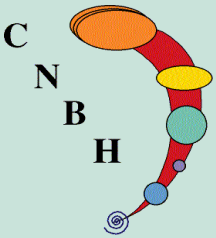


Equipe Audition, DEC, ENS 12-13 May 2006
NEW IDEAS IN HEARING

Auditory adaptation to the pitch and scale of communication sounds

Roy Patterson

Centre for the Neural Basis of Hearing,
Department Physiology, University of Cambridge, U.K.

Toshio Irino

Engineering Design, Wakayama University, Japan

email rdp1@cam.ac.uk

www.pdn.cam.ac.uk/cnbh

Colleagues and Collaborators

David Smith	speech sounds	CNBH
Tim Ives	speech sounds	CNBH
Ralph van Dinther	musical sounds	CNBH
Martin Vestergaard	speech sounds	CNBH
Tom Walters	animal sounds	CNBH

Hideki Kawahara, STRAIGHT

Engineering Design, Wakayama University, Japan

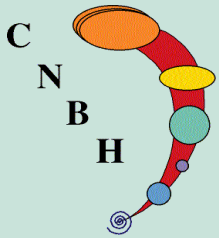
Contents: First half

I: The form of communication sounds:
pulse rate and resonance scale

II: The perception of resonance scale

III: The robustness of auditory perception to changes in pulse rate and resonance scale

IV: The auditory preprocessor adapts the analysis of sources to pulse rate and resonance scale, and produces a scale-invariant (or scale-covariant) representation of sound sources.



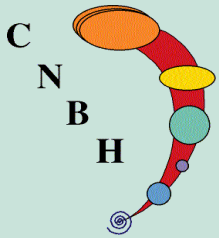
At first glance this would appear to imply that:

I: **Auditory processing is time shift invariant *and* scale invariant.**

This is not probably possible mathematically.

II: **The auditory preprocessor seems to be simultaneously invariant to changes in pulse rate and changes in resonance scale.**

This is probably not possible mathematically.

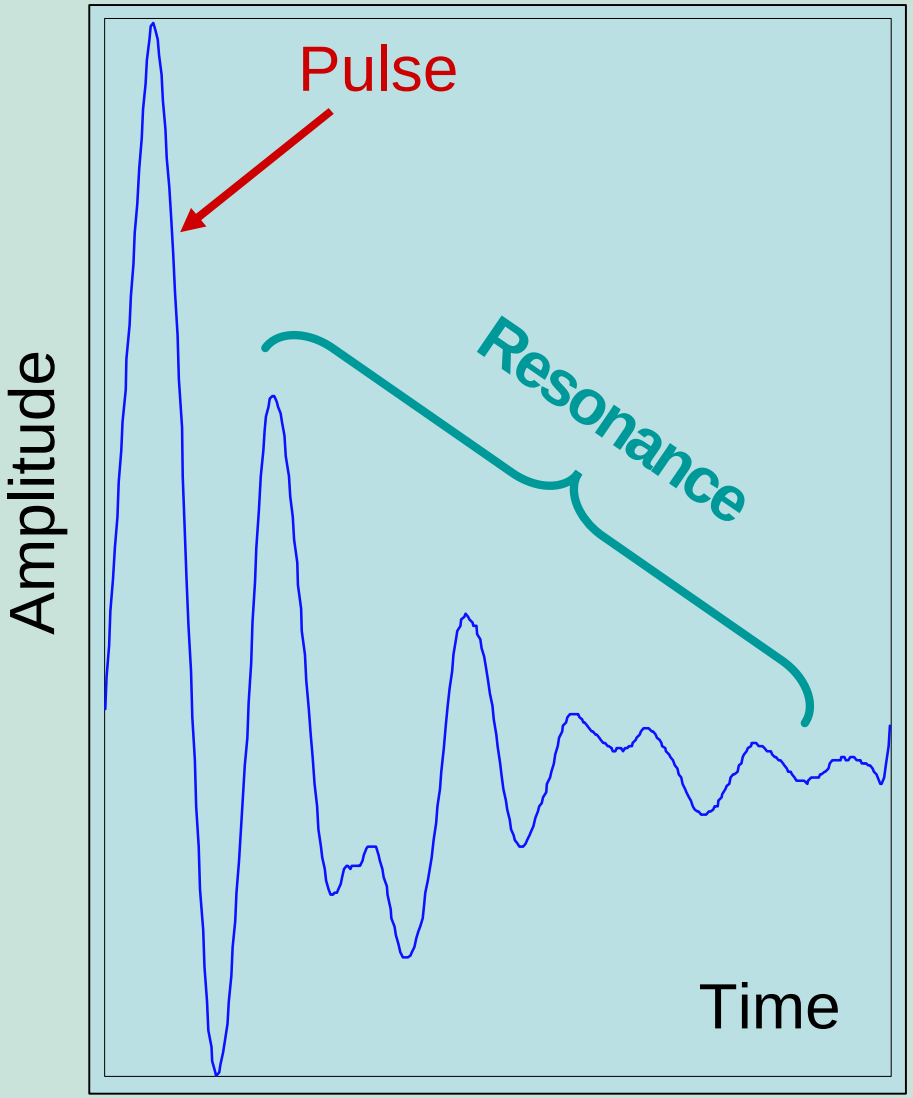


Contents: Second half

Explain how the auditory system sequences the operations to achieve the benefits of the different transforms without violating the laws of physics

Suggest that CASA and ASR will have to adopt the same approach if they are to become as robust as auditory perception

Sounds used to communicate at a distance,



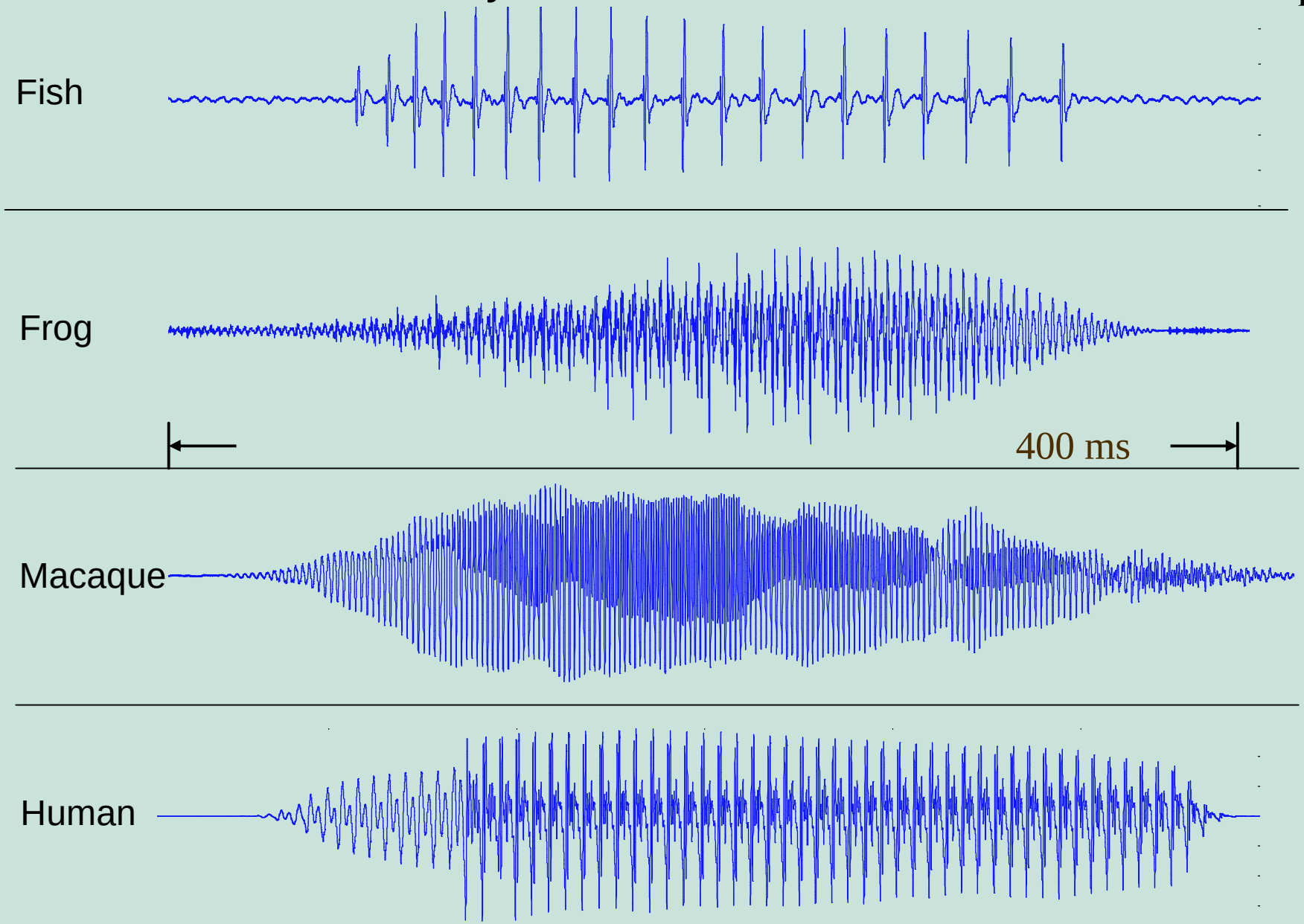
to declare territories and attract mates, are typically

