Speech and Audio Processing
University of Thessaly
Department of Electrical and Computer Engineering

Assignment 1 Report

Alexandra Gianni
STUDENT ID: 03382

Question 1

In this question, we will analyze two audio files, `sample1.wav` and `sample2.wav`, using spectrograms to extract key speech features such as the fundamental frequencies, formant frequencies, and the speaker's gender. The analysis is performed in Matlab using the `audioread` function and we continue by plotting the narrowband and wideband spectrograms.

Spectograms provide a time-frequency representation of an audio signal by applying the Short-Time Fourier Transform (STFT) over overlapping windows. This allows us to visualize how frequency content changes over time. Depending on the analysis window length, we can produce:

- **Wideband spectrograms:** A spectrogram that emphasizes temporal changes in the signal. They use short time windows (e.g. $5 - -10$ ms). They are well-suited for identifying transient speech features and formant structures.
- **Narrowband spectrograms:** A spectrogram that emphasizes frequency details in the signal. They use longer time windows (e.g. $25 - -30$ ms). They are effective for visualizing harmonic structures and the fundamental frequency.

The Fundamental frequency (F_0) is defined as the frequency at which the vocal folds vibrate when voiced speech sounds are made. It appears in *Narrowband spectrograms* as a series of evenly spaced horizontal lines (harmonics). Typically:

- Male speakers: $F_0 \approx 85 - 155$ Hz
- Female speakers: $F_0 \approx 165 - 255$ Hz

The difference in frequency depends on the speaker's gender and age—men generally have longer and thicker vocal folds, producing lower pitch and thus lower F_0, whereas women have shorter, thinner folds. Children have the highest F_0.

Furthermore, Formants are the resonant frequencies of the vocal tract and are crucial in characterizing vowel sounds. These are observed more clearly in *Wideband spectograms*. The first two formants, F_1 and F_2, are particularly important in analyzing vowel qualities.

- F_1 is inversely related to vowel height (e.g., low vowels have higherF_1).
- F_2 relates to vowel frontness (e.g., front vowels have higherF_2.

Together, F_0, F_1 and F_2 allow us to distinguish vowels and estimate speaker characteristics such as gender. Spectrogram analysis is therefore widely used in speech processing, phonetics, and speaker recognition applications.

Implementation and results

After reading and loading the files using `audioread`, both narrowband and wideband spectrograms were computed using MATLAB's `spectrogram` function.
Here are the spectrogram plots for both `sample1.wav` and `sample2.wav`. In both plots, I have identified the fundamental frequency and the formants have been marked.

Sample1.wav:

