

Physiologically plausible prosody modelling

Phil Garner

Idiap Research Institute

Contents

- Part 1: Some background
 - Applications
- Part 2: Prosody models
 - And the one we used
 - And how to apply it
- Part 3:
 - Emphasis transfer & synthesis

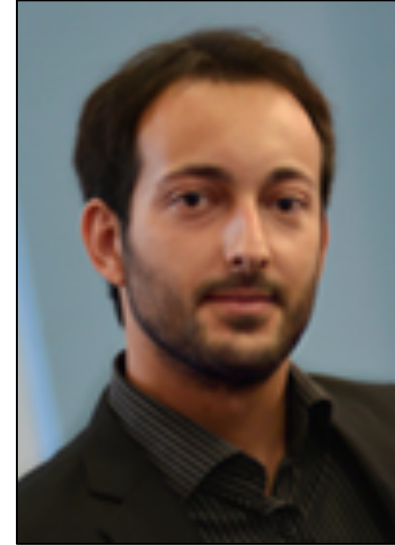
Acknowledgements



Pierre-Edouard Honnet
Idiap Research Institute, Switzerland



Branislav Gerazov, Aleksandar Gjoreski
St.s Cyril & Methodius Univ., Skopje, R. Macedonia



...also P-E's roommate Mélanie

Isn't this a hearing conference?

During the evolution of human speech, the articulatory motor system has presumably structured its output to match those rhythms the auditory system can best apprehend. Similarly, the auditory system has likely become tuned to the complex acoustic signal produced by combined jaw and articulator rhythmic movements. Both auditory and motor systems must, furthermore, build on the existing biophysical constraints provided by the neuronal infrastructure.

– Giraud and Poeppel (2012)
attributed to Heimbauer et al. (2011) and Liberman and Mattingley (1985)

Part 1

Application background

ITCHY FEET in Switzerland





© 2014 - Malachi Rempen

www.itchyfeetcomic.com

Text to speech synthesis

Festival Text-to-Speech Online Demo - Technical

Select a Voice  Type the text to synthesise (max 70 chars)

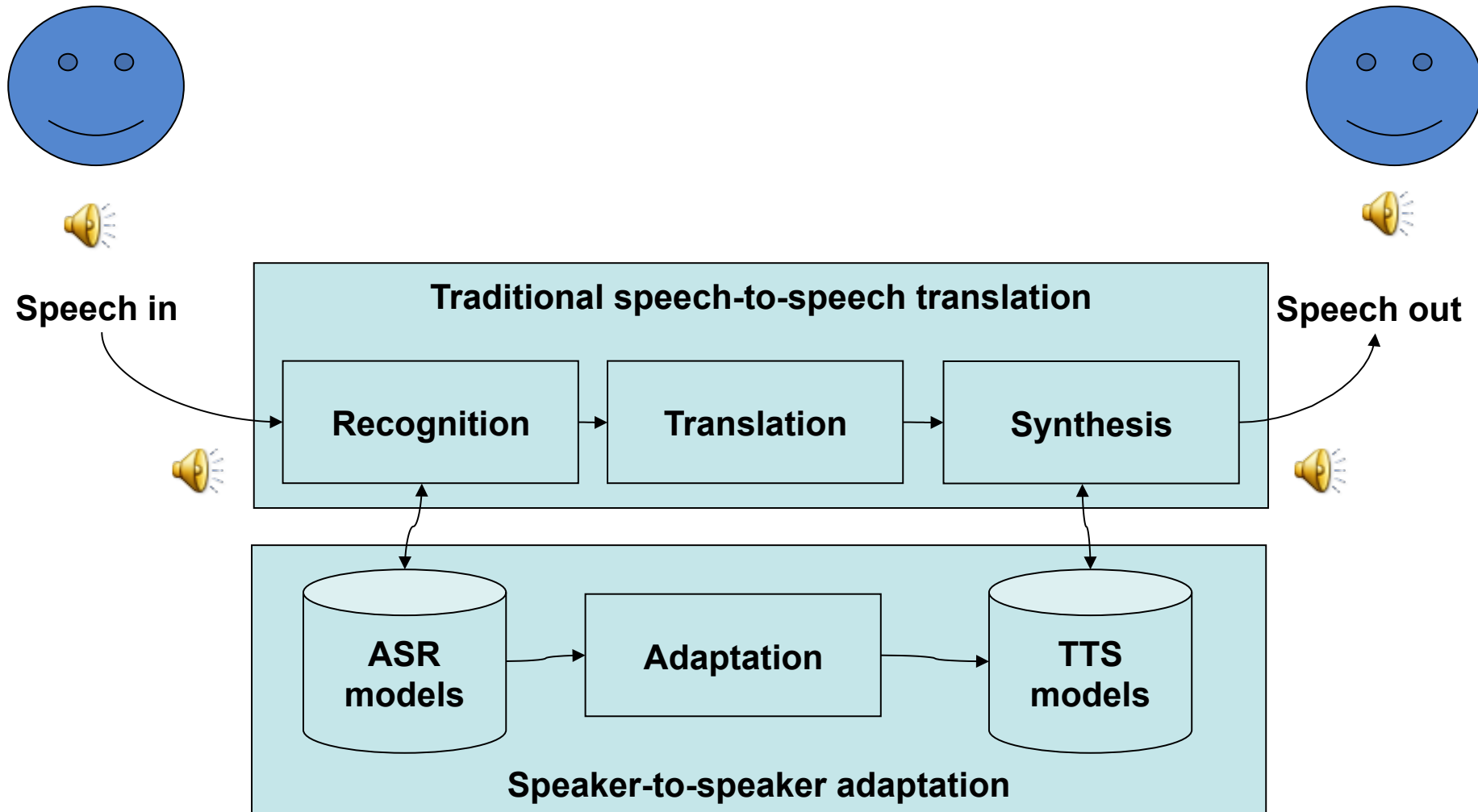
Nina (English RP female) 

- It works!
 - Basic TTS *in its raw form* is a solved problem
 - <http://www.cstr.ed.ac.uk/projects/festival/>