Pulse-Resonance Sounds

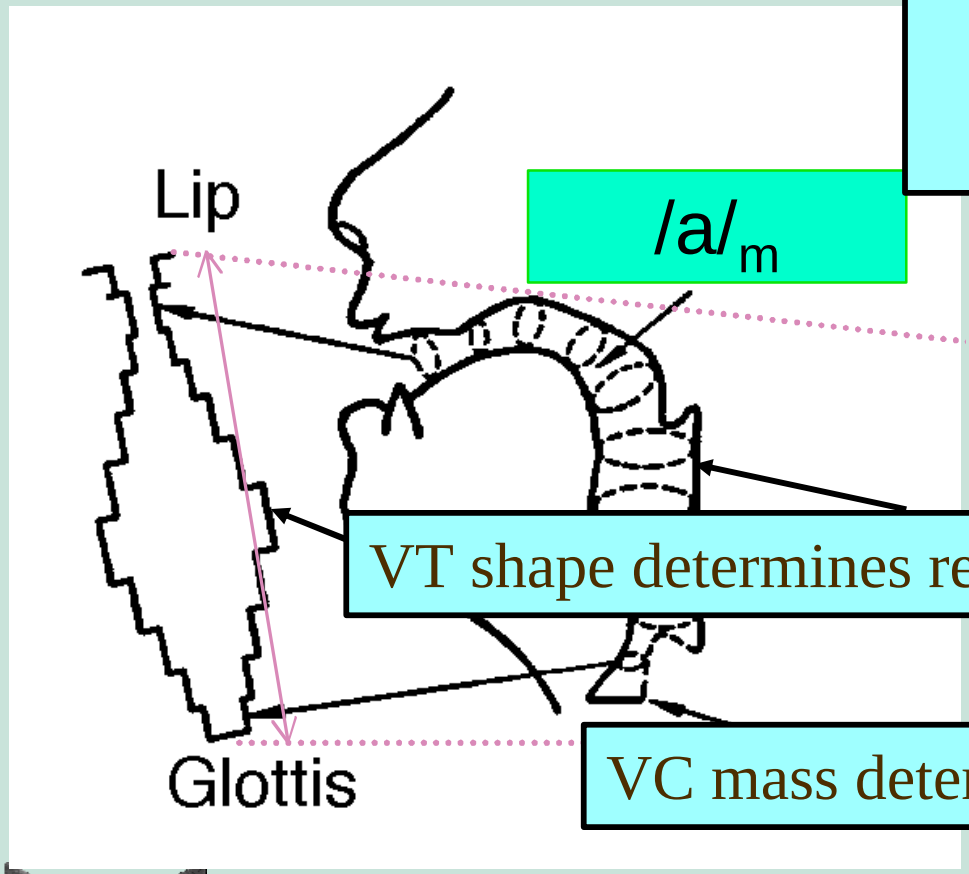
The **pulse** marks the start of the communication.

The **resonance** provides distinctive information about the shape *and* size of resonators in the sender's body.

Communication ‘syllables’ of four different animals



The information in speech sounds:



VT *length* determines resonance rate

/a/$_m$

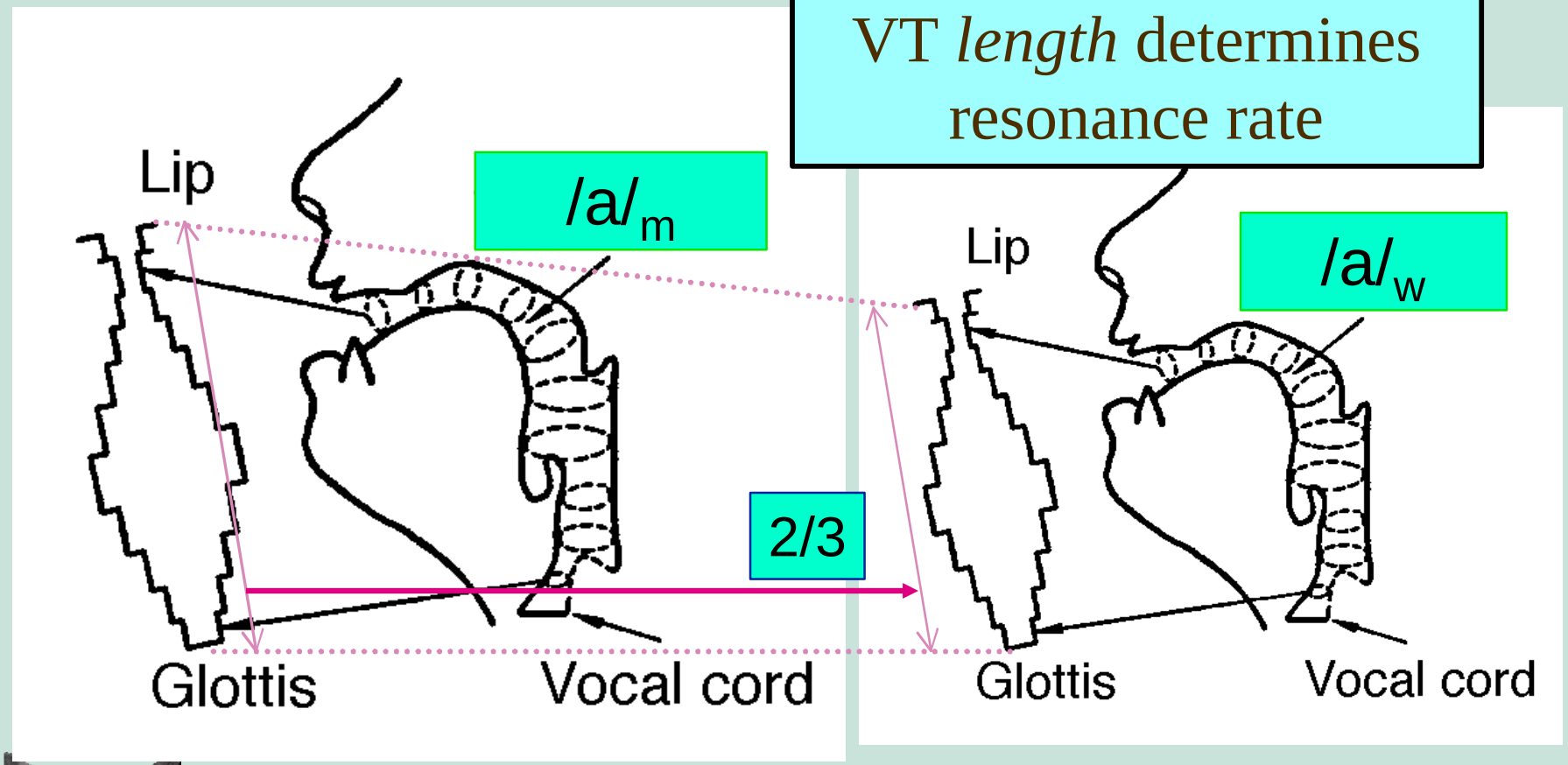
VT shape determines resonance shape (vowel type)

VC mass determines GPR (voice pitch)

Long vocal tract ~ Man

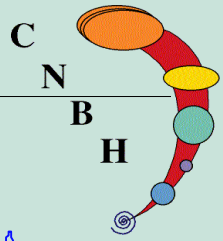
The information in speech sounds:

