Speech File Compression by Eliminating Unvoiced-Silence Components

Arda ŞAHİN

Department of Electrical and Electronics Engineering, İzmir Institute of Technology, İzmir, TURKEY

ardasahin@std.iyte.edu.tr

Mehmet Zübeyir ÜNLÜ

Department of Electrical and Electronics Engineering, İzmir Institute of Technology, İzmir, TURKEY

zubeyirunlu@iyte.edu.tr

OBJECTIVE

To separate and eliminate 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.

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
  - **Silence** parts which only contain the unwanted noise.
- Separating and eliminating the unvoiced and silence regions from voiced region yield compressing the speech signal.

Methodology

A. Eliminating the Unvoiced/Silence Regions with an Offset

- The separation process among voiced, unvoiced, and silence regions can be performed by using
  - Short Time Energy (STE) and

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

  - Zero Crossing Rate (ZCR)

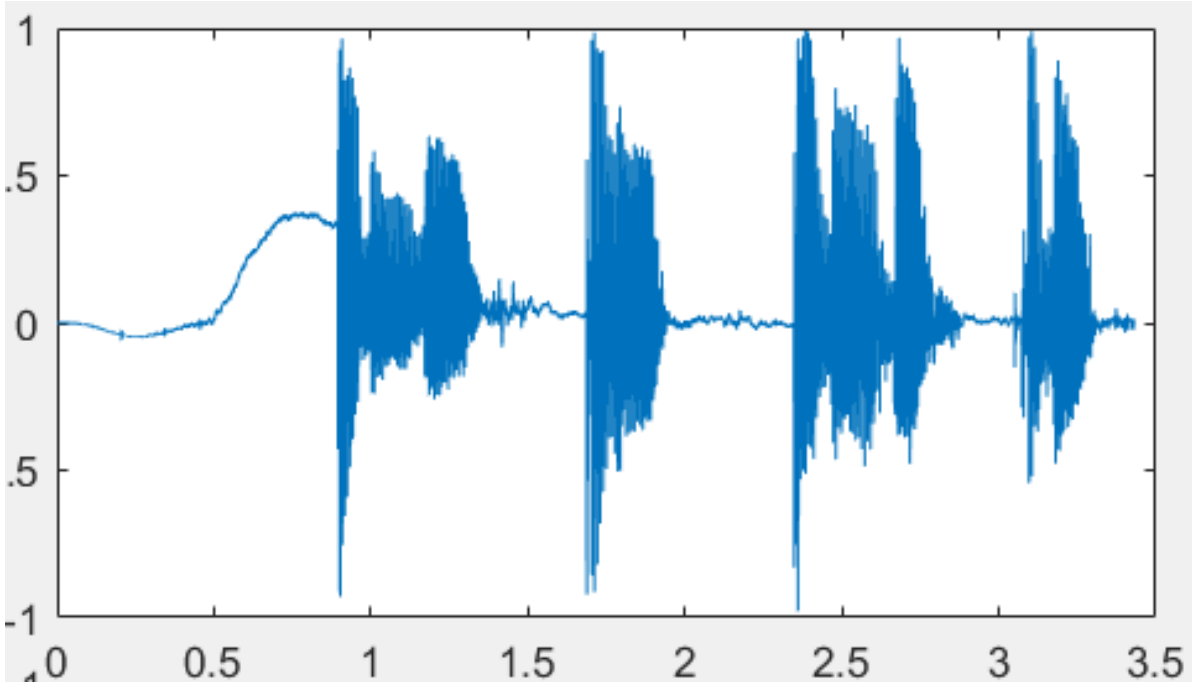
$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \qquad (2)$$

methods.

Methodology

A. Eliminating the Unvoiced/Silence Regions with an Offset

- The unvoiced and silence parts which have an offset should be removed before using the STE because STE cannot recognize these regions as unvoiced/silence regions since they have an offset.
- The figure shows a sample of a voice recording that has unvoiced/silence regions with an offset.

