A NEW MODEL FOR SPEECH COMPRESSION

ASENA EKŞİOĞLU

IŞIK UNIVERSITY
SEPTEMBER, 2021

A NEW MODEL FOR SPEECH COMPRESSION

ASENA EKŞİOĞLU
B.S., Electrical and Electronic Engineering, Işık University, 2016

Submitted to School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science in Electronic Engineering

IŞIK UNIVERSITY
SEPTEMBER, 2021

IŞIK UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

A NEW MODEL FOR SPEECH COMPRESSION

ASENA EKŞİOĞLU

APPROVED BY:

Assoc. Prof. Ümit Güz	Işık University
(Thesis Supervisor)	
Assoc. Prof. Hakan Gürkan	Bursa Technical University
(Thesis Co-Advisor)	
Assist. Prof. Ramazan Köprü	Işık University
Assoc. Prof. Cemal Hanilçi	Bursa Technical University

APPROVAL DATE: 20 / 09 / 2021

A NEW MODEL FOR SPEECH COMPRESSION

ABSTRACT

This thesis aims to propose a new low bit-rate speech compression method. In the proposed method, classical SYMPES (Systematic Procedure for Predefined Envelope and Signature Sequences) based predefined signature and envelope sequences are obtained using zero-cross and phoneme-based segmentation. Some disadvantages of the traditional SYMPES technique, like computational complexity, relatively long encoding times, and so on, are also reduced in the new version in order to obtain lower bit rates.

The new approach significantly reduces the bit-rate and yields high compression ratios with more intelligible speech quality than that of the classical SYMPES approach. Furthermore, in comparison to other traditional techniques such as the CELP (Code Excited Linear Predictive) coding algorithm, experimental findings demonstrate that at almost the equal bit rates, extremely promising speech quality is produced.

Key words: Low bit-rate speech coding, Speech compression, SYMPES, CELP, MOS.

SES SIKIŞTIRMA İÇİN YENİ BİR MODEL

ÖZET

Bu tezde, yeni bir düşük bit hızlı konuşma kodlama yöntemi önermek amaçlanmaktadır. Önerilen yöntemde, klasik SYMPES tabanlı önceden tanımlanmış imza ve zarf dizileri, sıfır geçiş oranı ve fonem tabanlı segmentasyon kullanılarak elde edilir. Daha düşük bit hızları elde etmek için klasik SYMPES yönteminin hesaplama karmaşıklığı, nispeten yüksek kodlama süreleri vb. gibi bazı dezavantajlar da yeni yaklaşımda önemli ölçüde azaltılmıştır.

Bahsedilen yeni yaklaşım, bit hızını önemli ölçüde azalttığı için klasik SYMPES yaklaşımından daha anlaşılır bir konuşma kalitesiyle yüksek sıkıştırma oranları sağlar. Ayrıca, deneysel sonuçlarda, CELP (Code Excited Linear Predictive) kodlama algoritması gibi diğer geleneksel yöntemlere kıyasla neredeyse aynı bit hızlarında çok daha umut verici bir konuşma kalitesinin elde edildiğini göstermektedir.

Anahtar Kelimeler: Düşük bit hızlı konuşma kodlama, Konuşma sıkıştırma, SYMPES, CELP, MOS.

ACKNOWLEDGEMENTS

Firstly, I am so thankful to have my supervisor, Ümit Güz for his endless help. Thanks to him and the rest of the other my supervisors (Hakan Gürkan, Burak Şişman, and Ebru Gürsu Çimen), I never hesitated to succeed and I always tried to improve myself in this journey with their priceless knowledge.

Also, I would like to thank you my dearest parent for supporting me in all circumstances and for their belief in me in my whole life and, to my lovely friends who motivate me with their positive energy.

Asena EKŞİOĞLU

To my family. . .

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ZC	: Zero Cross
CR	: Compression Ratio
LPC	: Linear Predictive Coding
MELP	: Mixed Excitation Linear Prediction
CELP	: Code Excited Linear Prediction
PCM	: Pulse Code Modulation
ADPCM	: Adaptive Differential Pulse Code Modulation
SYMPES	: Systematic Procedure for Predefined Envelope and Signature Sequences
MOS	: Mean Opinion Score
SNR	: Signal to Noise Ratio

CHAPTER 1

1. INTRODUCTION

Transmission and storage of speech signals are critical in modern communication systems, especially to make effective use of the restricted transmission bandwidth for compression and to store data in very small data spaces. Since there is often a tradeoff between lower bit rate and speech quality, the most important thing to remember in any of these applications is to keep the decoded speech's perceptual quality while lowering the bit rate or raising the compression ratio (Goldberg, 2019) (Guo & Kuo, 2007).

In this thesis, we present a new coding scheme consisting of two processes; ZC (zero-cross) based segmentation of the phonemes and construction of the phoneme-based signature and envelope vectors. Instead of set frame lengths such as 16 or 32 samples, the new approach divides speech signals into variable frame lengths based on the ZC lengths of the phonemes. The most important contribution of this work is to get the capability to create a variable rate encoder thanks to the variable length of the signature and envelope vectors using this new approach. This approach combines the advantages of the traditional SYMPES method with the benefits of ZC and phoneme-based segmentation, resulting in a new and more sophisticated coding scheme that achieves lower bit rate speech compression than the traditional SYMPES algorithm.

This thesis is organized as follows: In Chapter 2, the speech production mechanism and analysis of speech signals are explained. Then, in Chapter 3 definitions and types of speech compression are stated following with the explanation of speech coding techniques and their literature reviews. Additionally, evaluation metrics and computation of the performance parameters are explained. Subsequently, in Chapter 4, the classical SYMPES method is briefly explained and theoretical aspects of the newly proposed speech coding algorithm based on zero cross and phoneme-based

SYMPES are presented. Encoding and decoding schemes and algorithms are also illustrated in this section. After the expression of used fundamental information, the four proposed approaches are clearly explained. Then, experimental and comparative results of the speech coding methods including classical SYMPES and CELP, and the newly proposed method are presented in Chapter 5. And the conclusion of the thesis is given in Chapter 6.

CHAPTER 2

2. SPEECH PRODUCTION MECHANISM

2.1 Definition of Voice

The acoustic pressure wave produced by the voluntary movements of anatomical structures in the sound production system can be defined as a sound wave. The sound formation is a physical phenomenon. Sound is a special type of motion energy and is produced by the vibration of objects. When an object begins to vibrate, it also vibrates the surrounding air molecules. Additionally, these vibrations lead to speech waves which are occurred by the cause of pressure differences in air, directly (Rabiner & Schafer, 2007). The number of compressions per second that occurs during the formation of the sound gives the frequency. The human ear can hear sounds between 20 Hz and 20 kHz (Hansen, 2001).

Speech signals are non-stationary signals and if the speech signal is segmented, it can be seen that it consists of basic elements of 5-20 milliseconds (Matassini, 2001). Speech signals can be divided into two parts which are voiced or unvoiced. Here, the voiced parts are vowels that we know thereby, the unvoiced parts are the pronunciation of the remaining letters that are said to be silent. The energy of the voiced part is normally quite high compared to the unvoiced part.

2.2 Production of Speech

Regardless of the language spoken, all humans use the same anatomy to speak. The sounds that can produce are limited due to human anatomy. Only in some languages, laryngeal and nasal sounds are used more. Producing the speech signal can be roughly described as through the lungs pumping air first into the sound system and then to the mouth (Rabiner & Schafer, 2007). According to this definition, the lung

can be considered as the sound source, while the vocal system can be thought of as a filter that produces the speech signal by producing various sounds.

To understand how the vocal system converts air from the lungs into sound, it is necessary to understand the basic definitions. The phon is the smallest distinguishing element in the language. For example, Turkish is a phonetic language, each letter can be thought of as a phoneme. The limited-independent formed of phon sets are also called phonemes. There are two types of phonemes, voiced and unvoiced. The voiced phonemes are usually vowels, with high average energy levels and different resonant frequencies. The voiced phonemes are comprised of the periodic vibration of the vocal cords which are generated by the air that comes from the lungs (Wolfe, Garnier, & Smith, 2009). The frequency of the vibration determines the pitch of the sound produced. Air pulses formed as a result of these vibrations pass through the vocal system and formants occur at some frequencies. The sounds that are produced by women and children are usually of a higher pitch, hence a higher frequency. Essentially the sound vibration is an important parameter in terms of producing different types of sounds. The other component that affects voice production is the shape of the vocal system. Different sounds occur according to geometry and resonance frequencies. The vocal system consists of the larynx, tongue, nose, and mouth. Figure 2. 1 shows the anatomy of the human speech production mechanism.

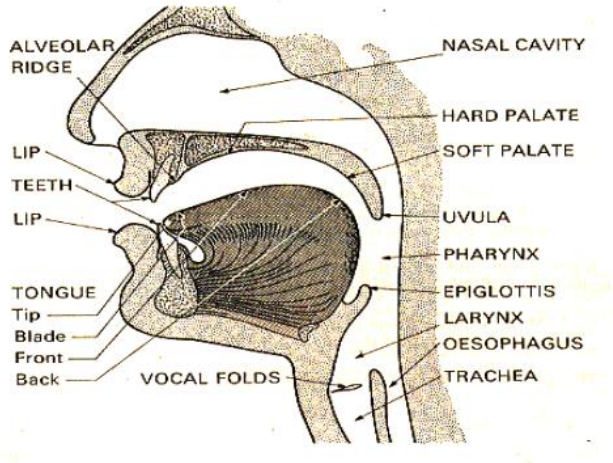

Figure 2. 1: Representation of the anatomy of human speech **(Campbell, 1997)**.

2.3 Analysis of Speech Signal

As previously mentioned, speech signals are divided into two groups, such as voiced and unvoiced signals. In contrast to voiced signal which is a periodic signal, the unvoiced signal is in the form of noise.

Moreover, to identify whether a speech segment is voiced or unvoiced, there are two fundamental checklists which are the zero-crossing rate and energy level of the signal (Grassi, 1998). All the researches that have been made so far show that voiced signals have a lower ZC rate than the unvoiced signals whereas the energy level is higher. Simple representation of both voiced and unvoiced speech signals are shown in Figure 2. 2 and the decision block diagram for labeling the signal as voiced or unvoiced is illustrated in Figure 2. 3.

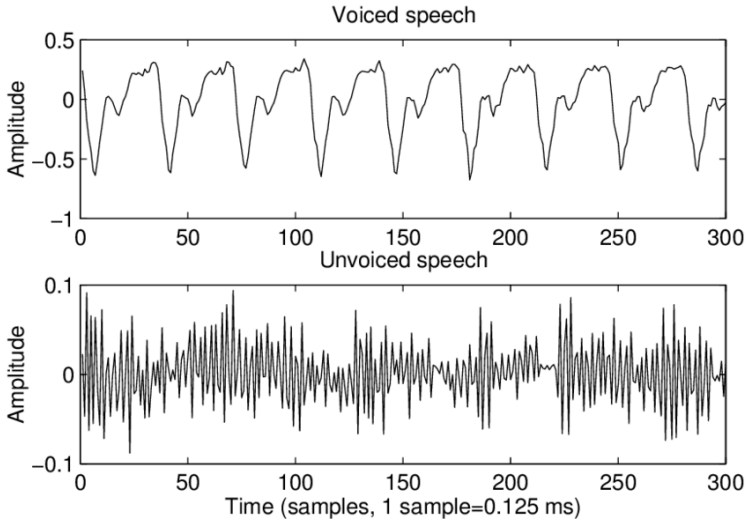

Figure 2. 2: Voiced/Unvoiced signal representation **(Grassi, 1998)**.

2.3.1 Voiced/Unvoiced (v/u) Decision of Speech Signal

Operations in speech signals are performed in very small-time intervals because the speech signals are similar to each other and, these operations' brief explanations are given below.

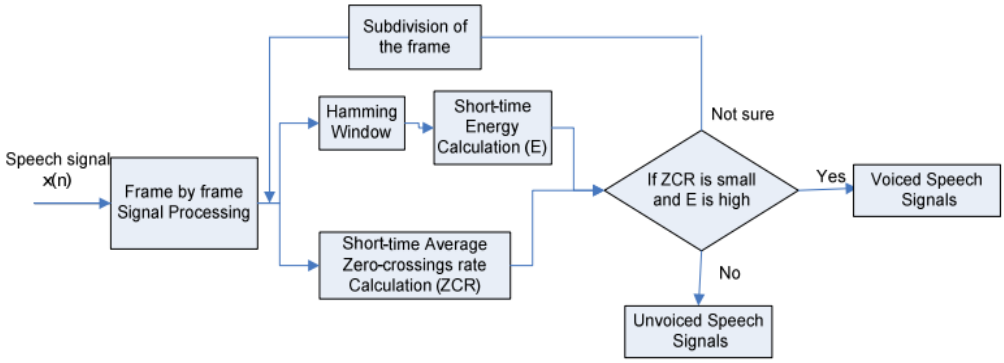

Figure 2. 3: The illustration of voiced/unvoiced classification with block diagram (Bachu, Kopparthi, Adapa, & Barkana, 2008).

2.3.1.1 Zero-Crossing (ZC) Rate

The zero-crossing rate is defined as the rate at which successive samples of a discrete-time signal have different mathematical signs. As shown in Figure 2. 4, the rate of occurrence of zero crossings reflects a simple measurement of the frequency content of the signal. Zero crossing rate in speech signals is measured by the number of times the amplitude value of the speech signal exceeds the zero value in a certain time interval or within a frame.

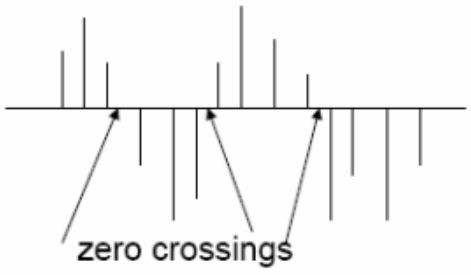

Figure 2. 4: Zero-Crossing representation of a discrete time signal (Bachu, Kopparthi, Adapa, & Barkana, 2008).

A mathematical expression of ZC rate can be given below in Equation (2.3) and (2.2).

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn\,[x(m-1)]| \tag{2.1}$$

where,

$$sgn[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \tag{2.2}$$

2.3.1.2 Energy Calculation of Speech Signal

In speech signals, depending on time amplitude vary. Short time energy value is much higher for voiced signals than the unvoiced signals. Therefore, short-time energy gives us important information about speech samples (Bachu, Kopparthi, Adapa, & Barkana, 2008). The formulation of short time energy calculation is given in Equation (2.3)

$$E(n) = \sum_{n=1}^{N} X_i^2(n) \tag{2.3}$$

Here, N, X(n), $X_i(n)$ and E(n) represent the frame length, the original speech file, the i-sample frame of speech file and the frame energy, respectively.

CHAPTER 3

3. SPEECH COMPRESSION

The limited capacity of digital storage media and digital transmission lines directed many researchers to express the speech signals with fewer required bits but acceptable quality. This is where the concept of speech compression or in other words speech coding emerged. Speech compression can be described as taking the basic information and its characteristics from the original speech signal in order to re-generate it later. With the help of compression, speech signals occupy less storage space and, less bandwidth in communication, shorter transmission times are gained (Jagtap, Mulye, & Uplane, 2015). There are two types of compression methods; lossless and lossy compression, respectively.

3.1 Lossless Compression

There is no loss of information in the data when using lossless compression techniques, as the name implies. The original data can be reconstructed from the completely compressed data if the data is lossless compressed. Lossless compression techniques are used in applications that do not tolerate any difference between the original and reconstructed data such as text compression. It is critical that the reconstructed data be identical to the original data. Different interpretations of this data set will result from even the tiniest variation. Huffman coding and Arithmetic coding can be given as an example of lossless compression (Warkade & Mishra, 2015).

3.2 Lossy Compression

Lossy compression takes the risk of some data losses. It is well known that the original data cannot be entirely recovered when compressed data is rebuilt using this approach. The compression can be applied with the elimination of pieces of information that the human ears have trouble hearing from the data (Ng, Choi, & Ravishankar, 1997). Depending on the quality expected from the reconstructed speech, a variable amount of information loss per sample can be taken into account. Linear Predictive Coding (LPC), Pulse Code Modulation (PCM), and Code Excited Linear Prediction (CELP) are some of the basic examples of lossy speech compression techniques.

