Tone and quantity (and prominence): the psychoacoustic trade-offs in prosodic signaling

Martti Vainio
Department of Digital Humanities
University of Helsinki

TAL 2018, Berlin, June 18-20, 2018

Collaborators

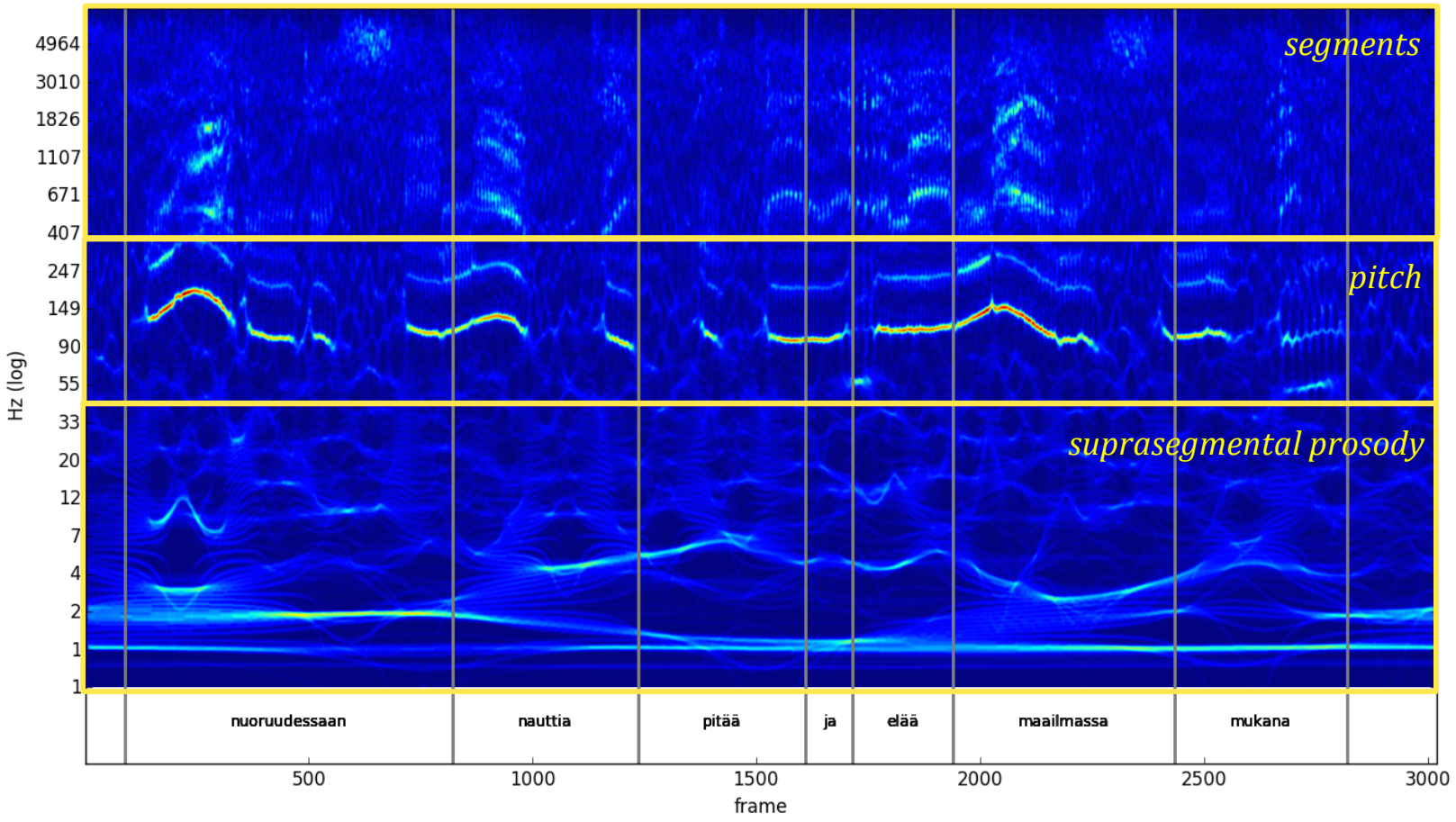
- **Juraj Šimko**, Department of Digital Humanities, University of Helsinki, Finland
- Antti Suni, ————————”————————
- **Daniel Aalto**, Institute for Reconstructive Sciences in Medicine, University of Alberta, Edmonton, Canada
- **Juhani Järvikivi**, Department of Linguistics , University of Alberta, Edmondon, Canada



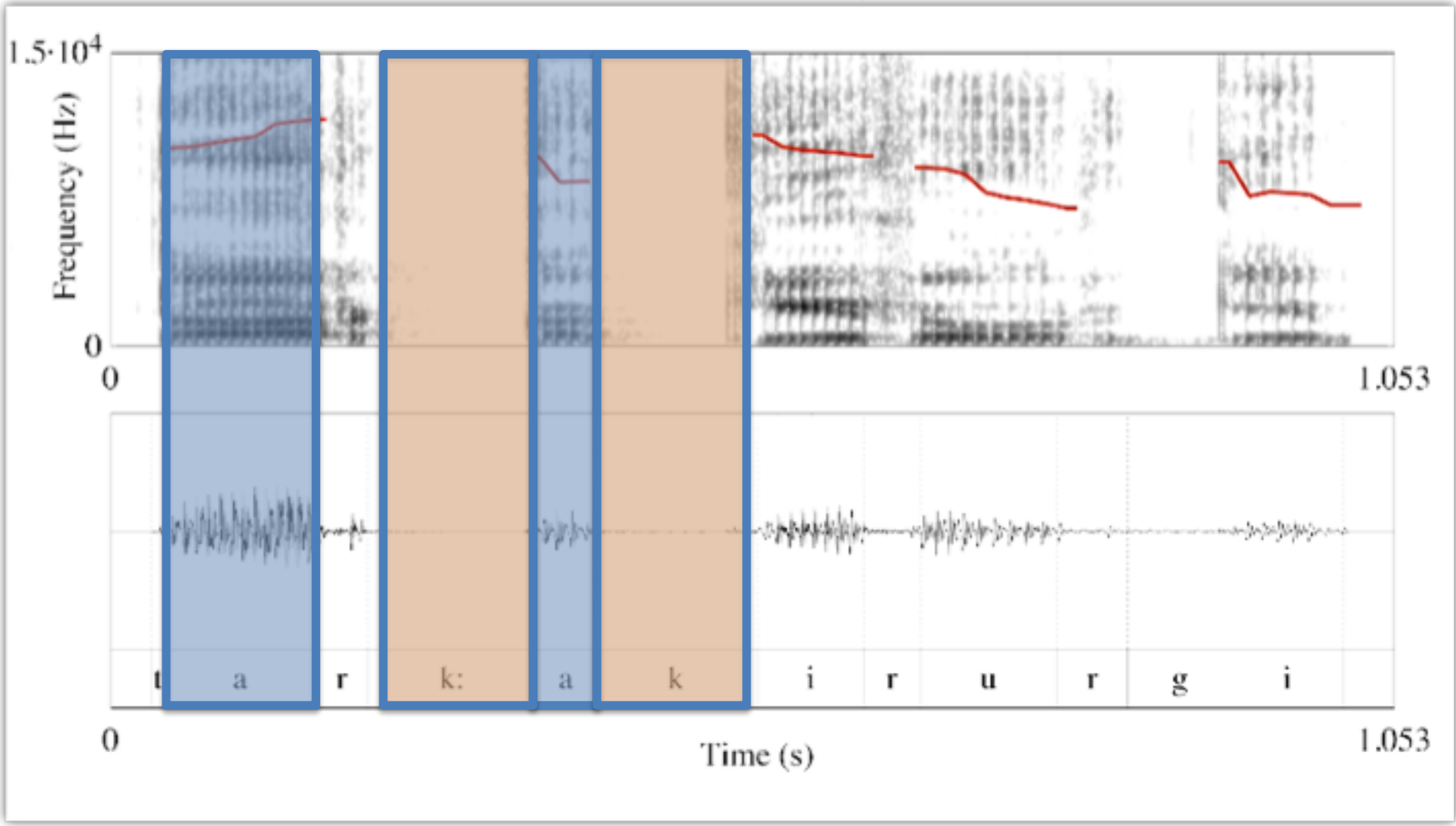

Tone, Quantity, and Prominence

- in speech, we frequently mark something as more important, standing out from the rest
- we can make some chunk of speech signal (corresponding to a syllable, word, phrase,…) more **prominent** by (generally) *increasing* pitch, intensity, duration
- how do the T&Q interact?
- prominence-based account of prosody

