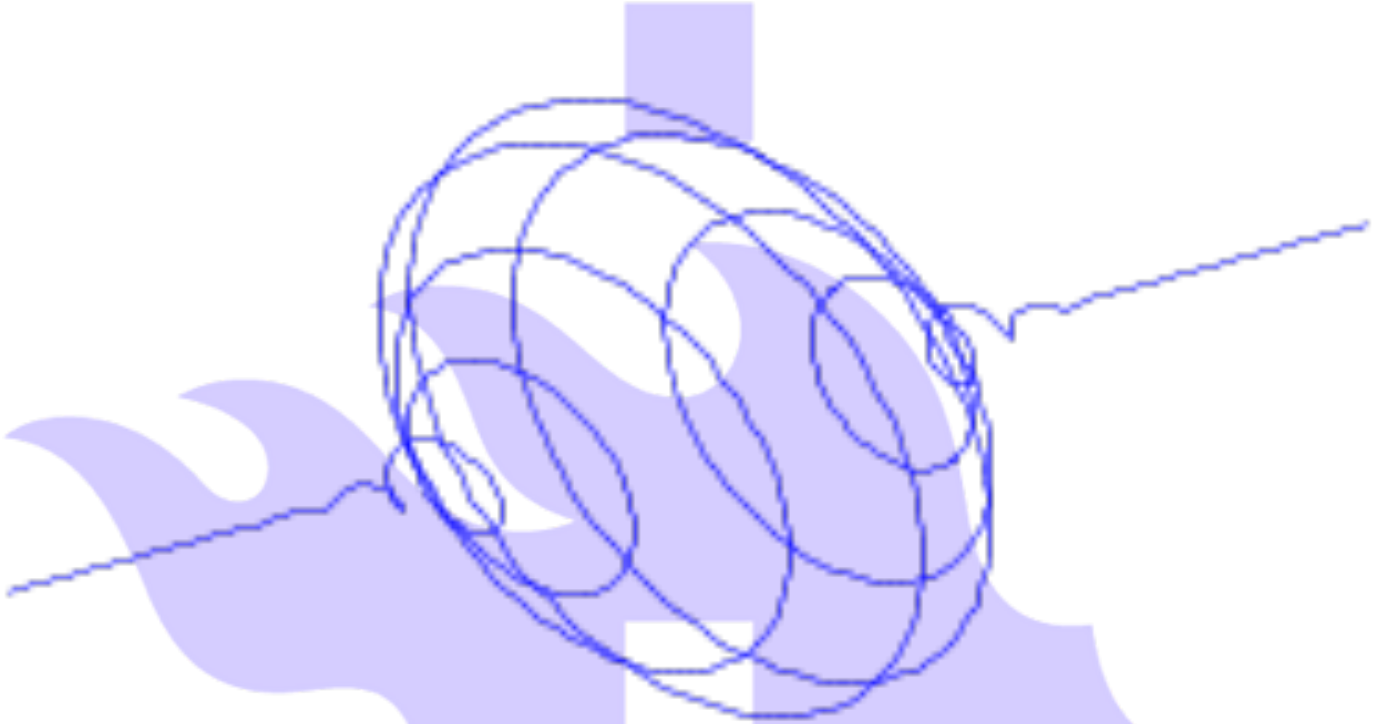


Continuous wavelet transform for speech research

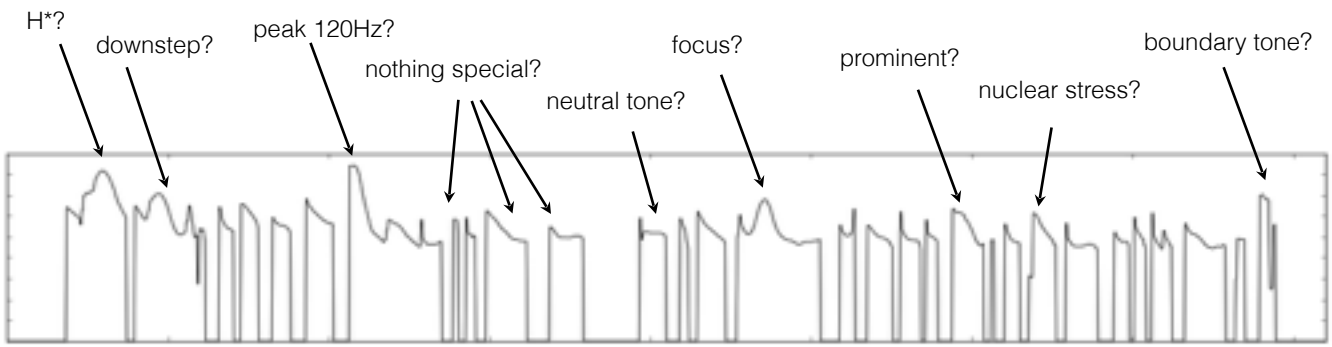
Martti Vainio, Juraj Šimko, Antti Suni

Linguistic Convergence Laboratory Meeting

HSE Moscow, November 27, 2017

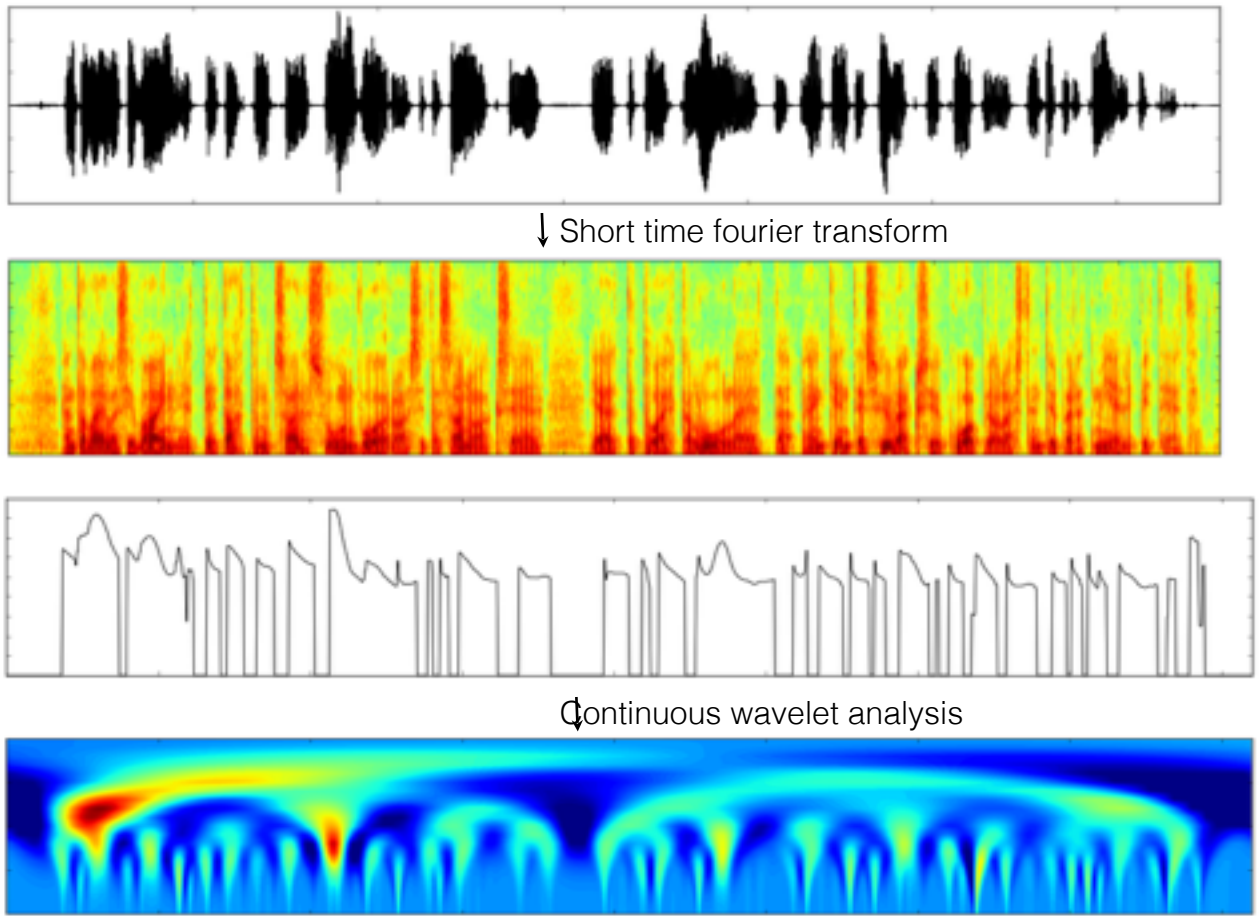
Introduction

- Prosodic signals, like f_0, are complex, containing information on syllable, word, phrase and utterance levels, with diverse functions.
- The information is encoded ***in parallel*** in one dimensional signal;
  - Automatic non-trivial prosodic analysis is difficult
  - Expert analysis requires a lot of subjectivity and effort
  - No generally agreed framework for analysis



Introduction

- Thus, what is sought after, is a representation for prosody, where the contribution of different phonological layers is distinguishable: **Continuous wavelet analysis**



Definition

time

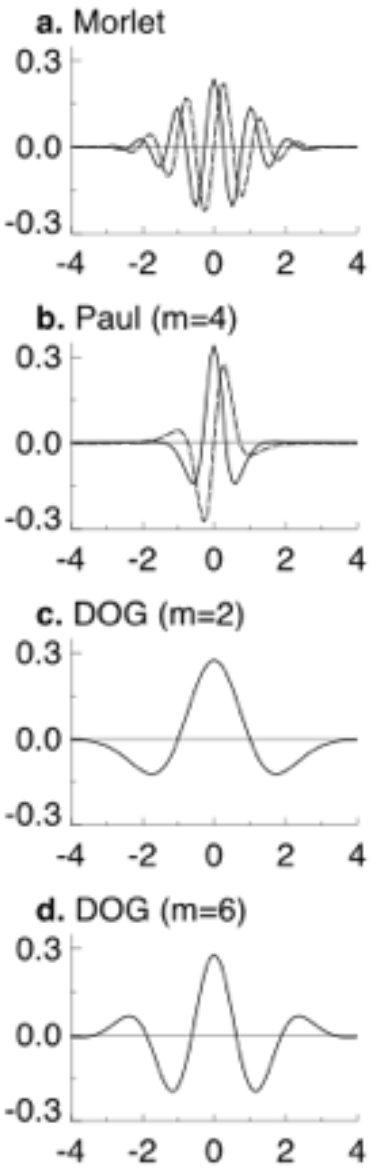
signal

wavelet

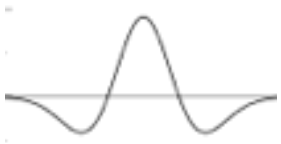
$$X(t,s) = \frac{1}{|s|^{1/2}} \int_{-\infty}^{\infty} x(\tau)\overline{\psi}\left(\frac{\tau - t}{s}\right)\, d\tau$$

scale

Mother wavelet



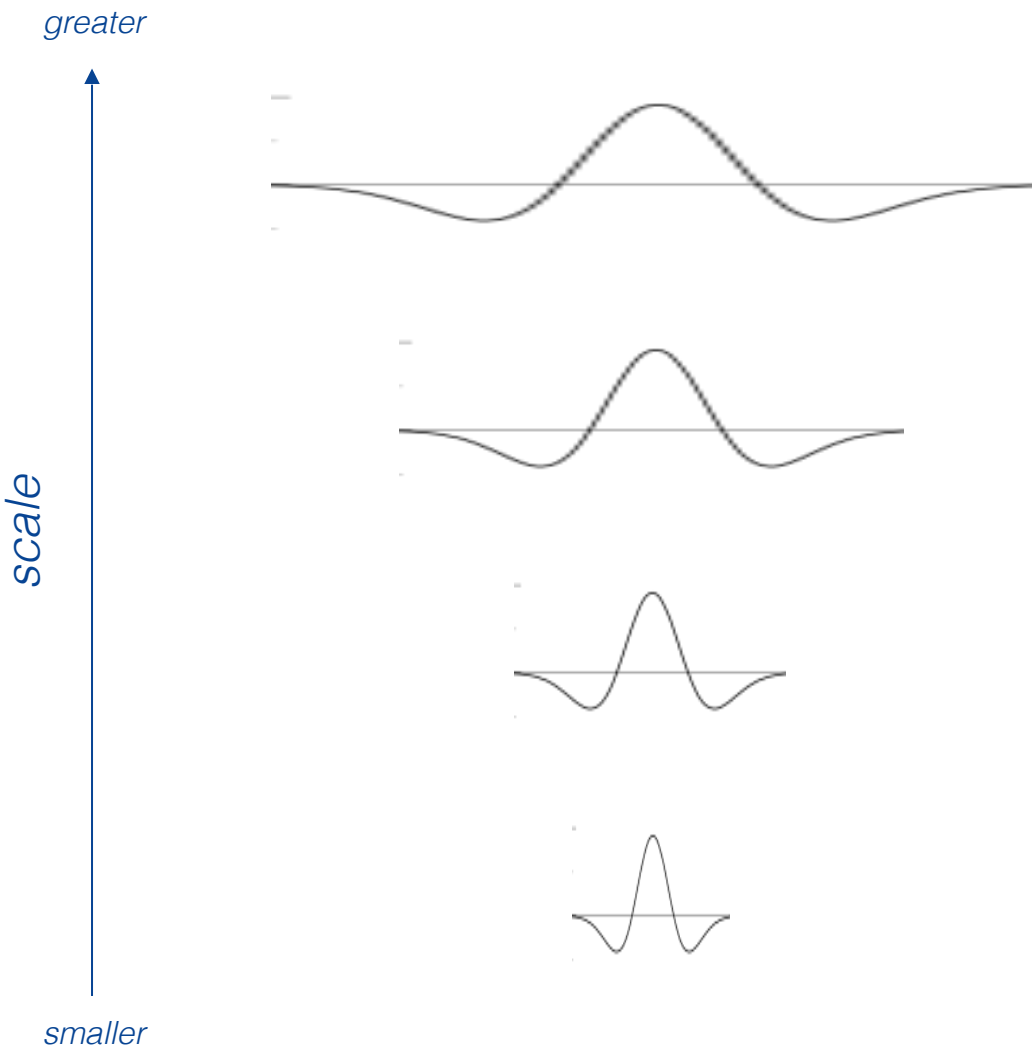
Wavelet: scaling



$$\psi\left(\frac{\tau - t}{s}\right)$$

scale

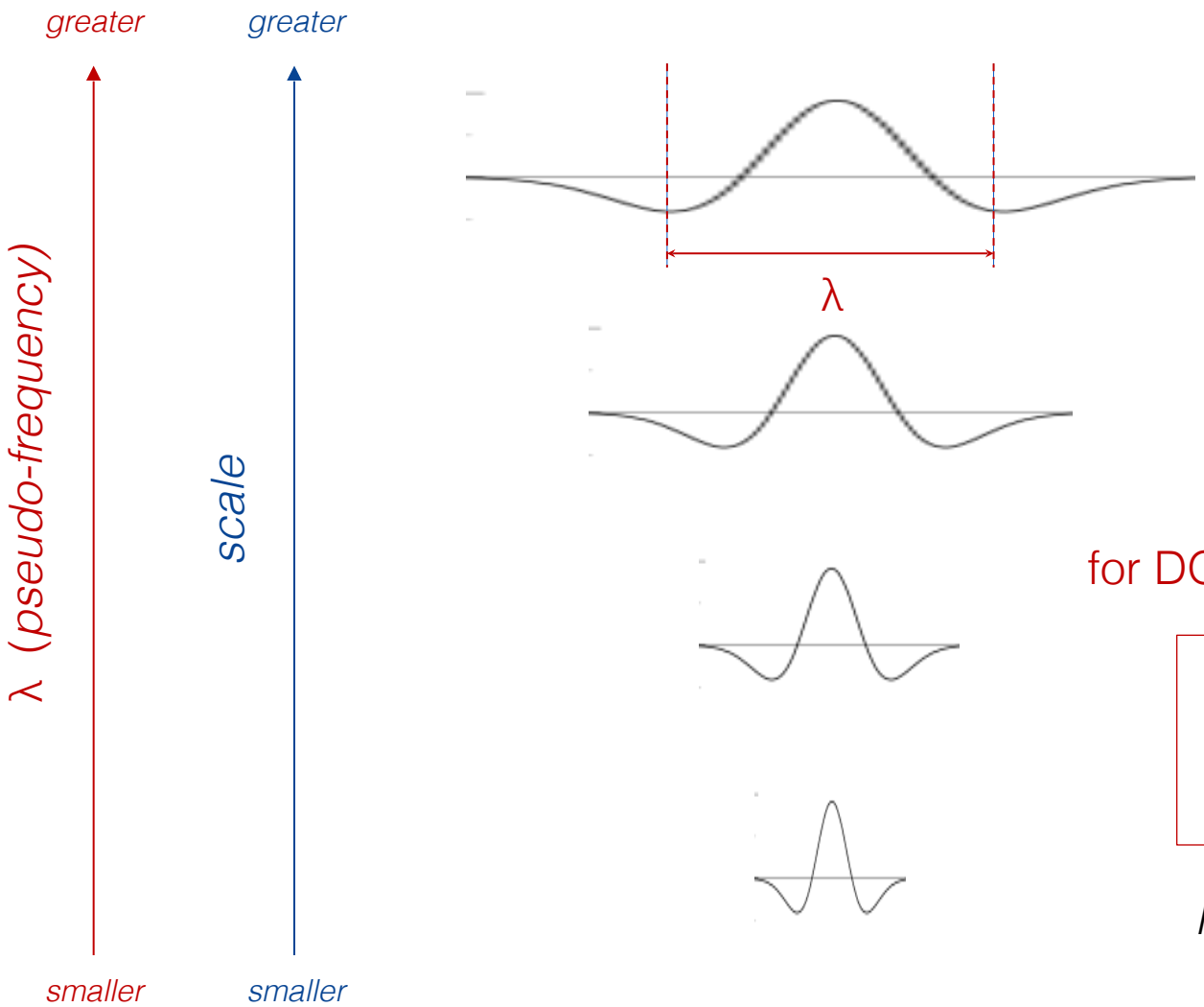
Wavelet: scaling



$$\psi\left(\frac{\tau - t}{s}\right)$$

scale

Wavelet: scaling



for DOG (m=2):

$$\lambda = \frac{2\pi s}{\sqrt{5/2}}$$

linear relationship

Convolution

$$X(t,s) = \frac{1}{|s|^{1/2}} \underbrace{\int_{-\infty}^{\infty} x(\tau)\overline{\psi}\left(\frac{\tau - t}{s}\right)\, d\tau}_{\text{convolution}}$$