3.3 Compression Techniques

Speech coding research started in 1939 with the pioneering work of H. Dudley of the Bell Telephone Laboratories. After observing the redundancy in the speech signal, Dudley implemented the first voice coder namely vocoder using the first analysis-synthesis method (Dudley, 1939) (—, The Vocoder, 1939). Speech compression has been and continues to be, a significant task in the field of digital speech processing (Shannon, 1993) (Spanias, 1994) (Gersho & Gray, 1993). Nowadays, for digital cellular communications, voice over Internet protocol (VoIP), voice response applications, and video conferencing systems, speech coding is a key technology (Chen & Thyssen, 2008) (Gibson, 2016).

Several new algorithms for coding (or compressing) speech signals have been proposed in the past. These algorithms are developed based on numerical, mathematical, statistical, and heuristic methodologies. Wide-band speech coding and narrow-band speech coding are the two primary types of speech coding. However, in general, speech coding standards for speech communication are concerned with the narrow-band speech which is sampled at 8KHz sampling frequency. Narrow-band speech coding standards can be grouped as voice coders (vocoders) or parametric coders, waveform coders, and hybrid coders (Deller Jr, Hansen, & Proakis, 1999) (Rabiner & Schafer, 2010) (Osman, Al, Magboub, & Alfandi, 2010).

3.3.1 Voice Coders

In the voice coders due to the different approaches to speech coding known as parameter coding or analysis-synthesis coding, there is no effort to reproduce the exact speech waveform at the receiver. Vocoders provide much lower data rates using a functional model (source filter or vocal tract) of the human speech production mechanism at the receiver (Makhoul, 1975). LPC (Linear Predictive Coding) and MELP (Mixed Excitation Linear Prediction) can be given as an example of a known voice coder and they use multiple parametric models to generate speech signals.

Low bit rates of 2.4 kbps are used in LPC methods like LPC-10E (FS1015). LPC is a source filter analysis-synthesis methodology that approximates speech generation as an excitation (a pulse or noise) passing through an all-pole resonant filter. LPC reduces the amount of data (frame) to a few filter coefficients. Many current speech processing systems employ LPC in a variety of applications, including coding, analysis, synthesis, and recognition. LPC is a satellite communication protocol with a high complexity range of 2.0-4.8 kbps (Raja, Jangid, & Gulhane). The LPC has an advantage since it refers to a simplified vocal-tract model and compares a source-filter model to the human speech production system. LPC analysis is usually more appropriate for modeling vowels that are periodic, except nasalized vowels. The LPC method has several drawbacks, one of which is its synthetic (not natural) reconstruction performance, particularly in a vowel or voiced portions of speech signals (Atal & Remde, 1982) (Tremain, 1982) (El-Jaroudi & Makhoul, 1991) (Kondoz, 1998) (Murthi & Rao, 2000) (Ekman & Kleijn, 2008) (Itakura, 1975).

In order to improve the performance of the low-bit-rate speech coders, alternative representations to LPC parameters are introduced called LSP (Line Spectrum Pairs) or LSF (Line Spectral Frequencies). This method proposes a new model and bit rate for the LPC-10 exploiting the LSP parameters (Hasegawa-Johnson, 2000) (—, Specifications for The Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction, 1980).

In 1996, an LPC-based method which is called MELP was chosen as the 2.4 kbps Federal Standard Vocoder by the United States, Department of Defense Digital Voice Processing Consortium (—, Specifications for The Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction, 1980).

MELP was bringing additional features such as mixed excitation, aperiodic pulses,

pulse dispersion filtering, adaptive spectral enhancement, and Fourier magnitude modeling when compared to the traditional LPC (Kohler, 1997). Although the

Reconstructed speech by the MELP has better quality than the LPC, wrong estimations of some features in the MELP model such as line spectrum frequencies, pitch frequencies, and voiced/unvoiced (V/U) decision may result in audible distortion and low estimation accuracy in the other parameters (Gavula, Scheets, Teague, & Weber, 2008). These errors frequently occur in the V/U decision and pitch estimation processes especially in the transition segments which contain both voiced and unvoiced parts of the speech (Wu, Jiang, & Li, 2009). On the other hand, it should be noted that pitch estimation and voicing level calculation are important in low bit-rate speech coding because they affect the quality of the reconstructed speech.

3.3.2 Waveform Coders

Waveform coders attempt to reproduce the original shape of the actual frame of the speech signal (Varghese & Ramesh, 2015). Therefore, the similarity is very important between the original and reconstructed shapes of frames.

Waveform coders are mainly based on the PCM coding scheme and these coders contain, PCM, DPCM (Differential Pulse Code Modulation), and ADPCM (Adaptive Differential Pulse Code Modulation) coders. PCM-based techniques such as ADPCM (G.726) yield much better perceptual quality over LPC-10E but demand higher bit rates of 32 or 16 kbps. Waveform coders have very low computational complexity and delay, but they require a large number of bits to maintain better speech quality (Oliver, Pierce, & Shannon, 1948) (Jayant, 1974) (Ramamoorthy & Jayant, 1984) (Draft Rec, 1988).

3.3.3 Hybrid Coders

Hybrid coders combine the benefits or some useful features of the voice coders and waveform coders in order to provide better PESQ (perceptual evaluation of speech quality) (Rix, Beerends, Hollier, & Hekstra, 2001) performance at low bit rates.

CELP (Code Excited Linear Predictive) is the most commonly used hybrid speech coder based on the principle of the LPC (Jage & Upadhya, 2016). The CELP coder, first proposed in 1985. It has been shown that the CELP coder is one of the most efficient ways of encoding speech at very low bit rates (Jage & Upadhya, 2016)

(Schroeder & Atal, 1985) (Campbell Jr, Tremain, & Welch, 1991). Most of the speech coding standards currently deployed in communication systems (i.e. mobile communication systems) are based on the CELP algorithm.

CELP model integrates vector quantization with prediction-based coding. The excitation signal in time domain analysis-by-synthesis speech coder is chosen by searching through a huge vector quantizer codebook to match the reconstructed speech waveform as nearly as possible to the original speech waveform. A complete search of all potential codebook excitation vectors necessitates a high computational complexity of the coder, which is sometimes unavailable even with contemporary digital signal processors (Kumar, 2007) (Jelinek, Eksler, Lemyre, & Lefebvre, 2007) (Chu, 2003) (Kleijn, Krasinski, & Ketchum, Fast Methods for The CELP Speech Coding Algorithm, 1990).

3.4 Some Recent Work for Speech Compression Techniques

In addition to conventional methods, many speech coding techniques have taken their place in the literature. In research (Bansal & Sircar, 2018), a new model of low bit-rate (4.11 kbps) speech coding technique is represented using a model of amplitude and frequency modulated (AFM) signal with the help of Fourier–Bessel series to extract the amplitude envelope (AE) and the instantaneous frequency (IF) to represent the signal.

In (Uddin, Ansari, & Naaz, 2016), different speech samples have been encoded with a lower number of bits as compared to the original voice samples devoid of much deterioration in the voice quality. In (Kleijn, ve diğerleri, 2021), it is claimed that generative modeling with 3 kbps codings for real-world speech signals has an acceptable computational complexity. Another low-bit-rate speech coder based on non-uniform sampling is presented in (Iem, 2015), which uses detection of inflection points (IP) to produce an SNR of about 5.27 dB at a data rate of 1.5 kbps.

In (Iem, 2015; Yarman, Güz, & Gürkan, 2006) (Güz, Gürkan, & Yarman, 2007), a novel method referred to as SYMPES (Systematic Procedure for Envelope and Signature Sequences) was introduced and implemented on the representation of the 1-D signals such as speech signals. In those works, comparative results of SYMPES and other traditional speech compression standards such as LPC and ADPCM (G.726) were also presented.

SYMPES was also applied in the compression of the bio-signals like ECG (Electrocardiogram) (Gürkan, Güz, & Yarman, Modeling of Electrocardiogram Signals Using Predefined Signature and Envelope Vector Sets, 2007), EEG (Electroencephalogram) (Gürkan, Güz, & Yarman, EEG Signal Compression Based on Classified Signature and Envelope Vector Sets, 2009) and EMG (Electromyogram) (Gürkan, Güz, & Yarman, A Novel Representation Method ror Electromyogram (EMG) Signal with Predefined Signature and Envelope Functional Bank, 2004) signals. In these studies, the signals are first examined in terms of their physical characteristics, and then signature and envelope functions are used to best characterize the signals. Signature functions were obtained by using the energy compaction property of the PCA (principal component analysis) (Jolli, 1993). PCA was also given an optimal solution via minimization of the error in the least mean square (LMS) sense.

A novel block-based image compression technique based on the creation of predefined block sets termed Classified Energy Blocks (CEBs) and Classified Pattern Blocks (CPBs) was proposed in a recent SYMPES-based effort. All of these distinct block sets were grouped under the Classified Energy and Pattern Blocks framework (CEPBs). The method's main stages were the construction of the CEPB, the encoding process (which included constructing the energy and pattern building blocks of the image to be reconstructed and obtaining the encoding parameters), and decoding (which included reconstructing the input image using the encoding parameters from the already located CEPB in the receiver part). It has been shown that the images were compressed from CR=20 up to CR=70 with very limited edge effect using the classical SYMPES Algorithm.

3.5 Evaluation Metrics

To assess the quality of speech signals that have been regenerated, there are two types of evaluation metrics. They are objective and subjective measurements, respectively.

3.5.1 Objective Measurement of Speech Signals

One of the common measurement methods in speech quality is Signal-to-Noise Ratio (SNR) measurement. In order to measure the objective quality of the reconstructed speech signals Segmental *SNR(SNR$_{seg}$)* is utilized. The *SNR$_{seg}$* is defined as the average of measurements of SNR over the frames and it is computed by

$$SNR_{seg} = \frac{1}{T_F}\sum_{j=0}^{T_F-1} 10\log_{10}\left[\frac{\sum_{n=m_j-K_F+1}^{m_j}[x(n)]^2}{\sum_{n=m_j-K_F+1}^{m_j}[x(n)-\hat{x}(n)]^2}\right] \quad (3.1)$$

T_F corresponds to the total number of frames, j is the frame index and K_F is the number of samples in each frame.

Let N be the total number of samples in the speech piece to be reconstructed. Then in Eq.

(3.1) $T_F = N/K_F$ corresponds to the total number of frames, j is the frame index; K_F is the number of sample in each frame. It should be noted that the indices $m_0, m_1, \dots, m_{T_F-1}\left(m_0 = K_F; m_j = jK_F\right)$ refer to the endpoints of each segment placed in the speech piece to be reconstructed.

3.5.2 Subjective Measurement of Speech Signals

To give a subjective aspect of speech signals, the Mean Opinion Score (MOS) test is evaluated. This test aims to classify the quality of the voice signals by a group of humans' judgment. The speech quality of the reconstructed signals is determined by a predefined scale from 1 (Bad) to 5 (Excellent) which is shown in Table 3. 1, and the average of these results gives the subjective opinion (MOS) about the reconstructed speech signals (ITU, 2016).

Table 3. 1: MOS Scaling Table **(ITU, 2016)**

MOS	Speech Quality
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

3.6 Computation of the Performance Parameters

The Compression Ratio *($CR_{Overall}$)*, Bit-per-sample *($BPSample_{Overall}$)* and Bit-per-second *($BPSecond_{Overall}$)* are used to evaluate the speech coding algorithm's performance. These parameters can be calculated by

$$CR_{Overal} = \frac{1}{N_{seg}} \sum_{k=1}^{N_{seg}} CR_{seg_k} \tag{3.2}$$

$$CR_{seg} = \frac{nbits \times L_{seg}}{b_{Total}} CR_{Overal} = \frac{1}{N_{seg}} \sum_{k=1}^{N_{seg}} CR_{seg_k} \tag{3.3}$$

$$b_{Total} = b_{phoneme} + b_G + b_{SE} \tag{3.4}$$

$$BPSample_{Overal} = \frac{nbits}{CR_{Overal}} \tag{3.5}$$

$$BPSecond_{Overal} = BPSample_{Overal} \cdot fs \tag{3.6}$$

where *nbits* and f_s denote the number of bits per sample and sampling frequency for the recorded speech signal, respectively. N_{seg}, L_{seg}, b_{total} and CR_{seg} the number of ZC segment, the length of the ZC segment, the total number of bits required to represent the current segment and compression ratio of the segment, respectively. Likewise, *($CR_{Overall}$)*, *($BPSample_{Overall}$)* and *($BPSecond_{Overall}$)* represent the overall compression ratio, overall bit per sample, and overall bit per second, respectively.

CHAPTER 4

4. PROPOSED APPROACHES IN THIS THESIS

In this chapter, before giving details of the proposed approaches, the classical SYMPES algorithm and the newly proposed speech coding method, which includes ZC and a phoneme-based segmentation technique, is interpreted. Then, the four different approaches are established to reach maximum compression results. The main difference is generated codebook sizes and forms among these approaches. These approaches' explanations are given in this chapter but their experimental results and comparison with other methods are stated in Section 5

4.1 Classical SYMPES Approach

In the classical SYMPES approach (Yarman, Güz, & Gürkan, 2006) (Güz, Gürkan, & Yarman, 2007) (Gürkan, Güz, & Yarman, Modeling of Electrocardiogram Signals Using Predefined Signature and Envelope Vector Sets, 2007) (Gürkan, Güz, & Yarman, EEG Signal Compression Based on Classified Signature and Envelope Vector Sets, 2009) (Gürkan, Güz, & Yarman, A Novel Representation Method ror Electromyogram (EMG) Signal with Predefined Signature and Envelope Functional Bank, 2004), a discrete-time mathematical model is proposed in order to best represent the equally divided speech frames (fixed frame lengths) into reasonable lengths of time of the speech signals. Such as the part of the speech signal $x(n)$'s discrete time domain representation is illustrated in Figure 4. 1. Here, N_F and L_F are stated as the frame number and the length of number of samples in each frame, respectively.

In classical SYMPES (Yarman, Güz, & Gürkan, 2006) (Gürkan, Güz, & Yarman, A Novel Representation Method ror Electromyogram (EMG) Signal with Predefined Signature and Envelope Functional Bank, 2004)

Scheme, for any time frame *i*, the sampled speech signal which is given by the vector X_i of length L_F can be approximated as

$$X_i \approx C_i E_K S_R \tag{4.1}$$

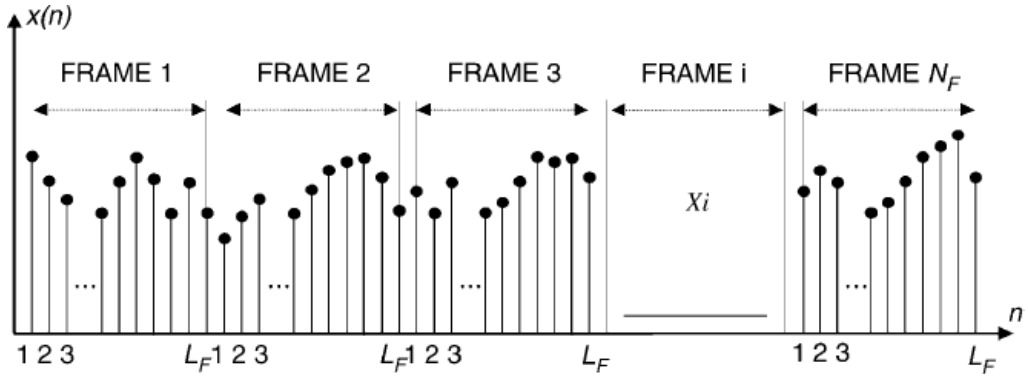

Figure 4. 1: A discrete signal's segmentation frame by frame **(Yarman, Güz, & Gürkan, 2006)**.

where,

- C_i is a real constant and it is called the gain factor.
- $K \in \{1, 2, ., ., ., N_E\}$ and $R \in \{1, 2, ., ., ., N_S\}$, all variables *K, R,* N_E and N_S are integers.
- $S_R^T = [S_{R1} \quad S_{R2} \quad \cdots \quad S_{RL_F}]$ is the Signature Vector which is generated utilizing the statistical behavior of the speech signals and the term C_iS_R contains almost full energy of X_i in the least mean square (LMS) sense.
- E_K is a $(L_F \times L_F)$ diagonal matrix with the formula $E_K = \text{diag}[e_{k1} \quad e_{k2} \quad \cdots \quad e_{kL_F}]$ and works as an envelope term on the quantity C_iS_R which also reflects the statistical features of the speech signal under consideration. PSS and PES contain S_R and E_K, which are correctly extracted.

