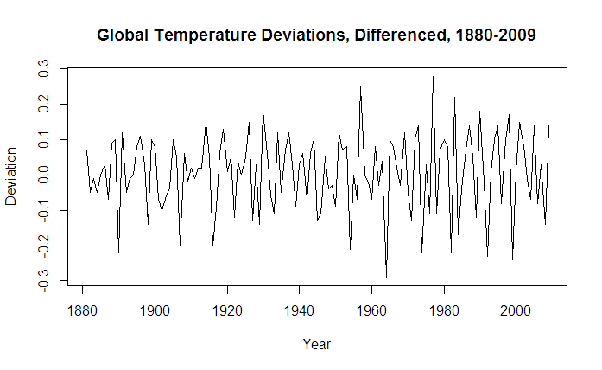
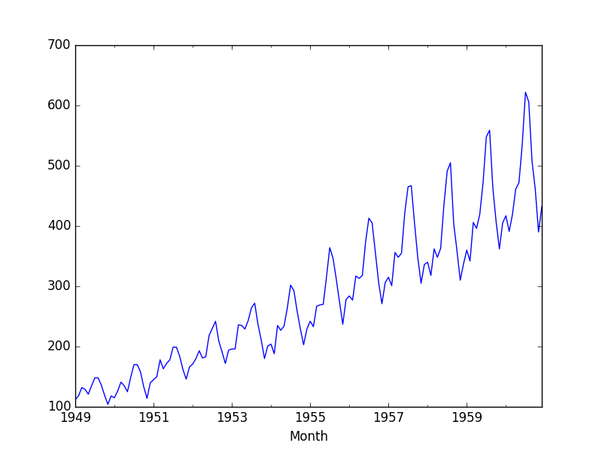
**Glossary of environmental sound**

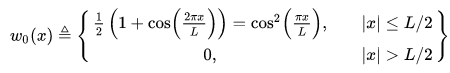
* **Environmental sound** - Environmental noise is defined as the unwanted or harmful outdoor sound created by human activity, such as noise emitted by means of transport, road traffic, rail traffic, air traffic, and industrial activity. It can be natural or artificial (Other than speech and music).
* **Environmental sound recognition -** The process of detecting and classifying the environmental sound is called as the environmental sound recognition.
* **Machine audition/computer audition -** Computer audition (CA) or machine listening is a general field of study of algorithms and systems for audio understanding by machine.
* **Example based search -** In example-based search the user, or the analyst circumvent query languages by using examples as input.
* **Content-based image retrieval (CNIR)/Query by image content (QBIC)** - It is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases.
* **Metadata** - Data that provides information about other data. Such as keywords, tags, or descriptions associated with the image.
* **Keyword-based audio retrieval -** Most of the audio retrieval systems are based on keyword/title/author search typed into the system by users. The system then searches for particular keywords and gives a list of entire audio files that potentially are relevant to the query.
* **Robot navigation -** Robot navigation means the robot's ability to determine its own position in its frame of reference and then to plan a path towards some goal location.
* **Speech -** The communication or expression of thoughts in spoken words. When people speak there are pauses between words, sentences, thoughts, etc.
* **Music** - Vocal or instrumental sounds (or both) combined in such a way as to produce beauty of form. Music (with multiple instruments) is fairly continuous. Even if one instrument stops another is probably still playing.
* **Stationary sound** - The wind provides stationary noise: it’s mean and variance does not appreciably change over the course of the day.
* **Non-stationary sound**  - Bug noises, bird chirps, traffic, etc. are nonstationary: they are in fact cyclical and the mean and variance can be modeled using a Fourier expansion. Bugs get noticeably louder at night; car noise gets louder around 7 am-9 am and 4 pm-6 pm, etc. If you \*remove\* these types of sound, you might get something close to stationary. Data points are often non-stationary or have means, variances, and covariances that change over time. Non-stationary behaviors can be trends, cycles, random walks, or combinations of the three.
* **Stationary time series** - In stationary time series the statistical measures such as the mean, standard deviation, autocorrelation are somewhat similar over time. It has no trend.