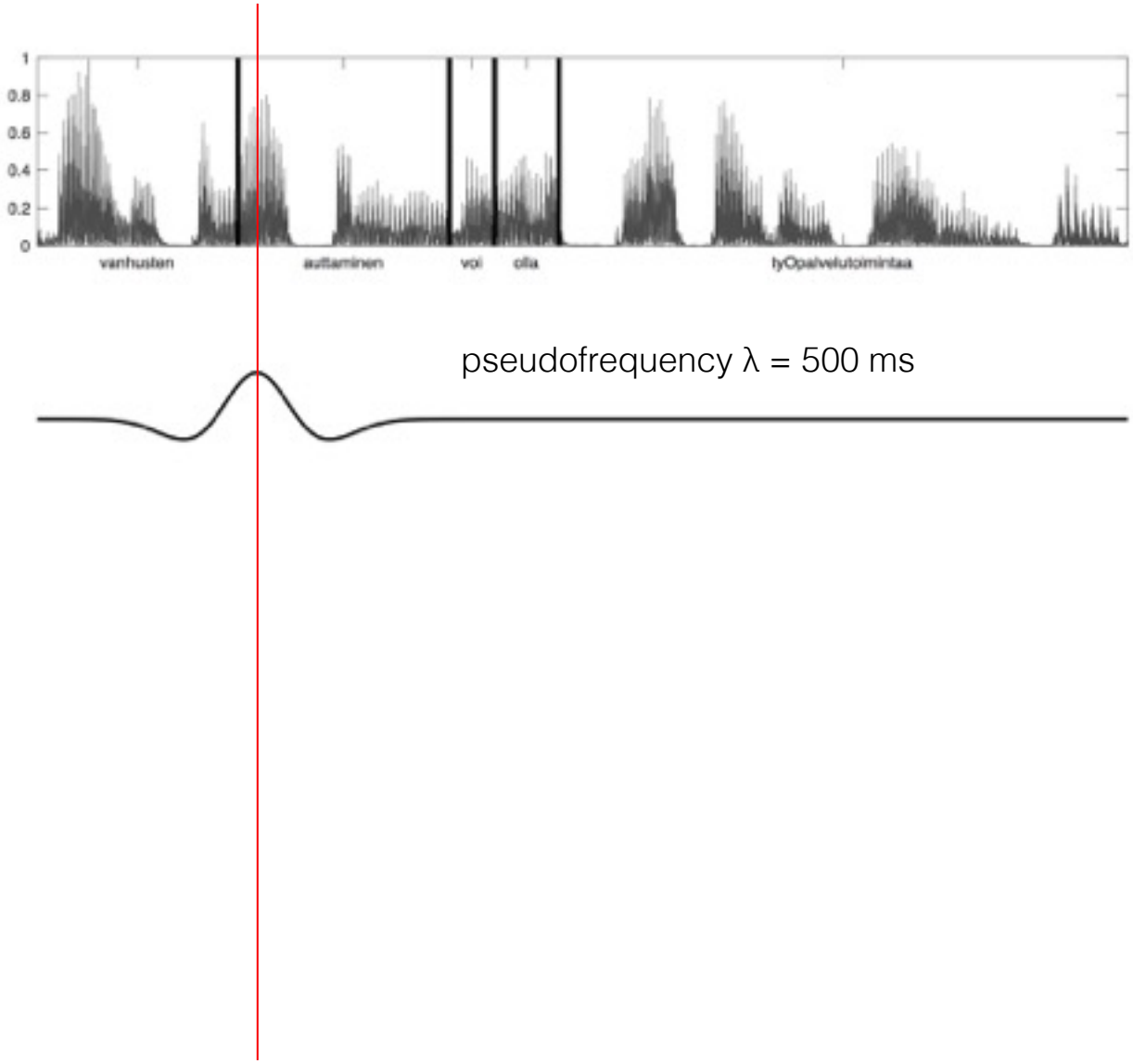
time

signal

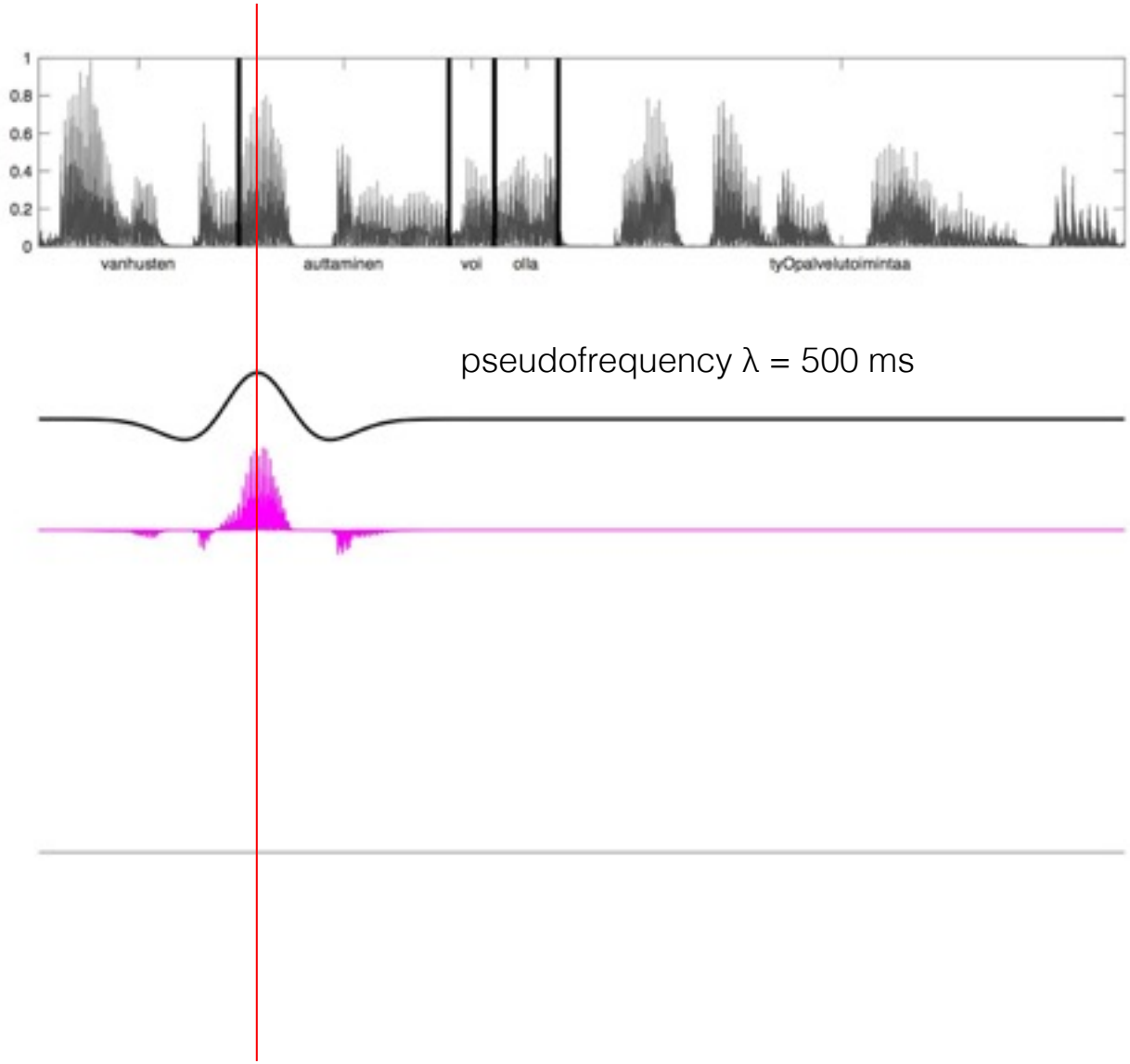
wavelet

scale

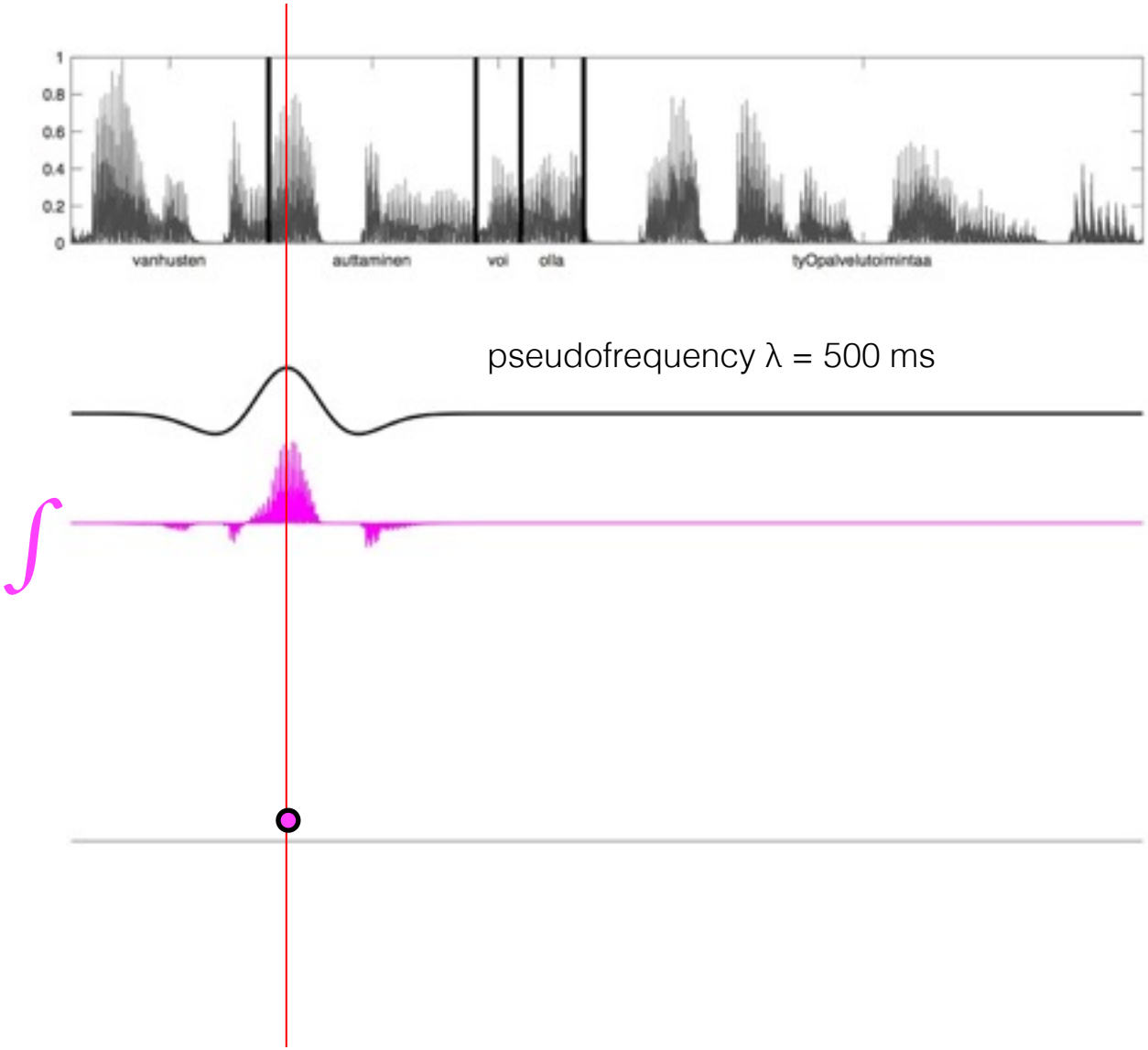
Convolution



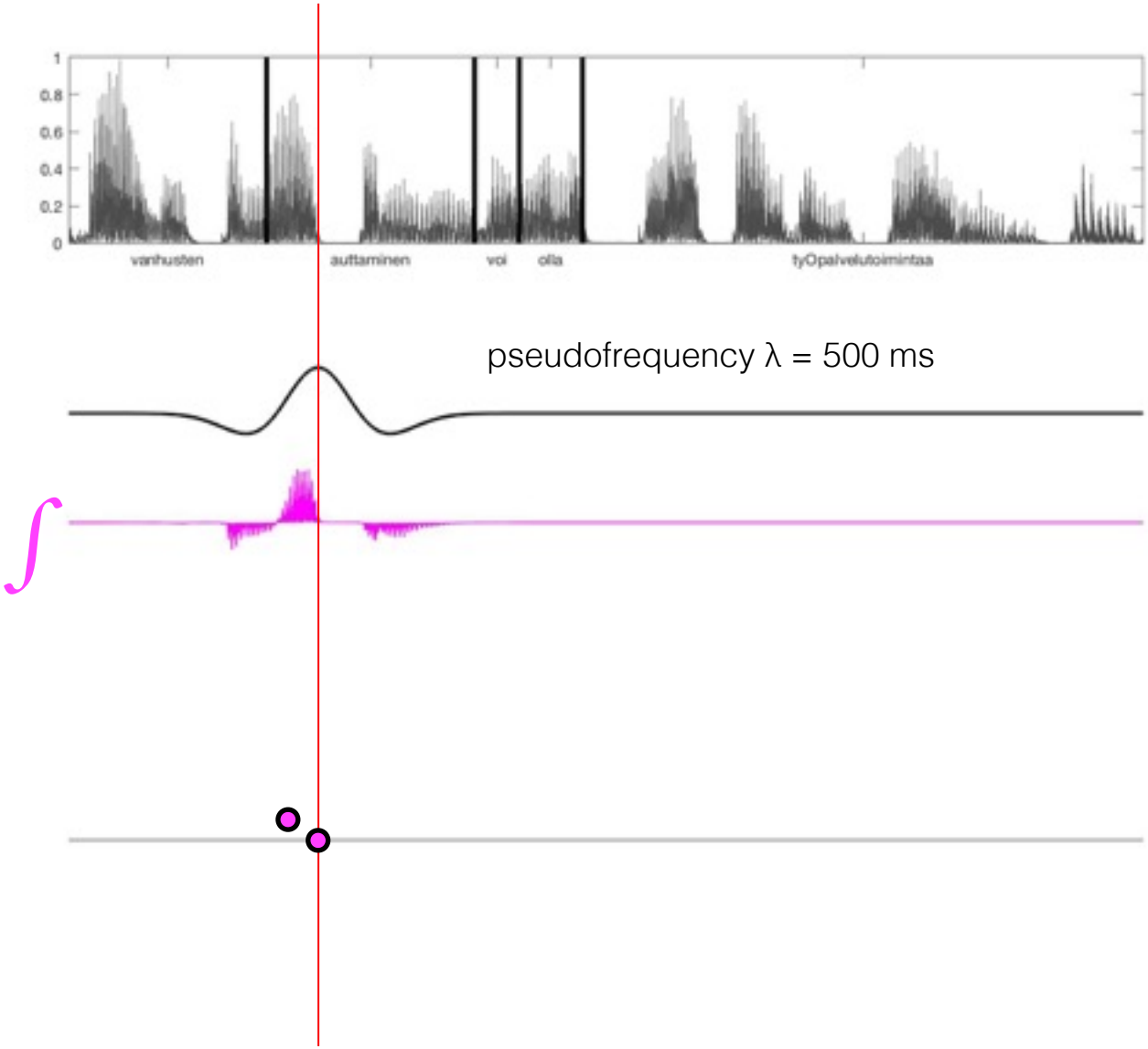
Convolution



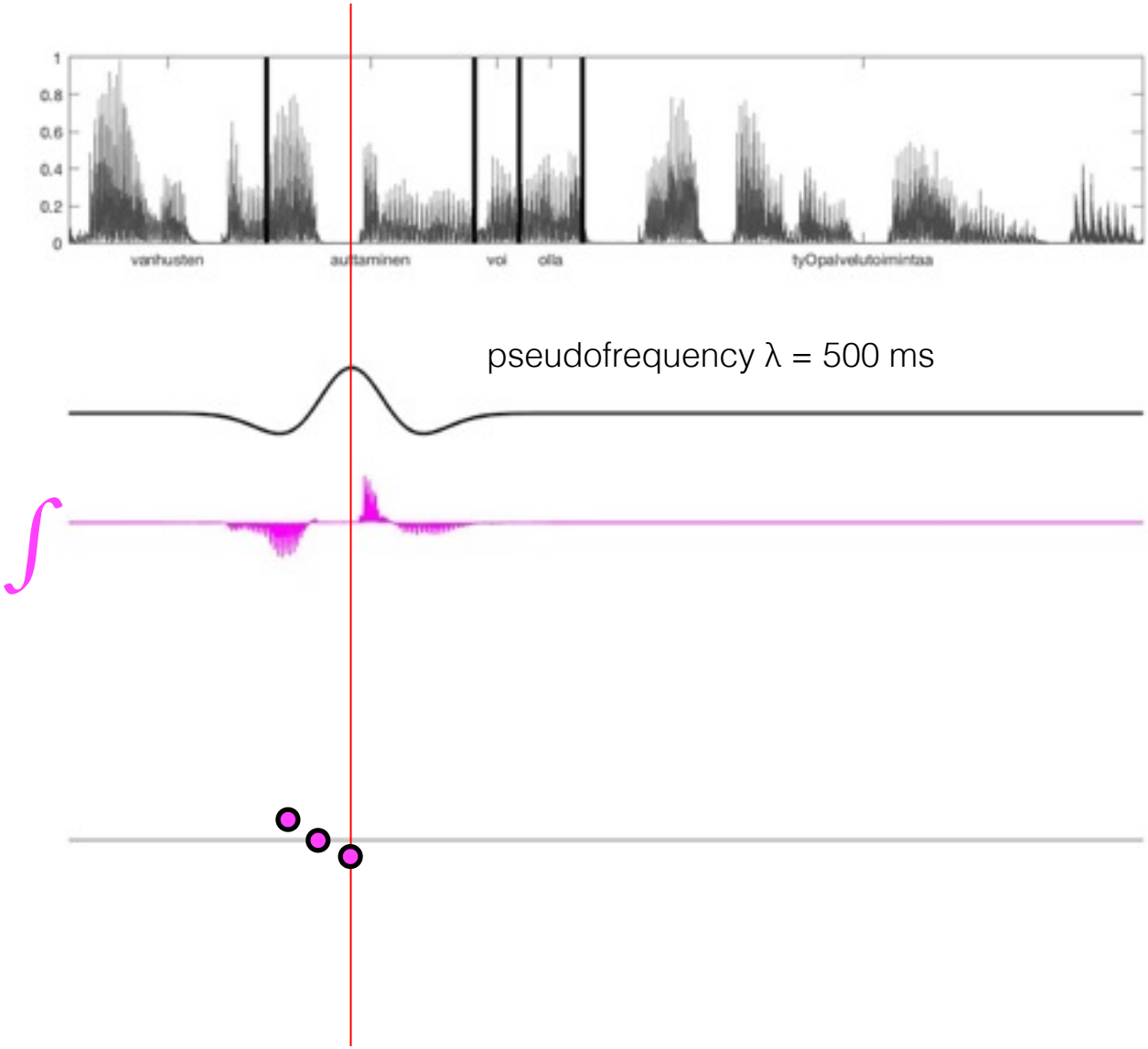
Convolution



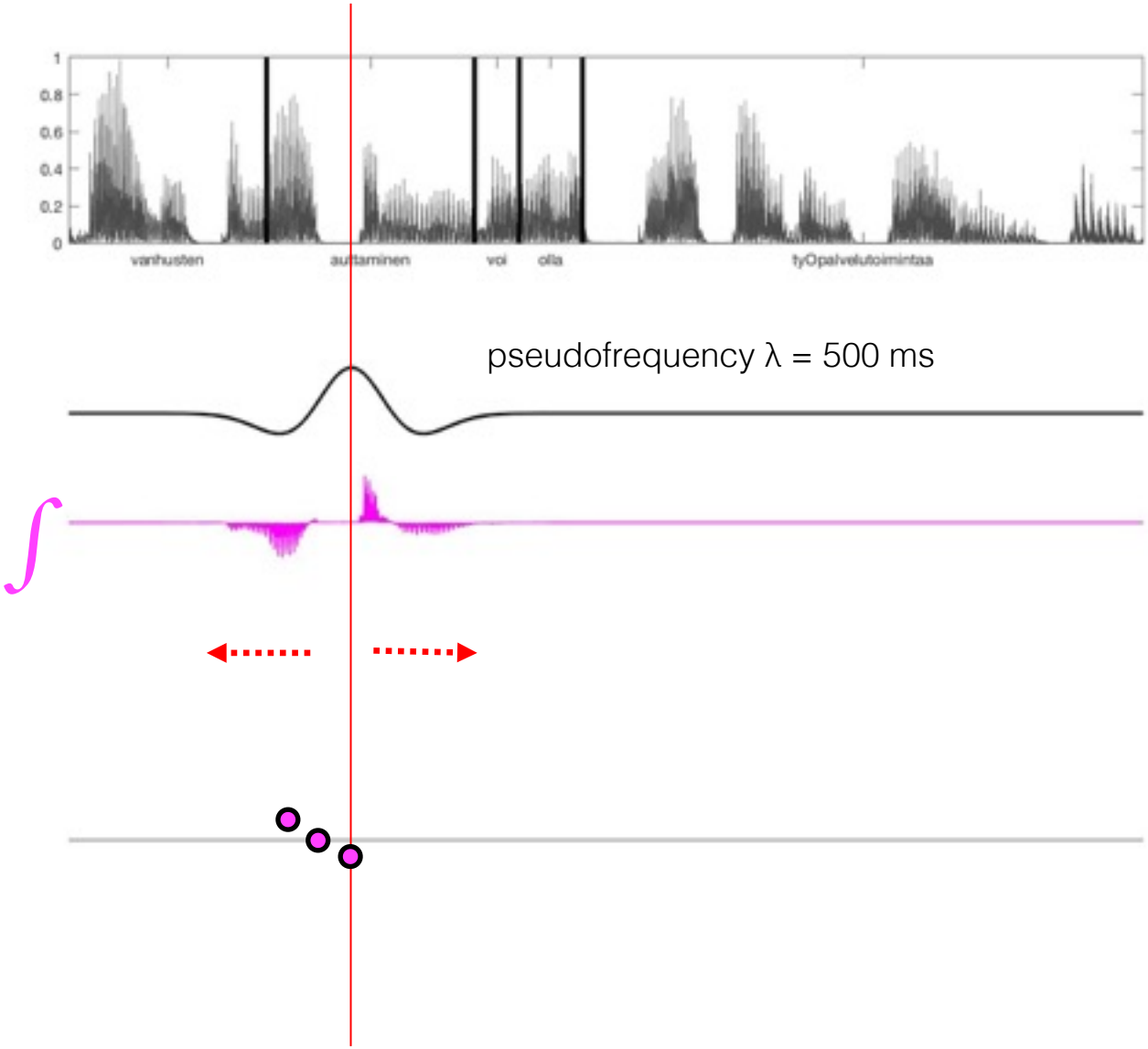
Convolution



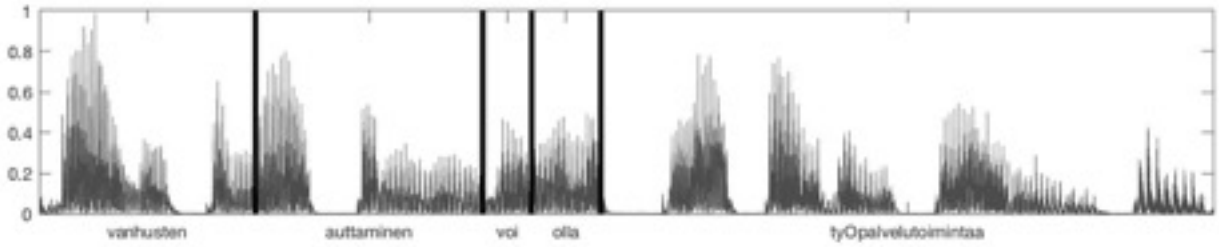
Convolution



Convolution