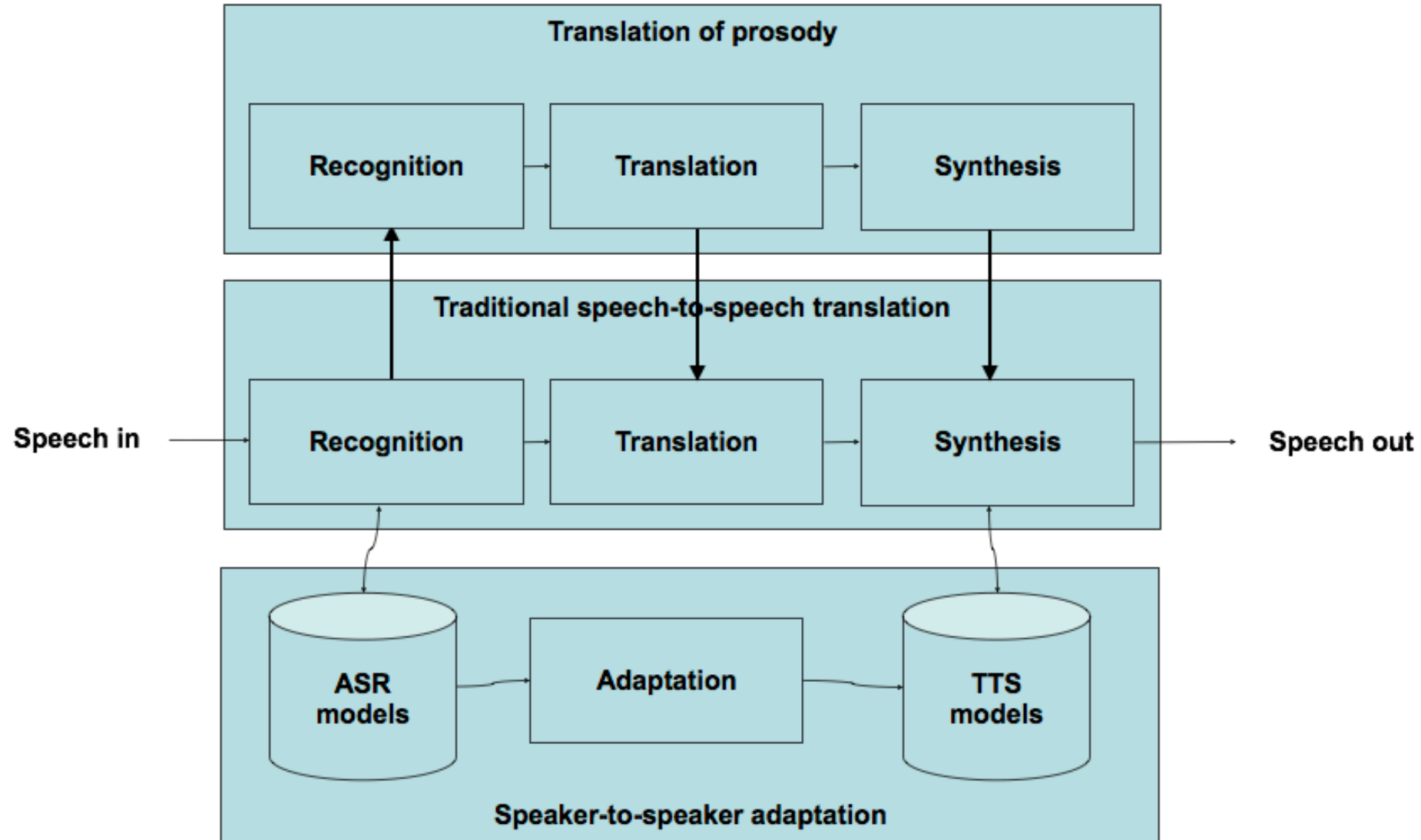
Speech to speech translation



The EMIME scenario



The SIWIS scenario



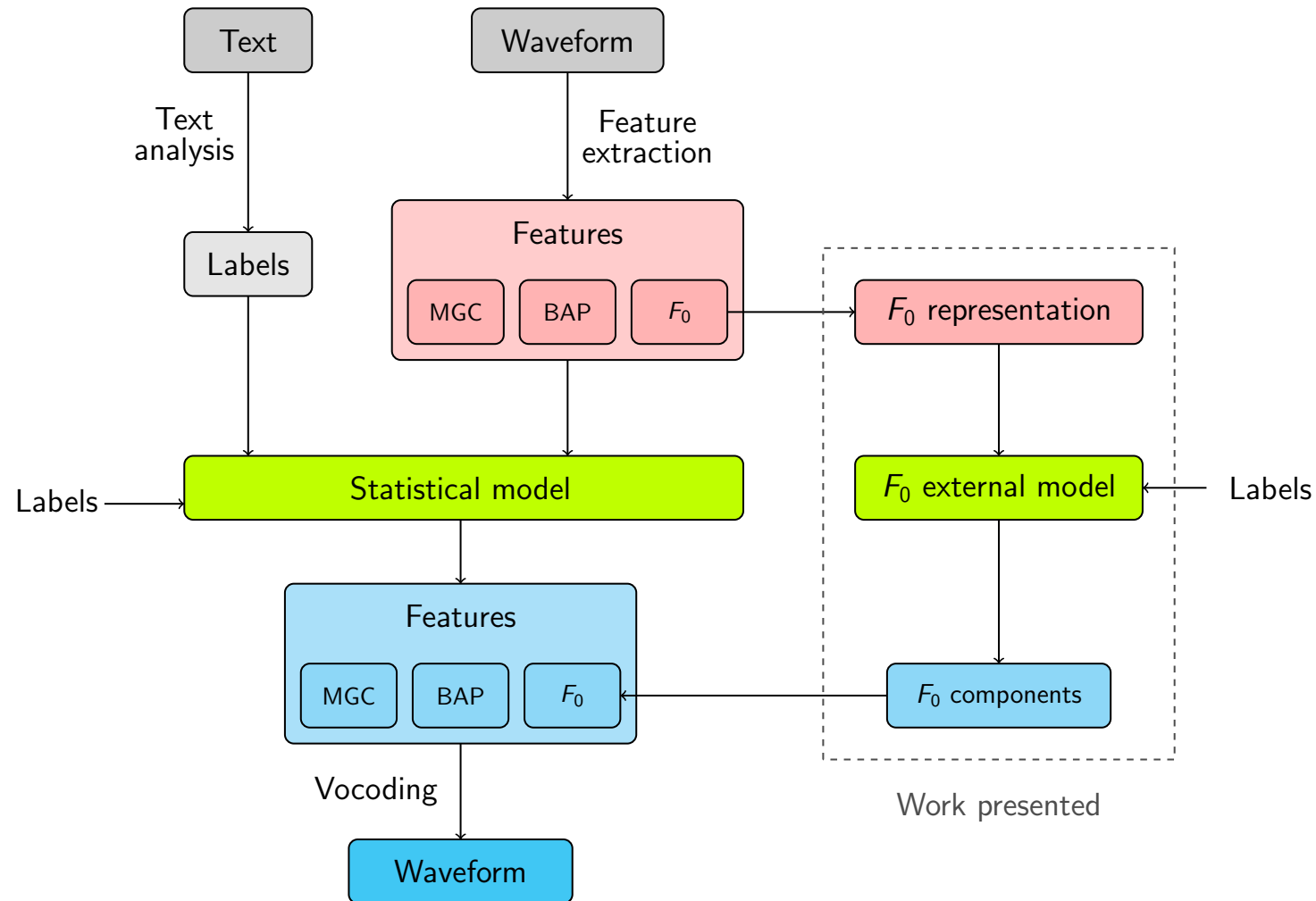
Part 2

Prosody models

Prosody

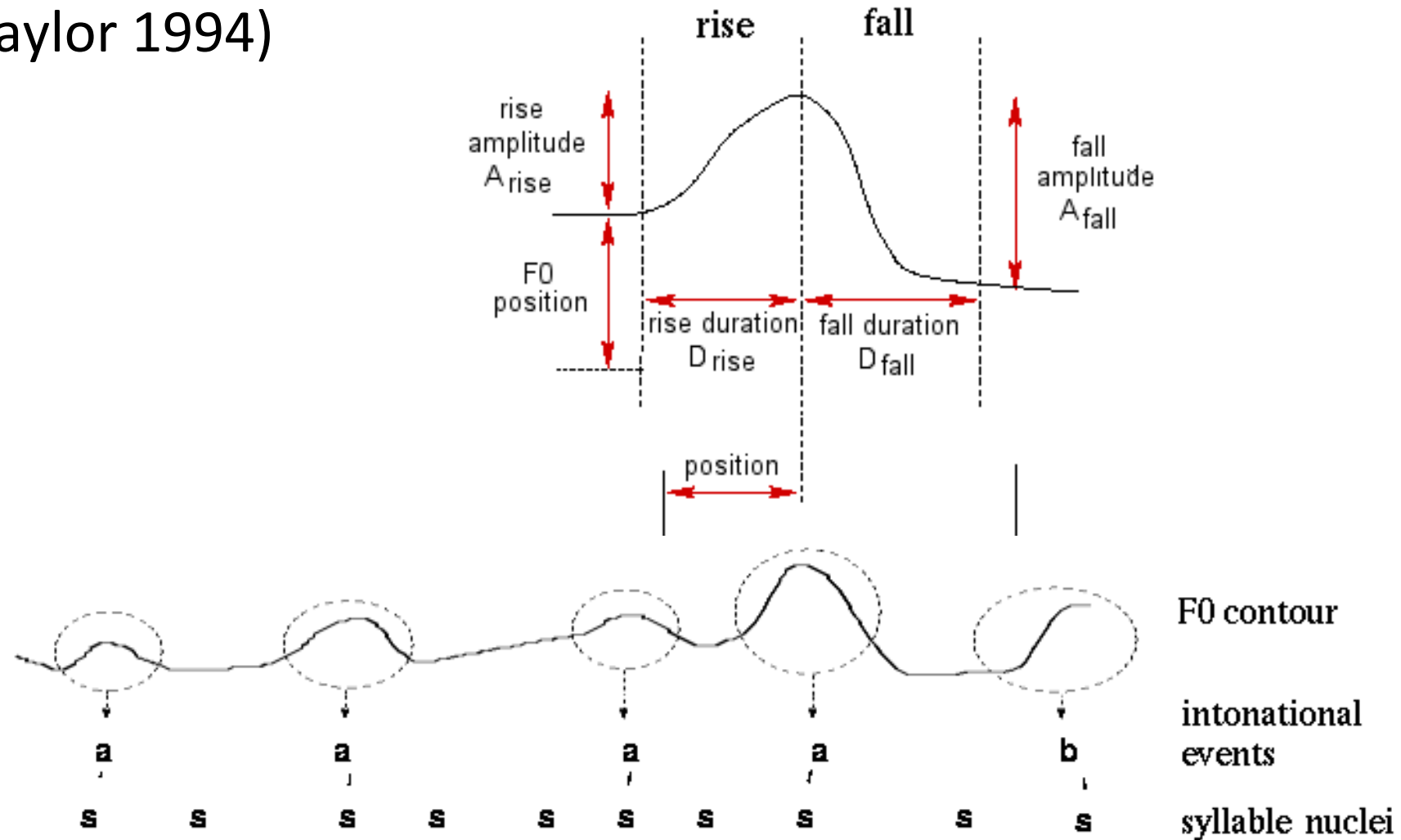
- Prosody is generally taken to have three components
 - Pitch (intonation)
 - Energy
 - Duration
- Pitch and energy are quite closely correlated
- In this work, we focus on *pitch*

More detailed (student PoV)



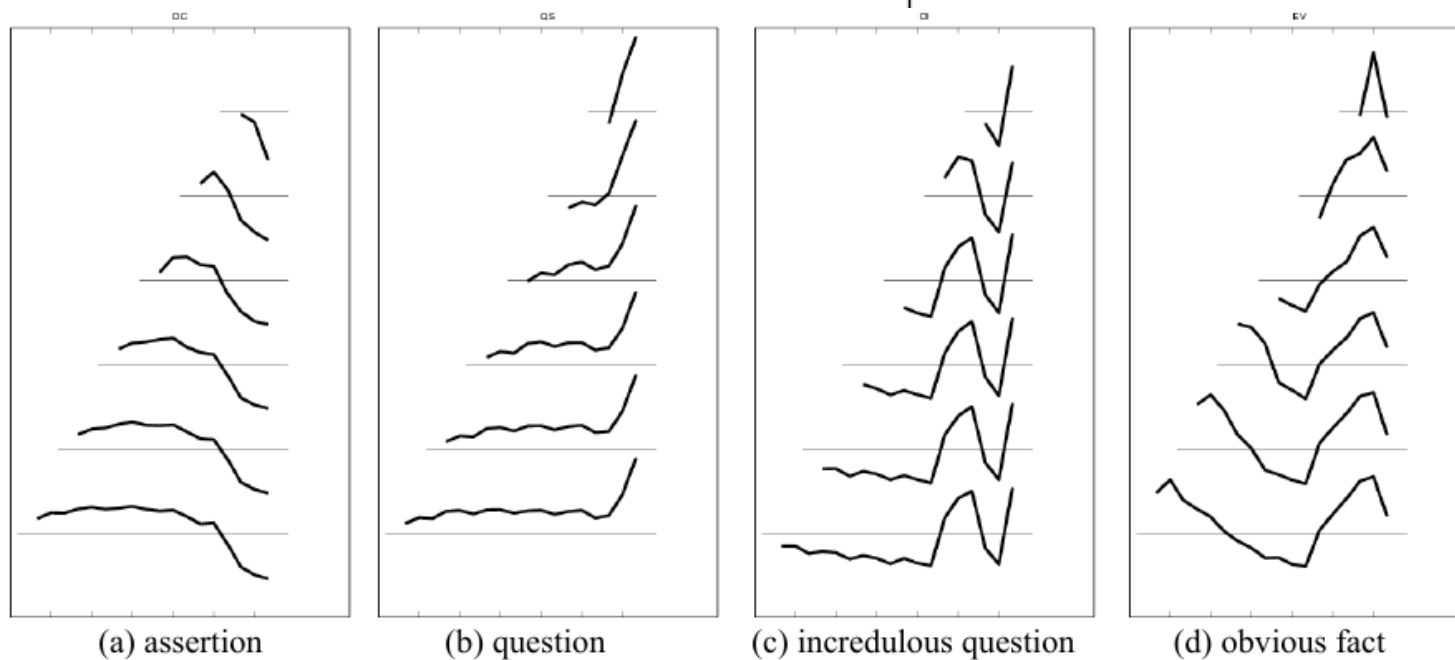
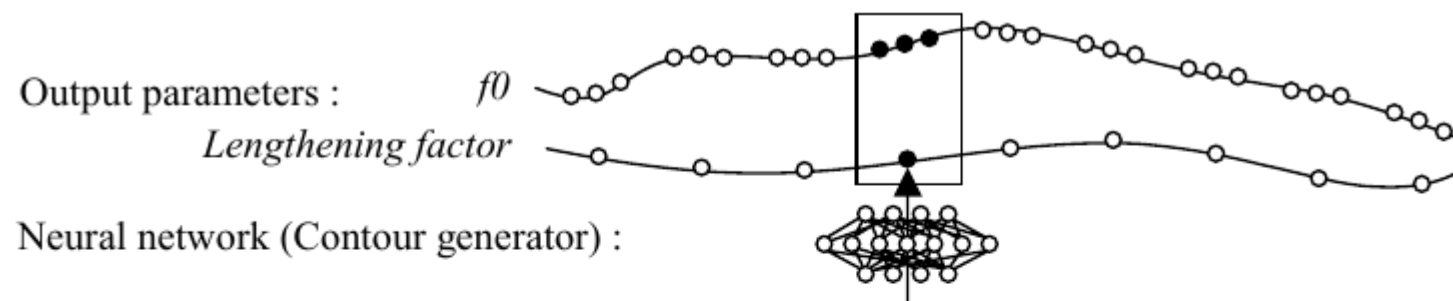
Intonation models

- Tilt (Taylor 1994)



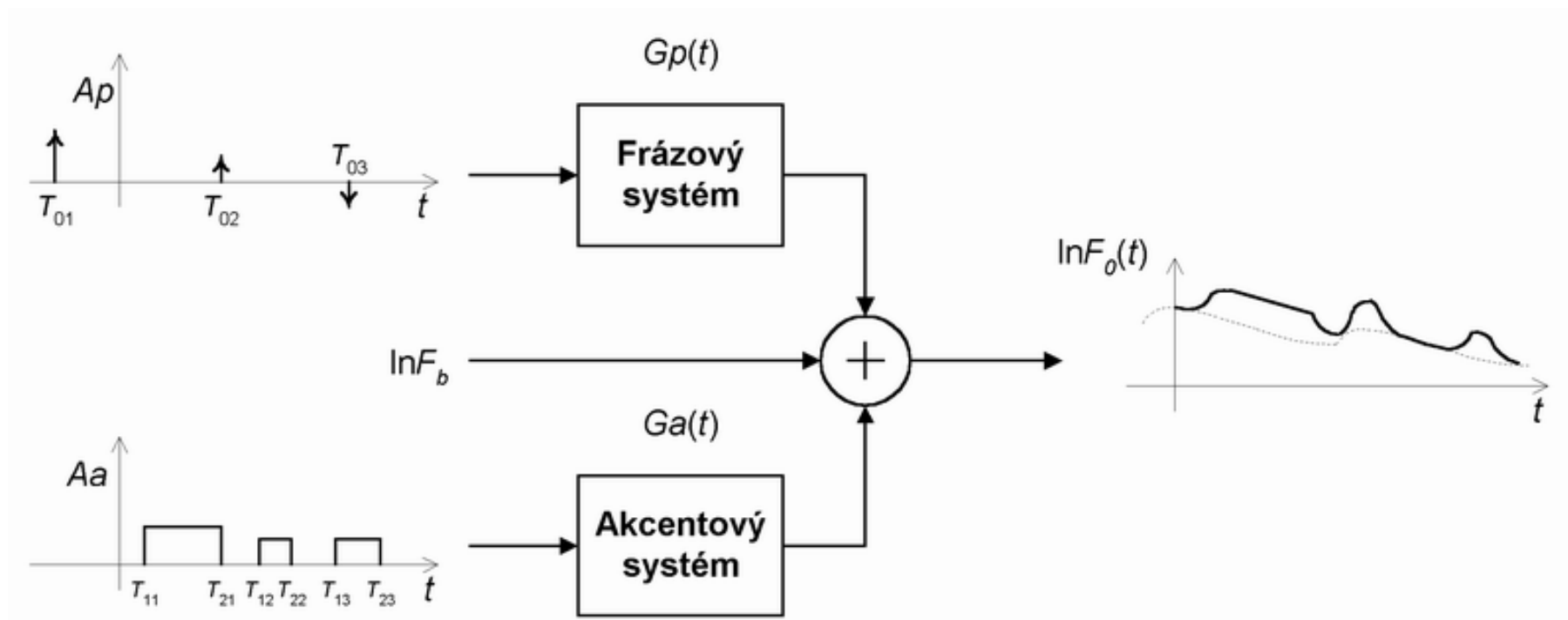
Intonation models

- SFC (Bailly & Holm)



