters such as weights and biases will naturally increase, creating a heavy processing burden (Bayat, 2020). This causes large-size operations with many variables to not be represented effectively with shallow network structures (Bengio, 2012).

Another factor that encourages the use of deep learning structure comes from the work of the human brain. While the nerve signal transmits visual information within the body, some measurements have been made to find the distance and time it travels (Bayat, 2020). The results of these measurements showed that even in a simple object recognition process, the number of layers of biological neurons involved in this process is approximately ten (Cakir, 2014).

Feature Extraction

Sound processing is the process of digitizing sound signals and processing them in a computer environment with numerical methods. This process starts with recording the sounds and transferring them to the computer environment. The operations performed in this stage are the efforts to express the sounds numerically in the best way (Bayat, 2020).

Voice recognition is a fundamental pattern recognition problem that has been very popular and wide-ranging in the last half-century. Sound files have a continuous sinusoidal wave structure and new methods are being developed to express certain characteristics within this structure. The common purpose of all these methods is to separate the data in the audio file from the other data in itself and to reveal their unique features. For this, it is very important to extract the feature of the sound file in speech recognition systems (Tiwari, 2010). One of the most powerful methods of extracting the properties of sound data is the MFCC (Mel-Frequency Cepstrum Coefficients) method (Dave, 2013; Bayat, 2020).

Mel frequency kepstrum method; It is a very popular and successful feature extraction method used in speech recognition systems (Meng et al, 2004). This method has been created by modeling the hearing process of the human ear. Studies on the human hearing process have shown that the human ear has high resolution and saturation against low-frequency sounds when compared to high-frequency sounds (Dave, 2013).

MFC coefficients are obtained by going through certain stages. There are many studies conducted with different coefficients in the literature. These coefficients are usually chosen as a result of certain tests or taking into account the characteristics of the audio data. In Figure 3, the steps of obtaining the MFC coefficients are given in a flow.

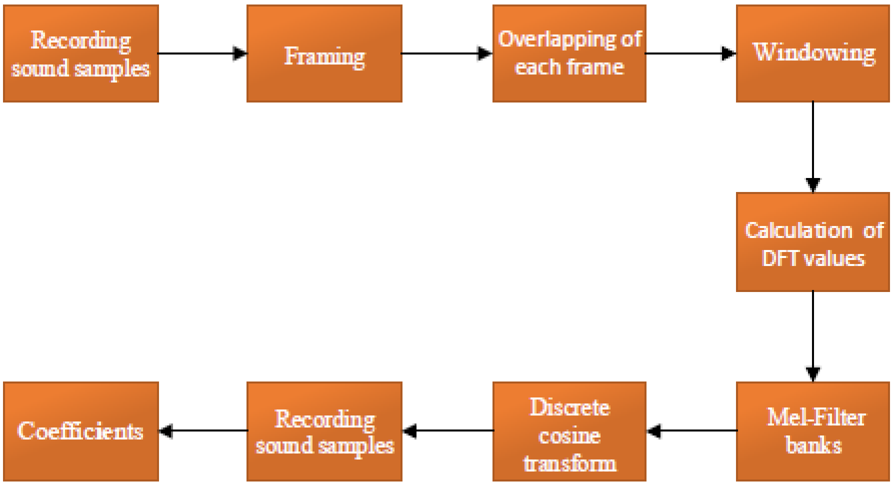

Figure 3. Flow diagram showing the process of obtaining the MFC coefficients (Bayat, 2020).

Materials and Working Framework

The study aims to distinguish the vowels from consonants from 29 letters; 26 different records were taken from each of the vowels and 30 different from the vowels. Records; It has 16-bit PCM and a sampling frequency of 11.025 kHz. In total, 240 records from vowels and 540 records from consonants were obtained. Sample lengths and time intervals of recordings from Turkish vowels and consonants are given in Table 1.

Table 1. Sample and time length intervals of sound recordings taken from vowels and consonants

Letters	Sample count range	Length(sec.)	Letters	Sample count range	Length(sec.)
A	2603-2946	0.236-0.267	M	716-3581	0.064-0.324
B	474-544	0.043-0.049	N	876-2503	0.079-0.227
C	494-1166	0.044-0.105	O	2353-3214	0.213-0.291
Ç	1243-2555	0.112-0.231	Ö	2351-2983	0.213-0.271
D	513-601	0.046-0.054	P	442-1968	0.040-0.178
E	2351-3151	0.213-0.285	R	1217-2570	0.110-0.233
F	2051-3209	0.186-0.291	S	1239-2800	0.112-0.254
G	472-858	0.043-0.077	Ş	1124-2652	0.101-0.241
Ğ	1084-2530	0.098-0.229	T	803-1476	0.072-0.133
H	574-2089	0.052-0.189	U	1985-3161	0.177-0.281
I	1995-2988	0.181-0.271	Ü	2195-3103	0.199-0.281
İ	2411-3360	0.218-0.304	V	1135-2565	0.103-0.232
J	1175-2564	0.106-0.232	Y	1046-2122	0.094-0.192
K	794-1622	0.072-0.147	Z	1355-2283	0.123-0.207
L	1007-1787	0.091-0.162			

Approximately 39% of the records were reserved for testing, while the rest were used in training. The records of the letters were taken from a single person and the system was designed as speaker-dependent. The records of the letters were obtained by extracting the letters from syllables so that their characteristics could be better understood and the letters were better picked up in the tests within the word. For example; For the example of the letter "s", the vowels were first and then followed by the vowels (as, es, ıs, is, os, ös, us, üs; sa, se, si, sı, so, so, sü, su) However, it has been separated from these records by removing the letter "s" in such a way that its characteristic is intact.).