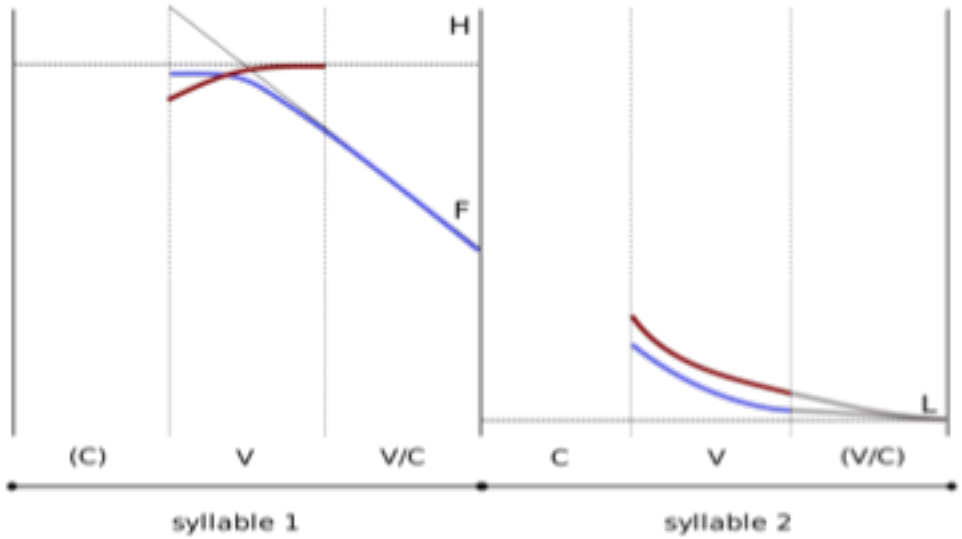
The prosodic speech signal

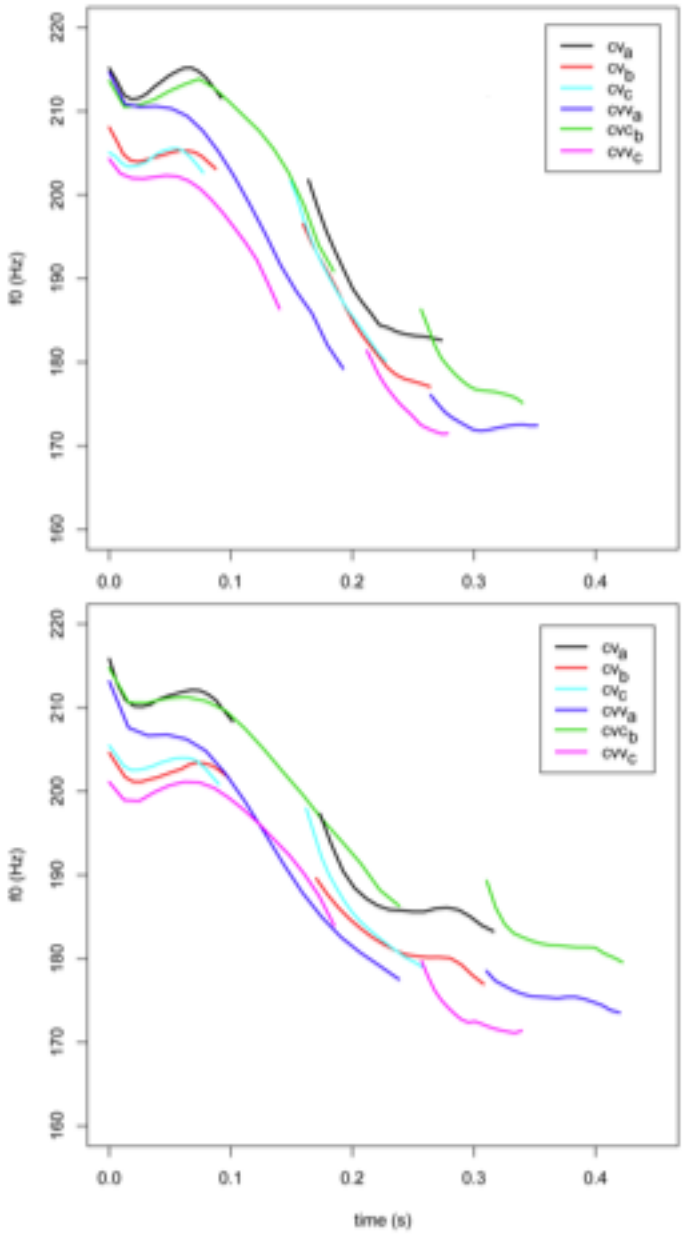


Quantity

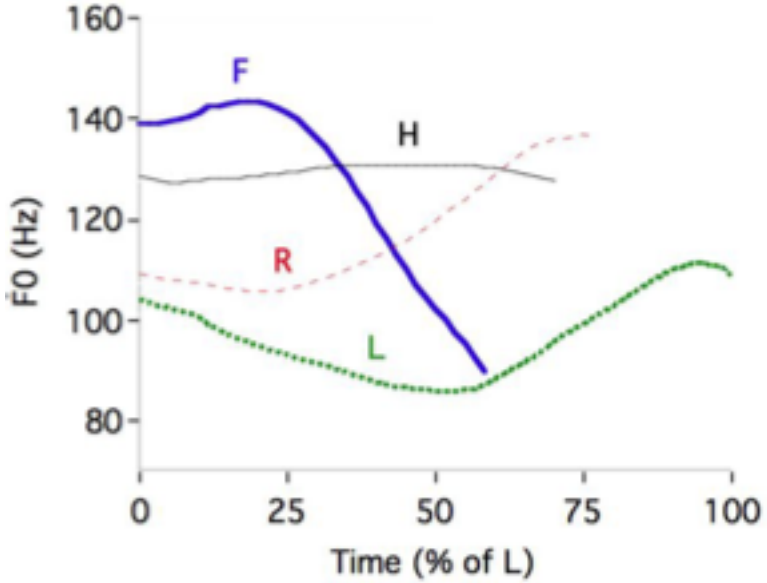


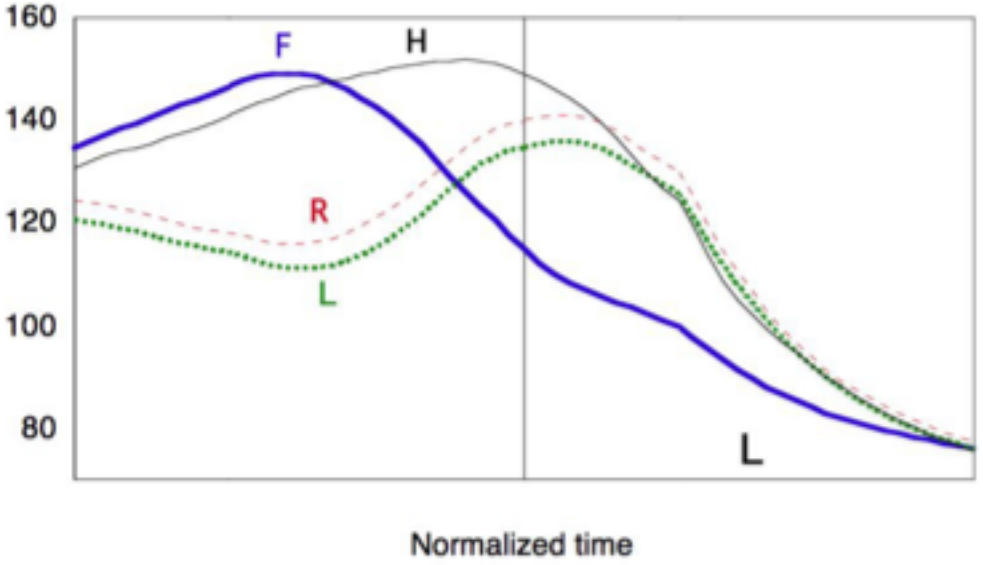
Quantity and f_0

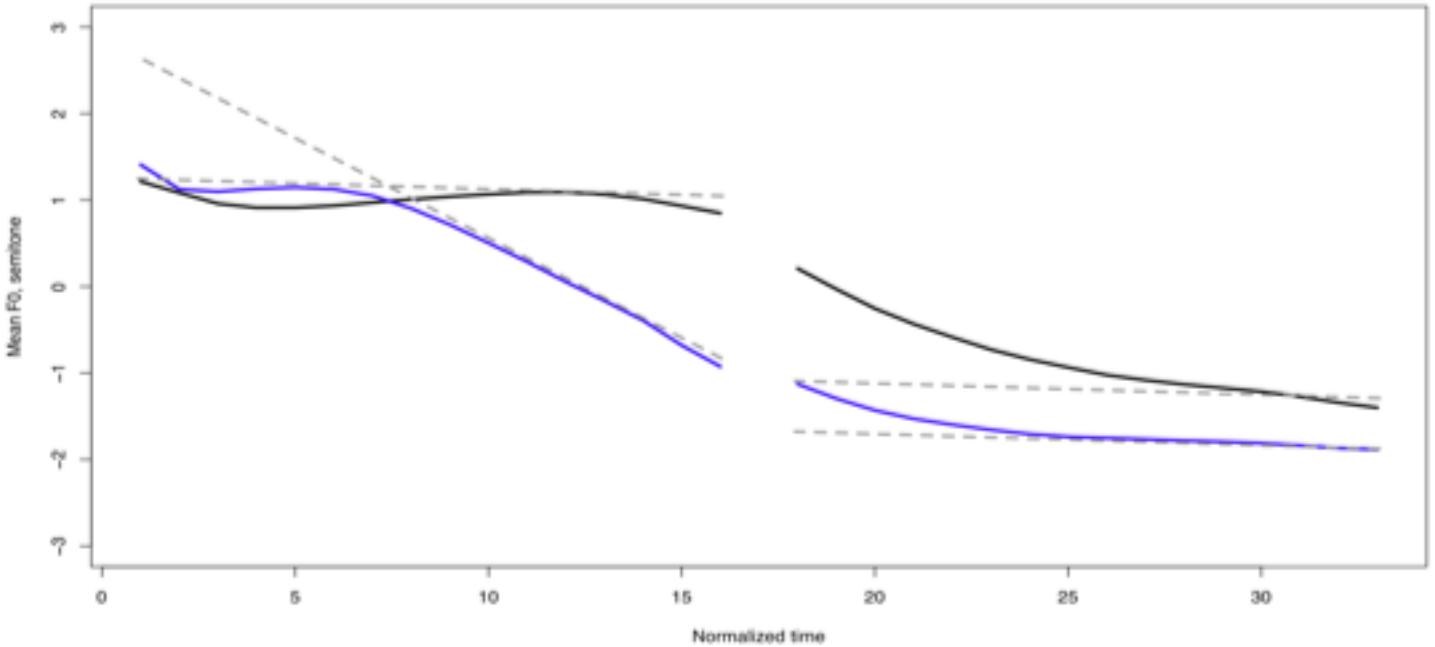


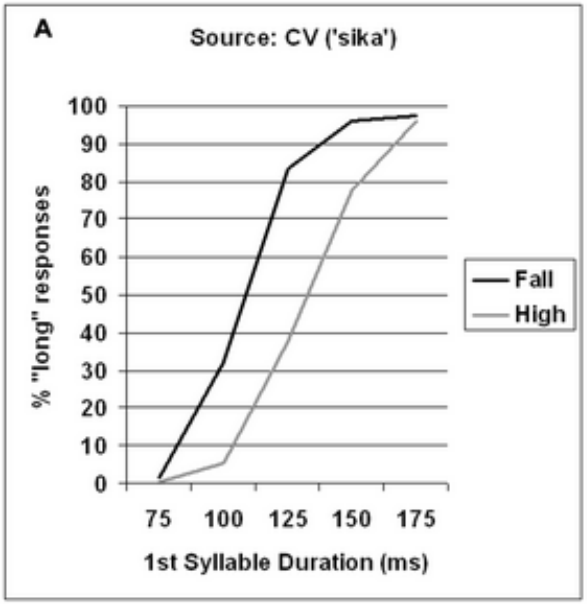


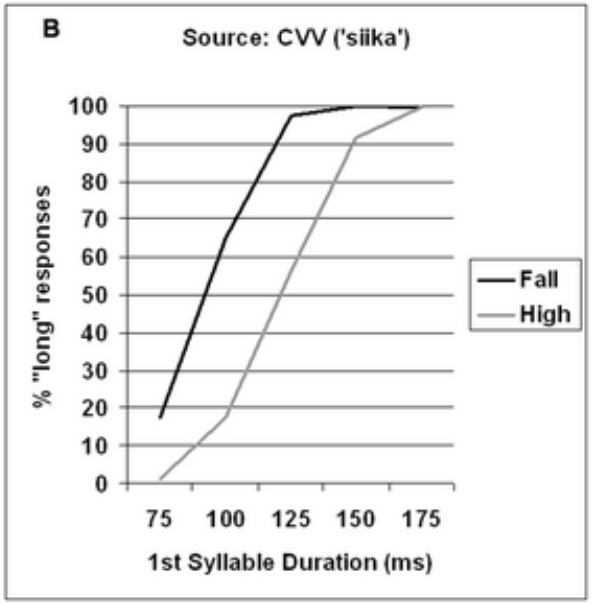
Martti Vainio, Juhani Järvikivi, Daniel Aalto, and Antti Suni, Phonetic tone signals phonological quantity and word structure *J. Acoust. Soc. Am. 128, 1313 (2010), DOI:10.1121/1.3467767, 2010*


160
140
120
100
80
F0 (Hz)
F
H
R
L
0
25
50
75
100
Time (% of L)


160
140
120
100
80
F
H
R
L
L
Normalized time


Mean F0, semitone
3
2
1
0
-1
-2
-3
0
5
10
15
20
25
30
Normalized time





Tone	Duration	Errors (%)	RT (ms)	Priming (ms)
High level	Long	5.2	583	+33
High level	Short	4.6	572	+44
Fall contour	Long	6.9	548	+68
Fall contour	Short	6.3	567	+49
Unrelated	Unrelated	11.5	616	-

Positive sign in the column for Priming indicates facilitation (in ms) compared to the unrelated control condition.

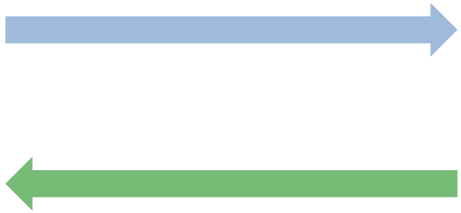
doi:10.1371/journal.pone.0012603.t003

Järvikivi J, Vainio M, Aalto D, Real-Time Correlates of Phonological Quantity Reveal Unity of Tonal and Non-Tonal Languages. *PLoS ONE* 5(9): e12603. doi:10.1371/journal.pone.0012603, 2010.

production ←→ perception

perceptual compensation theory: we know that low-tone syllables are longer, so when we hear equally long high and low pitch syllables, the higher pitch one appears longer (Gussenhoven)

low-tone syllables are generally (on average) **produced** *longer* than high-tone ones



if equally long, low-pitch tones are generally (on average) **perceived** as *shorter* than high-pitch ones

production compensation theory: we *hear* higher pitch syllables (sounds) as longer than lower pitch ones of equal duration, and to compensate for this perceptual effect we produce the lower ones a bit longer (Yu)

Problem 1

perceptual compensation theory: we know that low-tone syllables are longer, so when we hear equally long high and low pitch syllables, the higher pitch one appears longer (Gussenhoven)

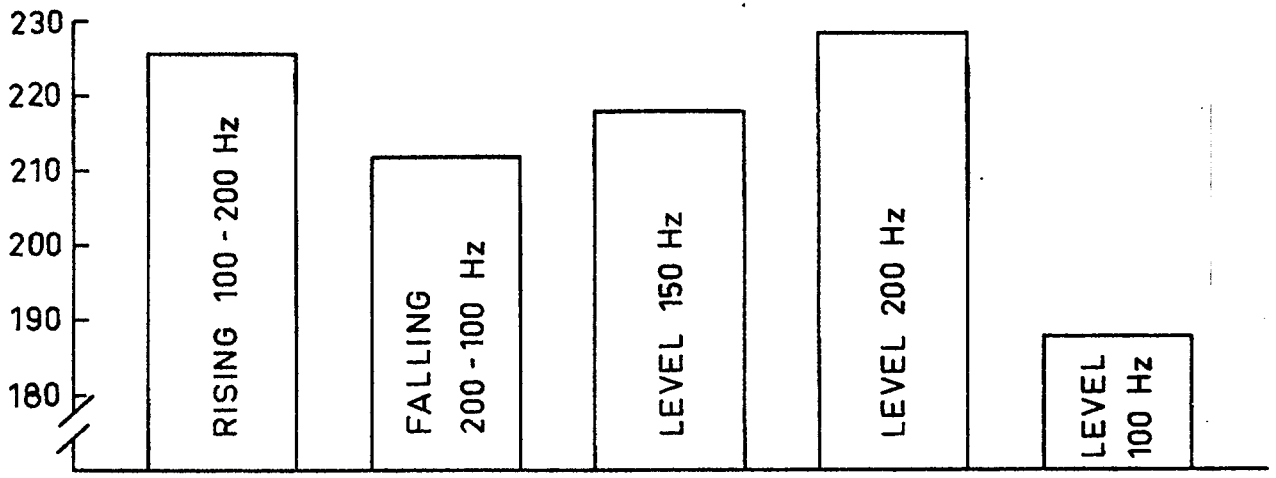
Can we attempt to decide between these two theories?

Is the fact that the higher sounds are perceived as longer than lower ones (of the same duration) based on our perceptual, auditory apparatus?

production compensation theory: we *hear* higher pitch syllables (sounds) as longer than lower pitch ones of equal duration, and to compensate for this perceptual effect we produce the lower ones a bit longer (Yu)

"we *hear* higher pitch syllables (sounds) as longer than lower pitch ones of equal duration"

- this is a well established fact (but, of course, we can't be sure if it isn't down to the "perceptual compensation effect")
- Burghardt (1972) has shown it for sinusoid tones
- Lehiste (1976), Rosen (1977) did similar thing for speech syllables



***Point of subjective equality**: Duration of a second /a/ sound at which it is perceived as of equal duration as a preceding 200 ms 150 Hz standard (from Rosen, 1977)*

"we *hear* higher pitch syllables (sounds) as longer than lower pitch ones of equal duration"

- shape of the tone makes a difference, too
- comparing level, falling and rising speech sounds: rising tones perceived as longer than falling ones, and both longer (?) than the level ones
- Lehnert-LeHouillier (2007) for speakers of Latin American Spanish, German, Thai and **Japanese**
- Yu (2010) for **English** speakers,
- Cumming (2011) for speakers of **Swiss German**, **Swiss French** and **French**
- Gussenhoven (2013) for **Dutch** and **Mandarin** speakers

- but speakers of all languages heard higher pitch syllables as longer!!!

Problem 2

- shape of the tone makes a difference, to
- comparing level, falling and rising sp… …ones perceived as longer than falling … (?) than the level ones
- Lehnert-LeHouillier (200… …n American Spanish, German, Thai and **Ja**
- Yu (2010) for **En**
- Cumming (2… …**Swiss German**, **Swiss French** and **French**
- Gu… …**Dutch** and **Mandarin** speakers
- b… … all languages heard higher pitch syllables as lon…

Are there any quantitative differences based on native language in terms of the effect of sound fundamental frequency on its perceived duration?