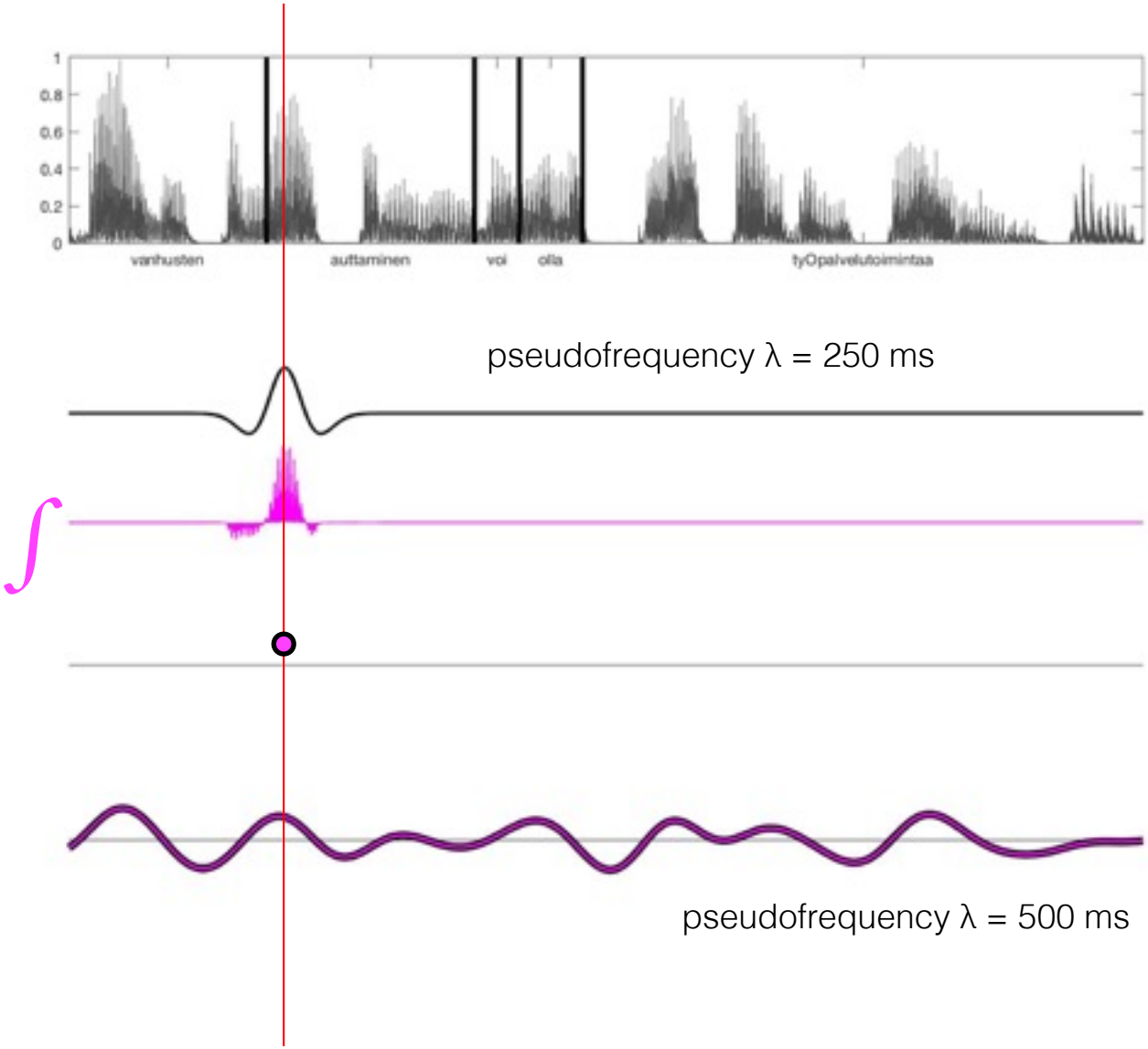


Convolution

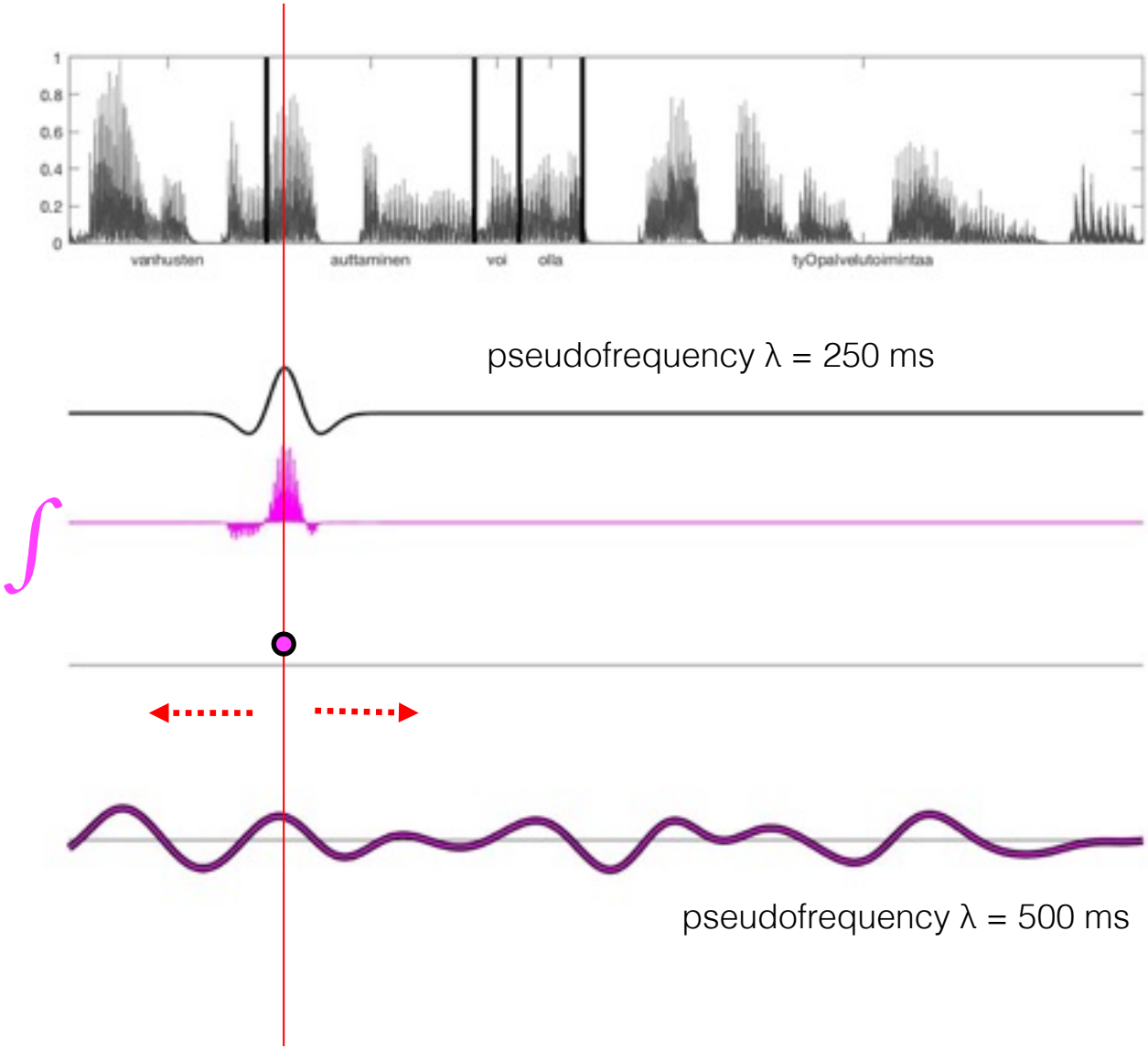


pseudofrequency λ = 500 ms

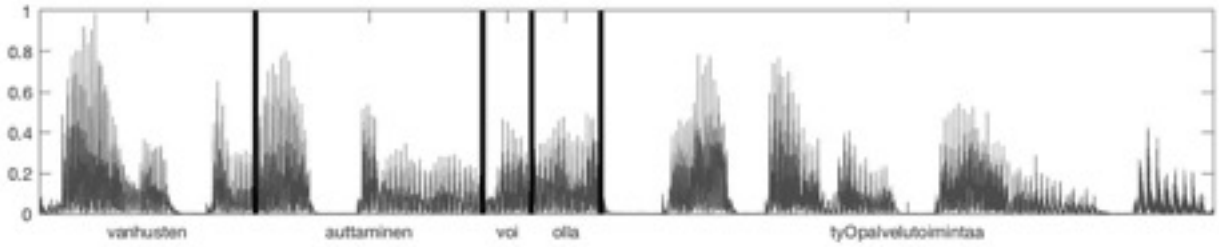
Convolution

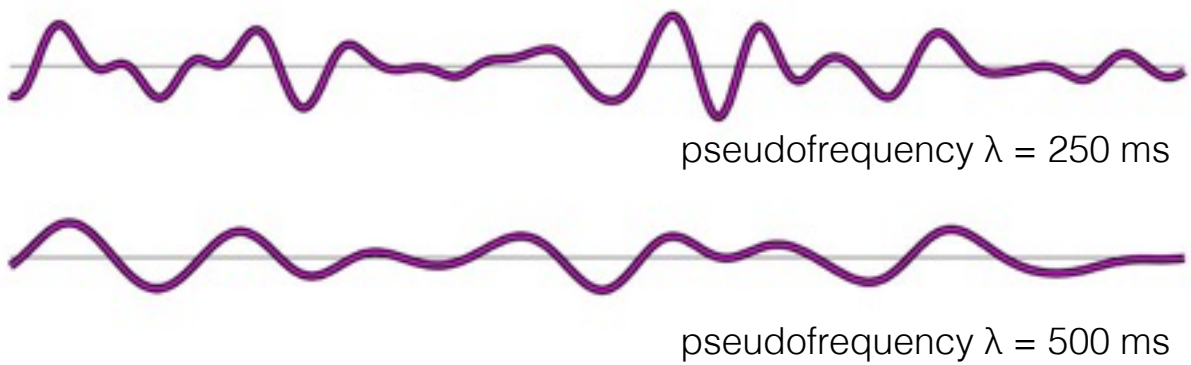


Convolution

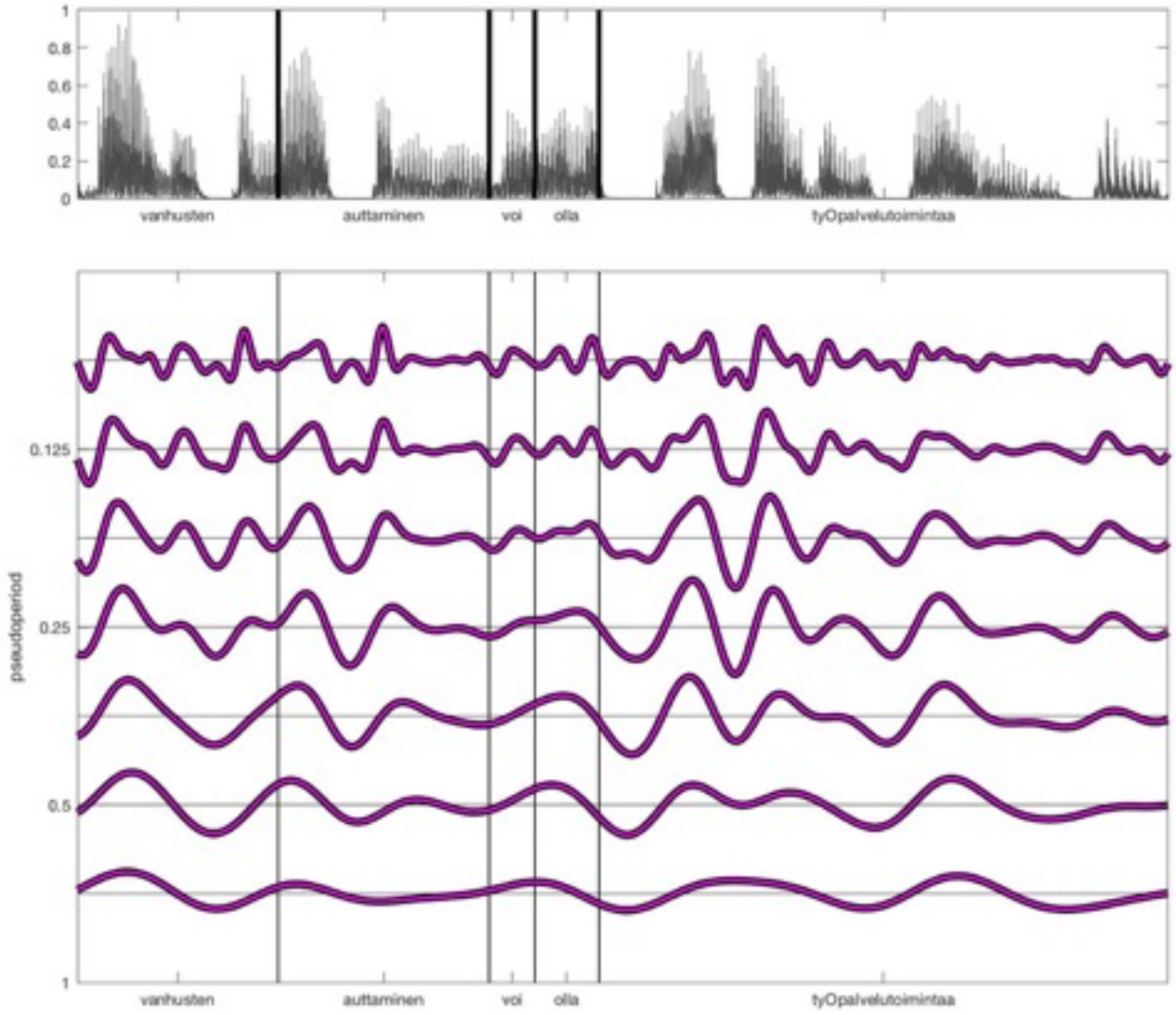


Convolution

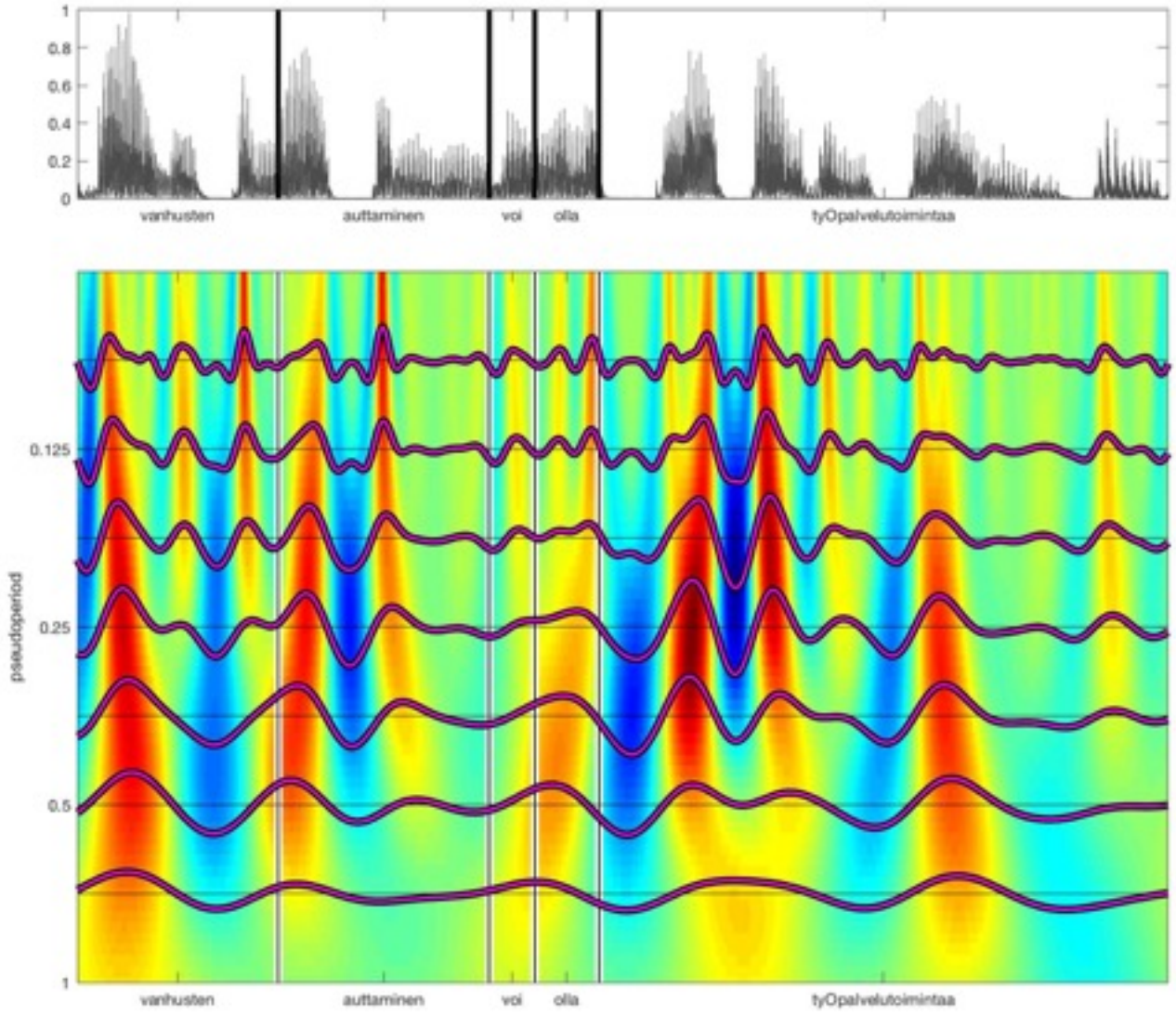




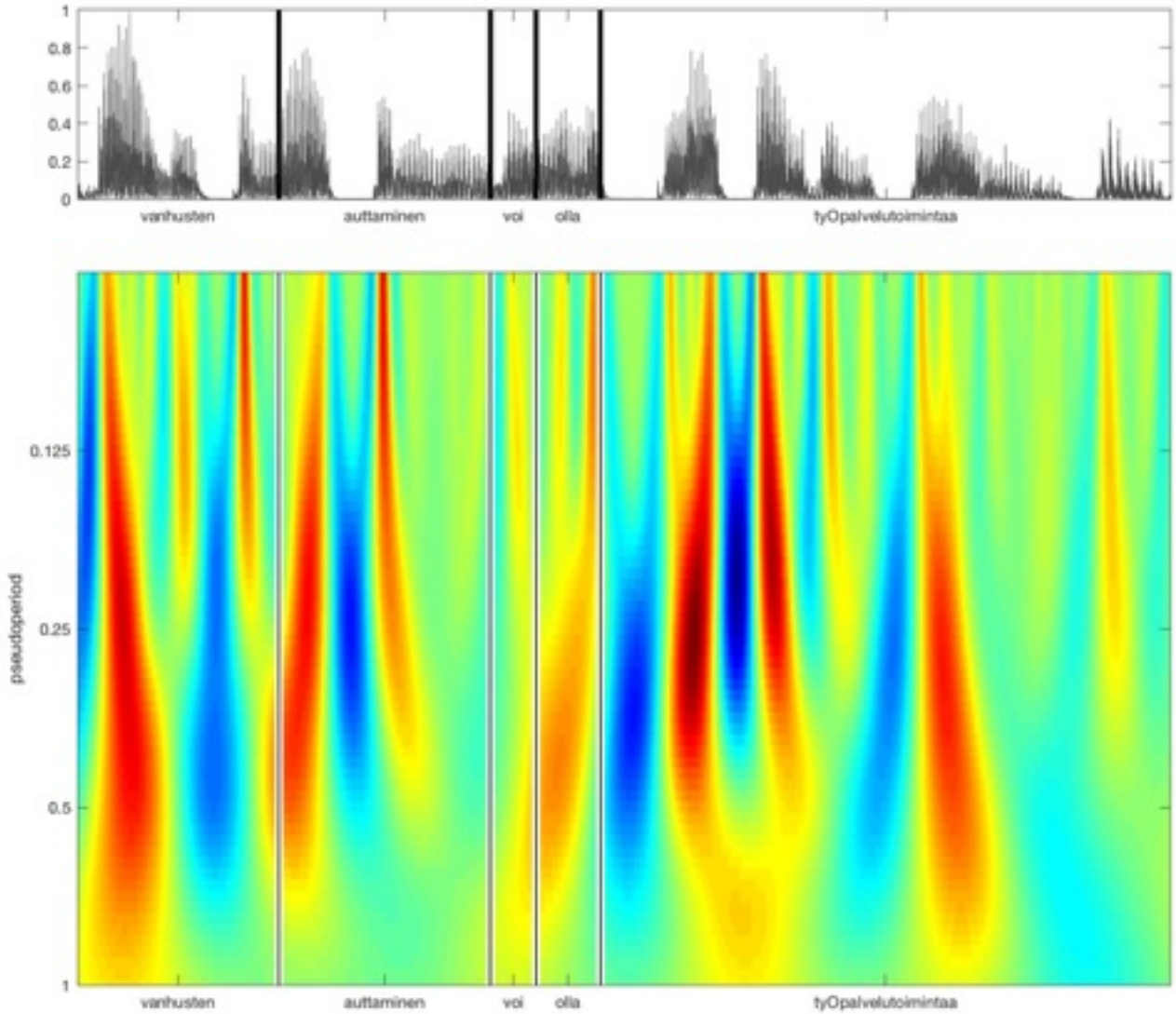
Continuous wavelet transform



Scalogram


1