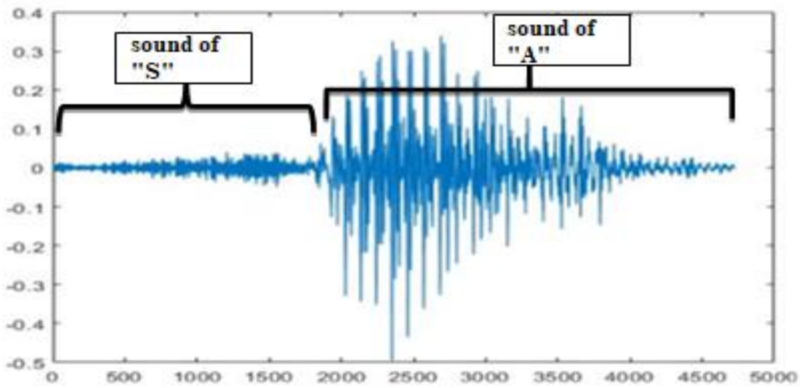

Figure 4. Graph of “Sa” sound recording

The records were then collected under a single matrix. Each column of the matrix was created to belong to one record of only one letter. The data were first passed through the Hamming windowing function by arranging each with floating frames and with a certain overlap ratio, and then the MFC coefficients were found and rearranged for network training in the same matrix format. The parameters used in the operations are shown in Table 2.

Table 2. Parameters used to extract the properties of audio data

Parameters	**Selectıons**
Feature Extraction Method	MFCC
Number of Coefficients	17 (In the first stage)
Windowing Function	Hamming
Length of Windowing Function	Up to Frame Size (440)
FFT Degree	512
Overlapping Ratio of Successive Frames	%75
Specified frame size (according to the tests)	440 (0.0399 sec.)

In the study, target matrices were created according to two classifications. Figure 5 shows the representation of the target matrix. The aim here is to distinguish vowels from consonants, as we mentioned earlier.

$$\begin{array}{c}\text{TARGET}\\ \text{MATRIX}\end{array} = \left\{\begin{array}{l} \overbrace{11111111111 \ldots 111}^{\text{VOWELS}}0000000000 \ldots 000 \\ 00000000000 \ldots 000\underbrace{1111111111 \ldots 111}_{\text{CONSONANT}} \end{array}\right\}$$

Figure 5. Target Matrix

Interface program and spelling algorithm

At the last stage of the study, the times of the vowels in the word were determined with the appropriate ANN and recorded in text files. These files were read with the interface program developed for simultaneous highlighting of the voice-text. The operation of the interface program is as follows:

1. Read the text file, write the word text on the screen.
2. Break the text of the word into syllables.
3. Select the audio file for the word.
4. Highlight the word and sound file simultaneously.

Figure 6. Voice-Text synchronizer interface.

The voice-text synchronizer interface program is given in Figure 6. In this program, after clicking the "Open text file" phrase and selecting the file, the text of the word in question is divided into syllables on the screen. Here, the hyphenation algorithm shown in Figure 7 has been developed to do this.

Considering the words, we use in our daily lives and the voiced text in this study, the spelling performance of up to four syllables was found to be sufficient.