Intensity-pitch confound

- (I forgot to tell you) louder sounds are also perceived as longer than quieter ones of the same duration

Btw, this is a more general phenomenon: for example, darker objects look heavier than brighter but otherwise identical objects (Walker, Francis and Walker, 2010)

- and lower pitch sounds (pure tones) are "objectively" quieter than higher pitched ones

Problem 3

- (I forgot to tell you) louder sounds are
  perceived as longer than quieter o
  same duration

Btw, this is a more general phenor
objects look heavier than brial
objects (Walker, Francis a

- and l (pure tones) are
  “ than higher pitched ones

Is the effect of fundamental frequency of sound on its perceived duration purely due to the fact that higher sounds are louder (and therefore perceived as longer)?

Three questions

Q1 *Is the fact that the higher sounds are perceived as longer than lower ones (of the same duration) based on our perceptual, auditory apparatus?*

Q2 *Are there any quantitative differences based on native language in terms of the effect of sound fundamental frequency on its perceived duration?*

Q3 *Is the effect of fundamental frequency of sound on its perceived duration purely due to the fact that higher sounds are louder (and therefore perceived as longer)?*

Q3

Is the effect of fundamental frequency of sound on its perceived duration purely due to the fact that higher sounds are louder (and therefore perceived as longer)?

1. Let's play people 2 sounds and ask which one they perceive as longer!

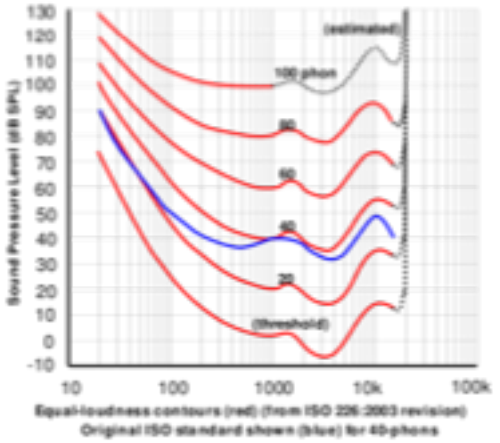
But what kind of sounds?

- maybe not directly speech sounds (WHY?)
- let's try to make the sounds such that pitch-loudness dependence is minimized (WHY?)

- let them vary in all dimensions: duration, f0 and intensity, plus the shape of f0 contour (WHY?)
- let's also ask other questions: e.g., which one is louder (WHY?)

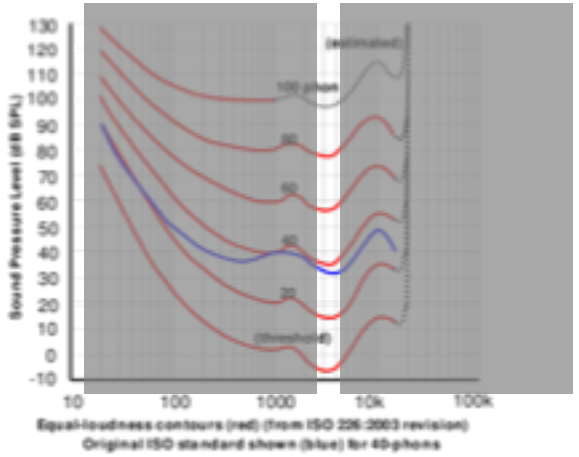
Minimizing pitch-loudness dependency

Let’s make sounds with no frequency components other than in the area where f0 influence matters the least!

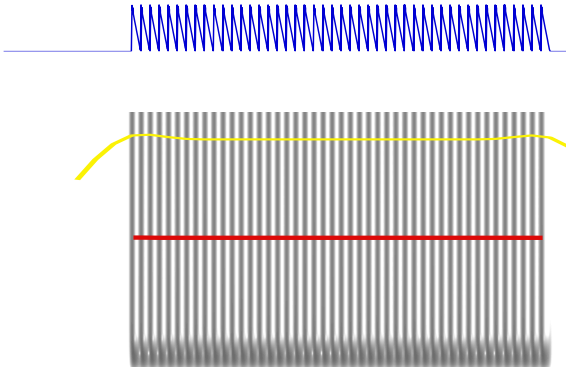


Minimizing pitch-loudness dependency

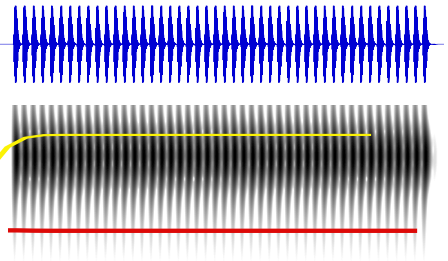
Let's make sounds with no frequency components other than in the area where f0 influence matters the least (and where our hearing is the most sensitive)!



1. Let's make a sawtooth wave with appropriate *frequency* and *duration*



2. Let's band-pass filter out everything apart from stuff around 3.2 kHz (gamma-filter)



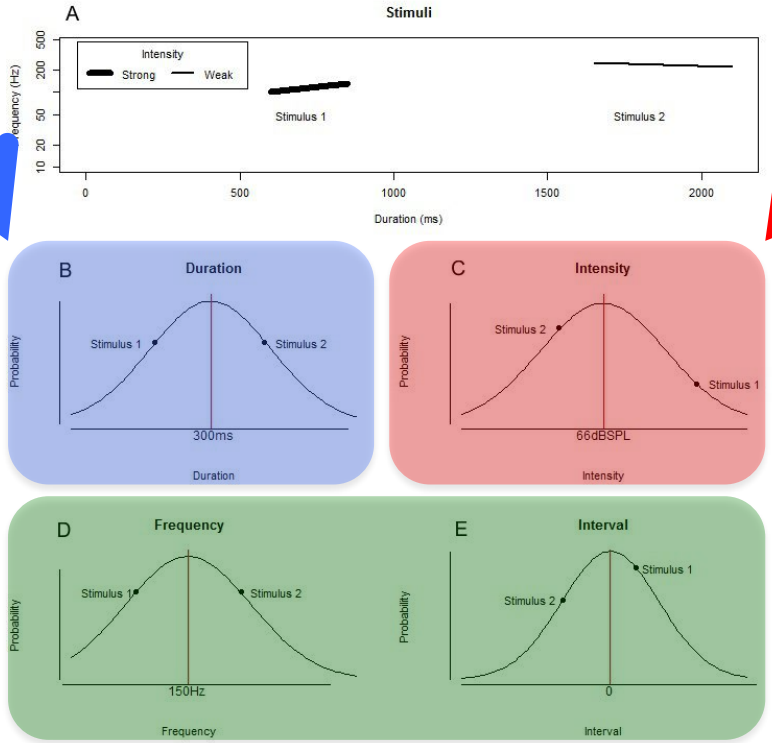
3. Let's normalize *intensity* and then adjust it to a required *level*

Also, sort out some technical issues such as cutting the sawtooth at a right spot…

Frequency, duration, intensity

We make a pair of sounds, with randomly selected parameters:

1. Let's make a sawtooth wave with appropriate ***frequency*** and ***duration***
2. Let's band-pass filter out everything apart from stuff around 3.2 kHz (gamma-filter)
3. Let's normalize ***intensity*** and then adjust it to a required ***level***



Which one is **longer**?

Which sound was longer?
First (press "a”) or second (press "x”)

This question was repeated around 300 times (with different pairs of sounds)

And we run this “torture” with around 15-20 subjects (depending on the experiment)

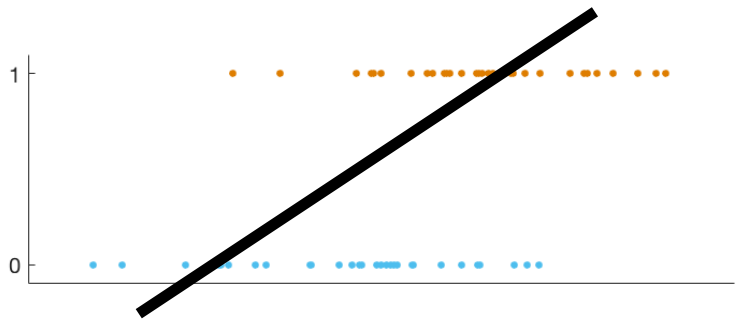
Sometimes we asked a different question / used different stimuli

Which one is longer?

We log all the parameters and responses:

phase	order	dd	a_dur	x_dur	dp	a_per	x_per	di	a_int	x_int	isi	gain1	gain2	response
2	2	-0.070862	0.29107	0.36193	0.0016019	0.0085616	0.0069597	1.3095	-2.1983	-0.88888	0.80431	-0.24753	1.526	0
2	3	0.061383	0.30236	0.24098	0.00045448	0.0059369	0.0054824	-2.9678	1.8569	-4.8247	0.8065	0.40022	1.5168	1
2	4	0.10608	0.3881	0.28202	-0.0011734	0.0053976	0.0065711	-0.67699	2.2505	2.9274	0.79024	-0.1708	3.5522	1
2	5	0.10127	0.37866	0.27739	-0.0006118	0.0080515	0.0086633	0.44069	-2.3487	-1.9081	0.79625	2.1553	-2.3784	1
2	6	0.00086168	0.33821	0.33735	-0.0028353	0.0063243	0.0091595	1.5117	7.8559	-6.3442	0.78924	5.4285	0.2026	1
2	7	-0.011383	0.35376	0.36515	0.00059849	0.0082264	0.0076279	-4.0152	-0.14983	4.1651	0.79986	-2.9441	0.95277	0
2	8	0.14374	0.40914	0.2654	-0.001108	0.0055319	0.0066399	-2.4242	1.2842	-3.7084	0.78106	1.3864	1.7181	1
2	9	-0.013447	0.27156	0.28501	0.0011371	0.009049	0.007912	0.43174	-1.435	-1.0033	0.80402	1.9978	-4.6574	0
2	10	0.082993	0.41653	0.33354	0.00321	0.010159	0.0069487	0.22936	-0.44307	0.21371	0.81991	-5.202	2.6955	0
2	11	0.12304	0.3266	0.20356	0.0011541	0.0056763	0.0045223	4.8113	6.9393	-2.128	0.79381	-3.0102	0.1527	0
2	12	0.0080045	0.3146	0.3066	0.0036034	0.0095578	0.0059544	-4.5775	3.0284	7.606	0.78373	-2.8943	3.5951	0
2	13	-0.12569	0.28823	0.41392	-0.0013968	0.0080104	0.0094072	0.1461	0.75456	-0.60846	0.81431	1.5316	2.053	0
2	14	-0.072721	0.34866	0.42138	-0.0023265	0.0056235	0.00795	-0.36164	-0.66857	-1.0302	0.79854	0.22259	4.435	1
2	15	0.00029478	0.39787	0.39757	-0.001491	0.0055264	0.0070175	-4.5857	0.87043	5.4561	0.81105	2.0851	-2.8007	1
2	16	0.017551	0.32413	0.30658	-0.000805	0.00601	0.006815	2.2052	2.8715	0.66634	0.79781	0.28547	-2.3947	1
2	17	-0.028413	0.19957	0.22798	-0.0004417	0.0066745	0.0071162	-0.096208	3.789	-3.8853	0.8011	3.0855	3.4877	1
2	18	-0.18084	0.23113	0.41197	0.0011186	0.0070033	0.0058847	-0.73532	-0.25484	-0.99016	0.79678	1.5357	0.20651	0
2	19	0.046009	0.2251	0.17909	-0.0013561	0.0050421	0.0063981	5.5129	6.8272	1.3143	0.79585	3.563	-0.94994	1
2	20	-0.0094558	0.21256	0.22202	1.21E-05	0.0088832	0.0088711	-2.0945	-5.3648	-7.4593	0.81729	-0.39508	-4.8469	1
2	21	-0.13898	0.22061	0.35959	0.0021559	0.0084951	0.0063393	-2.1753	2.4121	4.5875	0.81835	-0.041927	-0.17071	0
2	22	-0.088367	0.21308	0.30145	-0.0010867	0.0057589	0.0068456	0.77698	-2.6588	-1.8818	0.79203	2.9194	0.67056	1
2	23	-0.0026531	0.37259	0.37524	-0.002328	0.006013	0.008341	-0.047388	-3.8045	-3.8519	0.78987	0.069616	1.6046	0
2	24	0.1381	0.42107	0.28297	-0.0002428	0.0074014	0.0076442	5.3453	-5.9629	-0.61767	0.79998	2.3111	5.1989	1
2	25	-0.17535	0.27141	0.44676	0.0005264	0.0077664	0.00724	-3.8041	1.6766	5.4807	0.81239	-0.90795	5.9026	0
2	26	0.044036	0.35671	0.31268	0.0022828	0.0099191	0.0076363	-0.85052	1.3694	2.2199	0.81624	-4.6132	-1.2616	0
2	27	-0.19161	0.15138	0.34299	0.0020089	0.0065826	0.0045737	-1.0958	0.7586	-1.8544	0.81282	-2.6226	-3.1488	0
2	28	-0.034671	0.27209	0.30676	-0.0029715	0.0045516	0.0075231	-0.64642	-5.278	5.9245	0.79957	-0.58602	2.0385	0
2	29	-0.19209	0.20696	0.39905	-0.001782	0.0059155	0.0076975	-3.407	0.99084	4.3978	0.80203	-2.6298	-5.4828	0
2	30	0.078821	0.3744	0.29558	-0.0007245	0.0079657	0.0086902	-0.31339	0.19392	-0.50731	0.78567	-0.70306	1.1216	1
2	31	-0.12617	0.21424	0.34041	0.0024701	0.0076546	0.0051845	-6.4171	0.8247	-7.2418	0.7831	0.23817	-2.0789	0
2	32	-0.064943	0.21356	0.2785	0.0033443	0.0093944	0.0060501	4.6997	6.7329	-2.0332	0.80628	-2.0655	1.2013	0
2	33	0.0073469	0.34336	0.33601	0.0010853	0.0060451	0.0049598	-0.25507	4.413	4.6681	0.79533	2.0481	3.1025	0
2	34	-0.15397	0.25467	0.40864	-0.001789	0.0057877	0.0075768	-2.5376	0.066569	2.6041	0.80191	-0.95975	-0.0001426	0
2	35	-0.089524	0.21376	0.30329	0.0010532	0.0059477	0.0048945	2.4466	-4.6459	-2.1993	0.79746	-1.8188	-1.4977	0
2	36	-0.018866	0.2095	0.22837	-0.00167	0.0087393	0.010409	-0.5495	2.3438	2.8933	0.81352	-1.9102	-0.098592	1
2	37	0.079796	0.31723	0.23744	0.0023379	0.0081323	0.0057943	0.92641	-2.1527	1.2263	0.78492	-3.7532	2.8859	0
2	38	-0.0042177	0.27193	0.27615	-0.003955	0.0059104	0.0098654	0.52563	-1.2569	0.73126	0.79975	2.7192	1.8863	1
2	39	-0.019388	0.32998	0.34937	0.00081912	0.006111	0.0052919	2.4331	-3.2365	-0.80345	0.80103	-3.2774	-1.311	0
2	40	-0.028367	0.22109	0.24946	-5.82E-06	0.0071465	0.0071523	-1.484	2.6931	4.1771	0.78472	-0.85702	-0.8381	1