0.8
0.6
0.4
0.2
0
vanhusten
auttaminen
voi
olla
tyOpalvelutoimintaa
0.125
0.25
0.5
1
pseudoperiod
vanhusten
auttaminen
voi
olla
tyOpalvelutoimintaa

Scalogram


1
0.8
0.6
0.4
0.2
0
vanhusten
auttaminen
voi
olla
tyOpalvelutoimintaa
pseudoperiod
0.125
0.25
0.5
1
vanhusten
auttaminen
voi
olla
tyOpalvelutoimintaa

Computation of CWT

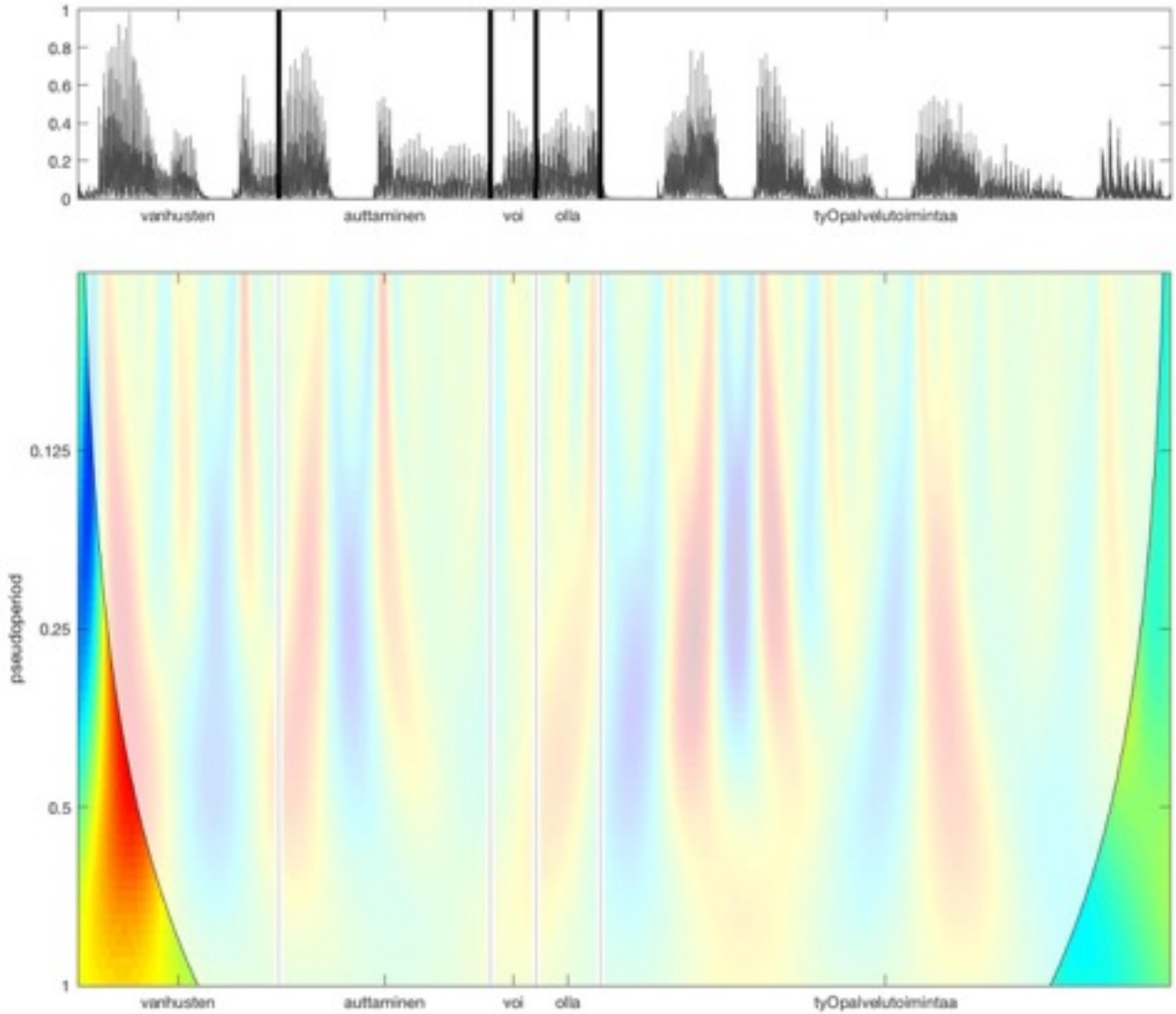
- is quite fast, uses Convolution Theorem

