

Eskisehir Technical University

Faculty of Engineering

Department of Electrical & Electronics Engineering

**SPEECH RECOGNITION USING LPC**

Ahmet Yiğit BİRCAN-29023668578

Ceren Mine ESEN-51610244520

Koray GÜNAL-25366549246

Zeki ÇOLAK-24436564622

**EEM301-SIGNALS AND SYSTEMS**

**OCAK 2021**

**ABSTRACT**

Humans have been interacting and communicating with each other throughout history. There are many types of communication used in the world. People have managed to communicate with some form of encryption, using the vocal cords as a medium. This encryption system is called voice communication. In time, it has become a necessity to ensure the active use of voice communication not only from person to person but also in electronic devices. Voice recognition systems are being developed for this.

In this study, it was aimed to write a few words voiced as text by investigating the general working principle of speech recognition systems. The voice recognition process takes place through certain stages. First of all, sample sound recordings must be added in order for the desired sound to be perceived by the system at the entrance, and the first step is performed by determining the expression here. In the second step, the necessary data carrying information is obtained by processing the recorded sounds and desired input. Finally, the desired sound at the input is compared and matched with the recorded sound files and the action corresponding to the detected expression is performed. There are many methods that can be used in each of these stages.

This speech recognition system developed as a word-based command-control system independent of the person; It is designed to perform 4 different commands by restricting the word capacity to 4 words. It is aimed to determine the sound correctly by recording 10 different sounds for each word in the system. Gaussian distribution is made for LPC coefficients of each sound recording. In this way, the possibility of giving wrong commands is minimized.

**INTRODUCTION**

From past to present, people have developed many methods to communicate. Today, communication is mostly done by technology. Written, audio or visual communication can be made with electronic devices such as phones and computers. Voice communication is especially preferred because it is easy and natural.

While people communicate among themselves with electronic devices, speech recognition systems are used to communicate easily and naturally with these devices. The purpose of speech recognition systems is to provide the voice communication that people commonly prefer between human and computer. Speech recognition systems provide fast and remote data entry to the computer and eliminate the need for additional devices such as mouse and keyboard. In this way, besides reducing the cost, it also provides convenience to individuals who cannot use additional devices.

As a result, speech recognition systems that come together with communication and technology that are frequently used in life; It takes an active role in daily life as it is fast, easy and natural.

Recognized Word

Input Signal

recording sound

comparison of sound

templates

pairing

sound processing

Speech Recognition System

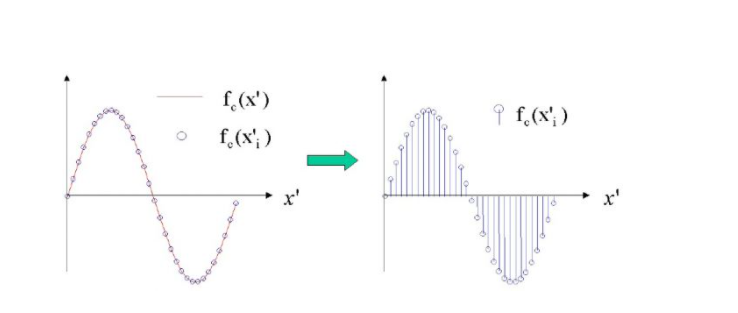
*Figure 1*

**TECHNIQUES USED IN SPEECH RECOGNITION PROCESS**

Speech recognition discipline in the voice recognition field is a system that tries to gain an important place in the developing technology process and it is the process of recognizing human voice by detecting it by a computer through a microphone. several different techniques are used during these processes. In this section; These techniques are explained by including the techniques used for each stage of the speech recognition process.

RECORDING VOICE AND DEFINING EXPRESSION: At this stage, the sound is recorded first.Sound has been recorded then it will go through various processes and be processed In order for these operations to be performed digitization of the sound is carried out.

DIGITIZATION OF SOUND: First, the function of digitizing the sound is realized. In order to use digital signal processing techniques, the analog signal, that is speech, must be represented in the form of serial numbers.