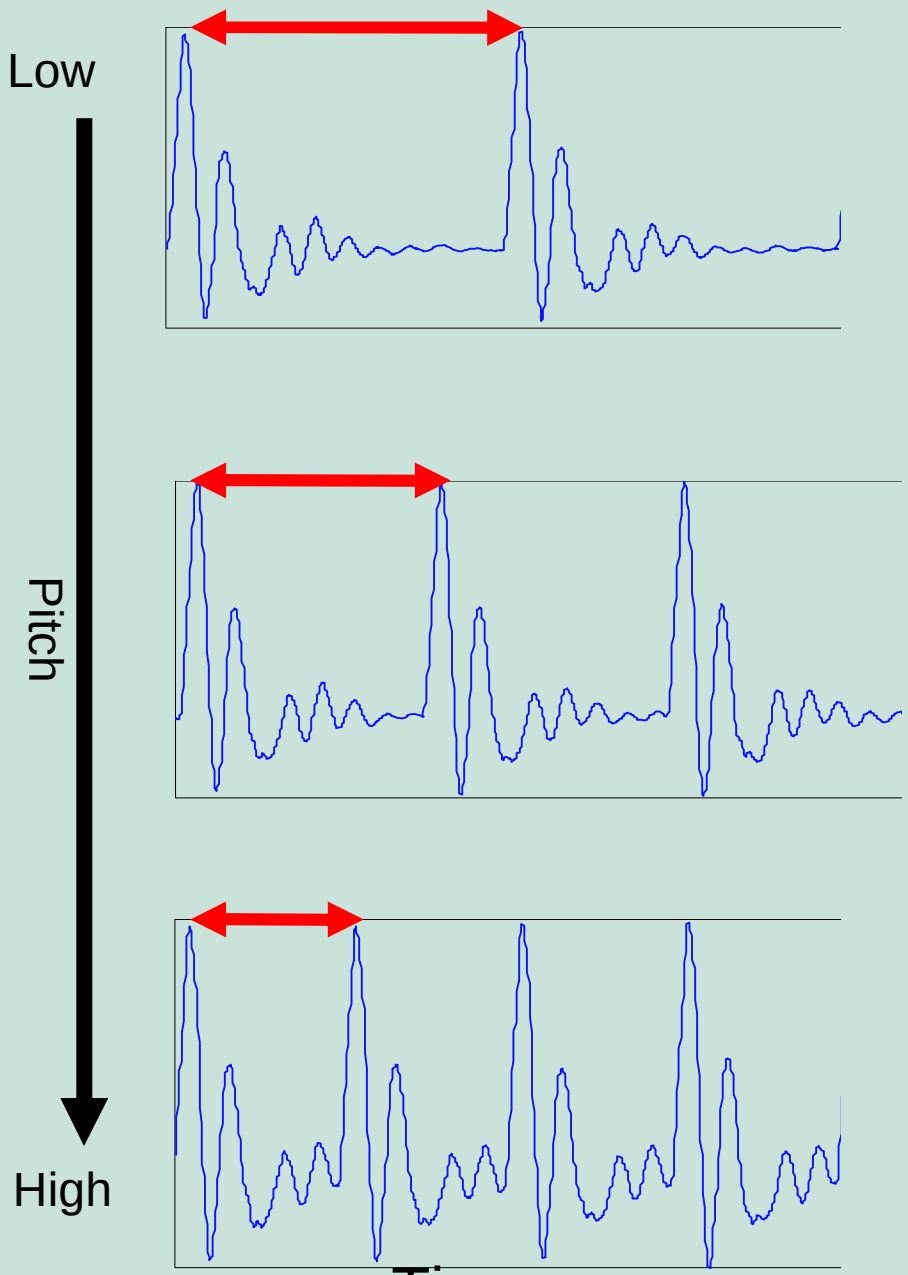
VT *length* determines resonance rate



Long vocal tract ~ Man

Shorter vocal tract ~ Woman


C
N
B
H


Low
Pitch
High
Time

Long
VTL
Short
Time

In natural communication sounds,
at the syllable level,
there are three important kinds of information:

resonance shape → the message

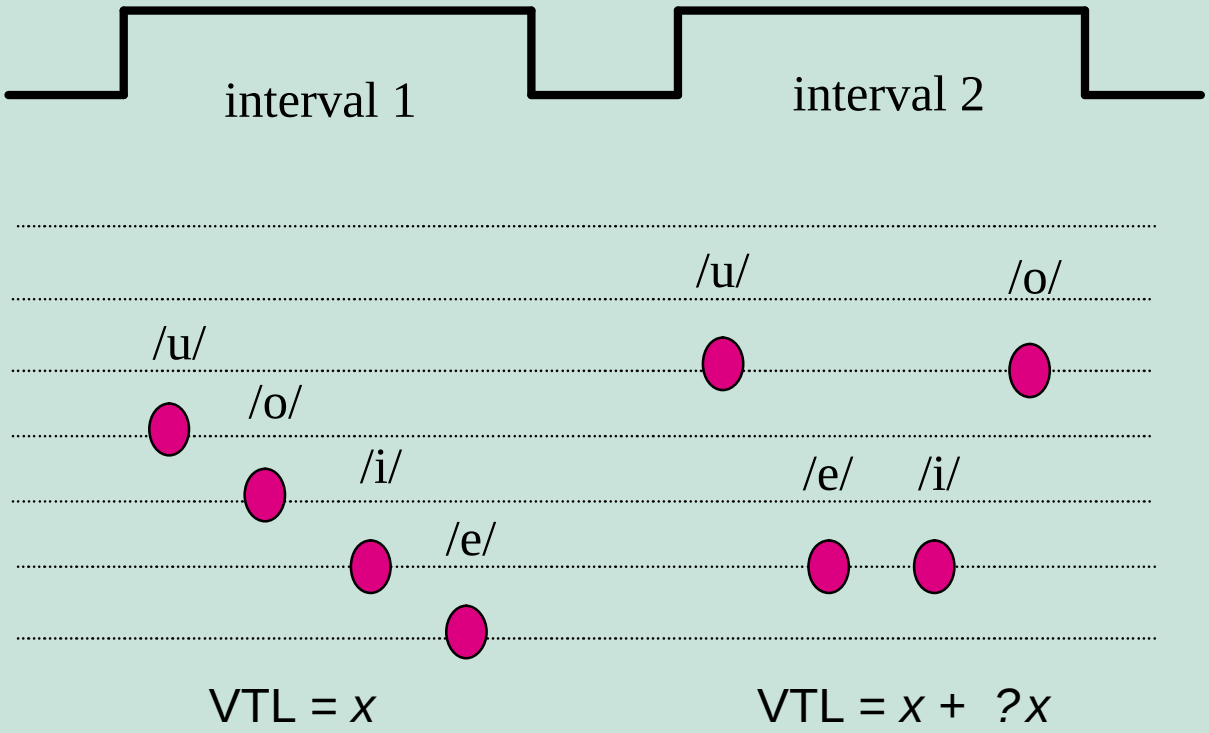
glottal pulse rate → pitch

resonance scale → resonator size and body size

Speaker-size discrimination task (vowels, or syllables)

Present two intervals of vowels and ask: *"Which is the smaller speaker?"*

- Different VTLs in the two intervals
- Rove level between intervals
- Different pitch contours between intervals
- Only consistent cue is the change in VTL

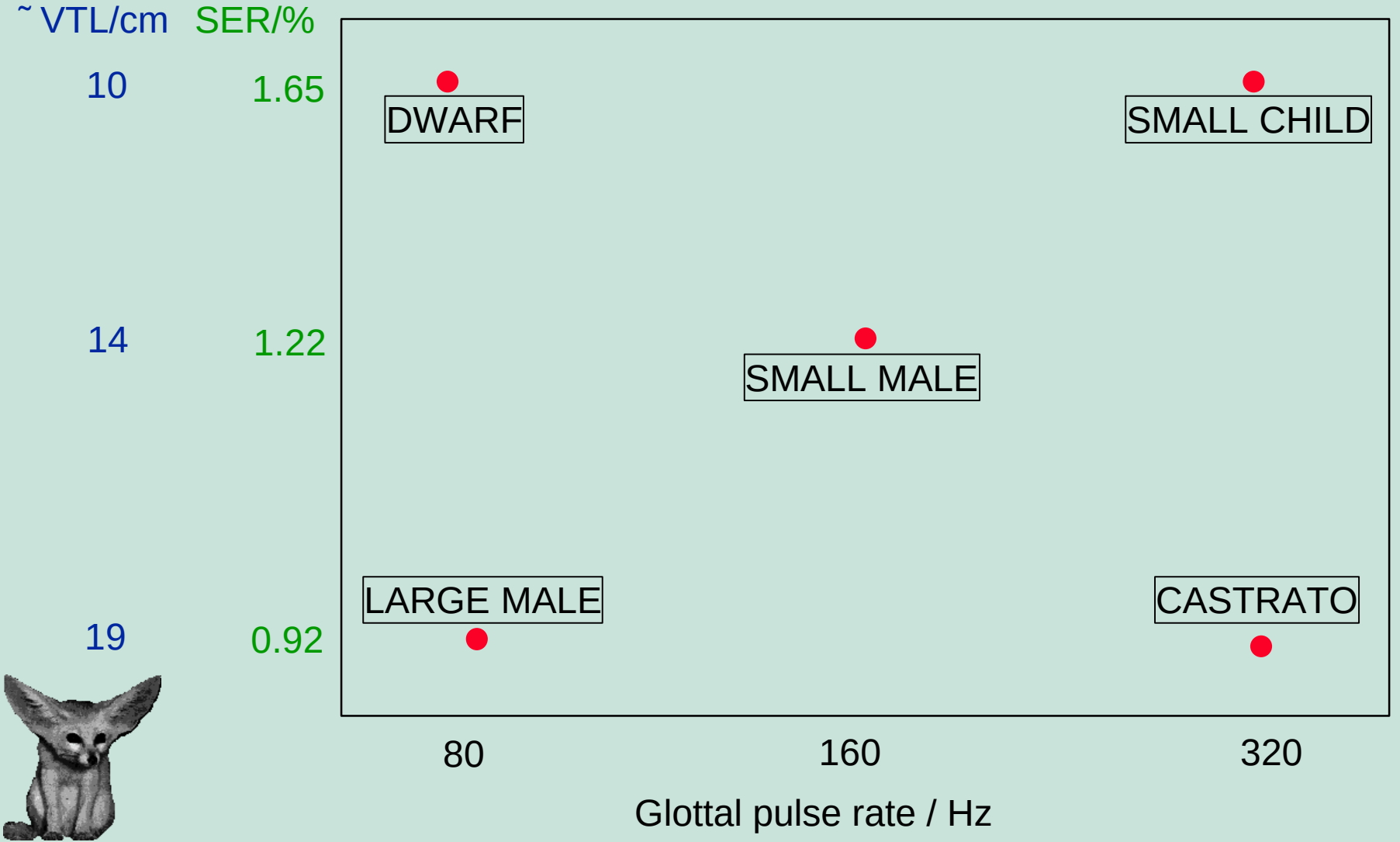


Smith, Patterson, Turner, Kawahara and Irino JASA (2005)