Intonation models

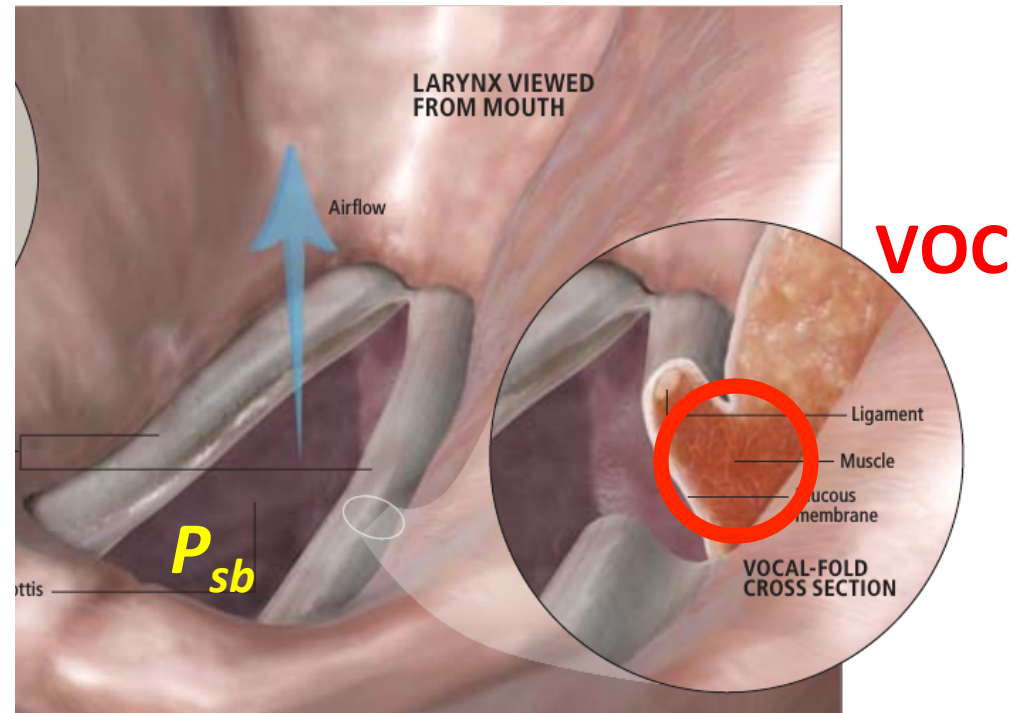
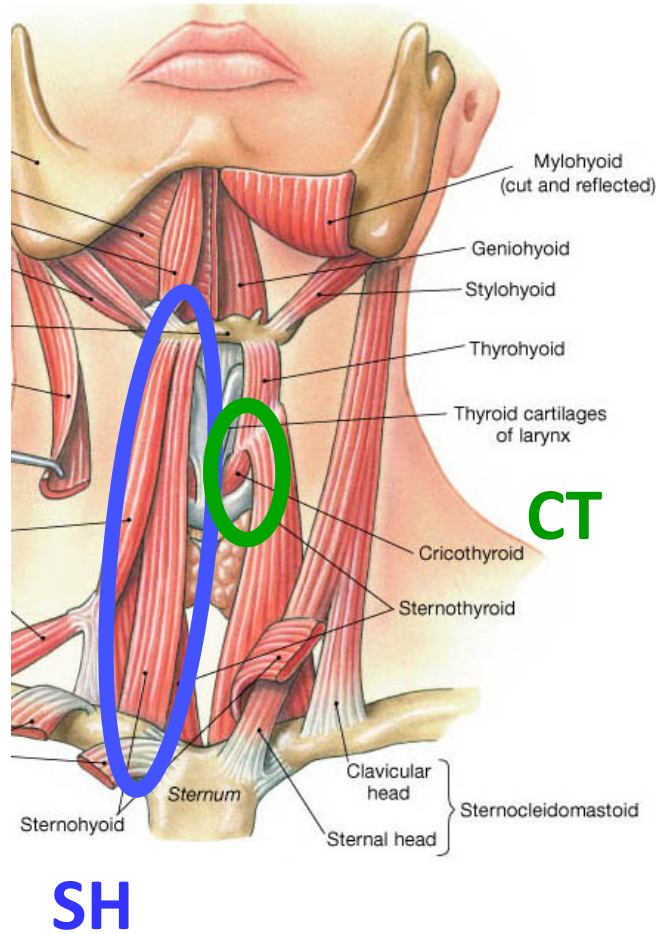
- Command response (Fujisaki)



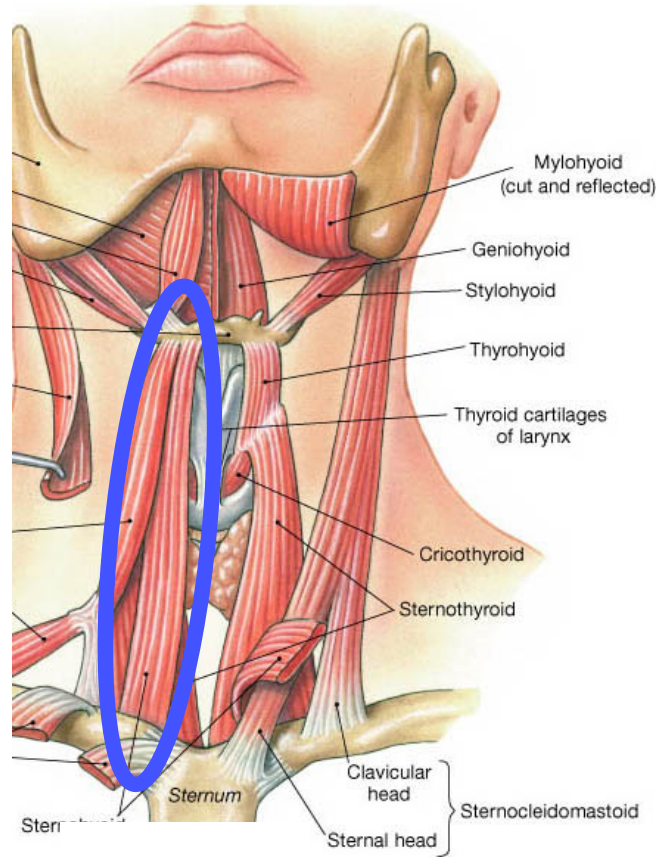
Physiology of intonation

- **Four physiological sources of pitch change** were identified by Strik (PhD, 1994) using elaborate measurements including electromyographic (EMG) recordings of different laryngeal muscles.
 - **Cricothyroid (CT)** muscle
 - rotates thyroid, stretches vocal cords, raising F_0
 - **Vocalis (VOC)** muscle
 - found within the vocal folds, decreases cord length, but increases tensile stress, net effect is rise in F_0
 - **Sternohyoid (SH)** muscle
 - one of three strap muscles, lowers larynx, decreasing vocal fold tension and F_0 .
 - **Subglottal pressure (P_{sb})**
 - linearly correlates to F_0 .