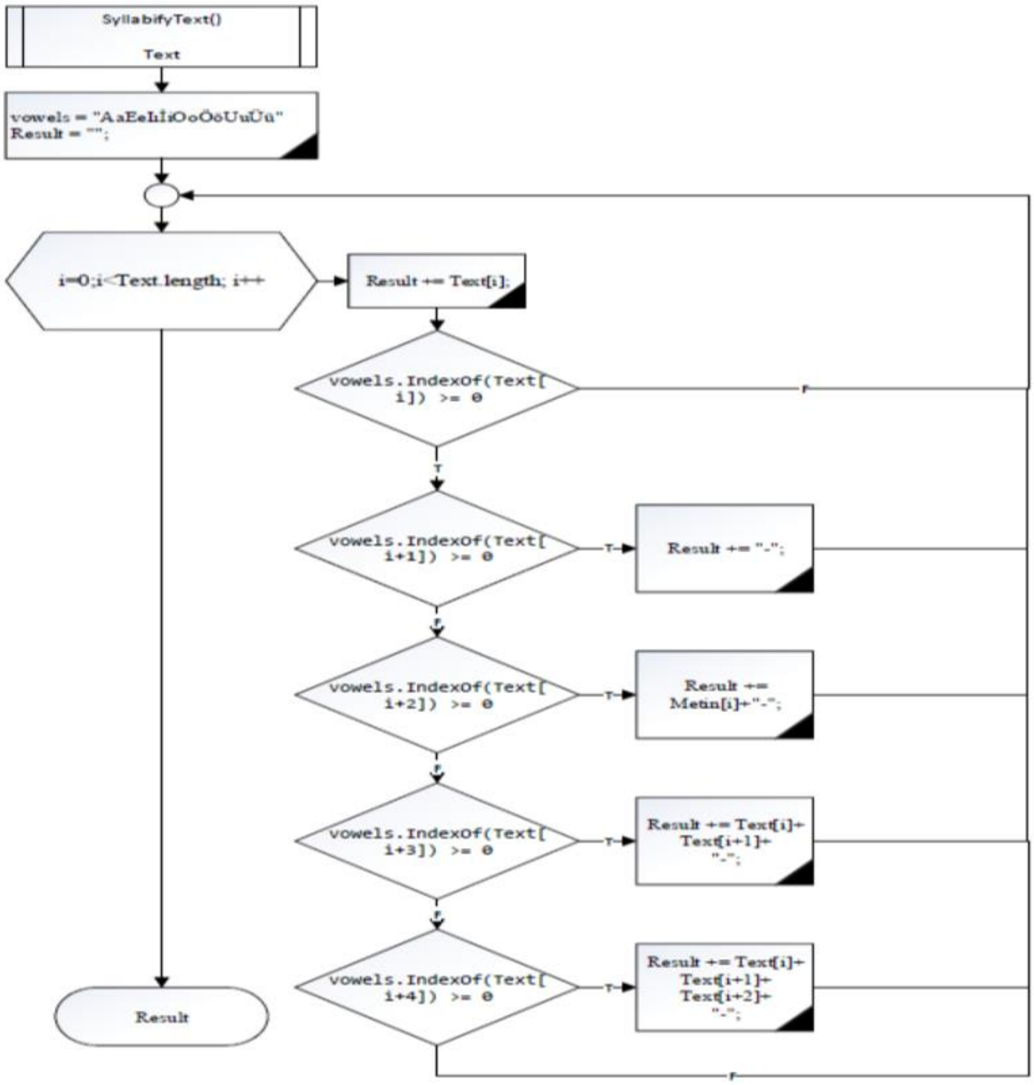

Figure 7. Hyphenation algorithm developed on C # .Net

RESULTS AND DISCUSSION

Selection of Network Parameters

The first stage of the study is to conduct lots of training and tests on two different network structures. The aim here is to roughly reveal the parameters in which the two network structures give the best performance. Network structures for one, two, and three hidden layers have been tested with different numbers of neurons and their performance has been tested by tests. Transfer, training, and performance functions for two ANN models are as shown in Table 3.

Table 3. Transfer, training and error functions used in ANN trainings

ANN models	Fırst hidden layer transfer function	Second hidden layer transfer function	Third hidden layer transfer function	Output layer transfer function	Training function	Performance function
Layer-recurrent neural network	Tansig	Tansig	Tansig	Purelin	Trainlm	“cross-entropy”
Pattern recognition network	Tansig	Tansig	Tansig	Softmax	Trainscg	“mean squared error"

As a result of the tests performed with Layer-Recursive and Pattern Recognition Networks with different layers and neuron numbers, the error, sensitivity, accuracy, and special factor are shown in Table 4. Networks have been trained only once. The network structure with the highest network performance was determined in both ANN structures and gradual tests were performed with different MFC coefficients and the performances of the two network structures were compared.

Table 4. Layer-Recursive and Pattern Recognition Neural Networks test results

Layer-recurrent neural network	Neurons	Training perform. (mean-square error)	Correct rate	Error rate	Sensitivity	Specificity
One hidden layer	10	0.01484	0.9502	0.0498	0.9491	0.9514
	15	0.0091246	0.9388	0.0612	0.9173	0.9624
	20	0.0095811	0.9592	0.0408	0.9560	0.9624
	25	0.0071028	0.9670	0.0330	0.9623	0.9718
	30	0.0065754	0.9435	0.0565	0.9424	0.9447
	35	0.0053399	0.9575	0.0425	0.9538	0.9612
Two hidden layers	10,5	3.22498e-5	0.9488	0.0512	0.9311	0.9680
	10,10	5.5396e-10	0.9628	0.0372	0.9528	0.9732
	15,5	3.6656e-11	0.9667	0.0333	0.9674	0.9660
	15,10	1.1226e-10	0.9603	0.0397	0.9419	0.9802
	20,5	1.2996e-13	0.9673	0.0327	0.9537	0.9816
	20,10	1.8323e-13	0.9595	0.0405	0.9505	0.9687
	20,15	1.1336e-11	0.9542	0.0458	0.9374	0.9722
	30,5	1.521e-11	0.9670	0.0330	0.9562	0.9783
	30,10	2.8939e-13	0.9637	0.0363	0.9559	0.9716
	30,15	8.7641e-14	0.9597	0.0403	0.9447	0.9758
	30,20	2.486e-10	0.9550	0.0450	0.9375	0.9739
Three hidden layers	10,5,5	1.6007e-14	0.9553	0.0447	0.9437	0.9674
	10,10,5	3.1416e-11	0.9586	0.0414	0.9446	0.9735
	10,10,10	2.9122e-13	0.9382	0.0618	0.9068	0.9746
	20,5,5	6.2002e-15	0.9623	0.0377	0.9543	0.9705
	20,10,5	9.2299e-12	0.9592	0.0408	0.9394	0.9807
	20,10,10	8.444e-12	0.9611	0.0389	0.9587	0.9636
	20,15,10	3.848e-12	0.9690	0.0310	0.9634	0.9746
	20,15,15	2.9957e-15	0.9572	0.0428	0.9430	0.9724
	20,20,10	6.0341e-13	0.9611	0.0389	0.9603	0.9620
	20,20,15	5.177e-14	0.9634	0.0366	0.9524	0.9749
	30,10,5	2.2712e-13	0.9670	0.0330	0.9522	0.9827
	30,10,10	4.6839e-12	0.9561	0.0439	0.9334	0.9812
	30,20,5	1.9102e-10	0.9533	0.0467	0.9344	0.9738
	30,20,10	1.8471e-11	0.9642	0.0358	0.9575	0.9711
	30,20,20	1.1853e-13	0.9665	0.0335	0.9537	0.9799
	30,30,10	1.0845e-10	0.9665	0.0335	0.9607	0.9723
	30,30,20	8.4561e-15	0.9544	0.0456	0.9355	0.9749

Table 4. Layer-Recursive and Pattern Recognition Neural Networks test results (continue)