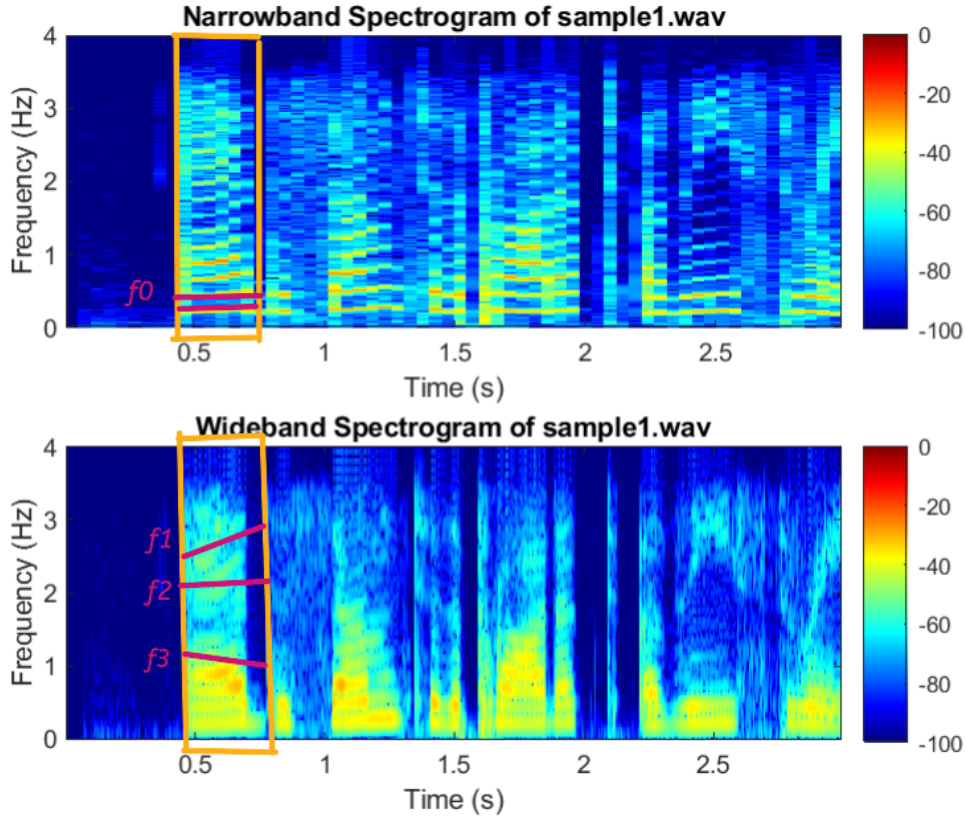

Figure 1: Narrowband and wideband spectrograms of `sample1.wav`

Sample2.wav:

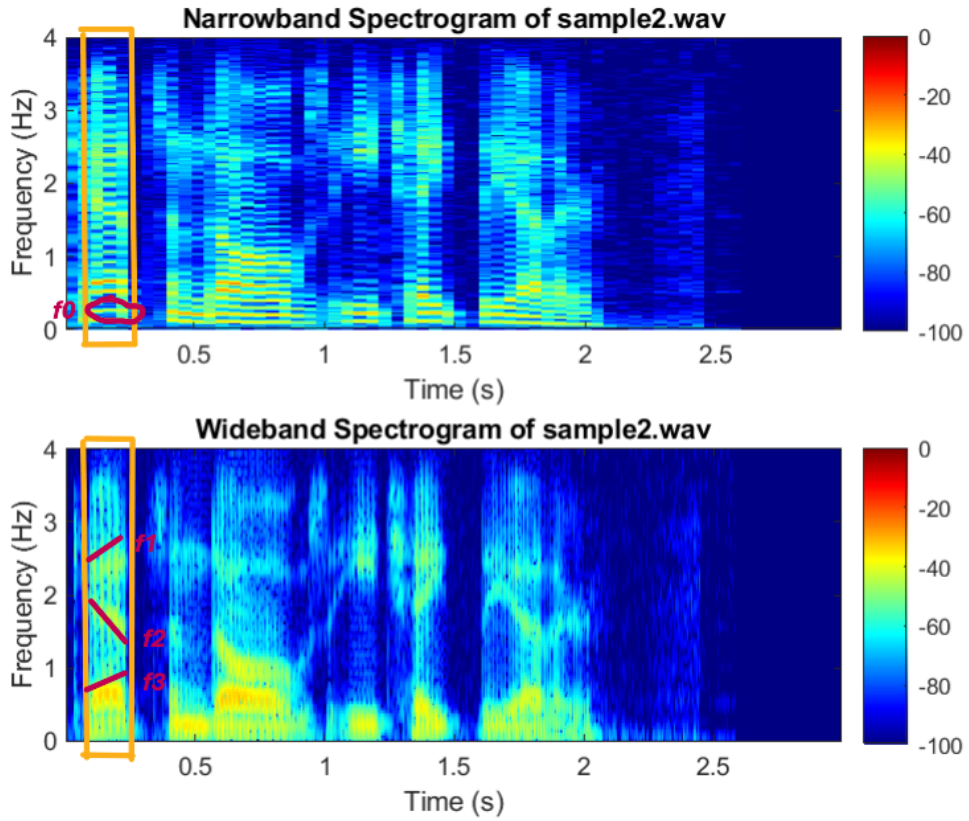

Figure 2: Narrowband and wideband spectrograms of `sample2.wav`

In the narrowband spectrograms, the fundamental frequency F_0 appears as the lowest horizontal band of high energy (bright color) that is stable over time.

- **Sample1:** The estimated F_0 is approximately **200 Hz**. This relatively high pitch suggests the speaker is likely a **woman**.
- **Sample2:** The estimated F_0 is approximately **80–130 Hz**. This lower pitch is consistent with a **male** speaker.

In the wideband spectrograms, we identify the formant frequencies of the vowels:

- `Sample1.wav` (vowel in ”add”): $F_1 \approx 700$ Hz, $F_2 \approx 1700$ Hz
- `Sample2.wav` (vowel in ”cats”):$F_1 \approx 750$ Hz, $F_2 \approx 1600$ Hz

These values are characteristic of the vowel /æ/, further supporting the phonetic content of the recordings.