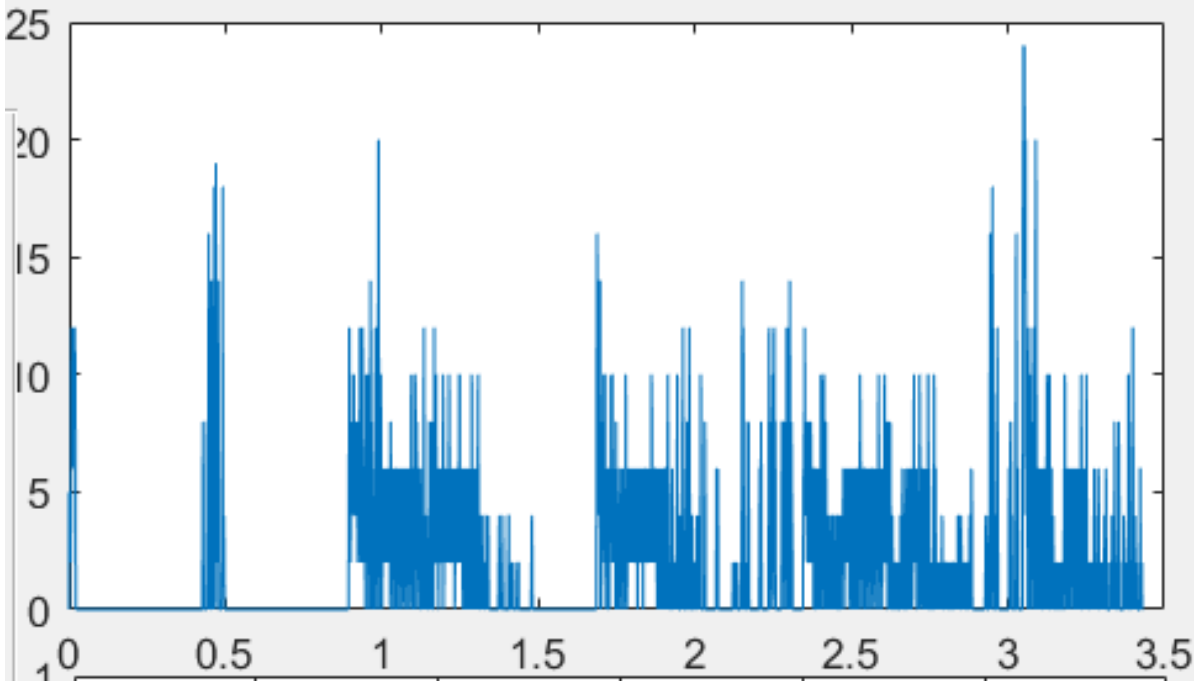


Methodology

A. Eliminating the Unvoiced/Silence Regions with an Offset

- It can be seen from the following figure 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, regions with zero ZCR will be eliminated if they are long enough.



Methodology

B. Adaptive Threshold Value for STE

- In order to determine the value of the threshold, first the STE of the speech signal is 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.