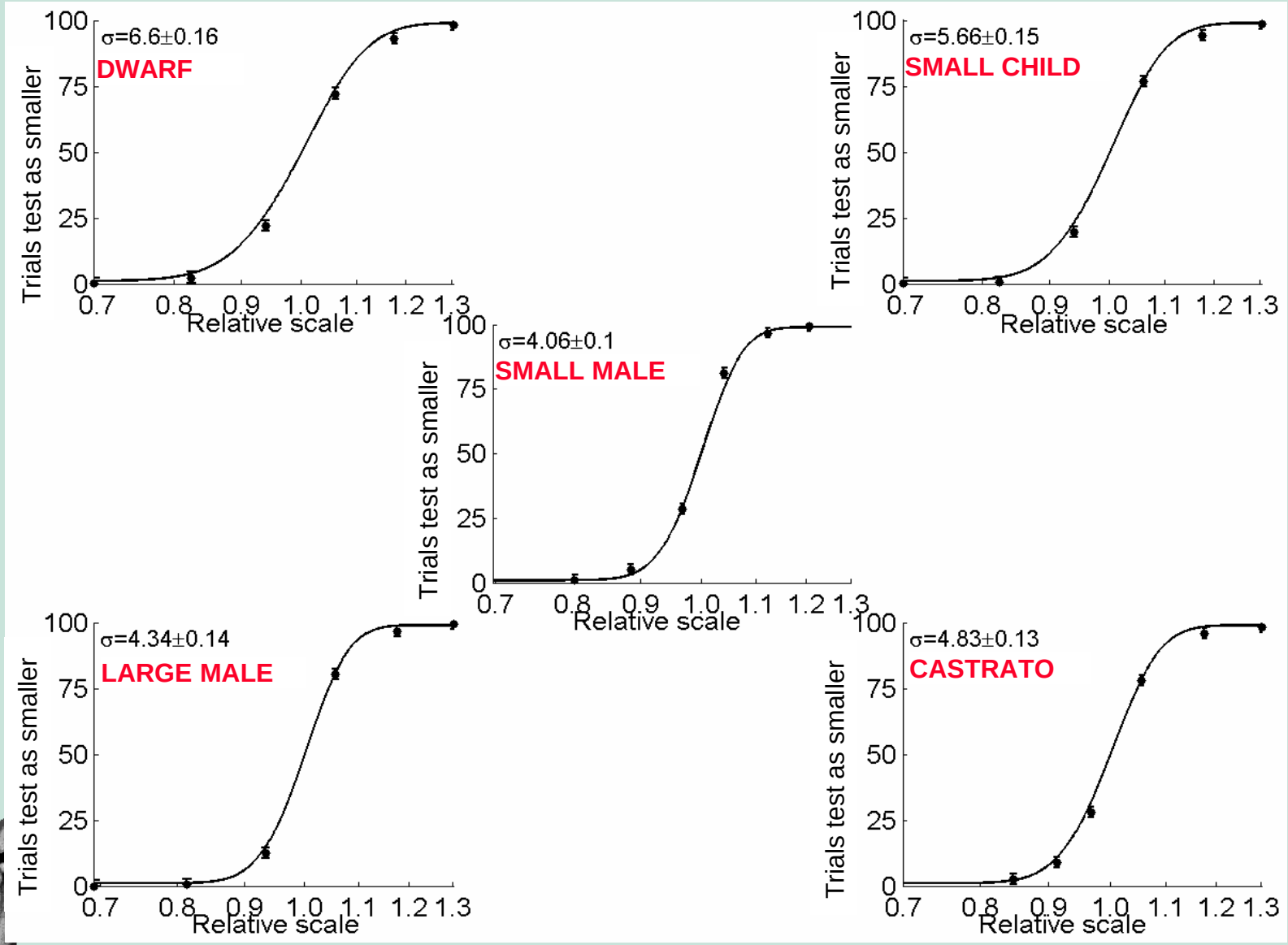
Ives, Smith and Patterson, JASA (2005)

Experiment

Measure size discrimination thresholds for different sized people



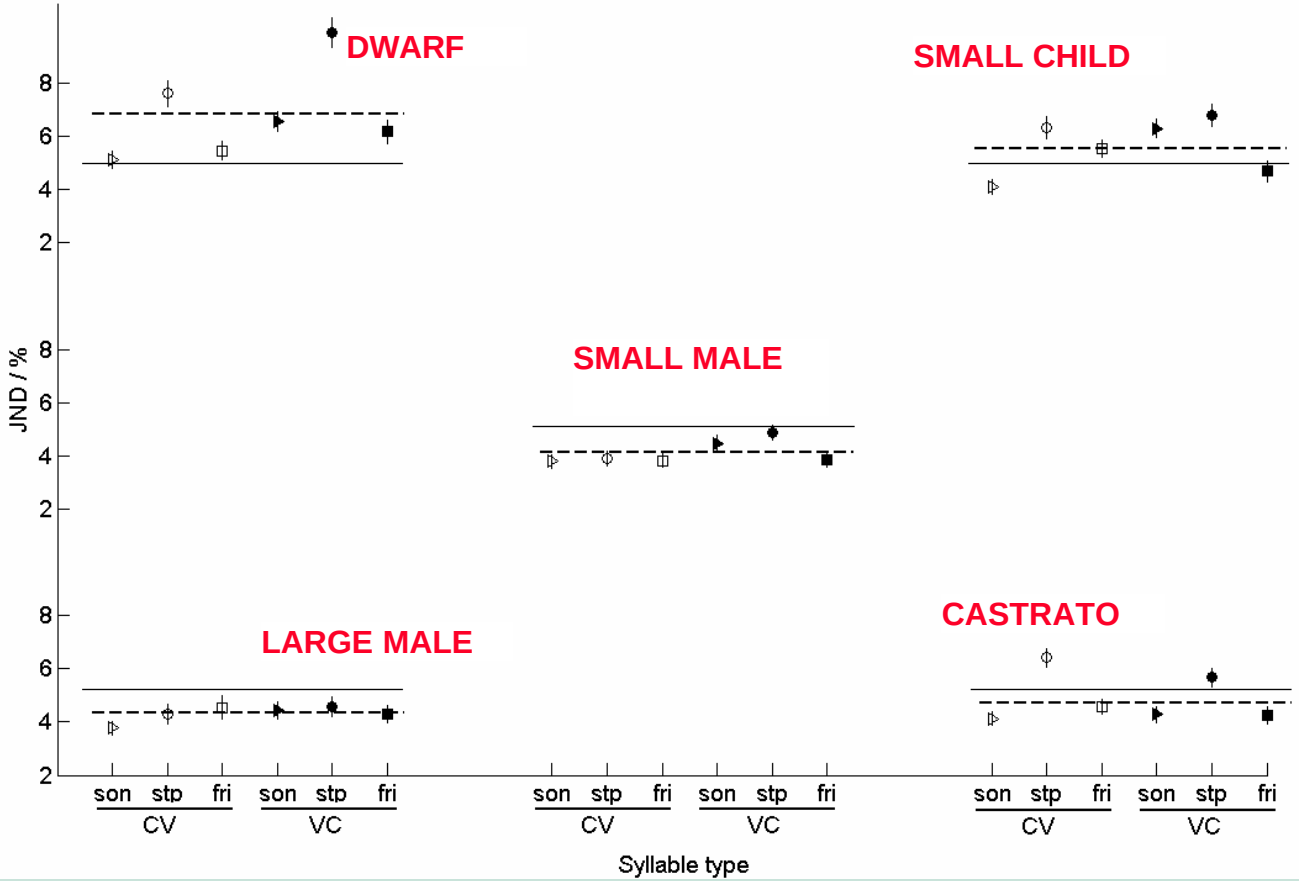
Results: all subjects, all stimuli



Smith, Patterson, JASA (2005)

Results: all subjects, all stimuli (Syllables)