Conclusion

By examining both spectrogram types, we successfully identified the fundamental and formant frequencies in each sample. Based on the frequency ranges, we determined the likely gender of each speaker. Spectrograms thus serve as powerful tools for acoustic speech analysis and speaker profiling.

Question 2

In this question, we explore the acoustic characteristics of the word ”asa” by recording, analyzing, segmenting, and resampling the audio signal using Matlab. We divide the signal into 200ms segments and plot it.

Speech signals are composed of both **voiced** and **unvoiced** sounds. Understanding the difference between these is fundamental in speech processing:

- **Voiced sounds** (e.g., vowels like /a/) are produced when the vocal folds vibrate. These vibrations result in periodic waveforms with higher amplitude and a clear structure. The signal typically shows regular oscillations.
- **Unvoiced sounds** (e.g., fricatives like /s/) are produced without vocal fold vibration, typically by turbulent airflow through a constriction in the vocal tract. These sounds appear as noisy, aperiodic segments with low amplitude and no clear pitch.

Segmenting the signal into equal time intervals helps in observing the temporal structure of speech. In this case, we divide the signal into segments of 200 ms each to analyze individual phonetic components.

Resampling is a common technique used to alter the playback speed of a signal. By changing the sampling rate, we can stretch or compress the time axis:

- Increasing the sampling period (reducing the sampling rate) slows down the signal.
- Decreasing the sampling period (increasing the sampling rate) speeds it up.

When slowing down the signal, we expect the duration to increase while the overall amplitude remains constant. This helps us better observe fine-grained signal features.

Implementation and results

The word ”asa” was recorded in Matlab using the `audiorecorder` function. The recorded waveform then was plotted, and the signal was divided into 200ms.

To identify the voiced and unvoiced portions of the signal, we examine the amplitude and periodicity of each segment:

- The vowel /a/ portions appear as smooth, periodic regions with prominent amplitude—indicative of voiced sounds.
- The /s/ portion in the middle shows a noisy, aperiodic pattern with lower amplitude, marking it as an unvoiced fricative.

This is the audio segments signal plot:

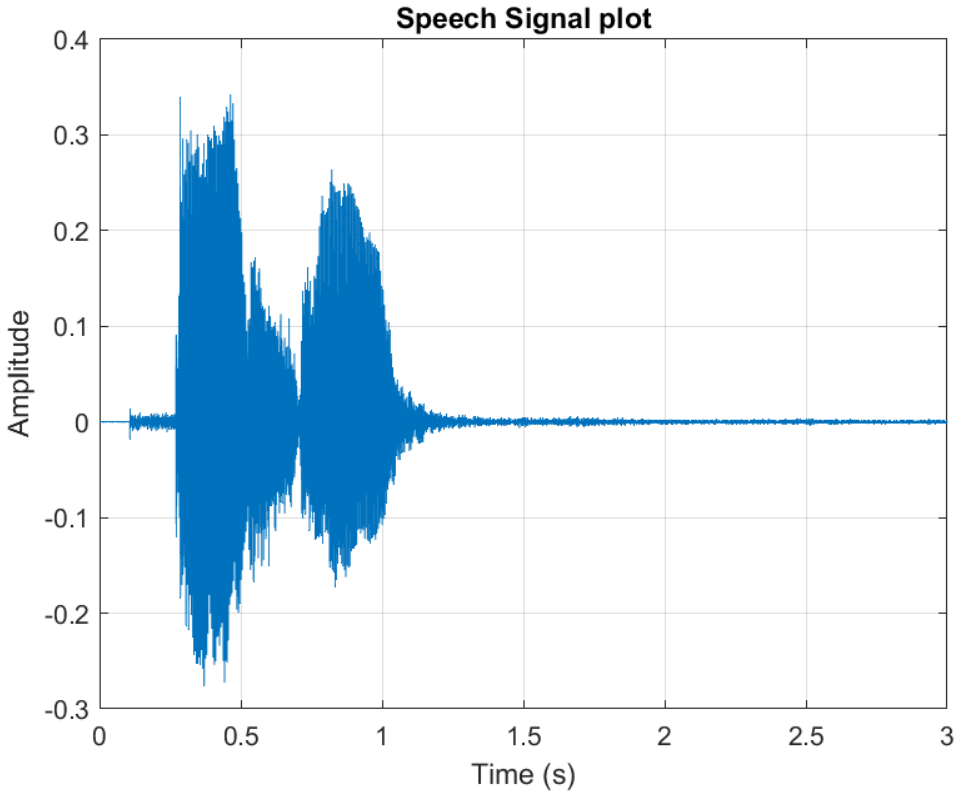

Figure 3: Full waveform of the recorded word “asa”

