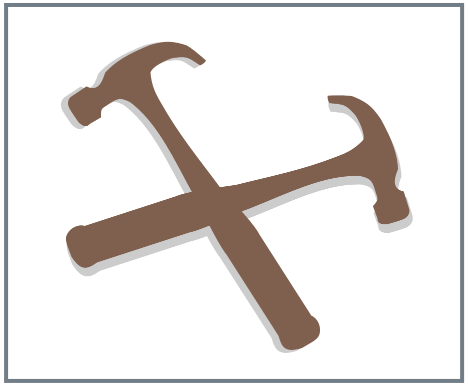
Harvard-MIT Division of Health Sciences and Technology
HST.723: Neural Coding and Perception of Sound
Instructor: Christophe Micheyl

Auditory scene analysis

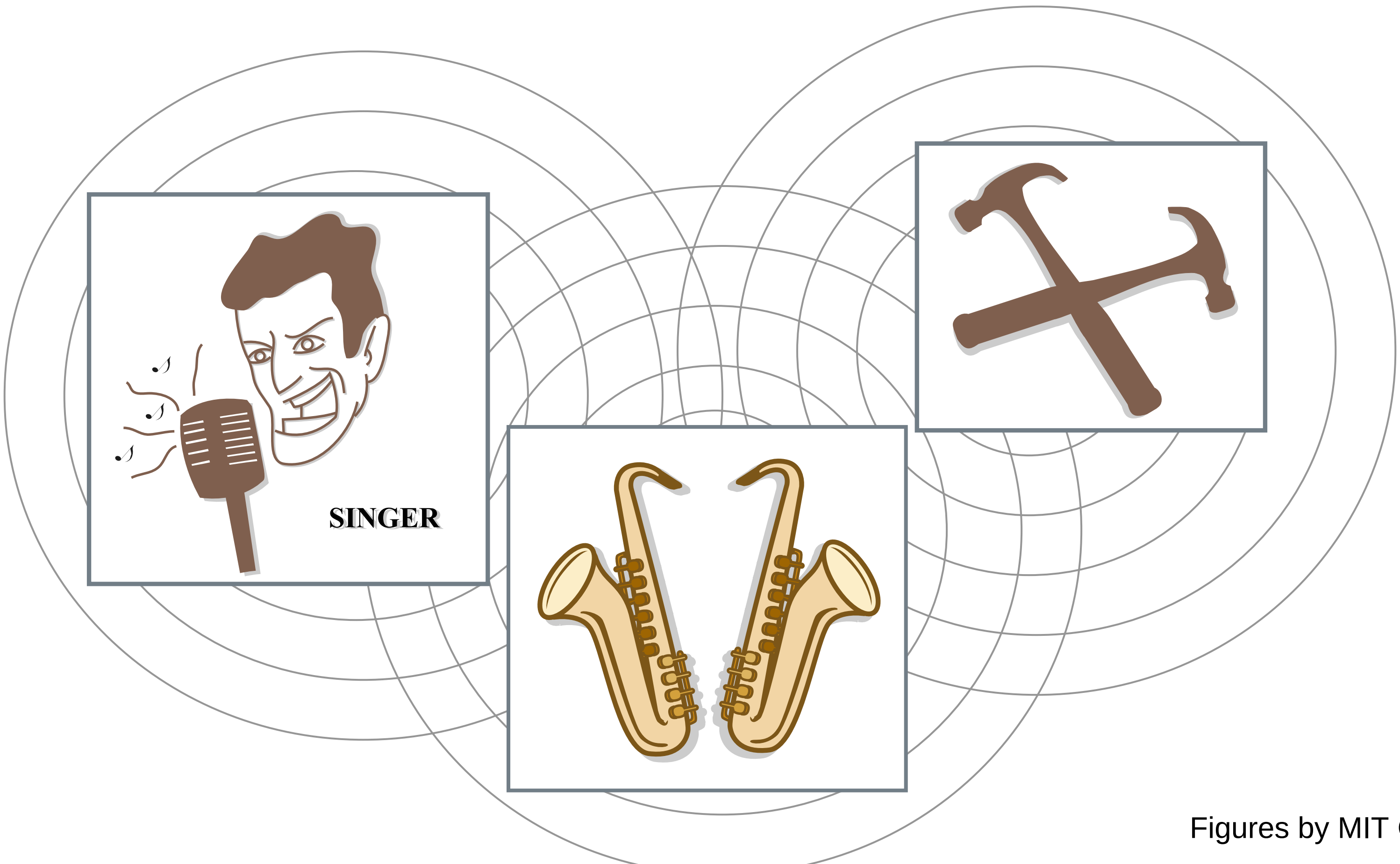
Christophe Micheyl

We are often surrounded by various sound sources.
Some of importance to us; others, a nuisance.



Figures by MIT OCW.

The waves from these sources mingle before reaching our ears.



Figures by MIT OCW.

The result is a complex acoustic mixture.

Figures removed due
to copyright reasons.

The auditory system must disentangle the mixture to permit (or at least facilitate) source identification

Figures removed due
to copyright reasons.

Solution:

Figures removed due
to copyright reasons.

Some of the questions that we will address:

- What tricks does the auditory system use to analyze complex scenes?

- What neural/brain processes subtend these perceptual phenomena?

- Why do hearing-impaired listeners have listening difficulties in the presence of multiple sound sources?

Why is this important?

-Understand how the auditory system works in ‘real-life’

(the system was probably not designed primarily to process isolated sounds)

-Build artificial sound-processing systems that can do ASA like us…

(speaker separation for speech recognition, instrument separation for music transcription, content-based indexing in audio recordings,…)

-… or help us do it better

(sound pre-processing for ‘intelligent’ hearing aids, enhanced speech-in-noise understanding,…)

Bottom-up and top-down mechanisms

- **Bottom-up (or 'primitive') mechanisms**

-partition the sensory input based on simple stimulus properties

-largely automatic (pre-attentive)

-probably innate or acquired early during infancy

- **Top-down (or 'schema-based') mechanisms**

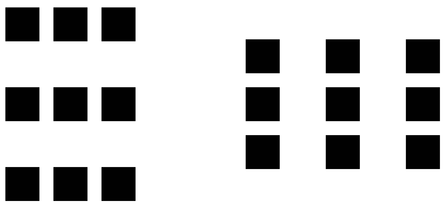
-partition the input based on stored object representations (prototypes)

-heavily dependent upon experience/knowledge

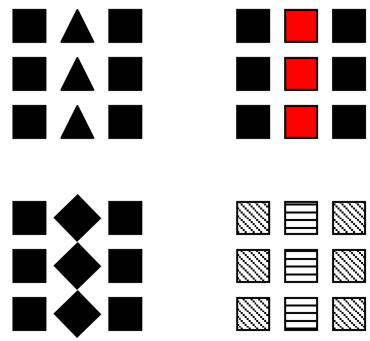
The basic laws of perceptual organization

courtesy of: the Gestalt-psychology school

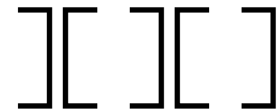
promixity



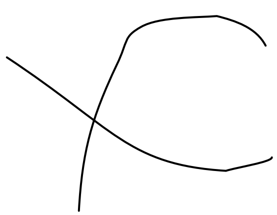
similarity



closure



continuity



etc…

Top-down

Figure removed due
to copyright reasons.

Sequential and simultaneous mechanisms

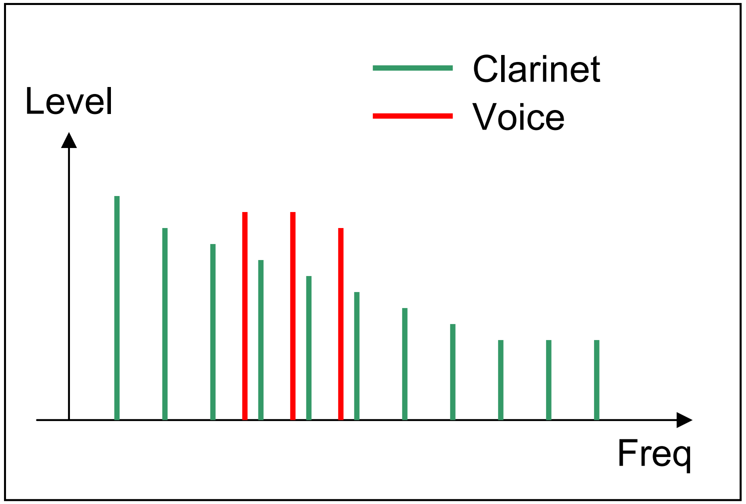
Sequential mechanisms
(auditory ‘streaming’)

Figures removed due
to copyright reasons.

Sequential and simultaneous mechanisms

Simultaneous mechanisms

Figure removed due to copyright reasons.



Outline

I. Simultaneous ASA processes

- Harmonicity
- Onset/offset
- Co-modulation

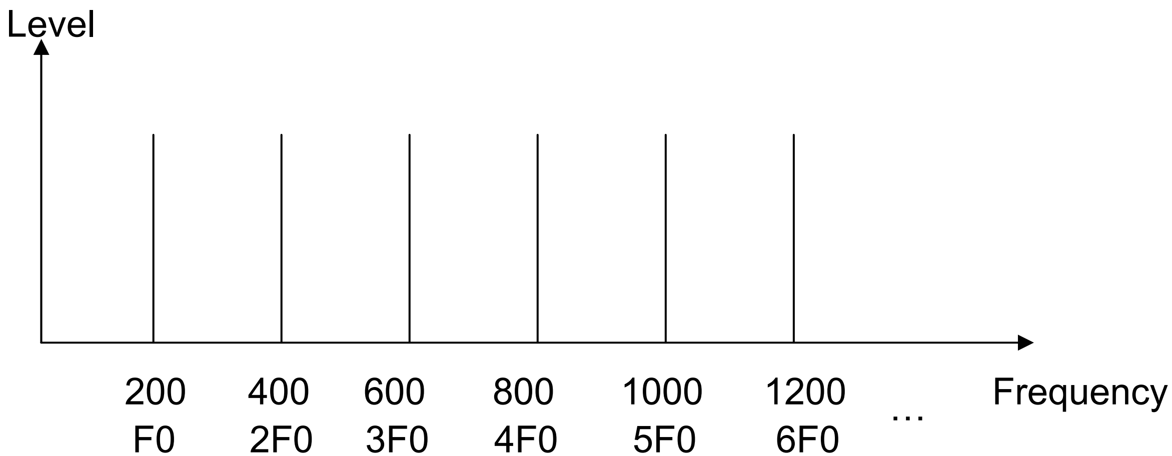
II. Sequential ASA processes

- Auditory streaming

Harmonicity

Many important sounds are harmonic

(vowels of speech, most musical sounds, animal calls,…)

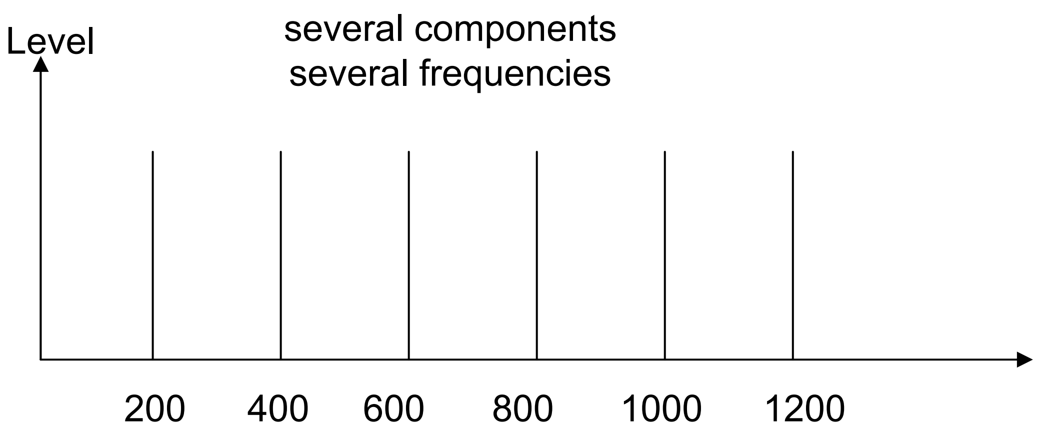


Does the auditory system exploit this physical property to group/segregate frequency components?

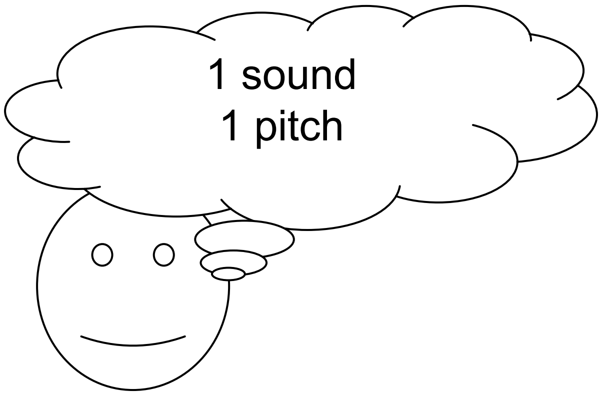
Harmonic fusion

Harmonic complexes are generally perceived as one sound

stimulus



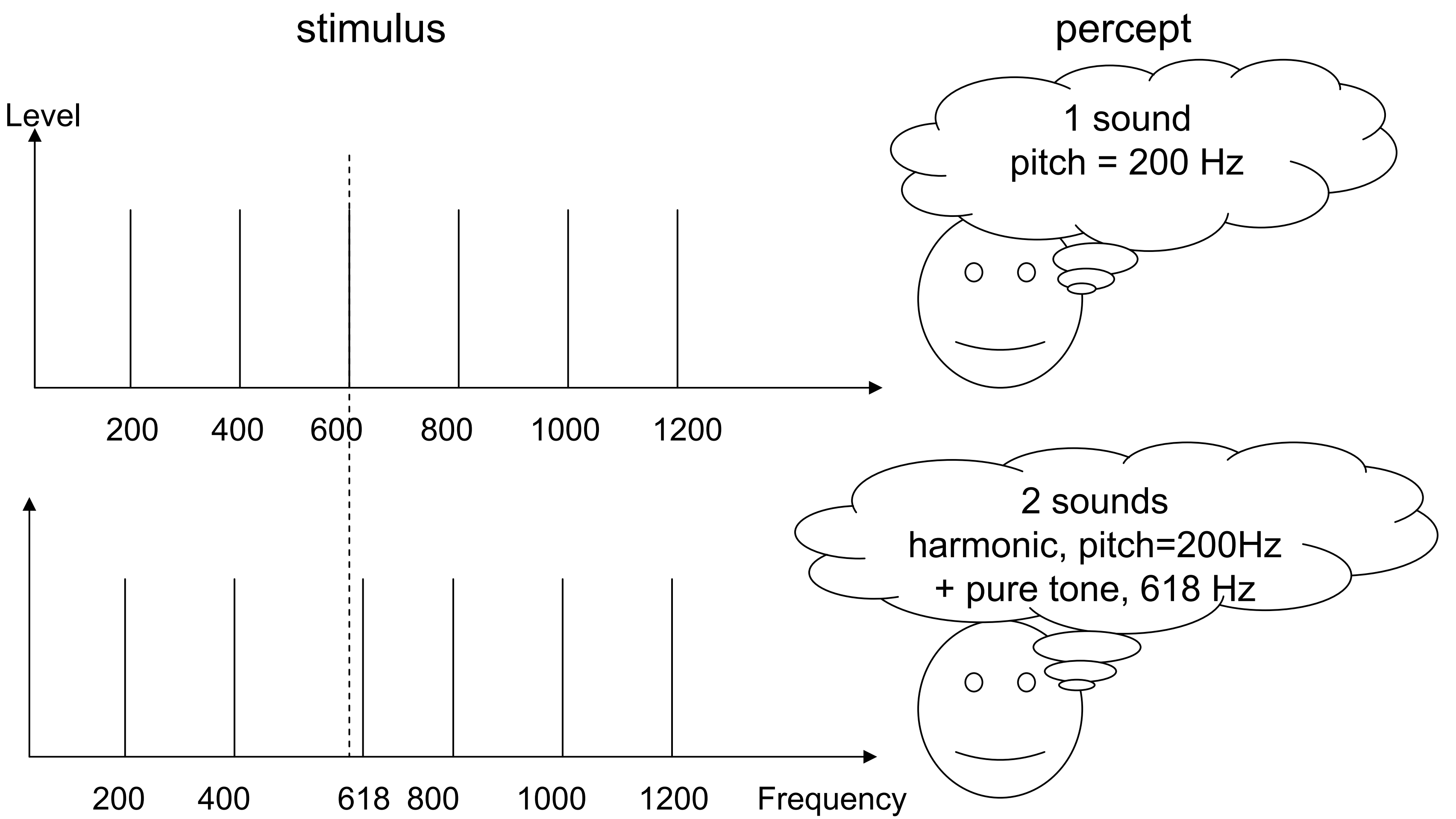
percept



Deviations from harmonicity promote segregation

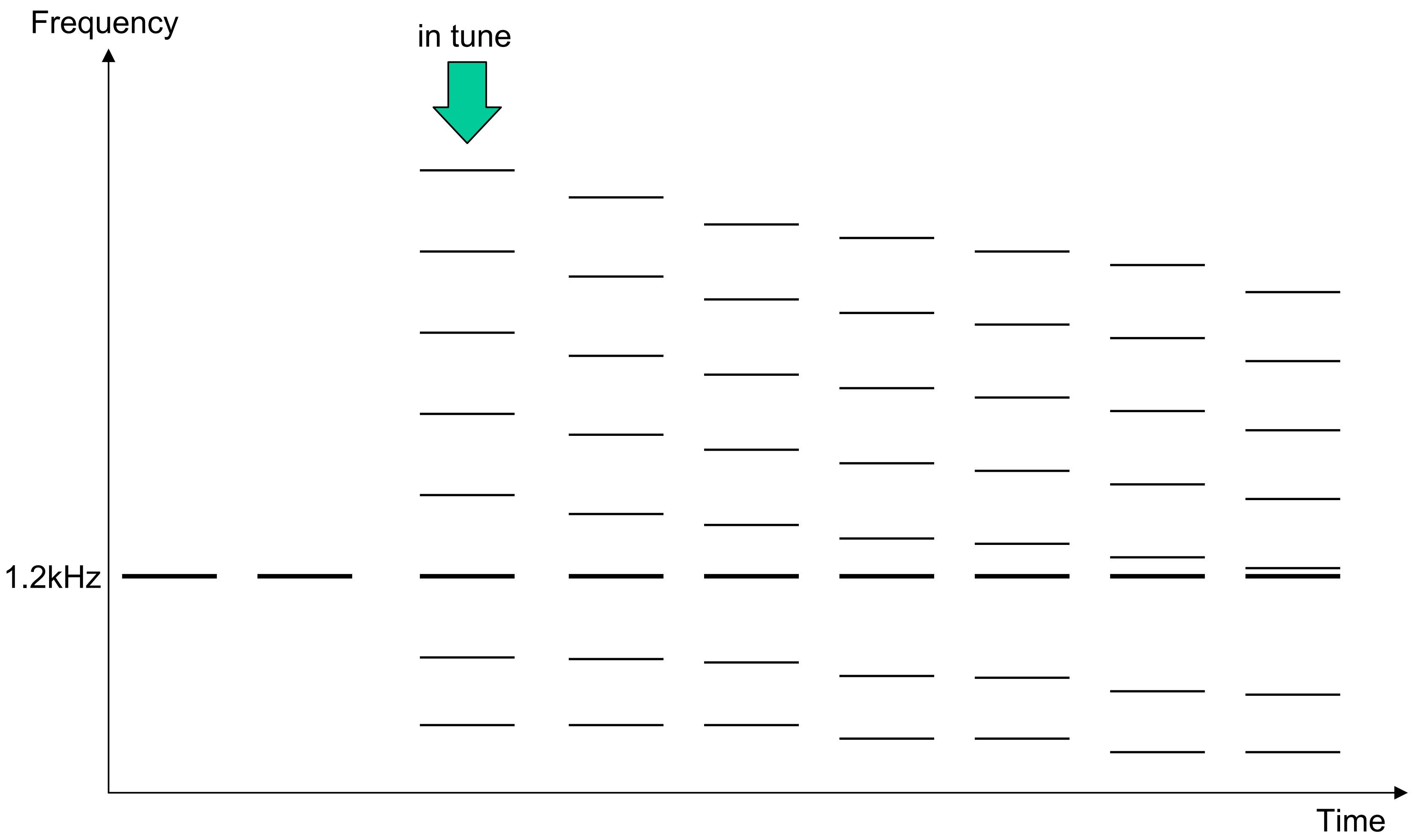
If a harmonic is mistuned by > 2-3%, it stands out perceptually

(Moore et al., 1985, 1986; Hartmann *et al.*, 1990)