Pattern recognition network (*trainscg*)	**Neurons**	**Training perform. (cross-entropy)**	**Correct rate**	**Error rate**	**Sensitivity**	**Specificity**
One hidden layer	10	9.3562e-07	0.9474	0.0526	0.9263	0.9707
	20	2.902e-07	0.9651	0.0349	0.9570	0.9733
	30	2.52e-07	0.9651	0.0349	0.9506	0.9804
	40	2.7563e-07	0.9709	0.0291	0.9590	0.9834
	50	2.7942e-07	0.9704	0.0296	0.9625	0.9784
	60	1.9065e-07	0.9692	0.0308	0.9624	0.9762
	70	3.04e-07	0.9645	0.0355	0.9481	0.9821
Two hidden layer	20,5	5.9388e-08	0.9631	0.0369	0.9450	0.9826
	20,10	8.6892e-08	0.9679	0.0321	0.9543	0.9822
	20,20	1.5134e-07	0.9665	0.0335	0.9537	0.9799
	30,10	8.5897e-08	0.9648	0.0352	0.9545	0.9755
	30,20	9.8748e-08	0.9723	0.0277	0.9596	0.9857
	40,10	1.0487e-07	0.9679	0.0321	0.9654	0.9703
	40,20	1.0262e-07	0.9651	0.0349	0.9501	0.9810
	40,30	7.774e-08	0.9690	0.0310	0.9619	0.9762
	50,10	1.4472e-07	0.9799	0.0201	0.9787	0.9810
	50,20	1.8098e-07	0.9726	0.0274	0.9637	0.9818
	50,30	6.0268e-08	0.9651	0.0349	0.9511	0.9799
	50,40	9.5682e-08	0.9723	0.0277	0.9704	0.9742
	50,50	4.6491e-08	0.9712	0.0288	0.9536	0.9901
Three hidden layers	20,10,10	2.4129e-08	0.9628	0.0372	0.9479	0.9787
	20,20,10	4.4833e-08	0.9704	0.0296	0.9540	0.9879
	30,10,10	2.9942e-08	0.9681	0.0319	0.9553	0.9816
	30,20,10	5.7483e-08	0.9676	0.0324	0.9557	0.9800
	30,20,20	3.6177e-08	0.9718	0.0282	0.9616	0.9823
	40,10,10	1.6496e-07	0.9589	0.0411	0.9515	0.9666
	40,20,10	4.0654e-08	0.9804	0.0196	0.9730	0.9881
	40,20,20	2.8066e-08	0.9737	0.0263	0.9710	0.9764
	40,30,20	3.4884e-08	0.9740	0.0260	0.9628	0.9857
	40,30,30	2.0738e-08	0.9712	0.0288	0.9595	0.9834
	50,10,10	3.0302e-08	0.9729	0.0271	0.9627	0.9835
	50,20,10	4.2468e-08	0.9754	0.0246	0.9695	0.9814
	50,30,10	2.2659e-07	0.9653	0.0347	0.9591	0.9717
	50,30,20	5.6888e-08	0.9785	0.0215	0.9646	0.9931
	50,30,30	4.5603e-08	0.9631	0.0369	0.9479	0.9792
	50,40,10	3.5215e-08	0.9734	0.0266	0.9653	0.9818
	50,40,20	3.7714e-08	0.9692	0.0308	0.9569	0.9822
	50,40,30	4.0281e-08	0.9679	0.0321	0.9644	0.9713
	50,40,40	3.8336e-08	0.9595	0.0405	0.9490	0.9703
	50,50,10	6.0878e-08	0.9720	0.0280	0.9637	0.9807
	50,50,40	3.0435e-08	0.9687	0.0313	0.9529	0.9856

In the tests performed for two different network structures, the network structures that give the highest performance were subjected to a separate test with different numbers of MFC coefficients and the results were observed. Test results are as shown in Table 5.

Table 5. Results of tests with different MFC coefficients

Layer-Recurrent Neural Network (trainlm)	Mfc coefficients	Performance function (mean-square error)	Correct rate	Error rate	Sensitivity	Specificity
Three hidden layers and neurons each of them (25-15-10)	10	2.8799e-12	0.9553	0.0447	0.9624	0.9484
	11	2.9343e-13	0.9561	0.0439	0.9467	0.9658
	12	8.1739e-13	0.9645	0.0355	0.9668	0.9623
	13	5.0523e-14	0.9463	0.0537	0.9554	0.9377
	14	9.9367e-14	0.9486	0.0514	0.9592	0.9384
	15	1.5955e-14	0.9555	0.0445	0.9532	0.9580
	16	1.1181e-14	0.9637	0.0363	0.9539	0.9738
	17	3.848e-12	0.9690	0.0310	0.9634	0.9746
	18	1.812e-09	0.9679	0.0321	0.9603	0.9756
	19	1.0547e-13	0.9743	0.0257	0.9684	0.9802
	20	1.471e-13	0.9595	0.0405	0.9461	0.9736
Pattern Recognition Neural Network (trainscg)	**Mfc coefficients**	**Performance function (cross-entropy)**	**Correct rate**	**Error rate**	**Sensitivity**	**Specificity**
Three hidden layers and neuron each of them (40-20-10)	10	3.2201e-08	0.9536	0.0464	0.9560	0.9512
	11	3.3926e-08	0.9600	0.0400	0.9536	0.9666
	12	4.2466e-08	0.9648	0.0352	0.9595	0.9701
	13	4.0485e-08	0.9637	0.0363	0.9662	0.9612
	14	7.1674e-08	0.9555	0.0445	0.9537	0.9574
	15	3.3127e-08	0.9614	0.0386	0.9650	0.9579
	16	3.0604e-08	0.9651	0.0349	0.9550	0.9755
	17	4.0654e-08	0.9804	0.0196	0.9730	0.9881
	18	3.8866e-08	0.9740	0.0260	0.9633	0.9852
	19	4.5827e-08	0.9732	0.0268	0.9602	0.9868
	20	4.1406e-06	0.9732	0.0268	0.9617	0.9851

As it can be understood from the table, Pattern Recognition Network gives the best performance with 17 MFC coefficients, while this is 19 in Layer-Recursive ANN. The details of the test results of the Pattern Recognition Network are shown in Table 6. Based on the accuracy values, the Pattern Recognition network was selected for vowel-unvoiced classification within word vectors.

