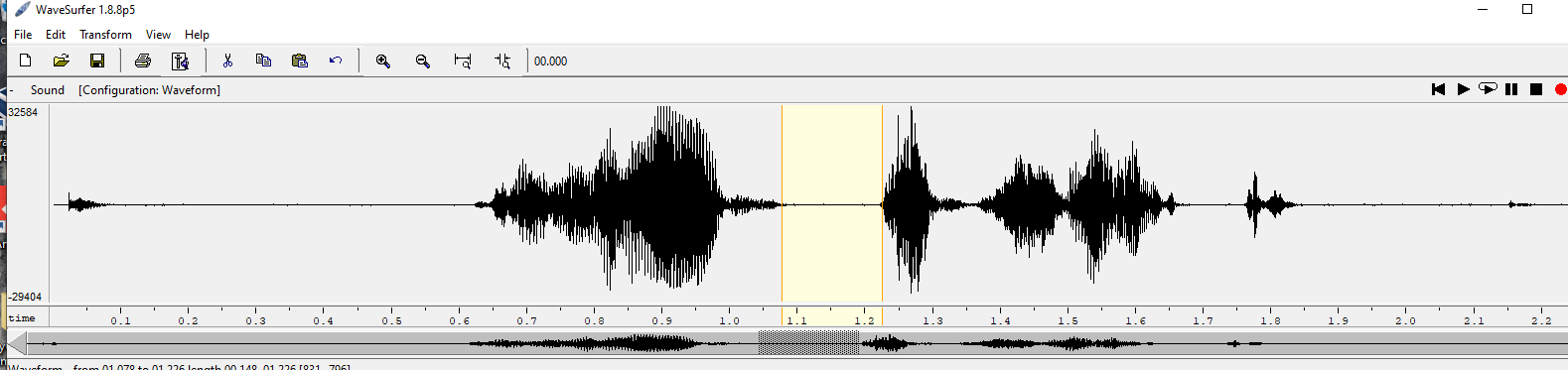
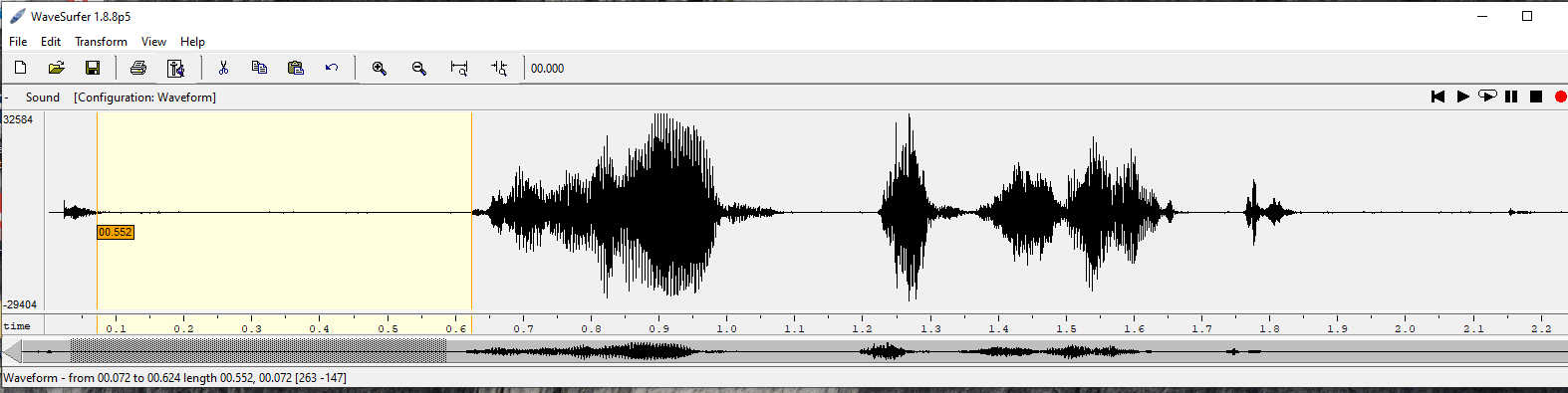