Demonstration

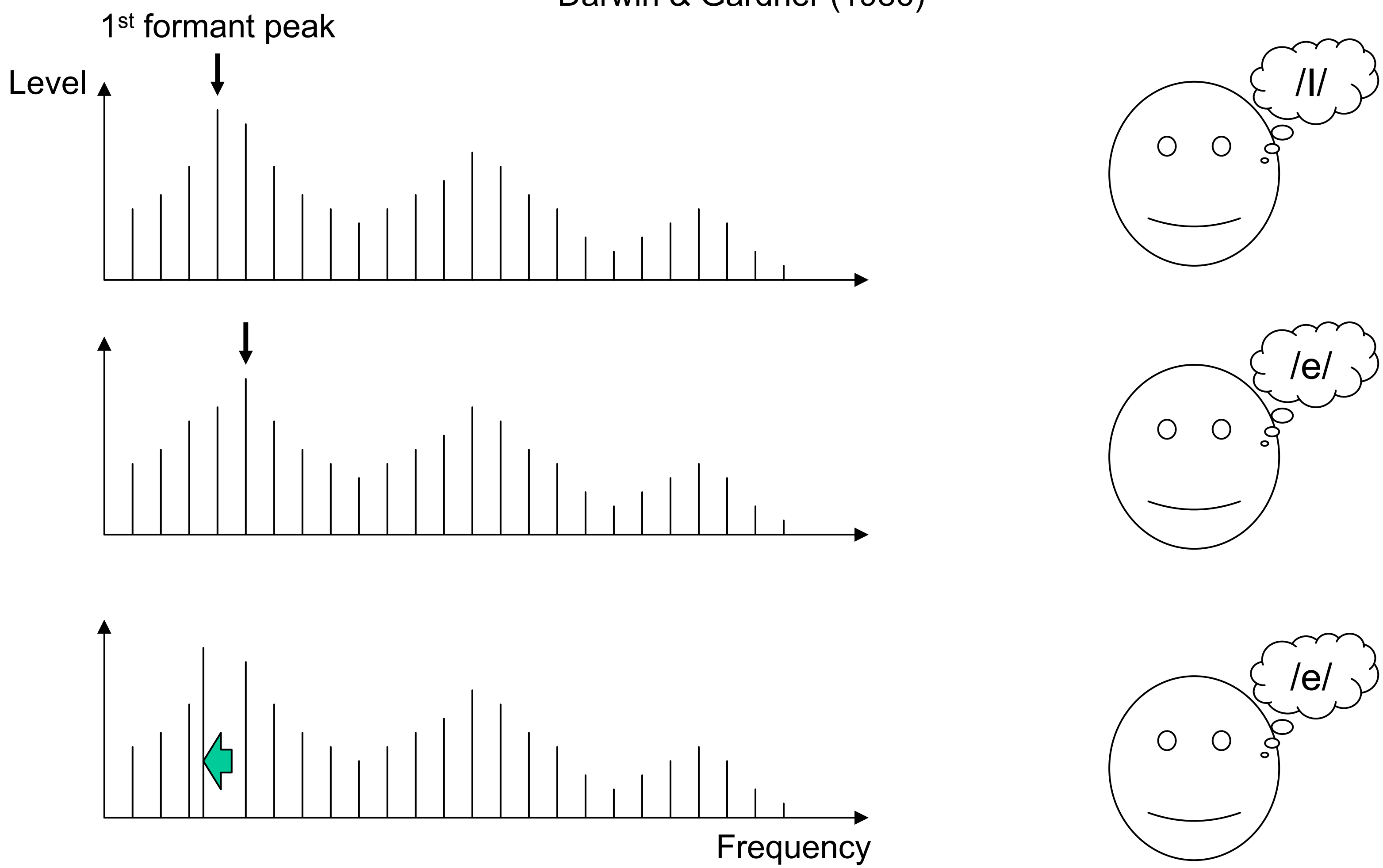
From:
Bregman (1990)
Auditory scene analysis
MIT Press
Demo CD



Influence of harmonic grouping/segregation on other aspects of auditory perception

Mistuning a harmonic near a formant can affect the perceived identity of a vowel

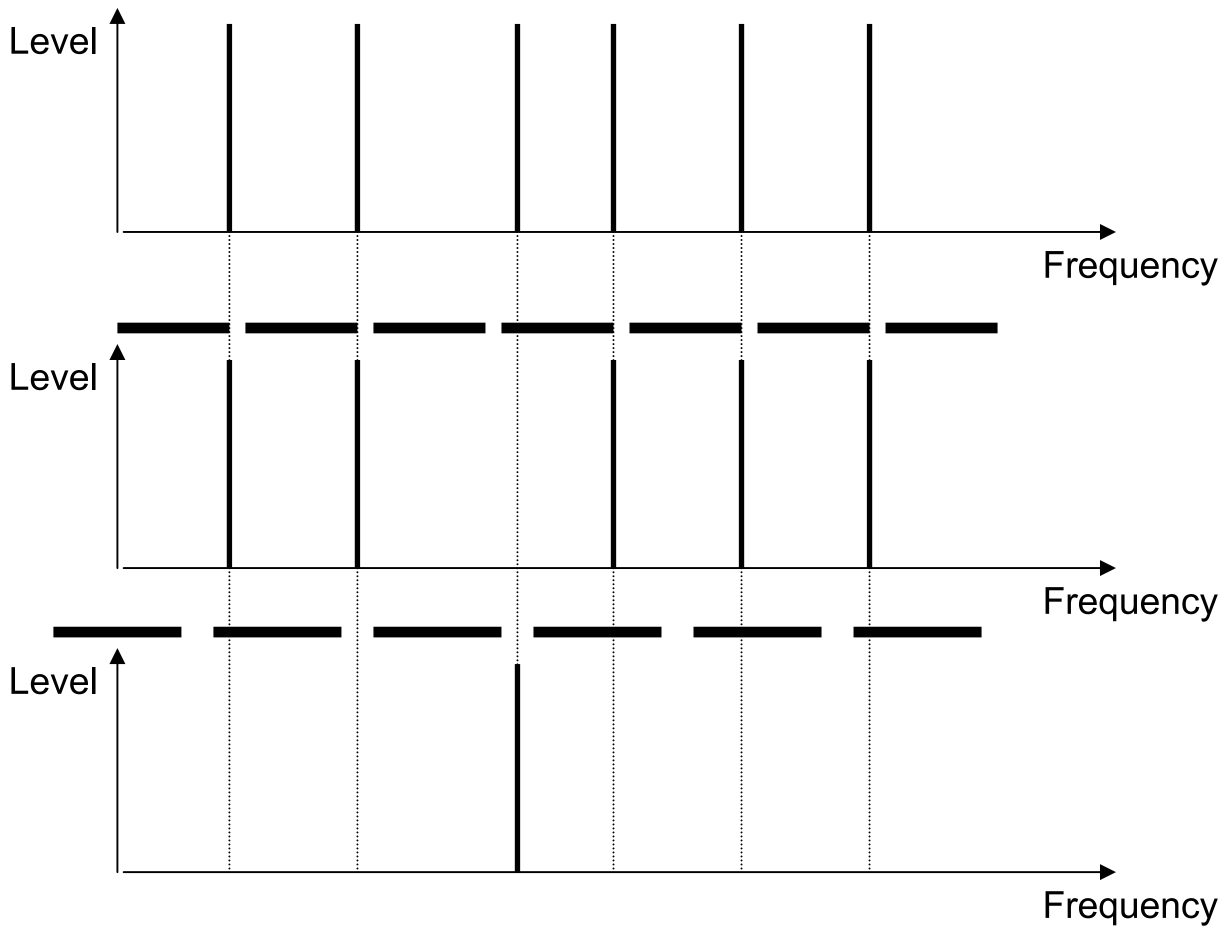
Darwin & Gardner (1986)



Mechanisms of harmonicity-based grouping?

Spectral: the harmonic sieve (Duifhuis et al., 1982)

Components that pass through the sieve are grouped; those that don't are excluded

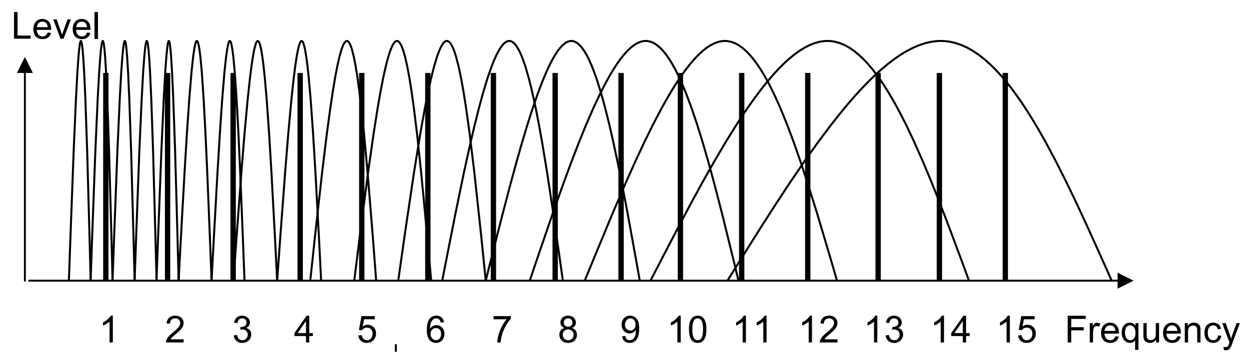


Actual mechanisms of harmonicity-based grouping?

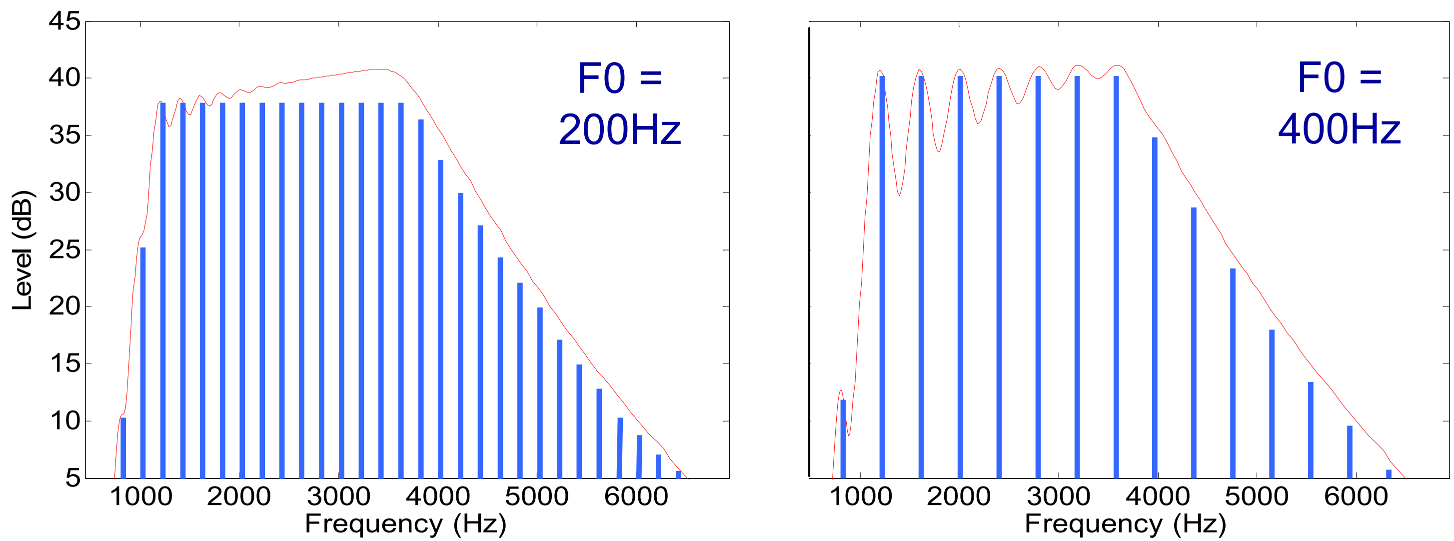
Harmonics above the 10th can generally not be heard out (Moore *et al.*, 1985)

This suggests a role of peripheral frequency selectivity, because harmonics above the 10th are generally unresolved in the cochlea:

The cochlea as a filter bank



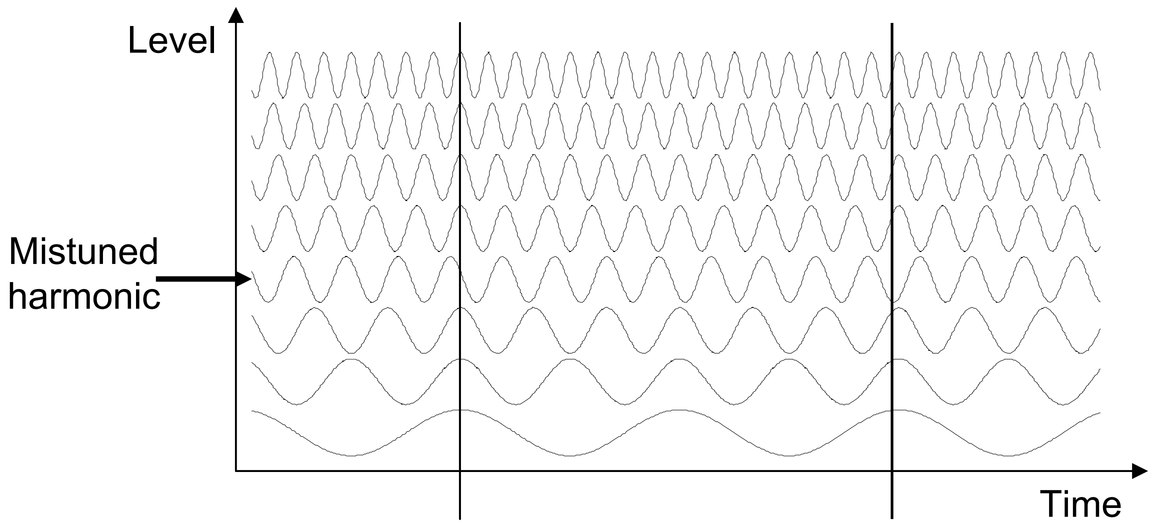
Simulated Spectral EPs:



Mechanisms of harmonicity-based grouping?

Temporal: across-channel synchrony (Roberts & Brunstrom, 2001)

Components that elicit synchronous neural responses are grouped



Above 2000 Hz, harmonics become increasingly harder to hear out

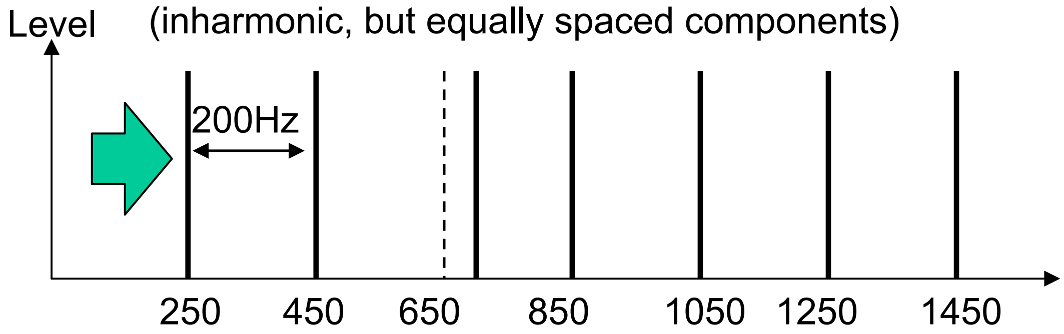
(Hartmann *et al.*, 1990)

This suggests a contribution of temporal mechanisms, because phase locking breaks down at high frequencies

An aside: harmonicity or equal spectral spacing?

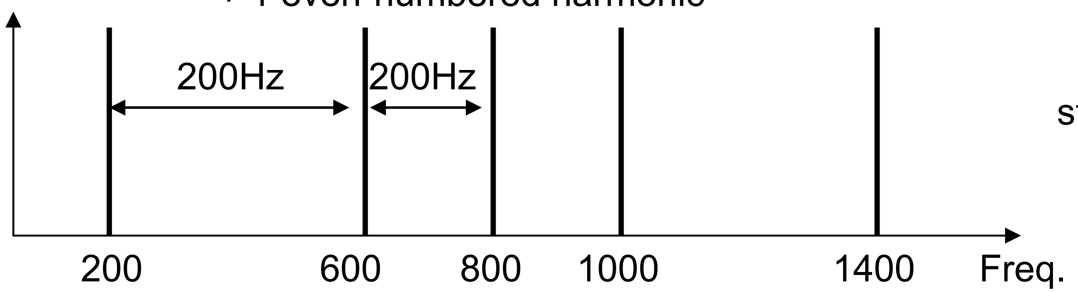
Grouping/segregation of spectral components is based not solely on harmonicity, but also on spectral spacing Roberts & Bregman (1991)

Frequency-shifted complex
(inharmonic, but equally spaced components)



Shifting the frequency of a component in a shifted complex makes it stand out

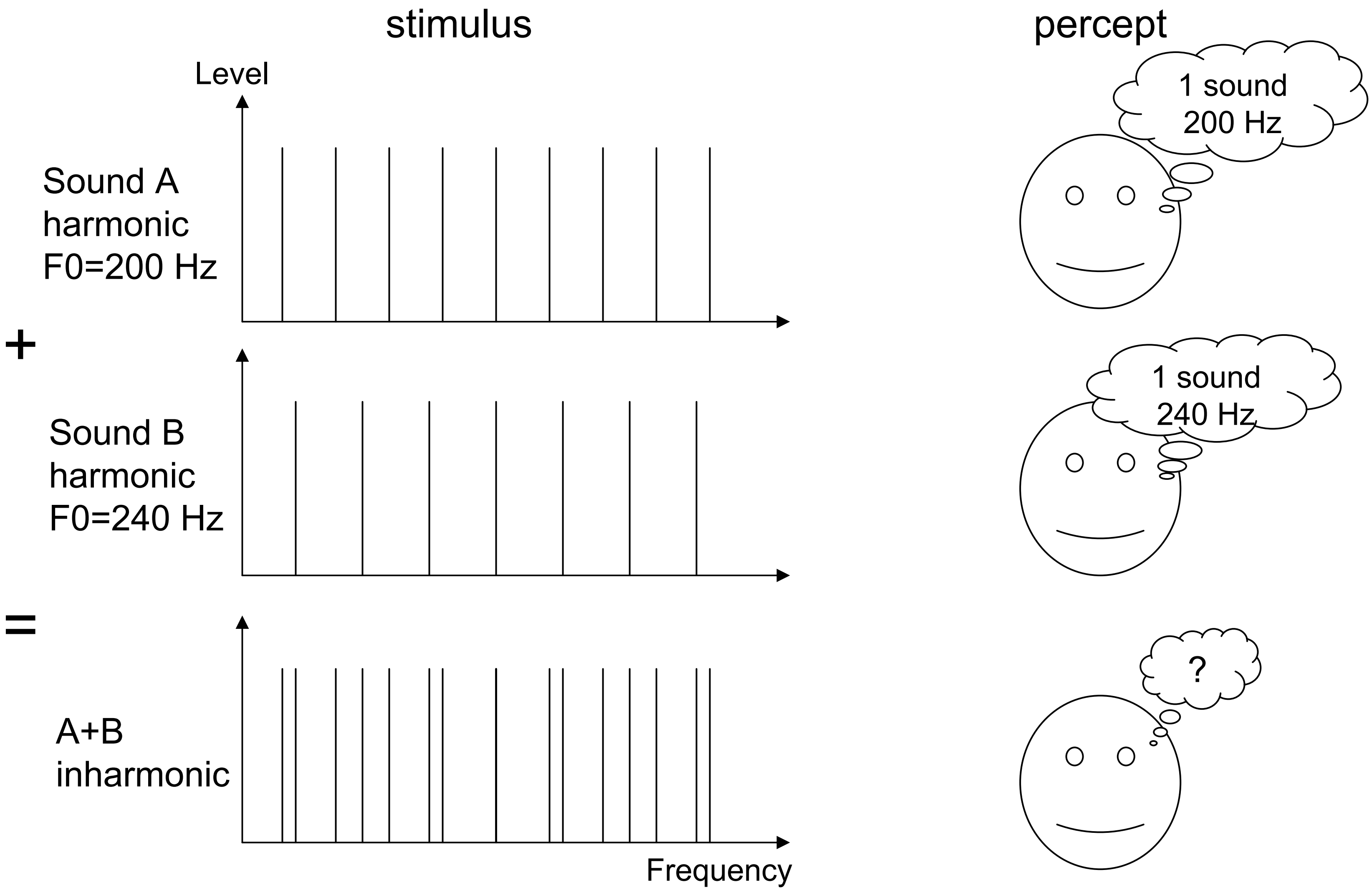
Odd-numbered harmonics
+ 1 even-numbered harmonic