Logistic regression



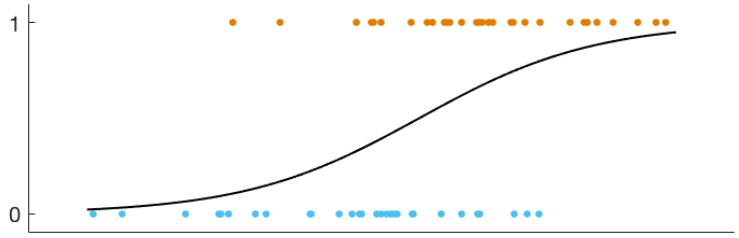
linear regression:

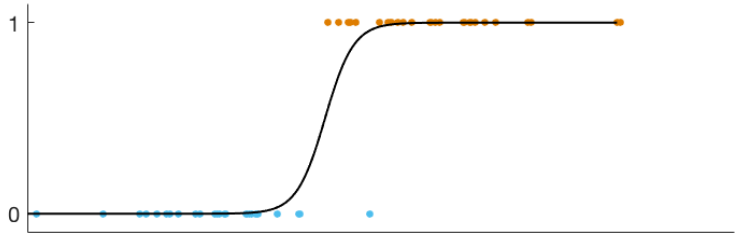
$$\text{resp} \approx a + b\text{x}$$

binomial (logistic) regression:

$$\text{resp} \approx \frac{1}{1 + \text{e}^{-(a + b\text{x})}}$$

Logistic regression





a = -0.43
b = 1.35

<

a = -0.21
b = 8.29

binomial (logistic) regression:

$$\text{resp} \approx \frac{1}{1 + e^{-(a + bx)}}$$

Logistic regression

In our case:

dependent variable: response

independent variables: Δdur, Δf0, Δinterval, Δintensity

our binomial (logistic) regression:

$$\text{resp} \approx \frac{1}{1 + e^{-(a + b_1\Delta\text{dur} + b_2\Delta\text{f0} + b_3\Delta\text{interval} + b_4\Delta\text{intensity})}}$$

let Δdur = 0.1 s, Δinterval = 0, Δintensity = 0

then $0.1b_1 = -b_2\Delta\text{f0}$

$$\Delta\text{f0} = -\frac{0.1b_1}{b_2}$$

difference in f0 corresponding to duration difference of 0.1 s

Experiment 1: Duration judgments

Table 3 Mixed effects model fitted to the responses of *duration discrimination* with frequency range difference calculated as the difference between the absolute values of the dynamic f_o ranges.

Effect	Size	Error	*z* value	p (MCMC)
Intercept	0.47	0.19	2.4	0.016
Duration difference	29	2.8	10	$2 \cdot 10^{-16}$
Intensity difference	0.073	0.018	4.1	$4 \cdot 10^{-5}$
Frequency difference	0.19	0.029	6.8	$1 \cdot 10^{-11}$
Frequency range difference	0.021	0.020	1.0	0.3

Δdur = 10 ms ≈ Δintensity = 4 dB
Δdur = 10 ms ≈ Δf0 = 1.5 st
Δintensity = 1 dB ≈ Δf0 = 0.38 st

Table 1 Mixed effects model fitted to the responses of *intensity discrimination* with frequency range difference calculated as the difference between the absolute values of the dynamic f_o ranges.

Effect	Size	Error	*z* value	p (MCMC)
Intercept	0.14	0.076	1.8	0.07
Duration difference	6.5	0.71	9.2	$2 \cdot 10^{-16}$
Intensity difference	0.34	0.055	6.2	$4 \cdot 10^{-10}$
Frequency difference	0.14	0.036	3.9	$1 \cdot 10^{-4}$
Frequency range difference	0.028	0.020	1.4	0.2

Δdur = 10 ms ≈ Δintensity = 0.2 dB
Δintensity = 1 dB ≈ Δdur = 52 ms
Δintensity = 1 dB ≈ Δf0 = 2.4 st

Q3

Is the effect of fundamental frequency of sound on its perceived duration purely due to the fact that higher sounds are louder (and therefore perceived as longer)?

no!

Dawson, Aalto, Šimko, Vainio (2017). The influence of fundamental frequency on perceived duration in spectrally comparable sounds, *PeerJ*

Experiment 2: Multiple languages

"Which one is longer?" experiment

speakers of four languages: Estonian, Finnish, Mandarin Chinese and Swedish

Finnish: a quantity language (2V, 2C)
no lexical tones

Estonian: a quantity language (3V, 3C)
no lexical tones

} *as Martti told you, long vowels have tonal elements (falling pitch)*

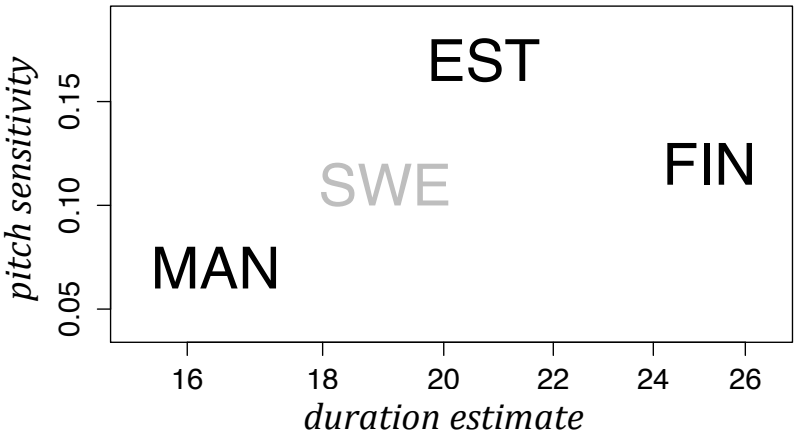
Mandarin: not a quantity language
lexical tones

Swedish: lexical quantity opposition (2V, 2C)
some tonal elements (2 "accents")

Šimko, Aalto, Lippus, Włodarczak, Vainio (2015). Pitch, perceived duration and auditory biases: Comparison among languages, *18th International Congress of Phonetic Sciences*, Glasgow, Scotland

Experiment 2: Multiple languages

	EST	FIN	SWE	MAN
interc.	0.20	0.18	0.25	0.56***
dur. dif.	20.7***	25.2***	19.0***	16.4***
f_0 dif.	0.17***	0.12***	0.11***	0.07***
Δf_0 dif.	0.05***	0.04***	0.03*	0.03***
$\lvert\Delta f_0\rvert$ dif.	-0.01	0.05**	0.02	0.06***
level dif.	0.15***	0.09**	0.09	0.07*



Finnish: a quantity language (2V, 2C)
no lexical tones

Estonian: a quantity language (3V, 3C)
no lexical tones

Mandarin: not a quantity language
lexical tones

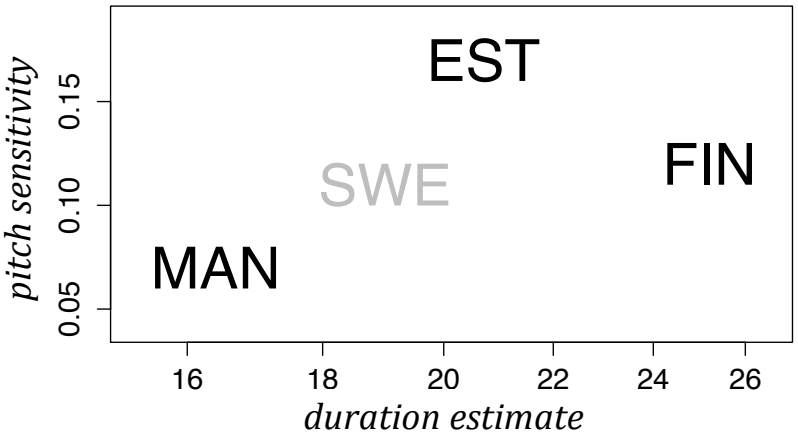
Swedish: lexical quantity opposition (2V, 2C)
some tonal elements (2 “accents”)

} *as Martti told you, long vowels have tonal elements (falling pitch)*

Šimko, Aalto, Lippus, Włodarczak, Vainio (2015). Pitch, perceived duration and auditory biases: Comparison among languages, *18th International Congress of Phonetic Sciences*, Glasgow, Scotland

Experiment 2: Multiple languages

	EST	FIN	SWE	MAN
interc.	0.20	0.18	0.25	0.56***
dur. dif.	20.7***	25.2***	19.0***	16.4***
f_0 dif.	0.17***	0.12***	0.11***	0.07***
Δf_0 dif.	0.05***	0.04***	0.03*	0.03***
$\lvert\Delta f_0\rvert$ dif.	-0.01	0.05**	0.02	0.06***
level dif.	0.15***	0.09**	0.09	0.07*



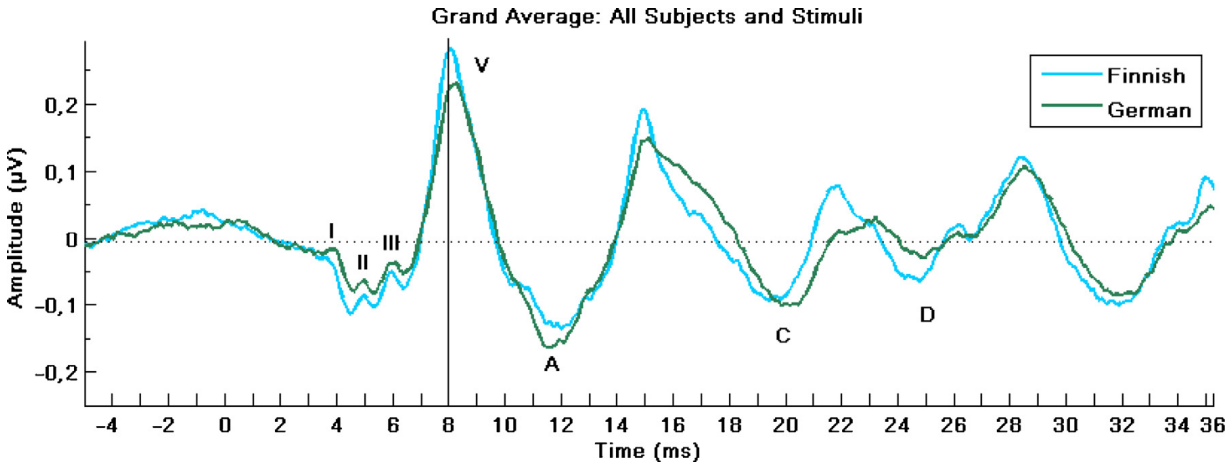
Q2 *Are there any quantitative differences based on native language in terms of the effect of sound fundamental frequency on its perceived duration?*

yes!

Experiment 3: Brain study

Speakers of two languages: 15 Finns and 15 Germans

listening to frog sounds for about 70 minutes each
they brain stem response was recorded by EEG



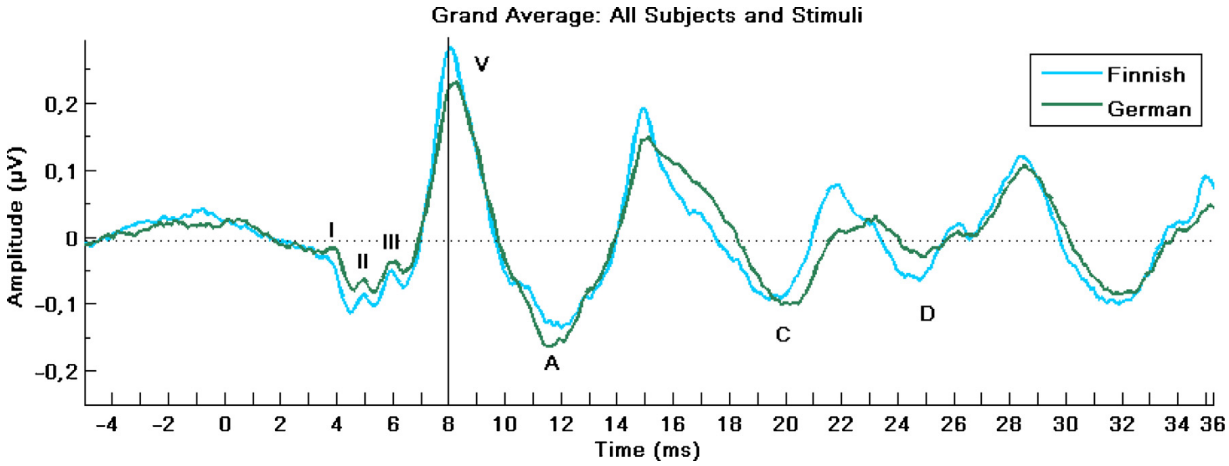
Dawson, Aalto, Šimko, Putkinen, Tervaniemi, Vainio (2016). Quantity language speakers show enhanced subcortical processing, *Biol Psych*

Experiment 3: Brain study

main effect of language on wave V response: a higher degree of precision in alignment of the onset response latencies due to populations of neurons firing in better synchrony for the Finns

Finns have "better", more precise encoding of timing in the inferior colliculus

neurally-based differences in frog sound perception determined by mother-tongue



Dawson, Aalto, Šimko, Putkinen, Tervaniemi, Vainio (2016). Quantity language speakers show enhanced subcortical processing, *Biol Psych*