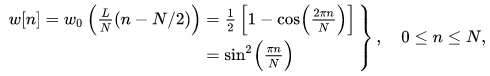
* **Non-stationary time series** - In such a time series the statistical measures such as the mean, standard deviation, autocorrelation shows a decreasing or increasing trend over time. It has a trend. The below plot shows an increasing trend.



* **Quotidian**  - Of or occurring every day; daily.
* **Phonetic structure**  - Phonetics is a branch of linguistics that studies how humans make and perceive sounds.
* **Databases** - A database is an organized collection of data, generally stored and accessed electronically from a computer system.
* **Hidden Markov Model**  - Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process – call it – with unobservable ("hidden") states.
* **Phonemes**  - In phonology and linguistics, a phoneme is a unit of sound that distinguishes one word from another in a particular language.
* **Temporal**  - Relating to time.
* **Melody**  - the aspect of musical composition concerned with the arrangement of single notes to form a satisfying sequence.
* **Rhythm**  - The measured flow of words and phrases in verse or prose as determined by the relation of long and short or stressed and unstressed syllables.
* **Spectral**  - of, relating to or made by a spectrum
* **Sequential learning -** Sequential learning is a type of learning in which one part of a task is learned before the next.
* **Framing based processing** - The process of splitting the input signal into multiple frames and processing each frame at a time.
* **Hanning window** - The Hann function of length L, used to perform Hann smoothing, is named after the Austrian meteorologist Julius von Hann, is a window function given by:



For digital signal processing, the function can be sampled symmetrically as:



where the length of the window is {\displaystyle N+1,}{\displaystyle N+1,} and N can be even or odd. It is also known as the raised cosine window, Hann filter, von Hann window.

* **Hamming window** -



A ‘window’ is a tool used to process discrete-time data and analyze the spectrum (frequency domain). In practice, only a subset of the time domain data can be processed. When a finite set of data is captured in the time domain, there are spectral artifacts due to this truncation. This is a ‘rectangular’ window. The key to understanding this relationship is,

x(n) = w(n)s(n) then transforming into the frequency domain,

X(f) = W(f) \* S(f)

Here the signal is s(n) in the time domain and the window is w(n). As this is transformed into the frequency domain, the multiplication in time results in convolution in frequency. So, the window, w(n), transforms to W(f) and this spectrum is convolved with the desired spectrum, S(f). If windowing is not used, then w(n) = rect(n/T) where T is the sample width.

w(n) = rect(n/T)

W(f) = sinc(n/T)

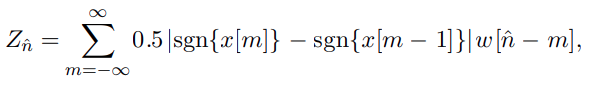
A sinc() function is convolved with the desired spectrum. The convolution ‘smears’ the desired information and makes it difficult to observe the desired spectrum, S(f).

So w(n) can be changed. As the ‘sharpness’ of the time domain waveform is reduced, the spectral content is reduced. There are many different window functions that have been employed to impact the shape of this convolution. All of these windows attenuate the ‘edges’ of the data relative to the middle of the data to reduce the W(f) spectral bandwidth.

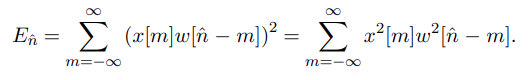
The Hamming window is one of these. Hanning is another. I have seen Blackman-Harris used the most but there are trade-offs. It is more important to understand what is going on so the user can recognize the limitations to the spectral measurement. A good tool to understand the window properties is to take a DFT (FFT) of only the window. This is W(f). This provides some intuition into the artifacts introduced in the process.

Another tool is to understand that if the data to be processed in the time domain is periodic, a sample rate could be chosen to align the zero crossings of the sinc() function with the sample rate. That is, an integer number of periods must be exactly captured with the sample size. This transforms the W(f) for a rectangular (no window if you will) into an impulse, so the output is only X(f).