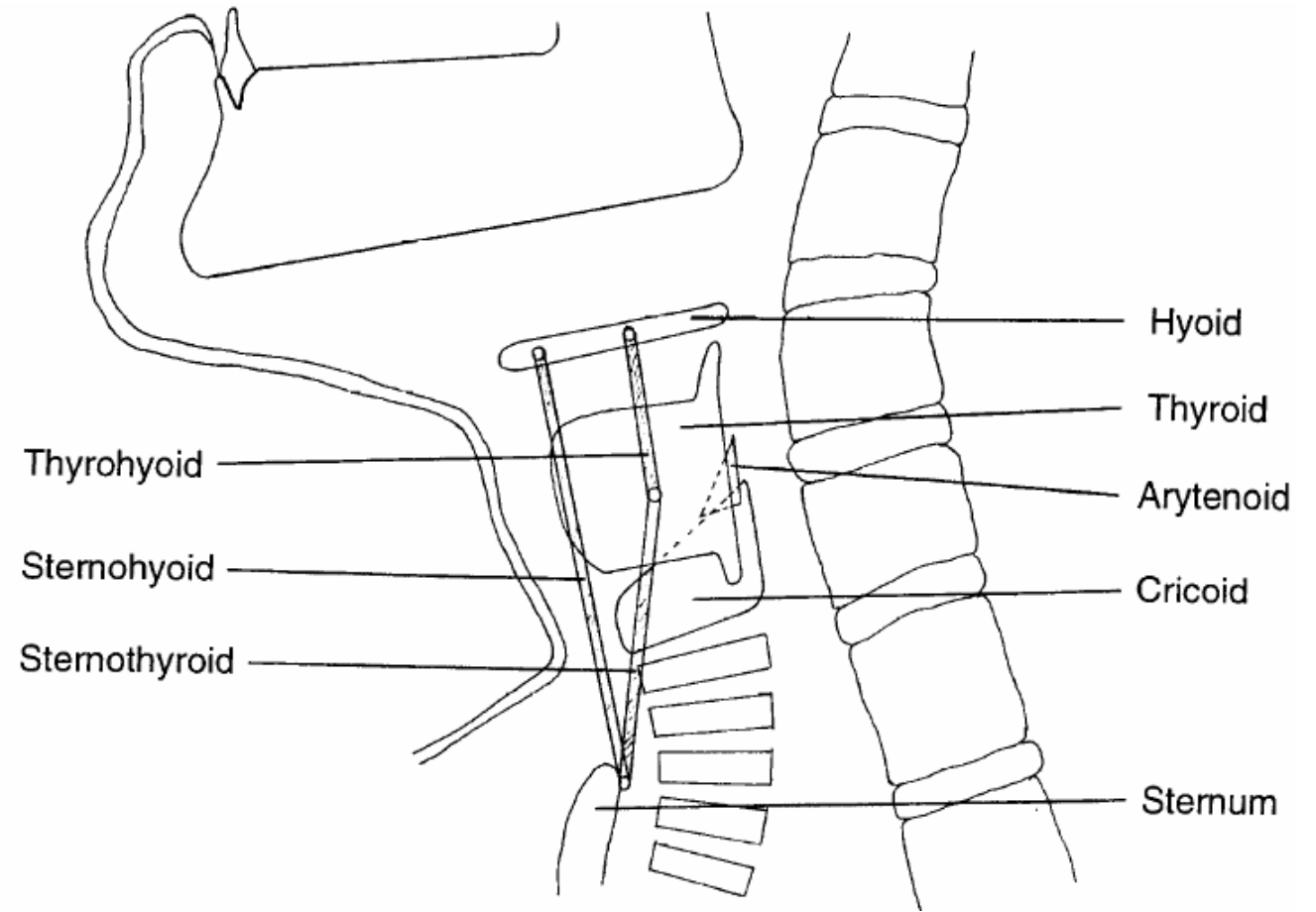
Physiology of intonation



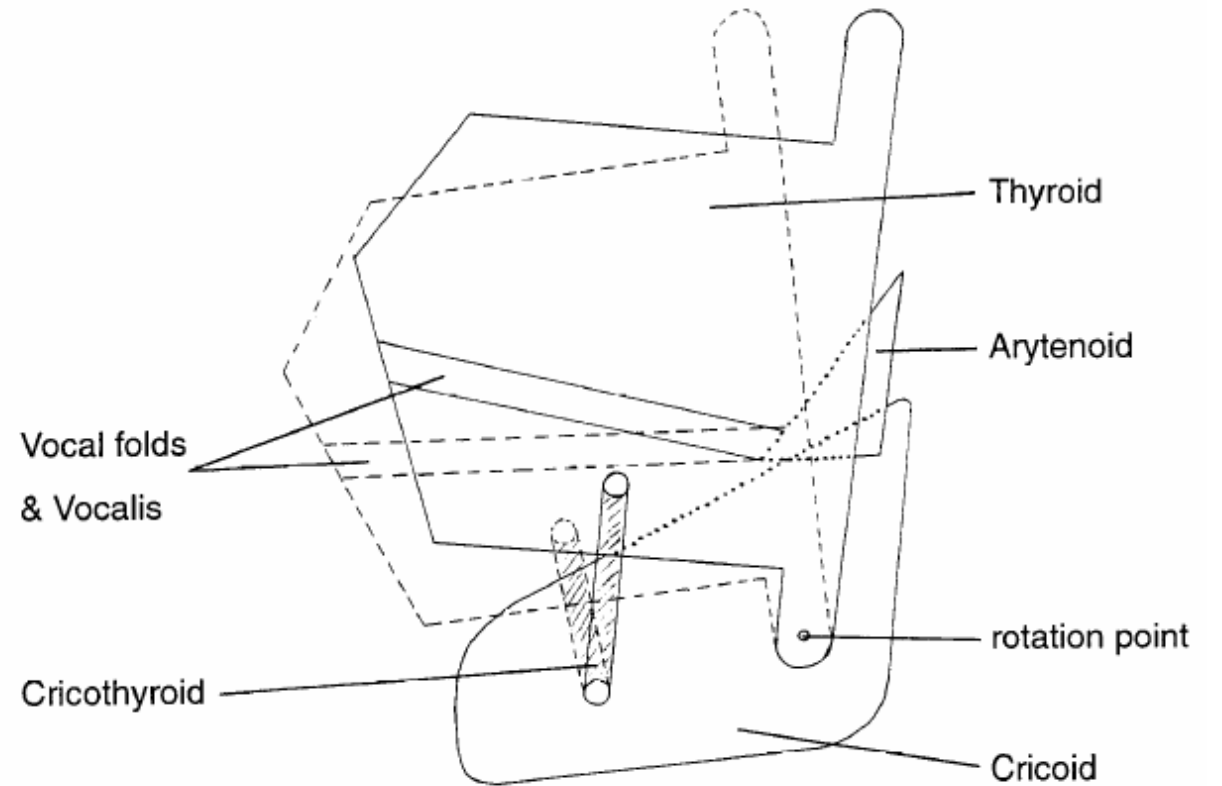
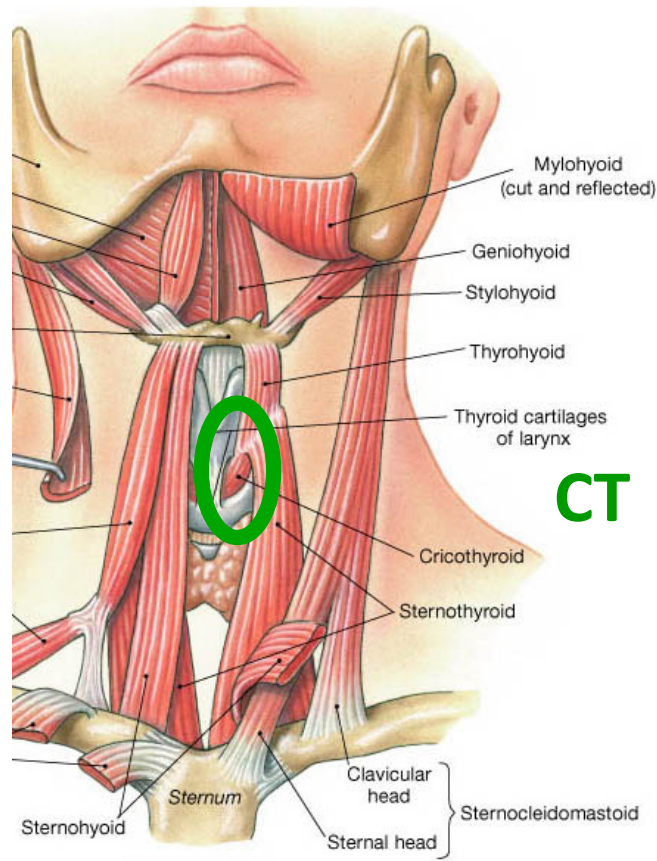
Physiology of intonation



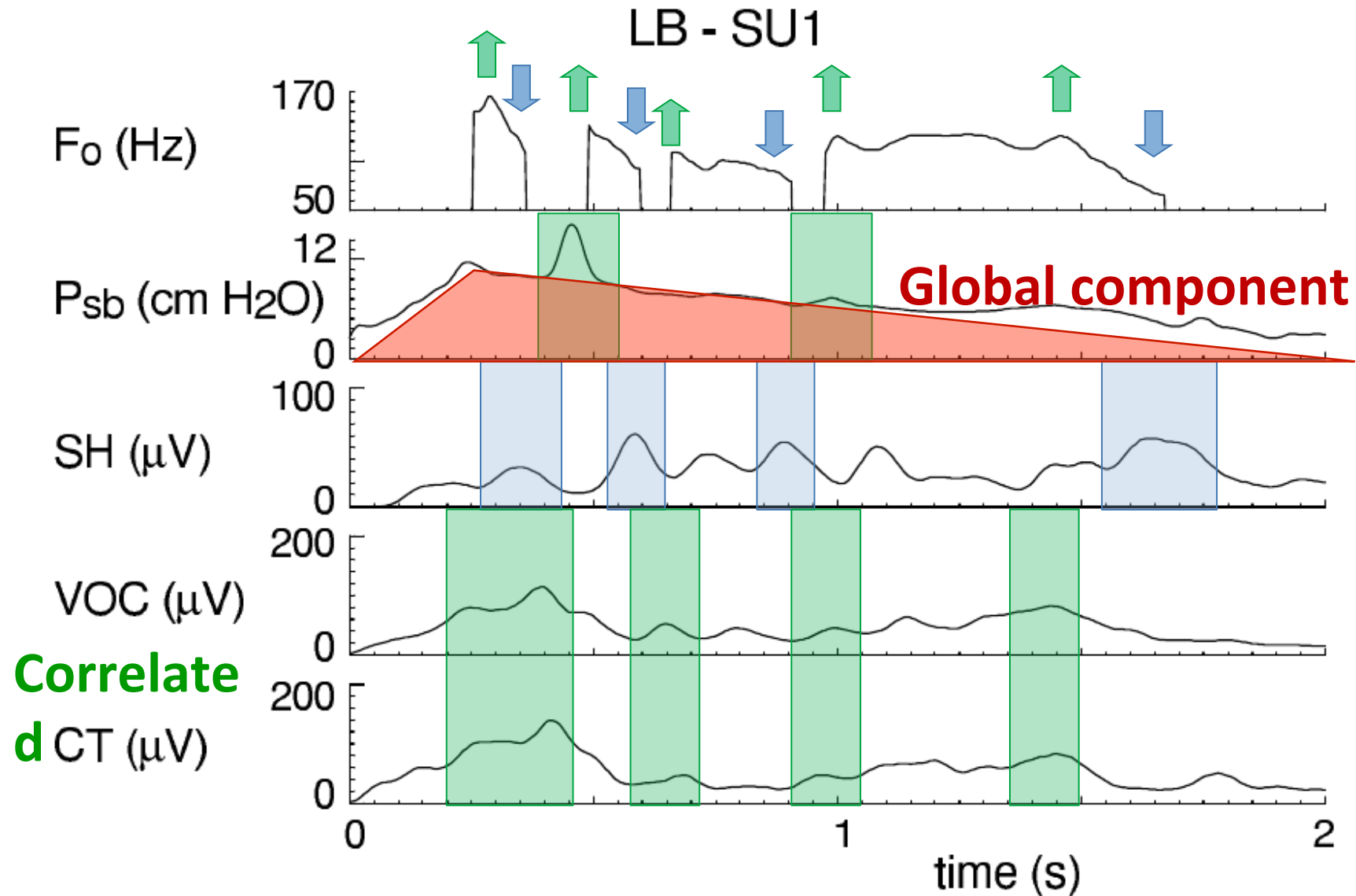
SH



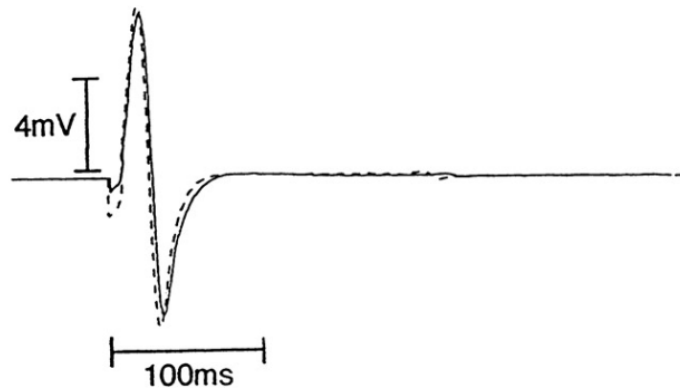
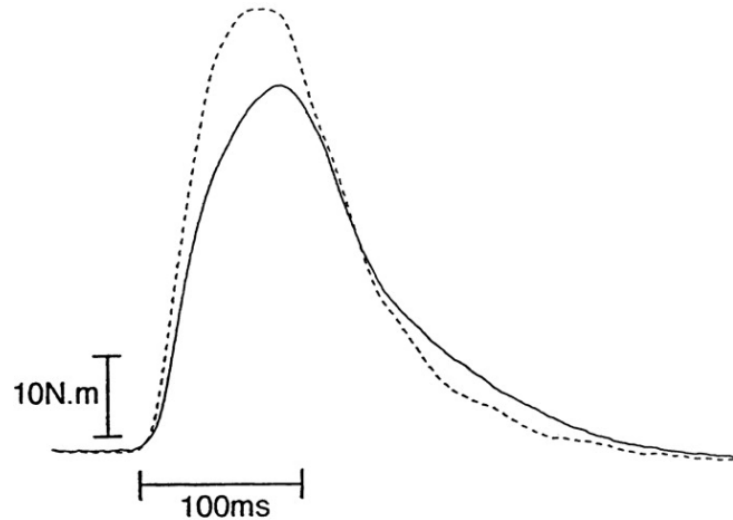
Physiology of intonation



Physiology of intonation

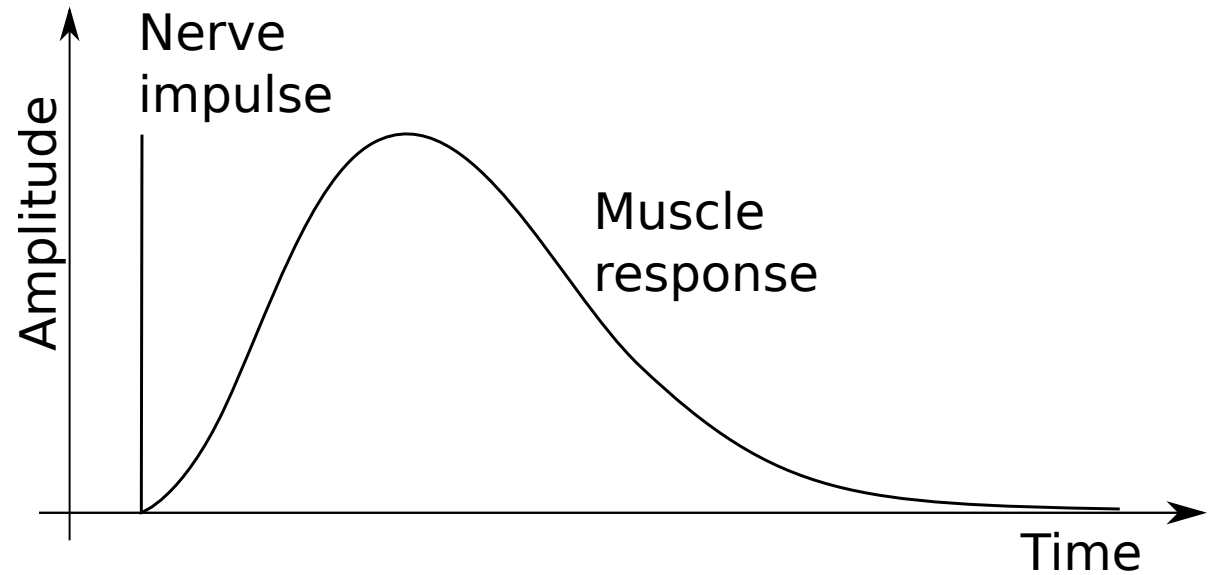
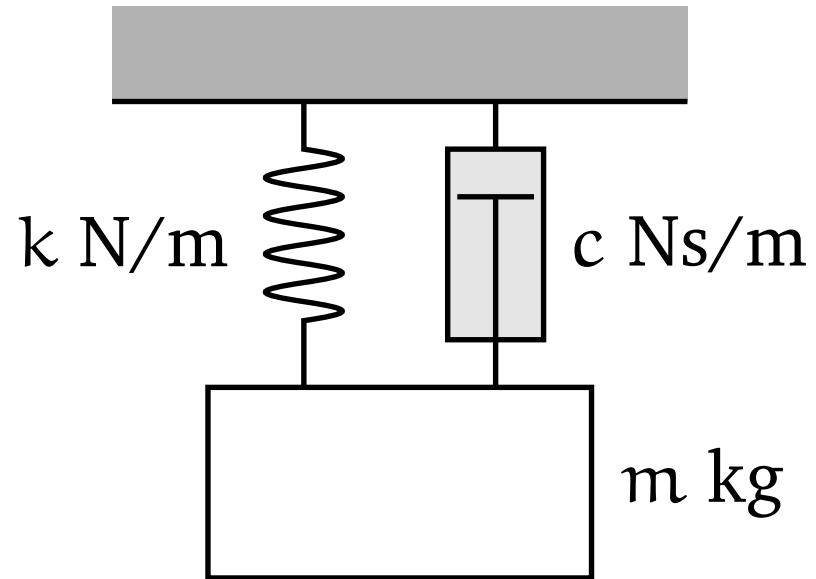


