Introduction

- It was designed for stationary signals, and became soon the basic analysis tool in signal processing.
- The time-frequency (TF) analysis is a time-dependent extension of classical Fourier-based methods.
- Research topic of the past 3 decades.

Reasoning

- Having two signals, one may be interested in how they are similar and/or close to each other.
- The angle is a good indication about their similarity.
  - small, the vectors point to the same direction
  - large, the angle is 90∘, the vectors are dissimilar, calling also *orthogonal*.

Definition

- One way how to express this similarity between two vectors x and y is using the inner product:

$$< x, y >= \sum_k x[k].y[k] \tag{1}$$

- The inner product is large for similar sequences
- and zero for orthogonal ones.

Properties

- If the two sequences are orthogonal, also their linear combination is orthogonal.
- Two sinusoids with $f_1 \neq f_2$ are orthogonal. Orthogonality plays an important role in Fourier transforms as it simplifies a lot of calculations.

Closeness

Closeness is an another interpretation of the inner product. It is defined using the square of the norm between x and y as:

$$|x-y|^2 = |x|^2 - 2 < x, y > + |y|^2 \tag{2}$$

where the closeness really depends only on the inner product of the sequences.

Definition

- The correlation between two sequences with shift j can be directly defined by their inner product:

$$R_{xy}[j] =< x[k], y[k+j] >= \sum_{i=0}^{N-1} x[i].y[i+j] \qquad (3)$$

- (Cross)correlation shifts one sequence in time and multiplies point by point and sums it.
  - The sum is large if many points are similar – close to each other.
  - The sum is small if many points are different – far from each other.

Autocorrelation

- An useful version of the correlation is when the second sequence is just shifted first sequence. Then, (auto)correlation R_{xx} shows periodicity of the sequence x (if it is periodic).
- It has the largest value always at the shift of $j = 0$ – when there is no shift.
- For periodic signals the peaks are at values of j, and they correspond to the period if the signal.

Definition

- The Fourier transform decomposes a signal into a family of simple sine waves
- Each of them is characterised by amplitude, frequency and phase.

$$X[n] =< x[k], e^{-j2\pi nk/N} > . \tag{4}$$

How to read definition?

- Tt is the inner product of the signal $x[k]$ (in the time domain) and the complex-valued sinusoid $(e^{\pm j\theta} = cos(\theta) \pm j.sin(\theta))$.
- $X[n]$ is a correlation of the signal $x[k]$ and a sinusoid of frequency n/N.
- m is the magnitude and the phase θ of a sine wave that is closest to the $x[k]$.

Frequency resolution

- Discretization of the frequency introduces a finite sets of frequencies used for the analysis.
- The frequency can take only values of all integer multiplies n of $2\pi/N$.
- The frequencies thus differ in a fixed difference called frequency resolution:

$$resolution(Hz) = \frac{samplingRate}{windowSize} \tag{5}$$

The length of analysis windows

Why not use long windows to increase the frequency resolution? We have to consider following balance:

- long windows increase frequency resolution but decrease time resolution
- short windows decrease frequency resolution but increase time resolution

Short-time Fourier transform

- When the signal is too long, a short-time Fourier transform (STFT) is used.
- The signal is split into overlapping quasi-stationary short-time signal using the windowing.
- The STFT transform is defined as

$$X_{STFT}[n,i] =< x[k], w[k-i]e^{-j2\pi nk/N} > . \quad (6)$$

- The STFT is a function of the frequency variable n and time shift i (specifying where the window is nonzero). The transform is fully invertible by

$$x[k] =< X_{STFT}[n,i], w[k-i]e^{-j2\pi nk/N} > . \quad (7)$$

Phase vocoder

- The frequency resolution of the Fourier transform is defined by the sampling rate and the window size.
- The resolution can be significantly increased by using the phase spectra.
- Frequency can be estimated from phases of a signal calculated at different times.

Analysis

- The analysis of the phase vocoder is based on locating peaks in the magnitude spectrum of two different (neighbouring) waveform segments (windows).
- The better frequency estimate f_n, using the phases of the compound partials θ_1 and θ_2 (phase spectra for the waveform segments differs):

$$\begin{aligned} \theta &= 2\pi f t \\ \Delta\theta &= 2\pi f \Delta t \\ &= 2\pi f_n \Delta t + n2\pi \\ f_n &= \frac{\Delta\theta - n2\pi}{2\pi\Delta t} \end{aligned} \tag{8}$$

Frequency estimate

- Finally, having two neighbouring windows with the magnitude peaks at t_1 and t_2 ($\Delta t = t_2 - t_1$)), the better estimate of the frequency of the peaks is defined by

$$f_n = \frac{(\theta_2 - \theta_1) - 2\pi n}{2\pi(t_2 - t_1)} \tag{9}$$

- where the phase θ_1 corresponds to the time t_1, and the phase θ_2 corresponds to the time t_2.