* **Features -** a distinctive attribute or aspect of something.
* **Audio frame** - An audio frame, or sample, contains amplitude (loudness) information at that particular point in time.
* **Feature vector -** A feature vector is a vector containing multiple elements about an object.
* **Majority voting rule** - Majority rule is a decision rule that selects alternatives that have a majority, that is, more than half the votes.
* **Correlation**  - Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5'' is less than the average weight of people 5'6'', and their average weight is less than that of people 5'7'', etc. Correlation can tell you just how much of the variation in peoples' weights are related to their heights.
* **Zero-crossing rate**  - The zero-crossing rate is the rate of sign-changes along with a signal, i.e., the rate at which the signal changes from positive to zero to negative or from negative to zero to positive.
* **Short time zero-crossing rate -**  The short-time zero-crossing rate is defined as the number of times the speech signal changes sign within a given time window.



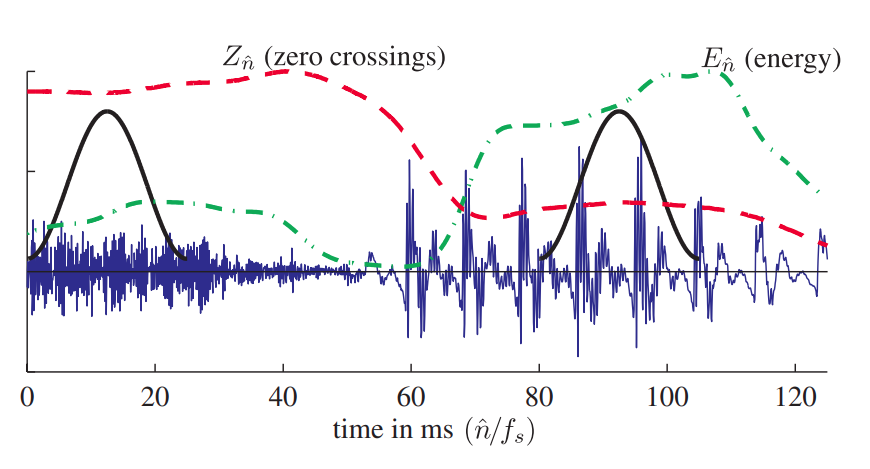
* **Short time energy** - Short time energy can be expressed as below:



* Short time energy and short-time zero-crossing rate is a simple but effective way to distinguish between voiced and unvoiced speech region.

Voiced region: higher energy, lower zero-crossing rate

Unvoiced region: lower energy, higher zero-crossing rate.



* **Spectral flux** - Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame.
* **Cepstral** - It is the result of a mathematical transformation in the field of Fourier Analysis. It serves as a tool to investigate periodic structures within frequency spectra.
* **Cepstral features**  - The cepstrum is a representation used in homomorphic signal processing, to convert signals combined by convolution (such as a source and filter) into sums of their cepstra, for linear separation. In particular, the power cepstrum is often used as a feature vector for representing the human voice and musical signals.
* **Mel frequency cepstral coefficient** - Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

MFCC is obtained in the below steps:

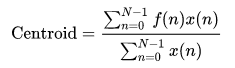
1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

* **Sampling rate**  - The Shannon Theorem or Nyquist Law requires that the sampling rate is higher than twice the maximum frequency in the digitized signal. Since speech is more or less located below 8kHz most speech corpora have a sampling rate of 16kHz minimum. Exceptions are telephone recording where the bandwidth is technically reduced to 300Hz - 3300Hz and usually a sampling rate of 8kHz is used. Since the audio CD standard was introduced with 44,1kHz also the dividers 22,05 kHz and 11,025 kHz are used because some audio devices do not process other sampling rates than these. This is also the reason why we recommend avoiding `exotic' sampling rates, whenever possible.

Laryngograph signals are usually sampled at the same frequency as the speech signal. Because of their low bandwidth speech movement data (e.g. EMA[4.5](https://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/footnode.html#foot341)) can be sampled at about 200 Hz.

* *Telephone Recording: 8kHz*
* -site or field recording: 16kHz, 22,05kHz
* signal: minimum 16kHz
* *EMA signals: 200Hz*
* **MPEG -7** - MPEG-7 is a multimedia content description standard. It was standardized in ISO/IEC 15938 (Multimedia content description interface). This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. MPEG-7 is formally called the “Multimedia Content Description Interface”.
* **Psychoacoustic -** it is the branch of science studying the psychological responses associated with sound (including noise, speech, and music).
* **Spectral centroid** - The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the center of mass of the spectrum is located. Perceptually, it has a robust connection with the impression of the brightness of a sound.