Table 6. Statistical performance of the Pattern Recognition Network

Confusion Matrix	Vowels	Consonants	Rate	Total
Vowels	1763	21	%98.82	1784
Consonants	49	1747	%97.27	1796
Rate	%97.30 Sensitivity	%98.81 Specificity	%98.04	-
Total	1812	1768	-	3580

Vowel-Consonant Separation and Determination of Time of Vowel Letters

After choosing the best network structure, the vowel-silent classification study was started for the predetermined words. Word vectors passed through the same signal processing process were inserted into the selected network and subjected to binary classification.

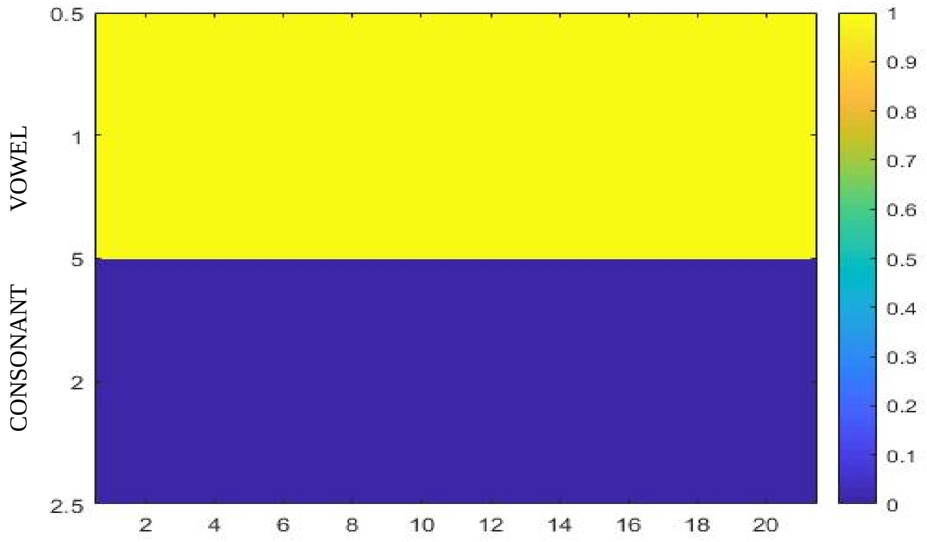

Figure 8. Visual printout of the result matrix of the letter "a" sent to the network.

Table 7. Result matrix for the letter "a"

Vowel	1	1	1	1	1	1
Consonan	1,08E-09	2,22E-10	2,58E-10	5,72E-10	4,69E-10	1,21E-09

In previous tests, network performance was checked by sending only vowel or consonant letters to networks (Figure 8). As seen in Table 8, the fact that it is shown with "1" in the result matrix shows that the vowel is recognized. In the vocabulary study, words were sent directly to the trained network and an output matrix was obtained according to the approximate values of "1" for vowels and "0" for consonants in the result matrix.

As can be seen in Figure 9 A and B, the results of the word vectors (vowel-consonants) sent to the network separately are evident. The visuals when the "Voice-Text Synchronizer" program developed with the separator works with these words are shown in Figures 10 and 11. Also, it has been observed that this clarity is slightly distorted when a single vector "hello world" audio file is sent to the trained network. But despite this, the time determination could still be made. Because the transition and starting times of vowels can be observed on the result matrix. Figure 12 shows how the starting time of vowels is calculated in the result matrices. In Equation 1, it is stated how it is calculated.