Muscle twitch



- Nerve impulses lead to muscle twitches
- Hypothesis
 - Prosody can be modelled using muscle twitches as fundamental atoms
- Similar to work of Plamondon (1995)
 - We use gamma rather than log-normal

Idealised muscle response

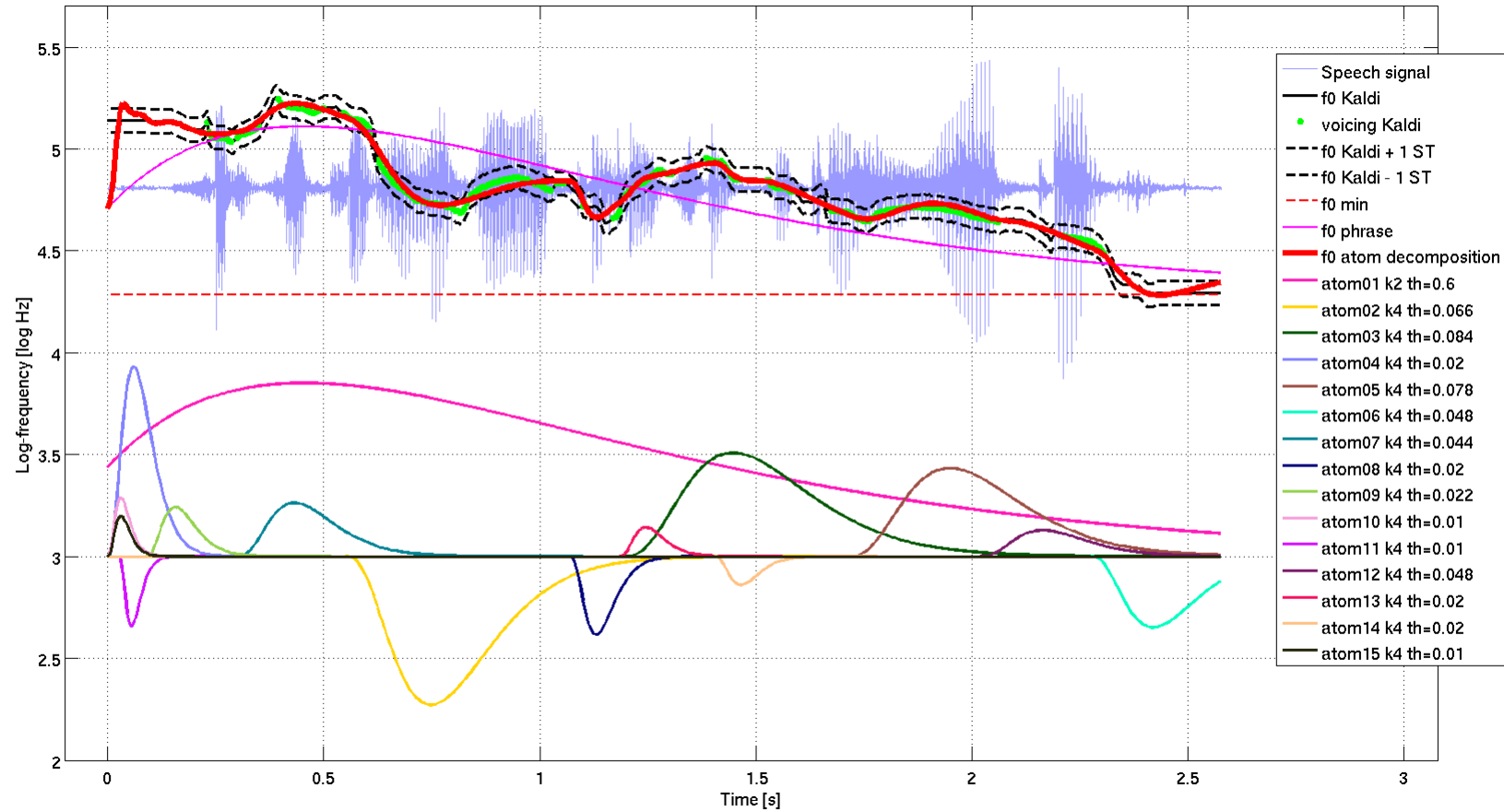


Critically damped second order system implies "gamma" form

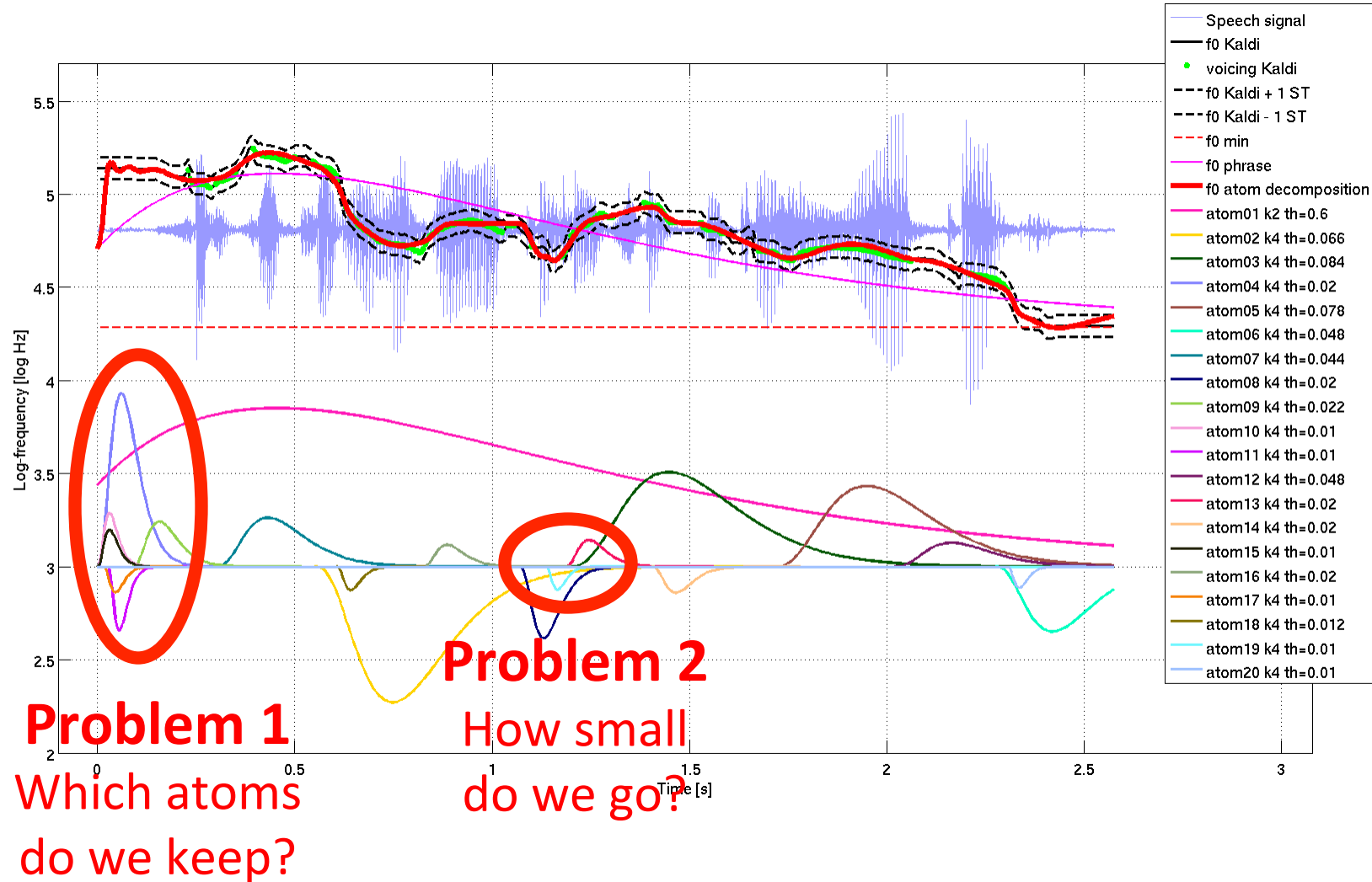
Matching pursuit (Mallat & Zhang, 1993)

- Explains a signal in terms of atoms from a dictionary
 - Greedy algorithm
 - Minimises RMS error
- Compromise between DFT and wavelet approaches
 - Removes the time-frequency dependence
- Arbitrary accuracy
 - We need a stopping criterion

Atom Decomposition Intonation Modelling



Atom Decomposition Intonation Modelling



Perceptual distance measures pt. 1

- Perceptual distance measures introduced by Hermes 1998.
 - Weighted RMS distance

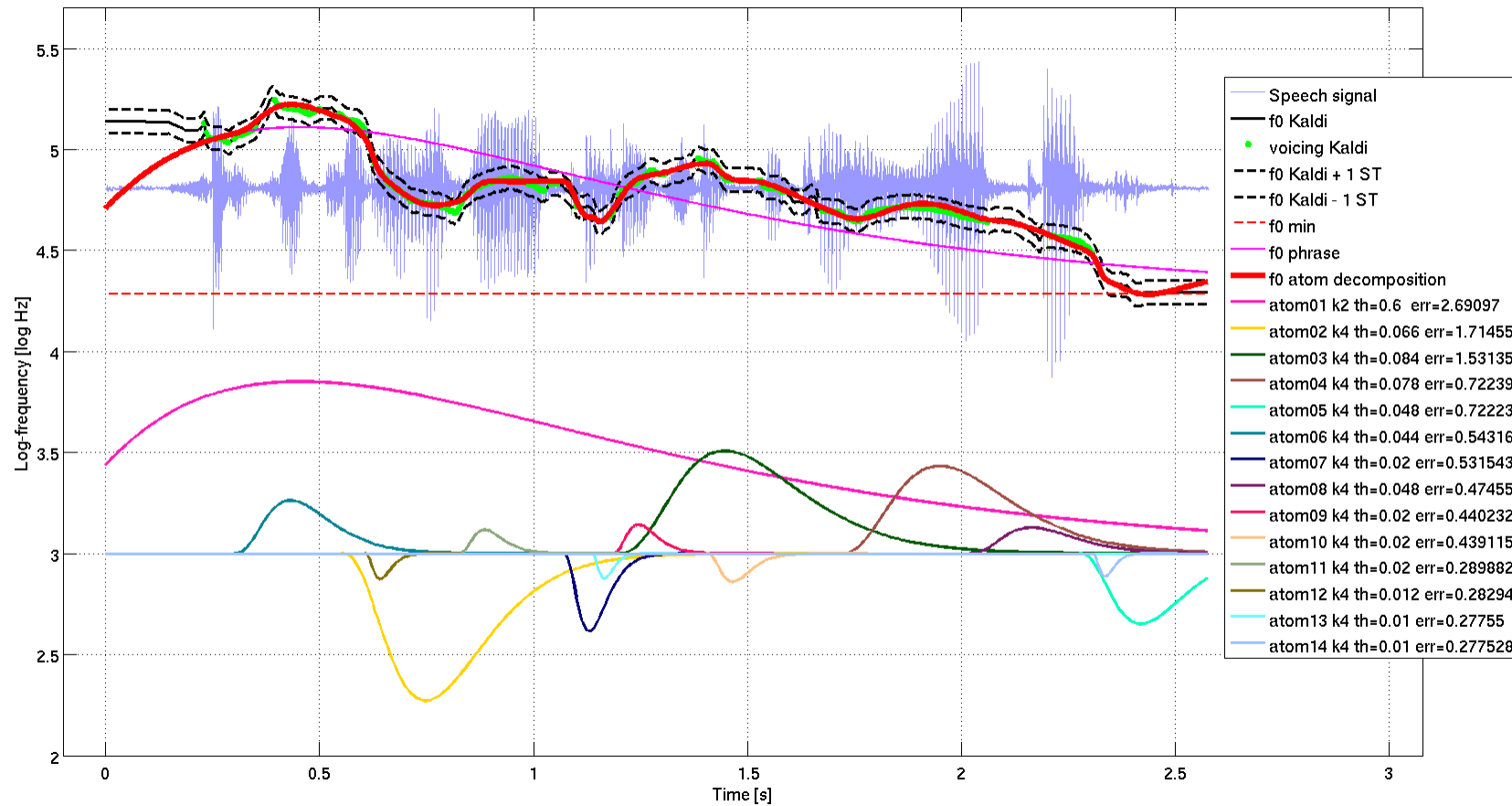
$$R = \sqrt{\frac{\sum_i w(i)(f_{synth}(i) - f_{orig}(i))^2}{\sum_i w(i)}}$$