The even-numbered harmonic stands out more than the neighboring odd-numbered harmonics

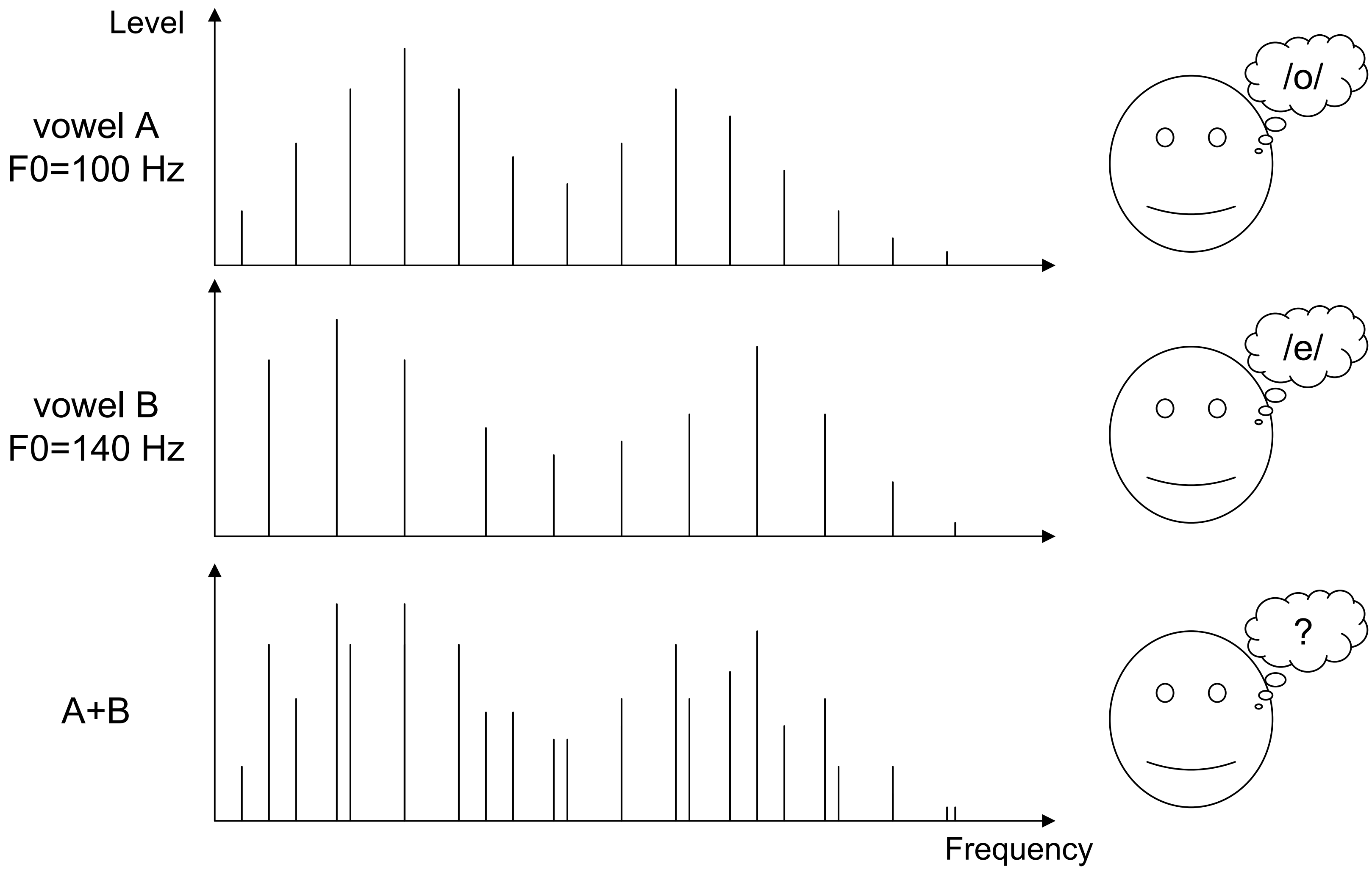
But the utility of a specific spectral-spacing-based grouping mechanism is questionable

F0-based segregation of whole harmonic complexes



Double vowels

Two (synthetic) vowels with different F0s played simultaneously

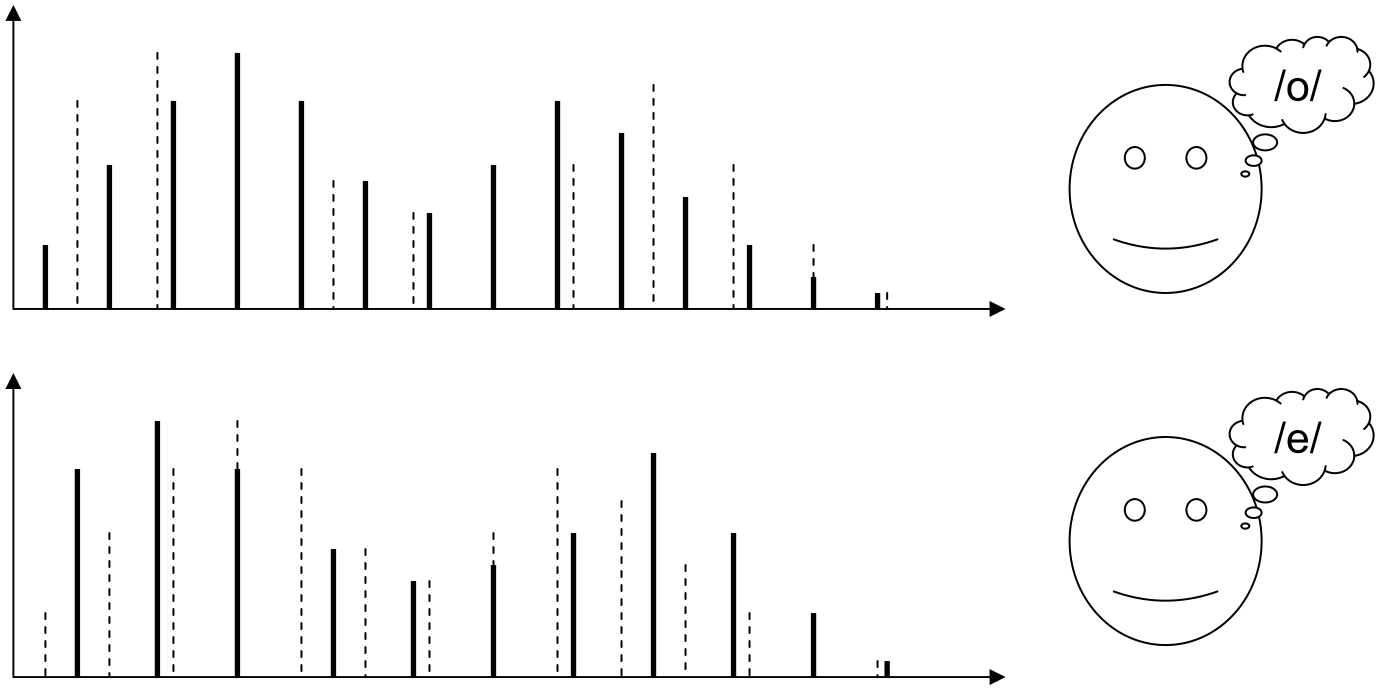


Double vowels

Can listeners use F0 differences to sort out the frequency components?

vowel A
F0=100 Hz

vowel B
F0=140 Hz



——— harmonics corresponding to one F0

- - - - - - - - harmonics corresponding to the other F0

Concurrent vowels

F0 differences facilitate the identification of concurrent vowels

(Scheffers, 1983; Assmann & Summerfield, 1990; …)

Figure removed due to copyright reasons. Please see: Assmann, and Summerfield. *J. Acoust. Soc. Am.* 88 (1990): 680-687.

(but note %-correct well above chance even with no F0 difference, providing evidence for a role of template-based mechanisms)

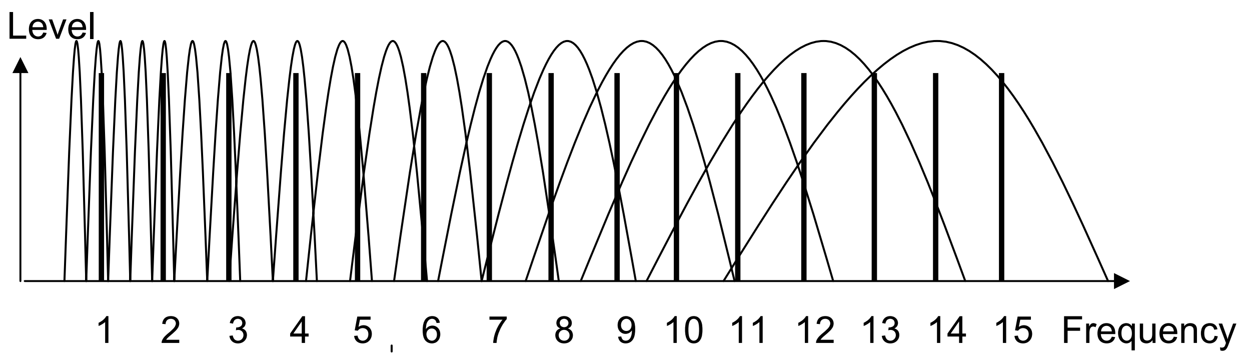
This also works with whole sentences (Brokx & Nooteboom, 1982)

Actual mechanisms of harmonicity-based grouping?

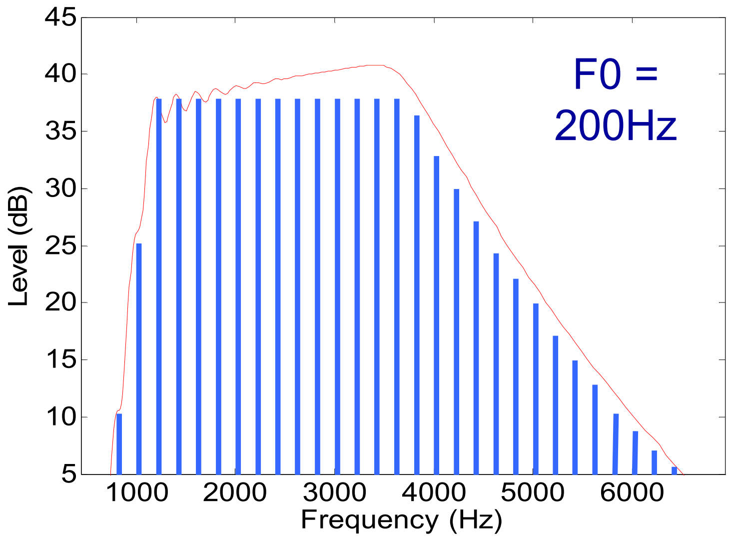
Harmonics above the 10th can generally not be heard out (Moore *et al.*, 1985)

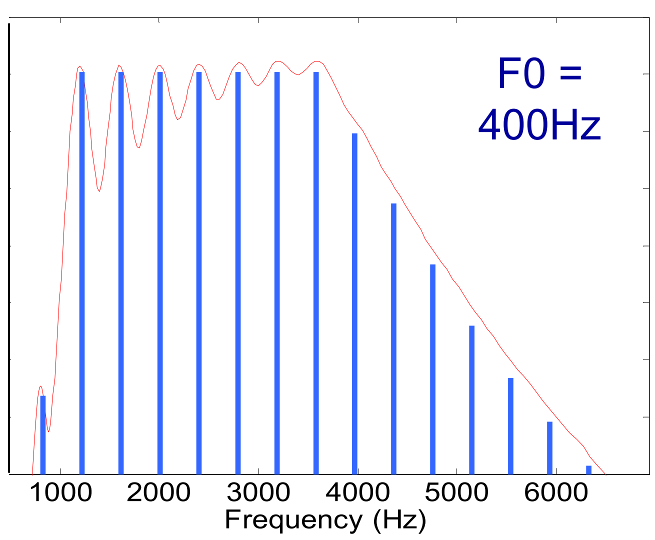
This suggests a role of peripheral frequency selectivity, because harmonics above the 10th are generally unresolved in the cochlea:

The cochlea as a filter bank



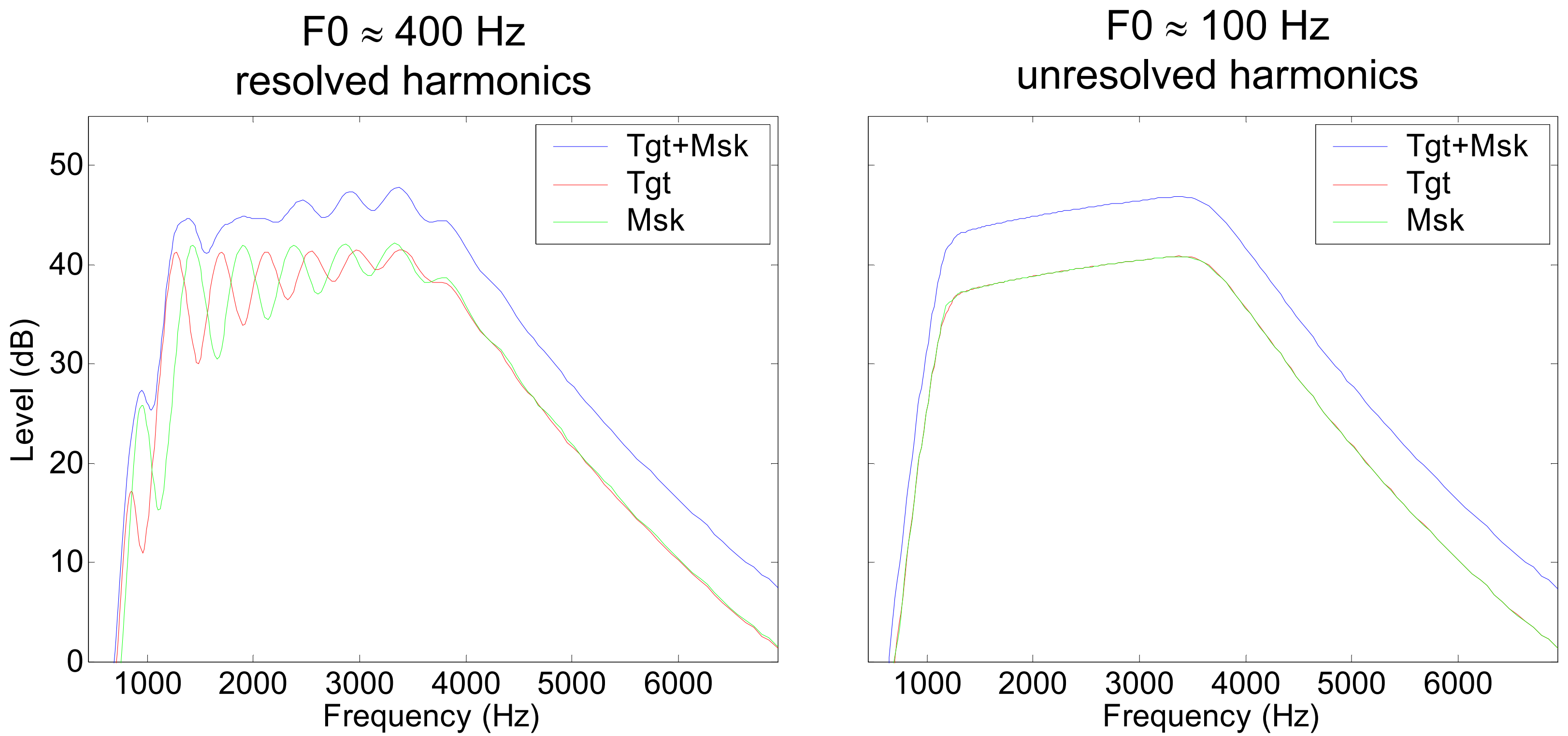
Simulated Spectral EPs:





Influence of frequency resolution on the F0-based segregation of concurrent complexes

Example simulated spectral excitation patterns
in response to harmonic complex target, maskers, and target+masker mixtures
at different F0s

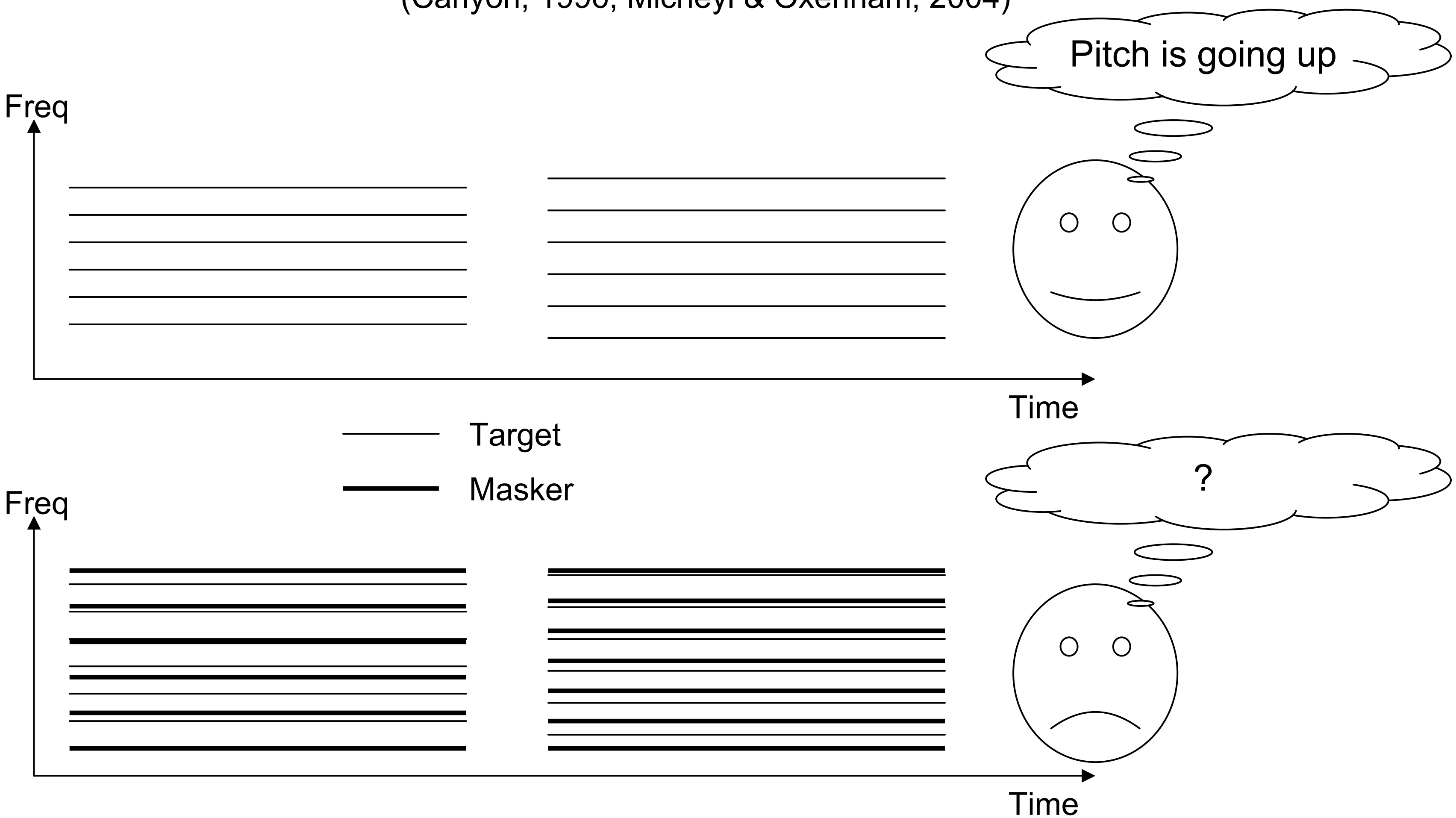


resulting EP displays
some peaks

resulting EP displays
no peaks

Influence of frequency resolution on the F0-based segregation of concurrent complexes

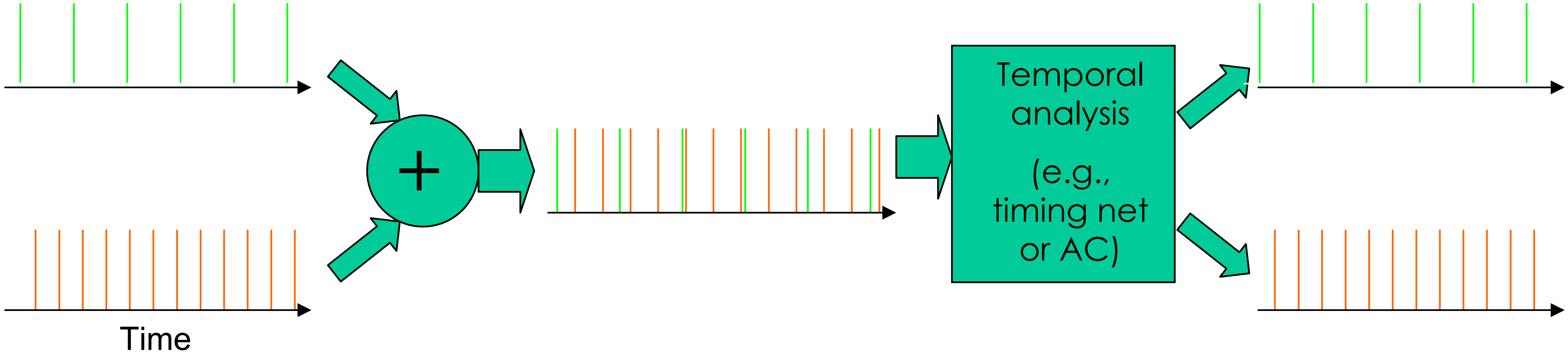
(Carlyon, 1996; Micheyl & Oxenham, 2004)