$$\boxed{f * g = iFFT\Big(c.FFT(f).FFT(g)\Big)}$$

- pads the signal (left and right) with zeros

- cone of influence

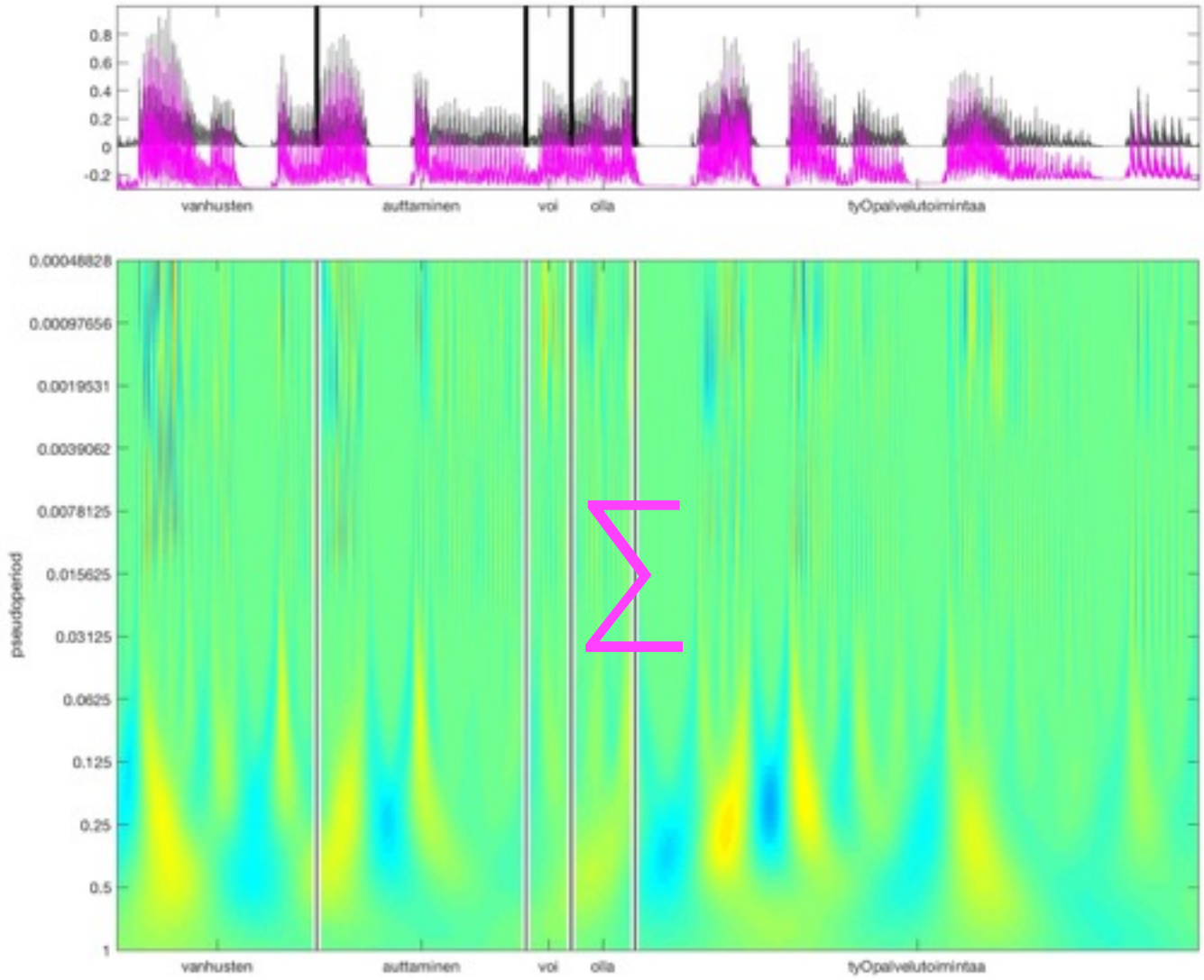
Cone of influence



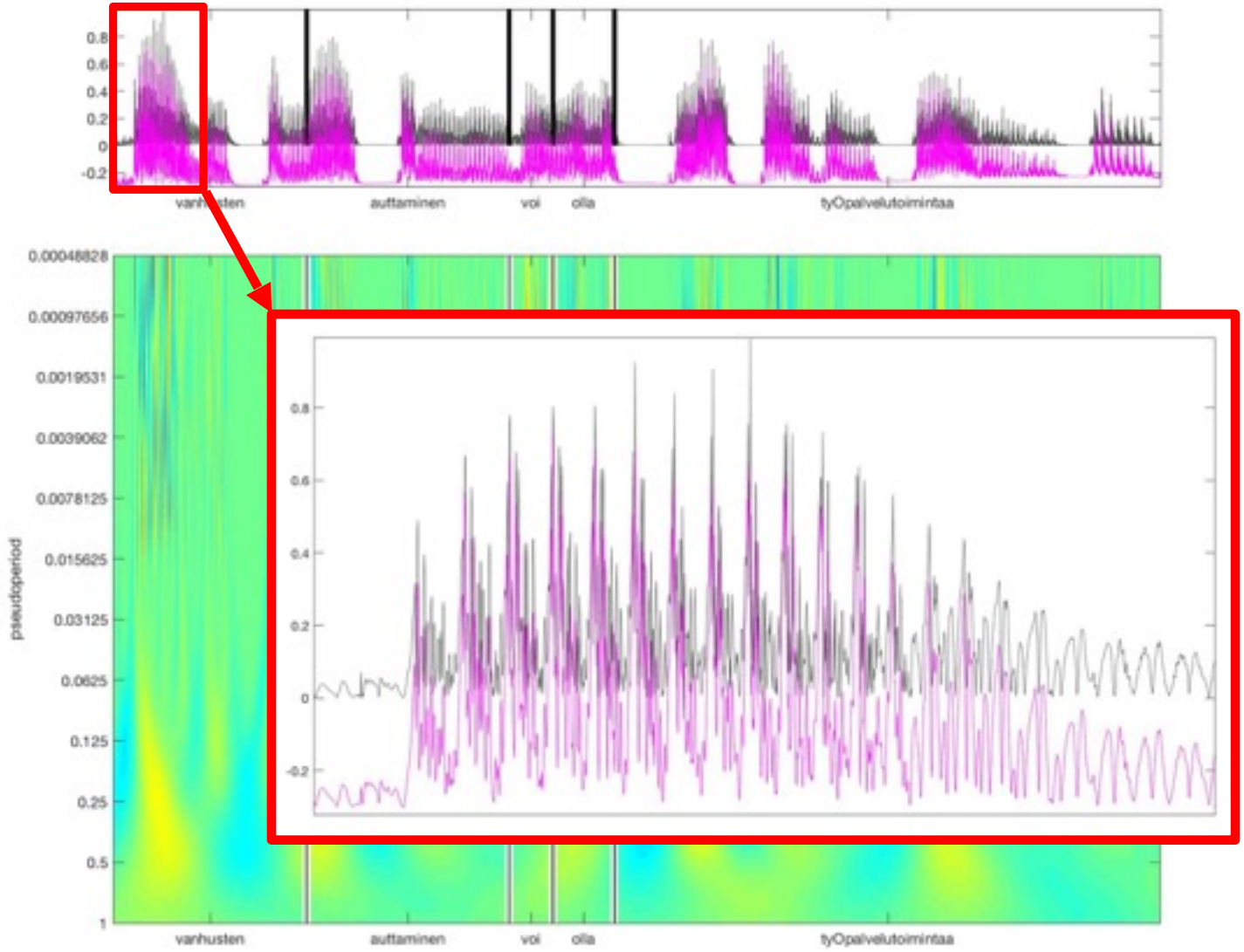
Computation of CWT

- packages in python, Matlab …
- sources:
  - Daubechies, I. (1992). *Then Lectures on Wavelets*, Society of Industrial and Applied Mathematics.
  - Torrence & Compo (1998). A practical guide to wavelet analysis, *Bulletin of the American Meteorological Society*, 79(1), pp. 61-78
  - Grinsted, Moore and Jevrejeva. (2004). Application of the cross-wavelet transform and wavelet coherence to geographical time series. *Nonlinear processes in Geophysics*, 11(5/6), pp. 561-566.
  - *and many many more*

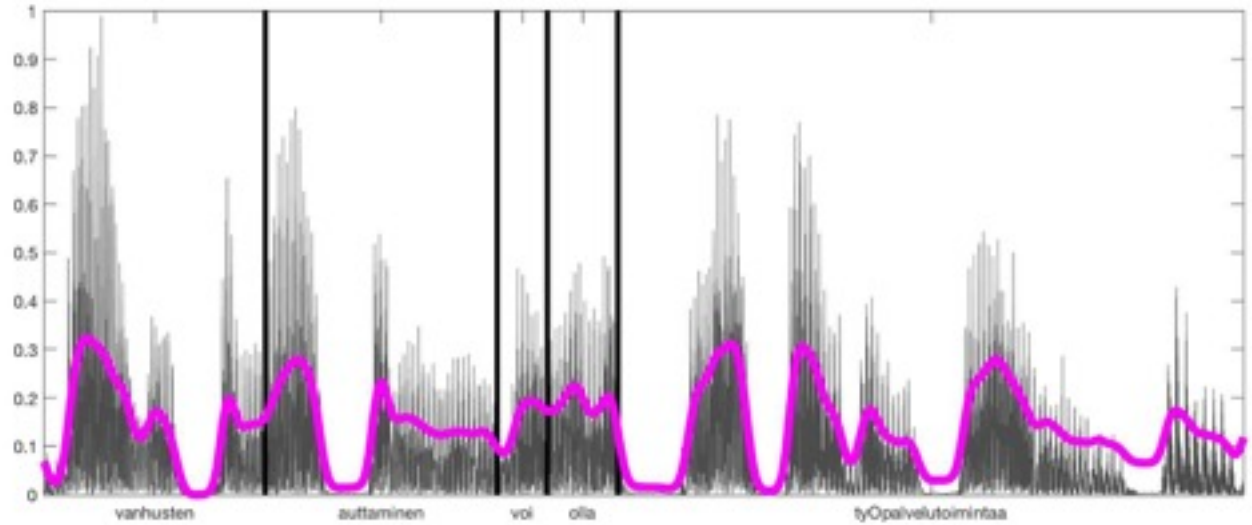
Decomposition: CWT is (almost) invertible



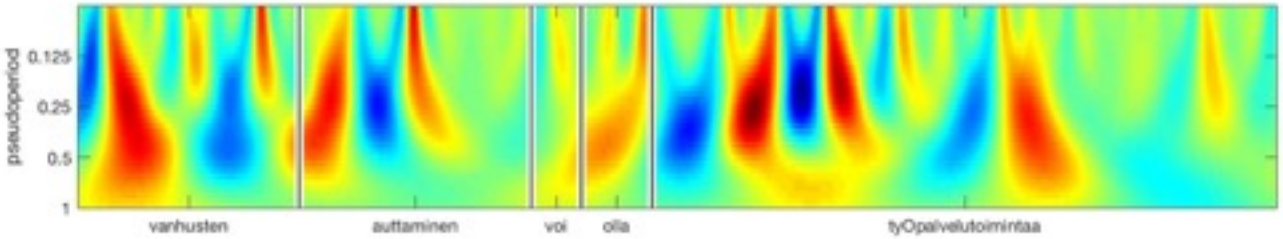
Decomposition: CWT is (almost) invertible



CWT as a filter

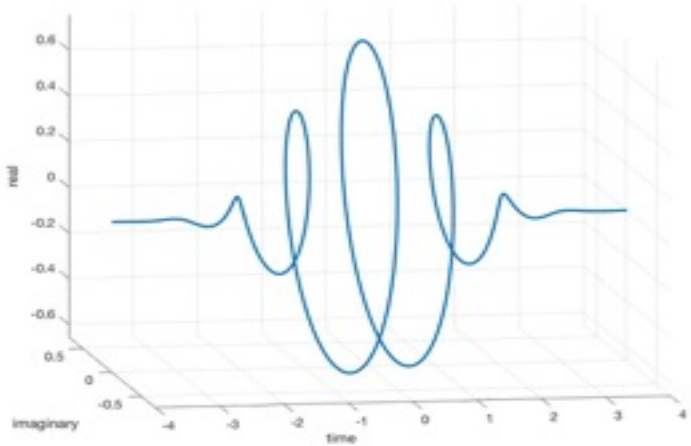


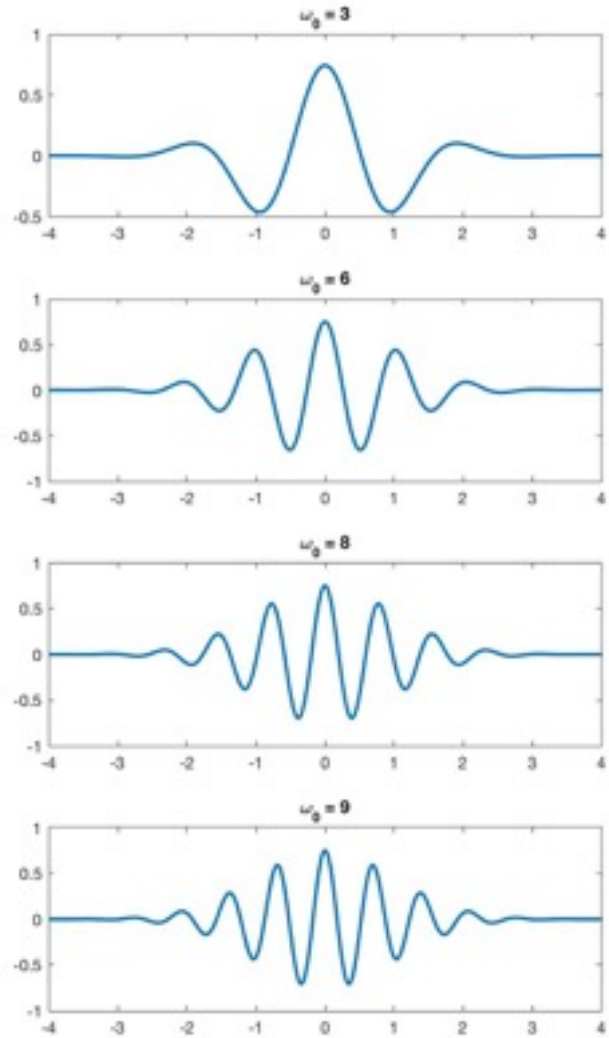
low-pass filter, approximation of signal energy



Morlet mother wavelet

- complex wavelet
- complex exponential ($e^{i\omega t}$)
  within Gaussian envelope