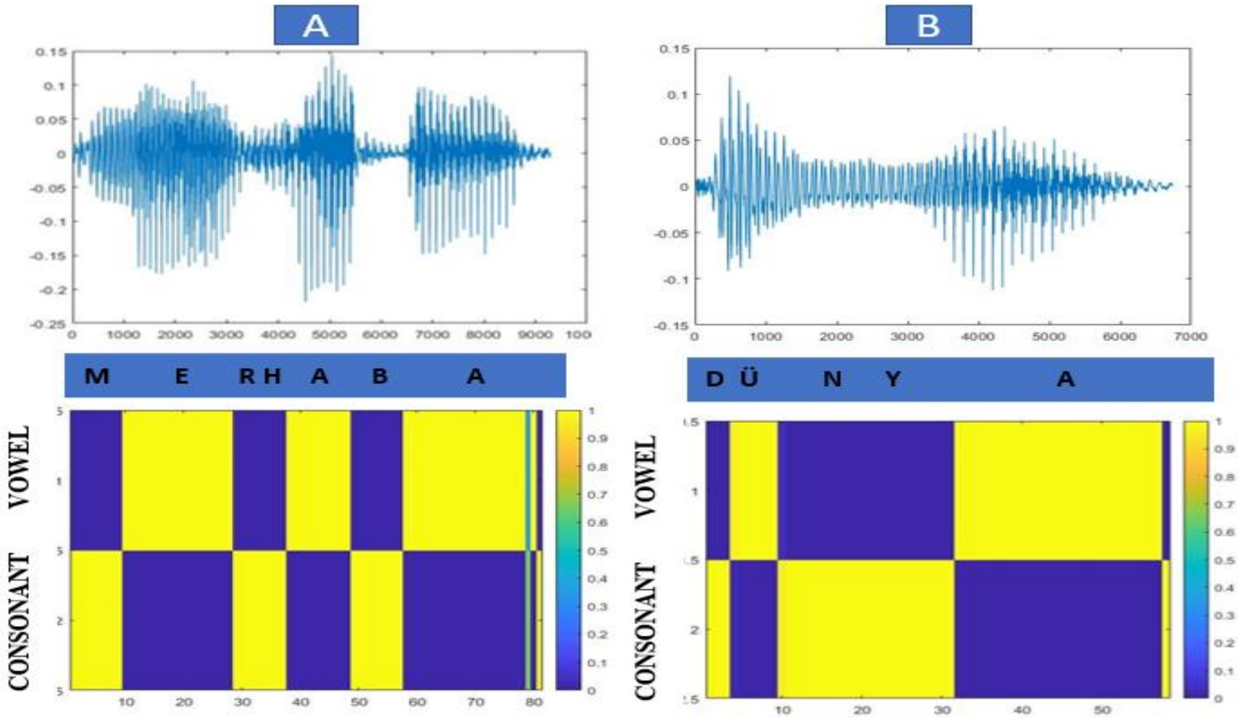

Figure 9. Audio files of the words "Merhaba" which means -hello, (A) and "Dünya”, means World in Turkish (B) and graphs of the network output results.

Figure 10. The screenshot of simultaneous emphasis of the sound-text of the word "Merhaba" with the interface program.

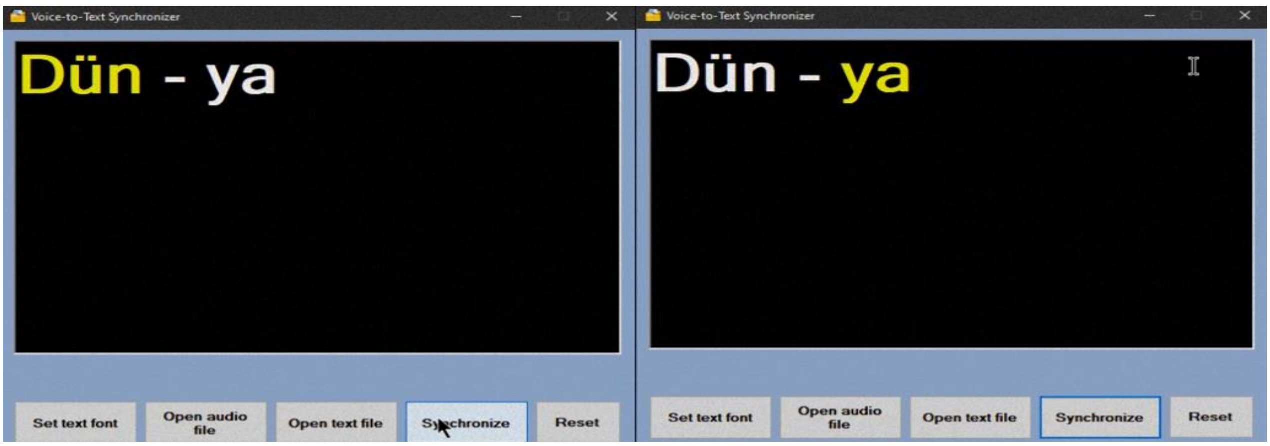

Figure 11. The screenshot of the simultaneous emphasis of the word "Dünya"

$$\overbrace{\frac{440}{11025}}^{\substack{length\ of \\ the\ first\ frame\ (sec.)}} + \overbrace{\left(\frac{440}{11025} - \frac{330}{11025}\right)}^{\substack{(Length\ of\ frame\ - \\ length\ of\ Overlapped\ frame\ (sec.)) \\ X\ (number\ of\ overlapped\ frames)}} \quad x9 \quad = \quad 0.1297\ sec. \tag{1}$$