Q1

Is the fact that the higher sounds are perceived as longer than lower ones (of the same duration) based on our perceptual, auditory apparatus?

it seems so

- the effect works for non-speech sounds
- even for non-speech sounds, it is language sensitive, reflecting non-trivial statistical properties of language, the needs of listeners to more or less precisely judge particular characteristics
- it seems to be neurally encoded, in very early stages of auditory processing (brain plasticity)
- in fact, the two hypotheses, the perceptual compensation and the production compensation don't seem to be so mutually exclusive anymore
- perhaps, properties of auditory apparatus and articulatory characteristics continuously reinforce each other during speech evolution
- in other words, evolution of speech (language?) seems to be in an important way determined by the properties of both auditory and articulatory apparatuses

More experiments and results

- Musical Finns

  Dawson, Aalto, Šimko, Vainio, Tervaniemi (2017). Musical Sophistication and the Effect of Complexity on Auditory Discrimination in Finnish Speakers, *Front. Neurosci*

- Which (frog) sound stands out (3 sounds)

  Tiia Ojala's Master thesis

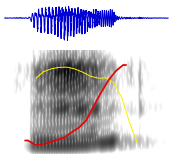
- Beyond frog sounds

  presently ongoing experiments "Experimental Phonetics" course

- and some more

  Aalto, Šimko, Vainio (2013). Pitch affects the duration judgments of non-speech sounds more for quantity-language speakers, *Front. Interspeech 2013*

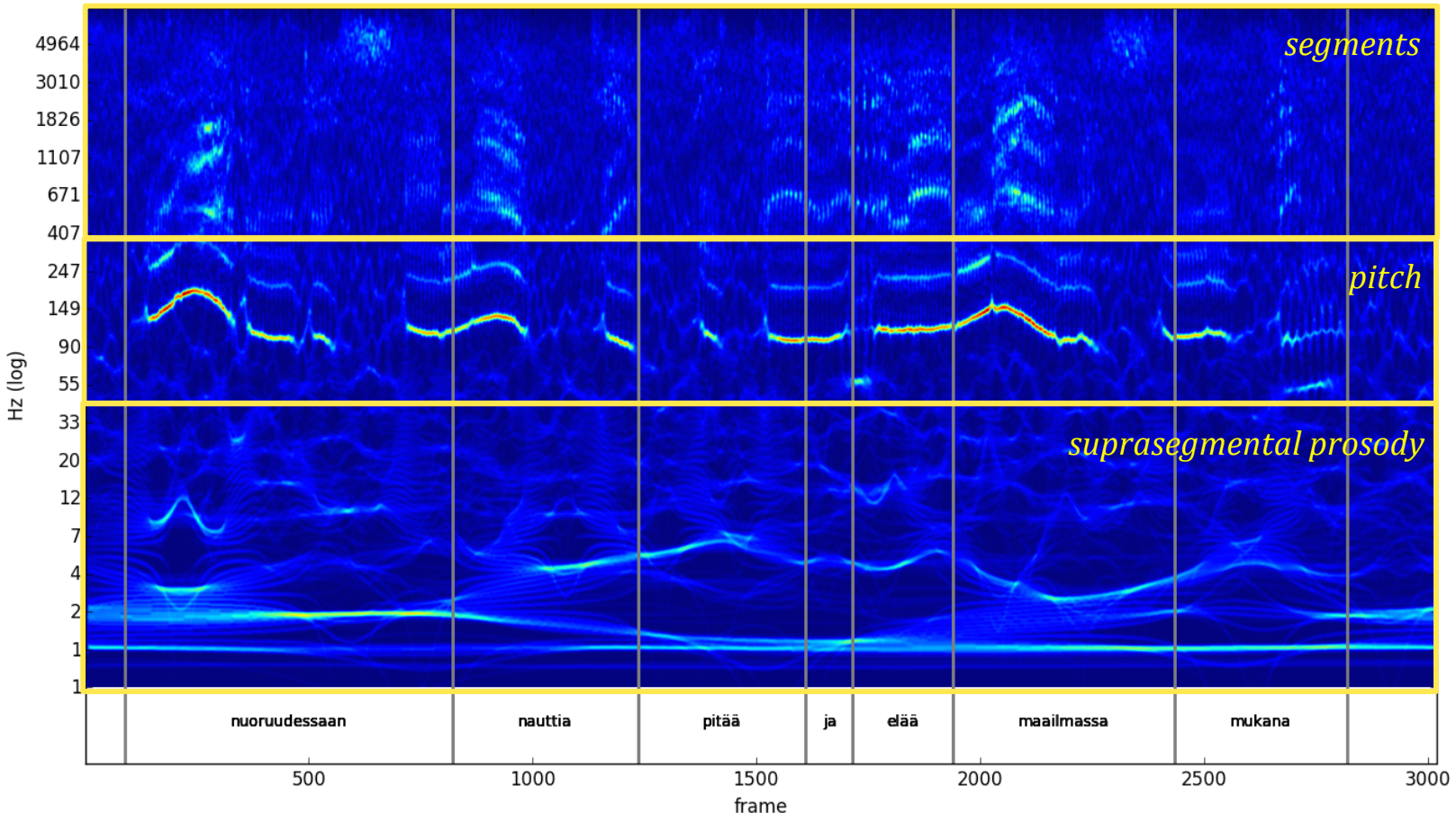
Back to prominence



- our auditory apparatus gets “confused” by different sound characteristics (duration, pitch, intensity)
- to some extent we can use them in a complementary fashion, replace one with another when the other is used for other purpose (Swedish accent)
- we can perhaps also use these trade-offs to make *chunks* not stand out (falling pitch in long Finnish, Estonian vowels)

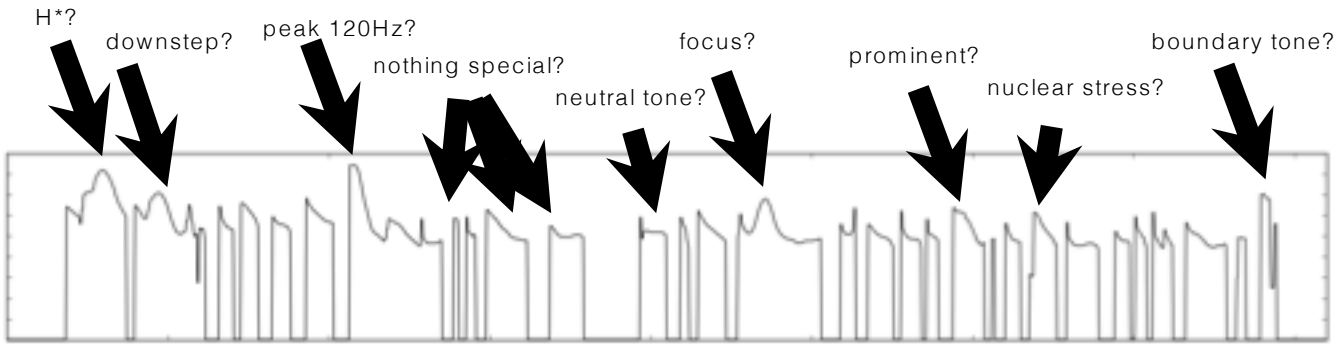
- in any case, understanding – and quantifying! – these trade-offs will help us to understand prominence and prosody. And quantify it: speech synthesis

Hierarchical scale—space analysis using the Continuous Wavelet Transform



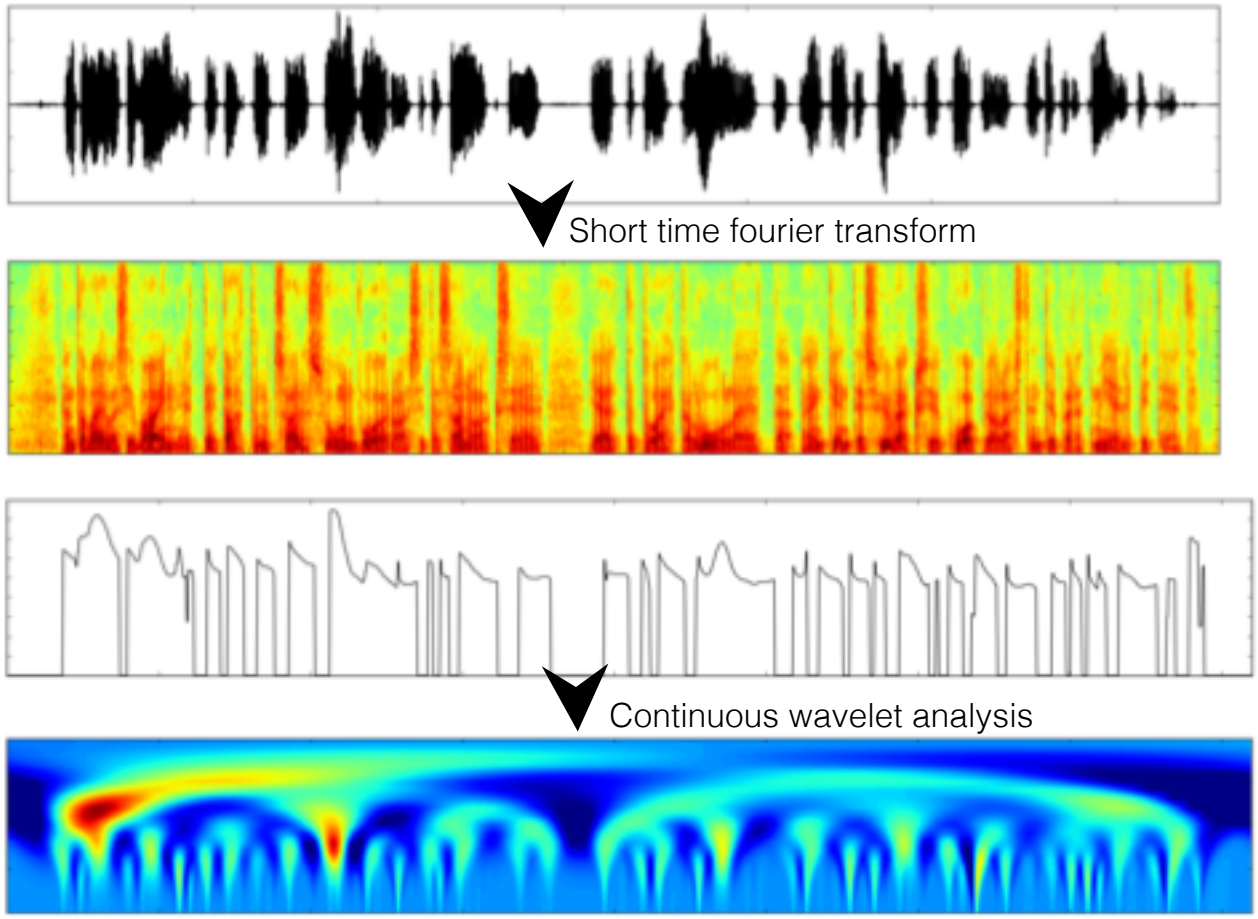
Introduction

- Prosodic signals, like *f0*, 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



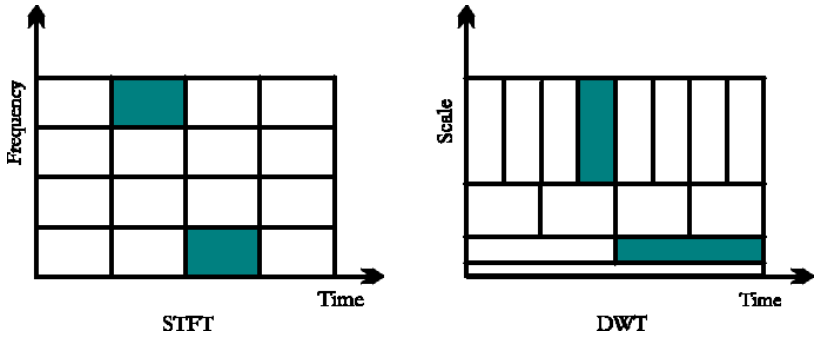
Introduction

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



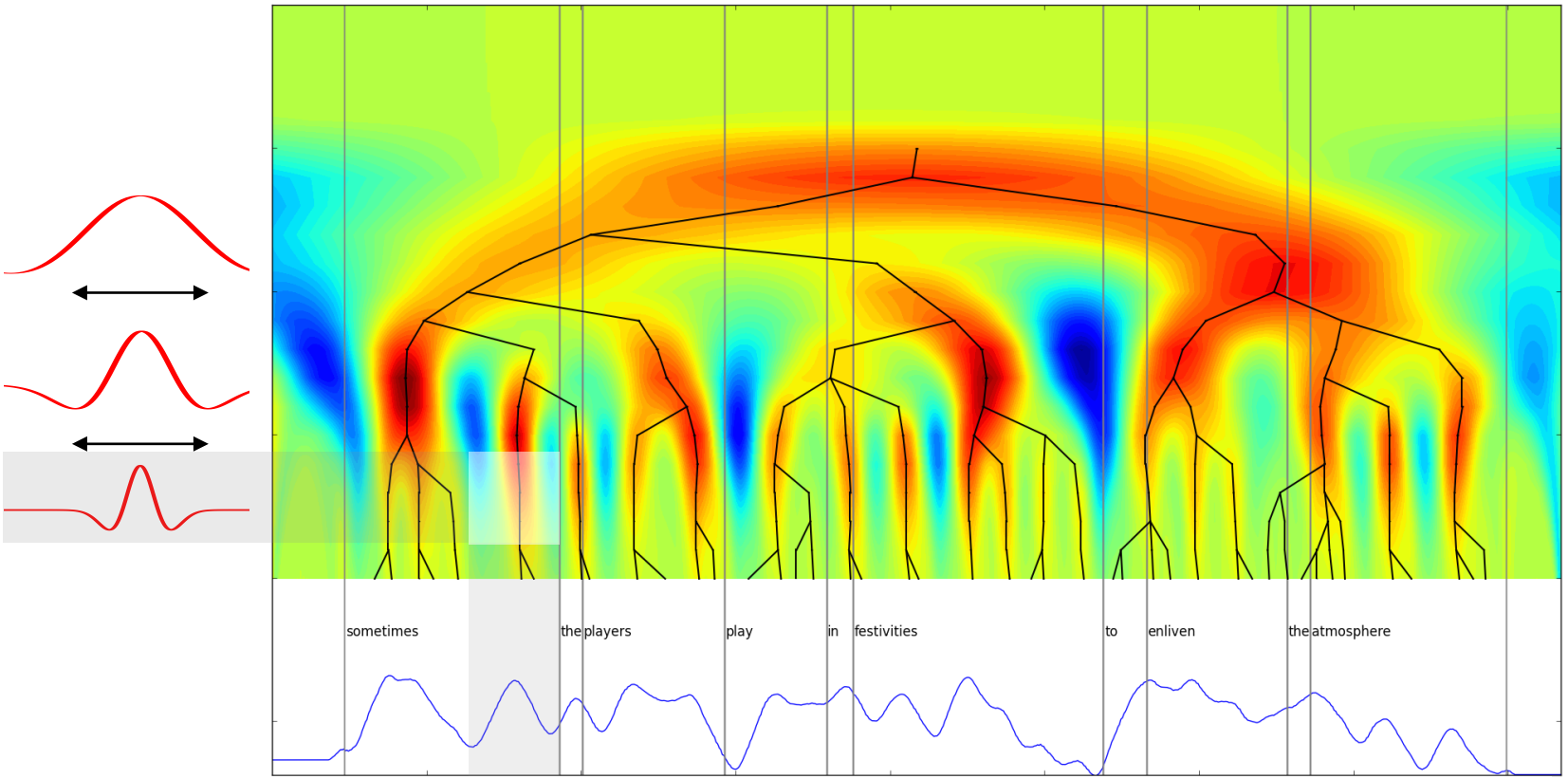
Continuous Wavelet Transform (CWT)