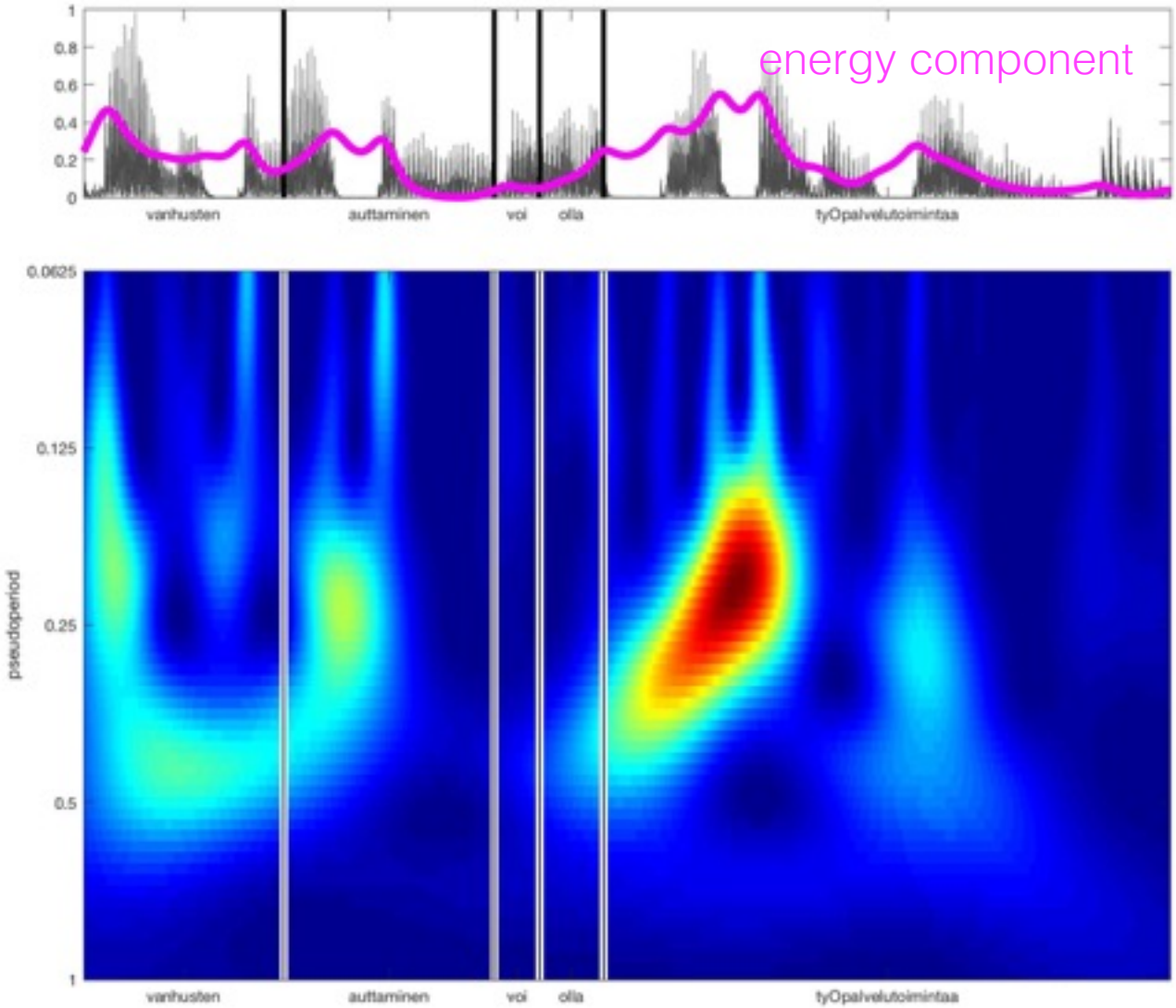




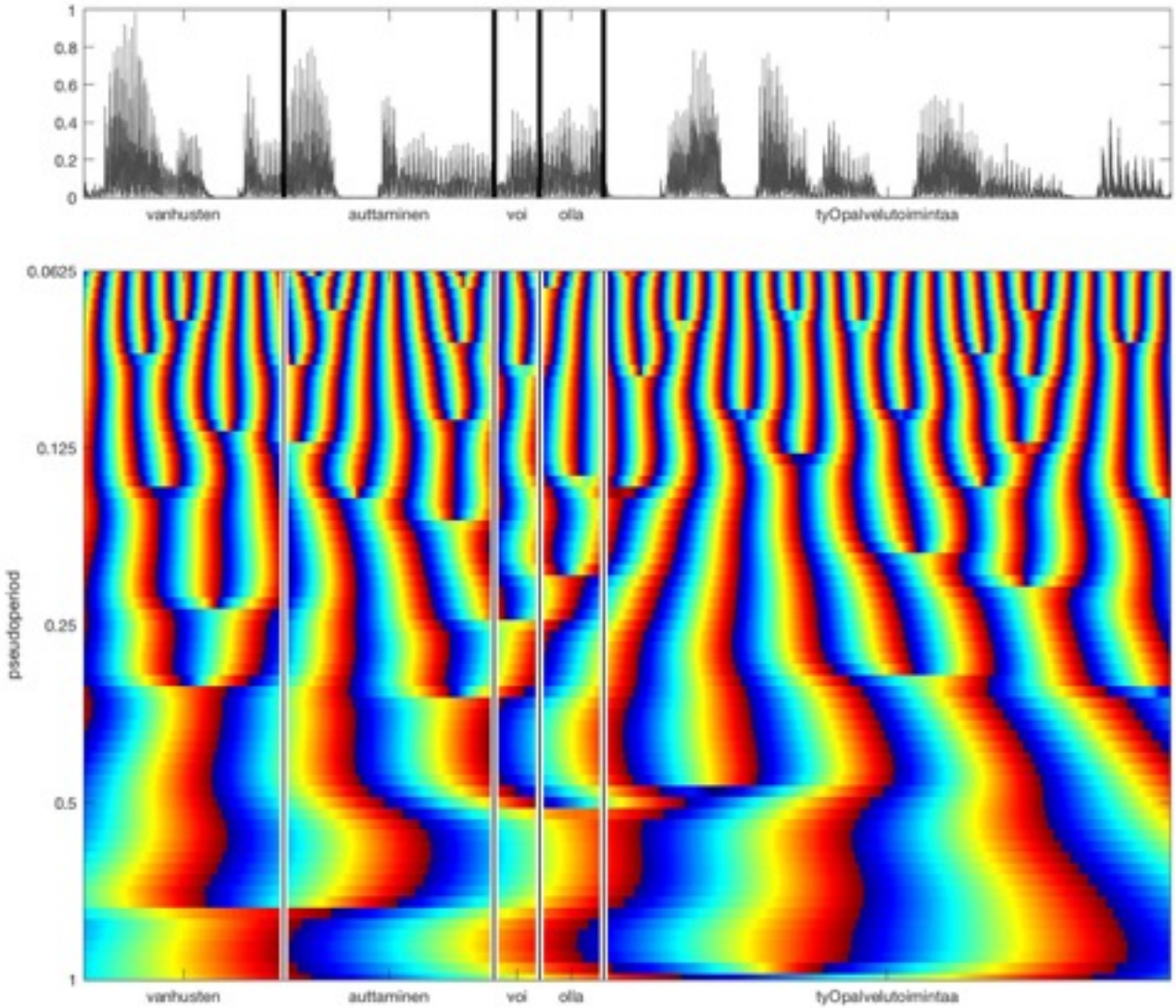
good temporal resolution

good frequency resolution

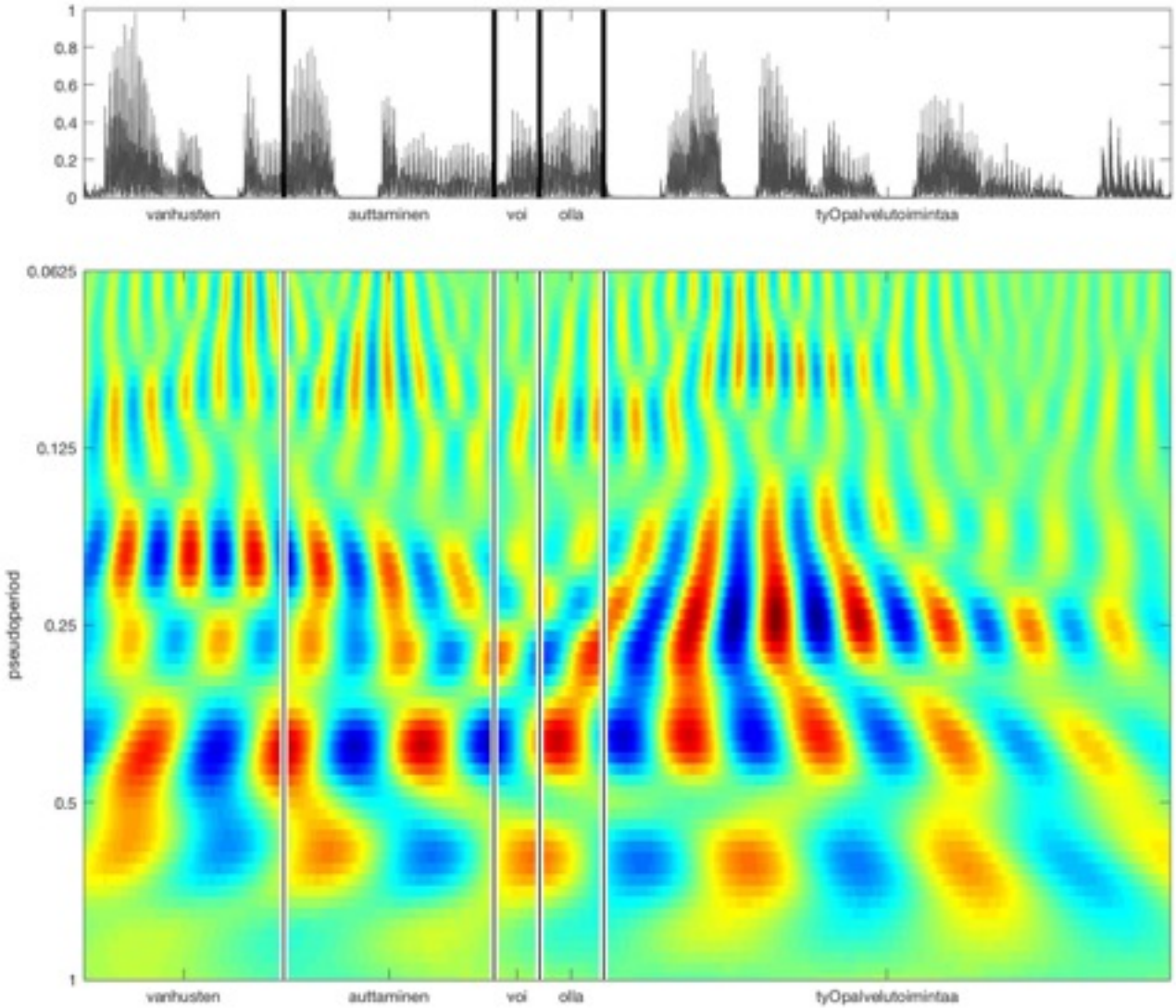
Energy – Morlet (amplitude): $\omega_0 = 3$



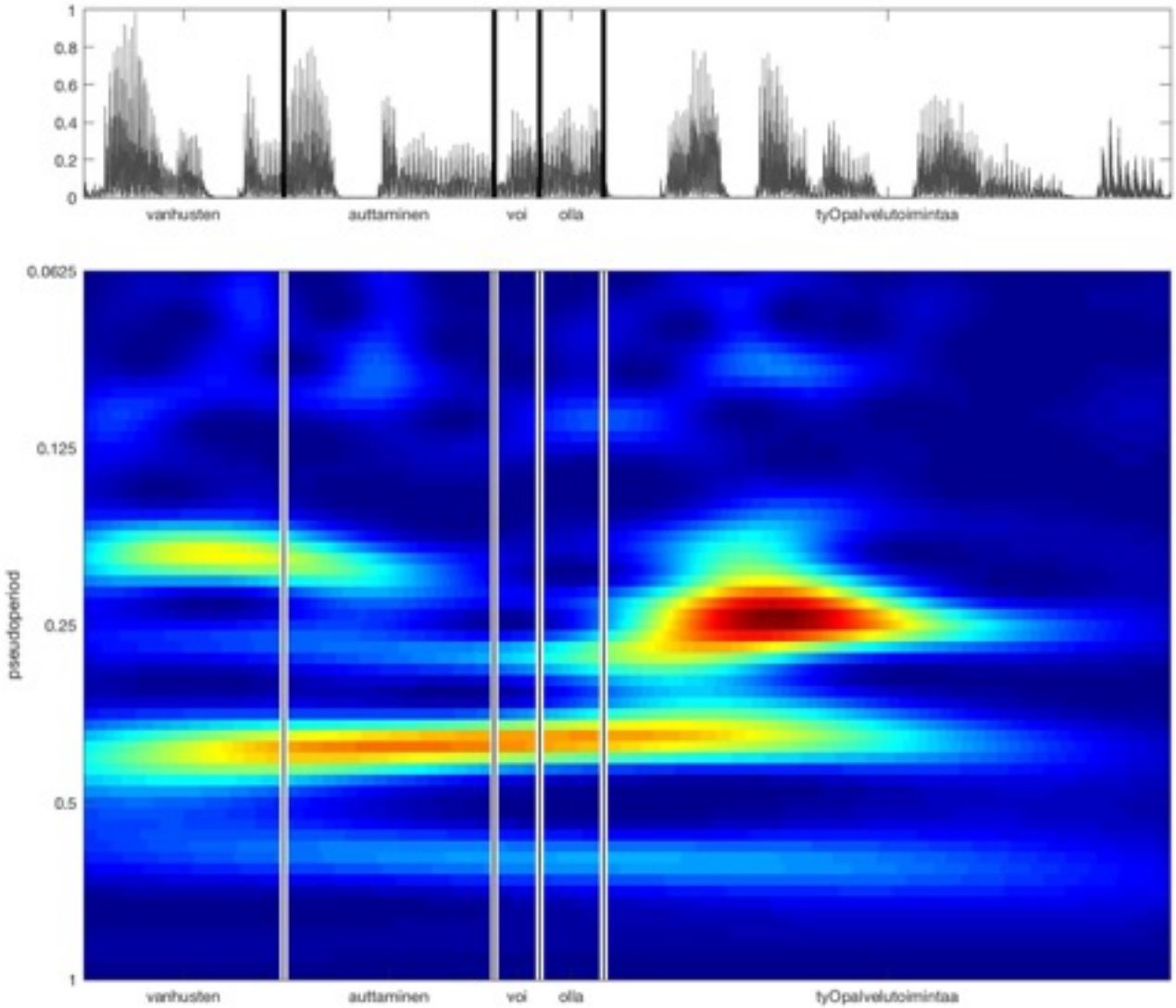
Phase – Morlet (amplitude): $\omega_0 = 10$



Morlet (real part): $\omega_0 = 10$

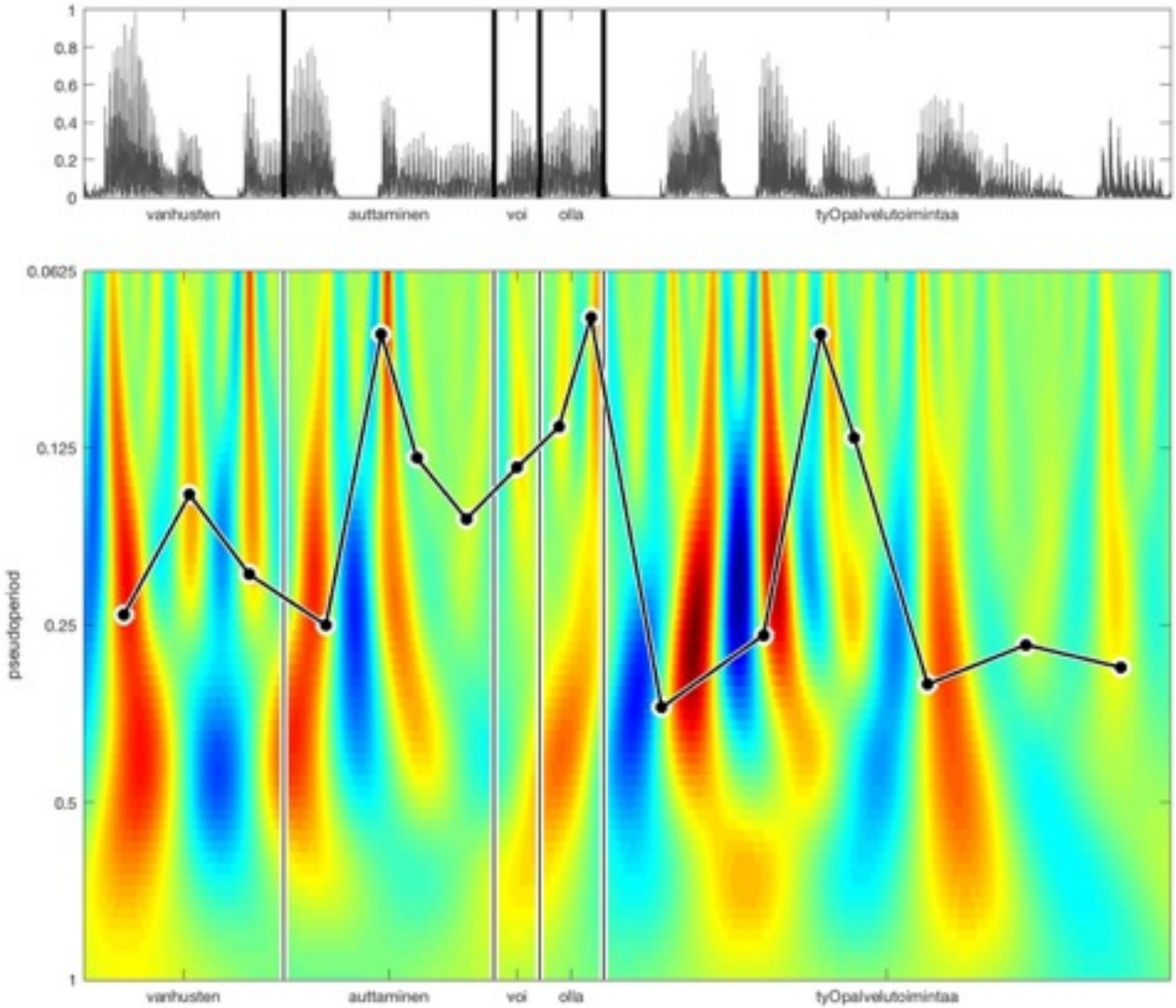


Energy – Morlet (amplitude): $\omega_0 = 10$

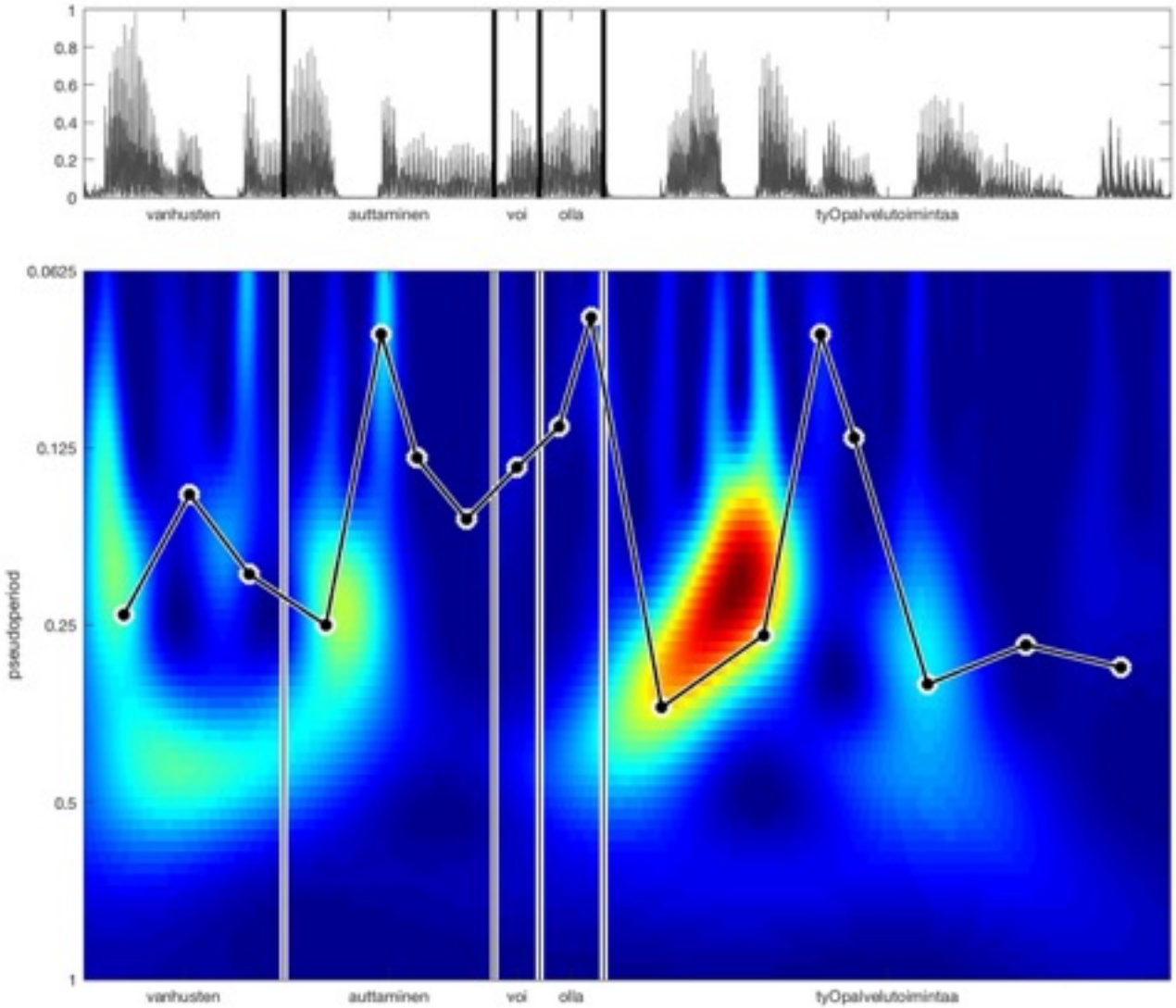


Continuous Wavelet Analysis
and
Speech

Syllable structure (Morlet, $\omega_0 = 3$)