F0-based segregation does not work if all frequency components are unresolved

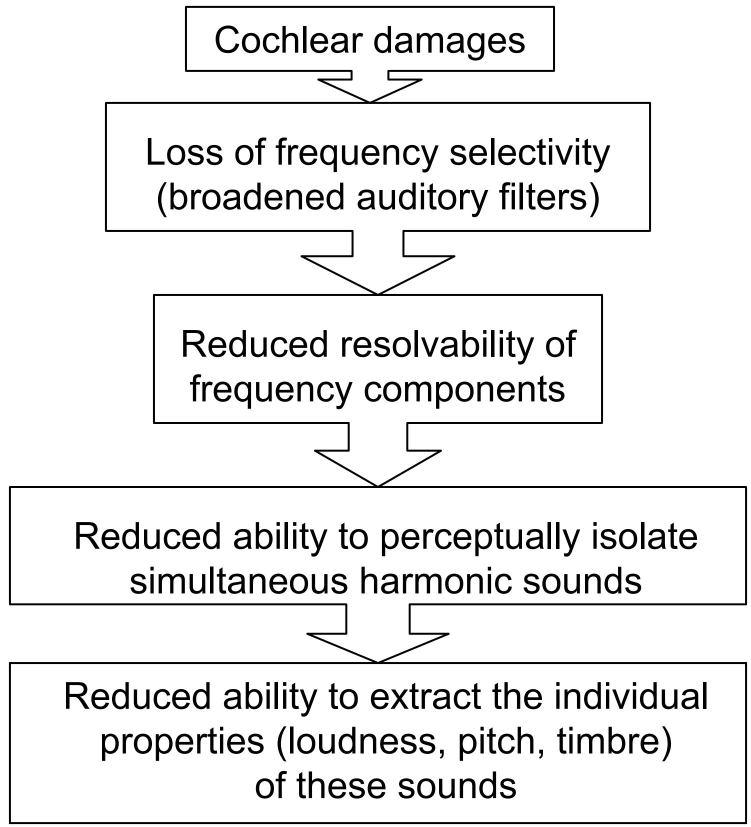
Influence of frequency resolution on the F0-based segregation of concurrent complexes

Yet, in principle, it is possible to segregate two periodic components falling into the same peripheral auditory filter using some temporal mechanism
(harmonic cancellation model, de Cheveigné et al., 1992; timing nets, Cariani, 2001)



Our results (Micheyl & Oxenham, 2004) and those of Carlyon (1996) indicate that the auditory system makes very limited (if any) use of this temporal strategy for segregating simultaneous harmonic complexes

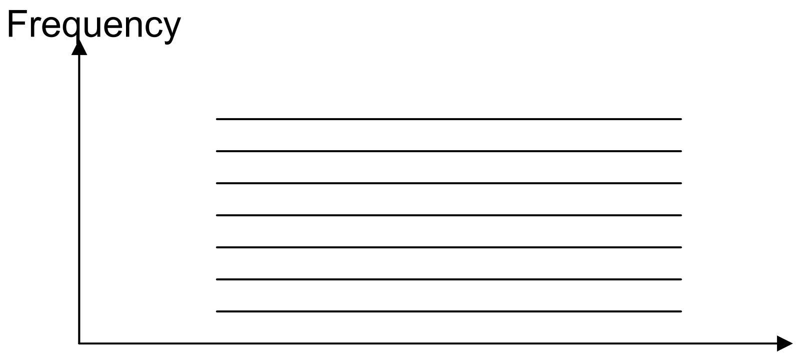
Implications for hearing-impaired listeners



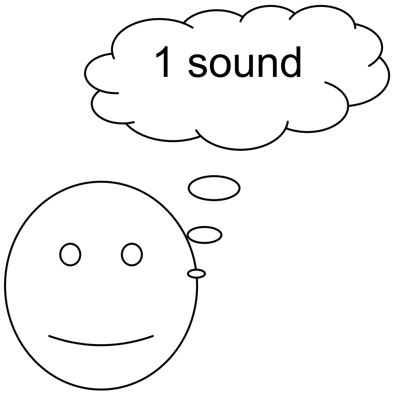
Onset time

Frequency components that start together tend to fuse together

stimulus

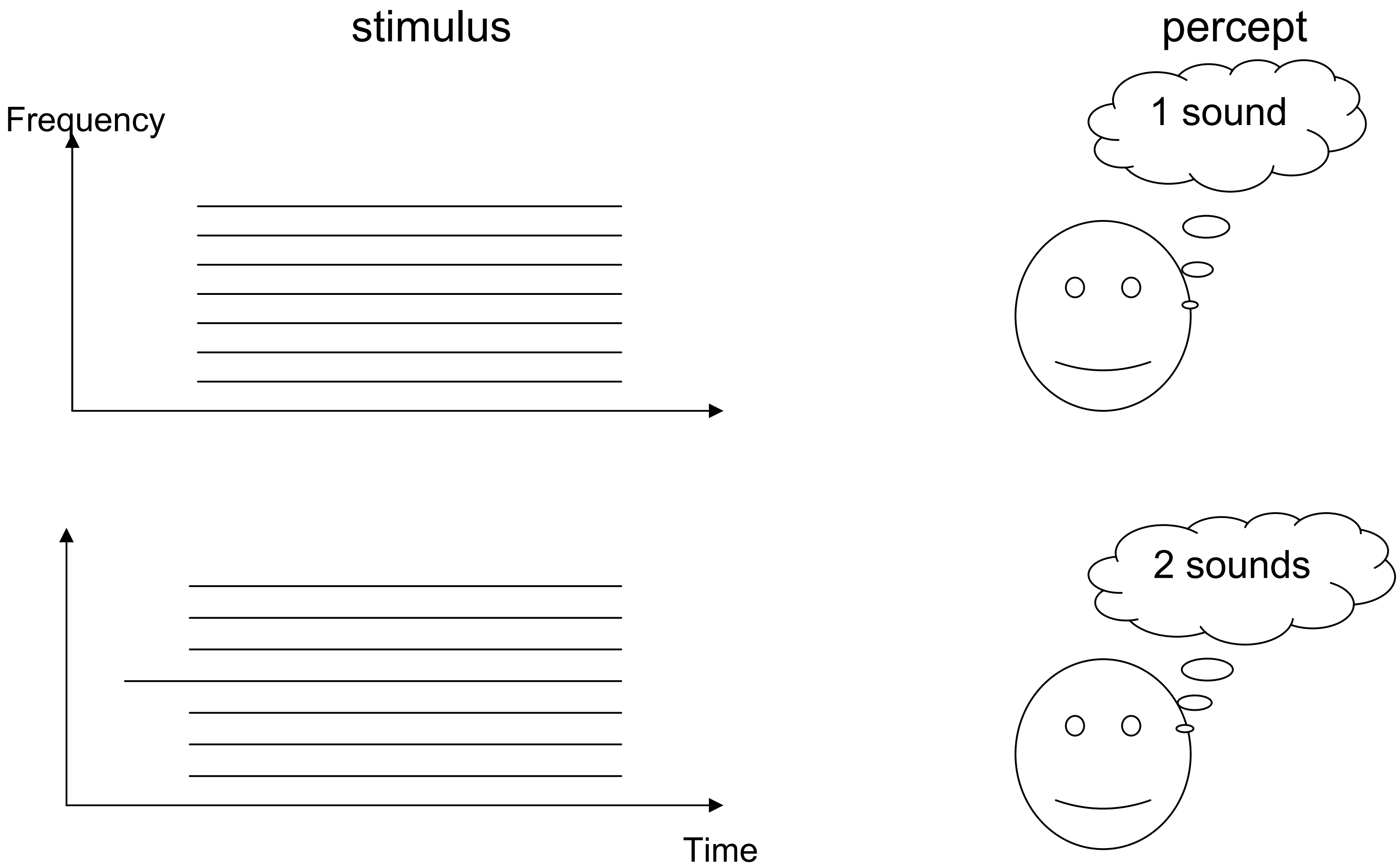


percept



Onset time

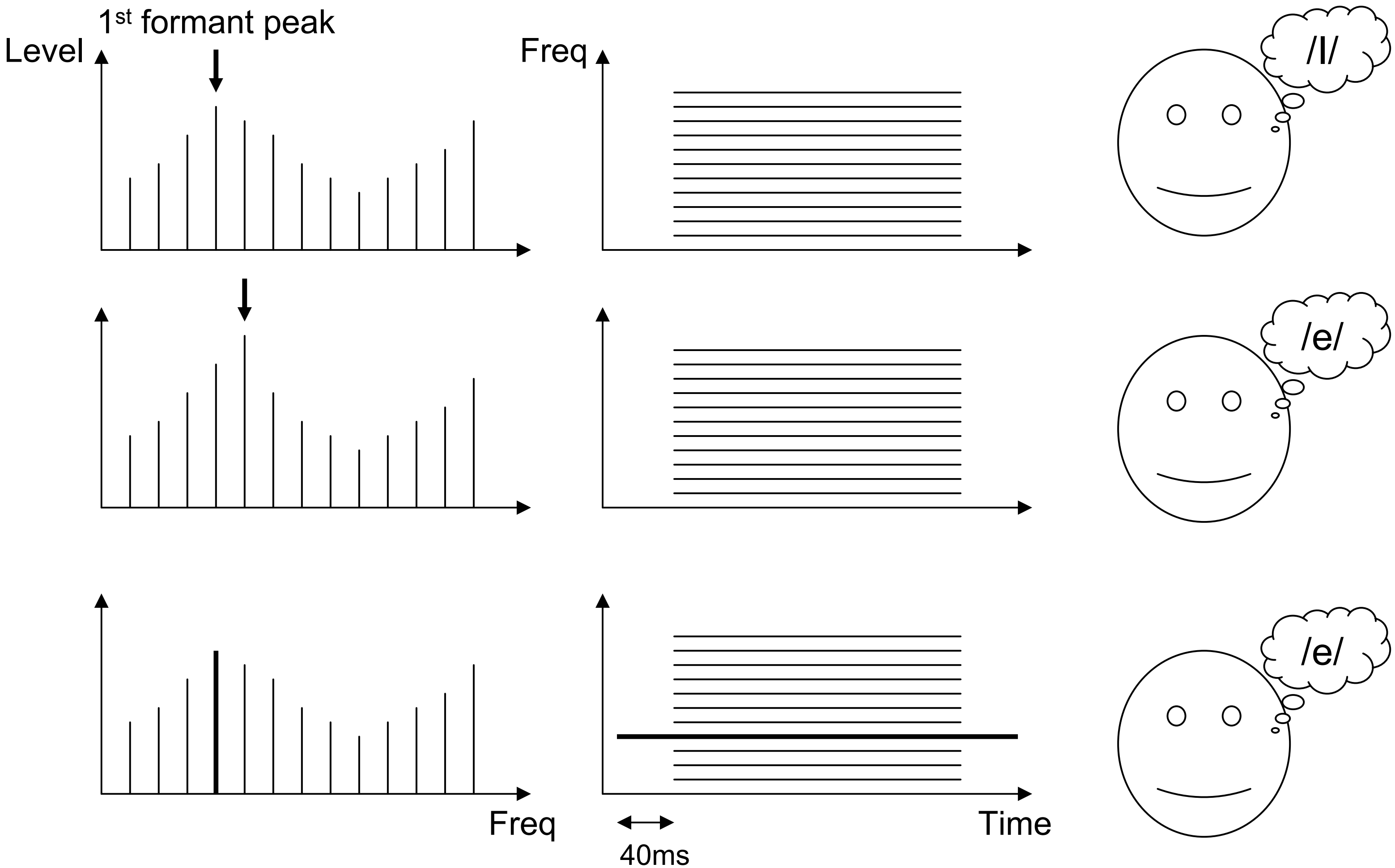
Onset asynchronies promote perceptual segregation



Influence of onset grouping/segregation on other aspects of auditory perception

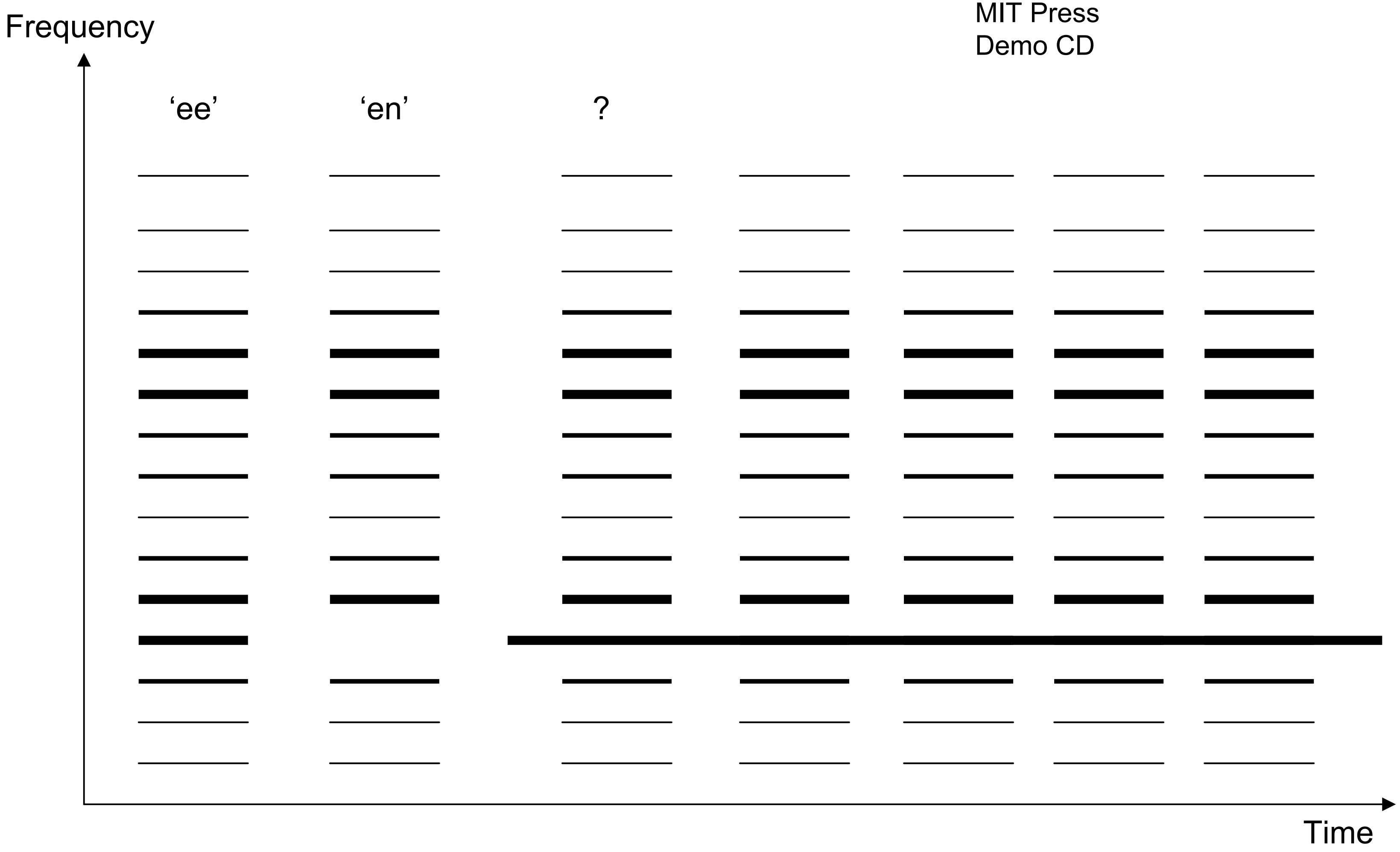
De-synchronizing a harmonic near a formant can affect perceived vowel identity

Darwin (1984); Darwin & Sutherland (1984)



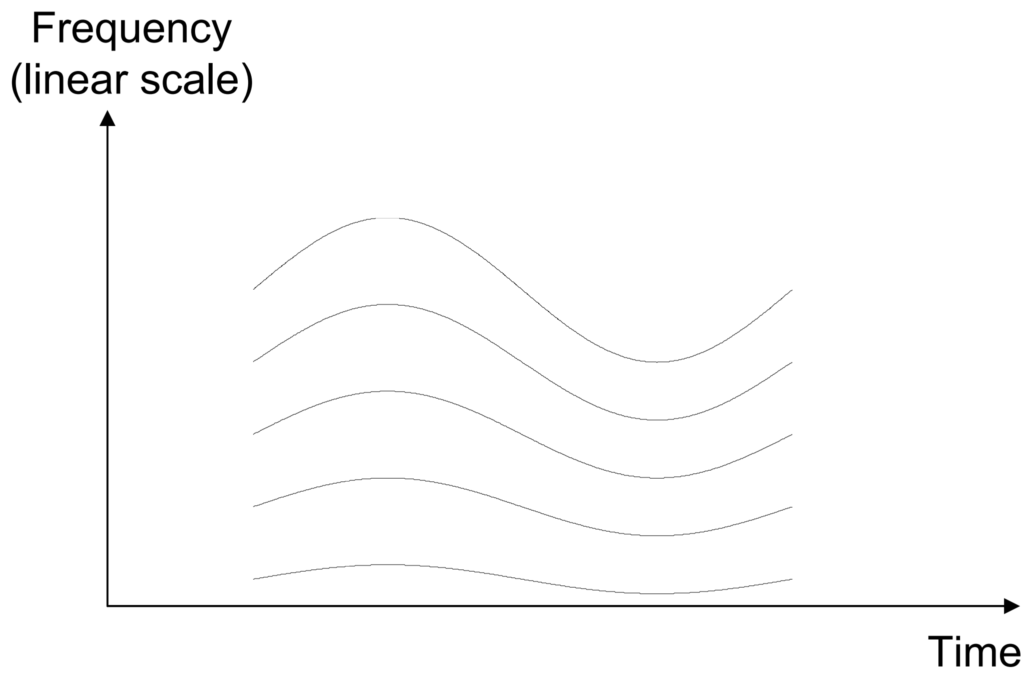
Demonstration of onset asynchrony and vowel identity

From:
Bregman (1990)
Auditory scene analysis
MIT Press
Demo CD



Co-modulation. I. Frequency modulation (FM)