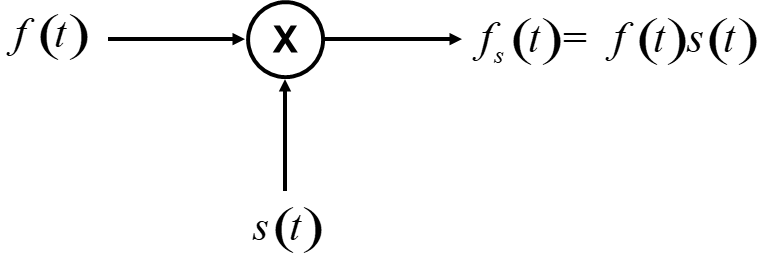


*Figure 2*

As can be seen in the figure2, the sound first transforms into an analog signal and then into a digitized signal.

SAMPLING: Before sampling, the sound digitizing function is performed first. In order to use digital signal processing techniques, the analog signal, ie speech, must be represented in the form of serial numbers. This is done by sampling the analog signal. In order to preserve the characteristics of the sound, sampling is performed at least twice the frequency of the highest frequency included in the audio signal to be converted. This; It is expressed as the sampling theorem. We also need to remove and clear the silence parts of the sound that we are recording so that we can see the sound signal in its purest form.

Sampling can be thought of as a quantization on the time axis. Sound, an acoustic wave, is sampled by interface devices (sound cards, etc.) after it is converted into electrical voltage variations in the microphone. For example, the sampling used in CDs is done at 44,100 samples / second and 16 bit words long, approximately twice the human hearing sound frequency limit of 20,000 Hz. This value means taking 44,100 samples per second from the analog signal and displaying each sample with 16 bits in digital environment.



*Figure 3*

Sampling determines the quality of the digital recording made. Higher frequency sampling will provide a higher quality conversion from analog to digital, and will take up a lot of space in terms of data size.

**SOUND PROCESSING**

It is imperative to process sound in a properly functioning speech recognition system. At this stage, the aim is to have information about the sound without disturbing the properties of the recorded sound. The processing of sound takes place in a number of specific steps. If it is necessary to write these steps according to the order of the process, the stages of windowing of the sound, normalization, filtering of the sound, encoding of the sound and spreading of the sound signal over time can be listed.

WINDOWING THE SOUND: The windowing process is partly like the filtering process. The aim is to make the starting and ending points of the audio signal more specific. It is not possible to manipulate any infinitely long string. For this reason, the starting and ending points of the audio signal must be determined. This process is called windowing of sound. In the simplest windowing technique, the part of the sign to be examined is multiplied by 1 and the unobserved interval outside by 0 [1]. There are many window functions in the literature. To give a few example to these; Hamming Window, Rectangular Window, Barlett Window, Hanning Window, Blackman Window and Kaiser Window are some of the windowing functions.

NORMALIZATION PROCESS:Normalization is performed in order to remove the negative effects of high amplitude values in the audio signal during eigenvector extraction in order to obtain the necessary data in the audio file accurately and without loss. Normalization process is obtained by dividing each of the samples in the file by the highest amplitude value [2].

[n]=

FILTERING THE SOUND: Filtering is done to separate the valuable data contained in the audio signals from the noise signals. The most basic of voice recognition systems are effective filtering processes. Filters are divided into many parts. These filters are classified in the chart below.

Noise signal: Signals in undesirable frequency range.

Filters

Digital Filter

Analog Filter

Low-Pass Filter

High-Pass Filter

Band-Pass Filter

Band-Stop Filter

All-Pass Filter

Frequency Based

Hardware Based

Active Filter

Passive Filter

*Figure 4*

CODING OF THE SPEECH SIGNAL: After the digitization of the sound, the coding of the sound is performed. Effective coding of the speech signal is an issue that should be emphasized, especially in digital communication systems. Since the sound signal is not deterministic, it cannot be defined with the help of analytical formulas and can generally be characterized with the help of statistical functions such as Power Spectral Density Function, Autocorrelation Function, Probability Density Function, Probability Density Function. Audio coding methods can be examined in three main classes: waveform coding, speech coding, and hybrid coding. Firstly, they are expressed as waveform coding aiming to recreate the sound and secondly as voice coders aiming to preserve the speech characteristics in the encoded signal.

Waveform encoders are generally designed not to be dependent on the signal. In such encoders, the signal at the decoder output is intended to be an exact copy of the encoded source signal. Such encoders allow encoding of another type of signal, such as signaling signals and music. Therefore, their coding efficiency is lower.

PCM (Pulse Code Modulation): The most common way to convert analog signal to digital is PCM. Sampling is taken from the analog signal at a certain frequency. Each sample is converted to a two-coded number (0's, 1's).