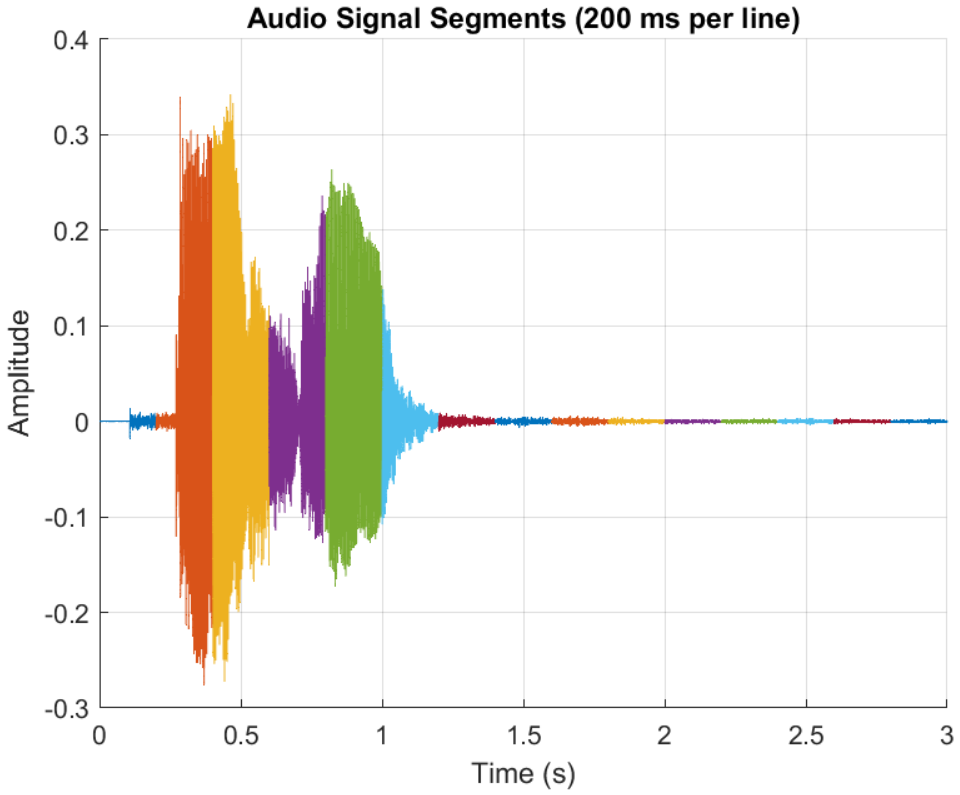

Figure 4: Signal divided into 200 ms segments

Voiced segments have a defined peak and are periodic with a damp. Unvoiced segments are irregular, do not have any peaks defined, nor a period. The following plot shows us exactly that. We can observe that voiced segments have a higher amplitude and the waveform has a regular and smoother pattern.