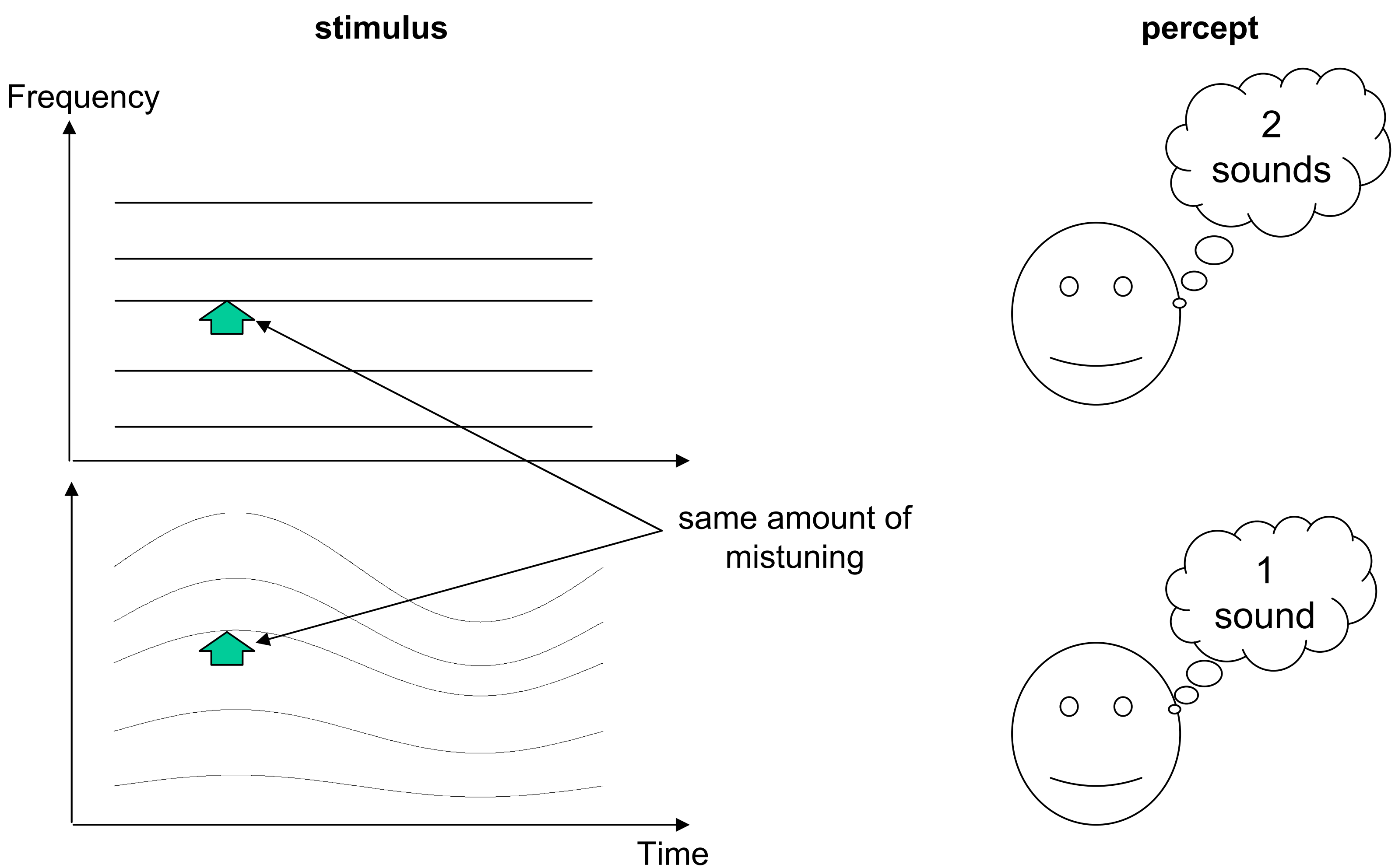
When the F0 of a harmonic sound changes,
all of its harmonics change frequency coherently



Co-modulation. I. Frequency modulation (FM)

Coherent FM promotes the fusion of harmonics

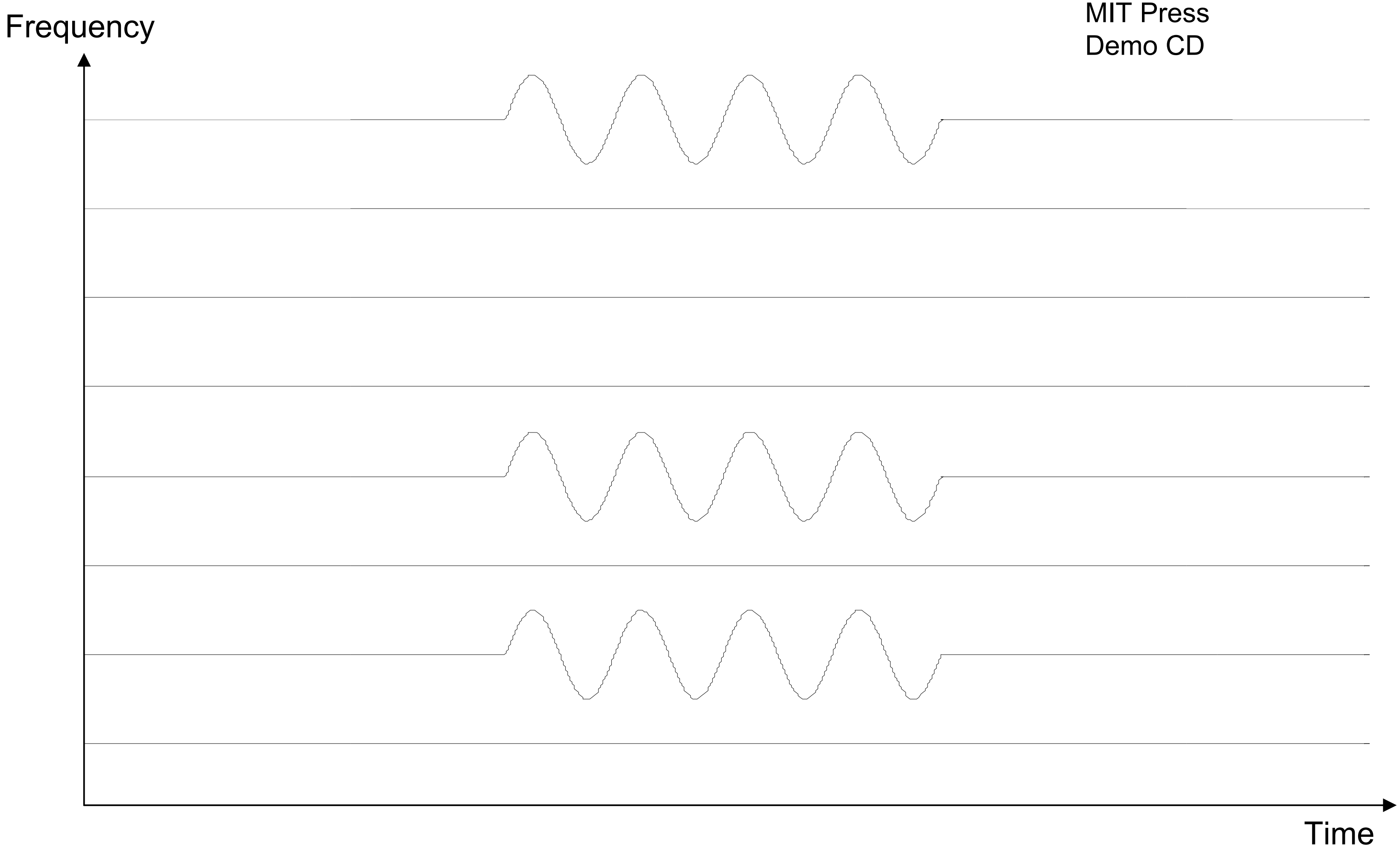
Darwin *et al.* (1994)



FM-based grouping - Demo 1

FM can make harmonics stand out

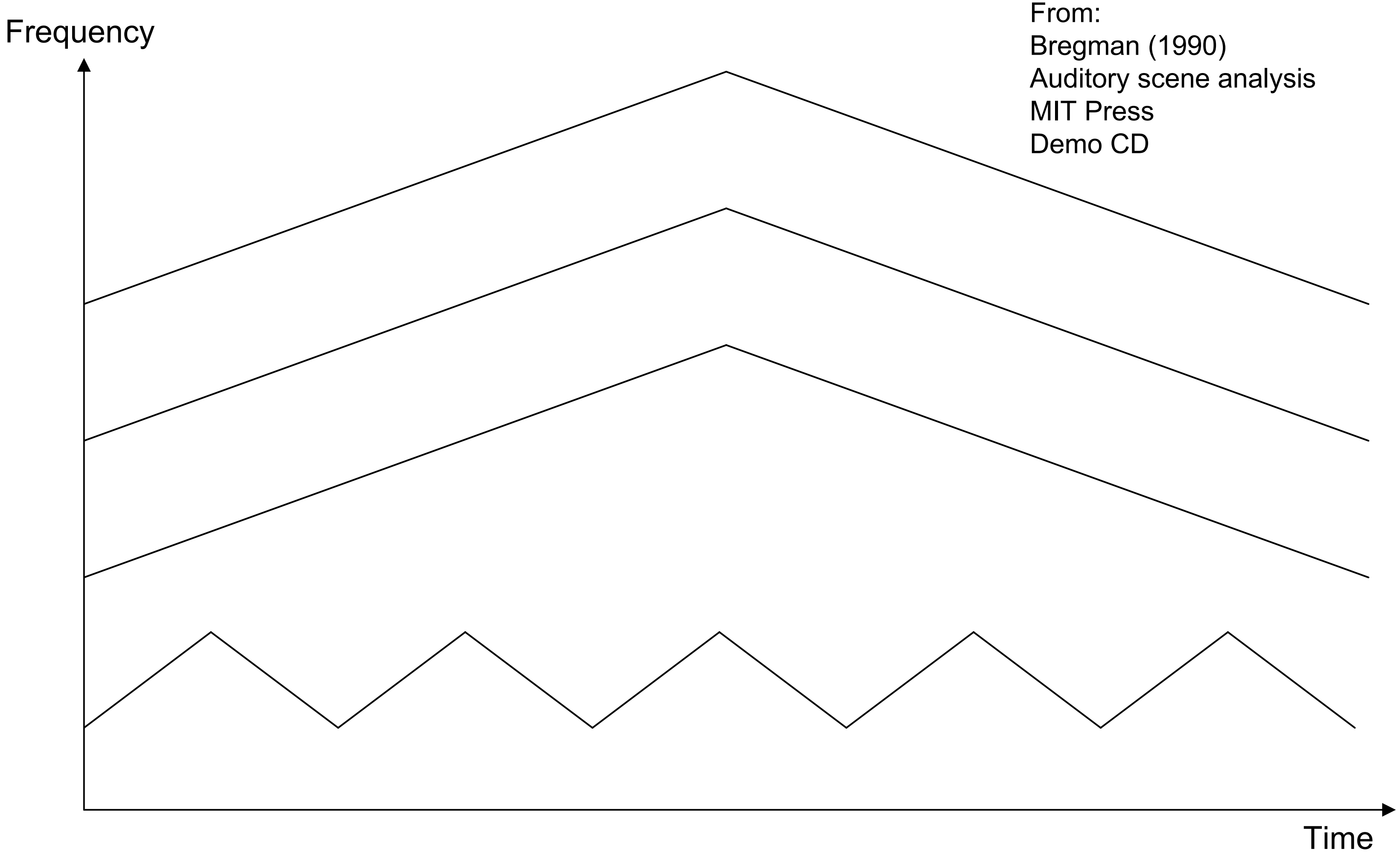
From:
Bregman (1990)
Auditory scene analysis
MIT Press
Demo CD



FM-based grouping - Demo 2

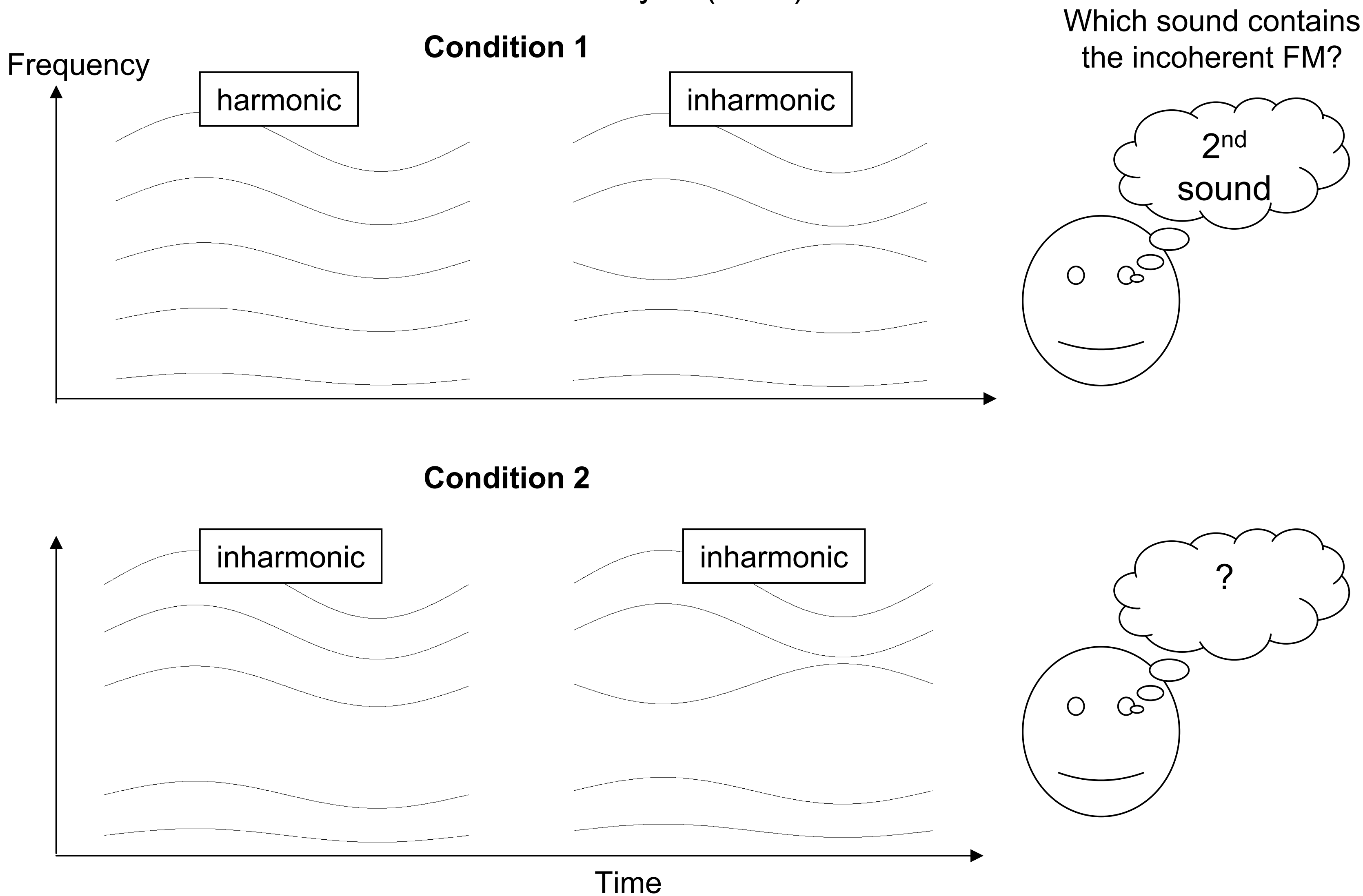
Incoherent FM promotes segregation

From:
Bregman (1990)
Auditory scene analysis
MIT Press
Demo CD



Is it FM or harmonicity?

Carlyon (1991)



Co-modulation. II. Amplitude modulation

Current evidence in favor of a genuine AM-based grouping/segregation mechanism is weak, at best

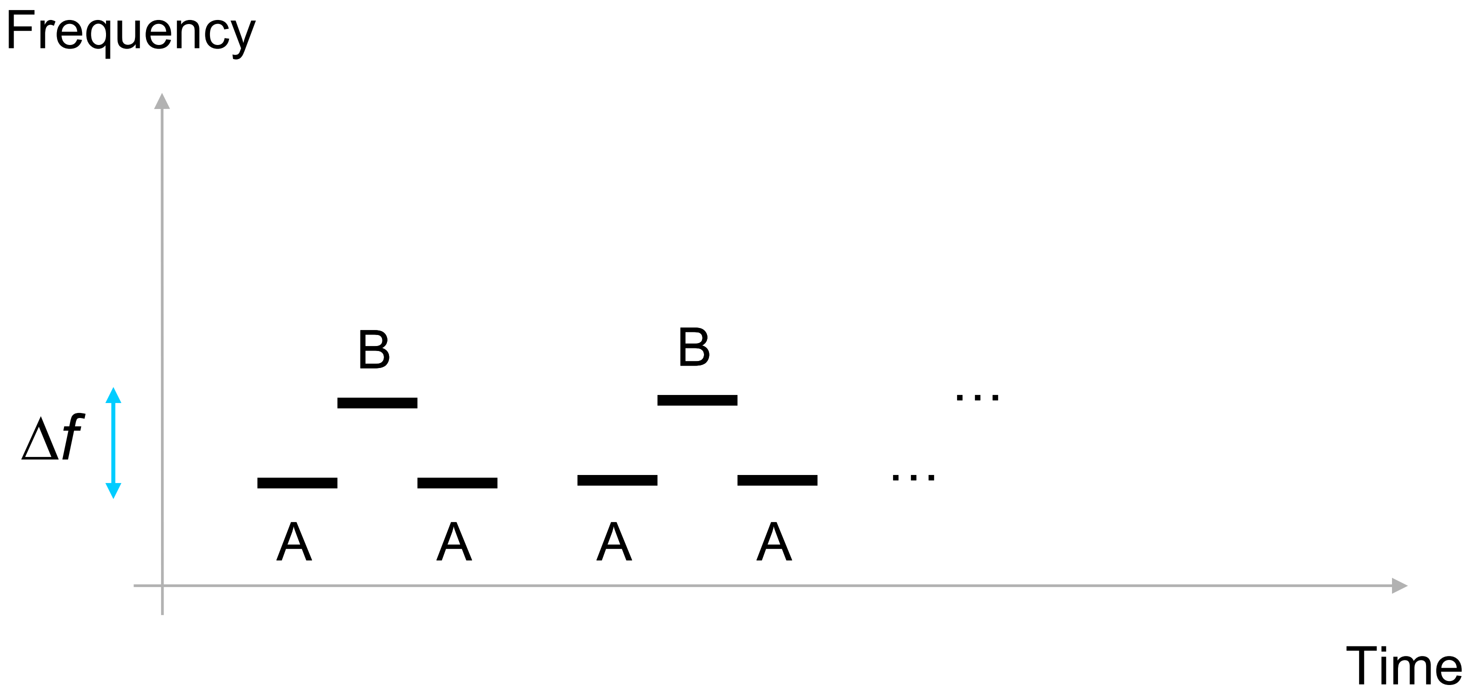
Out-of phase AM generally results in onset asynchronies
(leading to the question: is it really AM phase or rather onset asynchrony?)

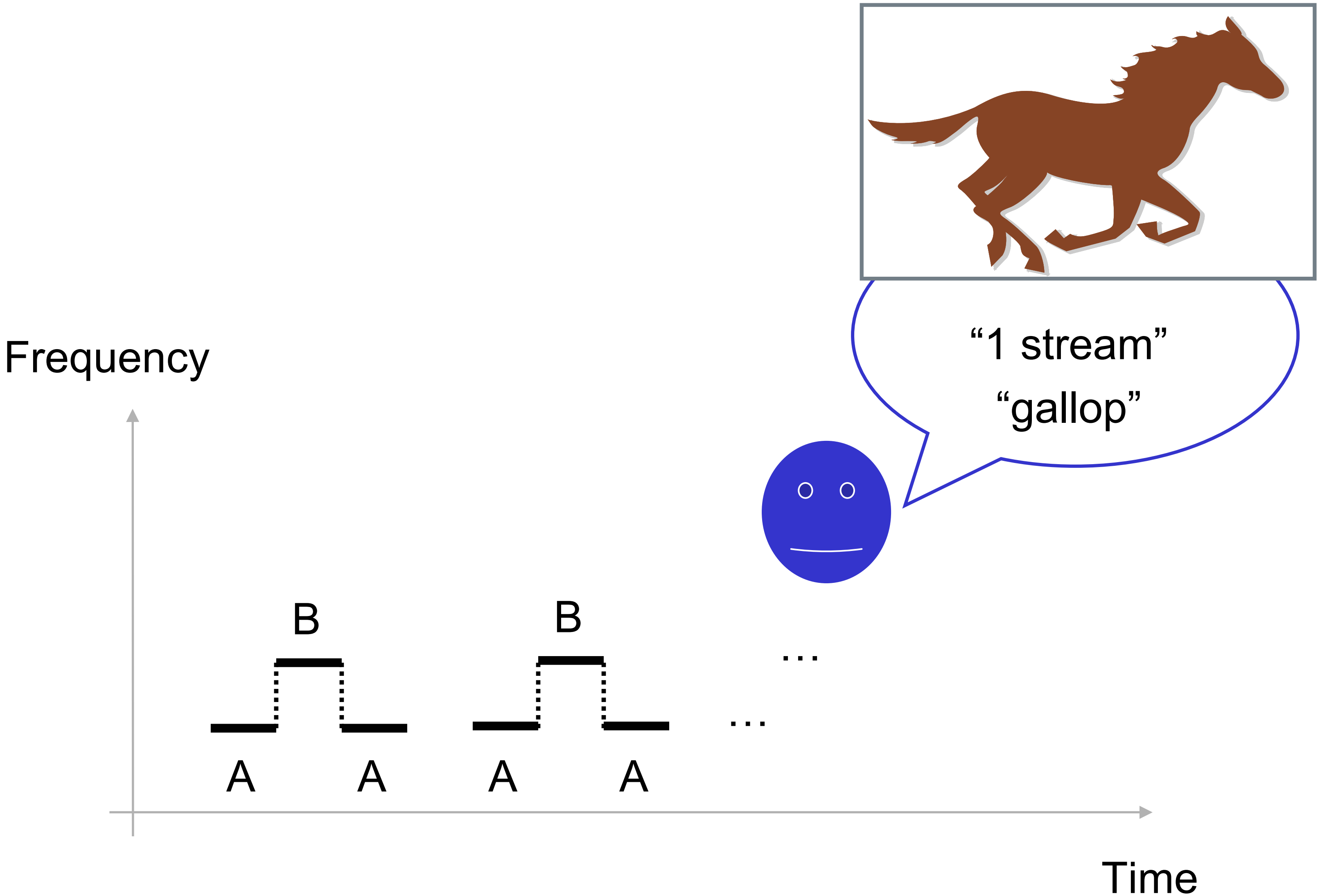
Out-of phase AM results in some spectral components being well audible while the others are not, at certain times
(leading to the question: is the pop-out due to AM or enhanced SNR?)