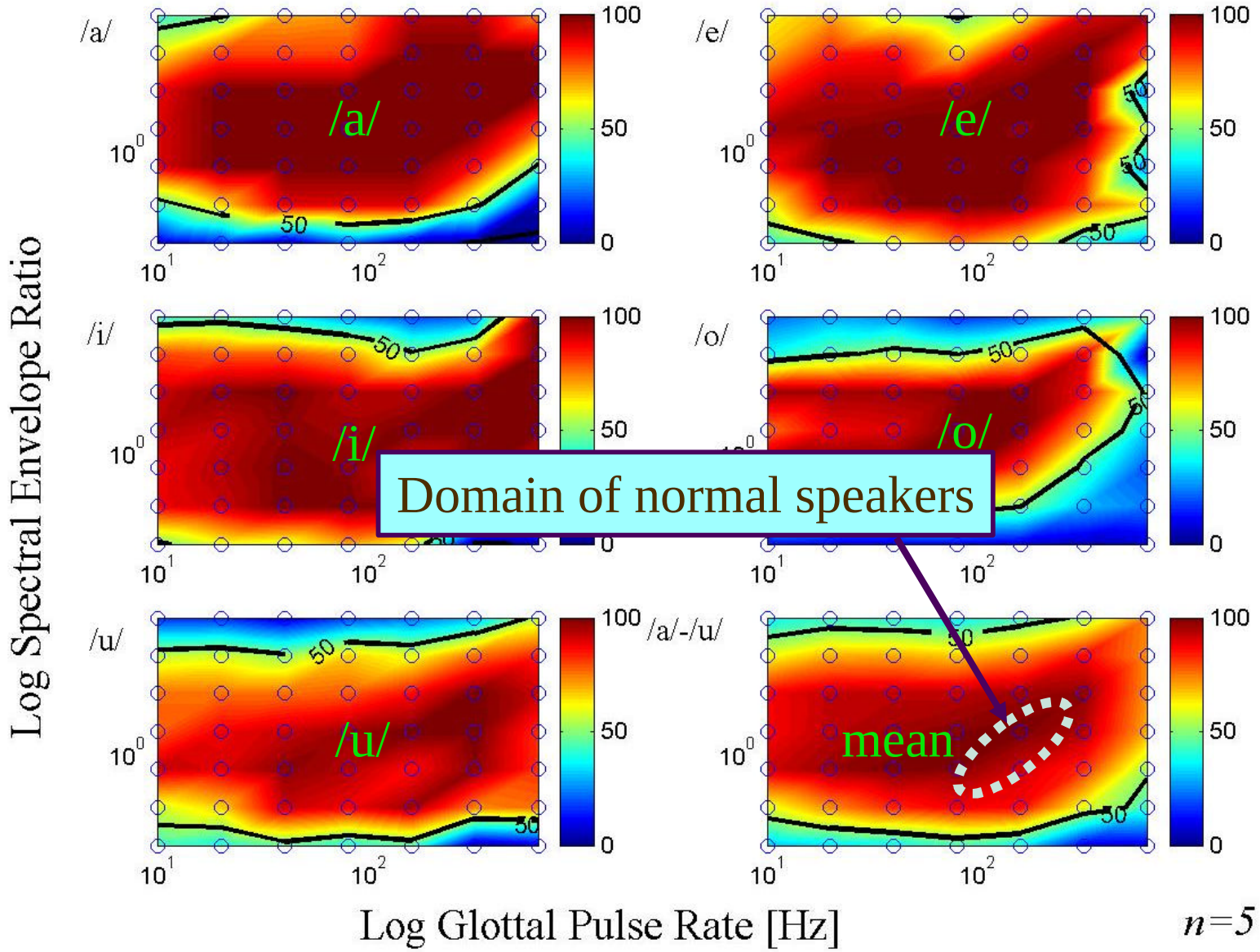
Ives, Smith and Patterson, JASA (2005)



average JND across syllable category for specific speaker type.

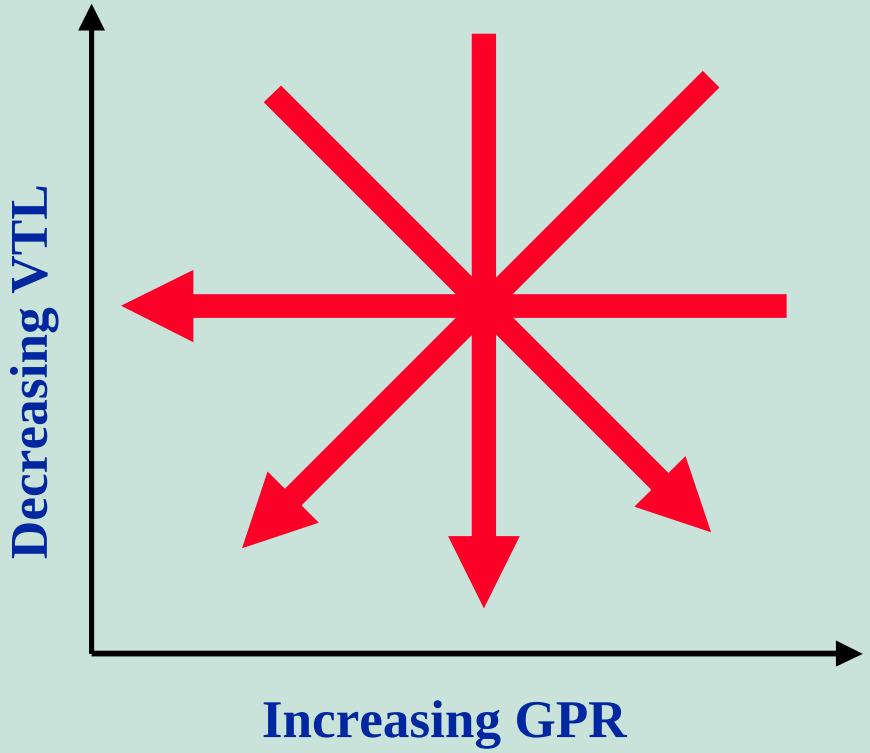
grand average JND for the experiment

Recognition of Scaled Vowels



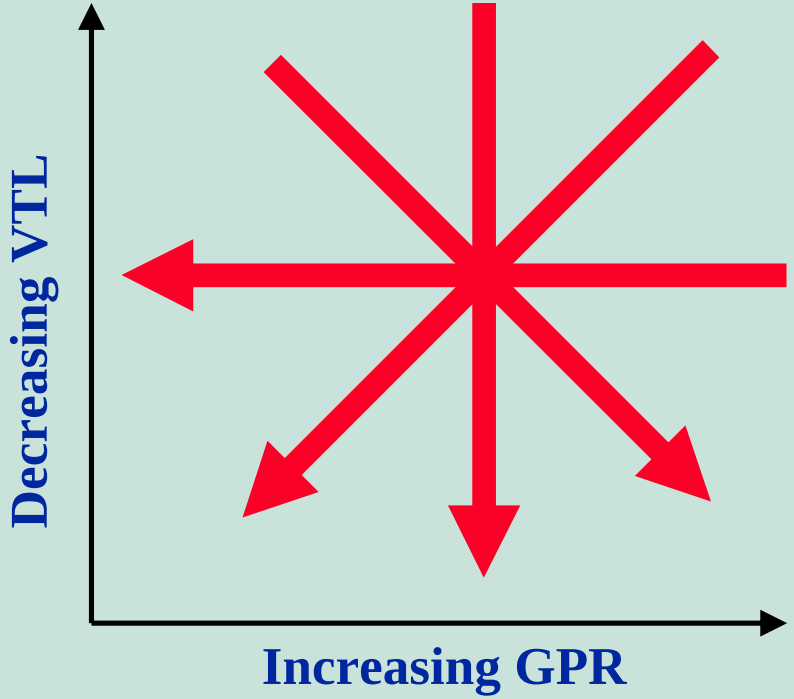
Smith, Patterson, Turner, Kawahara and Irino JASA (2005)

The effect of GPR and VTL on the perception of speaker size



Kawahara and Irino (2004). STRAIGHT, Kluer

Rana catesbeiana



Kawahara and Irino (2004). STRAIGHT, Kluer

Interim summary

The auditory system appears to adapt to the pulse rate and resonance scale of communication sounds automatically.

The processing appears to produce a carrier-invariant representation of the message.

So how do we do it ?
What are the physical constraints if the system is to be effectively shift *and* scale invariant, to both pulse rate *and* resonance scale.

The operations and their order:

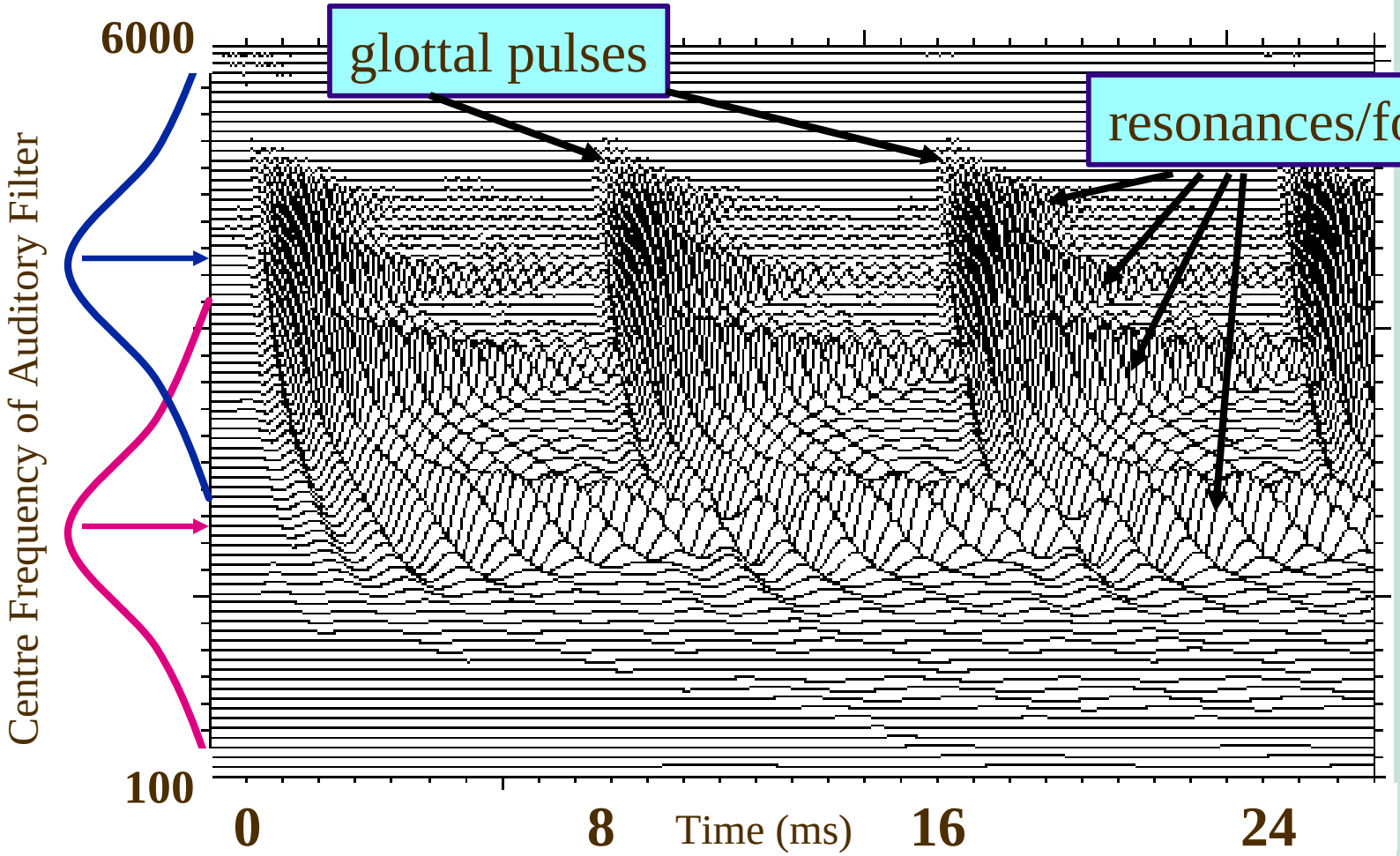
I: **Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)**

II: **Perform strobed temporal integration (STI) to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.**

[This STI process adapts the analysis to pulse rate.]