Re-synthesis

1. The re-synthesis is the same as with STFT method.
2. First invert Fourier transform is applied on the analysed signals.
3. Then put together with overlap and add.

Motivation

- Short-time Fourier transform is based on windowing the signal and its decomposition into a family of simple sine waves.
- Each of them is characterised by amplitude, frequency and phase.
- Wavelet transform can be seen analogically as a decomposition of the signal into other family of basis signals.
- However, the windows are incorporated directly into the basis signals.
- Variety of non-sinusoidal basis function exit, and they are called *"mother wavelets"*.

Definition

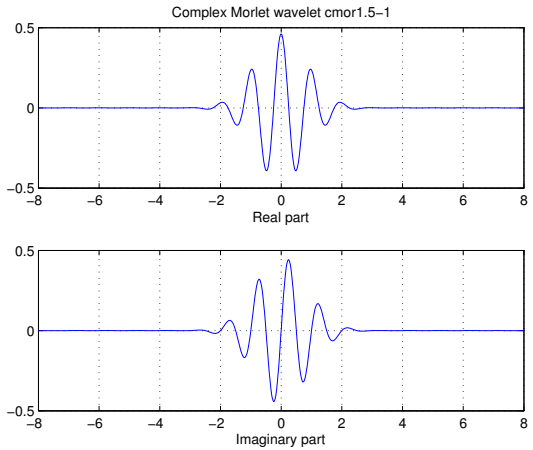
- The continuous wavelet transform W, defined similarly as Fourier transform, is the inner product of the continuous input signal and parametrized mother wavelet:

$$W(a,b) =< x(t), \psi_{a,b}(t) > . \tag{10}$$

- The b parameter shifts the wavelet in time (and it is analogous to the shift of the window in STFT).
- The a parameter is analogous to the stretches and compresses the wavelet (and it is analogous to use of sinusoids with different frequency and amplitude).

Example

- Let us suppose a complex Morlet function as a mother wavelet (shown at Fig. 19).



- Then, parametrized $\psi_{a,b}$ is defined as:

$$\psi_{a,b} = \frac{1}{\sqrt{|a|}} \psi \left(\frac{t-b}{a} \right), \tag{11}$$

where a is the scale and b is the shift of the mother wavelet ψ.

Why wavelets?

- The inner product says that it has large values as the signals are aligned (or similar). And it has small values for different (dissimilar - zero for the orthogonal) signals.
- That means that having prior information about the character of the signal might be useful for wavelet transform, and the inside into the signal may be better then with the Fourier transform.
- However it is valid also in the opposite way, the bad selection of the mother wavelet could result into less useful insight.

Differences with STFT

- The relation of the frequency of the waveform and duration of the wavelet.
- Typical wavelets contain the same number of cycles independently of the scale of the wavelet. Thus it allows more precise localisation of the high frequency components
  1. With higher scale - roughly frequency of the wavelet - the input signal is analysed with the same number of wavelet cycles.
  2. With lower scale: the lower frequency components are so wide in time and higher short in time.

Introduction

- The same tone played on a guitar and a piano sounds differently.
- This is valid for human voice as well – different timbre of voice, known also as a colour of voice or a voice quality.
- Estimation of the voiced excitation is referred as a glottal flow estimation. Glottal flow estimation is performed using source-tract decomposition.
- The most straight forward solution is to use source-filter deconvolution using LPC modelling.

Speech production

- The larynx has the size of the smallest piccolo
- An instrument consists of :
  1. A sound source that defines pitch and timbre
  2. Resonators that reinforce the F0
  3. Radiation surface / orifice for transfer into free air
- Human's air tube: 15 – 20 cm above larynx and 12 – 15 cm below. A trumpet has the tube about to 2 meters, a trombone 3 meters.

Source of a musical instrument

Design of a source of a musical instrument:

- For a reed or string to sustain its vibration, it must be elastic that is defined by its stiffness or tension T. To double its frequency f, one must quadruple the string tension.

$$f = \frac{\sqrt{\frac{T}{m/L}}}{2L} \tag{12}$$

  where m is mass string.
- Another mechanisms for changing frequency:
  1. Altering the length L of a string
  2. Skipping to another string:
- Players almost never can change both the length and tension simultaneously – singers to that!

