Sustainable Engineering and Innovation
Vol. 3, No. 1, January 2021, pp.11-14
https://doi.org/10.37868/sei.v3i1.119

ISSN 2712-0562
Original Research

Speech file compression by eliminating unvoiced/silence components

Arda Şahin[1], Mehmet Zübeyir Ünlü[1*]
[1] Department of Electrical and Electronics Engineering, İzmir Institute of Technology, İzmir, Turkey

*Corresponding author: zubeyirunlu@iyte.edu.tr

Abstract

The main objective of this study is to have the noise component of a speech signal eliminated and compressed by storing the locations and durations of silence regions. The separation between voiced, unvoiced, and silence regions is done by using the Short-Time Energy (STE) and Zero Crossing Rate (ZCR) methodologies. All operations in this study have been performed by using the User Interface (UI) developed on MATLAB®. These operations include voice recording, playing the recording, eliminating the unwanted regions, playing the modified recording, saving original and compressed files, and loading the recording compressed.

Keywords: Speech file compression, Zero Crossing Rate, Short Time Energy, Noise elimination

1. Introduction

A typical voice recording consists of three main regions: Voiced regions where the speech of interest is mainly stored on, unvoiced regions which contains the low amplitude sections from the source that is unrecognizable, and the silence parts which only contain the unwanted noise. Main objective of this study is to separate and eliminate the unvoiced and silence regions and compress the speech signal. The locations and durations of these signal parts will be stored for the reconstruction purpose also. This study also includes developing a User Interface (UI) created on MATLAB®. Typical view of the UI can be seen below.

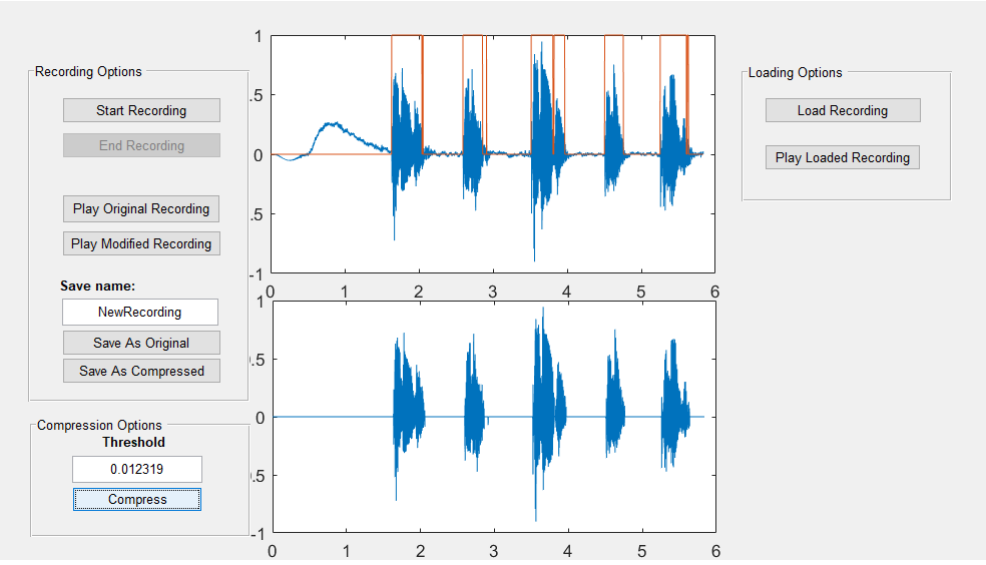

Figure 1. The user interface developed during the study

2. Methodology

The process is started by the user with recording. After recording step has been performed, the original speech waveform can be obtained as seen in Figure 1.

2.1. Eliminating the unvoiced/silence regions with an offset

The first course of action is eliminating the unvoiced and silence parts which have an offset. These parts should be removed before using the Short Time Energy (STE) because STE cannot recognize these regions as unvoiced/silence regions [1,2]. As an example, the figure below shows a sample of a voice recording that has unvoiced/silence regions with offset.

$$E_n = \sum_{m=n-L+1}^{n} [x(m)w(n-m)]^2 \quad (1)$$

Here, E is the energy of the signal, and $x(m)$ is the signal in above the formulas, L is the window length and $w(n\text{-}m)$ is the window function [1].