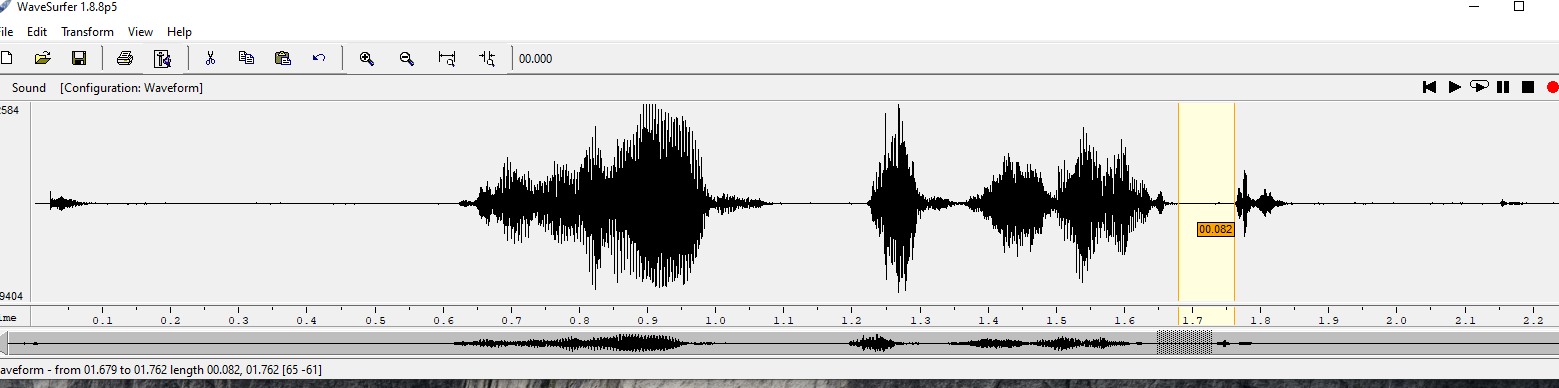
VOICED SIGNAL high pitch



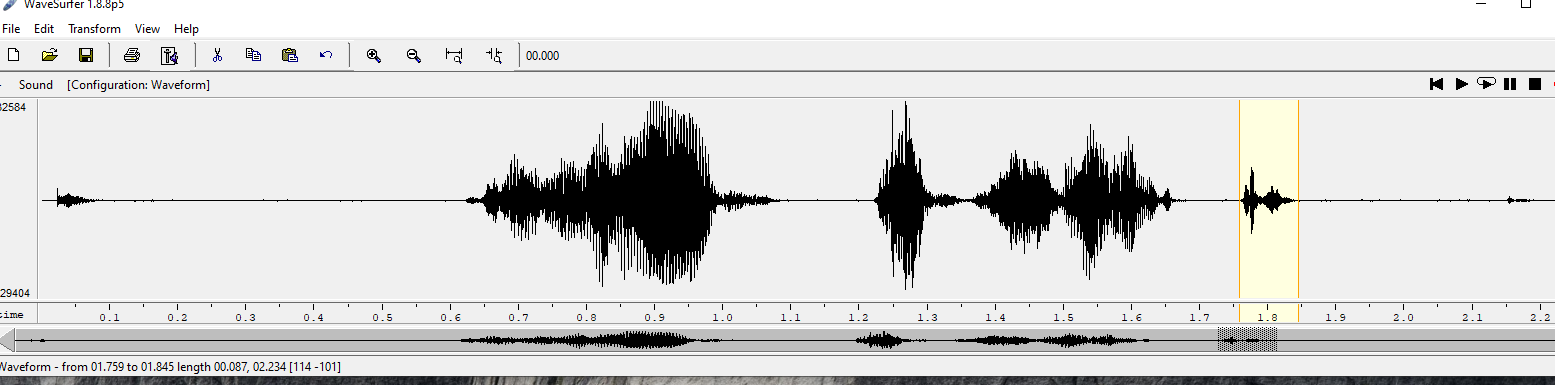
UNVOICED SIGNAL low pitch



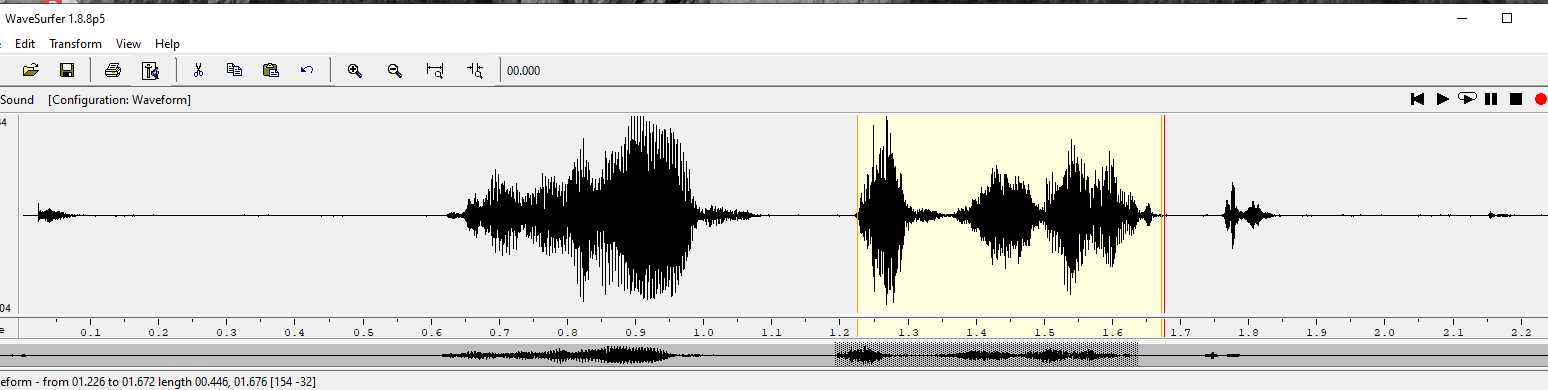
UNVOICED SIGNAL low pitch



UNVOICED SIGNAL low pitch



VOICED SIGNAL low pitch region

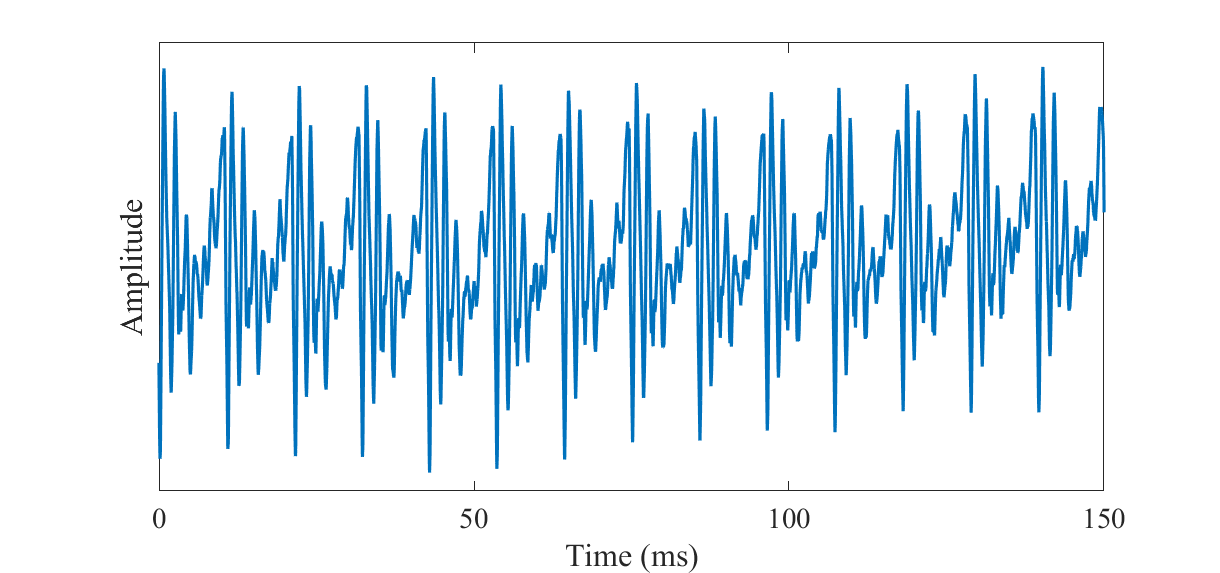


VOICED SIGNAL

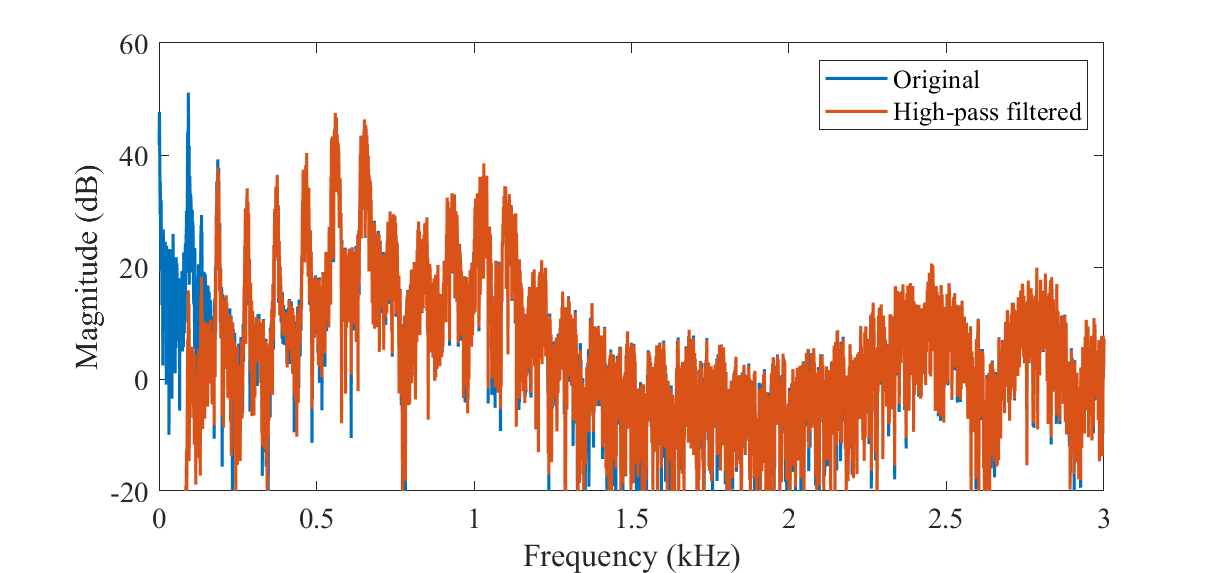
The fundamental frequency of a speech signal, often denoted by F0 or F0, refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. The oscillation originates from the vocal folds, which oscillate in the airflow when appropriately tensed. The fundamental frequency is defined as the average number of oscillations per second and expressed in Hertz. Since the oscillation originates from an organic structure, it is not exactly periodic but contains significant fluctuations. In particular, amount of variation in period length and amplitude are known respectively as jitter and shimmer. Moreover, the F0 is typically not stationary, but changes constantly within a sentence. In fact, the F0 can be used for expressive purposes to signify, for example, emphasis and questions.

Typically fundamental frequencies lie roughly in the range 80 to 450 Hz, where males have lower voices than females and children. The F0 of an individual speaker depends primarily on the length of the vocal folds, which is in turn correlated with overall body size. Cultural and stylistic aspects of speech naturally have also a large impact.