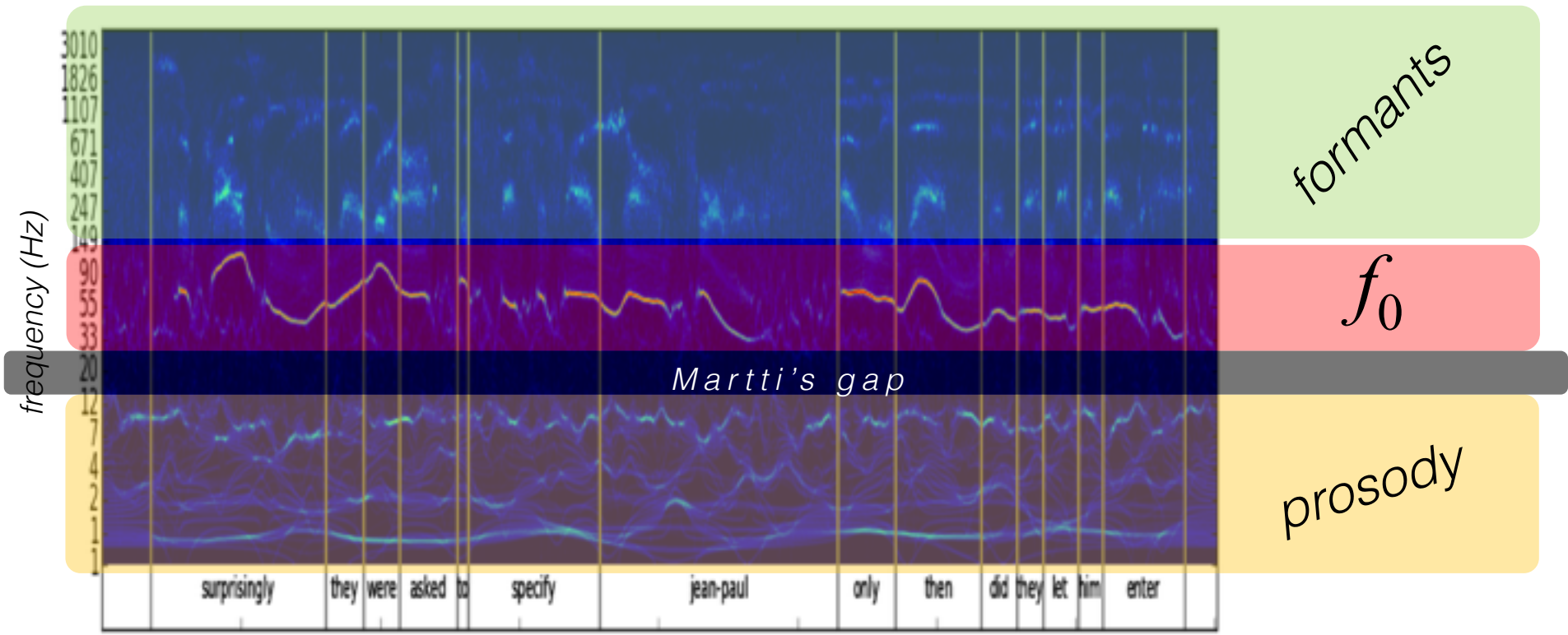
Prominence (Morlet, $\omega_0 = 3$)



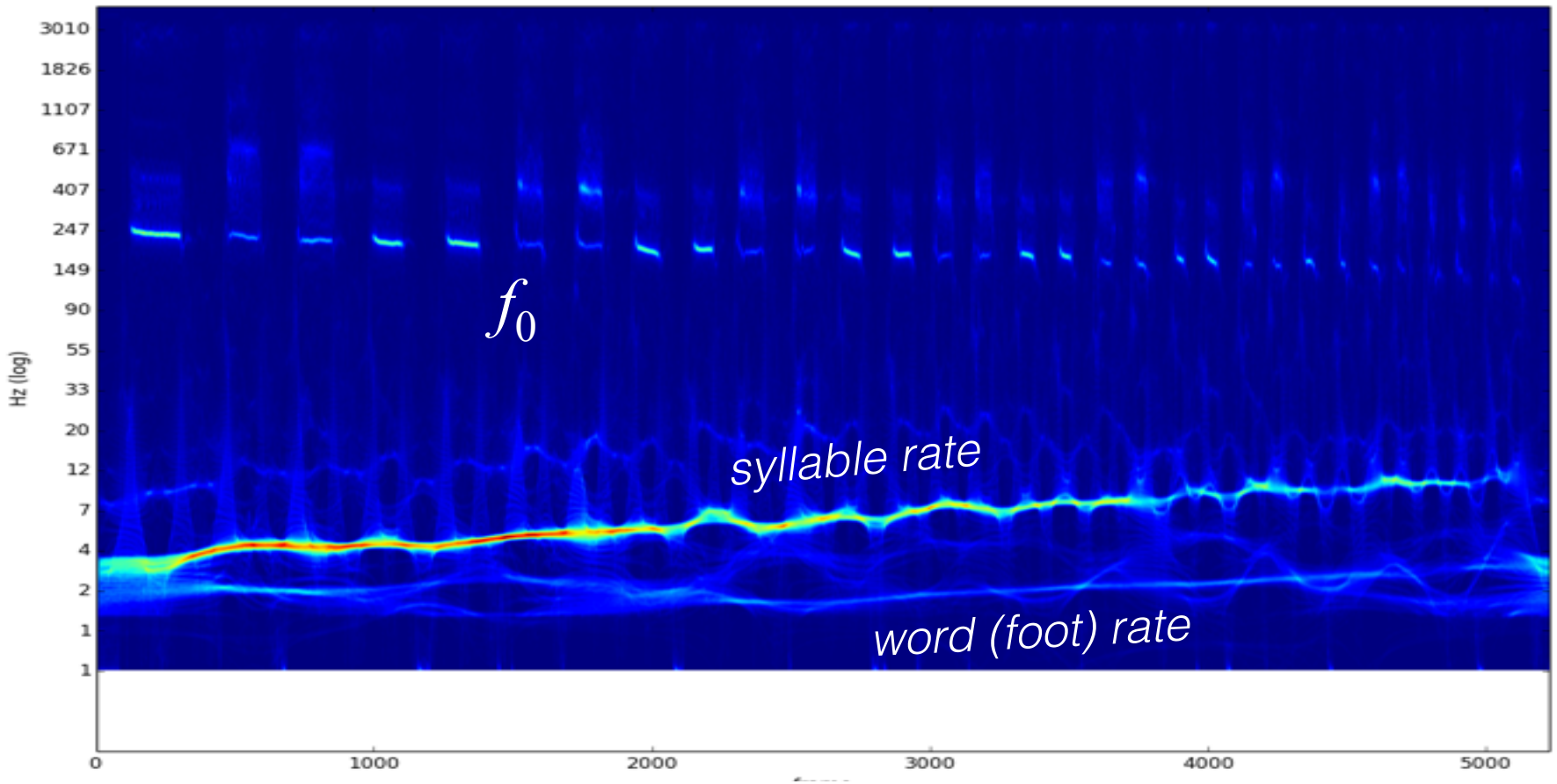
A “complete” (upside-down) scalogram

- quite heavily engineered: calculated instantaneous frequencies at each scales and then plotted in frequency domain (aka Hilbert spectrum)

- huge range of frequencies (compared to Fourier Transform)

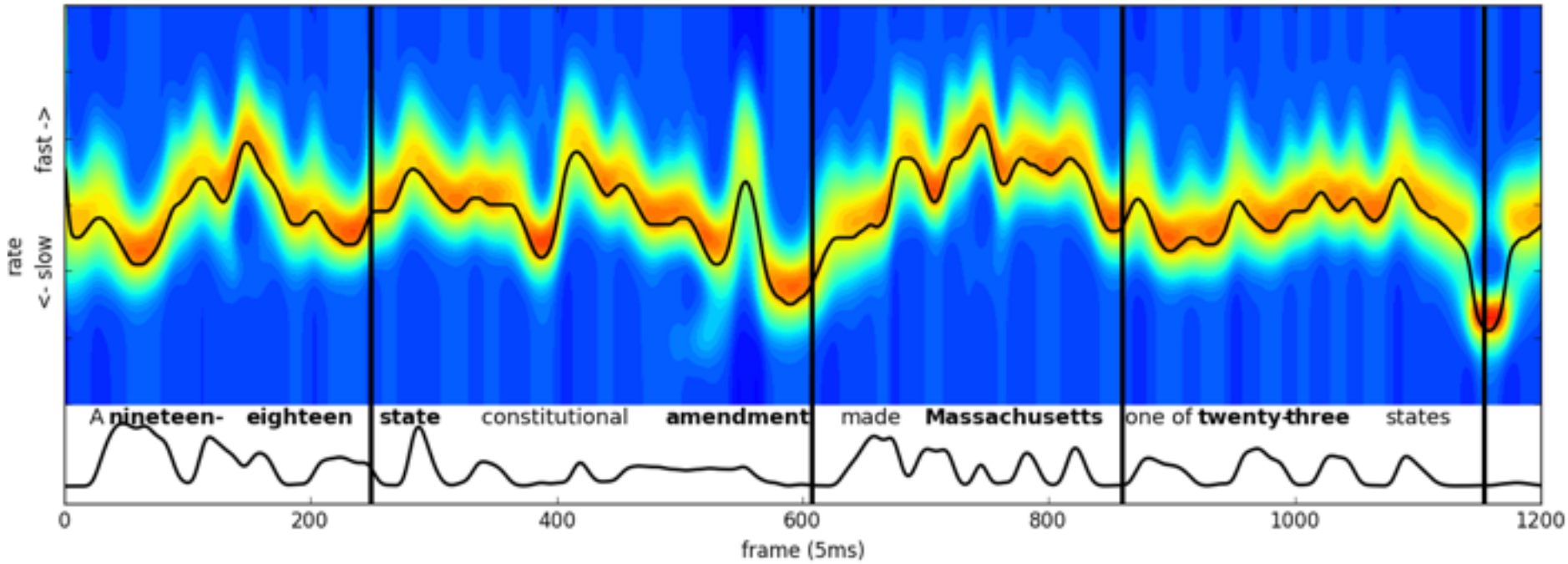


f_0 and speaking rate(s)



Speaking rate

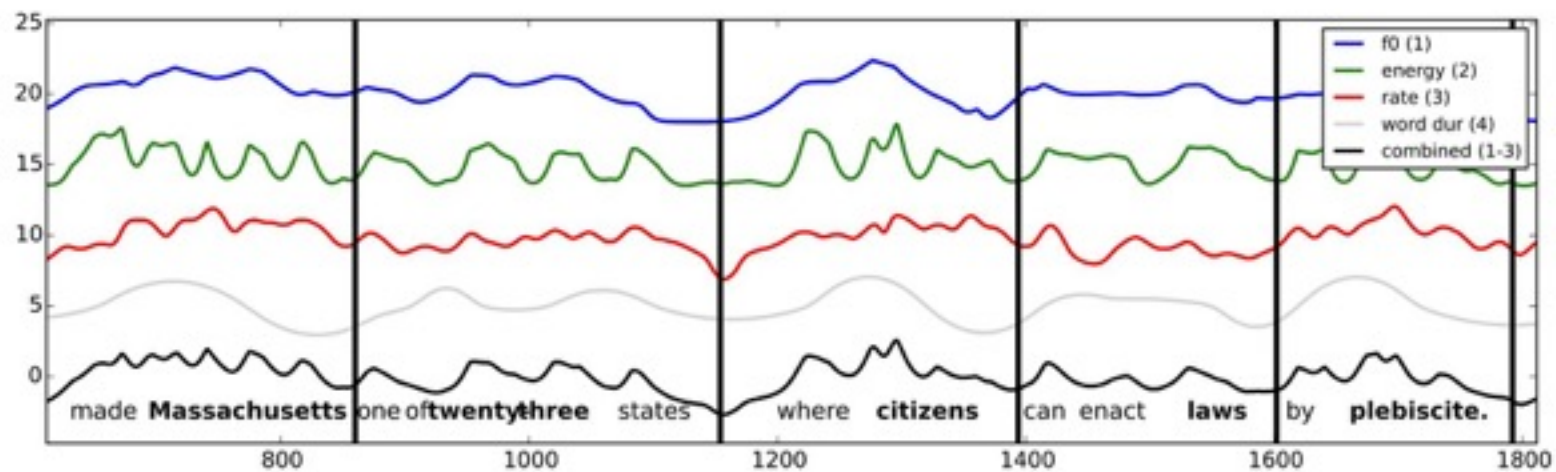
- band-pass energy signal, Morlet wavelet
- amplitude scalogram (abs), normalize per frame
- follow ridge in time by viterbi



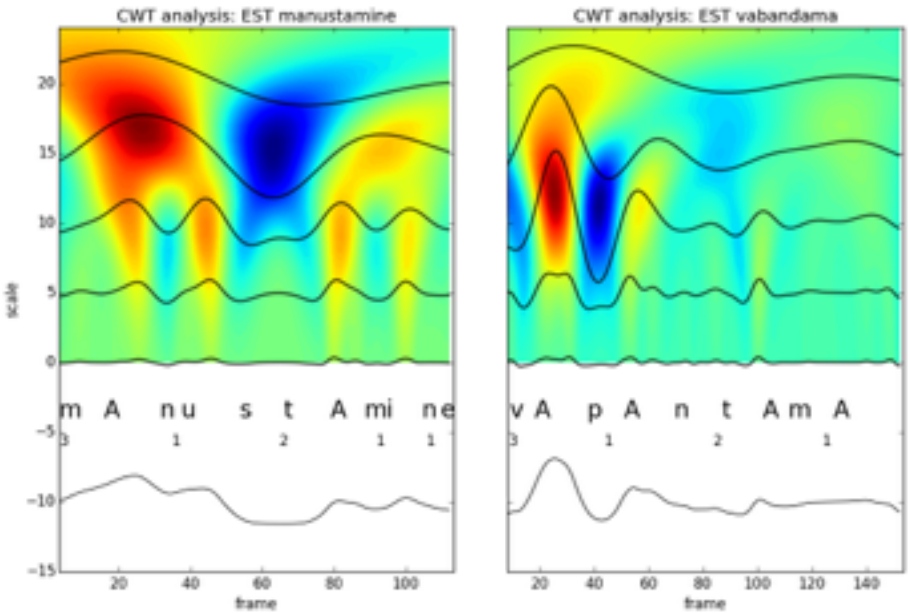
Different signals

- other speech signals cab be processed by CWT:
  - (interpolated) f_0
  - (interpolated) energy (perhaps obtained via CWT)
  - duration signal (interpolated durations)
  - speaking rate (obtained via CWT)