Source of the quitar

Skipping to another string:

Figure: Strings

Source of the human voice

Design of vocal folds:

- To get maximum increase in frequency, folds should increase the tension while decreasing the length. This requires unusual material:
  1. A ligament that looks like cords (non-linear)
  2. 90% of the volume of the vocal folds is muscle tissue. Muscle fibres can raise the tension even they are shortened.
  3. A mucous membrane for needed air-driven oscillation to occur

Vocal quality – phonation types

The two phonatory organs lungs and larynx create the voice source signal that adjust the pitch, loudness and voice quality. There are other forms of the vocal fold movements that do not fall clearly into the three primary states:

- Creaky voice: folds are very tense small portion oscillates, resulting in harsh-sounding voice.
- Vocal fry: folds are massy and relaxed, characterised by secondary glottal pulses close to and overlapping the primary glottal pulse within the open phase.
- Diplophonic voice: again secondary pulse but within the closed phase, away from the primary pulses.

Voice quality control

1. Glottal chink: partially opened glottis during the closed phase (in soft and breathy voice).
2. Sharp vs. soft voice : open phase of the glottal cycle is shorter vs. longer.
3. Glottal airflow is sinusoidal with open and closing phases are longer $\rightarrow$ weak harmonics.
4. When open phase is shorter the airflow are pulsating waves with rich harmonics.
5. Modal voices vs. falseto: all the vocal folds layers vibrate vs. only the edges of the folds vibrate (incomplete closure with reduced harmonic components).

Resonators of the human voice

Vocal tract (resonators):

- All source frequencies are integer multiplies (2:1, 3:1, 4:1, . . .) – resonator must be quite large; French horn tubing uncoils from 3.7 to 5.2 m.
- Singer's resonator: only about 17 cm long ??? including only the odd-integer multiplies (1,3,5) and cannot change the length.
- Vocal tract reinforces a cluster of harmonics simultaneously by using energy feedback process (pushing someone someone on a playground swing).
- The human voice system behaves nonlinearly:
  1. Linear system: the outputs are proportional to the inputs (such as LPC modelling)
  2. Nonlinear system: In a nonlinear feedback system, small changes can result in disproportionately large effects.

Linear prediction

- Linear prediction can be viewed as a generative model of the speech production:

$$y[k] = -\sum_{p=0}^{P} \alpha_p x[k-p] + r[k] \tag{13}$$

  where generated speech $y[k]$ is calculated as a linear prediction of p past samples and an LP residual $r[k]$.
- The two components represent the vocal tract model and the excitation.

Excitation signal

- Excitation signal can be calculated using inverse filtering. In the z domain it becomes:

$$\begin{aligned} r(z) &= \frac{y(z)}{H(z)} \\ &= y(z) - \sum_{p=1}^{P} \alpha_p z^{-p} \end{aligned} \tag{14}$$

- $H(z)$ has a number of pairs of complex-conjugate poles, more commonly referred to as resonances or formants.
- The inverse-filter has a pair of complex-conjugate zeroes, more commonly referred to as an anti-resonance, for every vocal tract formant in the frequency range of interest.