ADPCM (Adaptive Differential Pulse Code Modulation): It is a modulation technique used to convert an analog signal into its binary form. The PCM saves data in the .wav file using a sampling rate of 8 bits per second. ADPCM will use the holes between samples at a rate of 4 bits per second.

The working principle of the vocoder is based on modeling the speech signal generation mechanism. The excitation signal expressed by E (z) is filtered through the vocal organs acting as a spectral shaping filter with transfer function H (z) = 1 / A (z). Instead of obtaining an exact duplicate signal of the input at the decoder output, here the appropriate number of audio parameters is calculated so that it can adequately reproduce that signal.

LPC (Linear Predictive Coding): Linear prediction is basically based on the principle that sound can be modeled by the output of a linear, time-varying system induced by periodic impulse or random noise.

It can be formulated the relation between the input and output as,

H[z]

When the inverse z-transform is applied,

LPC works on the principle that the next sample can be approximated from a previous series of samples.

with the solution, p number of LPC parameters are calculated. Where p is the LPC encoder level; If a1, a2, ..., ap is expressed as LPC Parameters.

The p number of parameters calculated by the optimization of this equation corresponds to the encoding result by outputting a frame sample input to the LPC encoder.

**COMPARISON AND PAIRING**

The aim in this part of the Project is to create a series of possible words by comparing and matching the encoded and featured extracted audio signals with the stored templates in the system. The most common techniques in these operations are based on statistical methods. These are:

Hidden Markov Model (HMM)

Artificial Neural Networks (ANN)

Dynamic Time Warping (DTW)

HIDDEN MARKOV MODEL(HMM): This model is a widely used statistical method for classifying discreate time observed samples. It can simply be explained as defining the situations that can not be directly observed with the help of observable situations. The reason why it is frequently used in voice recognation systems is its rich mathematical structure

Beginning

0

a01

a11

q2

a12

a34

final

4

q1

a33

b2(o5)

a22

a23

q3

0

b2(o3)

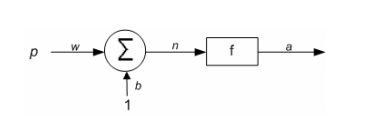
*Figure 5 Markov Chains Model*

Q: situations

A : transition possibilities

B: observation possibilities

ARTIFICIAL NEURAL NETWORKS(ANN): Artificial neural networks (ANN) are inspired by the human brain. It is based on the mathematical modeling of the learning process.ANN studies first started with the modeling of neurons, which are biological units of the brain.Today, it is applied in many areas. The artificial neural network basically the human brain its structure, neurons and the connections between them are taken. Artificial neural networks, the human brain It can be said that it is a simple model. The figure shows a single-entry neuron model.



*Figure 6*

In this neuron model: p represents the input to this neuron and a represents the output from this neuron.. Here w is the weight and b is the bias. Neuronoutput a, p multiplied by the weight of input p to this neuron, again the bias voltage in this neuronand the result (i.e., n) from the sum of output.

Biological nervous system: Artificial neural networks

Neuron Processor element

Dendrite Summation Function

Cell body Transfer Function

Artificial Axons Neuron Output

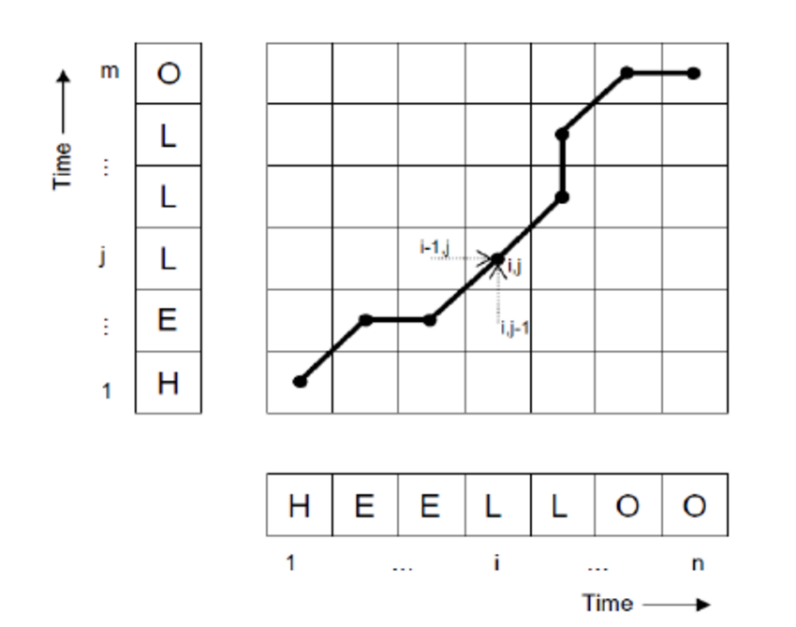
Synapse Weights

n data are entered into a cell (Xn data entry). All data is collected by multiplying the entered data by weights, and then bias is added, resulting in a clear judgment. Net input is passed through the activation function and a data output is obtained happens.