4.1.1 Generation of the "Predefined Signature Sequences (PSS)" and "Predefined Envelope Sequences (PES) and, Synthesis Process of Speech Signal"

In order to create PSS and PES, First the total number of frames is calculated taking into account N and L_F being an integer so that N_F is also an integer.

$$N_F = N \setminus L_F \tag{4.2}$$

Then, the speech signal is divided into samples, as illustrated in the Figure 4. 1 with the following formula

$$x(n) = \sum_{i=1}^{N} x_i \delta_i(n-i) \tag{4.3}$$

In

(4.3), the unit sample and amplitude of the *ith* sample of speech signal are represented as $\delta_i(n)$ and xi, respectively. Also, $x(n)$ can be shown in vector forms like below. Here, X is called as main frame vector (MFV).

$$X^T = [x(1)\ x(2) \bullet \bullet \bullet\ x(N)] = [x1\ x2 \bullet \bullet \bullet\ x_N] \tag{4.4}$$

Moreover, MFV is divided into frames of equal length, with 16, 24, 64, or 128 samples, for example (Yarman, Güz, & Gürkan, 2006). The frame vectors are used to obtain the Main Frame Matrix, which is denoted by M_F as given below equation (4.5).

$$M_F = \left[X_1 X_2 X_3 \cdots X_{N_F}\right] \tag{4.5}$$

where,

$$X_i = \begin{bmatrix} x_{(i-1)}L_F + 1 \\ x_{(i-1)}L_F + 2 \\ \vdots \\ x_i L_F \end{bmatrix}, \qquad i = 1,2,3,\cdots,N_F \tag{4.6}$$

Over and above, each frame sequence or vector Xi can be spanned by the orthonormal vectors $\{\phi_{ik}\}$ in a vector space like (4.7),

$$X_i = \sum_{k=1}^{L_f} c_k \phi_{ik}, \qquad k = 1,2,\cdots,L_F \tag{4.7}$$

In (4.7) c_k is the frame coefficients, and they are calculated as follows;

$$c_k = \phi_{ik}^T X_i, \qquad k = 1,2,\cdots,L_F \tag{4.8}$$

By the way, to calculate $\{\phi_{ik}\}$, the eigenvectors of the frame correlation matrix R_i are used with following equation (4.9).

$$R_i = E[X_i X_i^T] \begin{bmatrix} r_i(1) & r_i(2) & r_i(3) & \cdots & r_i(L_F) \\ r_i(2) & r_i(1) & r_i(2) & \cdots & r_i(L_F-1) \\ r_i(3) & r_i(2) & r_i(1) & \cdots & r_i(L_F-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_i(L_F) & r_i(L_F-1) & r_i(L_F-2) & \cdots & r_i(1) \end{bmatrix} \tag{4.9}$$

$$r_i(d+1) = \frac{1}{L_F} \sum_{j=[(i-1)L_F+1]}^{[iL_F-d]} x_j x_{j+d}, \quad d = 0,1,2,\cdots,L_F-1 \tag{4.10}$$

In (4.9), expected value of random variables represented with $E[\cdot]$. And previous studies (Yarman, Güz, & Gürkan, 2006) (Güz, Gürkan, & Yarman, 2007), showed that the non-negative eigenvalues (λ_{ik}) provide the following equation (4.11).

$$R_i \phi_{ik} = \lambda_{ik} \phi_{ik} \tag{4.11}$$

After that, the eigenvalues (λ_{ik}) are sorted and the eigenvector are created with maximum eigenvalues of each frame. Also, it can be concluded from these studies (Yarman, Güz, & Gürkan, 2006) (Güz, Gürkan, & Yarman, 2007), eigenvectors' first element (ϕ_{i1}) has the highest energy level of *ith* frame which is why ϕ_{i1} is represented as a signature vector and they have most useful information about the original speech signal to reconstruct them. Thus,

$$X_i \cong c_k \phi_{i1} \tag{4.12}$$

After the equation (4.12) is obtained, X_i can actually be determined with E_i which is a diagonal matrix for each frame with given in equation (4.13) and diagonal entries of the matrix (e_{ir}) are computed with the signature vector and frame vector's entries like in equation (4.14).

$$X_i = C_i E_i \phi_{i1} \tag{4.13}$$

$$e_{ir} = \frac{x_{ir}}{C_i \phi_{i1r}} \qquad (r = 1,2,3,\cdots, L_F) \tag{4.14}$$

After these computations are applied several speech signals frame by frame, and the similar patterns are eliminated with the Pearson correlation formula which is given in equation

(4.15), as in (Güz, Gürkan, & Yarman, 2007).

$$\rho_{YZ} = \frac{\sum_{i=1}^{L} (y_i z_i) - [\sum_{i=1}^{L} y_i \sum_{i=1}^{L} z_i]/L}{\sqrt{[\sum_{i=1}^{L} y_i^2 - (\sum_{i=1}^{L} y_i)^2/L][\sum_{i=1}^{L} z_i^2 - (\sum_{i=1}^{L} z_i)^2/L]}} \tag{4.15}$$

In equation

(4.15), ρ_{YZ} indicates the Pearson Correlation Coefficient (PCC). $Y = [y_1 \quad y_2 \quad \cdots \quad y_{L_F}]$ and $Z = [z_1 \quad z_2 \quad \cdots \quad z_{L_F}]$ are two sequences that will be compared. For the elemination process, PCC (ρ_{YZ}) is investigated. For instance, if ρ_{YZ} is equal to one ($\rho_{YZ} = 1$) then it can be said that two compared vectors are same. On the contrary, if ρ_{YZ} is equal to zero ($\rho_{YZ} = 0$) then it implies that these two vectors are uncorrelated. In this approach, compared vectors are assumed equal when $0.9 \leq \rho_{YZ} \leq 1$.

As a result, comparable signature and envelope sequence patterns are removed, leaving only unique signature and envelope sequences. The rest of the unique patterns create the so-called Predefined Signature Sequence set $PSS = \{S_R; R = 1, 2, \cdots, N_S\}$ and Predefined Envelope Sequence set $PES = \{E_K; K = 1, 2, \cdots, N_E\}$. Here N_S and N_E represent the total number of unique signature and envelope patterns, respectively. The sizes of the $S = \{S_R\}$ and $E = \{E_K\}$ are very important for searching time to determine optimum encoding parameters (C_i, *R, and K*) and achieve high compression ratio. In the classical SYMPES method, these sets are shared at the transmitter and receiver

parts of the communication system and this model results in substantial bandwidth reduction in transmission of the data. As a result, we can deduce that all these processes are shaped by the SYMPES approach.

Any speech signal can be reconstructed frame by frame once PSS and PES have been produced via $X_i = C_i E_K S_R$. To reconstruct the *ith* frame, the gain coefficient C_i and, the index numbers S and, K of S_R and E_K are extracted from PSS and PES, respectively are used. Details of the reconstruction process of the speech signals and steps are given below.

- Speech signal X is divided into frames X_i of length L_F. The main frame vector represents the original speech in this situation such as $M_F = [X_1 X_2 X_3 \cdots X_{N_F}]$ (4.5).
- To minimize the total error or distance, an appropriate signature vector S_R is drawn from the PSS for each frame i. For instance, the index number R is in $\tilde{R} = 1,2, \dots, R, \dots, N_S$ and it provides minimum error $\delta_R = min\{\|X_i - C_{\tilde{R}} S_{\tilde{R}}\|^2\} = \|X_i - C_R S_R\|^2$ (Yarman, Güz, & Gürkan, 2006) (Güz, Gürkan, & Yarman, 2007). Finally, the index number R which provides minimum error and indicates to S_R, in this scenario, is stored. Moreover, speech frame can be written as $X_i \approx C_R S_R$.
- Similar operations are applied as in the previous explanation about the index number R of the signature vector. In this time, convenient envelope sequence E_K is drawn from PES to make error or in other word distance is minimum for all $\tilde{K} = 1,2, \dots, K, \dots, N_E$. Thus, the index number K provides min error $\delta_R = min\{\|X_i - C_R E_{\tilde{K}} S_R\|^2\} = \|X_i - C_R E_K S_R\|^2$ and it is also stored to reconstruction process.
- After all these computations, since S_R and E_K are the best representation for frame X_i, it can be described by $X_i \approx C_R E_K S_R$ in a convenient way.
- The last step for the reconstruction process is to determine new gain factor C_i once the best E_K and S_R are obtained. A new gain factor can be found by (4.16) to reduce the distance between the vectors X_i and $C_K E_K S_R$ even further in the LMS sense.

$$C_i = \frac{(E_K S_R)^T X_i}{(E_K S_R)^T (E_K S_R)} \tag{4.16}$$

- At that rate, the global minimum error which is given in equation (4.18) is obtained and the frame sequence is approximated by $X_{Ai} = C_i E_K S_R$.

$$\delta_{\text{Global}} = \|X_i - C_i E_K S_R\|^2 \quad (4.17)$$

- Finally, to reconstruct speech signal as $x_A(n) \approx x(n)$, approximated frame vectors (X_{Ai}) are stored under the approximated main frame matrix
  $M_{AF} = [X_{A1} X_{A2} X_{A3} \cdots X_{AN_F}]$.

To sum up, the traditional SYMPES technique is divided into three stages: (1) Generate the codebook, (2) Encode the speech signal to be reconstructed by constructing the predefined signature and envelope vectors and obtaining the best encoding parameters, (3) Decode (reconstruction) the speech signal using the encoding parameters from the codebook which is already located in the receiver part.

4.2 Zero-Cross and Phoneme-based SYMPES

4.2.1 Production of the Codebook for ZC and Phoneme-based SYMPES

For both the codebook and the speech signal to be rebuilt according to the zero-cross lengths defined for each phoneme in the newly recommended approach, predefined signature and envelope vectors are produced in distinct frame lengths (Sisman, Gürkan, Güz, & Yarman, 2013).

All phonemes are split based on the ZC lengths for each voice file supplied by each speaker once phoneme level identification is acquired at the ASR (Automatic Speech Recognition) system's output. The segmentation stage determines the ZC lengths for each phoneme, which might range from 1 to 512.

The ZC lengths of all segments that match each phoneme are used to classify them. An average segment waveform is generated for each segment class from 1 to 512. In the last stage, signature and envelope vectors are calculated for each average segment waveform, and the codebook (predefined signature and envelope sequences based on

ZC lengths of the phonemes) is built. Codebook's producing process is depicted in Figure 4. 2.

The Turkish language contains 29 letters in total, including 21 consonants and 8 vowels (/a/,/e/,/ı/,/i/,/o/,/ö/,/u/,/ü/). Turkish contains 29 phonemes in general since it is a language that is written as it is spoken.

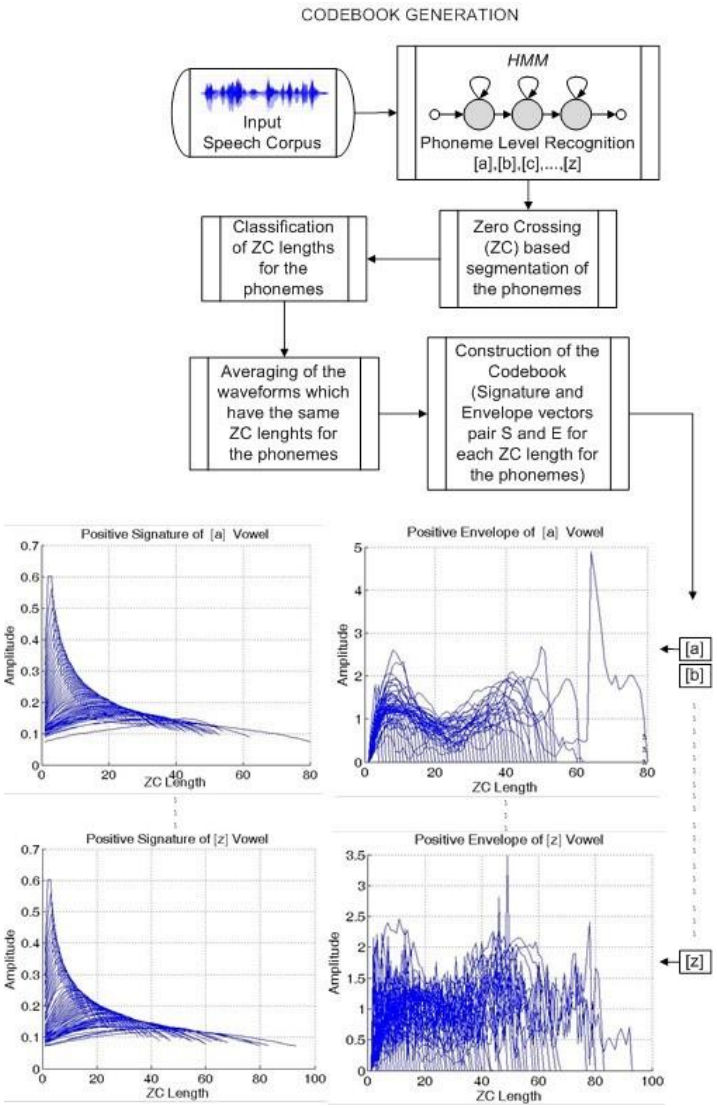

Figure 4. 2: Codebook Generation Process **(Sisman, Gürkan, Güz, & Yarman, 2013)**

4.2.2 Encoding Process of ZC and Phoneme-based SYMPES

In the encoding stage shown in Fig.4.3, the ASR system uses HTK (Hidden Markov Toolkit) to determine the phonemes and phoneme duration of the speech signal to be reconstructed. The zero cross lengths of all phonemes are used to partition them. Following the segmentation procedure, the current segment's zero-cross length is compared to the zero-cross lengths of similar phonemes in the codebook.

The SYMPES algorithm is used to identify the encoding parameters, gain factor G, and index number of the suitable signature and envelope vector combination based on a matching process for matched zero-cross lengths.

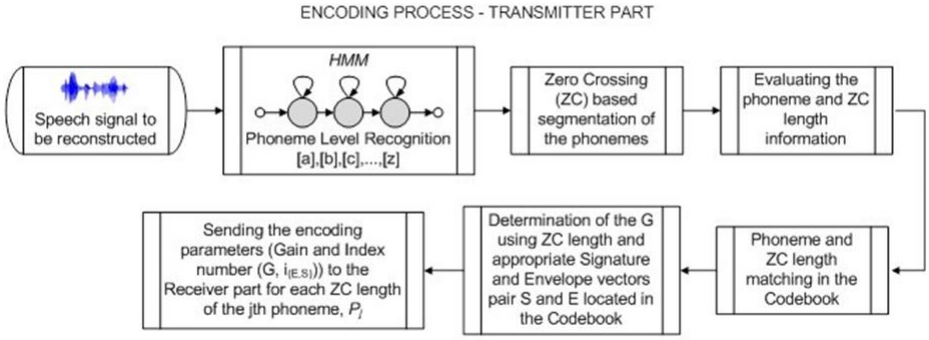

Figure 4. 3: The Transmitter Part's Encoding Process **(Sisman, Gürkan, Güz, & Yarman, 2013)**

4.2.3 Decoding Process of ZC and Phoneme-based SYMPES

The parameters that were received, gain factor *G*, for each ZC length of the phoneme P_j, the index number of the relevant signature and envelope vector pair supplied from the transmitter portion is utilized in the decoding stage to reconstruct the speech signal phoneme by phoneme using a mathematical model given by,

$$ZCL_{P_j}^{i} = (G \times E \times S)_{P_j}^{i} \tag{4.18}$$

The decoding process' scheme is displayed in **Figure 4. 4**

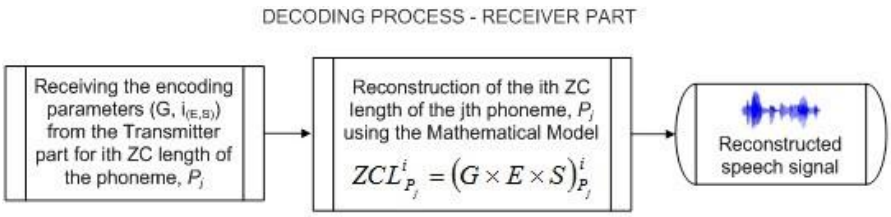

Figure 4. 4: The Receiver Part's Decoding Process (Sisman, Gürkan, Güz, & Yarman, 2013)

4.3 Proposed Approaches

4.3.1 Approach 1

In the first experiment, instead of generating a general codebook for all letters, in total 29 codebooks are generated for each Turkish letter. To construct these codebooks, each output assumed as a reliable ASR system of recorded speech files is used. These outputs include the specific letter and its duration according to the given input speech file. For each consonant, zero-crossing points and their indices are calculated. While the difference between these indices is not equal to one, envelope and signature vector pair and the gain coefficient for that part of the speech signal are computed. Hence, a codebook is generated when all phonemes are scanned by this method. Thus, the ZC length of each phoneme is computed and their envelope, signature vectors, and gain coefficients are added in the specific codebook according to ZC length.