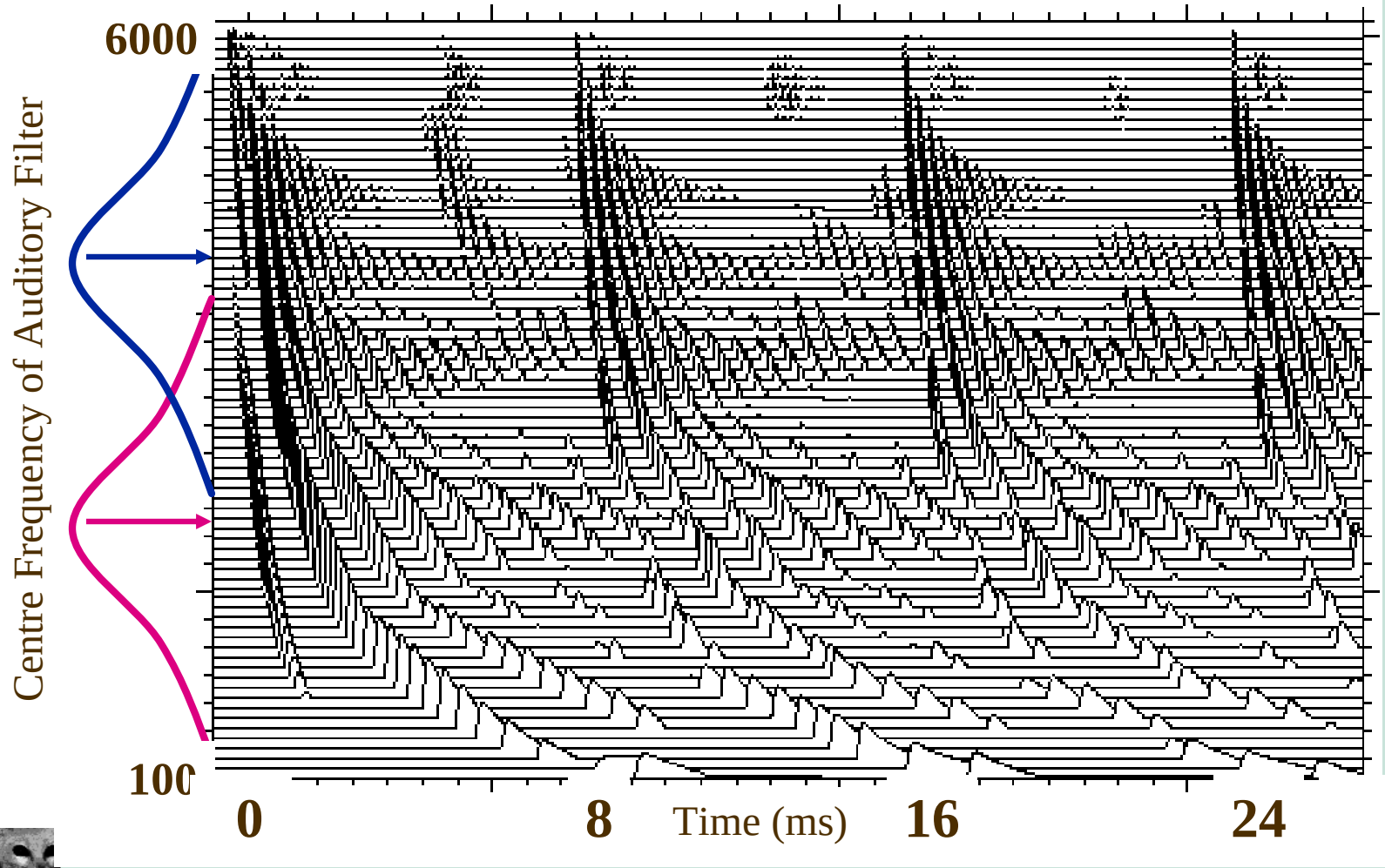
III: **Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message (or a scale invariant image of the message).**

basilar membrane motion in response to the /ae/ in ‘hat’

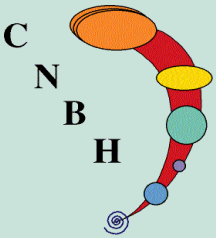


tonotopic ‘place’ in the co

Simulation of the neural activity pattern in response to four periods of the /ae/ in ‘hat’



tonotopic ‘place’ in the cochlea



The operations and their order:

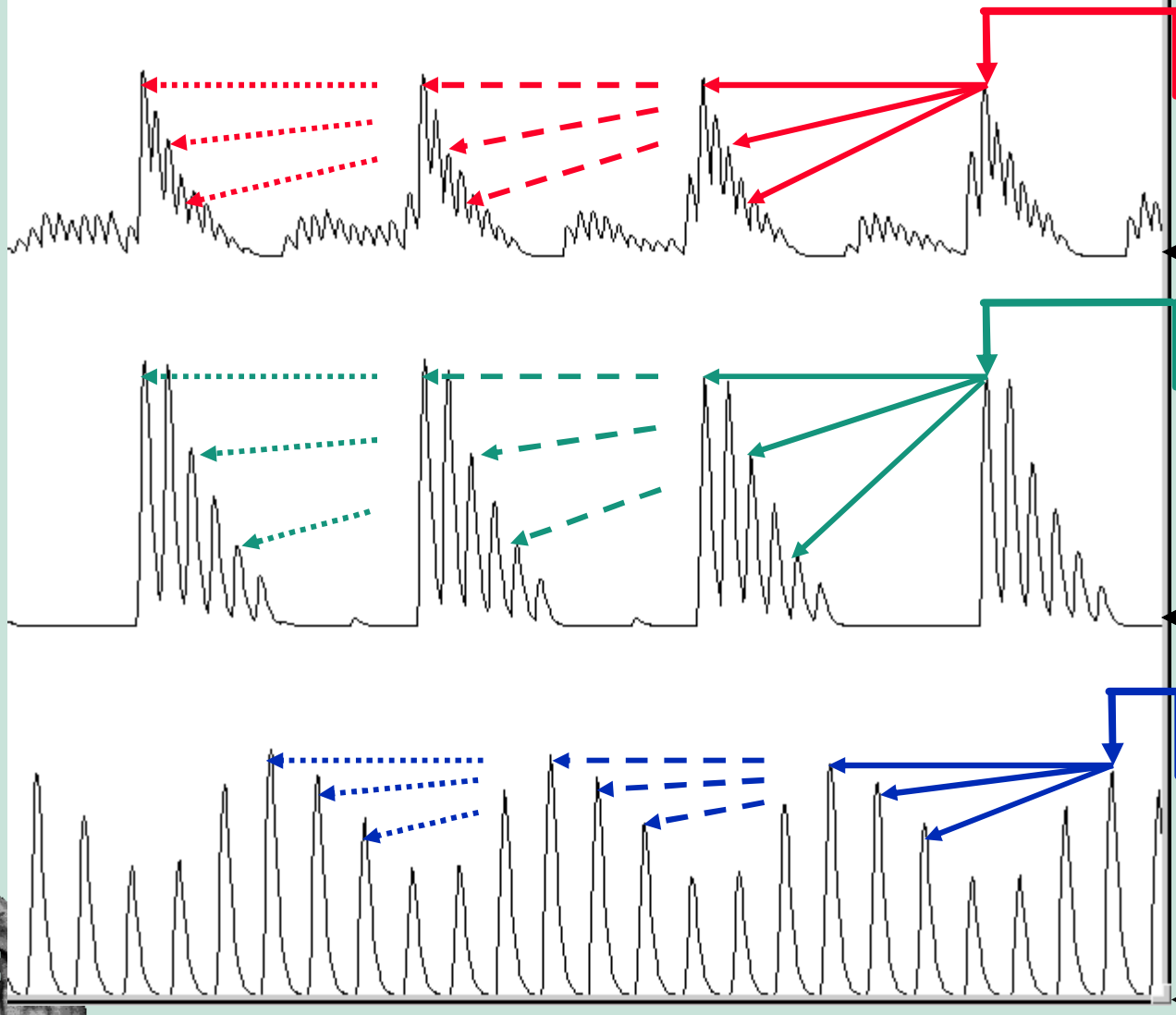
I: Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)

II: Perform strobed temporal integration (STI) to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.