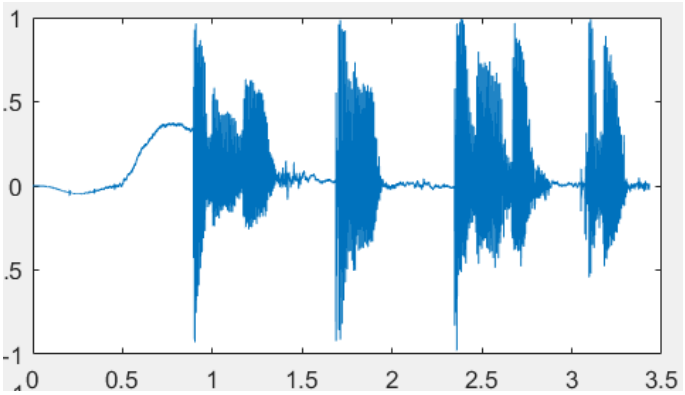

Figure 2. A sample of a voice recording

The main tool for eliminating these regions is Zero Crossing Rate (ZCR) [1-3]. Basically, ZCR of a speech signal indicates the number of crossings that have been done from zero for a specific windowed signal. For the windowing, a rectangular window has been used with a width of "Sampling Rate/1000", which corresponds to 1 millisecond of a speech recording.

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \quad (2)$$

where sgn($x[m]$) is the sign function of the speech signal. An example of ZCR graph of a speech signal can be seen in Figure 2. We can also see from the Figure 3 that the number of crosses in unvoiced/silence regions with offset are zero because their offset prevents them from crossing the zero point. With this knowledge, the regions with zero ZCR will be eliminated if they are long enough.

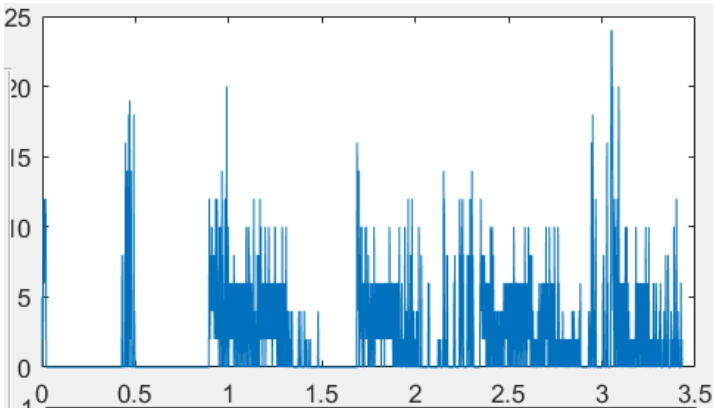

Figure 3. ZCR of a speech signal

2.2. Adaptive threshold value for STE

The threshold value can manually be selected by the user. But to give a reference, the threshold value is initially selected by the program. This way, the voiced, unvoiced, and silence regions can be decently separated without knowing the amplitude of the noise.

In order to determine the value of the threshold, first the STE of the speech signal has been obtained with a fixed threshold value. The value of threshold for this part is selected higher than the typical amount to make sure that all the noise component of the signal falls below the threshold value. After the decision, the parts below the threshold will significantly contain unvoiced and silence regions.

Next the mean value for all STE points below the threshold region are obtained. This value will be assumed to be the average STE for the noise component. Finally, this value will be multiplied by 1.5 to cover the "above average" parts of the noise region.

2.3. Separating the unvoiced/silence regions

After the threshold value is obtained, the STE of the speech signal will once again be obtained but this time, the threshold will be determined by the currently selected threshold value. The STE of the speech sample on Figure 2 can be analyzed in Figure 4.

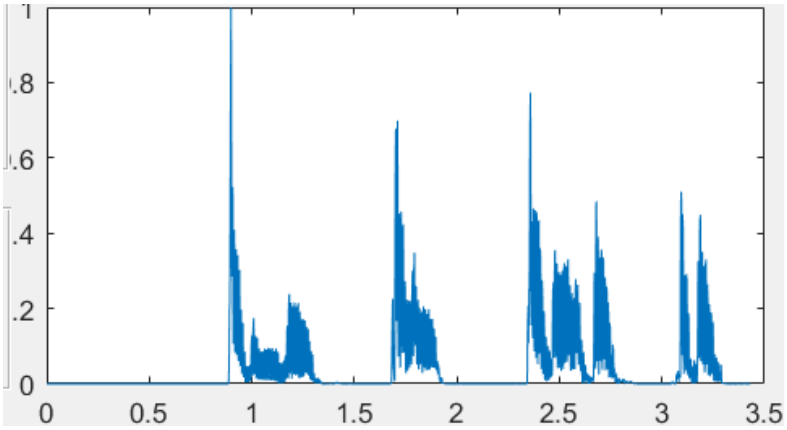

Figure 4. STE of the speech signal

For this sample, the length of the window is smaller than the ideal length since the ripples of the STE are too large. These ripples may cause the oscillatory decisions at the output, which is unwanted. In order to prevent this, new sections must be longer than 10ms to be decided as the new section. The results are presented in Figure 5. The areas contained inside red rectangles are considered as the voiced regions [4]. After the separations, the amplitudes of all unvoiced and silence regions will be assigned to zero. The finalized waveform can be seen in Figure 6.