The aim in an artificial neural network is to minimize errors. The method for this is for all training data is to minimize the sum of the squares of the errors that occur. This is an optimization is the problem. Many optimization methods have been developed and applied in this regard.As a result, the aim of the training of the artificial neural network is that the weights and coefficients are determined by the solution of the optimization problem. In order to do this, weights and biases must first be within a certain range. Assigned as random values. Step by step, with optimization problem solution, weights and the biases are updated with new values. Consequently, these weights and biases values are determined and the function is modeled.

Artificial neural networks to create a relationship between inputs and outputs often for pattern recognition or classification they are used. The purpose of voice recognition systems is since the phoneme and the determination of words from there, the processing of the sound signal the result is that artificial neural networks of feature vectors corresponding to the audio signal. Classification can be achieved with the help. Artificial neural networks in speech recognition systems usage is carried out in this way.

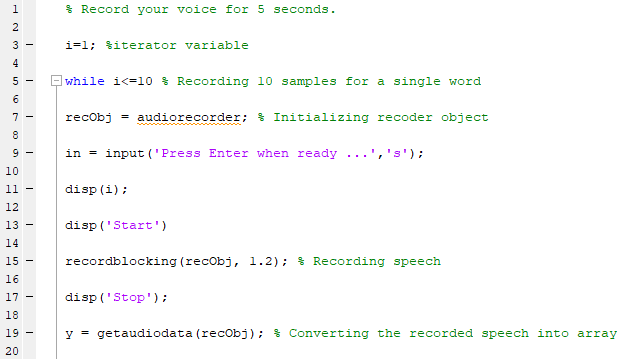
DYNAMIC TIME WARPING(DTW): Dynamic time bending tries to bring the different sounds of a word closer together by spreading or narrowing the sound over time. This method, which is frequently preferred in word-based speech recognition systems, compares the times of the voice received from the user with the word templates stored in the system



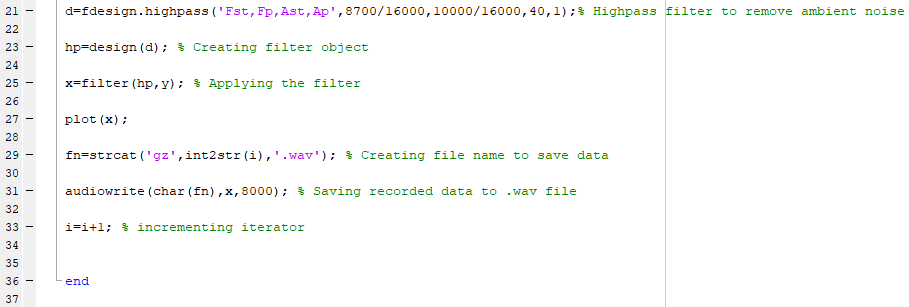
*Figure 7*

In this figure 7 , the sample consisting of M samples placed on the y-axis one audio signal and another consisting of n samples placed on the x-axis comparison of the signal is shown. The aim; of two series placed perpendicular to each other and reach to coordinate (n,M) from coordinate (1,1) with the easiest way. Each sample on the x-axis is compared with that on the y-axis and calculated in the matrix. In this way, when the matrix is ​​completed, the black lines infer how the two signals look alike.

The application of the speech recognition system created using LPC in Matlab is given in detail below.



In the first stage, it is necessary to create a sound file directory so that the speech recognition system can make a comparison with the sound detected. Sample sounds were recorded to create this index. In order to record the same voice recording ten times, the While loop was started by defining the i variable from 1 to 10.. In order to record sound, the audiorecorder function embedded in the matlab program is assigned to a string. This string is specified as recObj in the above function. After assigning a command with the input function, the disp function defined with the i variable starts the voice recording with “start” by indicating which sound recording has been taken. With the Recordblocking function, it is specified how long the recObj file will take in seconds. Finally, with the exit of the "stop" command, the audio recording is completed and the recorded file is digitized with the getaudiodata function.