Glottal waveform examples

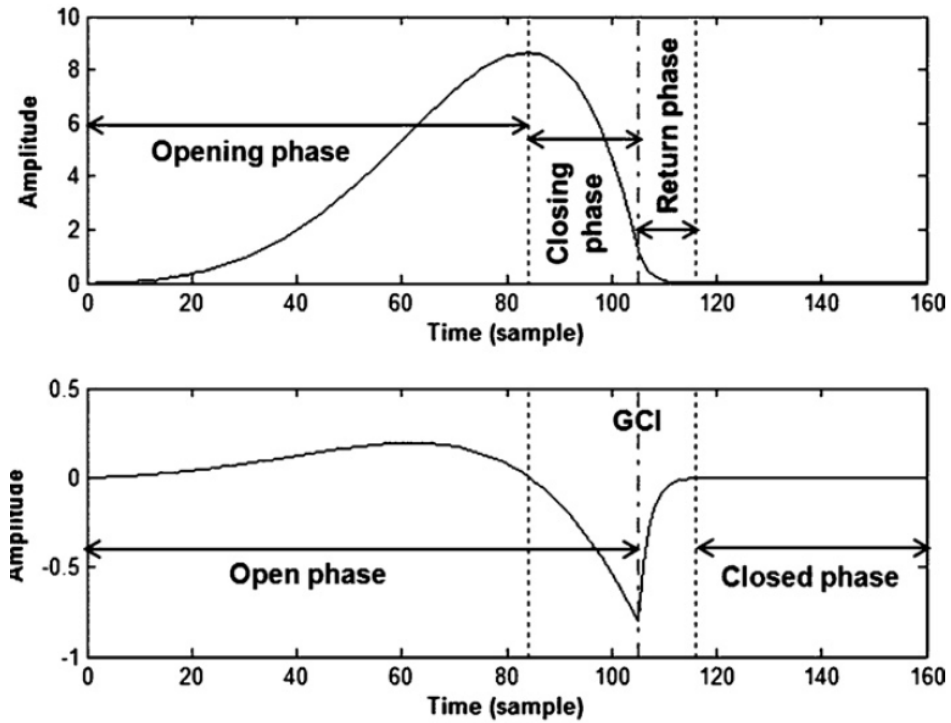


Figure: Glottal waveform examples according to the Liljencrants–Fant model: (top) glottal flow example, (bottom) glottal flow derivate example. Glottal closure events (GCI) are also indicated.

Glottal-synchronous speech processing

- If the filtering is performed during the closed phase of the glottal source cycle, the method is referred as *closed phase inverse filtering*.
- In other words, the excitation is estimated using glottal-synchronous speech processing such parts of the speech signal where glottal folds are closed.
- This event is associated with significant excitation and large LP residual, referred as *glottal closure instant*.
- During closed vocal folds, the generated speech can be modelled just by $H(z)$, and so the excitation estimation should be more precise (information related to the formants is removed).

Hilbert Envelope-Based Method

- Smooth LP residual first. Most common smoothing is the Hilbert envelope of the LP residual, defined as a magnitude of analytic signal $r_a[k]$:

$$\begin{aligned} h_e[k] &= \mid r_a[k] \mid \\ &= \mid r[k] + r_h[k] \mid \\ &= \sqrt{r^2[k] + r_h^2[k]} \end{aligned} \tag{15}$$

- $r_h[k]$ is the Hilbert transform computed as inverse Fourier transform of imaginary part of Fourier transform of the $r[k]$:

$$r_h[k] = IFFT[jDFT[r[n]]] \tag{16}$$

- The peaks of the $h[k]$ signal represent the glottal closure instants.

Zero Frequency Resonator-Based Method

1. Remove DC of the input speech signal $s[k]$

$$x[k] = s[k] - s[k-1]. \tag{17}$$

2. Pass $x[k]$ through a cascade of two ideal digital resonators at 0 Hz

$$\begin{aligned} y_1[k] &= x[k] + 2y_1(k-1) + y_1[k-2] \\ y_2[k] &= y_1[k] + 2y_2[k-1] + y_2[k-2] \end{aligned} \tag{18}$$

3. Compensate the exponential trend of y_2 by a mean-subtraction operation

$$y[k] = y_2[k] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[k+m] \tag{19}$$