Bit allocation table for the encoding parameters of the proposed algorithm is represented in Table 4. 1

Table 4. 1: Bit Allocation Table for Approach 1.

Coding Parameter	Number of bits
$b_{phoneme}$	5
b_G	5
b_{SE}	9
$b_{Total} = b_{phoneme} + b_G + b_{SE}$	19

In the Table 4. 1. $b_{phoneme}$, b_G, and b_{SE} are the numbers of the bits needed to represent the number of the phoneme ($2^{b_{phoneme}} \leq 29$), the gain coefficient G_i, and the index number of the signature and envelope pair (N_S,N_E) in the codebook, respectively.

While testing one of the speech signals, the other four speech files are used to construct the codebook. Therefore, in this experiment 3 groups of files are investigated and 3 different codebooks which belong to that specific letter are generated by averaging the result of their envelope, signature vectors, and gain coefficient values.

Comparison of the proposed method (Approach 1) with the CELP and classical SYMPES algorithm is given in Table 5. 3. Original and reconstructed speech signals via approach 1 for female and male speakers are illustrated in Figure 4. 5 and Figure 4. 6 respectively.

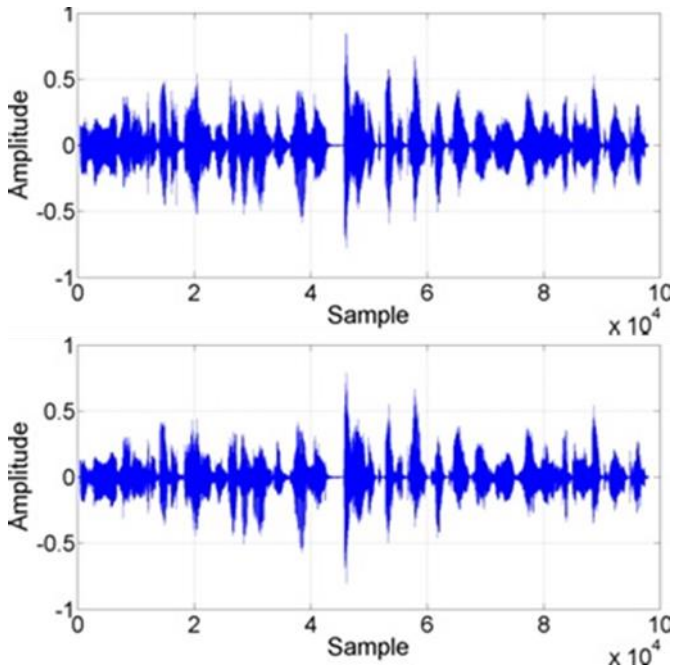

Figure 4. 5: Visual representation of original and reconstructed speech signals via approach 1 for Turkish female speaker.

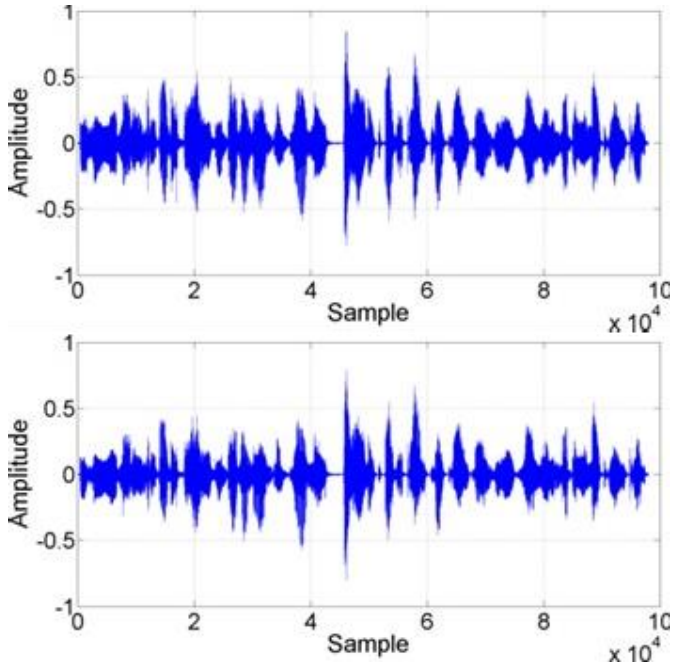

Figure 4. 6: Visual representation of original and reconstructed speech signals via approach 1 for Turkish male speaker.

4.3.2 Approach 2

In the second approach, 8 codebooks are constructed from 8 vowels. In addition to this 1 general codebook is generated using one consonant which represents all the 21 consonants. Bit allocation table for the encoding parameters of the proposed algorithm is characterized in Table 4. 2. Comparison of the proposed method (Approach 2) with the CELP and classical SYMPES algorithm is given in Table 5. 3. Original and reconstructed speech signals via approach 2 for female and male speakers are presented in Figure 4. 7 and Figure 4. 8 respectively.

Table 4. 2: Bit Allocation Table for Approach 2.

Coding Parameter	Number of bits
$b_{phoneme}$	3
$b_{U/V}$	1
b_G	5
b_{SE}	9
$b_{Total} = b_{U/V} + b_{phoneme} + b_G + b_{SE}$	18

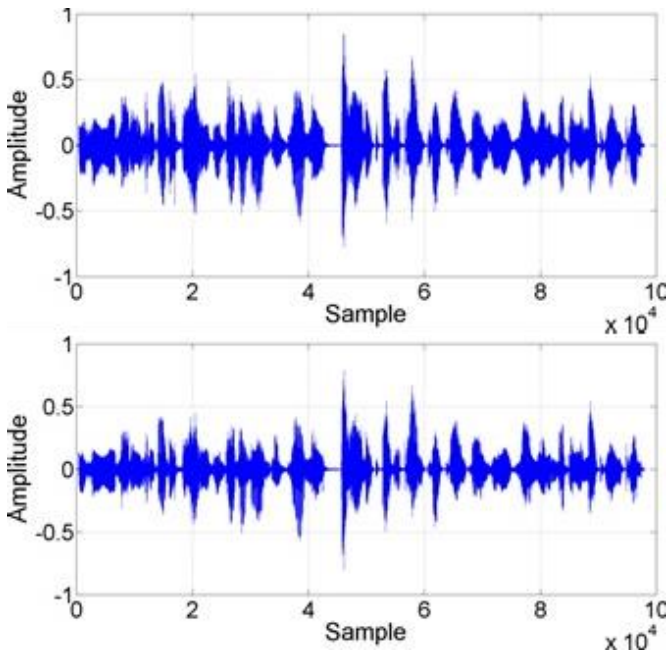

Figure 4. 7: Visual representation of original and reconstructed speech signals via approach 2 for Turkish female speaker.

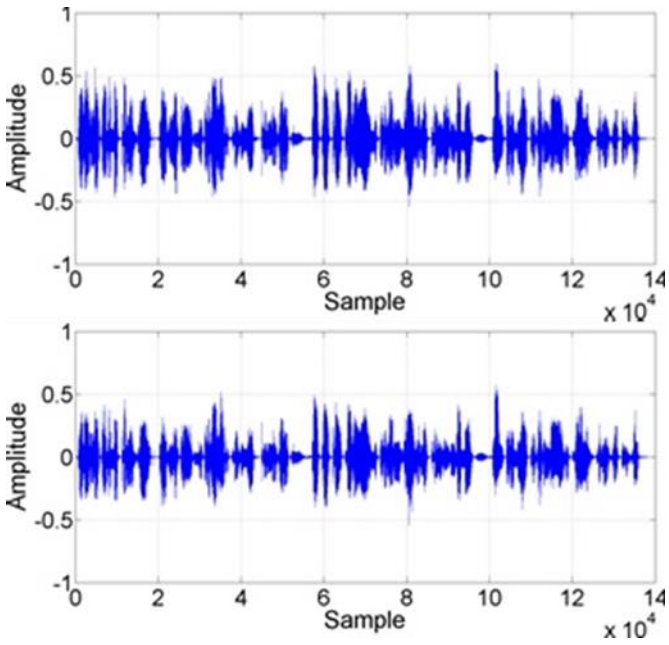

Figure 4. 8: Visual representation of original and reconstructed speech signals via approach 2 for Turkish male speaker.

4.3.3 Approach 3

In the third approach, 1 general codebook is created using one vowel which represents all the 8 vowels. In addition to this, another codebook is generated using one consonant which represents all the 21 consonants. Bit allocation table for the encoding parameters of the proposed algorithm is characterized in Table 4. 3.

Comparison of the proposed method (Approach 3) with the CELP and classical SYMPES algorithm is given in Table 5. 3. Original and reconstructed speech signals via approach 3 for female and male speakers are depicted in Figure 4. 9 and Figure 4. 10 respectively.

Table 4. 3: Bit Allocation Table for Approach 3.

Coding Parameter	Number of bits
$b_{phoneme}$ (or $b_{U/V}$)	1
b_G	5
b_{SE}	9
$b_{Total} = b_{phoneme}$ (or $b_{U/V}$) $+ b_G + b_{SE}$	15

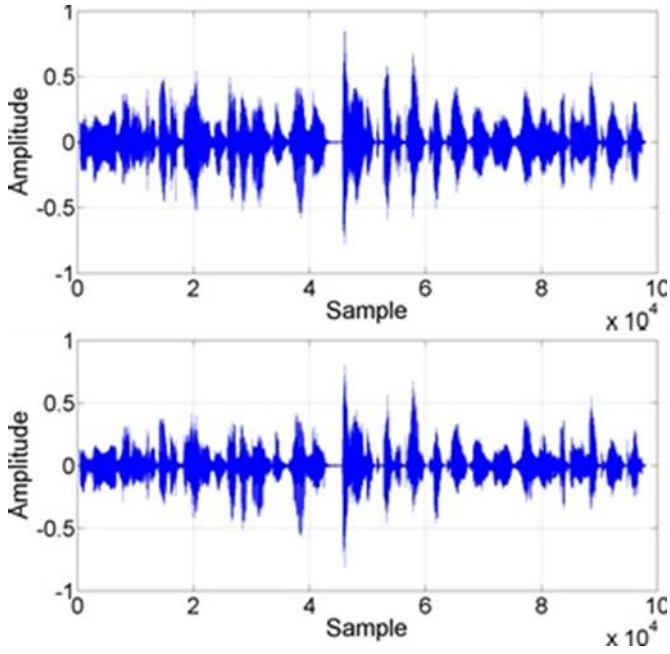

Figure 4. 9: Visual representation of original and reconstructed speech signals via approach 3 for Turkish female speaker.

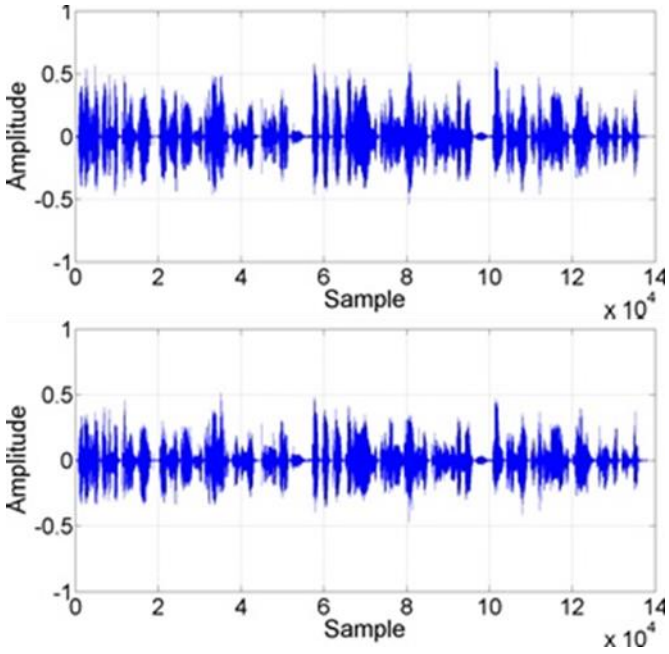

Figure 4. 10: Visual representation of original and reconstructed speech signals via approach 3 for Turkish male speaker.

4.3.4 Approach 4

In the fourth approach, the classical SYMPES method is used for all consonants and 1 general codebook is constructed for all vowels using zero cross and phoneme based SYMPES. In this approach, voiced parts of the speech signals are coded by zero-cross and phoneme-based SYMPES while unvoiced parts are coded by the classical SYMPES algorithm.

For voiced parts we use 1 bit for V/U decision ($b_{U/V}$ = 1 bit), 9 bits for predefined signature and envelope vectors based on zero cross and phoneme based SYMPES ($b_{U/V}$) = 9 *bits*) and, 5 bits for Gain factor (b_G = 5 *bits*). In this case, b_{Total} = $b_{phoneme}$ (or $b_{U/V}$) + b_G + b_{SE} = 1 + 9 + 5 = 15 *bits* are needed to code the voiced parts.

In order to code the unvoiced parts, we use the classical SYMPES algorithm which exploits fixed frame lengths instead of zero cross and phoneme-based frames. Therefore, the bit allocation table will be different than stated above. In this case, 1 bit for V/U decision ($b_{U/V}$ = 1 *bit*), 5 *bits* for gain factor (b_G = 5 *bits*) and a certain number of bits which is required for the number of predefined signature and envelope vectors in the codebooks determined by the ZC and phoneme-based SYMPES.

Bit allocation table for the encoding parameters of the proposed algorithm is characterized in Table 4. 4 and Table 4. 5. In the classical SYMPES codebook generation step, n' letter is used to generate the codebook which represents all the consonants. Several ways are used depending on the frame length, the number of signatures, and the number of envelopes in this approach.

The purpose of approach 4 is that combines the benefits of the classical SYMPES method and ZC and phoneme-based SYMPES algorithm. The advantage of the classical SYMPES method is high (SNR_{seg}) values for the unvoiced parts of the reconstructed signals even at low bit rates. The disadvantage of the classical SYMPES method is relatively high computational complexity. On the other hand, ZC and phoneme-based SYMPES method has lower computational complexity and very good quality especially in the voiced parts of the reconstructed signals.

For the fourth approach, the overall results and some of the original and reconstructed speech signals are given according to the different variables such as N_F , N_S, N_E, etc., in together with the following Figure 5. 2 - Figure 5. 47 that are illustrated in Section 5.3.

Table 4. 4: Bit Allocation Table (for Voiced Parts) for Approach 4 (ZC and Phoneme Based SYMPES Part).

Coding Parameter	Number of bits
$b_{phoneme}$ (or $b_{U/V}$)	1
b_G	5
b_{SE}	9
$b_{Total} = b_{phoneme}$ (or $b_{U/V}$) $+ b_G + b_{SE}$	15

Table 4. 5: Bit Allocation Table (for Unvoiced Parts) for Approach 4 (Classical SYMPES Part).

Coding Parameter	Number of bits
$b_{phoneme}$ (or $b_{U/V}$)	1
b_G	5
b_S	$2^{b_S} \leq N_S$
b_E	$2^{b_E} \leq N_E$
$b_{Total} = b_{phoneme}$ (or $b_{U/V}$) $+ b_G + b_S + b_E$	1+5+$b_S + b_E$

CHAPTER 5

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1 Training and Testing Data Sets

In our experiments, we recorded the speech files from the Voice of America (VOA) Turkish Broadcast News (BN) Channel (http://voanews.com/turkish). We separated the data as the training data and the test data. The training data is used in the codebook generation process. The test data is used to evaluate the performance of the proposed method. The detailed information about the training data and test data are given in Table 5. 1 and Table 5. 2. The speech files consist of 6 speakers and 6 recorded sentences uttered by each speaker. The files are sampled at 16 Khz, 16 bits using Praat (Boersma & D., 2016).

We used Hidden Markov Model Toolkit (HTK) in order to constitute forced alignments (Boersma & D., 2016) (Young & Young, 1993) (Young, et al., 2006). First, we parameterized speech waveforms in terms of Mel-Frequency Cepstral Coefficients (MFCC) provided by the HCopy tool of HTK. In addition, a list of spoken words and dictionary is constructed from Segment Time Mark (STM) files, i.e. corresponding transcription files. HVite tool of HTK pro- vides forced alignments in 100 Nanoseconds sense based on words and phonemes by using time labeled MFCC vectors, corresponding recognized word list, and a lexicon. Visual representation of the Force alignment process is given in Figure 5. 1

Table 5. 1: Test Data of Proposed Approaches.