- Weighted correlation

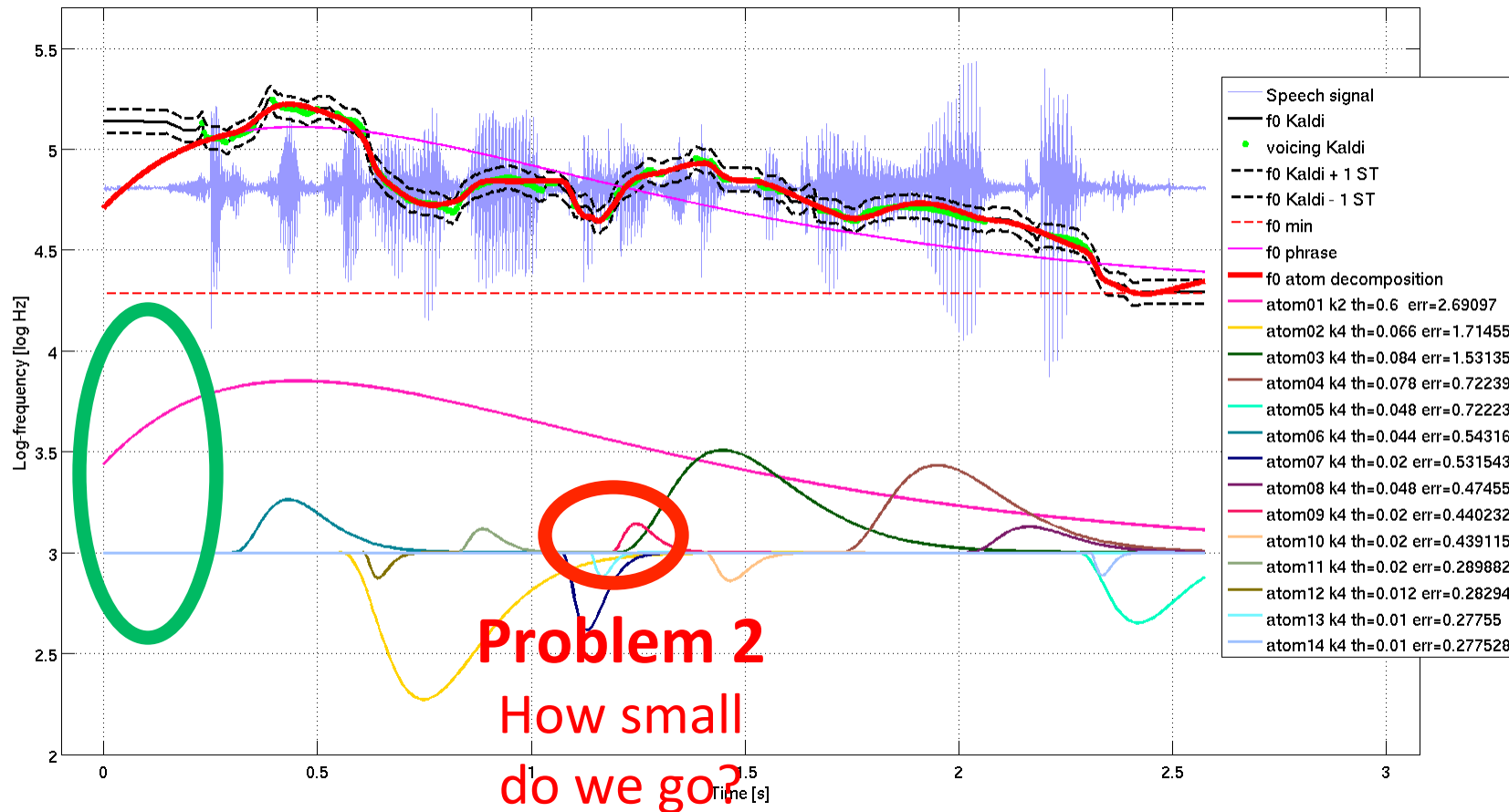
$$r = \frac{\sum_i w(i) f_{synth}(i) f_{orig}(i)}{\sqrt{\sum_i w(i) f_{synth}(i)^2 \sum_i w(i) f_{orig}(i)^2}}$$

- Continuous pitch provides weights

Atom Decomposition Intonation Modelling



Atom Decomposition Intonation Modelling



Perceptual distance measures pt. 2

- Psychoacoustic measurements have found human sensitivity of pitch change, termed the just-noticeable difference (JND), to be about 1 Hz for complex tones below 500 Hz, i.e.
 - **0.14 ST** for **men** (120 Hz average) and
 - **0.08 ST** for **women** (220 Hz average).
- In contrast to this, Hart (JASA, 1981) showed that only pitch movements above **3 ST** are of importance in communication.

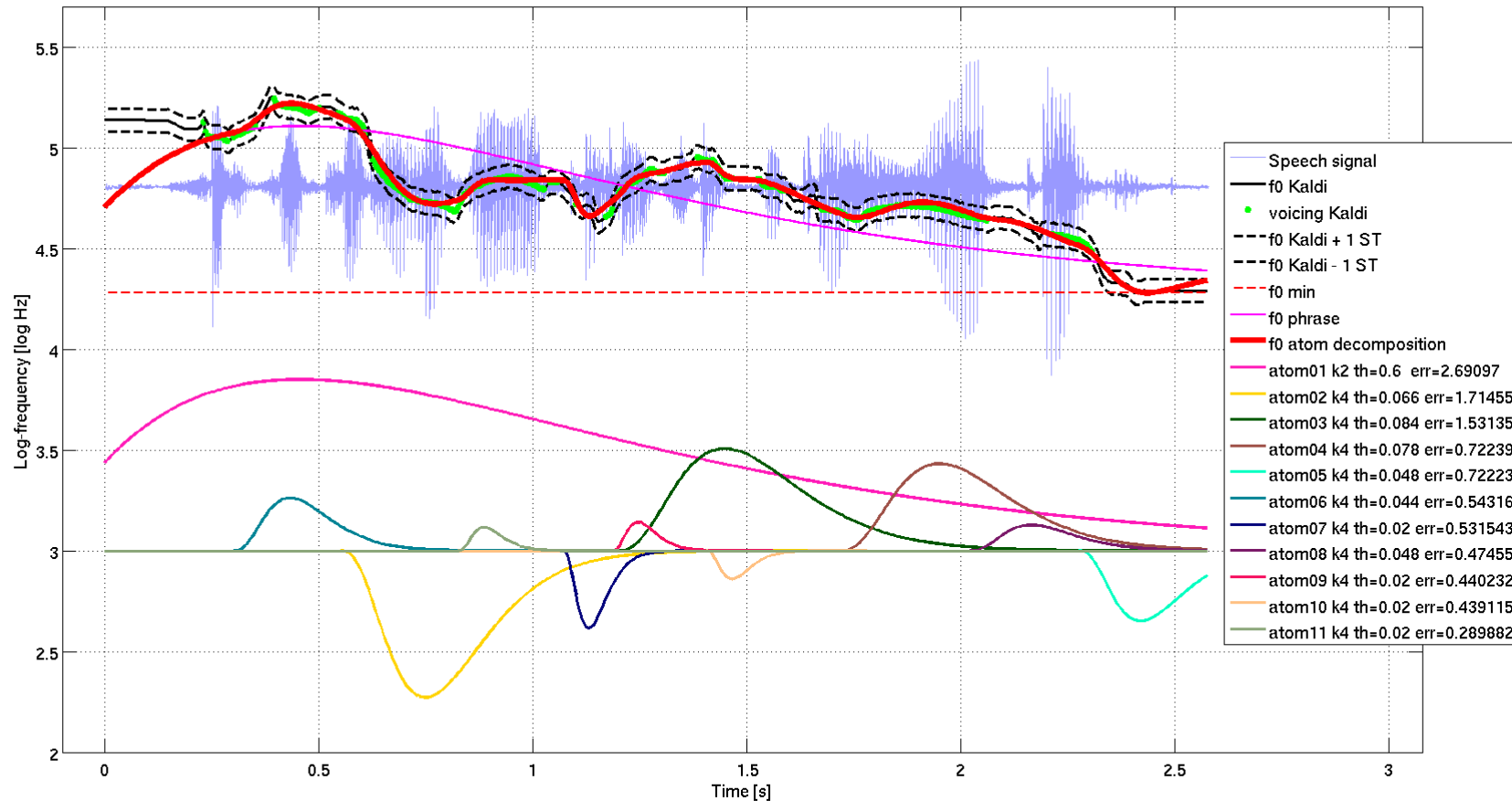
Perceptual distance measures pt. 2

Category	Weighted RMS [ST] men / women	Weighted correlation	Perceptual category
1	< 0.42 / 0.23	> 0.978	absence of perceptual differences
2	< 0.60 / 0.33	> 0.946	presence of audible differences
3	< 0.77 / 0.43	> 0.896	presence of clearly audible differences
4	< 0.91 / 0.50	> 0.827	presence of linguistic difference
5	> 0.91 / 0.50	< 0.827	completely different

(Hermes, 1998)