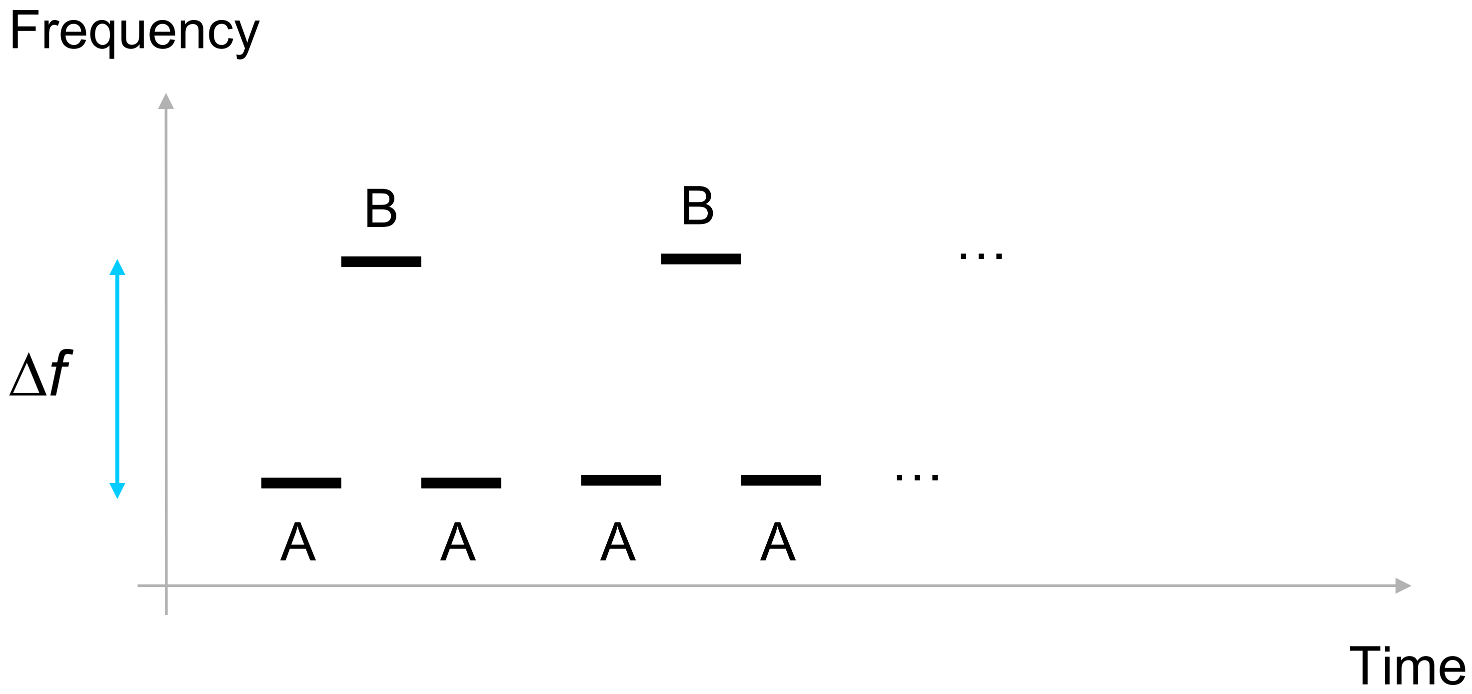
Auditory streaming

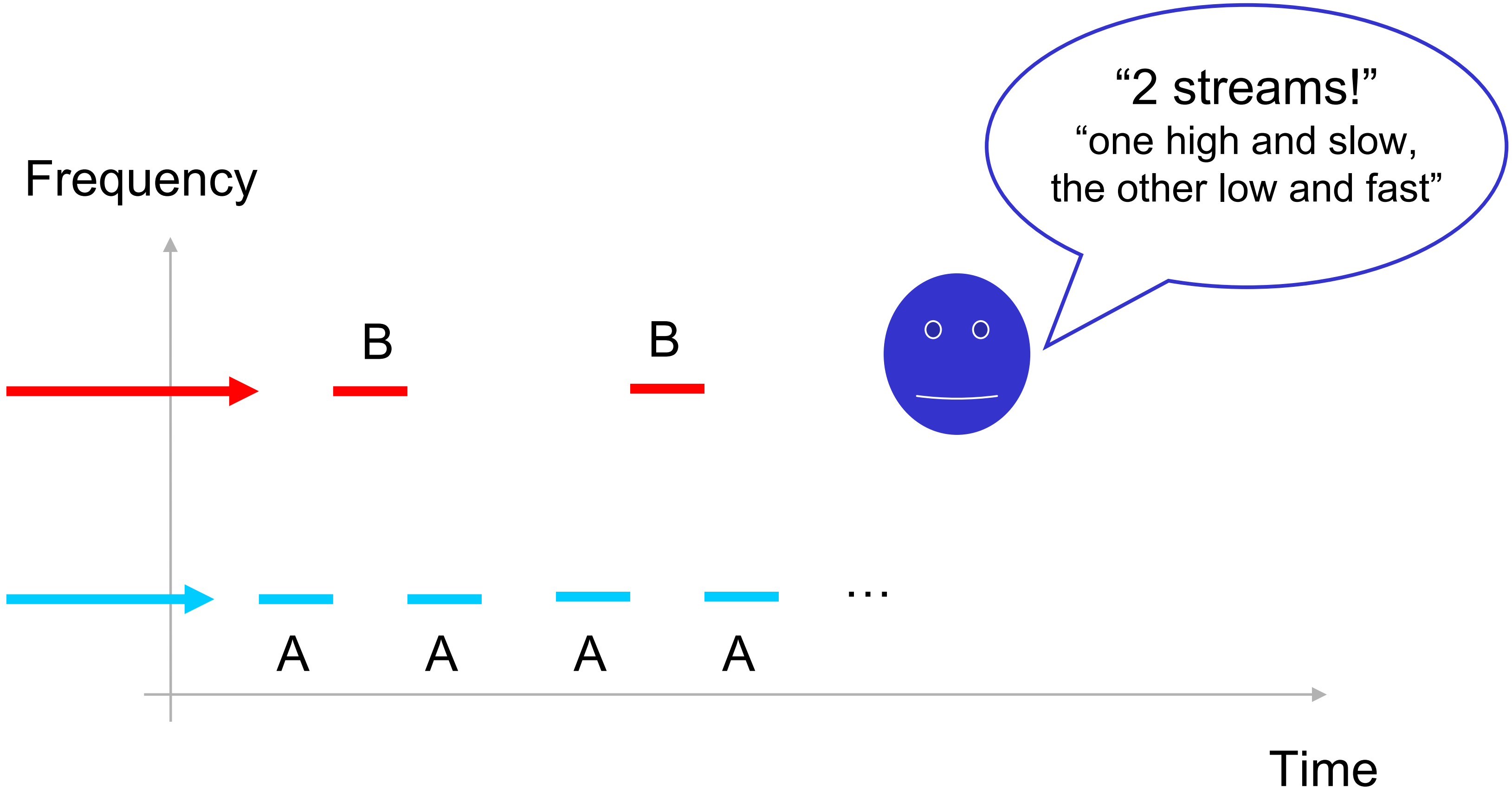
What is it?

Description and demonstration of the phenomenon

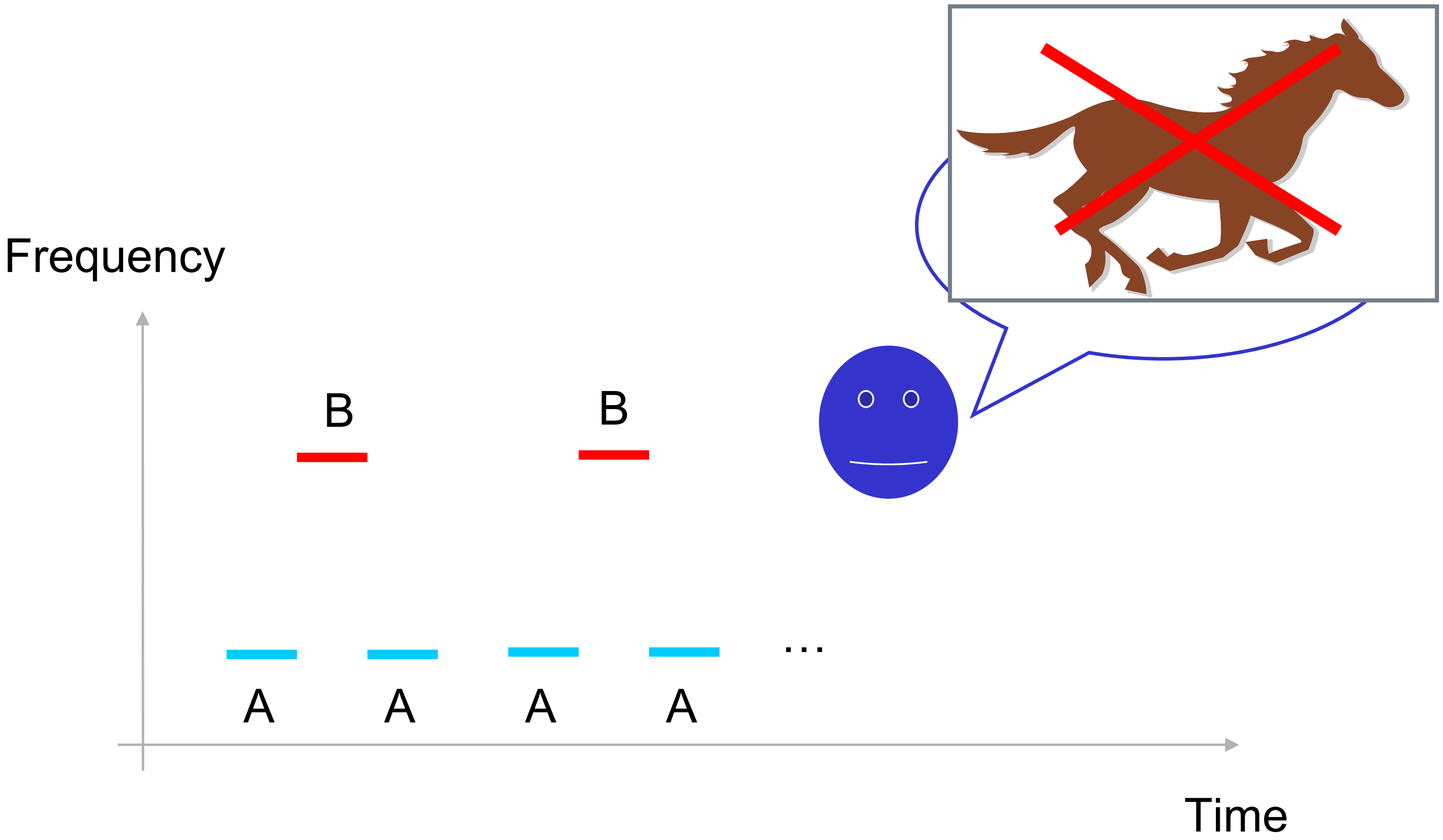

Frequency
Δf
B
B
A
A
A
A
...
...
Time


“1 stream”
“gallop”
Frequency
B
B
...
A
A
A
A
...
Time

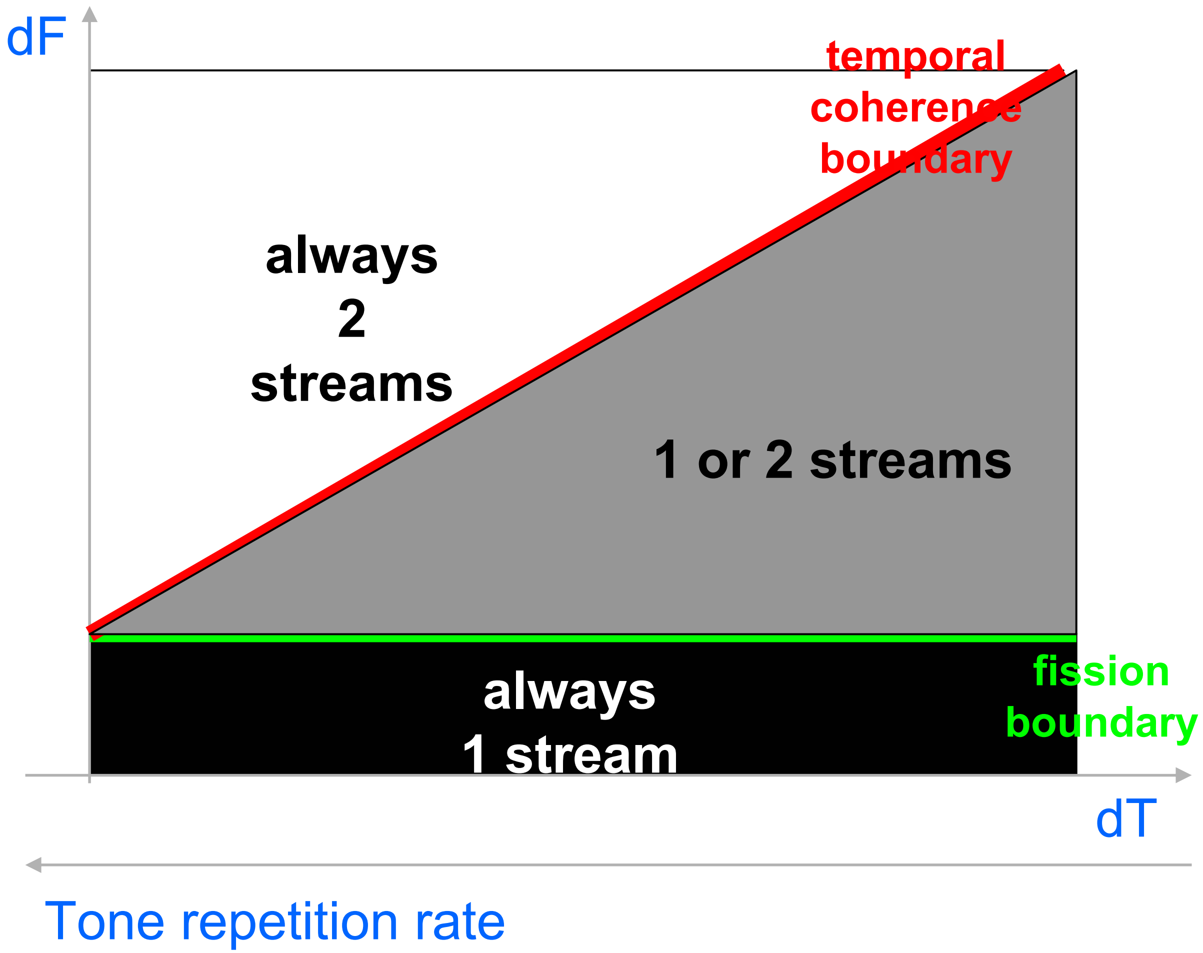

Frequency
B
B
...
Δf
...
A
A
A
A
Time

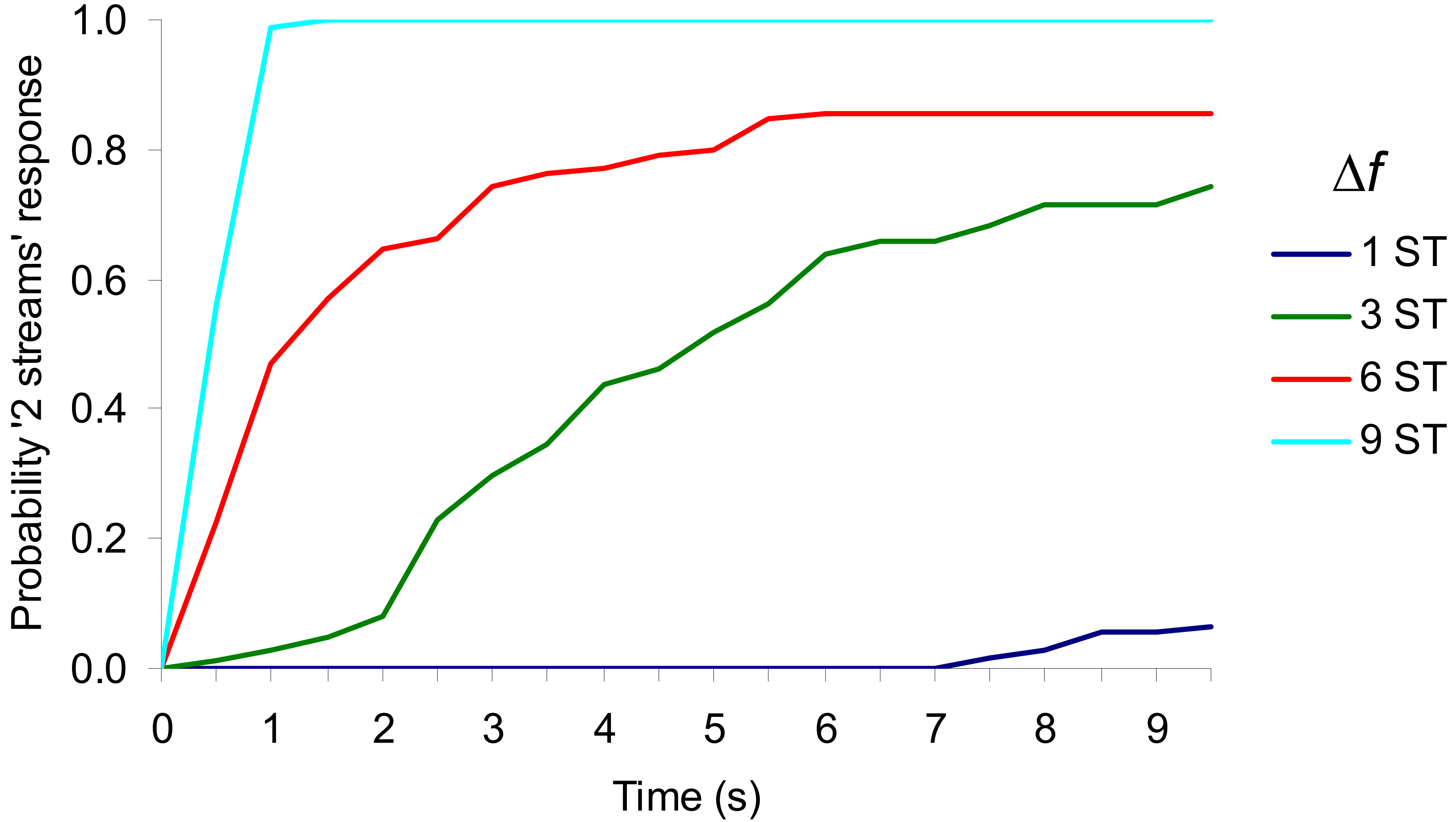


Frequency
B
B
A
A
A
A
...
Time
"2 streams!"
"one high and slow,
the other low and fast"