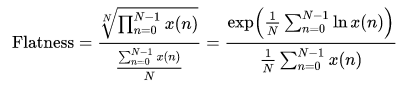
It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights:



where x(n) represents the weighted frequency value, or magnitude, of bin number n, and f(n) represents the center frequency of that bin.

* **Spectral flatness** - Spectral flatness or tonality coefficient, also known as Wiener entropy,[3][4] is a measure used in digital signal processing to characterize an audio spectrum. Spectral flatness is typically measured in decibels and provides a way to quantify how tone-like a sound is, as opposed to being noise-like.

The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum, i.e.:



where x(n) represents the magnitude of bin number n.

* **SVM**  - Support vector machine
* **KNN**  - K nearest neighbor
* **Scoring rule** - In decision theory, a score function, or scoring rule, measures the accuracy of probabilistic predictions.
* **Fisher’s discriminant ratio** - The Fisher's linear discriminant (FLD) ratio projects high dimensional data onto a line and performs classification of pixels.
* **Linear discriminant analysis** - The philosophy behind the linear prediction is that a speech sample can be approximated as a linear combination of past samples. Then, by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones over a finite interval, a unique set of predictor coefficients can be determined.
* **Autoregressive model** - In statistics, econometrics, and signal processing, an autoregressive (AR) model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc.
* **Linear prediction cepstral coefficients (LPCC) -** Linear prediction cepstral coefficients (LPCC) are cepstral coefficients derived from LPC calculated spectral envelope [11]. LPCC are the coefficients of the Fourier transform illustration of the logarithmic magnitude spectrum [30, 31] of LPC. The cepstral analysis is commonly applied in the field of speech processing because of its ability to perfectly symbolize speech waveforms and characteristics with a limited size of features.
* **Souce filter model** - The source-filter model represents speech as a combination of a sound source, such as the vocal cords, and a linear acoustic filter, the vocal tract. While only an approximation, the model is widely used in a number of applications such as speech synthesis and speech analysis because of its relative simplicity.
* **Code-excited linear prediction (CELP) -** Code-excited linear prediction (CELP) is a linear predictive speech coding algorithm originally proposed by Manfred R. Schroeder and Bishnu S. Atal in 1985. It is also used in MPEG-4 Audio speech coding. CELP is commonly used as a generic term for a class of algorithms and not for a particular codec.
* **Pitch** - The quality of a sound governed by the rate of vibrations producing it; the degree of highness or lowness of a tone.
* **Fast Fourier Transform (FFT) -** A fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The DFT is obtained by decomposing a sequence of values into components of different frequencies.
* **Filter bank** - In signal processing, a filter bank is an array of band-pass filters that separates the input signal into multiple components, each one carrying a single frequency sub-band of the original signal. One application of a filter bank is a graphic equalizer, which can attenuate the components differently and recombine them into a modified version of the original signal. The process of decomposition performed by the filter bank is called analysis (meaning analysis of the signal in terms of its components in each sub-band); the output of the analysis is referred to as a subband signal with as many subbands as there are filters in the filter bank. The reconstruction process is called synthesis, meaning the reconstitution of a complete signal resulting from the filtering process.
* **Discrete cosine transform** - The discrete Fourier transform (DFT) transforms a complex signal into its complex spectrum. However, if the signal is real as in most of the applications, half of the data is redundant. In the time domain, the imaginary part of the signal is all zero; in the frequency domain, the real part of the spectrum is even symmetric and the imaginary part odd. In comparison, Discrete cosine transform (DCT) transforms is a real transformation that transforms a sequence of real data points into its real spectrum and therefore avoids the problem of redundancy. Also, as DCT is derived from DFT, all the desirable properties of DFT (such as the fast algorithm) are preserved.
* **Autocorrelation function (ACF)** - Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies.
* **Power spectral density (PSD)** - A Power Spectral Density (PSD) is the measure of the signal's power content versus frequency. A PSD is typically used to characterize broadband random signals.
* **Discrete wavelet transform (DWT)** - In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time).