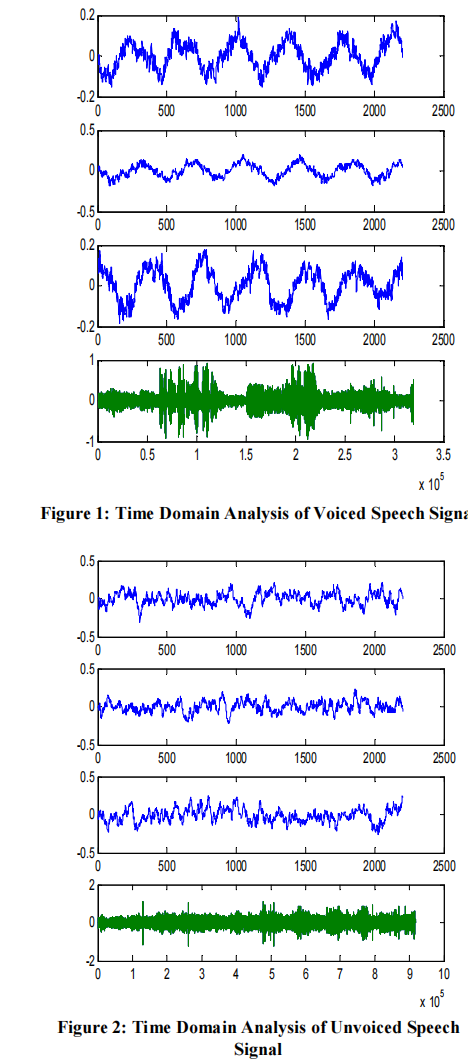
The fundamental frequency is closely related to pitch, which is defined as our perception of fundamental frequency. That is, the F0 describes the actual physical phenomenon, whereas pitch describes how our ears and brains interpret the signal, in terms of periodicity. For example, a voice signal could have an F0 of 100 Hz. If we then apply a high-pass filter to remove all signal components below 450 Hz, then that would remove the actual fundamental frequency. The lowest remaining periodic component would be 500 Hz, which correspond to the fifth harmonic of the original F0. However, a human listener would then typically still perceive a pitch of 100 Hz, even if it does not exist anymore. The brain somehow reconstructs the fundamental from the upper harmonics



The spectrum of a speech signal with a fundamental frequency of approximately F0=93Hz (original) and a high-pass filtered version of it such that the fundamental frequency has been removed (high-pass filtered).



A speech signal with a fundamental frequency of approximately F0=93Hz [aaa.wav](https://wiki.aalto.fi/download/attachments/149890776/aaa.wav?version=1&modificationDate=1597386007734&api=v2) and a high-pass filtered version of it such that the fundamental frequency has been removed [aaa\_highpass.wav](https://wiki.aalto.fi/download/attachments/149890776/aaa_highpass.wav?version=1&modificationDate=1597386029009&api=v2).



**Zero-Crossings Rate**

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero,. Speech signals are broadband signals and interpretation of average zero-crossing rate is therefore much less precise. 3 However, rough estimates of spectral properties can be obtained using a representation based on the shorttime average zero-crossing rate

The model for speech production suggests that the energy of voiced speech is concentrated below about 3 kHz because of the spectrum fall of introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced

**Short-Time Energy**

The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations The choice of the window determines the nature of the short-time energy representation. In our model, we used Hamming window. The hamming window gives much greater attenuation outside the bandpass than the comparable rectangular window. h(n) = 0.54 − 0.46cos(2πn /(N −1)), 0 ≤ n ≤ N −1

h(n) = 0 , otherwise

The attenuation of this window is independent of the window duration. Increasing the length, N, decreases the bandwidth,If N is too small, En will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, En will change very slowly and thus will not adequately reflect the changing properties of the speech signal