Test Data Durations		
Speaker No	Sentence	Duration of the file (sec)
1	1	4
1	2	4
1	3	7
1	4	8
1	5	2
1	6	10
2	1	9
2	2	3
2	3	3
2	4	4
2	5	4
3	1	4
3	2	5
3	3	4
3	4	3
3	5	4
4	1	6
4	2	3
4	3	4
4	4	7
4	5	4
5	1	6
5	2	3
5	3	5
5	4	6
5	5	3
5	6	3
5	7	6
6	1	11
6	2	8
6	3	3
6	4	6
6	5	7
6	6	7
Average		5,1765

Table 5. 2: Training Data for the Codebook Generation.

Training Data Duration			
	Duration of the file(sec)	Total number of Samples	Gender
Speaker 1	120:09:00	115.364.480	Female
Speaker 2	9:32:00	9.163.040	Male
Speaker 3	108:53:00	104.561.408	Male
Speaker 4	25:46:00	24.750.880	Male
Speaker 5	20:29:00	19.676.800	Female
Speaker 6	13:54:00	13.356.000	Male
Total:	298:43:00	286.872.608	

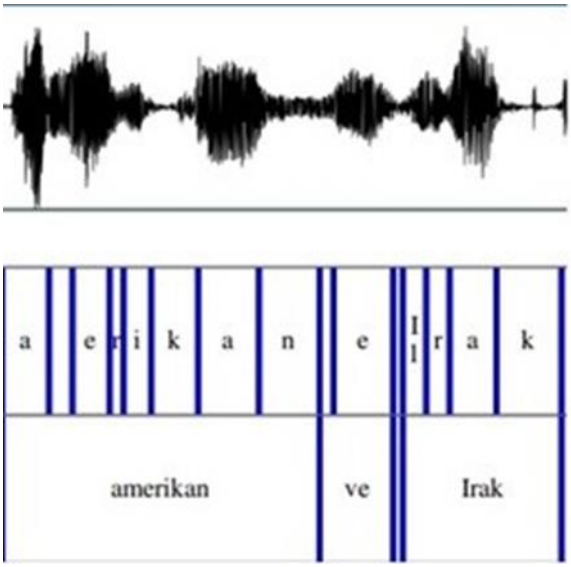

Figure 5. 1: Visual Representation of Forced Alignment Process.

5.2 Visual Representations of Reconstructed and Original Speech Signals' for Approach 4

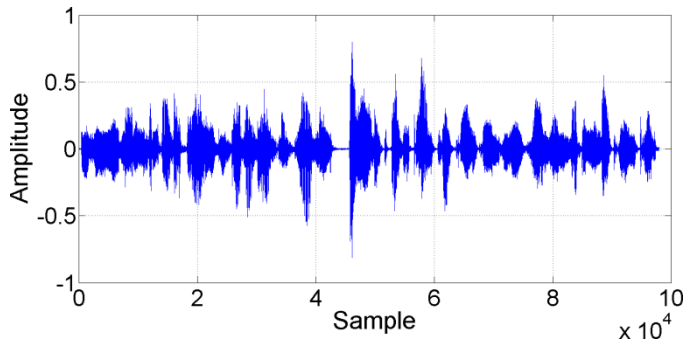

Figure 5. 2: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:8192.

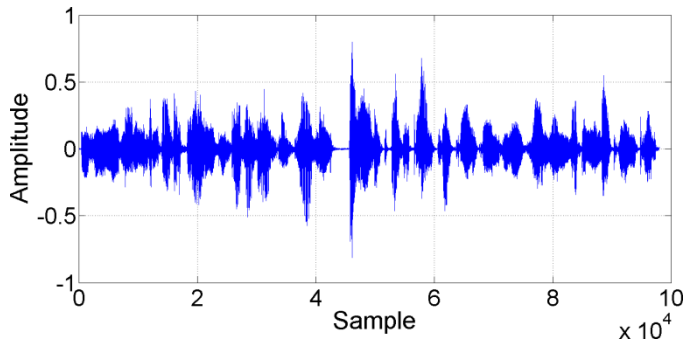

Figure 5. 3: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:4096.

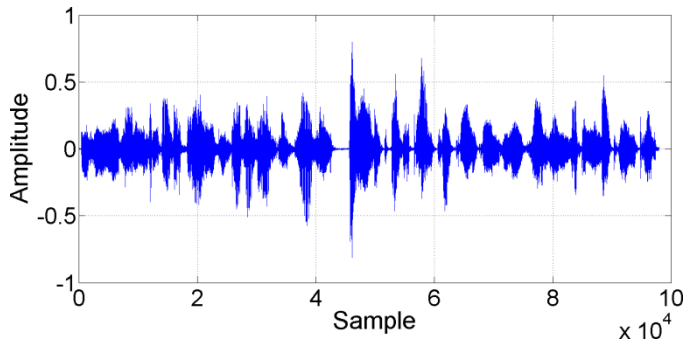

Figure 5. 4: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:2048.

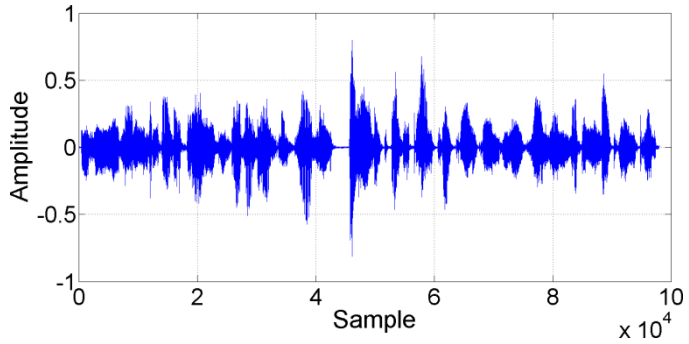

Figure 5. 5: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:1024.

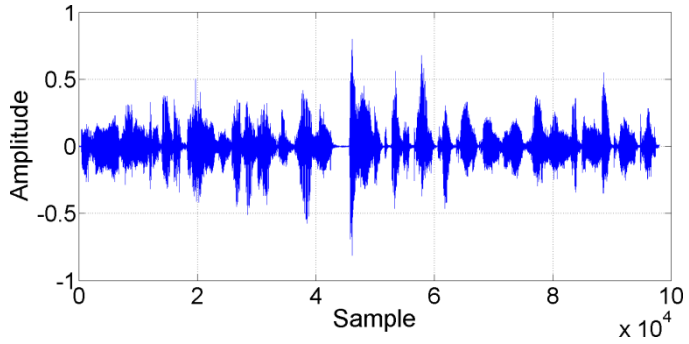

Figure 5. 6: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:512.

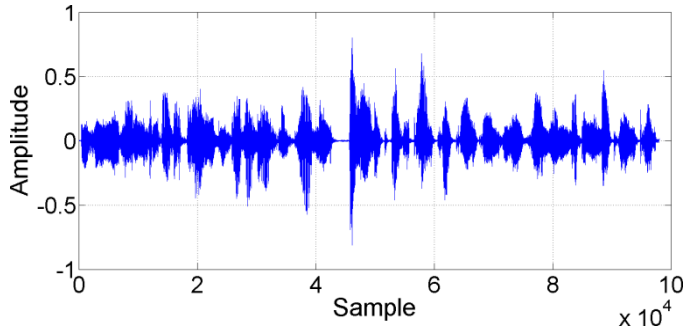

Figure 5. 7: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:32, number of Envelope:256.

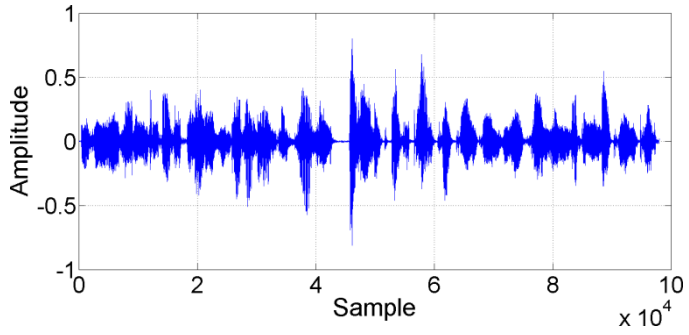

Figure 5. 8: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:16, number of Envelope:128.

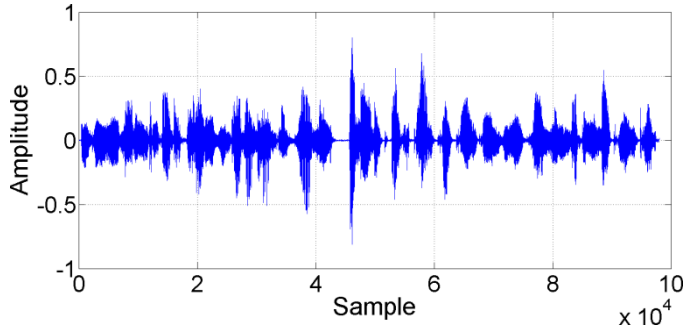

Figure 5. 9: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:16, number of Signature:8, number of Envelope:64.

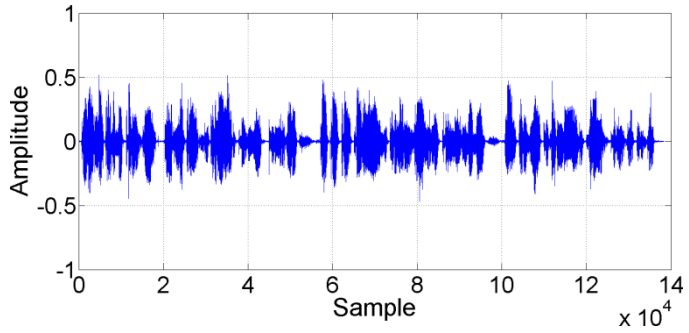

Figure 5. 10: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:8192.

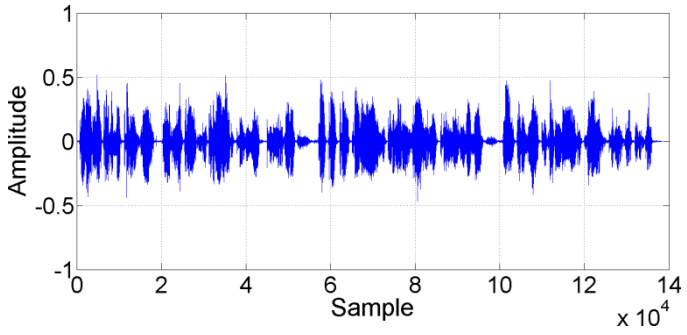

Figure 5. 11: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:4096.

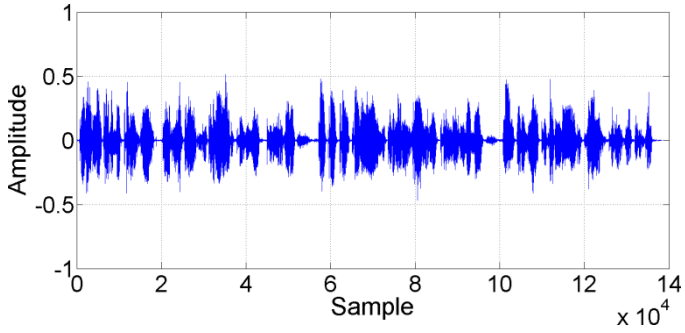

Figure 5. 12: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:2048.

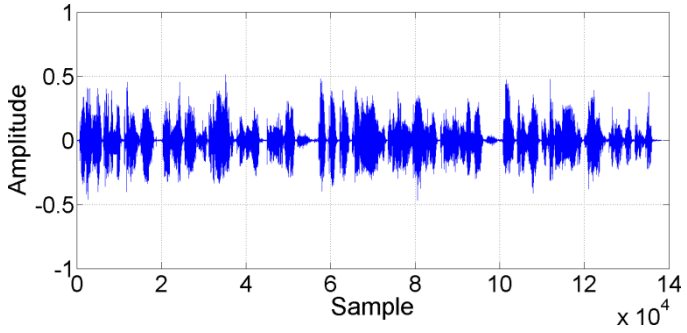

Figure 5. 13: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:1024.

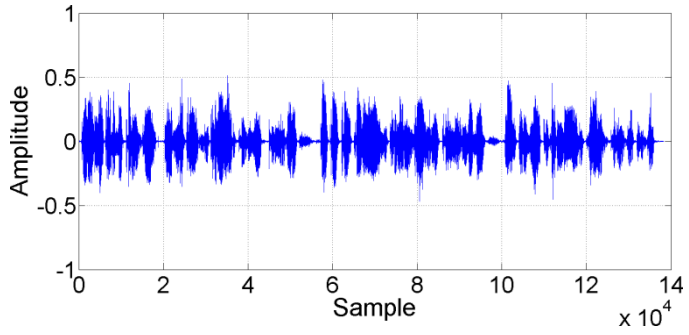

Figure 5. 14: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:512.

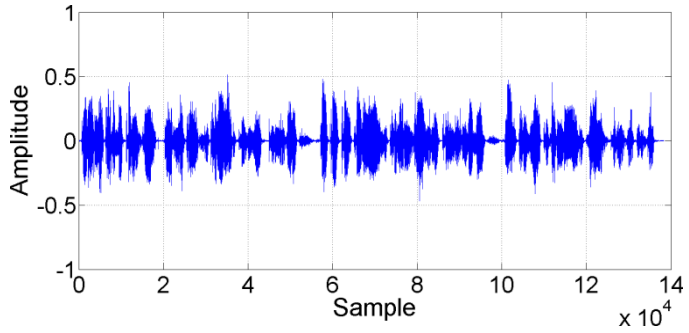

Figure 5. 15: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:32, number of Envelope:256.

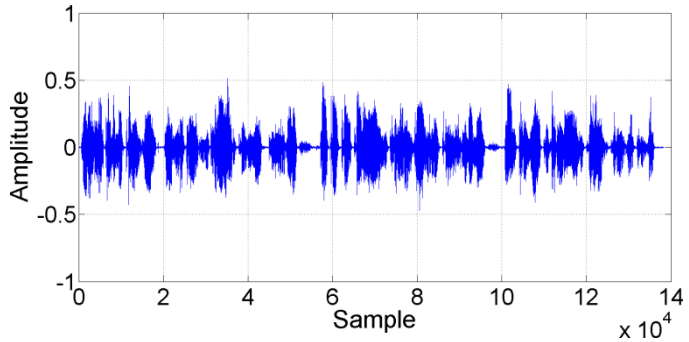

Figure 5. 16: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:16, number of Envelope:128.

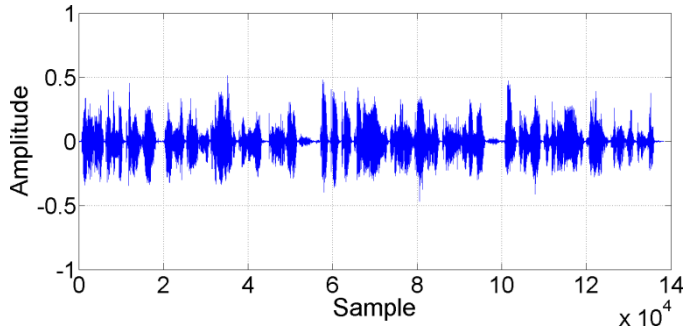

Figure 5. 17: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:16, number of Signature:8, number of Envelope:64.

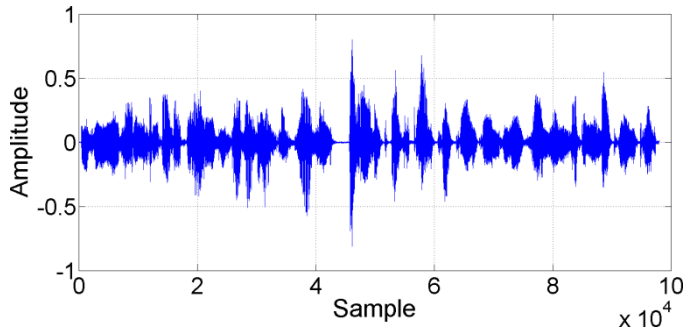

Figure 5. 18: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:8192.

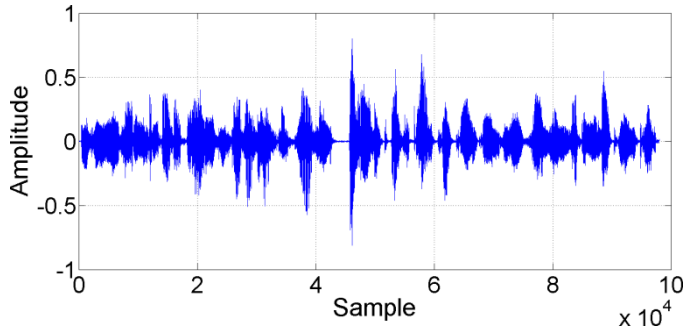

Figure 5. 19: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:4096.