- Two dimensional time-scale representation of a one dimensional signal, similar to Short Time Fourier Transform
- Frequency-adaptive resolution - better time resolution in high frequencies and better frequency resolution in low frequencies



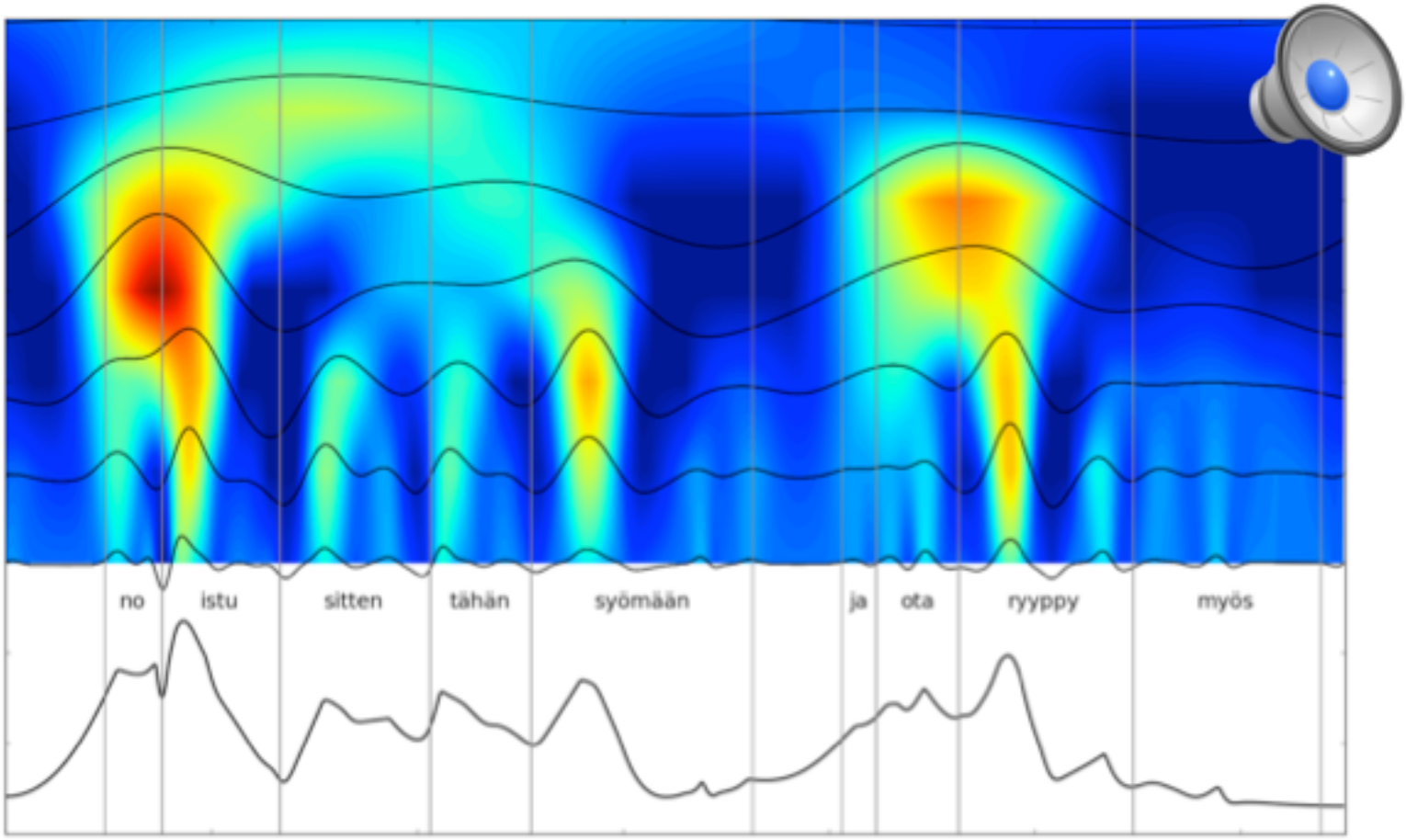
Decomposition: original signal is the sum of the components

Continuous Wavelet Transform (CWT)

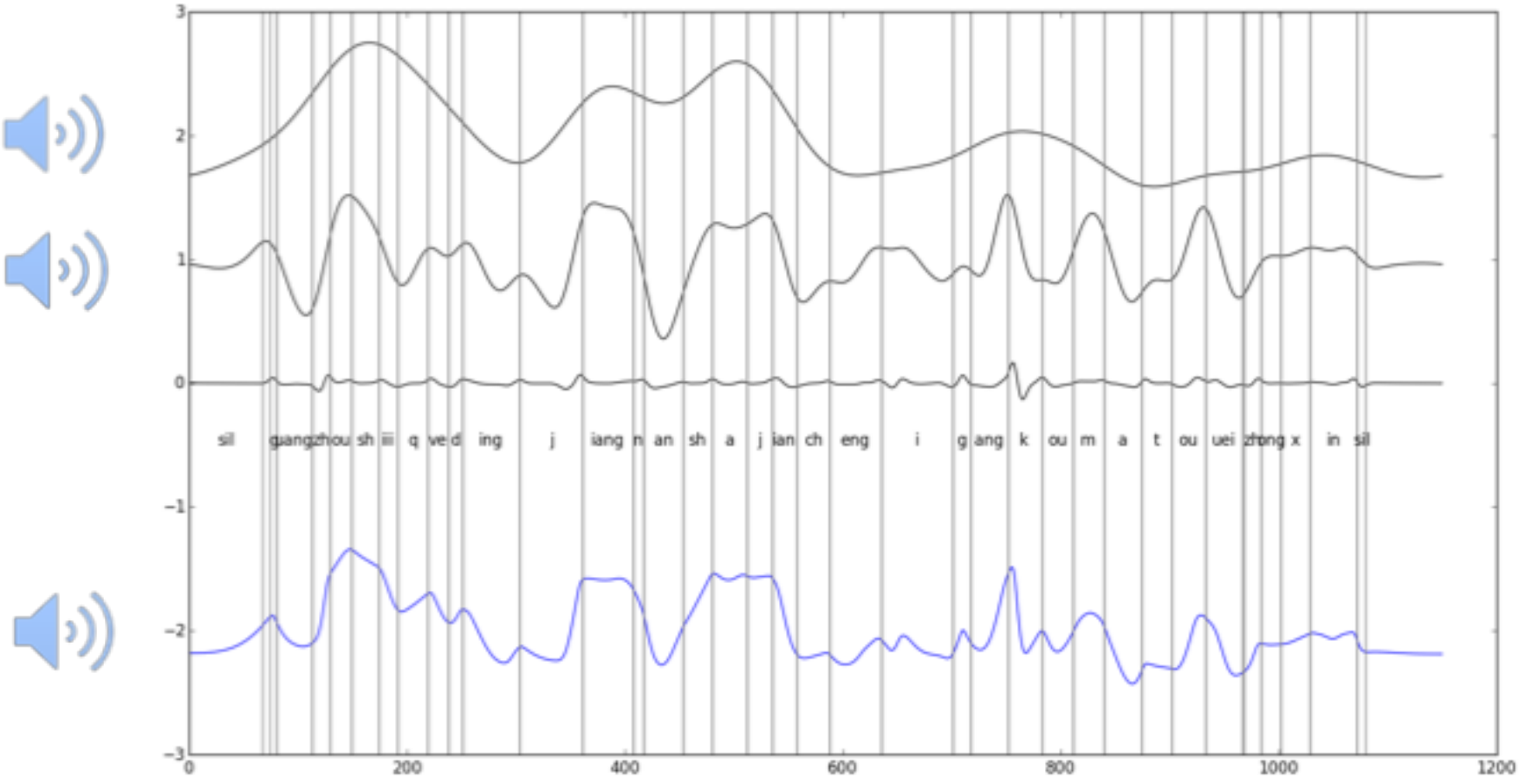


Continuous wavelet analysis - example

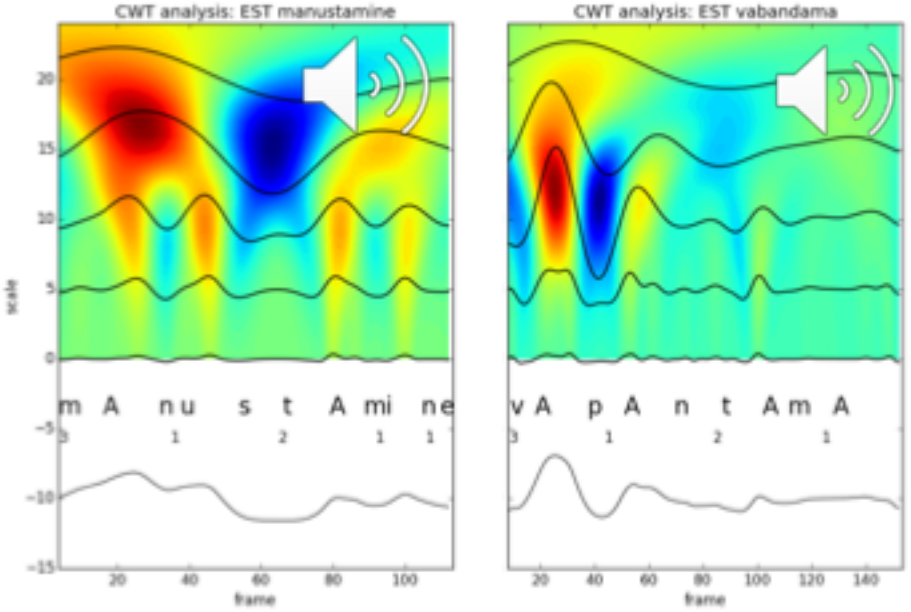
- Scales stacked and colored (peaks red, valleys blue): scalogram