Frequency
B
B
...
A
A
A
A
Time

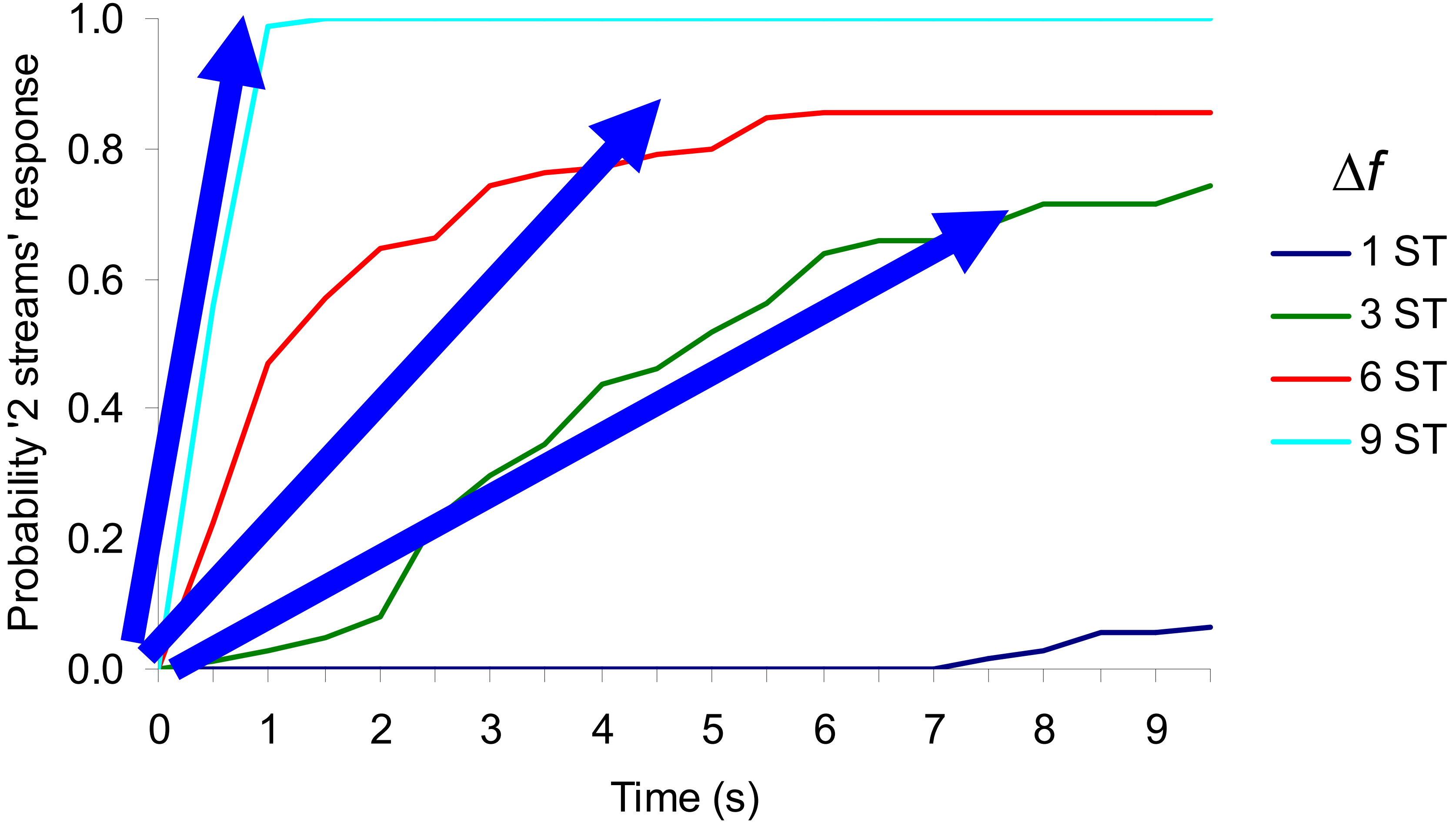
A basic pre-requisite for any neural correlate of streaming: depend on both dF and dT





Probability '2 streams' response
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
Time (s)
Δf
1 ST
3 ST
6 ST
9 ST

Build-up



1.0
0.8
0.6
0.4
0.2
0.0
Probability '2 streams' response
0
1
2
3
4
5
6
7
8
9
Time (s)
Δf
1 ST
3 ST
6 ST
9 ST

Traditional explanations for the build-up

« Neurophysiological » explanation

Neural adaptation of coherence/pitch-motion detectors

(Anstis & Saida, 1985)

« Cognitive » explanation

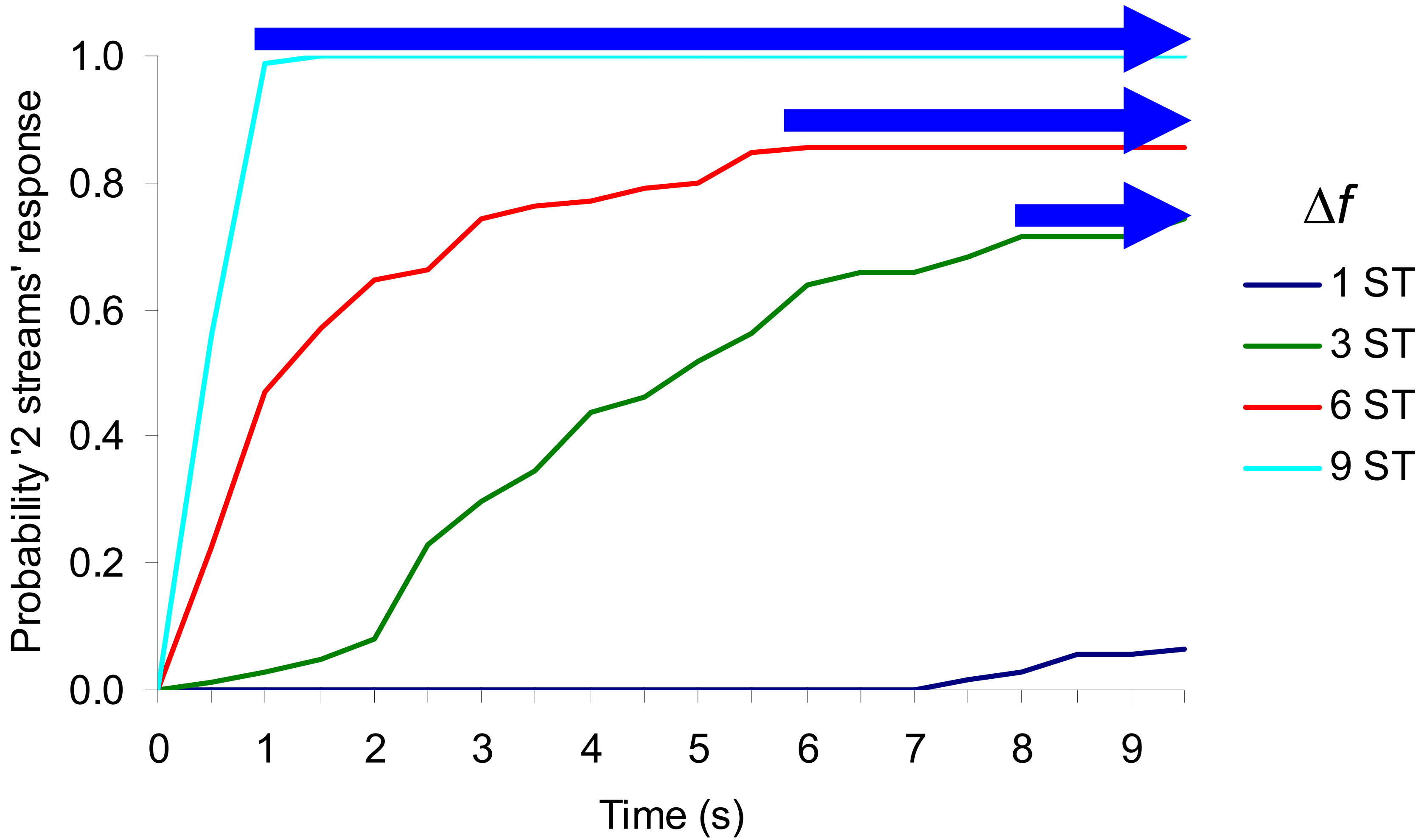
The default is integration (1 stream);

the brain needs to accumulate evidence that there is more than 1 stream

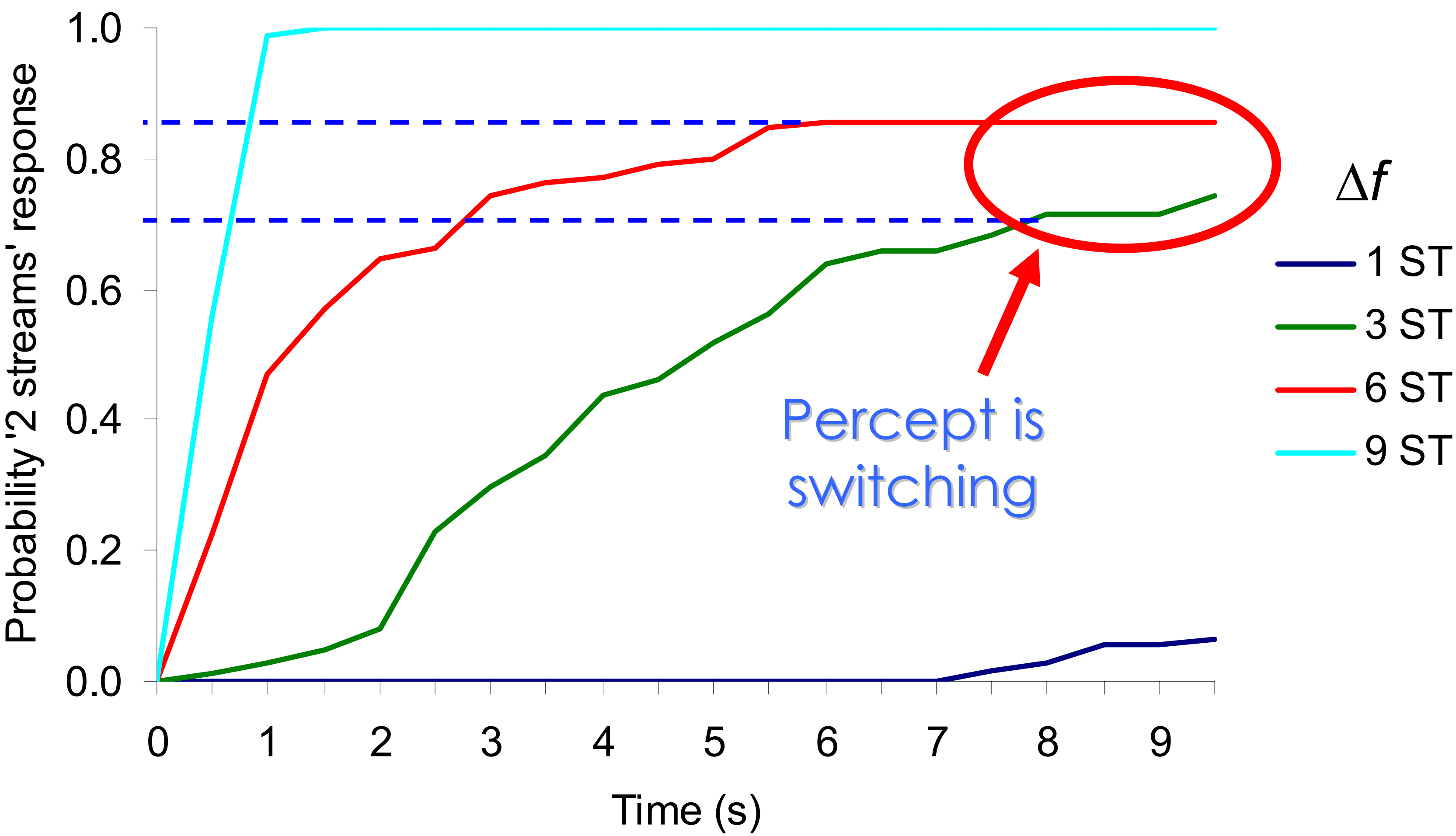
before declaring « 2 streams »

(Bregman, 1978, 1990,…)

Asymptote



1.0
0.8
0.6
0.4
0.2
0.0
Probability '2 streams' response
0
1
2
3
4
5
6
7
8
9
Time (s)
Δf
1 ST
3 ST
6 ST
9 ST



1.0
0.8
0.6
0.4
0.2
0.0
Probability '2 streams' response
0
1
2
3
4
5
6
7
8
9
Time (s)
Δf
1 ST
3 ST
6 ST
9 ST
Percept is switching

Ambiguous stimuli and bi-stable percepts

Necker's cube

Rubin's vase-faces

Figures removed due to copyright reasons.

have been used successfully in the past
to demonstrate neural/brain correlates of visual percepts

e.g., Logothetis & Schall (1989), Leopold & Logothetis (1996),..

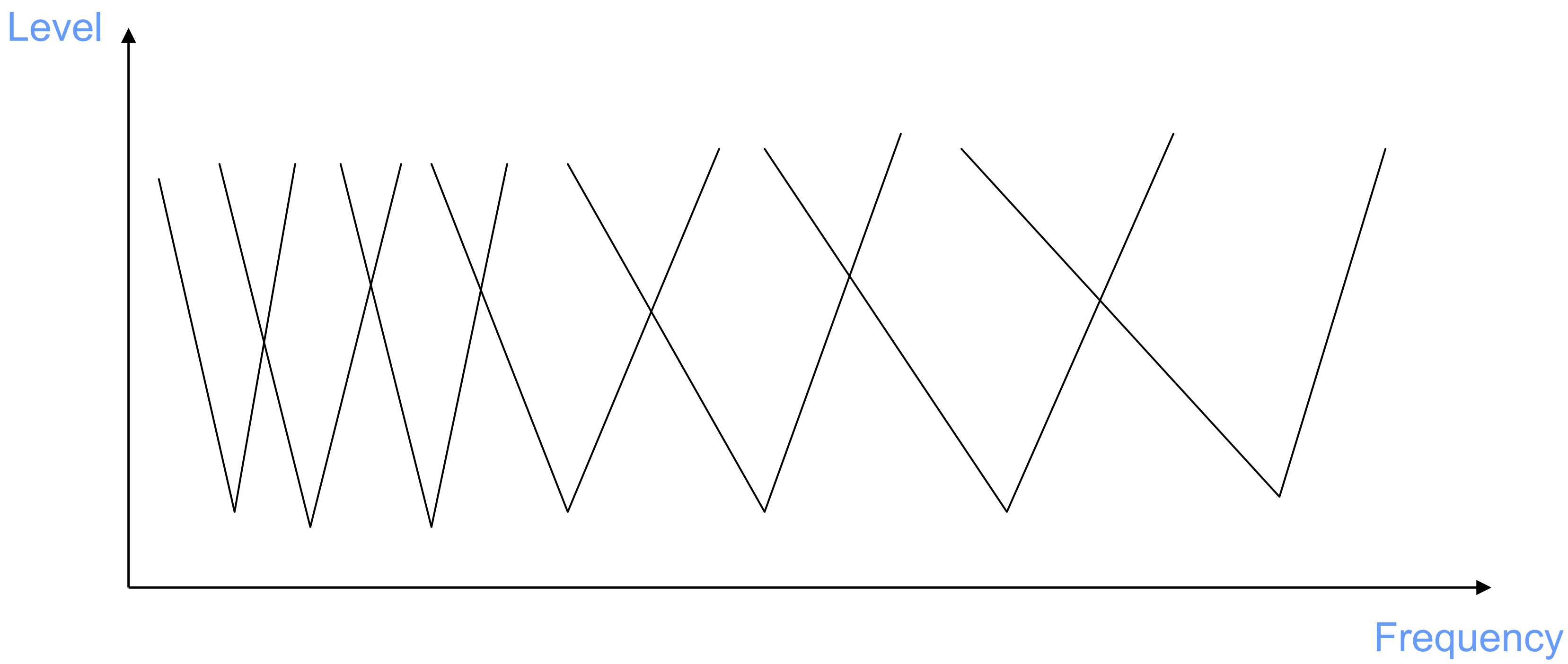
Streaming

How does it work?

Theories and computational models

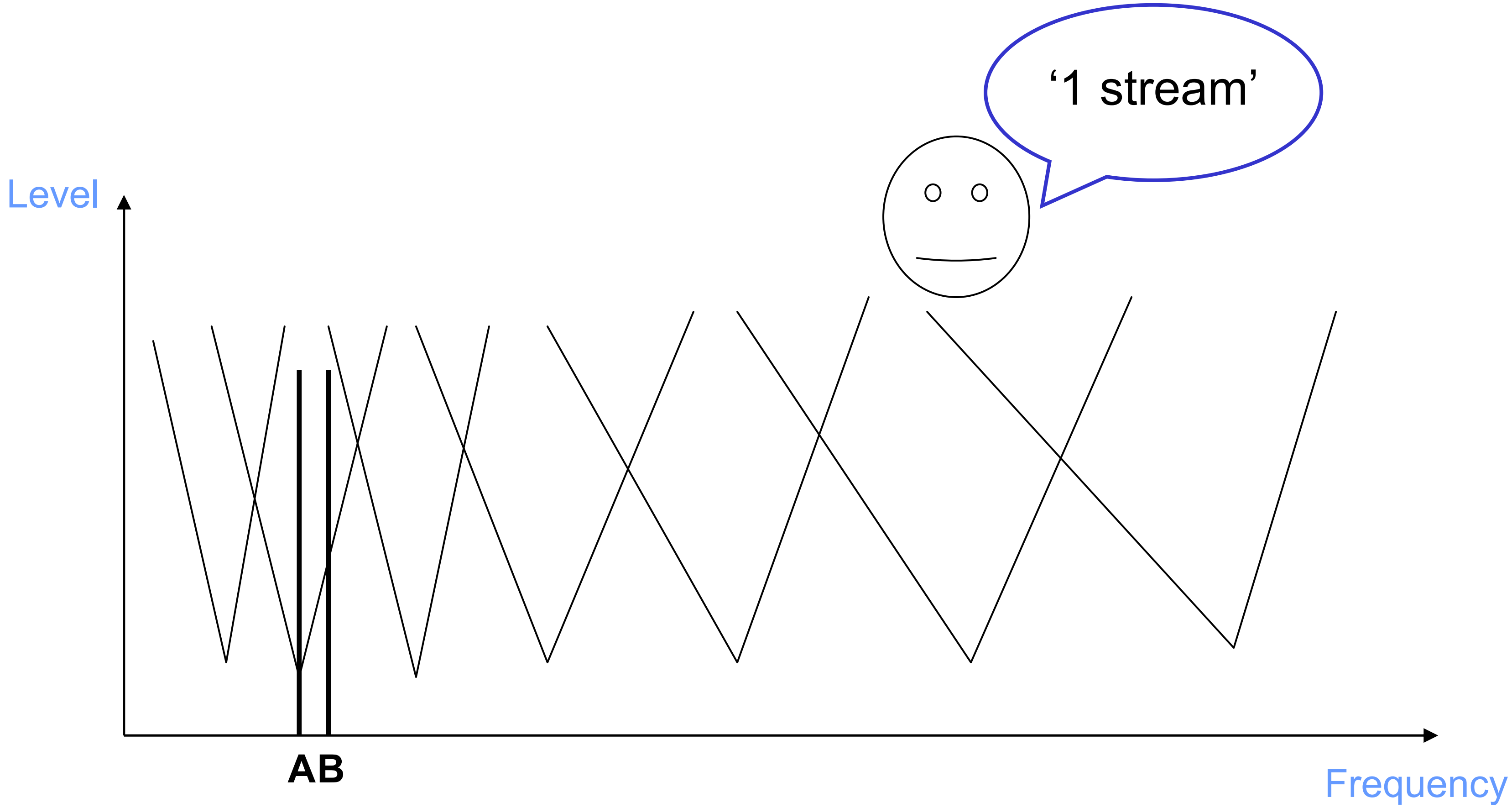
The channeling theory

Hartmann and Johnson (1991) Music Percept.



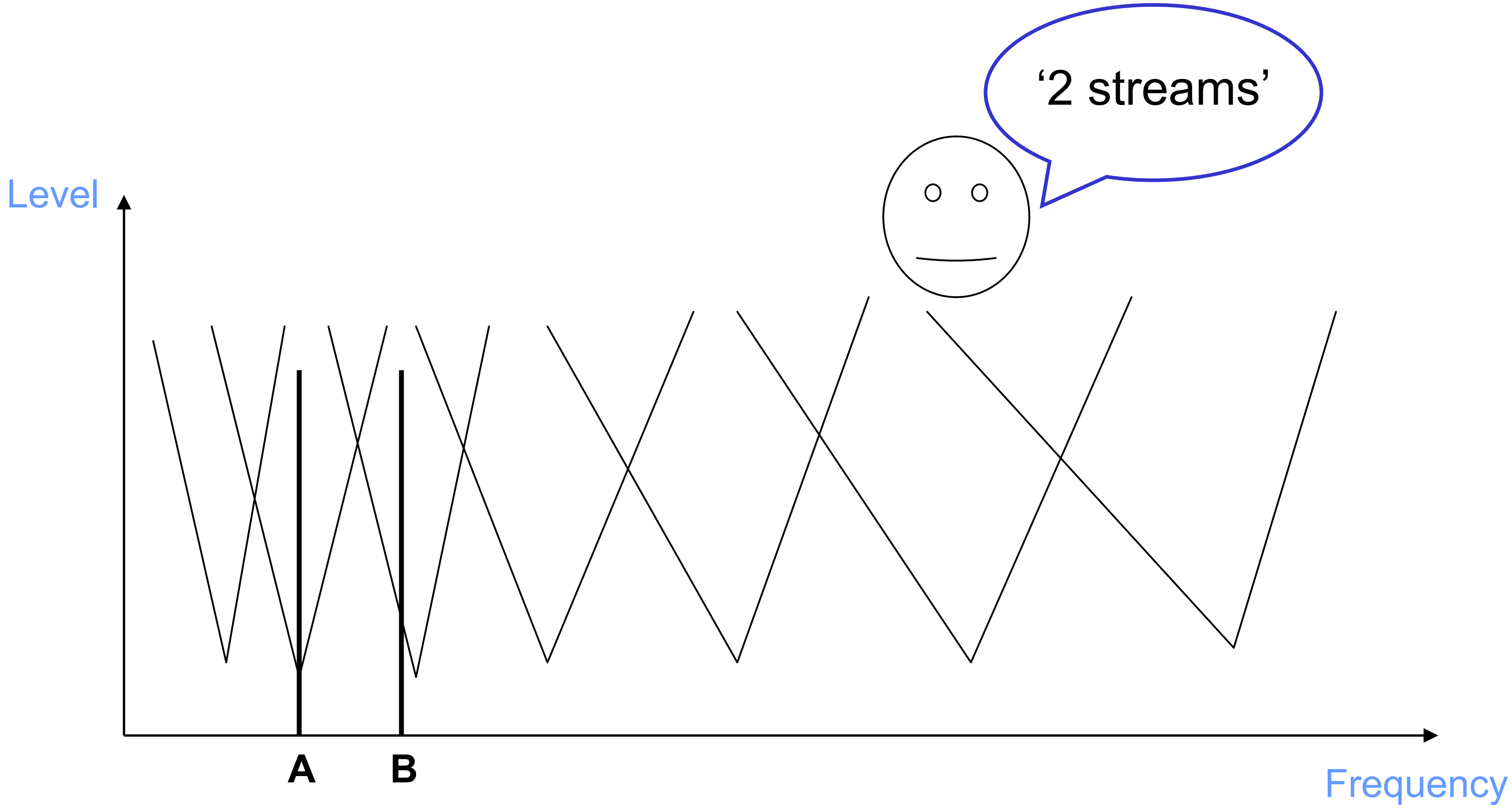
The channeling theory

Hartmann and Johnson (1991) Music Percept.