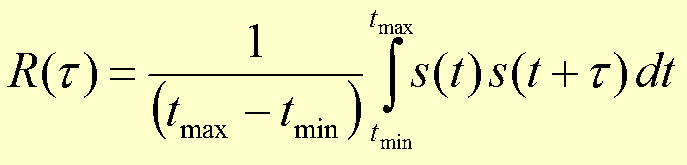
**Autocorrelation**

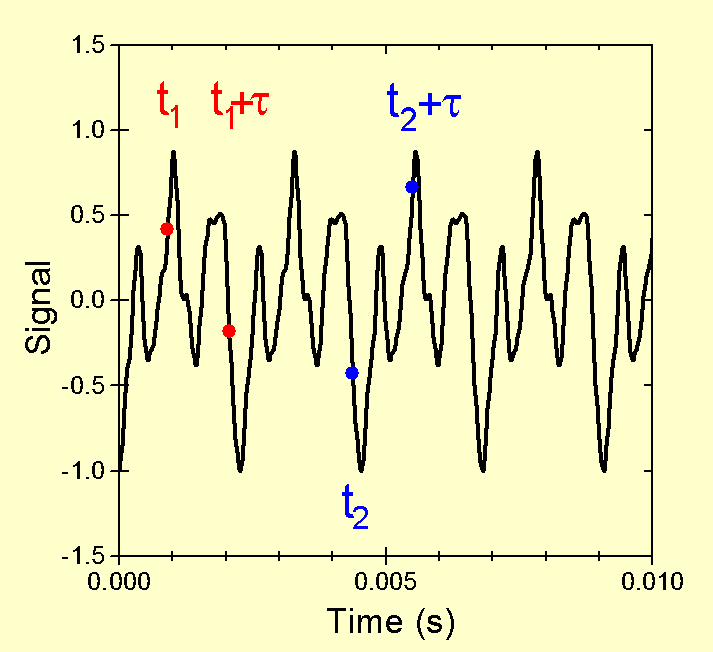
"Autocorrelation" is used to compare a signal with a time-delayed version of itself. If a signal is periodic, then the signal will be perfectly correlated with a version of itself if the time-delay is an integer number of periods. That fact, along with related experiments, has implicated autocorrelation as a potentially important part of signal processing in human hearing.

Mathematically, the autocorrelation corresponding to a delay time τ is calculated by

* (1) finding the value of the signal at a time t,
* (2) finding the value of the signal at a time t + τ,
* (3) multiplying those two values together,
* (4) repeating the process for all possible times, t, and then
* (5) computing the average of all those products.

The process can be repeated for (all) other values of τ, resulting in an autocorrelation which is a function of the delay time τ.

Mathematically, for a continuous signal, s(t), the autocorrelation, R(τ) is calculated using:  
 .



Pitch is defined as *that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale* (American Standards Association). There are a number of theories of pitch perception and these have given rise to computational models which implement them. These models have 3 stages:

1. *peripheral processing*
2. *feature analysis*
3. *pitch determination*

For pitch perception models that use temporal information some mechanism for identifying periodicities in the signal for use in the feature analysis stage is required. This demonstration does precisely this.

The usual method for deciding if a signal is periodic and then estimating its period is the autocorrelation function:  
IMG_256  
Essentially, all that is happening is the signal *x(t)* is being convolved with a time-lagged version of itself. To obtain a useful set of results, the autocorrelation function is computed over a range of lag values.  
It is an important property of the autocorrelation function that it is itself periodic. For periodic signals the function attains a maximum at sample lags of 0, +-P, +-2P, etc. where P is the period of the signal.

One major limitation of the autocorrelation function is that it can retain too much information present in the signal. In speech, numerous peaks present in the autocorrelation function are due to damped oscillations of the vocal tract response. If these peaks happen to be bigger than the peaks due to periodicity, the simple procedure of picking the largest peak to be the period will fail.

Therefore, the signal needs to be pre-processed in some way to make the periodicity more prominant while suppressing other features which may cause distracting peaks. Such pre-processing techniques are sometimes called *spectrum flatteners*. Many techniques have been proposed but **centre clipping** appears the best for this situation.

Centre clipping works by clipping a certain percentage of the waveform. Let Amax be the maximum amplitude of the signal and CL be the clipping level. CL is a fixed percentage of Amax (say 30%). Therefore, the output from the center clipper is as follows:

y(n) = x(n)-CL [x(n)>CL]  
y(n) = 0 [x(n)<=CL]