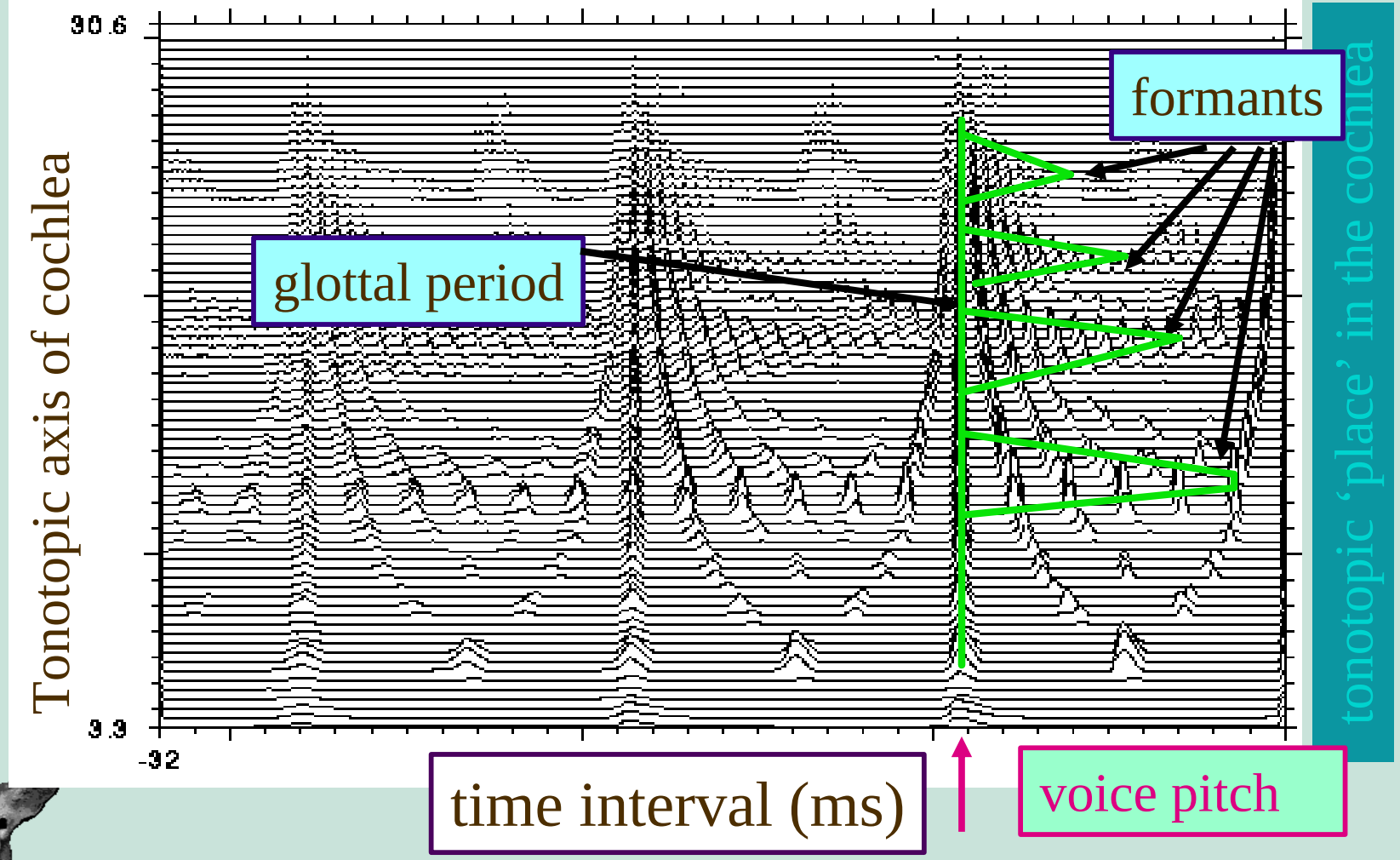
[This STI process adapts the analysis to pulse rate.]

III: Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message
(or a scale invariant image of the message).

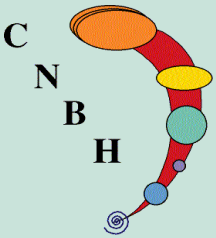
STI in three NAP channels



Simulation of the stabilised auditory image of the /ae/ in ‘hat’



Patterson et al. JASA (1995)



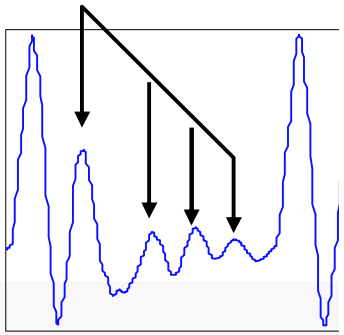
The operations and their order:

I: Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)

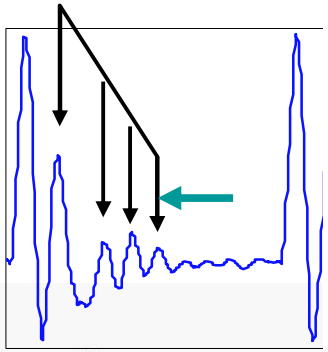
II: **Perform strobed temporal integration to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.**

[This STI process adapts the analysis to pulse rate.]

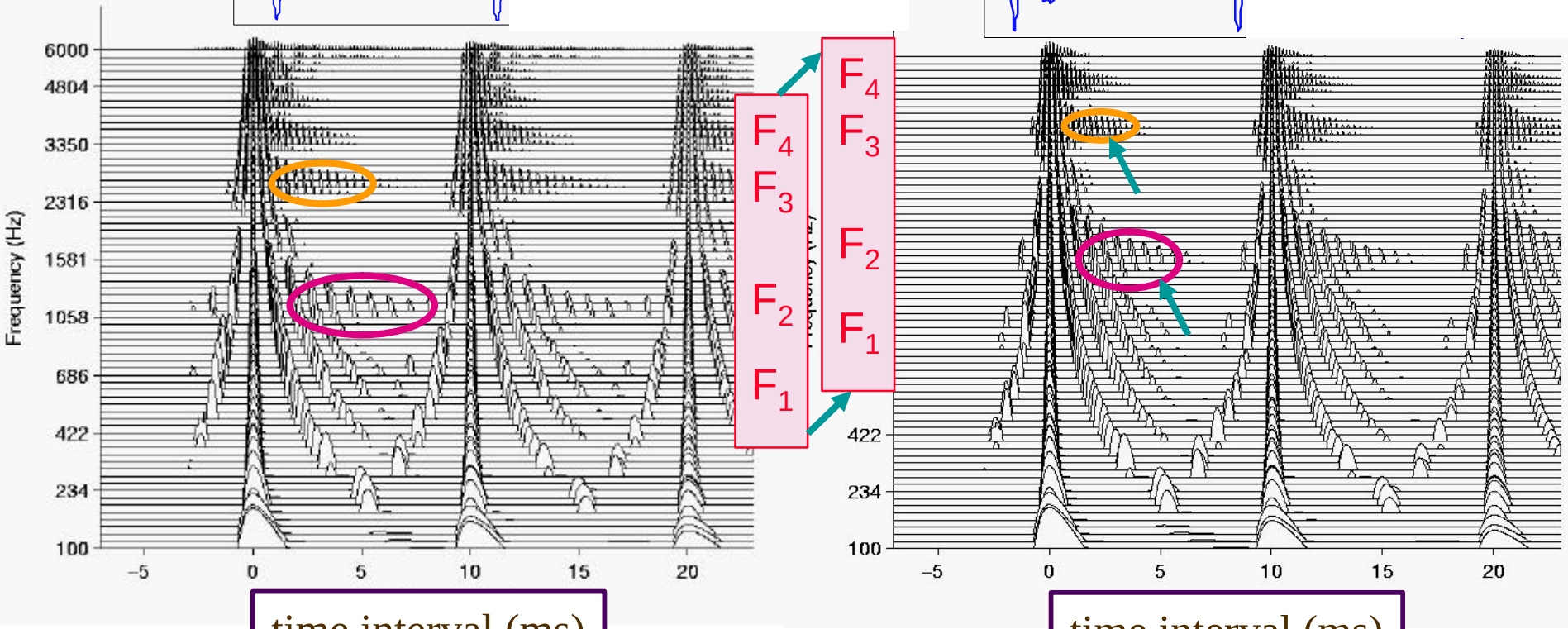
III: Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message (or a scale invariant image of the message).