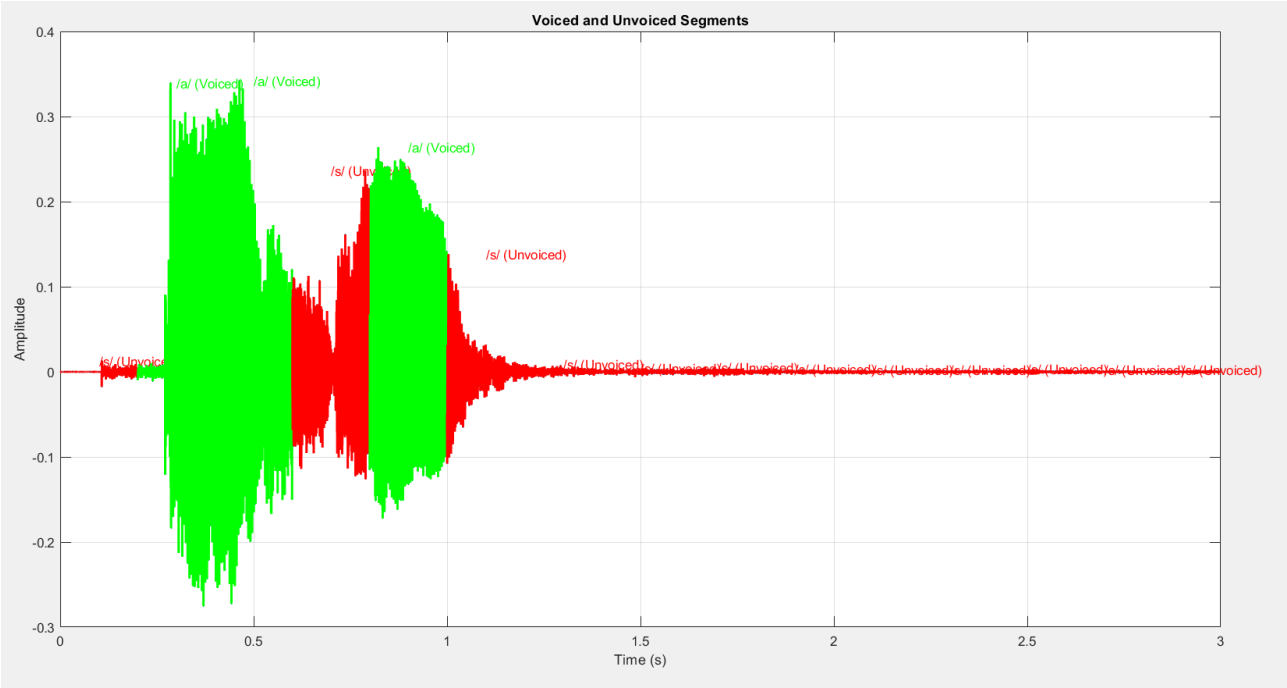

Figure 5: Labeled waveform showing voiced and unvoiced segments

For the final part, the audio signal was slowed down by using resampling and by alternating the frequency. We are going to double the frequency, so the period will be reduced by half.

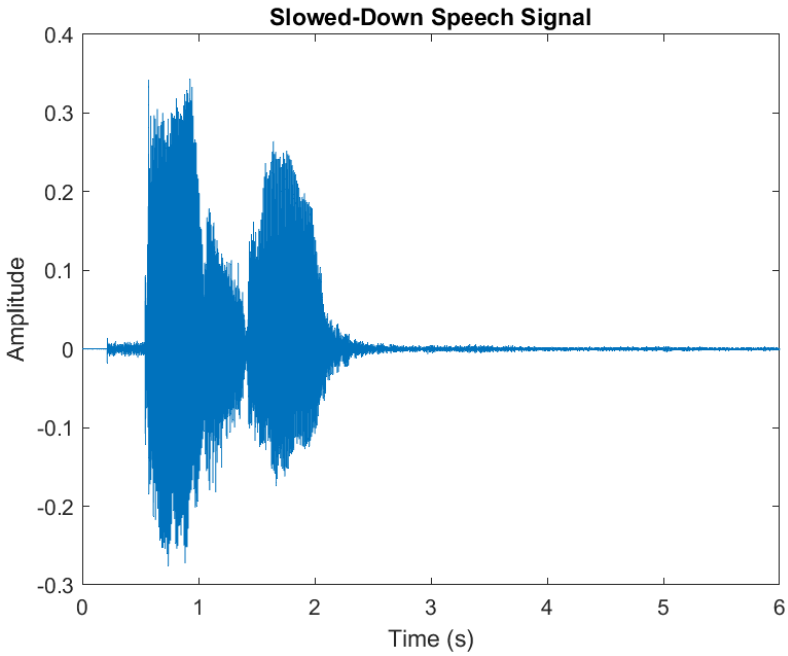

Figure 6: Slowed-down version of the recorded speech signal

We can observe that the signal is indeed slowed down, because on the original signal the speech ends at approximately $t = 1 - 1.5s$, while on the slowed down it is doubled $t = 2 - 3s$. The slowed-down version lasts approximately twice as long, thus the time of computing has increased. However, the amplitude remains consistent in both signals, confirming that resampling alters the time domain without affecting the signal's energy.

Conclusion

Through waveform analysis and resampling, we identified the acoustic properties of the word "asa", distinguishing voiced and unvoiced segments and observing how resampling affects the temporal characteristics of speech. This exercise demonstrates core principles of speech signal analysis in both the time and frequency domains.