Prosodic hierarchy: The case of lexical tone

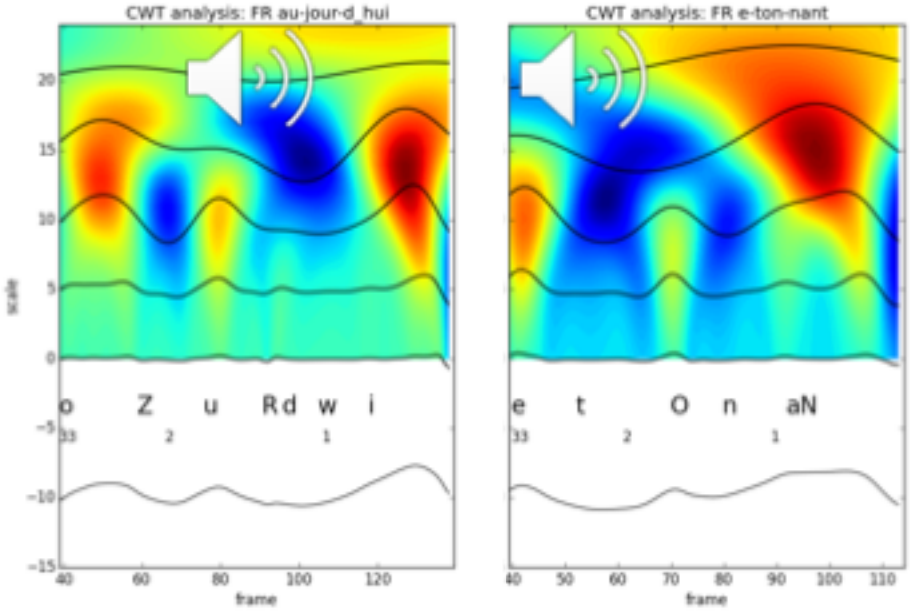


Prominence detection: lexical stress



Estonian: word initial

French: word final



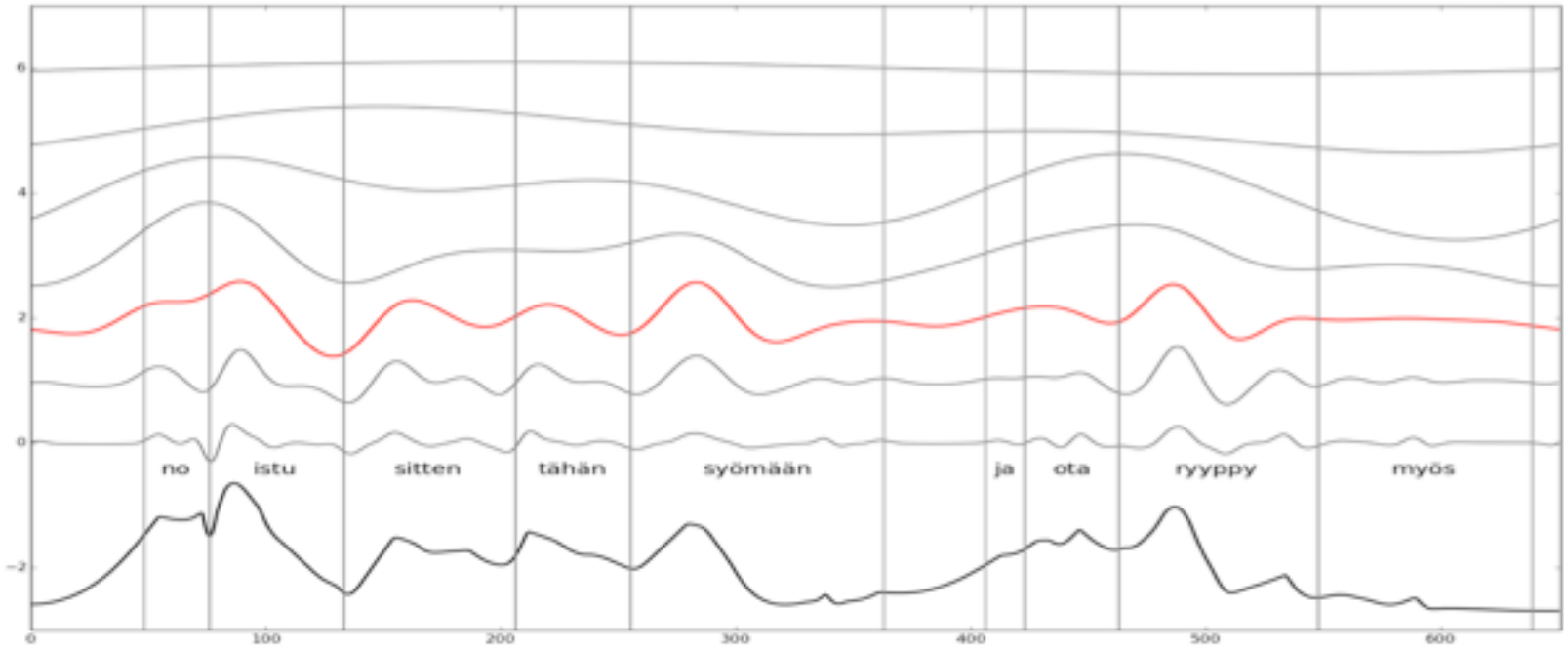
Eriksson, A., Suni, A., Vainio, M., & Šimko, J. (2018). The acoustic basis of lexical stress perception. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 70-74).

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

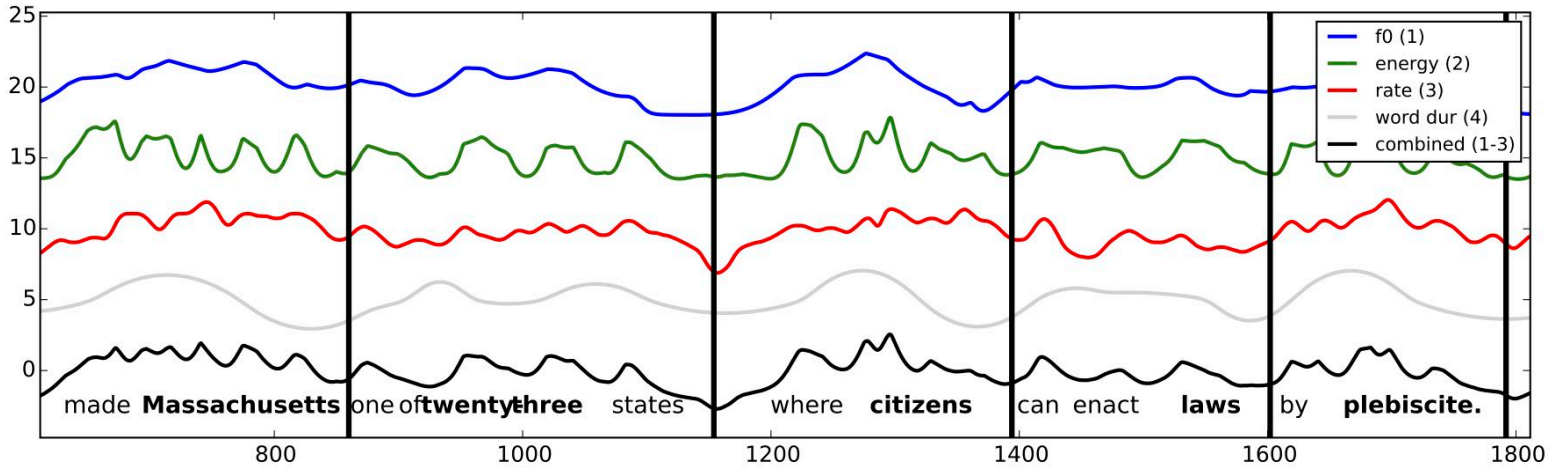
3. Prominence = the maximum peak of the word



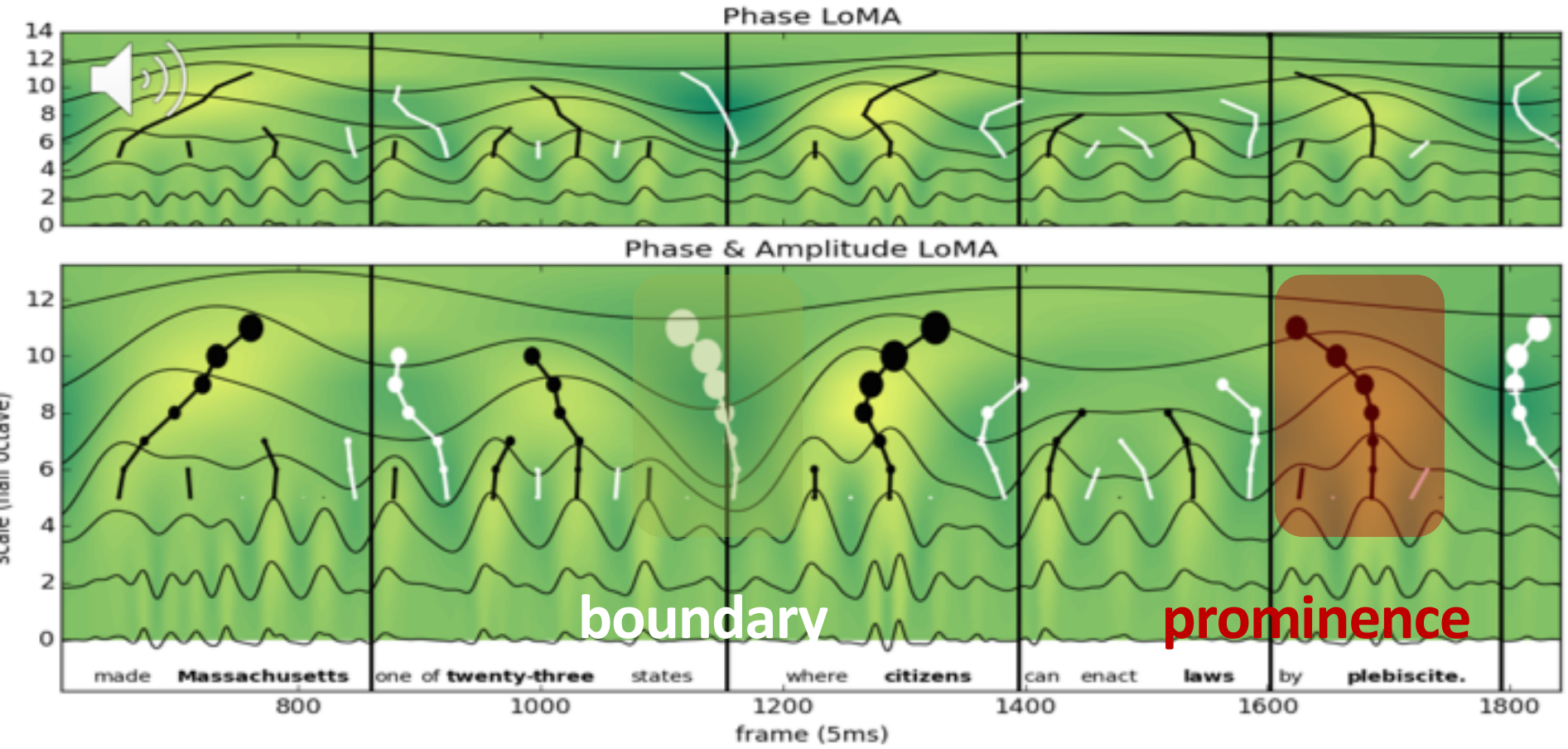
Different signals

- other speech signals cab be processed by CWT:
  - (interpolated) $f0$
  - (interpolated) energy (perhaps obtained via CWT)
  - duration signal (interpolated durations)
  - speaking rate (obtained via CWT)

