

Digital Signal Processing

Seminar for exam

*Derive Short Term Fourier Transform (STFT)
of woman's and man's sound record*

January 2021

Author: Ivan Nikolov, 63190378

1. Introduction

Two of the vocal properties are intensity (loudness) and frequency (pitch). It is well known that the pitch of a man's voice in average falls under low frequencies, whereas the pitch of a woman's voice is usually higher. Pitch and frequency are proportional to one another.

Data is given in the form of four records (man speaking, man screaming, woman speaking and woman screaming).

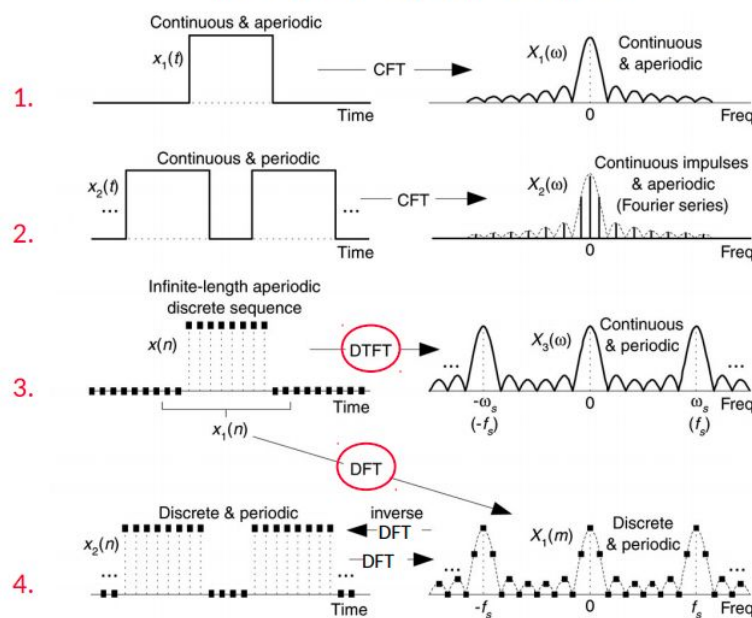
The goal of this seminar is to test using STFT(short time Fourier transform) if the properties defined above apply to the given records.

2. Methods

Fourier (1807) "... all periodic continuous signals can be represented by a sum of properly chosen sinus signals ..."

In order to get information of which frequencies a signal is composed, a Fourier transformation of the signal is needed. There are four types of Fourier transforms(picture below):

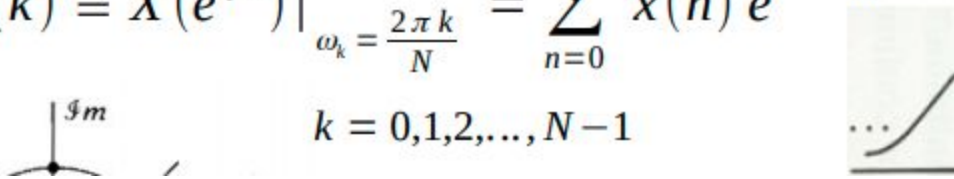
Fourier transforms



In digital signal processing DTFT and DFT are used. DFT (discrete Fourier transform) of a signal is calculated using the following formula.

$$X(k) \equiv X(e^{j\omega}) \Big|_{\omega_k = \frac{2\pi k}{N}} = \sum_{n=0}^{N-1} x(n) e^{-j(\frac{2\pi k}{N}) \cdot n}$$

$k = 0, 1, 2, \dots, N-1$



Spectrum is defined by N distinct frequency samples

The DFT divides (or samples) the interval $[0 \dots 2\pi]$ into N equal steps to get $X(k)$.

However, just from looking the frequency spectrum we could not see the amplitude (in our case how loud) a frequency is. That can be accomplished by calculating the amplitude spectrum of the signal which is given by the following formula.

$$|X(e^{j\omega})| = \sqrt{X_R^2(e^{j\omega}) + X_I^2(e^{j\omega})} \quad |X(k)| = \sqrt{X_R^2(k) + X_I^2(k)}$$

Although, the amplitude spectrum is a powerful tool to analyse one signal, some of its properties do not suit all types of signals and need to be discussed.

Amplitude spectrum of a signal only reflects which frequencies exist during the total observation interval, because the Fourier transform integrates frequency components over the total observation interval.

- Stationary signals
 - * Their frequency content does not change over time
 - * Amplitude spectrum is a suitable representation
 - Non-stationary signals
 - * Their frequency content does change over time
 - * Frequent changes are bringing important information (speech, music, other signals, ...)
- * It is important to provide information on when in time different frequencies of a signal occur
- Perform frequency analysis on short signal segments and move over signal
- = Short Term Fourier Transform (STFT) - spectrogram.