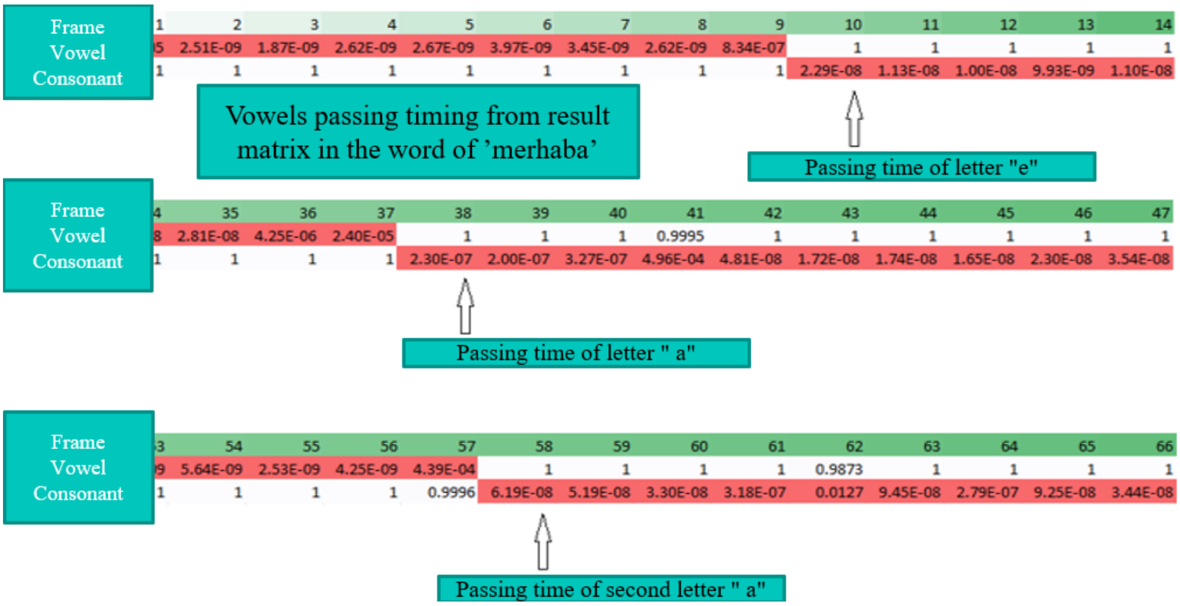

Figure 12. Transition times of vowels in the word "Merhaba".

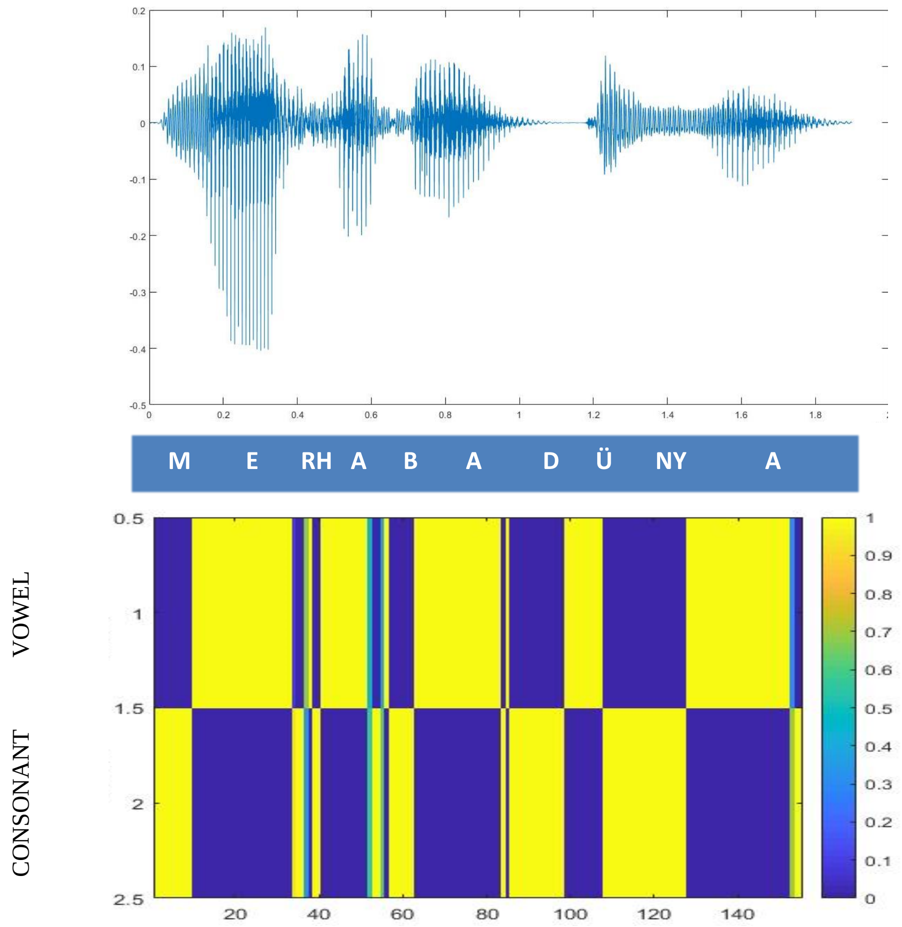

Figure 13. Graphical representation of audio and network outputs belonging to the word group "Merhaba Dünya"

As can be seen in Figure 13, the length of the audio file can be seen as a factor that makes the distinction within the word difficult. Because when the words are handled separately, this clarity is distorted in word groups while facing a clear graphic.

Other words and word groups used during the study are expressed in Table 8 The table includes the vowels in the word and the times of these vowels. The voice-to-text synchronization of all the words (and word groups) shown in the list has been successfully achieved.

Table 8. Words and word groups used in the study

Words and phrases	Number of the vowels	Tımes of voice letters (sec.)
Merhaba	3	0.139-0.419-0.618
Bilim	2	0.199-0.528
Dünya	2	0.099-0.329
Türkiye	3	0.126-0.529-0.752
Vatan	2	0.159-0.558
İstanbul	3	0.069-0.289-0.518
Günaydın	3	0.079-0.279-0.638
Millet	2	0.259-0.738
Cumhuriyet	4	0.089-0.628-0.937-1.167
Özgürlük	3	0.069-0.429-0.858
Ay yıldız	3	0.069-0.439-0.937
Merhaba dünya	5	0.139-0.448-0.678-1.027-1.326

CONCLUSION

As can be seen in studies with different ANNs, vowels and consonants were defined with an accuracy of 97-99% and gave successful results in tests. Although the tests for the distinction of vowel and consonants based on letters gave very clear and distinct results, the tests conducted with word vectors and especially with word groups remained far from this clarity. One of the biggest reasons for this is that when the network is tested with words, the system falls into the scope of an independent study from the text, that is, the network trained with letters is required to recognize letters with different harmonies within the words (Dede, 2008; Bayat, 2020). Another important reason is that the harmony of the vowels in the words in Turkish is closely related to the difference from word to word and in each vocalization style of the speaker (Kılıc, 2015). For these reasons, it makes it difficult to detect vowels or consonants from words. Vowels were successfully detected in 12 vectors containing words and word groups, and voice-text synchronization was achieved. Since the aim of this study was to show that capturing vowels and syllabic-scale word-sound emphasis can be done simultaneously, no study was conducted with a large word group (Bayat, 2020).