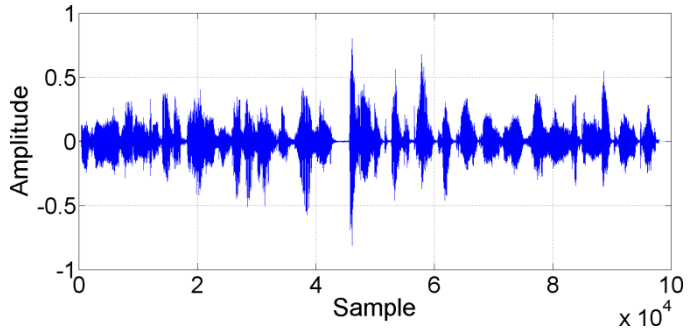

Figure 5. 20: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:2048.

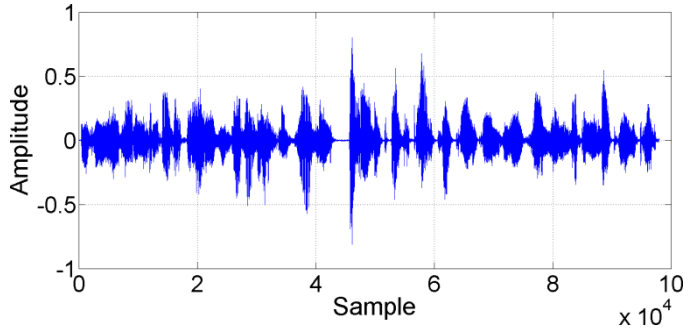

Figure 5. 21: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:1024**.**

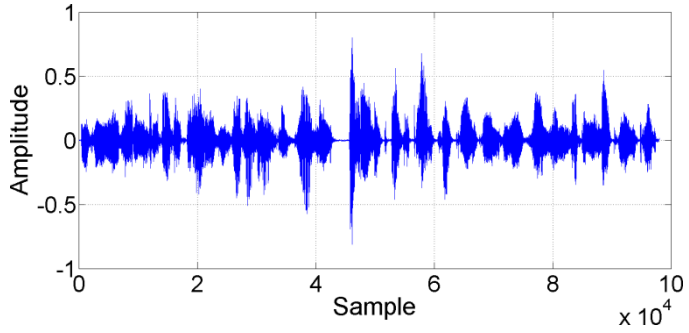

Figure 5. 22: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:512.

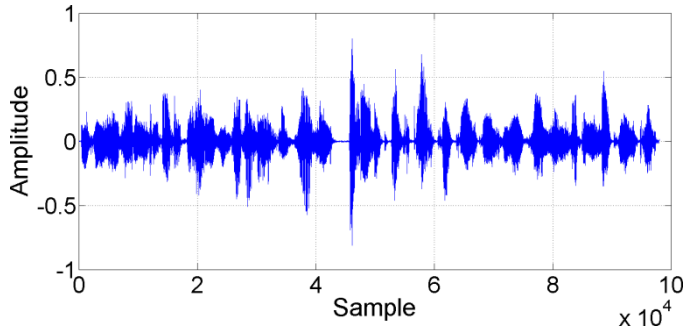

Figure 5. 23: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:32, number of Envelope:256.

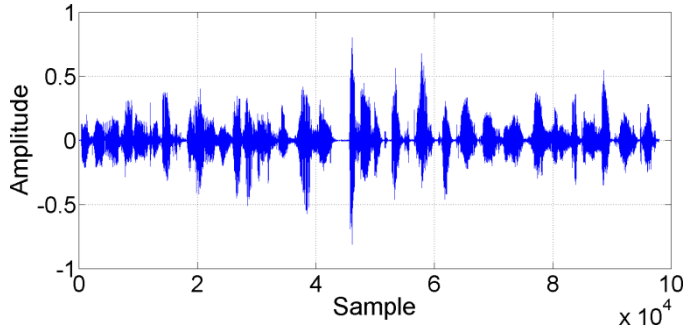

Figure 5. 24: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:16, number of Envelope:128.

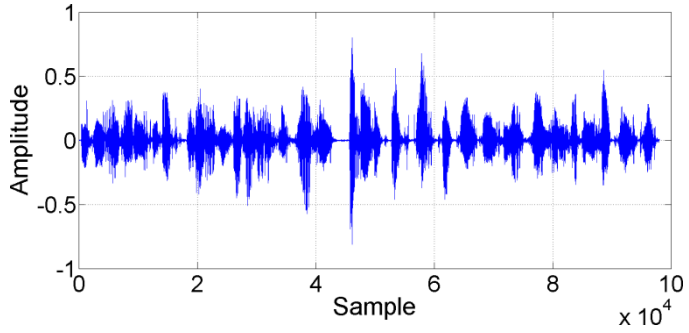

Figure 5. 25: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:32, number of Signature:8, number of Envelope:64.

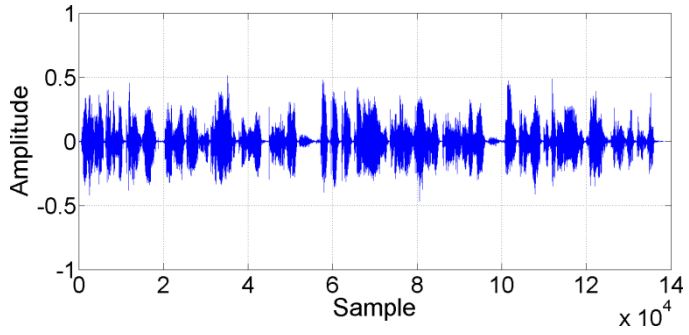

Figure 5. 26: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:8192.

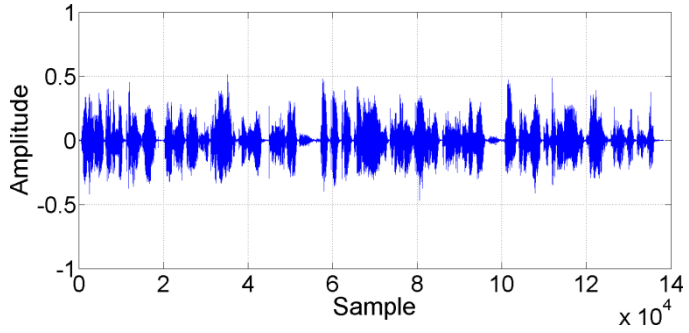

Figure 5. 27: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:4096.

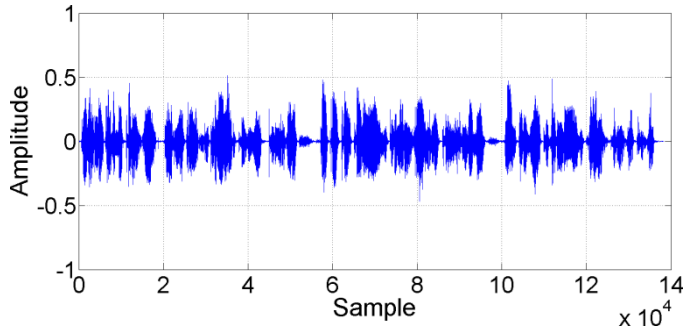

Figure 5. 28: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:2048.

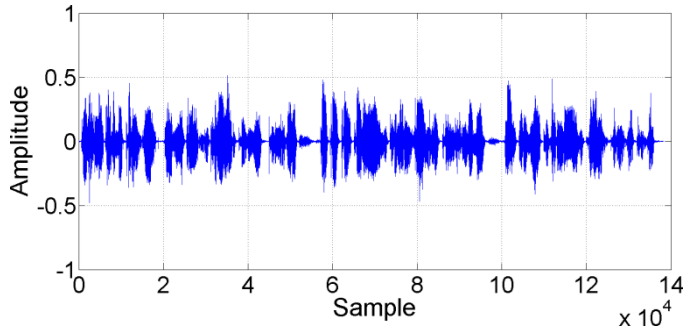

Figure 5. 29: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:1024.

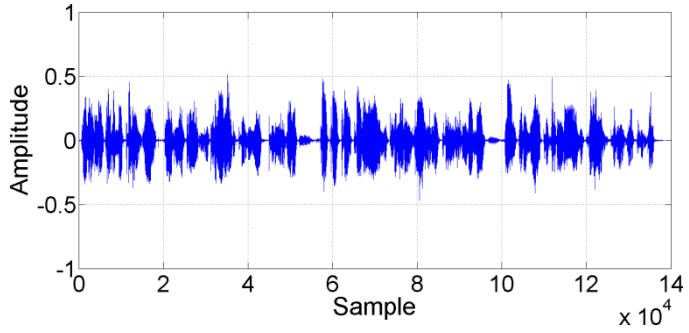

Figure 5. 30: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:512.

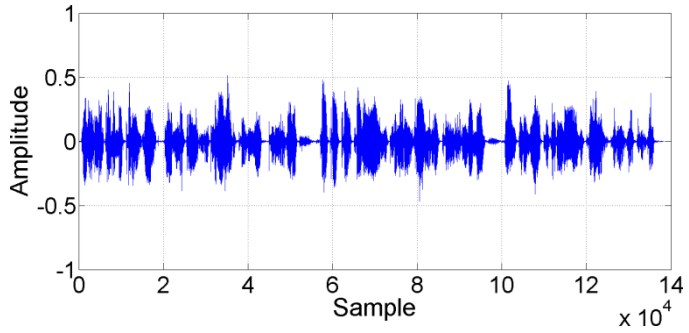

Figure 5. 31: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:32, number of Envelope:256.

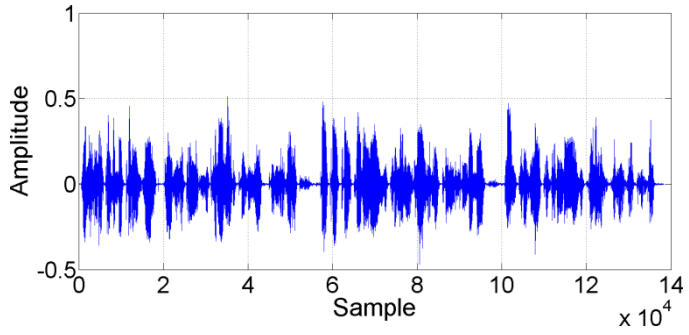

Figure 5. 32: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:16, number of Envelope:128**.**

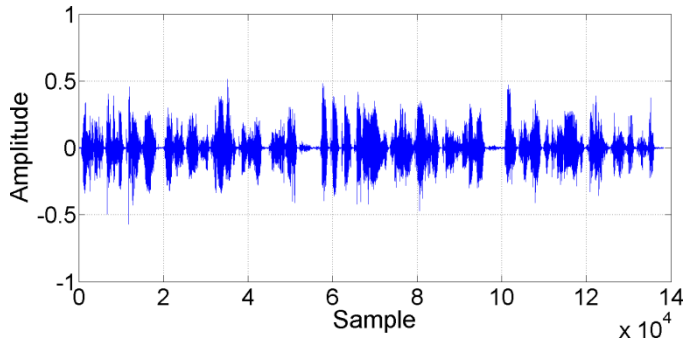

Figure 5. 33: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:32, number of Signature:8, number of Envelope:64.

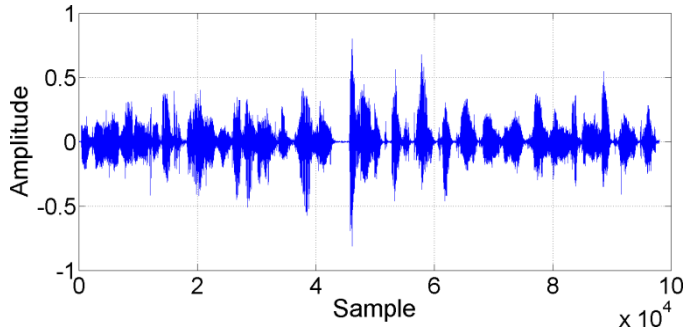

Figure 5. 34: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:32, number of Envelope:4096.

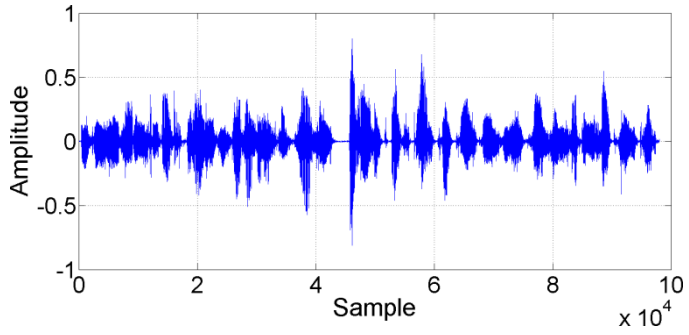

Figure 5. 35: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:32, number of Envelope:2048.

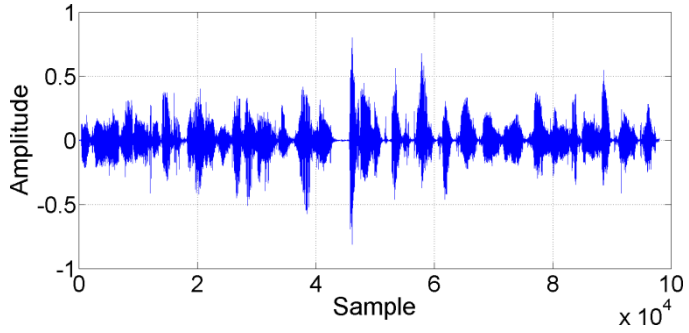

Figure 5. 36: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:32, number of Envelope:1024.

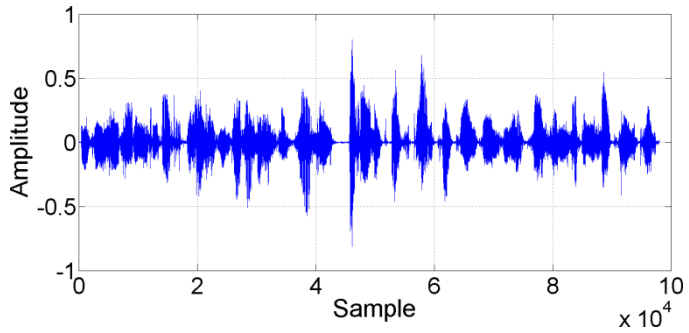

Figure 5. 37: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:32, number of Envelope:512.

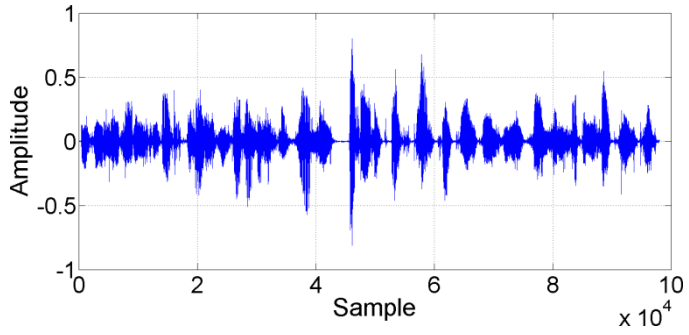

Figure 5. 38: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:32, number of Envelope:256.

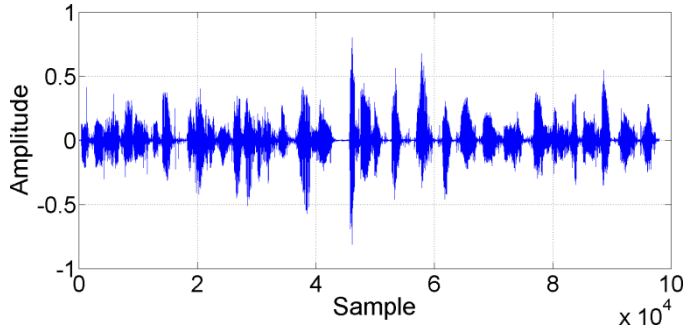

Figure 5. 39: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:16, number of Envelope:128.