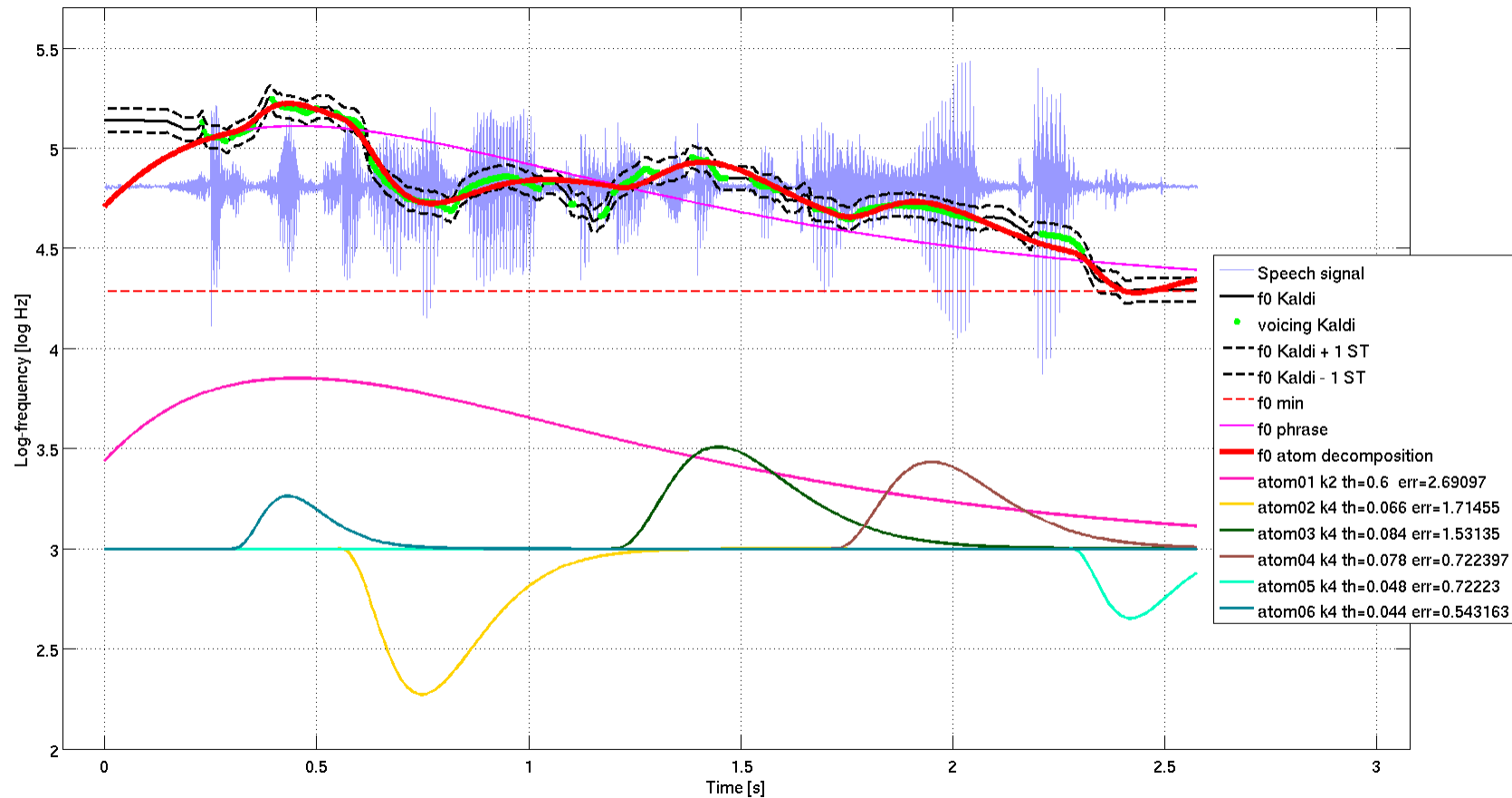
Atom Decomposition Intonation Modelling

- Cat 1 - Absence of perceptual differences



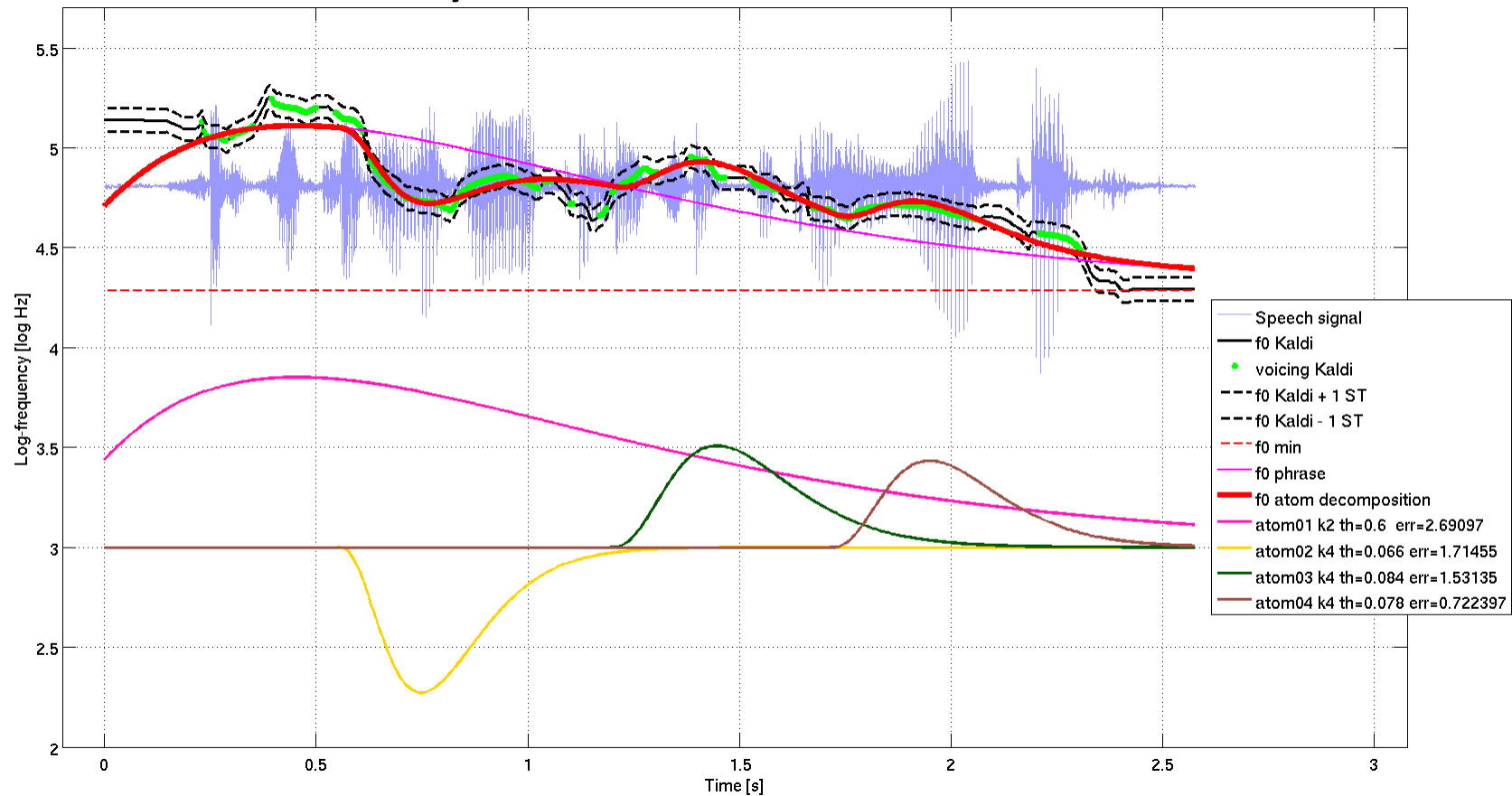
Atom Decomposition Intonation Modelling

- Cat 2 - Presence of audible differences



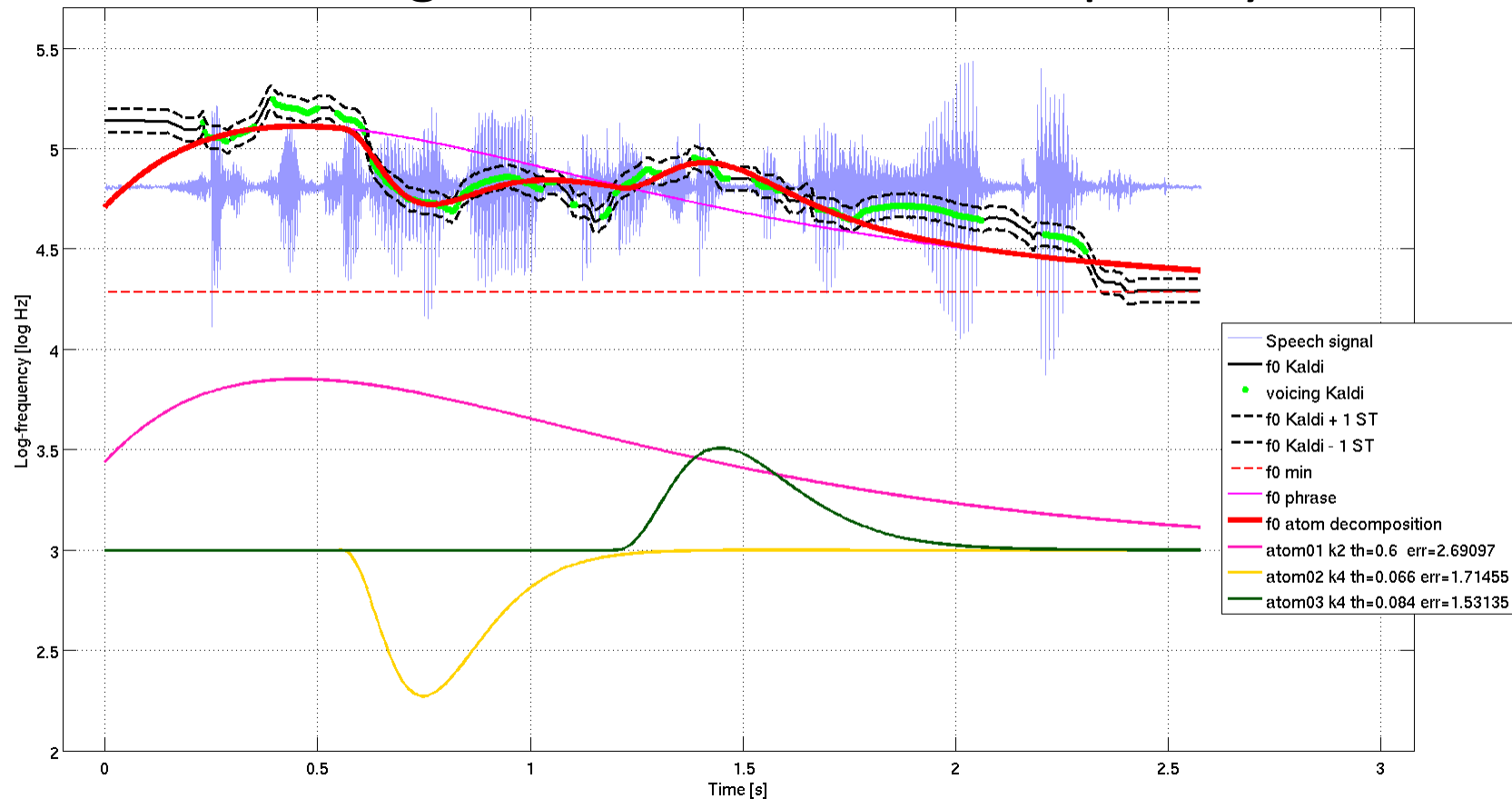
Atom Decomposition Intonation Modelling

- Cat 3 - Presence of clearly audible differences

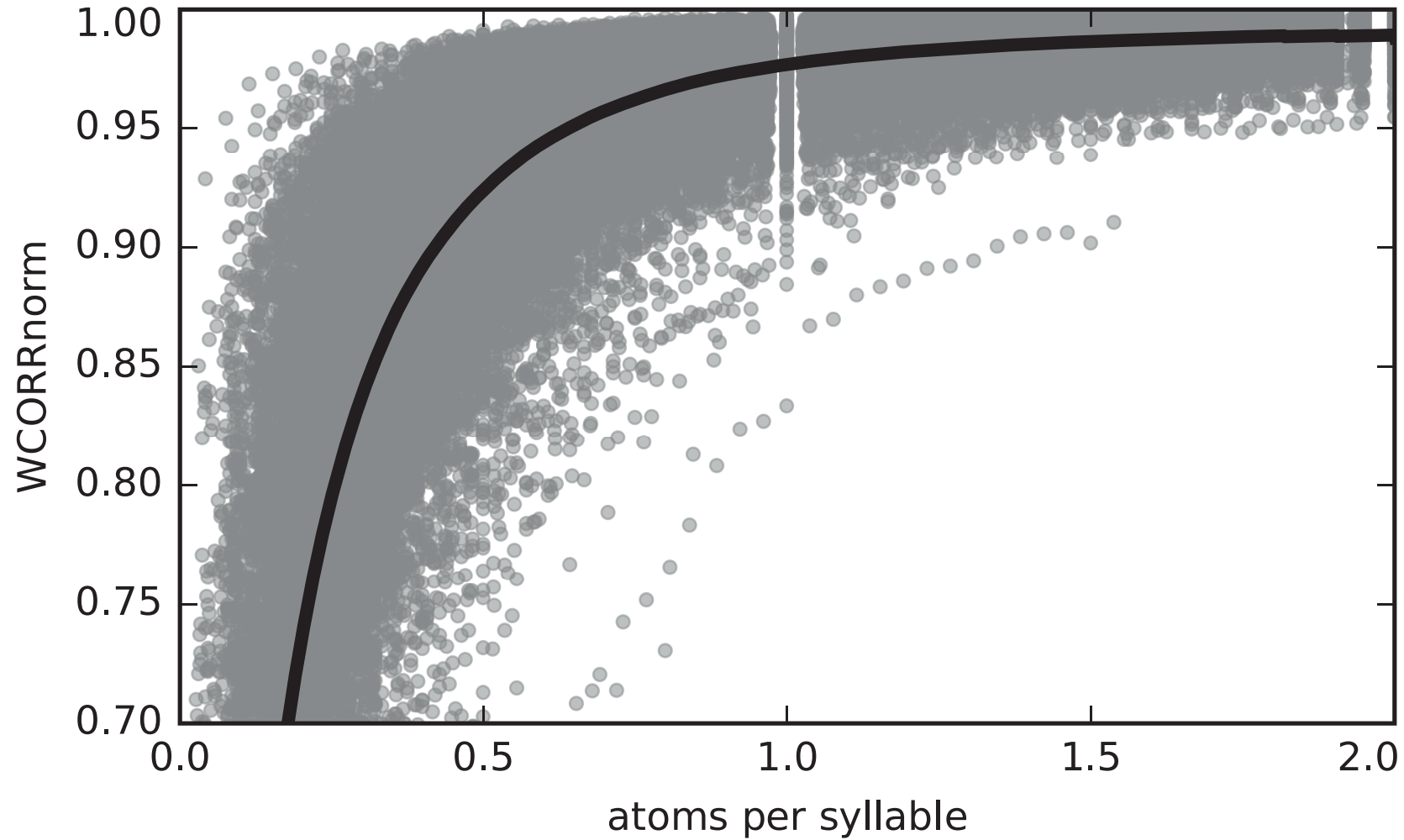


Atom Decomposition Intonation Modelling

- Cat 4&5 Presence of linguistic difference & Completely different



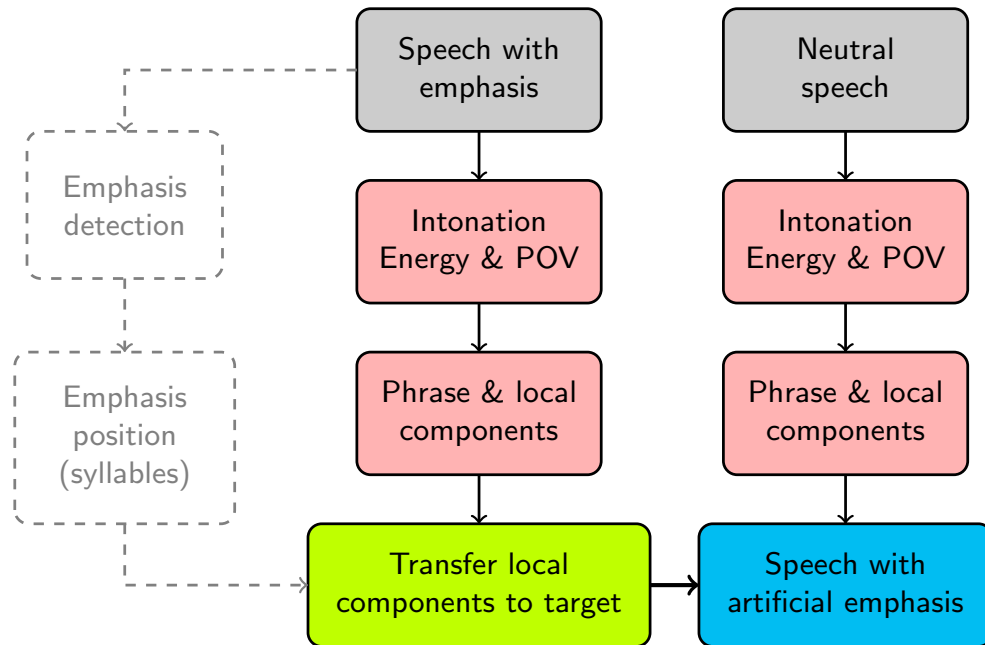
The atoms are probably modelling syllables