In this section, a high pass filter specification object (d) is created by applying the 'Fst, Fp, Ast, Ap' values ​​with the fdesign.highpass function.

Fst — frequency at the end of the stop band. Specified in normalized frequency units. Also called Fstop.

Fp — frequency at the start of the pass band. Specified in normalized frequency units. Also called Fpass.

Ast — attenuation in the stop band in decibels (the default units). Also called Astop.

Ap — amount of ripple allowed in the pass band in decibels (the default units). Also called Apass.

Using these values, a filter is created with the help of design function. To eliminate ambient noise with the "filter" function, the High Pass filter is assigned to the variable y. The file name has been created thanks to the strcat function. Saved as .wav file with audiowrite function.

0

F

pass

A

f(Hz)

Fs/2

Mag. (dB)

A

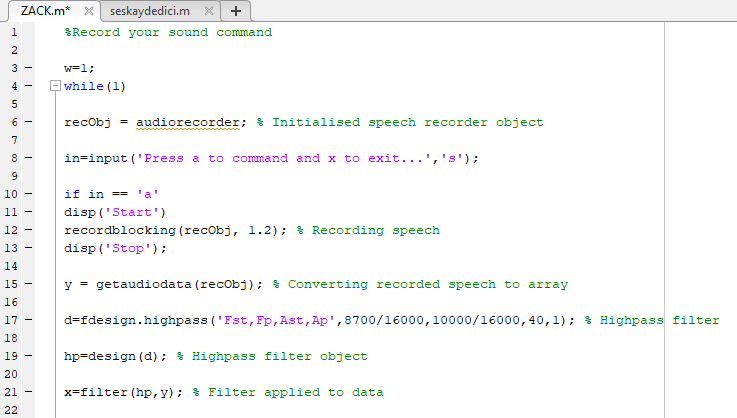
stop

pass

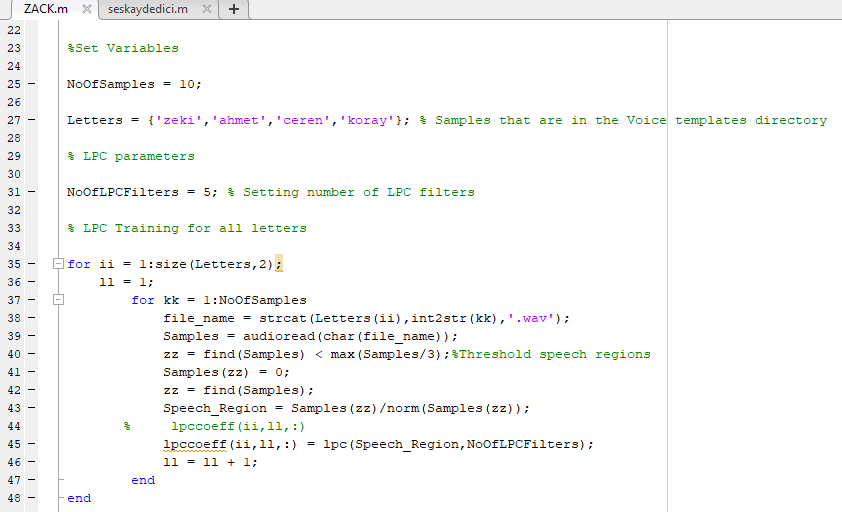
F

stop

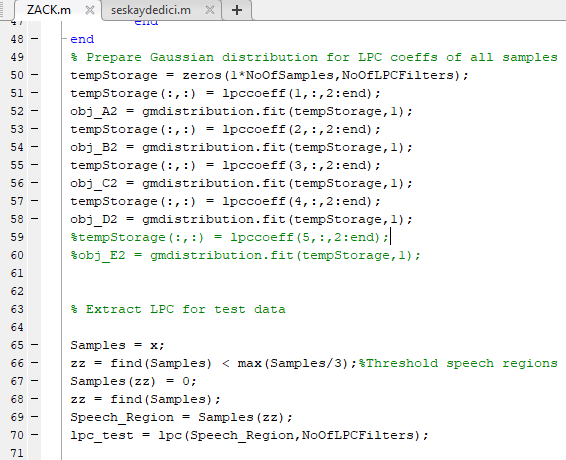
0



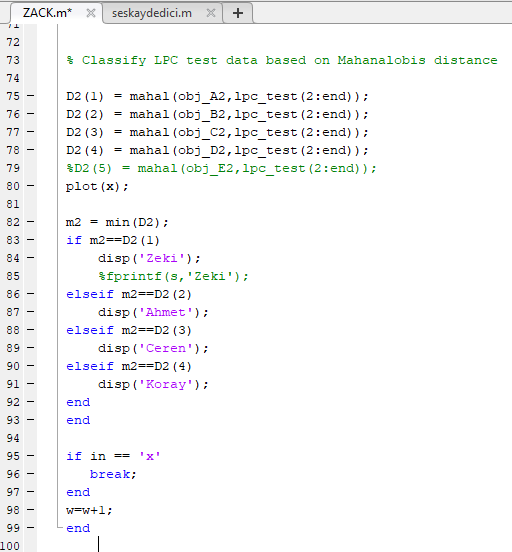
The audio processing part immediately starts with the audio recording, so that you can compare it with pre-recorded sounds. All steps performed in the sound recorder are repeated.



A for loop has been created to examine previously uploaded audio files to the system. To determine the correct file in this for loop, the sample file names specified in the "letters" section were used and the number of repetitions of the same sounds with the NoOfSamples section was used. At the same time, the sound coefficients were obtained by using the lpccoeff function for LPC by performing normalization in the for loop for each sound.



After temporarily stored in tempStorage with the obtained coefficients, it is inserted into the gmdistribution function and uses the Expectation Maximization (EM) algorithm to generate a sample of the parameters in the Gaussian mixture model with maximum probability estimates. The filtered input audio file is added to the LPC function after normalization.