4. The positive zero crossings of the trend-removed signal y[k] will give the locations of he glottal closure instants.

Zero frequency resonators

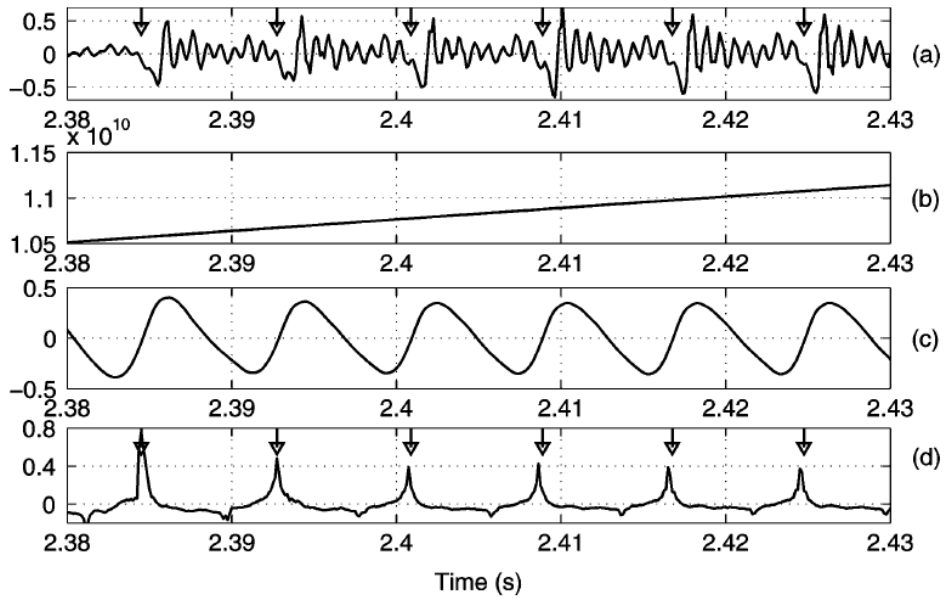


Figure: Zero frequency resonator-based extraction of glottal closure events (arrows). (a) A speech segment. (b) Output $y_2[k]$ of the two digital resonators at 0 Hz. (c) Compensated exponential trend $y[k]$. (d) Reference DEGG signal.

DEGG is differential electroglottograph signal, a measure of the impedance between the vocal folds.

Iterative adaptive inverse filtering

- The iterative adaptive inverse filtering (IAIF method) is based on iterative refinements of both the vocal tract and glottal source components.
- Implemented by TKK Aparat `sourceforge.net/projects/aparat`

Mixed-phase decomposition

- The input signal $x[k]$ windowed by a proper GCI-synchronous window can be represented in z-domain $X[z]$ as a set of zeros:

$$\begin{aligned} X[z] &= \sum_{n=0}^{N-1} x[n] z^{-n} \\ &= x[0] z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \\ &= x[0] z^{-N+1} \prod_{k=1}^{M_o} (z - Z_{max,k}) \prod_{k=1}^{M_i} (z - Z_{min,k}) \end{aligned} \tag{20}$$

- $Z_0, Z_1, \ldots, Z_{N-1}$ are roots (zeros) of the corresponding z-transform.

Mixed-phase decomposition

- The zeros $Z_{max,k}$ are due to maximum-phase (i.e., anticausal) components of speech and they fall outside of the unit circle. They are related to glottal open phase.
- The zeros $Z_{min,k}$ are due to minimum-phase (i.e., causal) components of speech and they fall inside of the unit circle. They are related to the vocal tract impulse response.
- Mixed-phase decomposition can be thus achieved by analysis of the speech in the z-domain with unit circle as the discriminant boundary.
- This analysis is analogous to complex cepstrum decomposition, which is in addition much faster.

In time domain

In the time domain, most important parameters are related to the closing phase of the glottal cycle:

- Normalized amplitude quotient (NAQ) – a parameter closely related to the closed quotient
- excitation strength – corresponds mostly to the rate of closure of the vocal folds in each glottal cycle.
- Basic shape parameter R_d, "the most effective single measure for describing voice qualities" (Fant'95)

$$R_d = (U_0/E_e)(F_0/110) \tag{21}$$

where (U_0/E_e) is effective glottal pulse declination time (in milliseconds, 0.5 – 1 ms for both male and female vowels). R_d is proportional to the fundamental frequency averaged 110 Hz.

In frequency domain

In frequency domain, most useful parameters are:

- glottal formants (frequency and bandwidth)
- H1–H2 parameter: a ration between magnitudes of glottal source spectrum at fundamental and second harmonic frequencies.
- harmonic-to-noise – the ratio between the sum of the amplitudes of harmonics and the amplitude at the fundamental frequency.

Re-synthesis from LP residual

- Time-frequency transforms allow to modify transformed signal and re-synthesis using the inverse transform.
- Analogically, one can decompose the signal into source and filter components, modify them, and re-synthesise the signal by applying vocal tract filter on the source.
- Using LP residual as source parameters would be much easier than glottal flow signal.

Scaling Hilbert envelope

- It is worth to use Hilbert envelope $h_e[k]$ of the LP residual $r[k]$. Hilbert envelope is a magnitude of analytic signal $r_a[k]$, and both are related through a phase of the analytic signal:

$$\cos(\theta[k]) = \frac{\Re(r_a[k])}{\mid r_a[k] \mid} = \frac{r[k]}{h_e[k]}. \tag{22}$$

- Scaling of the Hilbert envelope directly impacts of the slope of the vocal fold's closing phase.
- The sharper slope (the shorter closing time), the stronger excitation, and the more significant change of voice quality.
- The modified LP residual can by reconstructed by

$$r[k] = h'_e \cos(\theta[k]) \tag{23}$$