•

Because the human voice is not composed of one frequency and is a non-stationary signal, it would be better to use a Matlab tool that computes STFT or spectrogram.

STFT is calculated by:

- Dividing the signal in short segments
- Calculating the amplitude spectrum for each segment
- Composing time series of the spectra
 - If the spectra is squared (power spectra), then we get a **spectrogram**.

On the spectrogram the x-axis typically represents the time, and the y-axis the frequencies. On the z-axis (usually represented with a colour), the amplitudes are shown.

The Matlab function *specgram* gives us the spectrogram of the signals. Specgram takes four arguments (signal, length of FFT, sampling frequency, window, overlap). To see how the results differ depending on the parameters two setups were used.

Setup 1:

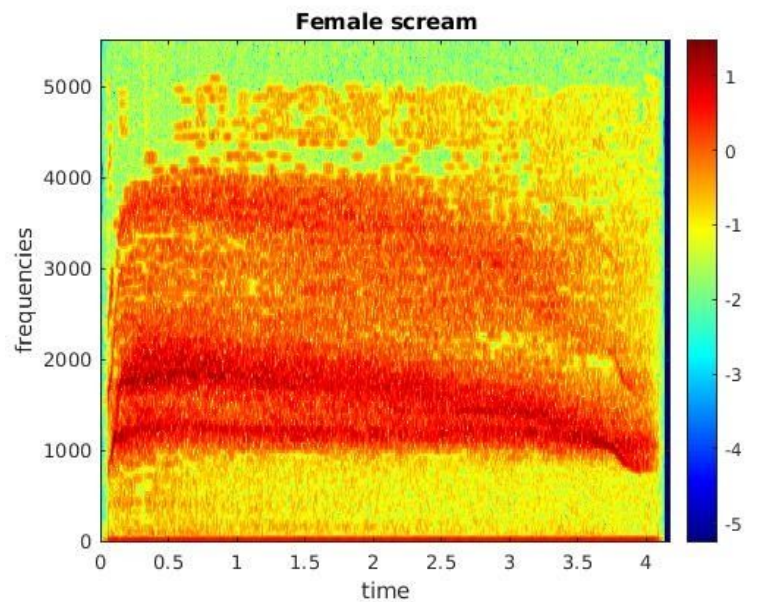
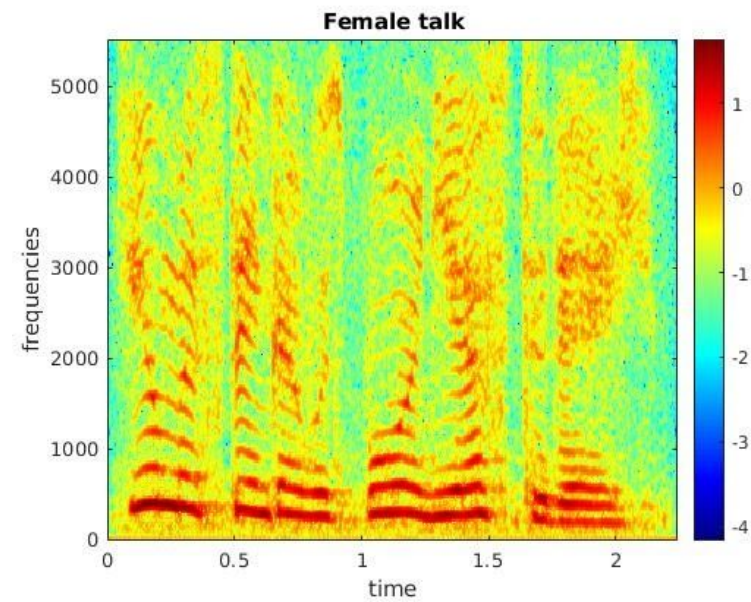
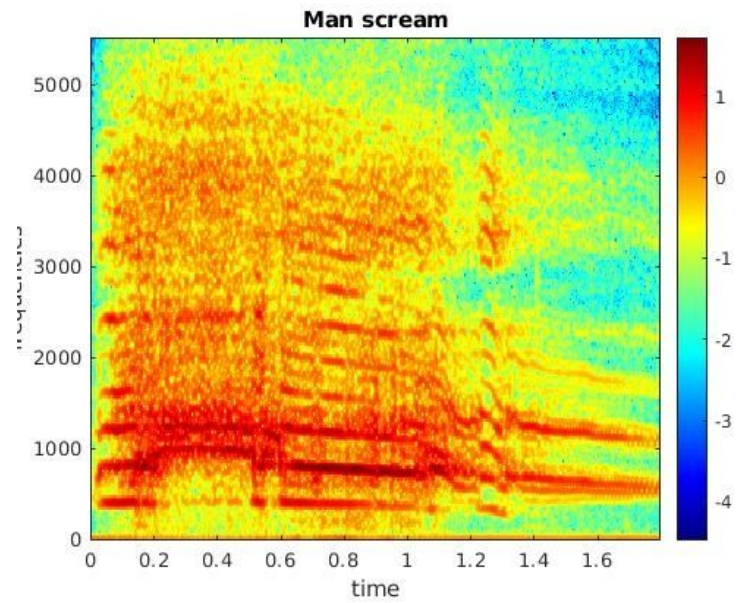
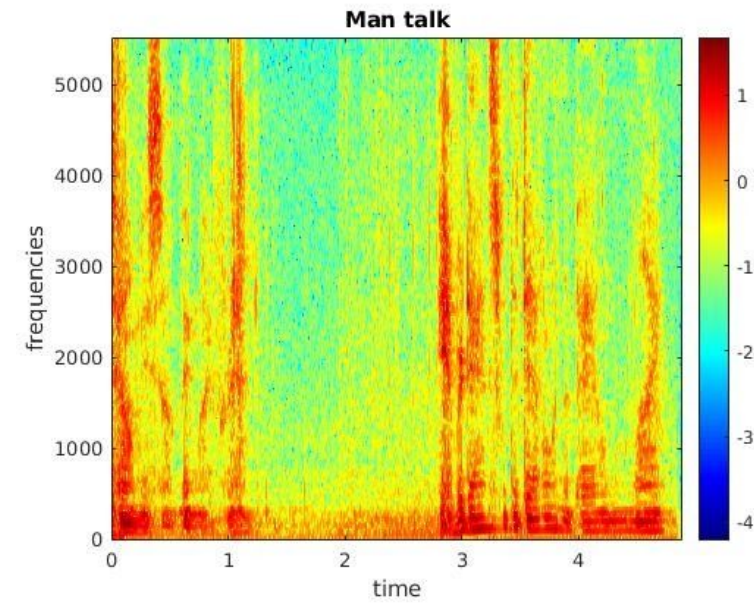
- $R = 256$;
- $N = 512$;
- $L = 35$;
- $\text{window} = \text{hamming}(R)$;
- $\text{overlap} = R - L$