Methodology

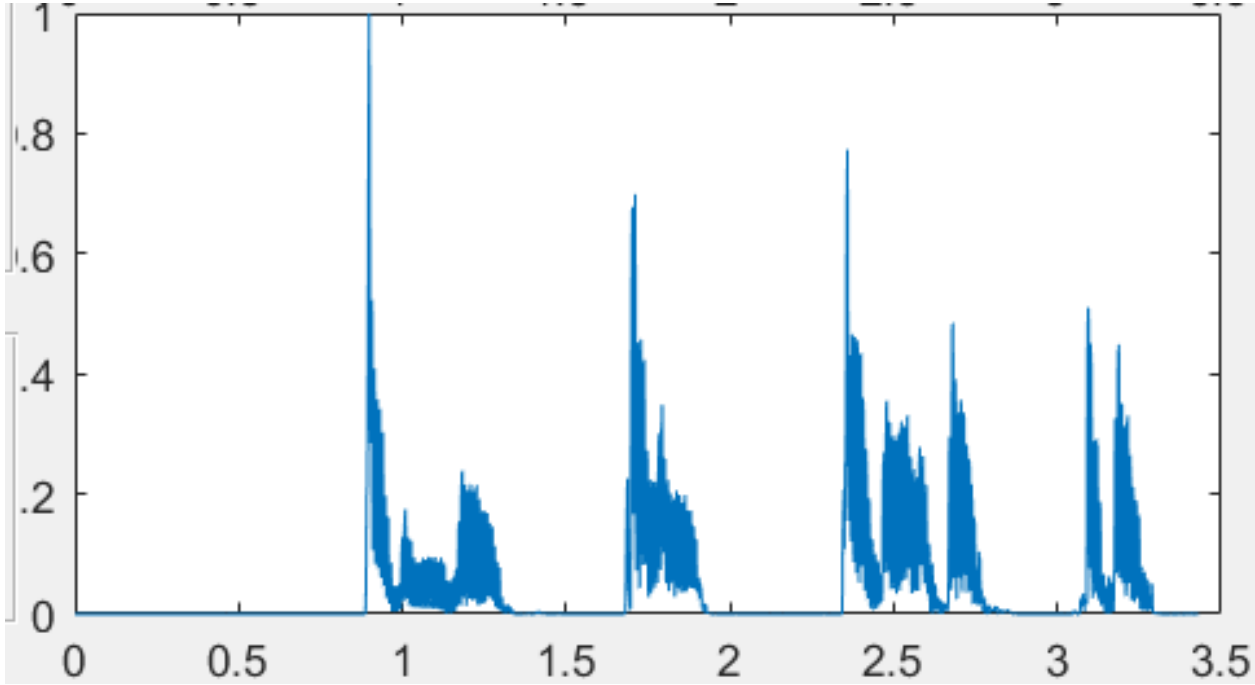
B. Adaptive Threshold Value for STE

- 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.

Methodology

C. Separating the Unvoiced/Silence Regions

- After the threshold value is obtained, the STE of the speech signal will once again be calculated 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 seen in the following figure.



Methodology

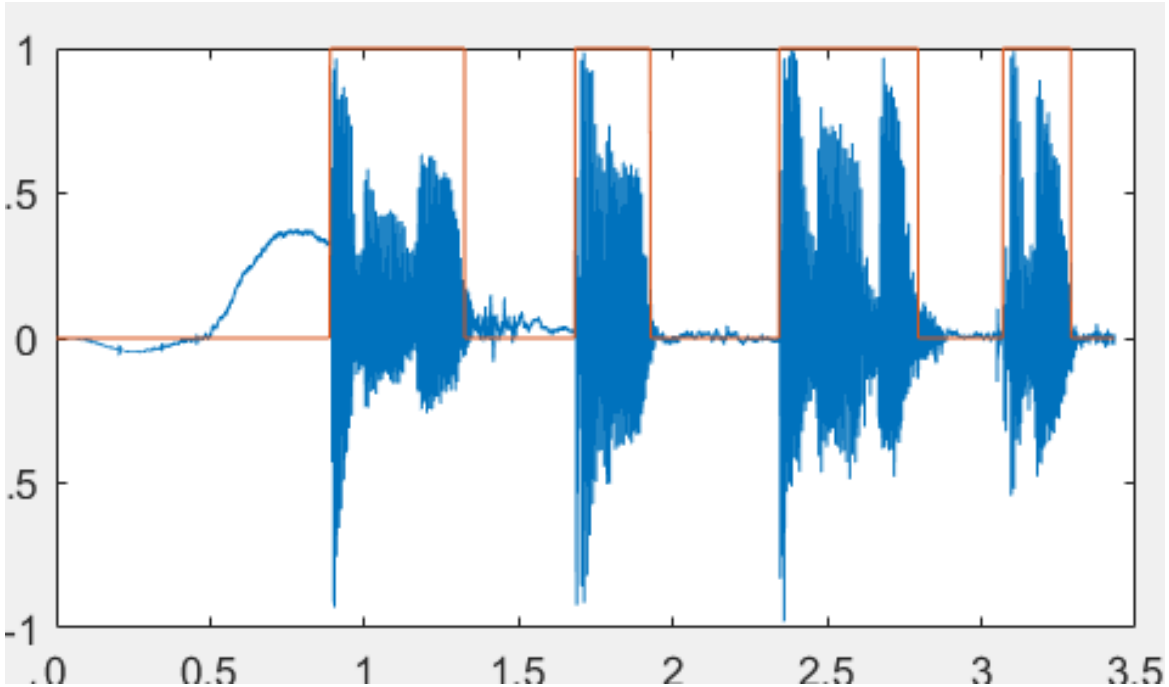
C. Separating the Unvoiced/Silence Regions

- 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.