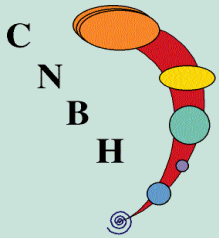


long VT



short VT





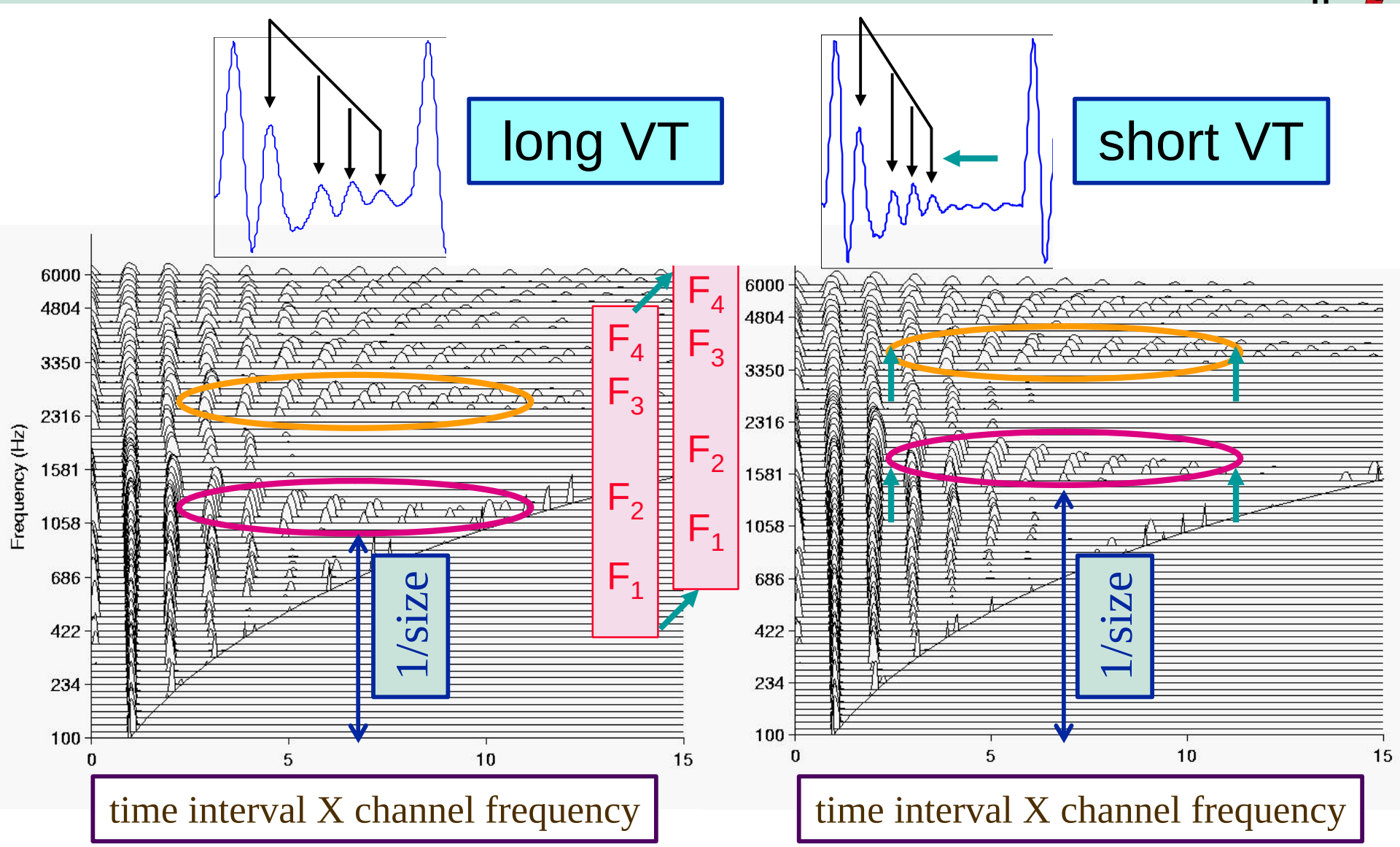
The operations and their order:

I: **Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)**

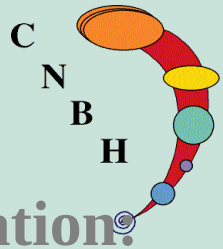
II: **Perform strobed temporal integration (STI) to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.**

[This STI process adapts the analysis to pulse rate.]

III: **Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message**
(or a scale invariant image of the message).



Irino and Patterson, Speech Communication (2002)



The operations and their order:

I: Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)

II: Perform strobed temporal integration to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.

[This STI process adapts the analysis to pulse rate.]

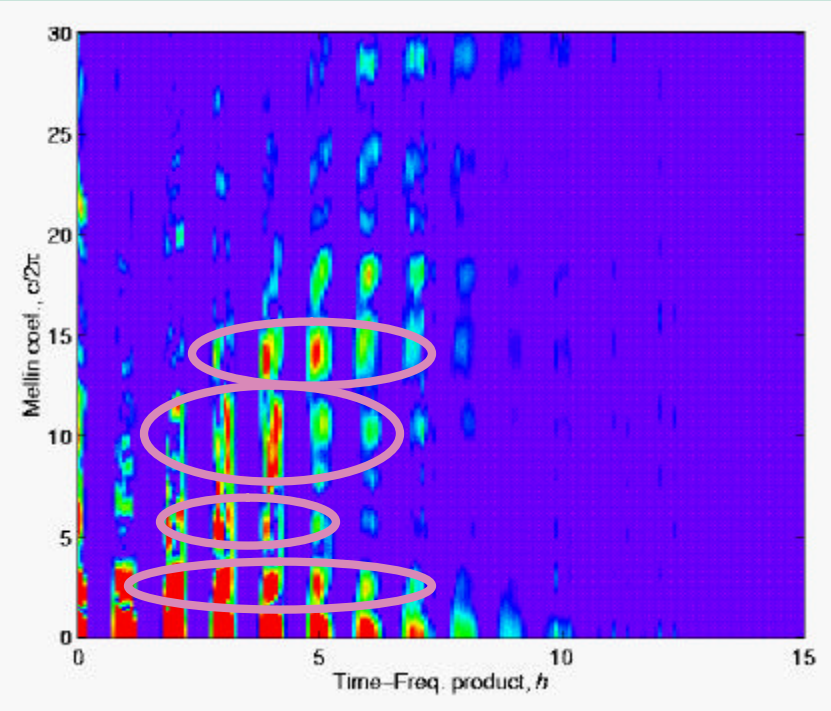
III: Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message
(or a scale invariant image of the message).

The message within these 2-D images is pulse-rate adapted and the period attached to the pulse is resonance-scale invariant.

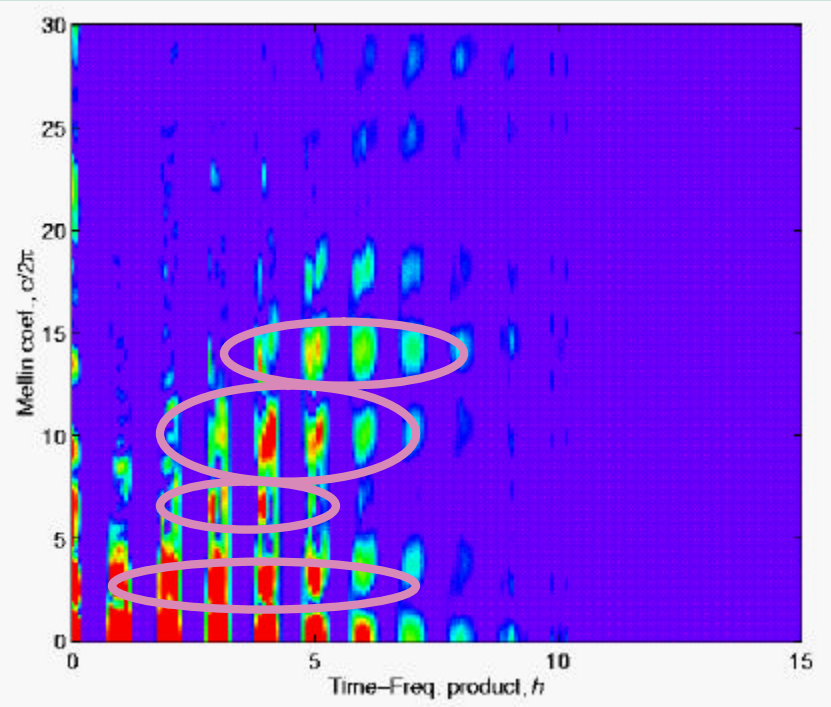
The message in a sequence of images produced, say, by an animal syllable is time shift invariant as a whole. That is, if you repeat the sound again later you get the same set of images.

Mellin Image of vowel ‘a’

‘a_m’

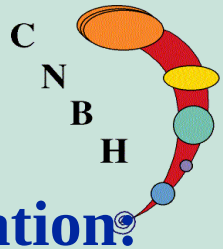


‘a_f’



- The two ‘a’s produce virtually the same image.

Irino and Patterson, Speech Communication (2002)



The operations and their order:

I: **Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)**

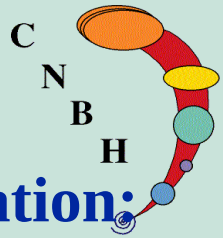
II: **Perform strobed temporal integration (STI) to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.**

[This STI process adapts the analysis to pulse rate.]

III: **Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message (or a scale invariant image of the message).**

The message within these 2-D images is pulse-rate adapted and the period attached to the pulse is resonance-scale invariant.

The message in a sequence of images produced, say, by an animal syllable is time shift invariant as a whole. That is, if you repeat the sound again later you get the same set of images.



Conclusions:

I: Perform a spectral analysis that preserves scale information: a wavelet transform (log-frequency with proportional BW)

II: Perform strobed temporal integration to create a representation with a *time-interval dimension* that preserves the scale information of messages less than 30 ms in duration, and which defines zero for the scale transform.

III: Apply a warping operator to produce a (pulse-rate adapted) scale-covariant image of the message (or a scale invariant image of the message).

The auditory system integrates these processes with its binaural analysis and its primitive channel grouping process.

The five processes are applied together as a group to identify the existence of sources in the auditory scene, before the messages of the sources are analysed by more central pattern recognition processes.