Setup 2:

- $R = 512$;
- $N = 1024$;
- $L = 500$;
- $\text{window} = \text{hamming}(R)$
- $\text{overlap} = R - L$;

3. Results

Results using setup 1:



(the scale of the amplitudes is logarithmic($\log_{10}(x)$))

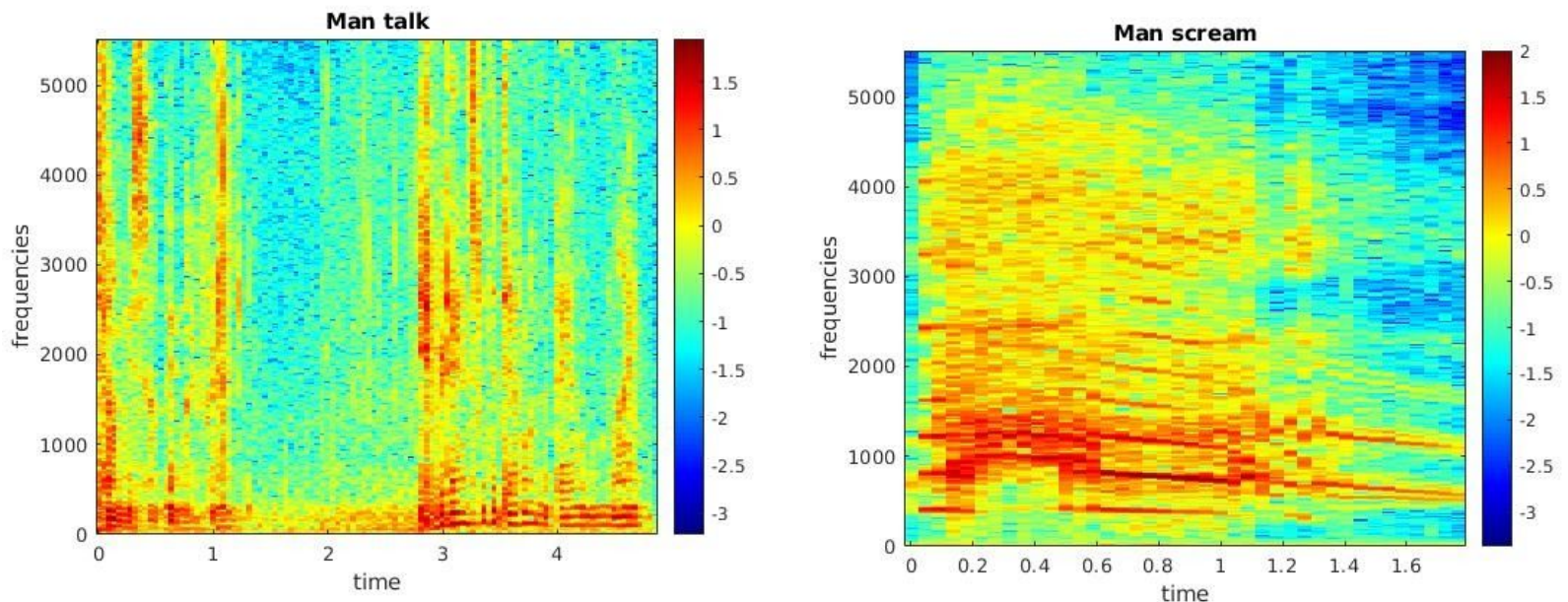
From the spectrogram we can see that the most red areas(high amplitude) in the male voices are lower than the females. This phenomenon is more visible in the screaming recordings and less visible in the normal speaking recordings. The speaking recordings have different words and those words have different vowels and consonants, so the difference in frequencies is less visible.