Methodology

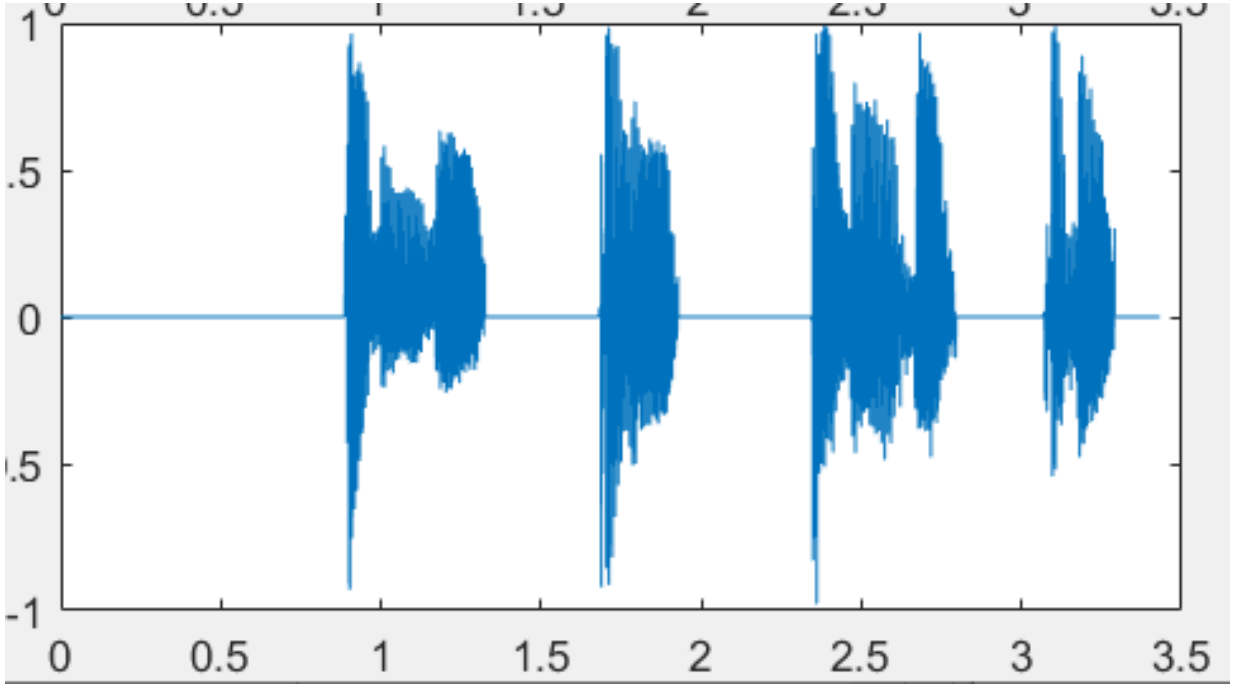
C. Separating the Unvoiced/Silence Regions

- The decision result can be seen below. The areas contained in the orange rectangles are considered as the voiced regions.



C. Separating the Unvoiced/Silence Regions

- After the separation, the amplitudes of all unvoiced and silence regions will be assigned as zero. The finalized waveform can be seen in following figure.