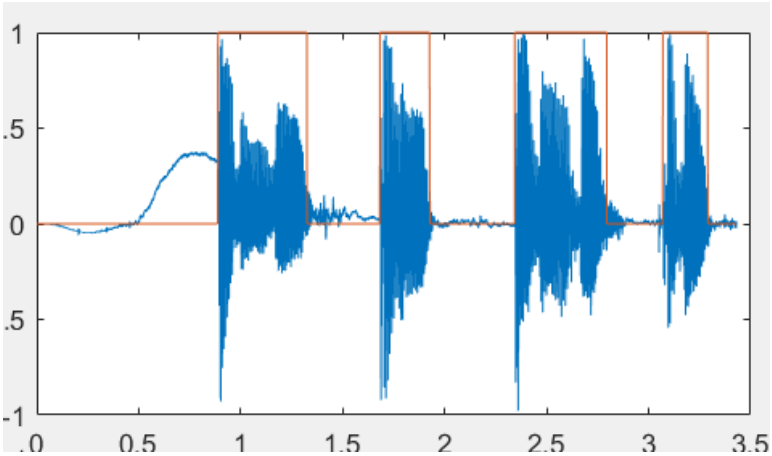

Figure 5. Separated regions of the speech signal

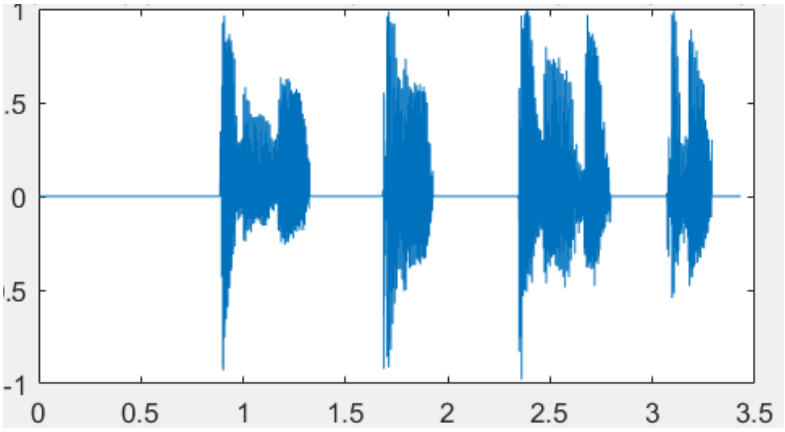

Figure 6. Finalized form of the speech signal

2.4. Wave file compression

After the voiced and unvoiced regions are separated, first the information about the number of trailing zeros in the signal (it has to have the length of at least 20 samples to be worth compressing) will be written at the start of the wave file. Next, information regarding the starting locations and the length of the trailing zeros will be written right next to the actual data.

After these points have been written, all trailing zeros will be removed from the voice recording. This way, our wave file is compressed with the amount that depends on the length of the silence regions.

2.5. Wave file reconstruction

After reading the compressed wave file, first initial information will be extracted from the audio data. Then, at the initial locations given by the initial information, trailing zeros will be re-placed to the signal. The final signal will be a combination of voiced regions and silences [5].

3. Conclusions

Despite the fact that there are much better lossy audio/speech compressions like .mp3 format, which can compress at least %75 of the raw speech file, this project was a great way to implement some of the concepts that have been thought on “Speech Processing” lectures.

There are some alternatives that may improve the resulting speech quality of this study. First, the length of the window can also be adaptive instead of a fixed length. This would minimize the probability of oscillatory decisions. One of the other alternatives is also changing the silence parts with artificial noise components. This may prevent the speech sound truncated at silence parts.

References

[1] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall Inc., 2011.

[2] T. F. Quatieri, *Principles of Discrete - Time Speech Processing*, Prentice Hall Inc, 2002.

[3] R. G. Bachu, S. Kopparthi, B., Adapa, B. D. Buket, Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy,” in *Advanced Techniques in Computing Sciences and Software Engineering*, Dordrecht: Springer, 2010, pp. 279-282.

[4] Z. Goh, K-C. Tan, B. T. G. Tan, “Kalman filtering speech enhancement method based on a voiced-unvoiced speech model,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, No. 5, pp. 510-524, 1999.

[5] W. B. Kleijn, J. Haagen, “Transformation and decomposition of the speech signal for coding,” *IEEE Signal Processing Letters*, vol. 1, no. 9, pp 136-138, 1994.