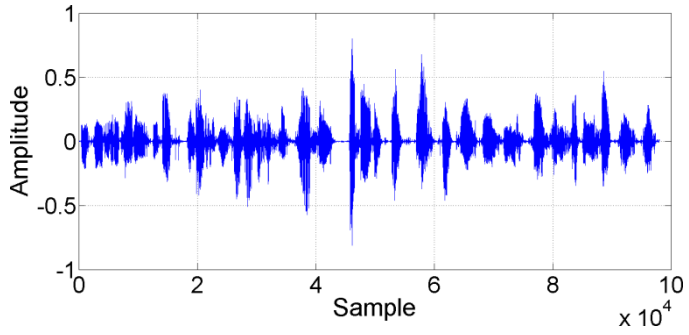

Figure 5. 40: Visual representation of reconstructed speech signals with Approach 4 for Turkish female, Frame length:64, number of Signature:8, number of Envelope:64.

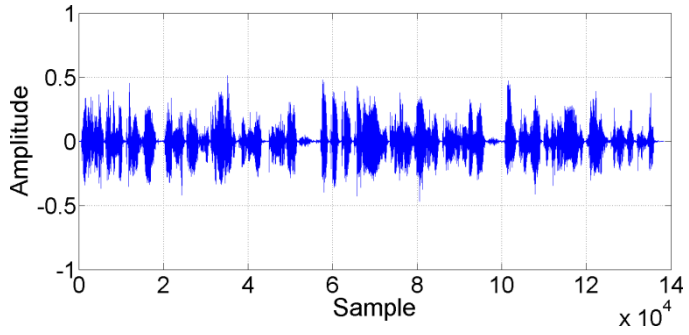

Figure 5. 41: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:32, number of Envelope:4096.

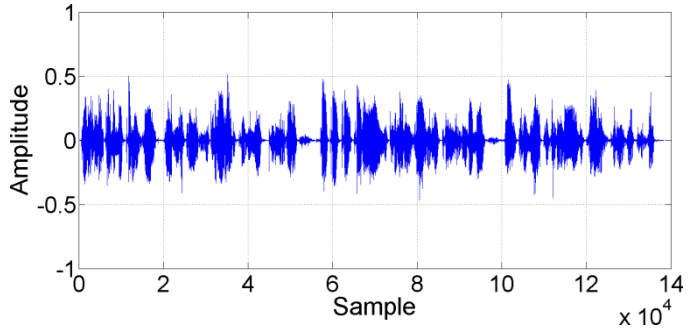

Figure 5. 42: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:32, number of Envelope:2048.

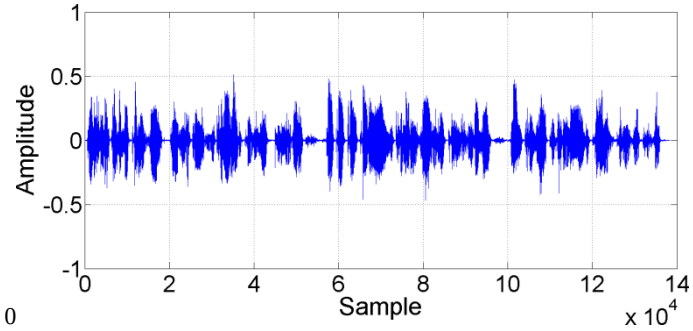

Figure 5. 43: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:32, number of Envelope:1024.

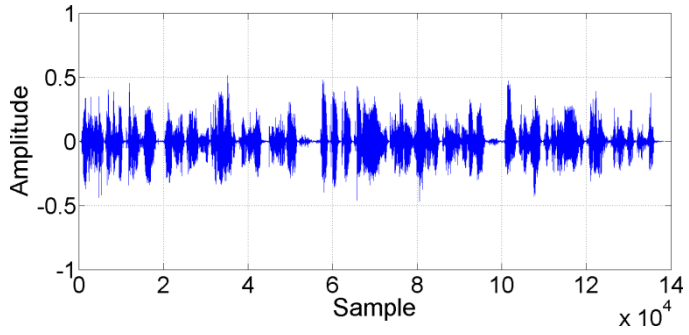

Figure 5. 44: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:32, number of Envelope:512.

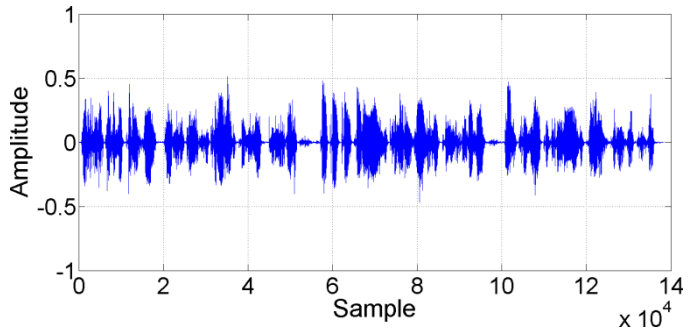

Figure 5. 45: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:32, number of Envelope:256.

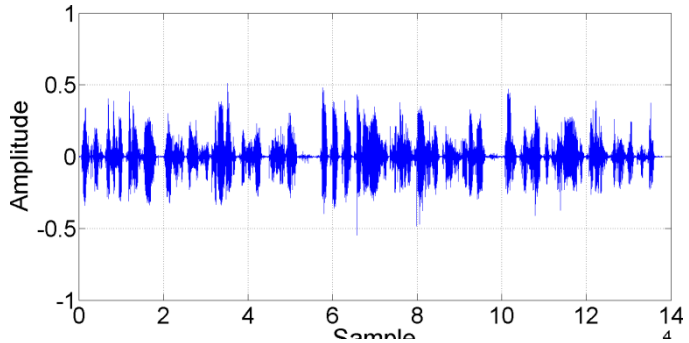

Figure 5. 46: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:16, number of Envelope:128.

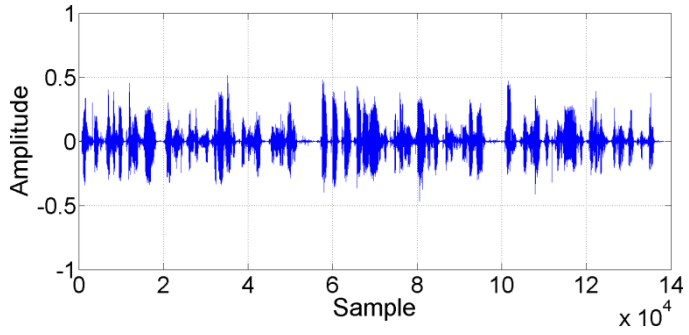

Figure 5. 47: Visual representation of reconstructed speech signals with Approach 4 for Turkish male, Frame length:64, number of Signature:8, number of Envelope:64.

5.3 Objective Experimental Results and Discussion of ZC and Phoneme-Based SYMPES Approaches

The results were created by visually and computationally comparing the newly proposed speech coding method with both the classical SYMPES method and the CELP algorithm. The figures of the original and reconstructed speech signals of both male and female speakers at different compression ratios are illustrated in this thesis. Therefore, as it can be easily seen from these figures, the differences in the compared signals are insignificant.

Table 5. 3: Comparison of the Proposed Methods with the CELP and the CLASSICAL SYMPES Algorithm.

ALGORITHM	Elapsed Time (sec) (Encoding+Decoding)	CR	Bit/sample	KBit/sec	Seg SNR
CELP (9 KBit/sec)	2, 8753	28, 44	0, 5625	9	3, 5164
CELP (16 KBit/sec)	2, 7764	16, 00	1, 0000	16	8, 5818
CELP (32 KBit/sec)	3, 3248	8, 00	2, 0000	32	8, 6139
SY MPES (9 KBit/sec − Framelength : 64)	214, 7783	28, 44	0, 5625	9	6, 2305
SY MPES (16 KBit/sec − Framelength : 32)	140, 8419	16, 00	1, 0000	16	8, 8829
SY MPES (32 KBit/sec − Framelength : 16)	380, 0377	8, 00	2, 0000	32	13, 4171
ZC and Phoneme based SY MPES (26 Kbit/sec) (Approach 1)	8, 4231	9, 91	1, 6459	26	9, 7997
ZC and Phoneme based SY MPES (25 Kbit/sec) (Approach 2)	7, 0872	10, 46	1, 5593	25	9, 6582
ZC and Phoneme based SY MPES (20, 8 Kbit/sec) (Approach 3)	6, 6018	12, 55	1, 2994	20, 8	9, 5533

In order to give an objective sum up of proposed approaches, Table.5.3 and 5.4 are made according to speech quality (SNRseg), compression ratio (CR), and elapsed time criterias at all bit rates, 9; 16; 32 Kbps.

Therefore, if Table.5.3 is investigated, following results can be concluded for Approach 1,2 and 3.

Although the compression ratio of Approach 1 and Approach 2 are not the same (but similar) when comparing the other two methods (CELP and classic SYMPES), it can be said that they have a better SNR_{seg} value than the CELP method and, they also provide a much lower elapsed time opposite of the classic SYMPES algorithm which gains an advantage over Approach 1 and Approach 2 for SNRseg.

- Although, Approach 2 is slightly better than Approach 1 in terms of elapsed time and compression ratio, it worse in terms of SNRseg. That's why there is a tradeoff between the transmission speed and the hearing quality of the speech signal since higher SNRseg value provides better speech quality and minimizing the elapsed time lead to transmission be fast.
- On the other hand, when Approach 3 is compared with other methods in table 5.3, it shows real supremacy over all the previous methods in terms of all evaluation conditions.
- In Approach 4, the zero-cross and phoneme-based SYMPES method which has lower computational complexity and very good quality especially in the voiced parts of the reconstructed signals and, the classical SYMPES method which has high SNRseg values for the unvoiced parts of the reconstructed signals even at low bit rates, but relatively high computational complexity are used to reconstruct the speech signal.
- The advantages of these methods are combined in Approach 4.
- It can be deduced from Table.5.4, the Approach 4 performs well than CELP, classic SYMPES, and all other three zero-cross and phoneme-based approaches, because the (SNRseg) value remains at the same levels (SNRseg = 10.35) while the CR is maximum level (CR = 24.4).
- Also, the elapsed time of Approach 4 is comparable to the CELP for transmitting speech signals fast.

Table 5. 4: Overall Results of the Approach 4.

N_F	N_S	N_E	Encoding Time	Decoding Time	CR	Bit/sample	Bit/sec	SegSNR	SNR
	32	8192	0.8841	323.342	113.756	14.095	225.513.369	105.253	89.065
	32	4096	0.8501	178.520	116.309	13.786	220.578.120	105.197	88.657
	32	2048	0.8627	107.351	119.105	13.464	215.417.281	105.134	87.958
16	32	1024	0.8570	71.065	122.180	13.123	210.014.840	104.979	86.789
	32	512	0.8561	53.247	125.579	12.772	204.353.216	104.860	85.207
	32	256	0.8562	44.093	129.356	12.401	198.413.058	104.740	83.200
	16	128	0.8815	39.542	138.325	11.601	185.609.453	104.068	68.424
	8	64	0.8850	37.229	149.857	10.713	171.403.993	104.025	67.278
	32	8192	0.9247	185.638	152.529	10.554	168.866.340	104.662	78.969
	32	4096	0.9071	109.971	155.966	10.324	165.185.606	104.617	78.410
	32	2048	0.9046	71.894	159.730	10.083	161.335.933	104.574	77.689
32	32	1024	0.9049	53.186	163.870	0.9832	157.305.281	104.474	76.343
	32	512	0.9046	43.215	168.447	0.9568	153.080.423	104.408	74.906
	32	256	0.9093	38.803	173.532	0.9290	148.646.795	104.291	72.583
	16	128	0.8951	35.689	185.608	0.8693	139.087.203	103.732	56.399
	8	64	0.8998	34.488	201.135	0.8030	128.475.789	103.579	51.953

Tablo 5. 4: (continued.)

N_F	N_S	N_E	Encoding Time	Decoding Time	CR	Bit/sample	Bit/sec	SegSNR	SNR
	32	4096	0.9032	80.983	189.170	0.8564	137.031.400	104.202	69.239
	32	2048	0.9022	59.442	193.750	0.8366	133.848.136	104.166	68.435
	32	1024	0.9045	47.051	198.789	0.8157	130.514.705	104.129	67.486
64	32	512	0.9043	41.441	204.358	0.7939	127.020.094	103.969	64.089
	32	256	0.9086	38.685	210.546	0.7710	123.352.172	103.853	60.867
	16	128	0.9067	36.882	225.243	0.7215	115.441.379	103.673	54.010
	8	64	0.9026	36.016	244.139	0.6666	106.656.677	103.577	50.370

5.4 Subjective Experiment (MOS Test) for Approach 4

Besides giving a result of computational aspects of the proposed method, MOS was also evaluated for subjective measurement to compare the new method (Approach 4) with the classical SYMPES and the CELP algorithm, two different original speech signals (female and male, respectively) were chosen from test data set and their three different reconstructed versions (with CELP or Classical SYM- PES or Approach 4) were selected according to their compression ratio ($CR \cong 16$ and $CR \in [24,28.44]$). Before a total of fourteen speech signals was randomly listened by fifty inexperienced listeners according to MOS listening test (ITU, 2016) to in a same suitable environment with using headphone, they were then instructed to pay close attention to the samples and consider distinctions between them when assigning evaluations based on their speech quality on a scale of 1 (Bad) to 5 (Excellent). To give a MOS, the results of the appraisal of fifty people were aver- aged and these results are illustrated in Table 5. 6 and for two different speech files, respectively. Then, the average evaluation of this measurement is given in Table 5. 5

Table 5. 5: Overall MOS results for Approach 4.

MOS Results for;	Original Speech File	CELP (CR=16)	Classic SYMPES (CR=16)	Approach 4 (CR=15.973)	CELP (CR=28.44)	Classic SYMPES (CR=28.44)	Approach 4 (CR=24.4139)
First Speech	4,714286	3,54	4,08	4,6	2,24	2,38	2,5
Second Speech	4,72	3,92	3,98	4,18	2,42	2,6	2,9
AVG:	4,717143	3,73	4,03	4,39	2,33	2,49	2,7

As a result of the MOS listening test (ITU, 2016), it can be easily deduced that the new approach is beaten the other conventional two methods once again concerning subjective measurement.

Table 5. 6: The MOS results for the first speech.

#of Listener	Original Speech File	CELP (CR=16)	Classic SYMPES (CR=16)	Approach 4 (CR=15.973)	CELP (CR=28.44)	Classic SYMPES (CR=28.44)	Approach 4 (CR=24.4139)
1	5	3	4	4	2	2	3
2	5	4	4	5	3	3	3
3	4	4	5	4	2	2	2
4	5	3	4	5	1	2	3
5	4	3	4	4	3	3	4
6	5	4	3	4	2	1	2
7	5	4	4	5	4	2	3
8	5	3	4	5	2	2	3
9	5	3	4	5	1	2	2
10	4	4	5	4	2	3	2
11	5	3	4	5	3	2	3
12	4	3	5	4	2	1	2
13	5	4	4	5	2	3	2
14	5	3	4	5	1	2	1
15	4	4	5	4	4	2	3
16	5	4	5	4	2	1	3
17	5	4	4	5	2	2	2
18	5	3	4	5	1	2	2
19	4	3	3	4	2	3	3
20	4	3	4	5	1	2	2
21	5	2	4	5	2	3	2
22	4	3	4	5	2	2	3
23	5	4	4	5	1	3	2
24	4	4	5	4	3	2	2

Tablo 5. 6: (continued.)

#of Listener	Original Speech File	CELP (CR=16)	Classic SYMPES (CR=16)	Approach 4 (CR=15.973)	CELP (CR=28.44)	Classic SYMPES (CR=28.44)	Approach 4 (CR=24.4139)
25	5	3	4	5	3	4	4
26	5	4	5	4	3	2	2
27	5	3	4	5	2	2	3
28	4	4	4	4	3	3	3
29	5	3	4	4	2	3	3
30	5	4	4	5	2	2	3
31	5	4	4	5	3	3	2
32	5	3	3	4	2	2	2
33	4	3	4	5	2	2	2
34	5	4	4	5	2	3	4
35	5	5	4	5	3	2	3
36	5	4	3	4	1	3	2
37	4	3	4	5	2	1	1
38	5	4	4	5	3	3	2
39	5	3	4	4	2	4	3
40	4	4	4	5	2	2	2
41	5	3	4	4	2	3	3
42	5	4	3	5	2	2	2
43	5	4	4	5	2	2	1
44	5	4	4	5	3	3	2
45	5	3	4	4	3	2	3
46	4	4	5	5	3	4	3
47	5	3	4	5	4	3	3
48	5	4	5	4	3	3	2
49	5	4	4	5	2	2	3
50	5	4	4	4	1	2	3
AVG:	4,7142857	3,54	4,08	4,6	2,24	2,38	2,5

Table 5. 7: The MOS Results for the second speech.