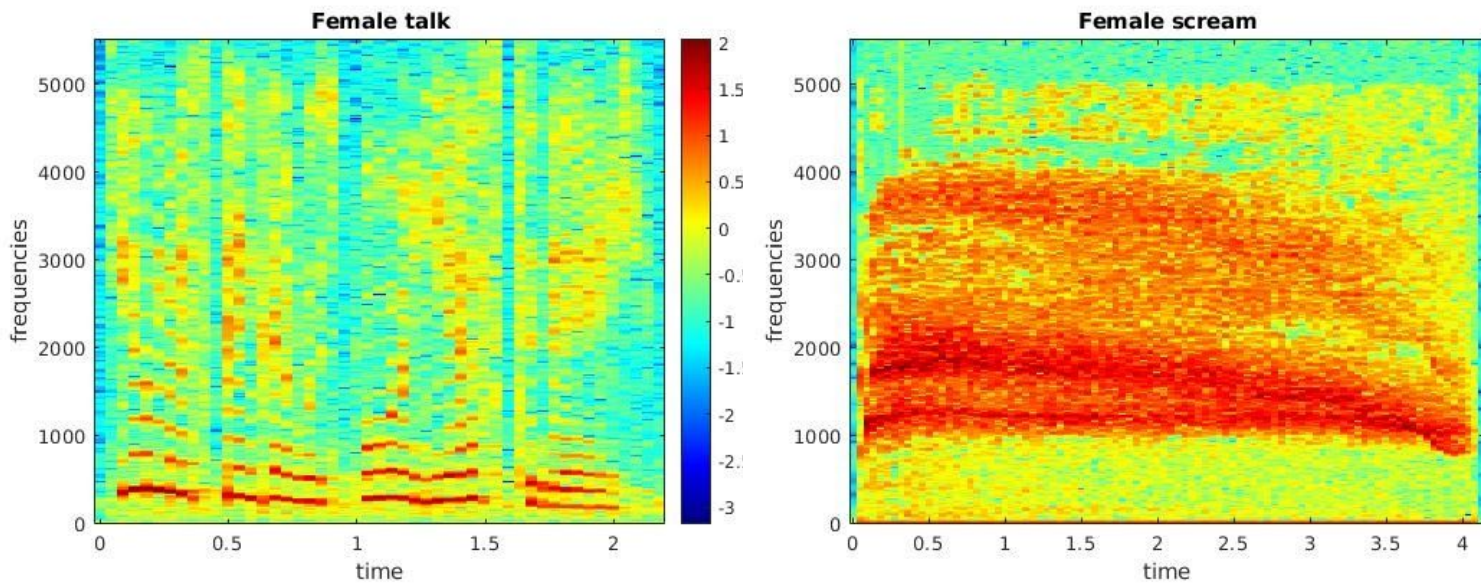
On the other hand, in the screaming records is more or less the same version of the sound performed by a man and a woman. In these two examples we can see that the frequencies are further apart and the frequencies from the female's voice are much higher than the male. From looking the spectrograms we can also see the pauses between words(high amplitudes are absent in those parts).

In this setup $L = 35$, so the frequency resolution is low while the resolution in the time space is high.

To increase the frequency resolution, we need to increase L (setup 2).

Results using setup 2:





With this approach we can now clearly spot the different frequencies that are close to one another because we increased the frequency resolution, and can confirm that the female's voice peaks are lower than male's on all examples. However, with increasing the resolution in frequency space, we lowered the resolution in 'time-space'.

4. Discussion

The results confirm the hypothesis that male voice has lower pitch(frequency) than female voice. In addition, we can compare more recordings to confirm this. Improvements could be made in the type of recordings(for example a man and a woman saying the same phrase). Also, before computing the spectrogram, we can apply a filter to eliminate any background noise that could affect the final result.