Part 3

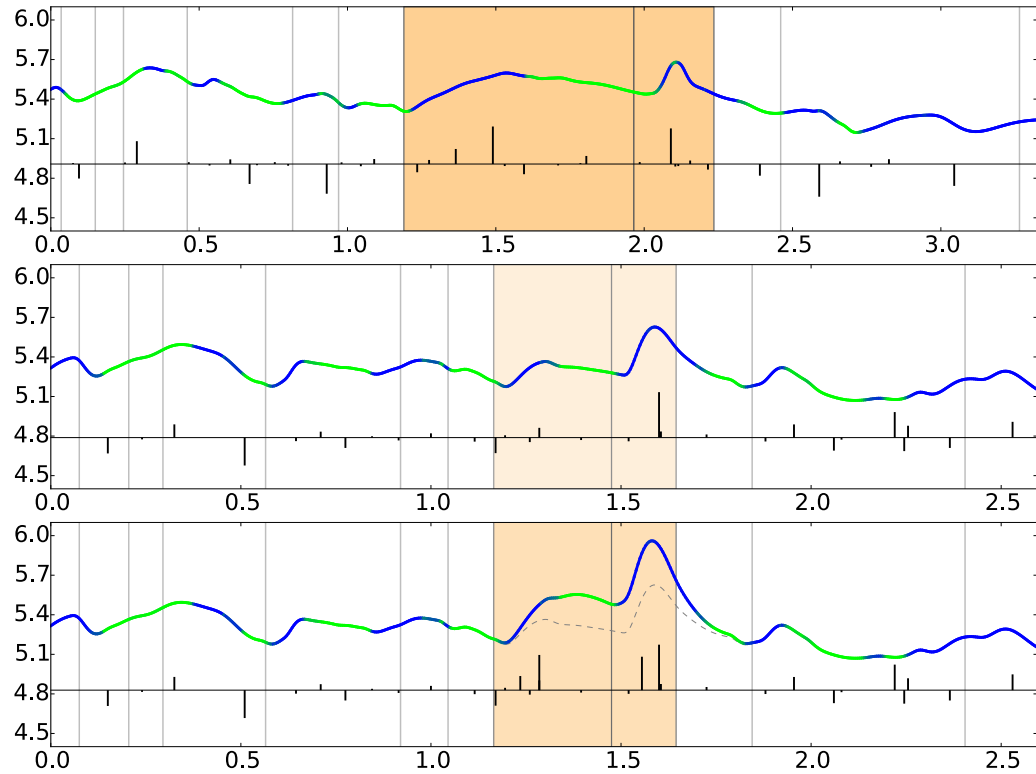
Emphasis transfer & synthesis

Transfer of local components to neutral word



- Can we transfer emphasis from an emphasised word to a neutral one?
 - No model involved
 - Test basic feasibility

Transfer of local components to neutral word



Real

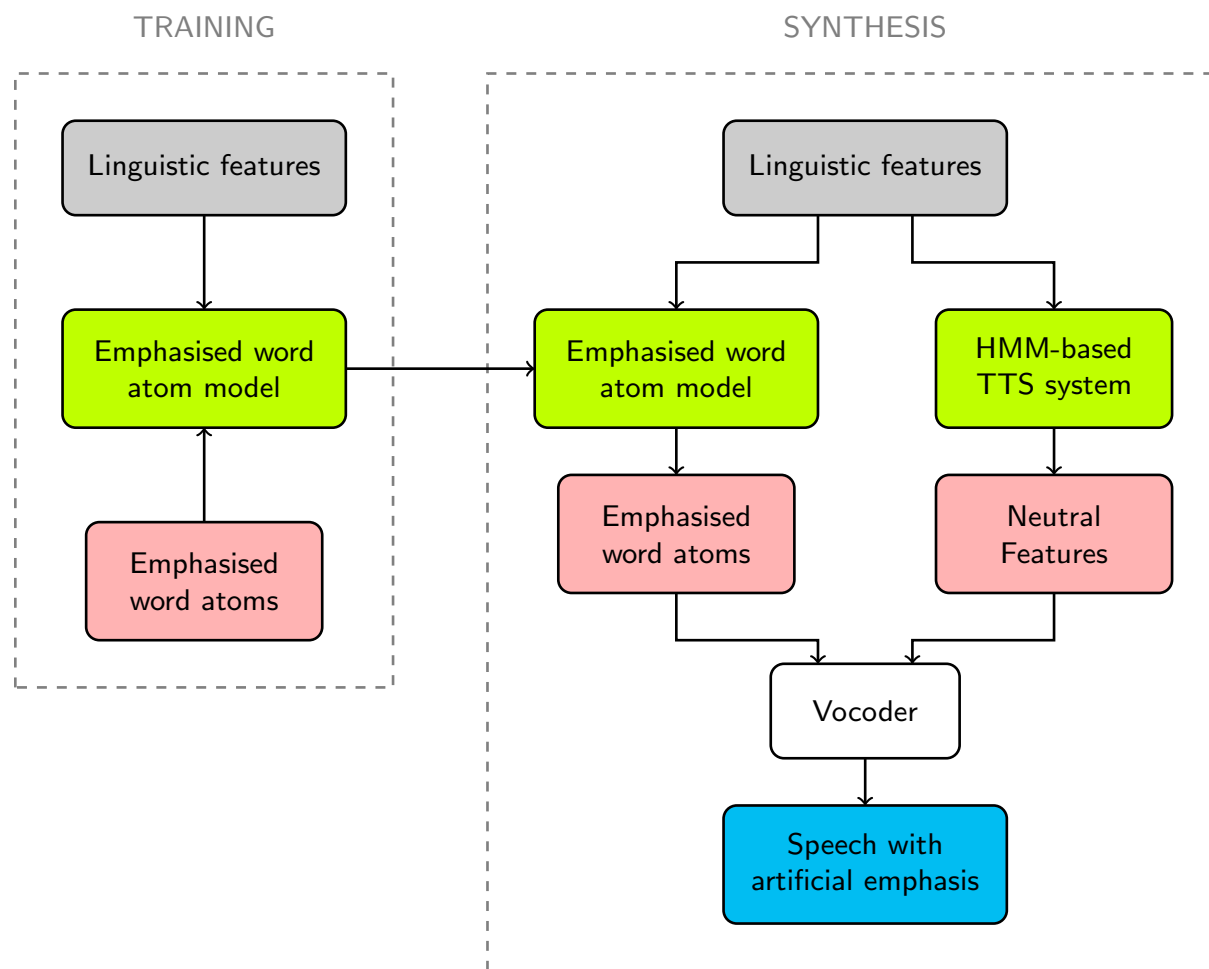


Neutral



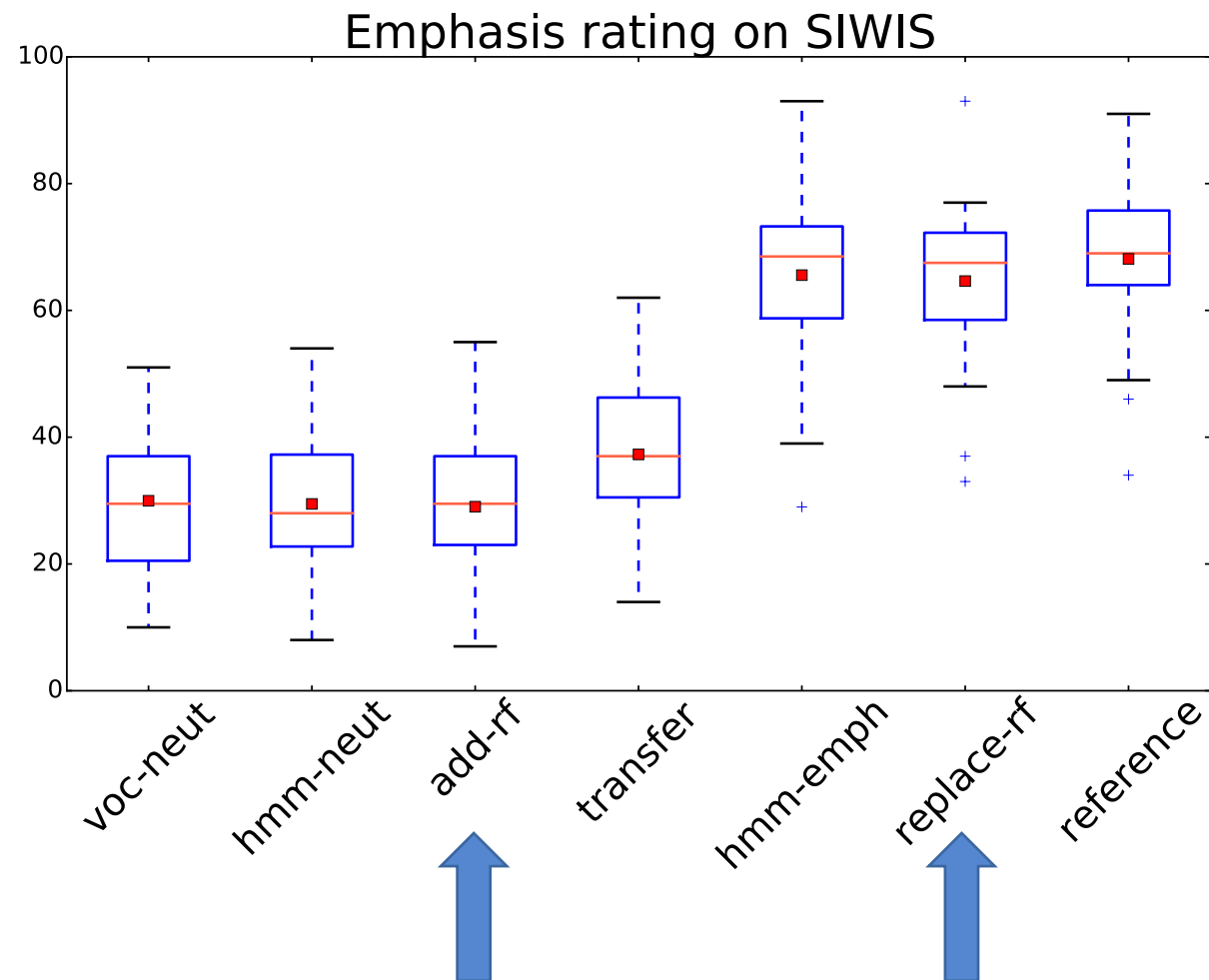
Transfer

Word level intonation for emphasis synthesis



- Build a generic emphasis model
 - Random forests
- Add to or replace the neutral prosody
- Implicit hypothesis:
 - Add: two different processes
 - Replace: indistinguishable processes

Listening test results



- MUSHRA test
 - Multiple Stimuli with Hidden Reference and Anchor
- Replacement of prosody works
 - Emphasis is an integral part of the normal prosody

Other validation

- Delić et al. (2016)
 - Found high correlation between high and low tonal events in the ToBI system and positive and negative atoms respectively.
- Szaszák et al. (2016)
 - Also used atom decomposition for emphasis detection.
- Cernak and Honnet (2015)
 - Combining stress and syllable modulation peaks to detect emphatic words

Issues

- Muscle twitches are probably shorter than syllables
 - We're probably modelling functional groups
- Pitch and energy are not independent
 - Should be modelled together
 - Energy is more dynamic
- The muscle model is too simplistic
 - 6th order works better than 2nd order
 - No plausible explanation for 6th order

Conclusions

- Intonation can be modelled in a physiologically plausible manner
 - We can't yet say how close this is to the actual mechanism
- The resulting model is “local”
 - It's good for transfer of elements to different utterance
 - You can add / replace components without continuity difficulties
- It's possible to test rudimentary hypotheses about the generation mechanism