#of Listener	Original Speech File	CELP (CR=16)	Classic SYMPES (CR=16)	Approach 4 (CR=15.973)	CELP (CR=28.44)	Classic SYMPES (CR=28.44)	Approach 4 (CR=24.4139)
1	5	4	4	5	2	3	3
2	4	3	4	4	3	2	3
3	5	4	5	4	2	3	3
4	5	4	3	4	1	2	2
5	4	3	4	5	3	2	3
6	5	4	4	4	2	3	4
7	4	3	3	3	2	2	3
8	5	4	4	4	2	2	2
9	5	5	5	4	2	3	3
10	5	5	4	3	3	2	2
11	5	4	5	3	2	2	3
12	4	5	4	4	3	3	2
13	5	4	3	4	2	2	3
14	4	4	4	4	3	2	3
15	5	5	5	4	3	3	3
16	5	4	5	5	3	2	2
17	5	5	4	4	2	2	2
18	4	4	5	4	3	3	2
19	4	4	5	5	4	4	3
20	5	3	4	4	3	3	3
21	5	4	3	4	2	3	2
22	4	5	4	5	3	4	3
23	5	4	4	4	2	3	3
24	5	4	4	5	3	2	2

Tablo 5. 7: (continued.)

#of Listener	Original Speech File	CELP (CR=16)	Classic SYMPES (CR=16)	Approach 4 (CR=15.973)	CELP (CR=28.44)	Classic SYMPES (CR=28.44)	Approach 4 (CR=24.4139)
25	5	3	4	4	2	3	3
26	5	4	3	4	2	2	3
27	4	5	4	5	3	3	2
28	5	4	3	4	2	3	4
29	5	3	4	4	1	3	3
30	4	4	4	4	3	2	3
31	5	4	3	4	2	3	4
32	5	3	4	5	2	3	3
33	4	4	4	4	1	2	3
34	5	4	3	5	2	2	3
35	4	3	4	4	3	2	3
36	5	4	4	3	1	2	3
37	4	4	3	4	2	3	4
38	5	3	4	4	3	2	3
39	4	4	3	4	3	2	3
40	5	3	4	4	3	2	3
41	5	4	4	3	2	2	2
42	5	4	5	4	3	3	4
43	5	5	4	5	3	2	3
44	5	4	4	5	2	3	3
45	5	4	4	5	2	3	3
46	5	4	4	4	2	4	4
47	5	3	5	5	3	3	4
48	5	4	4	4	3	3	3
49	5	3	4	5	3	3	3
50	5	4	4	4	3	3	2
AVG:	4,72	3,92	3,98	4,18	2,42	2,6	2,9

CHAPTER 6

6 CONCLUSION

As a contribution of this study, it can be stated that there is a new speech compression algorithm whose performance has been improved than the classical SYMPES method with the integration of ZC and phoneme-based segmentation. In order to give a convincing conclusion, the results were created by comparing the newly proposed speech coding method with both the previous SYMPES method and the CELP algorithm, which is used and accepted by everyone.

As given in the previous sections in this study, the figures of the original and the reconstructed speech signals of both male and female speakers at different compression ratios are illustrated. As it can be easily seen from those figures, these two signals are very close to each other. Therefore, it can be said that the reconstructed version of the original speech signal looks similar to the input signal.

Also, considering the speech quality criteria (SNR_{seg}) at all bit rates, 9, 16, 32 kbps, it can be deduced the new proposed zero-cross and phoneme-based SYMPES algorithm performs well than CELP. It has been also demonstrated that all the results obtained using approach 4 are superior to that of the results of the other 3 zero-cross and phoneme-based approaches and the results of the other conventional methods such as CELP and classical SYMPES algorithm at the same bit rates.

Finally, even when the subjective measurement is evaluated, the results prove that Approach 4 provides a better hearing quality for fifty listeners than CELP and the classic SYMPES algorithm according to specific compression ratios ($CR \cong 16$ and $CR \in [24,28.44]$).

REFERENCES

(1939). The Vocoder. Bell. Labs. Rec. Retrieved 9 2021

(1980). Specifications for The Analog to Digital Conversion of Voice by 2,400 Bit/Second Mixed Excitation Linear Prediction. U.S. Department of Defense. Retrieved 9 2021

Atal, B., & Remde, J. (1982). A New Model of Lpc Excitation for Producing Naturalsounding Speech at Low Bit Rates. *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing. 7*, pp. 614-617. IEEE. Retrieved 9 2021

Bachu, R., Kopparthi, S., Adapa, B., & Barkana, B. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, (pp. 1-7). Retrieved 9 2021

Bansal, M., & Sircar, P. (2018). Low Bit-Rate Speech Coding Based on Multicomponent AFM Signal Model. *International Journal of Speech Technology*, 783-795. Retrieved 9 2021, from https://www.springer.com

Boersma, P., & D., W. (2016, 6). Praat, Doing Phonetics by Computer (Computer Program). *6*. Retrieved 9 2021, from http://www.praat.org/

Campbell Jr, J. P., Tremain, T. E., & Welch, V. (1991). The Federal Standard 1016 4800 Bps CELP Voice Coder. In *Digital Signal Processing*, (Vol. 1, pp. 145-155). Retrieved 9 2021

Campbell, J. P. (1997). Speaker Recognition: A tutorial. *Proceedings of the IEEE, 85*(9), 1437-1462. Retrieved 9 2021

Chen, J.-H., & Thyssen, J. (2008). Analysis-By-Synthesis Speech Coding. In *Springer Handbook of Speech Processing* (pp. 351-392). Springer. Retrieved 9 2021

Chu, W. C. (2003). *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders.* New York: Wiley. Retrieved 9 2021, from https://www.wiley.com/en-us

Deller Jr, J., Hansen, J., & Proakis, J. (1999). *Discrete-Time Processing of Speech Signals Wiley.*

Draft Rec, I. t. (1988). G721, 32 Kbps Adaptive Differential Pulse Code Modulation (ADPCM). ITU.

Dudley, W. (1939). Remarking Speech. *11*, 169. Retrieved 9 2021, from https://asa.scitation.org/doi/10.1121/1.1916020

Ekman, L. A., & Kleijn, W. B. (2008). Regularized Linear Prediction of Speech. *IEEE Transactions on Speech and Audio Processing, 16*(1), 65-73. Retrieved 9 2021

El-Jaroudi, A., & Makhoul, J. (1991). Discrete All-Pole Modeling. *IEEE Transactions on Signal Processing, 39*(2), 411-423. Retrieved 9 2021

Gavula, B., Scheets, G., Teague, K., & Weber, J. (2008). The Perceptual Quality of Melp Speech over Error Tolerant IP Networks. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1633-1636). IEEE. Retrieved 9 2021

Gersho, A., & Gray, R. M. (1993). *Vector Quantization and Signal Compression.*

Gibson, J. D. (2016). Speech Compression. *Information, 7*(2), 32. Retrieved 9 2021

Goldberg, R. (2019). *A Practical Handbook of Speech Coders.* CRC press. Retrieved 9 2021

Grassi, S. (1998). Optimized Implementation of Speech Processing Algorithms. Ph.D. Dissertation, Université de Neuchâtel. Retrieved 9 2021

Guo, Y.-F., & Kuo, G.-S. (2007). A New Defined Lower Bit Rate AMR Mode in GSM and WCDMA Networks. *2007 IEEE 65th Vehicular Technology Conference-VTC2007-Spring* (pp. 725-729). IEEE. Retrieved 9 2021

Gürkan, H., Güz, U., & Yarman, B. S. (2004). A Novel Representation Method ror Electromyogram (EMG) Signal with Predefined Signature and Envelope Functional Bank. *2004 IEEE International Symposium on Circuits and Systems. 4*, pp. 69-72. IEEE. Retrieved 9 2021

Gürkan, H., Güz, U., & Yarman, B. S. (2007). Modeling of Electrocardiogram Signals Using Predefined Signature and Envelope Vector Sets. *EURASIP Journal on Advances in Signal Processing*, 1-12. Retrieved 9 2021

Gürkan, H., Güz, U., & Yarman, B. S. (2009). EEG Signal Compression Based on Classified Signature and Envelope Vector Sets. *International Journal of Circuit Theory and Applications, 37*(2), 351-363. Retrieved 9 2021

Güz, U., Gürkan, H., & Yarman, B. S. (2007). A New Method to Represent Speech Signals via Predefined Signature and Envelope Sequences. *EURASIP Journal on Advances in Signal Processing*, 1-17. Retrieved 9 2021

Hansen, C. H. (2001). Occupational Exposure to Noise: Evaluation, Prevention and Control. World Health Organization. In *Fundamentals of Acoustics* (pp. 23-52). Retrieved 9 2021

Hasegawa-Johnson, M. (2000). Line Spectral Frequencies are Poles and Zeros of The Glottal Driving-Point Impedance of a Discrete Matched-Impedance Vocal Tract Model. *The Journal of the Acoustical Society of America, 108*(1), 457-460. Retrieved 9 2021

Iem, B.-G. (2015). A Low Bit Rate Speech Coder Based on The Inflection Point Detection. *International Journal of Fuzzy Logic and Intelligent Systems, 15*(4), 300-304. Retrieved 9 2021

Itakura, F. (1975). Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals. *The Journal of the Acoustical Society of America, 57*(1), 35. Retrieved 9 2021

ITU. (2016). Mean Opinion Score (MOS) Terminology. *Recommendation ITU-T P.800.1. ITU-T Telecommunication Standardization Sector of ITU Geneva*. ITU-T P.800.1. ITU-T Telecommunication Standardization Sector of ITU Geneva. Retrieved 9 2021

Jage, R., & Upadhya, S. (2016). CELP and MELP Speech Coding Techniques. *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). 5*, pp. 1398-1402. IEEE. Retrieved 9 2021

Jagtap, S., Mulye, M., & Uplane, M. (2015). *Speech Coding Techniques* (Vol. 49). Retrieved 9 2021

Jayant, N. S. (1974). Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers. *Proceedings of the IEEE, 62*(5), 611-632. Retrieved 9 2021

Jelinek, M., Eksler, V., Lemyre, C., & Lefebvre, R. (2007). Classification-Based Techniques for Improving The Robustness of CELP Coders. *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers* (pp. 1480-1484). IEEE. Retrieved 9 2021

Jolli, J. (1993). Principal Component Analysis. *Springer Series in Statistics*. Retrieved 9 2021, from https://www.springer.com/series/692

Kleijn, W. B., Krasinski, D. J., & Ketchum, R. H. (1990). Fast Methods for The CELP Speech Coding Algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38*(8), 1330-1342. Retrieved 9 2021

Kleijn, W. B., Storus, A., Chinen, M., Denton, T., Lim, F. S., Luebs, A., . . . Yeh, H. (2021). Generative Speech Coding with Predictive Variance Regularization. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6478-6482). IEEE. Retrieved 9 2021

Kohler, M. A. (1997). A Comparison of The New 2400 Bps Melp Federal Standard With other Standard Coders. *1997 IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 1587-1590). Retrieved 9 2021

Kondoz, A. M. (1998). *Digital Speech: Coding for Low Bit Rate Communication Systems.* John Wiley & Sons. Retrieved 9 2021

Kumar, R. P. (2007). High Computational Performance in Code Exited Linear Prediction Speech Model Using Faster Codebook Search Techniques. *2007 International Conference on Computing: Theory and Applications (ICCTA'07)* (pp. 458-462). IEEE. Retrieved 9 2021

Makhoul, J. (1975). Linear Prediction: A Tutorial Review. *Proceedings of the IEEE, 63*(4). Retrieved 9 2021

Matassini, L. (2001). *Signal Analysis and Modelling of Non-Linear Non-Stationary Phenomena.* Germany: Bergische Universitat Gesamthochschule Wuppertal. Retrieved 9 2021

Murthi, M. N., & Rao, B. D. (2000). All-Pole Modeling of Speech Based on The Minimum Variance Distortionless Response Spectrum. *IEEE Transactions on Speech And Audio Processing, 8*(3), 221-239. Retrieved 9 2021

Ng, W. K., Choi, S., & Ravishankar, C. (1997). Lossless and Lossy Data Compression. In *Evolutionary Algorithms in Engineering Applications* (pp. 173-188). Springer. Retrieved 9 2021

Oliver, B. M., Pierce, J. R., & Shannon, C. E. (1948). The Philosophy of PCM. *Proceedings of the IRE, 36*, 1324-1331. Retrieved 9 2021

Osman, M. A., Al, N., Magboub, H. M., & Alfandi, S. (2010). Speech Compression Using Lpc and Wavelet. *2010 2nd International Conference on Computer Engineering and Technology. 7*, pp. 92-99. IEEE. Retrieved 9 2021

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, 77*(2), 257-286. Retrieved 9 2021

Rabiner, L., & Schafer, R. (2007). *Introduction to Digital Speech Processing* (Vol. 1). Retrieved 9 2021

Rabiner, L., & Schafer, R. (2010). Theory and Applications of Digital Speech Processing. Pearson. Retrieved 9 2021, from https://plc.pearson.com/en-GB

Raja, M. M., Jangid, M. P., & Gulhane, S. M. (n.d.). *International Journal of Engineering Sciences & Research Technology Linear Predictive Coding*.

Ramamoorthy, V., & Jayant, N. S. (1984). Enhancement of ADPCM Speech by Adaptive Postfiltering. *AT&T Bell Laboratories Technical Journal, 63*(8), 1465-1475. Retrieved 9 2021

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual Evaluation of Speech Quality (Pesq)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs. *2001 IEEE International Conference on Acoustics, Speech and Signal Processing. 2*, pp. 749-752. IEEE. Retrieved 9 2021

Schroeder, M., & Atal, B. (1985). Code-Excited Linear Prediction (CELP): Highquality Speech at Very Low Bit Rates. *ICASSP'85. IEEE International Conference on Acoustics, Speech and Signal Processing. 10*, pp. 937-940. IEEE. Retrieved 9 2021

Shannon, C. E. (1993). Collected Papers, Edited by Nja Sloane and Ad Wyner. Retrieved 9 2021

Sisman, B., Gürkan, H., Güz, U., & Yarman, B. S. (2013). A New Speech Coding Algorithm Using Zero Cross and Phoneme Based Sympes. *International Symposium on Signals, Circuits and Systems ISSCS2013* (pp. 1-4). IEEE. Retrieved 9 2021

Spanias, A. S. (1994). Speech Cding: A tutorial Review. *Proceedings of the IEEE, 82*(10), 1541-1582. Retrieved 9 2021

Tremain, T. E. (1982). The Government Standard Linear Predictive Coding Algorithm: Lpc-10. *Speech Technology*, 40-49. Retrieved 9 2021

Uddin, S., Ansari, I. R., & Naaz, S. (2016). Low Bit Rate Speech Coding Using Diffrential Pulse Code Modulation. *Advances in Research*, 1-6. Retrieved 9 2021, from https://www.journalair.com

Varghese, P., & Ramesh, P. (2015). Advanced Voice Excited Linear Predictive Coding with Noise Reduction. In *International Journal of Current Engineering and Technology* (Vol. 5, pp. 1887-1889). Retrieved 9 2021

Warkade, P., & Mishra, A. (2015, 5). Lossless Speech Compression Techniques: A Literature Review. *3*(Issue-3). Retrieved 9 2021, from https://www.academia.edu/36347557/Lossless_Speech_Compression_Techniques_A_Literature_Review

Wolfe, J., Garnier, M., & Smith, J. (2009). Vocal tract resonances in speech. *HFSP Journal, 3*(1), 6-23. Retrieved 9 2021

Wu, C., Jiang, H., & Li, B. (2009). An Improved Melp Speech Coder. *2009 International Conference on Information Technology and Computer Science. 2*, pp. 130-133. IEEE. Retrieved 9 2021

Yarman, B. S., Güz, U., & Gürkan, H. (2006). On The Comparative Results of Sympes: A New Method of Speech Modeling. *AEU-International Journal of Electronics and Communications, 60*(6), 421-427. Retrieved 9 2021

Young, S. J., & Young, S. (1993). *The HTK Hidden Markov Model Toolkit: Design And Philosophy*. Retrieved 9 2021, from https://www.researchgate.net/publication/263124034_The_HTK_Hidden_Markov_Model_Toolkit_Design_and_Philosophy

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., . . . Povey, D. (2006). *The HTK Book.* Cambridge University Engineering Department. Retrieved 9 2021, from

https://www.researchgate.net/publication/289354717_The_HTK_Book_version_35a

CURRICULUM VITAE