- or even a combination thereof…

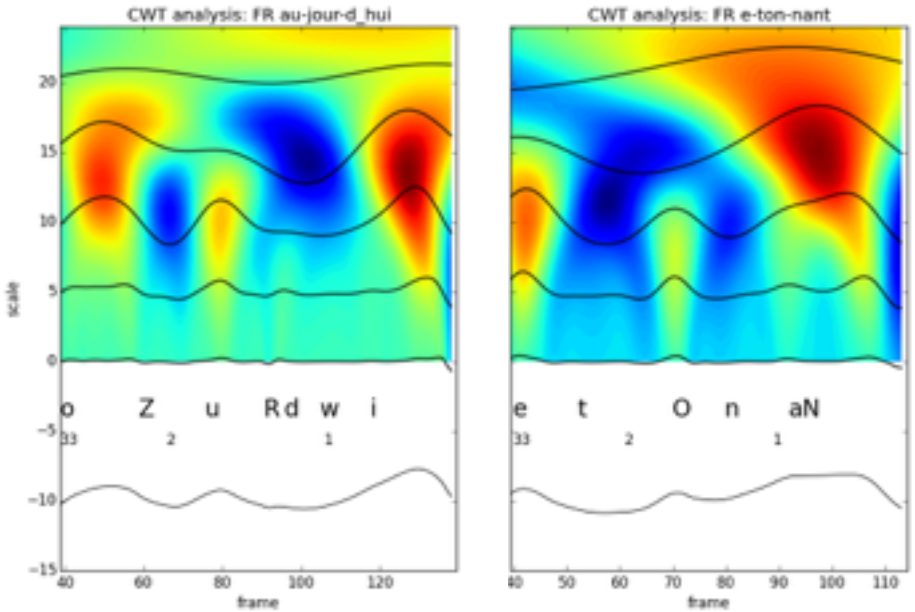


Prominence detection: lexical stress



Estonian: word initial

French: word final

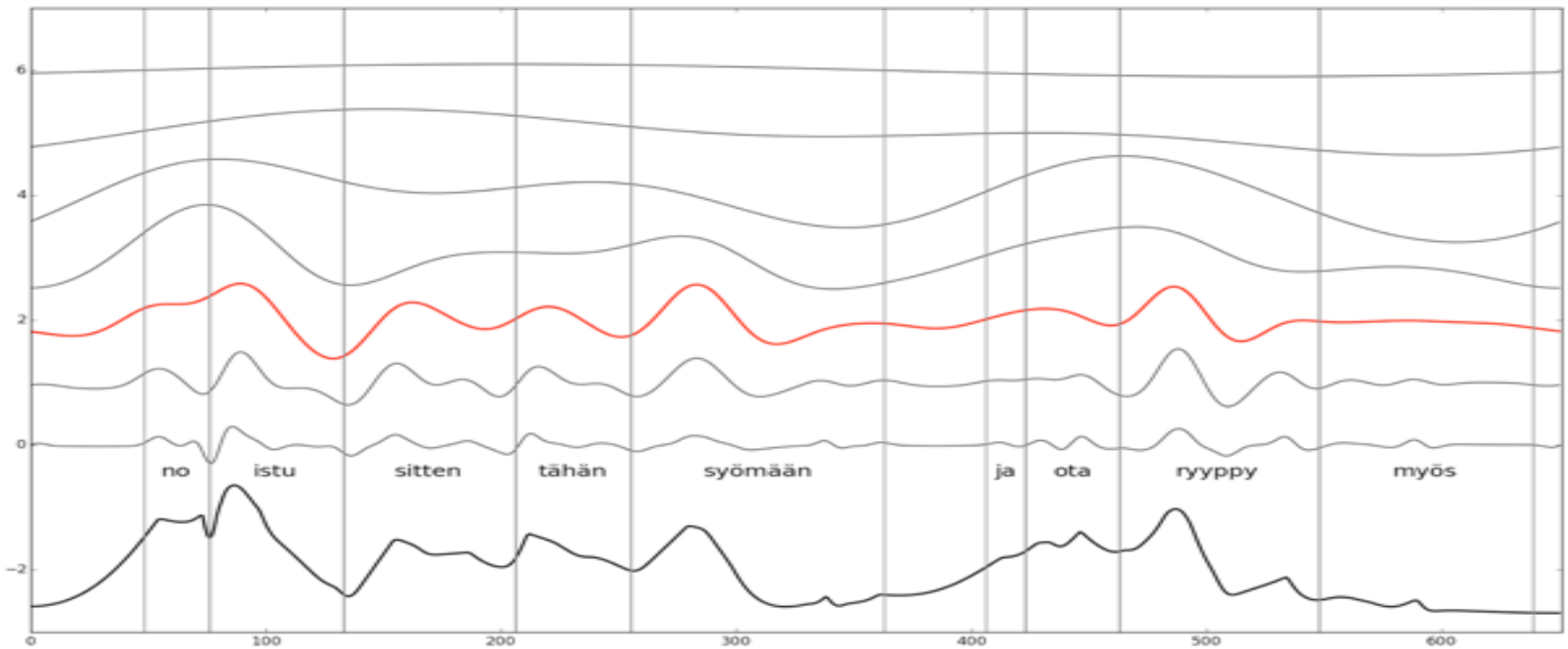


Prominence detection

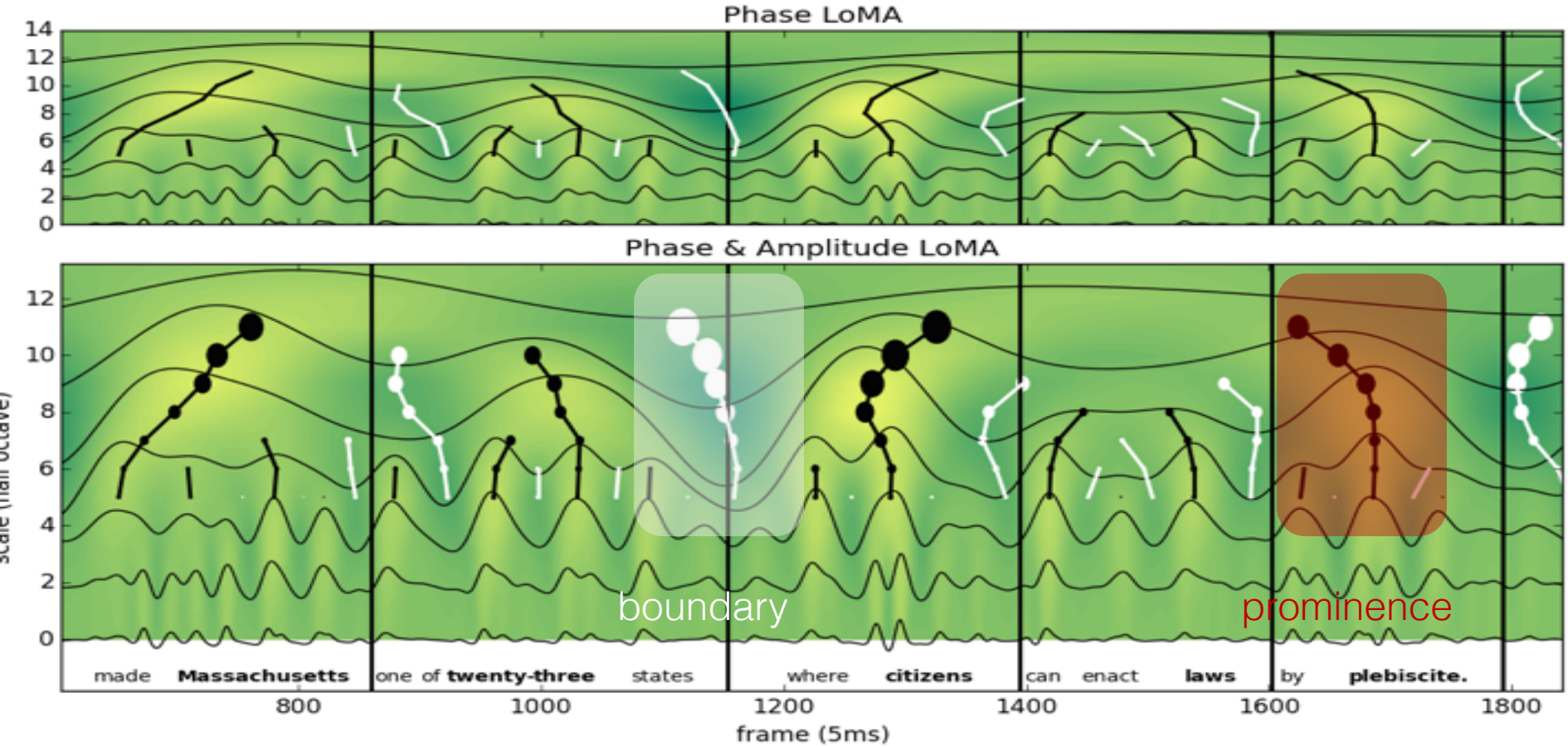
1. Perform wavelet analysis on interpolated f_0

2. Select the scale with the closest match of number of peaks and number of words in the utterance

3. Prominence = the maximum peak of the word



Combining prominence and boundary detection



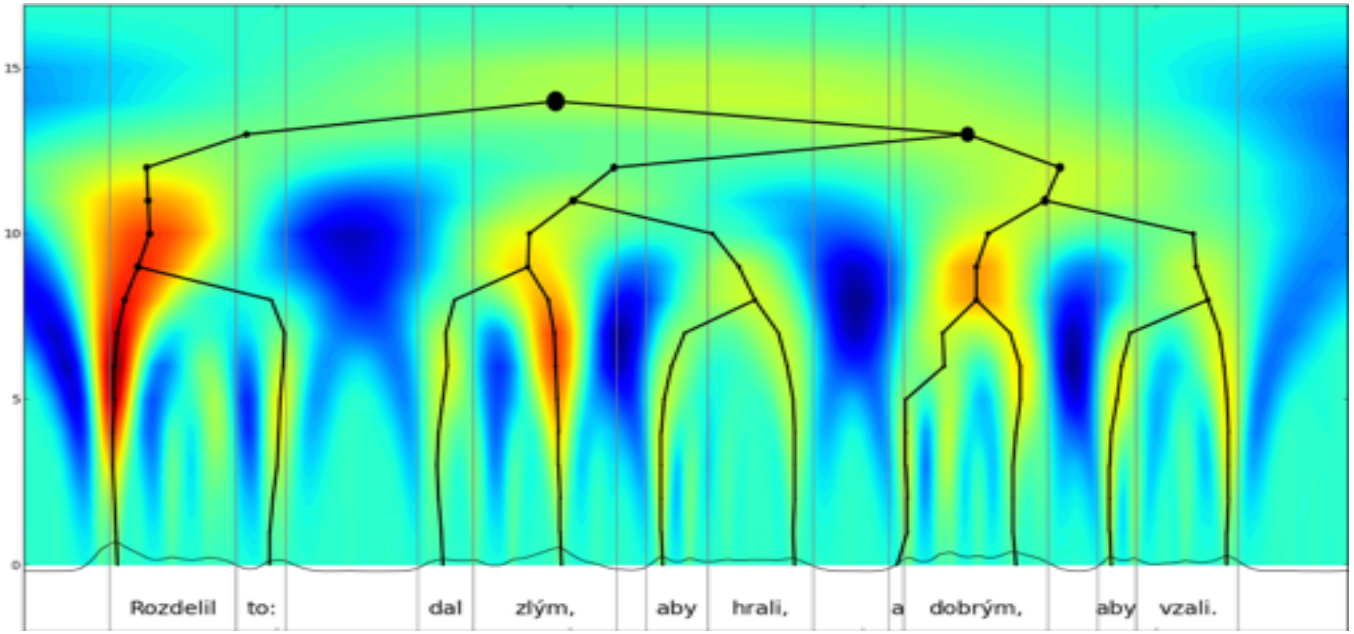
- Suni, Šimko, Aalto & Vainio (2016). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*
- Suni, Šimko & Vainio (2016). Boundary detection using Continuous Wavelet Analysis. *Proc.Speech Prosody*, Boston
- Suni & Kocharov (soon).

Prosodic hierarchy

- Speech (and language) is structured hierarchically
- Utterances consist of (temporally) nested phrases > (phonological) words > syllables > speech sounds > acoustic events
- CWT allows for a time-frequency localisation of these structural components in a speech signal and can reveal how they are nested

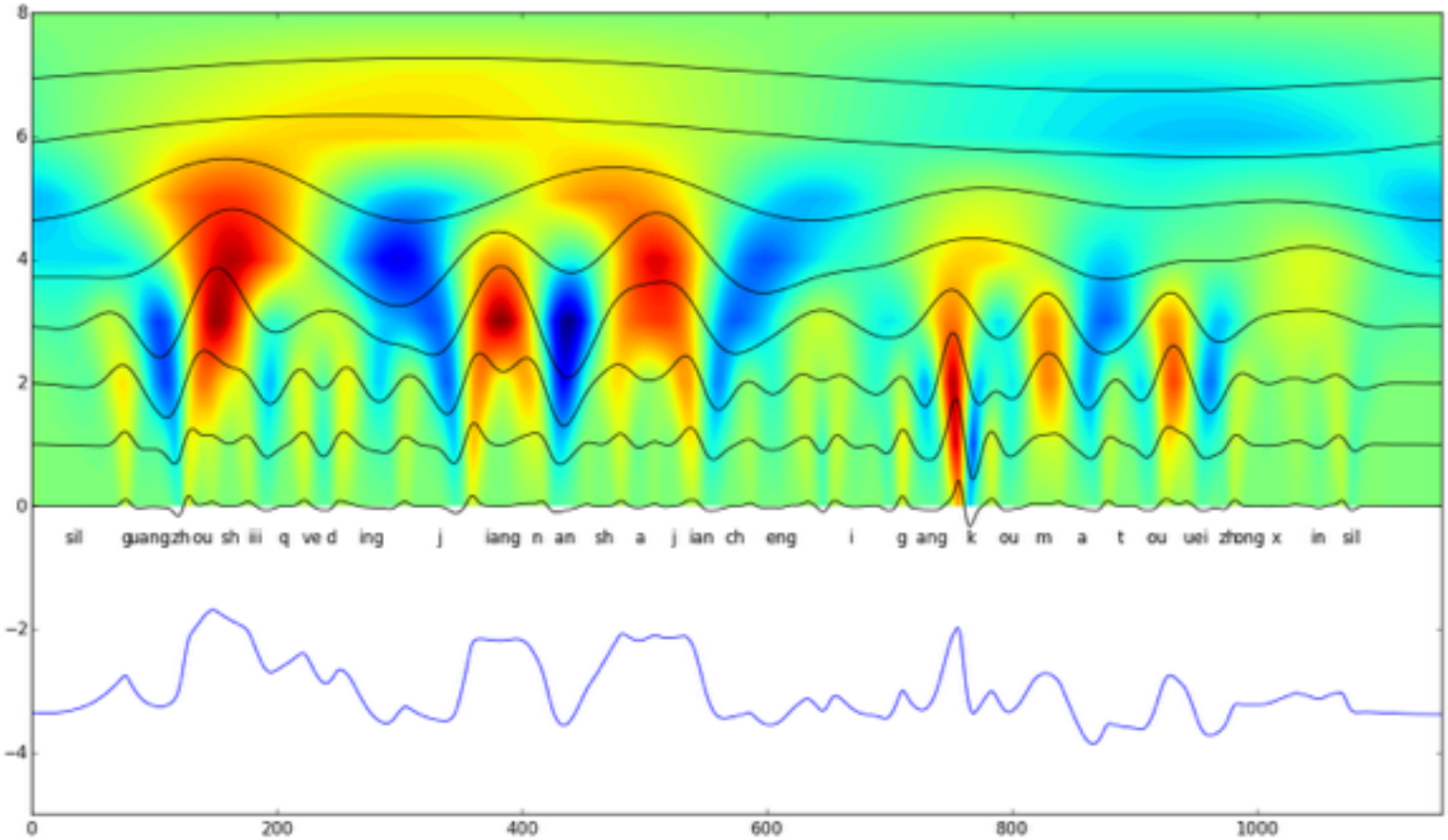
[[Rozdelil to:] [[[dal zlým] [aby hrali]] [[a dobrým] [aby vzali.]]]]

[*He*] divided it: [*he*] gave [*to*] baddies to play and [*to*] goodies to take.

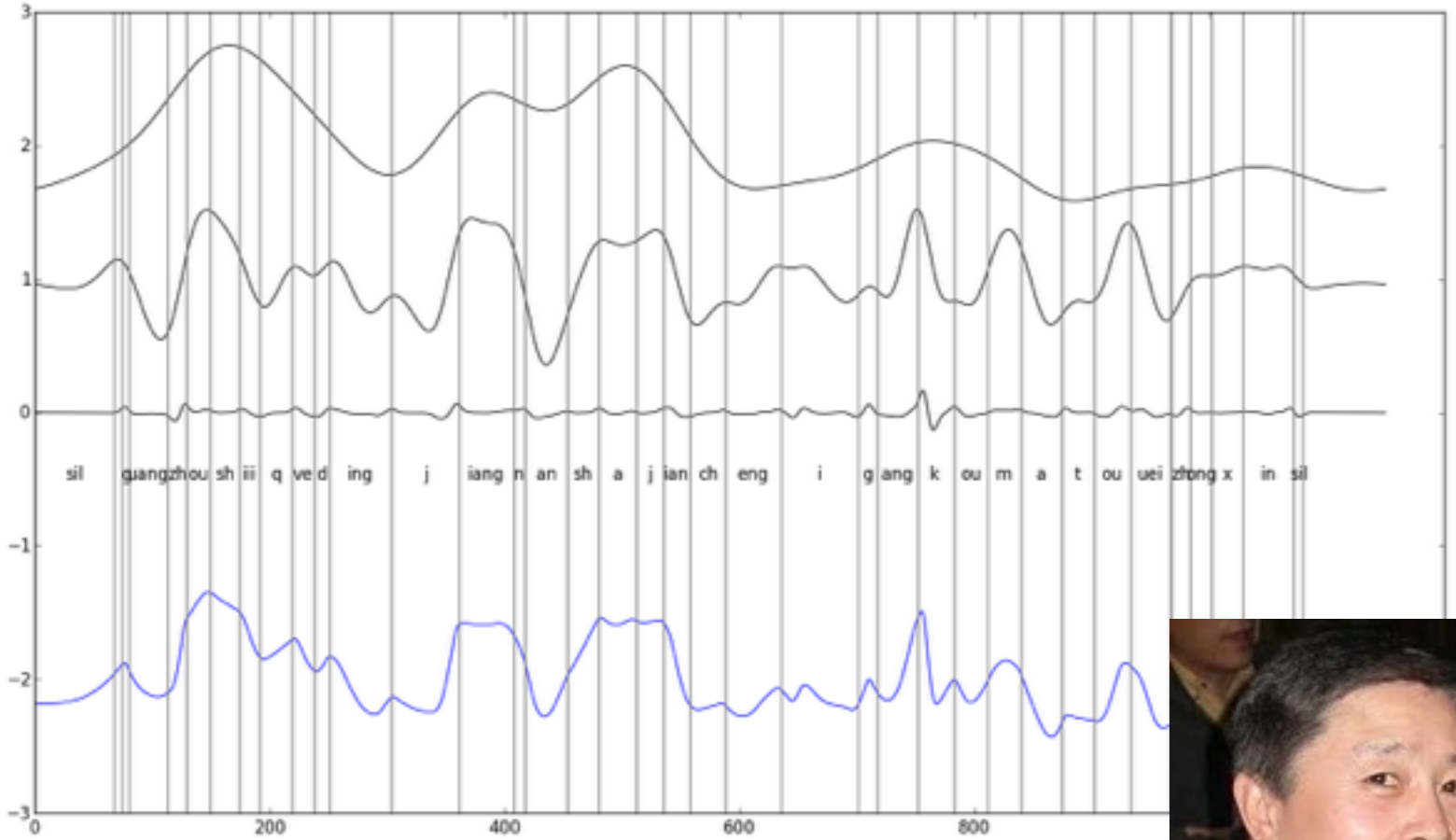


Dannenberg, Vainio, Suni & Werner. (2015). Prosodic and syntactic segmentation of spontaneous speech: A preliminary study. In *Proceedings of ICPhS*, Glasgow

Prosodic hierarchy: The case of lexical tone



Prosodic hierarchy: The case of lexical tone



Thank you!


Thank you!