D. 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 depends on the length of the silence regions.

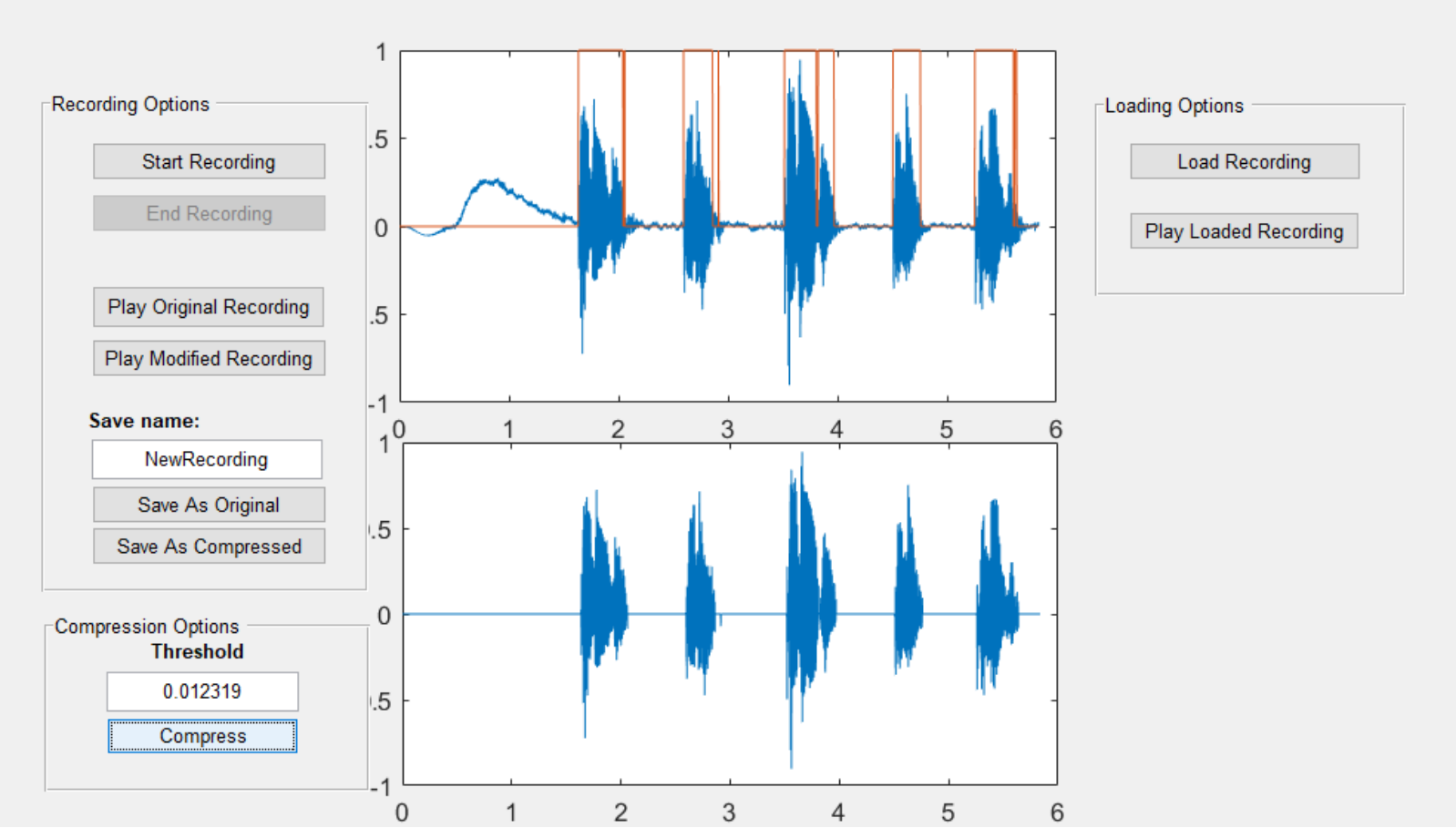
E. Wave File Reconstruction

- 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 replaced to the signal.
  - The final signal will be a combination of voiced regions and silences

Experimental Results

Developing a GUI

- This study also includes developing a UI that has been created on MATLAB® software. Typical view of the UI can be seen below:



Conclusion

- Despite the fact that, there are much better lossy audio compressions like .mp3 format, which can compress at least %75 of the wave file. This project was a great way to implement some of the concepts that have been thought on the “Speech Processing” lectures.
- There are some alternatives that may have improved the resulting audio quality of this project. First, the length of the window can also be adaptive instead of a fixed length. This would have minimized the probability of oscillatory decisions.
- One of the other alternatives are 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 & Barkana, Buket. (2010). Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. 10.1007/978-90-481-3660-5-47.

[4] Goh, Z., Tan, K.-C. & Tan, B. T. G. (1999), 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.

[5] Kleijn, W. B. & Haagen, J. (1994), Transformation and decomposition of the speech signal for coding, IEEE Signal Processing Letters, Vol. 1, No. 9, pp 136-138.

Questions?

Thanks for Listening!

Questions?