With the upper sound sample with LPC coefficients, the input sound file from the LPC function is inserted into the 'mahal' function. 'Mahal' distance measure determines the discrete value by considering the standard deviations of these two variables and measures the proximity of the sounds. As a result, the word with the match is printed.

**CONCLUSION**

Voice recognition system is an application developed to provide fluent and natural conversation between people and technology. This project demonstrates how this can be done in the simplest way possible. In the research, various sound recordings were transferred to digital media, using statistical techniques such as Gaussian Distribution and LPC as the encoder. It was aimed to compare the sample sound and the inputs received from the user and to result in the most appropriate matching. By developing the system, it can be used in areas such as smart home systems, the use of electronic devices by people with disabilities and security systems.

**REFERENCE**

[1] [Hyung-Suk Kim](https://ccrma.stanford.edu/~hskim08/), “Linear Predictive Coding is All-Pole Resonance Modeling”, Stanford University. URL:https://ccrma.stanford.edu/~hskim08/lpc/

[2] Fikret ARI, “SES KODLAMA TEKNİKLERİNİN SERBEST ORTAM OPTİKSEL İLETİŞİME UYGULANMASI”, Ankara, 2006.

[3] Daniel Jurafsky & James H. Martin, “Speech and Language Processing”. URL: <https://web.stanford.edu/~jurafsky/slp3/A.pdf>

[4] Süleyman Çeven, Raif Bayır, 2020, “Ortam Sesinden İnsan Sesinin Ayrıştırılması için Filtre Geliştirilmesi”, Avrupa Bilim ve Teknoloji Dergisi.

[5] Pavel Senin, “Dynamic Time Warping Algorithm Review”, Honolulu, 2008.

[6] Özkan YILMAZ, Yaşar KARTAL, Y.Buğra ÖZER, “Ses Etiketleme”, İstanbul.

[7] Hüseyin ÇAKIR, Bekir OKUTAN, “Ses Kontrollü Web Tarayıcı”, Bilişim Teknolojileri Dergisi, 2011

[8] Nursel YALÇIN, "KONUŞMA TANIMA TEORİSİ VE TEKNİKLERİ", Kastamonu Eğitim Dergisi, Mart 2008

[9] <http://eeminucer.blogspot.com/2015/06/ksa-ksa-orneklemesampling-nedir.html>

[10] https://elektronikhobi.net/analog-sinyallerin-orneklenmesi-sampling-ve-nyquist-teoremi/

[11] <https://www.pc-audiophile.com/sayisal-ses-digital-audio-pcm-ve-dsd-hakkinda-ozet/>