- or even a combination thereof…

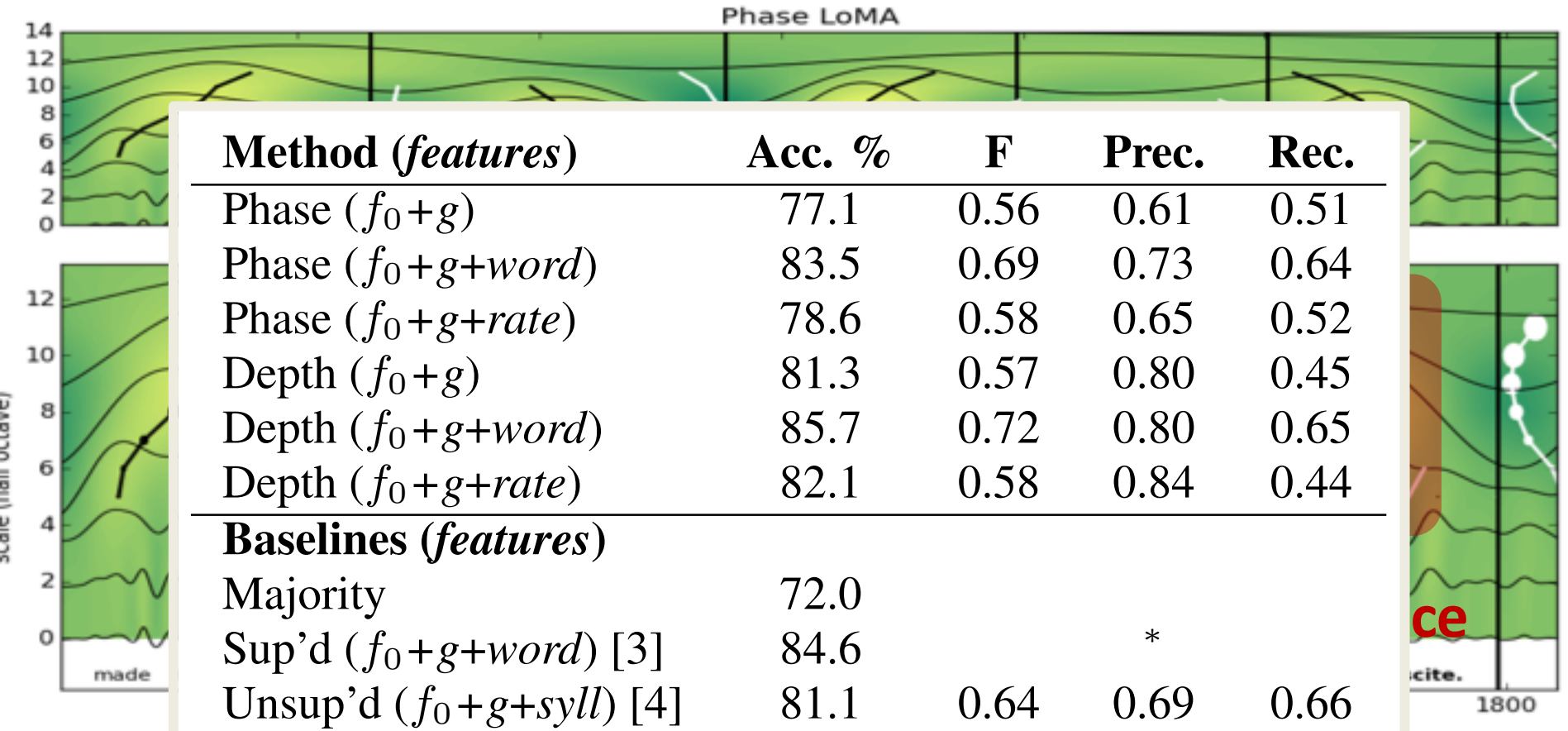


Combining prominence and boundary detection



- 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

Combining prominence and boundary detection

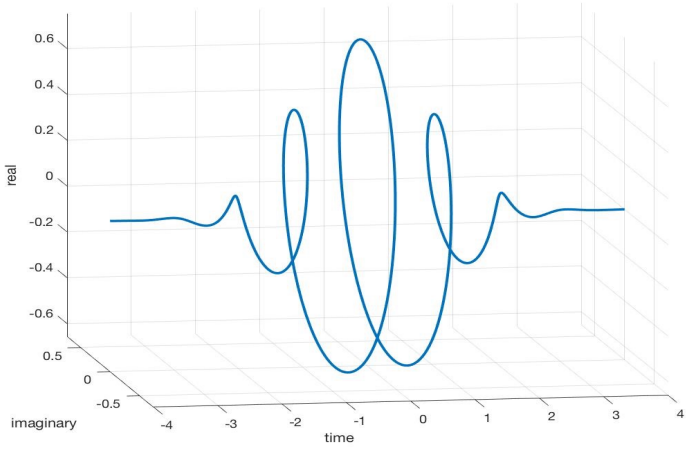


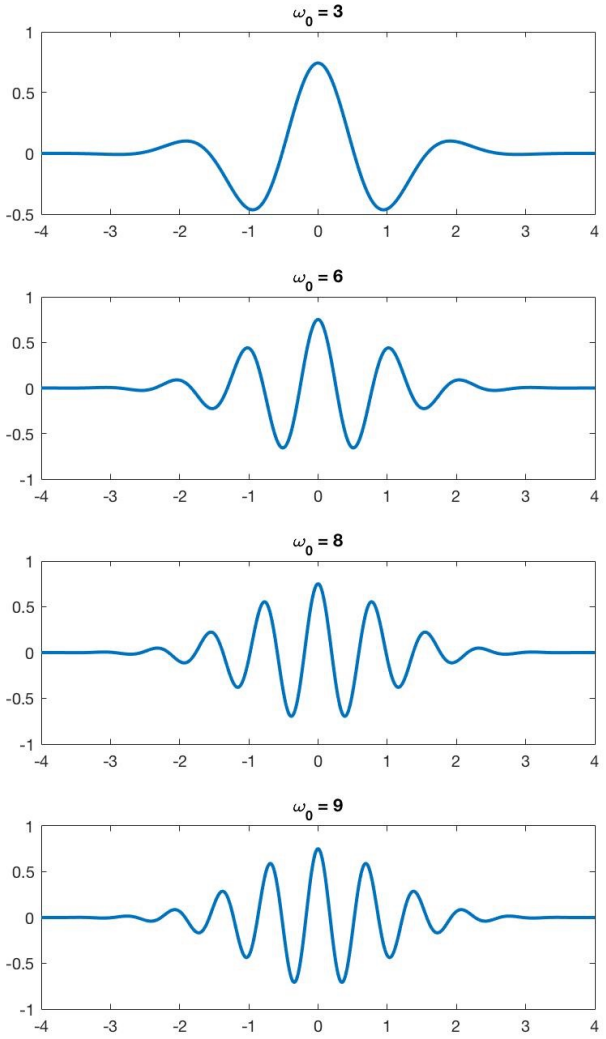
Method (*features*)	**Acc. %**	**F**	**Prec.**	**Rec.**
Phase (f_0+g)	77.1	0.56	0.61	0.51
Phase (f_0+g+*word*)	83.5	0.69	0.73	0.64
Phase (f_0+g+*rate*)	78.6	0.58	0.65	0.52
Depth (f_0+g)	81.3	0.57	0.80	0.45
Depth (f_0+g+*word*)	85.7	0.72	0.80	0.65
Depth (f_0+g+*rate*)	82.1	0.58	0.84	0.44
Baselines (*features*)				
Majority	72.0			
Sup'd (f_0+g+*word*) [3]	84.6		*	
Unsup'd (f_0+g+*syll*) [4]	81.1	0.64	0.69	0.66

- Suni, Simko, 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

Morlet mother wavelet

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

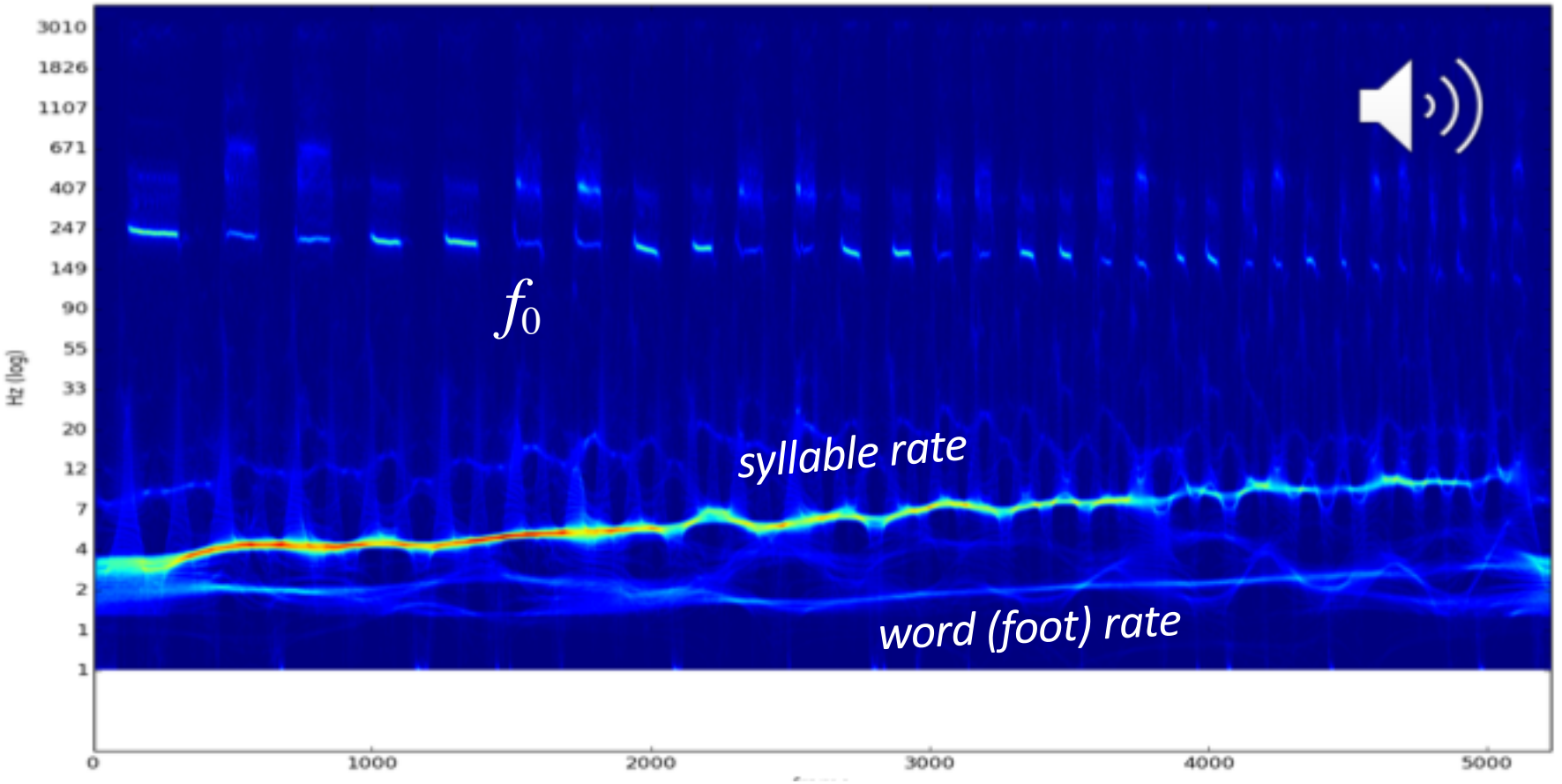




good temporal resolution

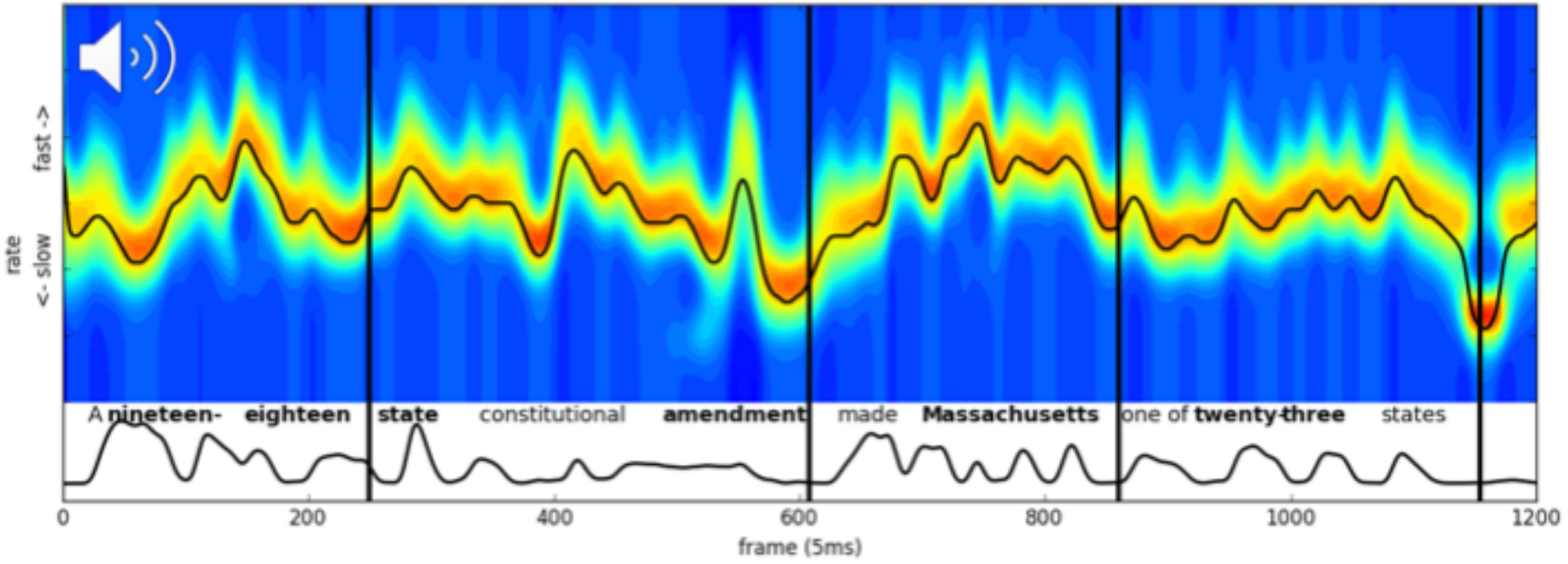
good frequency resolution

f_0 and speaking rate(s)



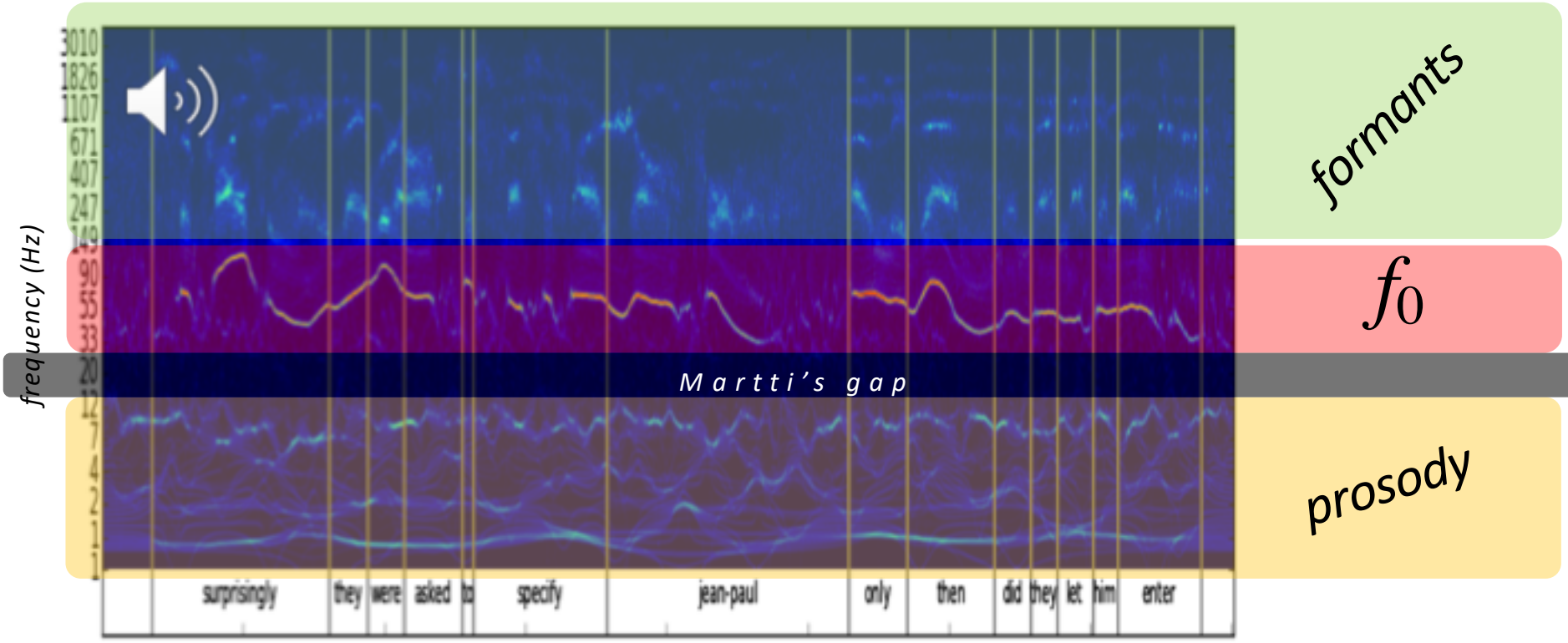
Speaking rate

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



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)



More stuff…

- Speech synthesis parameters

  Suni, Aalto, Raitio, Alku & Vainio. (2013). Wavelets for intonation modeling in HMM speech synthesis, In: *Proc. SSW8*, Barcelona

- f_0, intensity and breathing

  Šimko, Włodarczak, Suni, Heldner & Vainio. (2016). Coordination between f0, intensity and breathing signals. In *Proc. Nordic Prosody*, Trondheim

- Lexical stress

  Eriksson, A., Suni, A., Vainio, M., & Šimko, J. (2018). The acoustic basis of lexical stress perception. In Proc. 9th International Conference on Speech Prosody 2018 (pp. 70-74).

- Adaptation of intonation in Lombard speech

  Šimko, Suni & Vainio. (2016). Wavelet-based adaptation of pitch contour to Lombard speech, In *Proc. Speech Prosody*, Boston

- Digital Language Typology

  Šimko, Suni, Hiovain & Vainio. (2017). Comparing languages using hierarchical prosodic analysis , In *Proc. Interspeech 2017*, Stockholm

  Hiovain, K., Suni, A., Vainio, M., & Šimko, J. (2018). Mapping areal variation and majority language influence in North Sámi using hierarchical prosodic analysis. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 577-581).

  Włodarczak, M., Šimko, J., Suni, A., Vainio, M. (2018) Classification of Swedish dialects using a hierarchical prosodic analysis. *Proc. 9th International Conference on Speech Prosody 2018*, 304-308, DOI: 10.21437/SpeechProsody.2018-62.