While carrying out all these studies, some ideas were gained about increasing the performance of the system. One of the most important issues in voice recognition systems is the issue of keeping voice recordings and training data as wide and rich as possible (Sirigos et al, 1996). For this reason, the richness of training data and the quality of voice recordings significantly affect network performance (Kılıc, 2015). More advanced ANN models to be used with a discriminator and adaptations of different classifier combinations can be a guide in improving system performance (Vafeiadis et al, 2017). The recently developed hybrid systems suggest that they will pave the way for new solutions and successful results in artificial intelligence problem solving (Wang et al, 2006).

The use of different classifier and sound features can make the system give higher results (Gupta, 2004; Vafeiadis et al, 2017). Operating such a system with high efficiency will pave the way for more efficient creation of phoneme-based speech recognition, phoneme-text or text-sound systems in the future. Such a system contains content that can be utilized in many areas from education to communication. To summarize, the way to synchronize the text of any voice or visual made manually today will be made possible automatically and with high precision (Bayat, 2020).

ACKNOWLEDGEMENTS

Endless thanks to my mother, Birten BAYAT, who has always supported me throughout my education life.

Conflict of Interest

The article authors declare that there is no conflict of interest between them.

Author's Contributions

The authors declare that they have contributed equally to the article.

REFERENCES

Bayat, H.İ., 2020. Identification of vowel-non vowel letter with artificial neural network and sound-text synchronization at syllable level (Master thesis), Gaziosmanpaşa University, Institute of science and technology, Tokat, Turkey.

Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning. In Proceedings of ICML workshop on unsupervised and transfer learning, pp. 17-36.

Cakir, E., 2014. Multilabel sound event classification with neural networks. (Master thesis), Tampere University of Technology, Faculty of Computing and Electrical Engineering, Finland.

Çakır, M.Y., 2017. Real-time high-quality voice recognition. (Master thesis), İstanbul Sabahattin Zaim University, Institute of science and technology, İstanbul, Turkey

Cosi, P., Bengua, Y. and De Maria, R., 1990. Phonetically-based multi-layered neural networks for vowel classification. Speech Communication, 1(9), pp. 15-19.

Dave, N., 2013. Feature extraction methods LPC, PLP and MFCC in speech recognition. Internatıonal journal for advance research ın engıneerıng and technology, 4(1), 5 pp.

Dede, G., 2008. Speech recognition with artificial neural networks (Master thesis), Ankara University, Institute of science and technology, Ankara, Turkey.

Elman, L. J., 1990. Finding structure in time. Cognitive Science, 2(14), pp. 179-211.

Güloğlu, T., 2014. Speech recognition for Turkish phonology using wavelet techniques.(Master thesis), Dokuz Eylül University, Graduate School of Natural and Applied Sciences, İzmir.

Gupta, M., Jin, L. and Homma, N., 2004. Static and dynamic neural networks: from fundamentals to advanced theory. John Wiley & Sons.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, 2nd Edition, Prentice-Hall, pp. 823, Ontario, Canada.

Hinton, G., Osindero, S. ve Teh, Y. W., 2006. A fast-learning algorithm for deep belief nets.Neural computation, 18(7), pp. 1527-1554.

Kılıç, E., 2015. The effects of Turkish vowel harmony in word recognition. (Master thesis), DePaul University, The Department of Psychology Collage of Science and Health, Chicago,Illinois, USA.

Kohonen, T., 1987. State of the art in neural computing. In Proceedings, IEEE First International Conference on Neural Networks, pp. 179-190, San Diego, USA

McCulloch, W. S. and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), pp. 115-133.

Meng, Y., Lee, T., Ching, P.C. and Zhu, Y. 2004. Speech recognition on DSP: issue on computational efficiency and performance analysis. Microprocessors and Microsystems, 30(3), pp. 155-164.

Önder, M., In Printing. Elmas-Hece Engineering in Quran Education

Parlaktuna, O., Cakici, T., Tora H. and Barkana, A., 1994. Vowel and consonant recognition in Turkish using neural networks toward continuous speech recognition. Mediterranean Electrotechnical Conference, Antalya, Turkey.

Sirigos, J., V. Darsinos, N. Fakotakis and G. Kokkinakis, 1996. Vowel-non vowel decision using neural networks and rules. Proceedings of Third International Conference on Electronics, Circuits, and Systems, Rodos, Greece.

Tiwari, V., 2010. MFCC and its application in speaker recognition. International Journal on Emerging Technologies, 1(1), pp. 19-22.

Üstün, S.V., 1997. Recognition of vowels in Turkish using artificial neural networks. (Master thesis), Yıldız Technical University, Institute of science and technology, İstanbul, Turkey

Vafeiadis, A., Kalatzis, D., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L. And Hamzaoui, R., 2017, November. Acoustic scene classification: From a hybrid classifier to deep learning.

Wang, J.C., Wang, J.F., He, K.W. and Hsu, C.S., 2006, July. Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor. In The 2006 IEEE International Joint Conference on Neural Network Proceedings, pp. 1731-1735

Yalçın, N., 2006. Developing a software for teaching initial reading writing to class student of primary education using speech recognition technology. (Doctoral thesis), Institute of science and technology, Ankara, Turkey.

Yavuz, E. and Topuz, V., 2010. Recognation of Turkish vowels by probabilistic neural network using Yule-Walker AR method. International Conference on Hybrid Artificial Intelligence Systems, Berlin, Heidelberg, Germany.