The channeling theory

Hartmann and Johnson (1991) Music Percept.



Streaming

How does it really work?

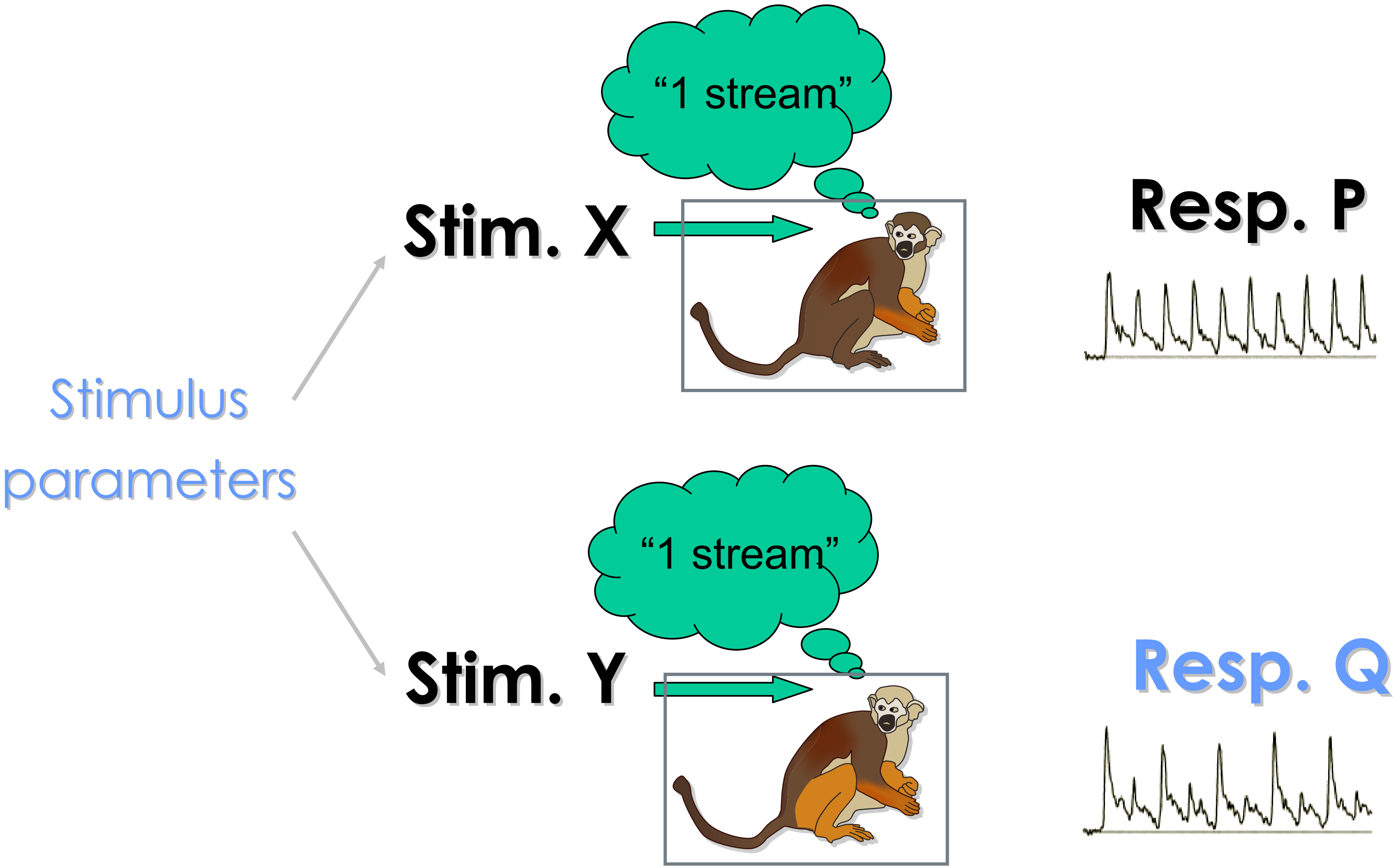
Neural mechanisms

Behavioral evidence that streaming occurs in

- **monkey** (Izumi, 2002)

- **bird** (Hulse et al., 1997; McDougall-Shackleton et al, 1998)

- **fish** (Fay, 1998)



Stimulus parameters
Stim. X
"1 stream"
Resp. P
Stim. Y
"1 stream"
Resp. Q

Single/few/multi-unit intra-cortical recordings

Monkeys: Fishman et al. (2001) Hear. Res. 151, 167-187

Bats: Kanwal, Medvedev, Micheyl (2003) Neural Networks

Figures removed due to copyright reasons.
Please see: Fishman et al. (2001)

At low repetition rates, units respond to both on- and off-BF tones

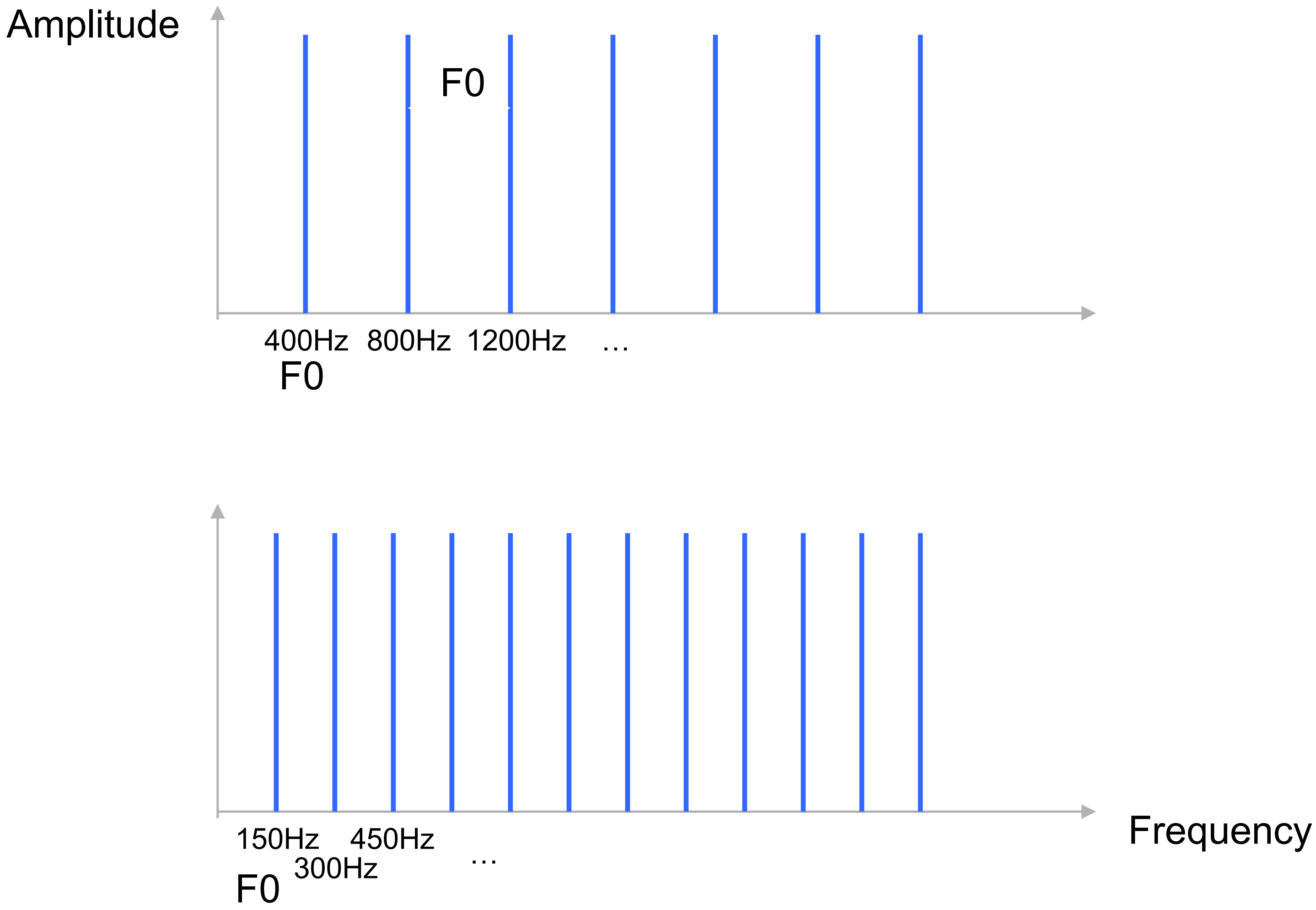
At high repetition rates, only on-BF tone response is visible

Is peripheral chanelling the whole story?

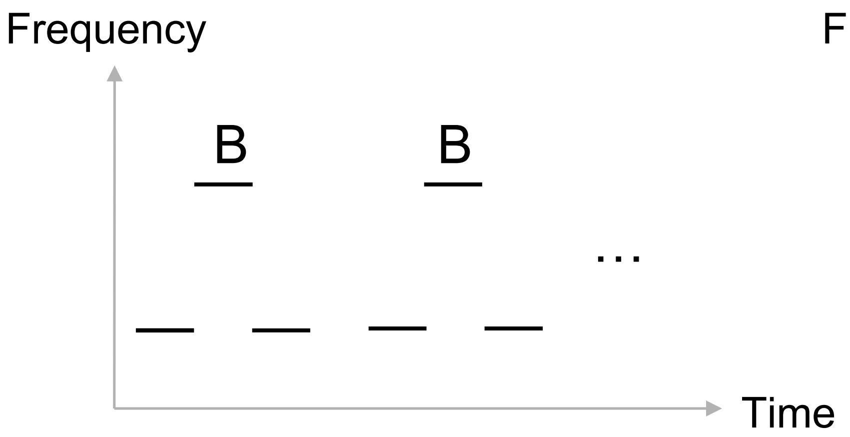
Sounds that excite the same peripheral channels can yield streaming

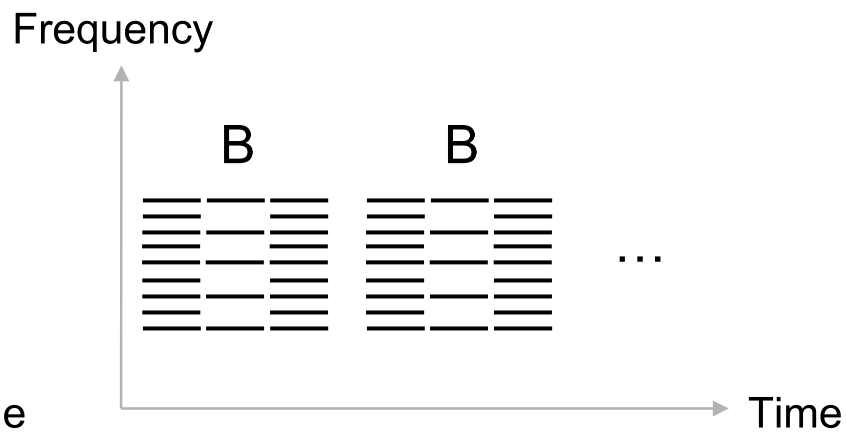
Vliegen & Oxenham (1999)
Vliegen, Moore, Oxenham (1999)
Grimault, Micheyl, Carlyon et al. (2001)
Grimault, Bacon, Micheyl (2002)
Roberts, Glasberg, Moore (2002)
...

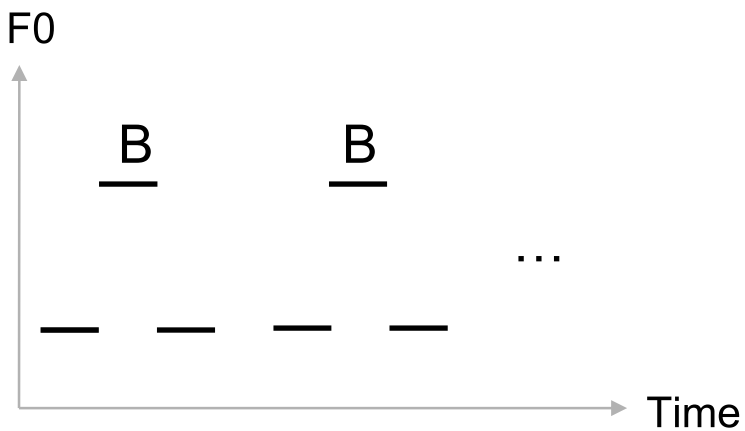
Streaming with complex tones



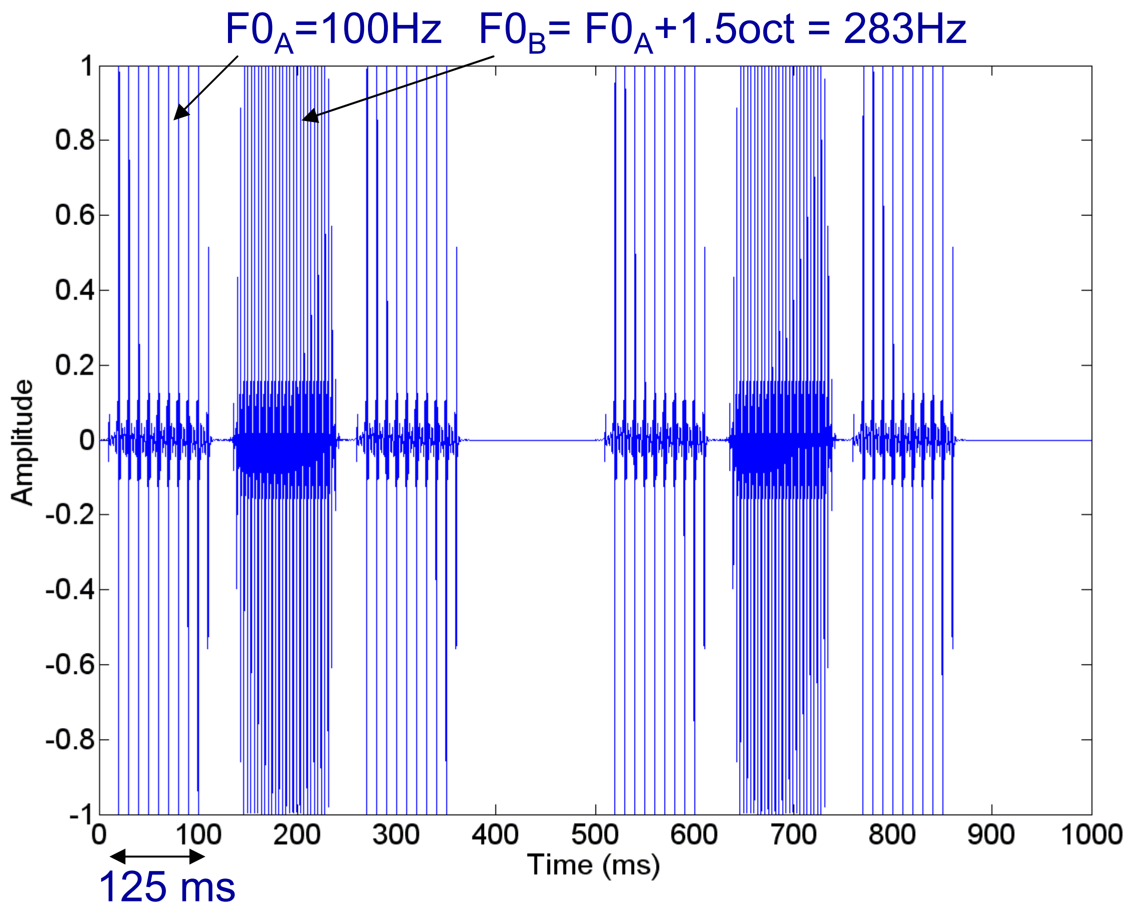
Streaming based on F0 differences



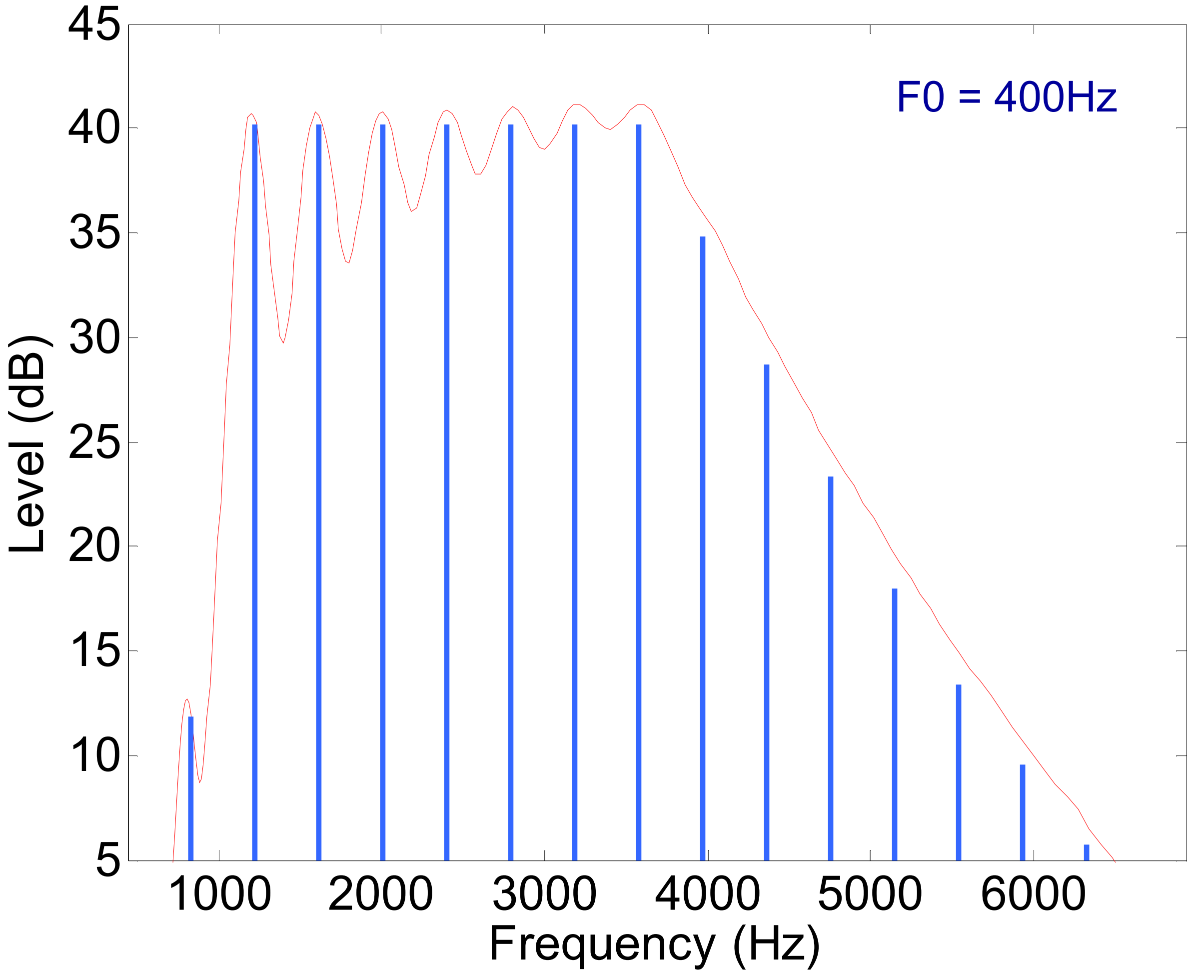




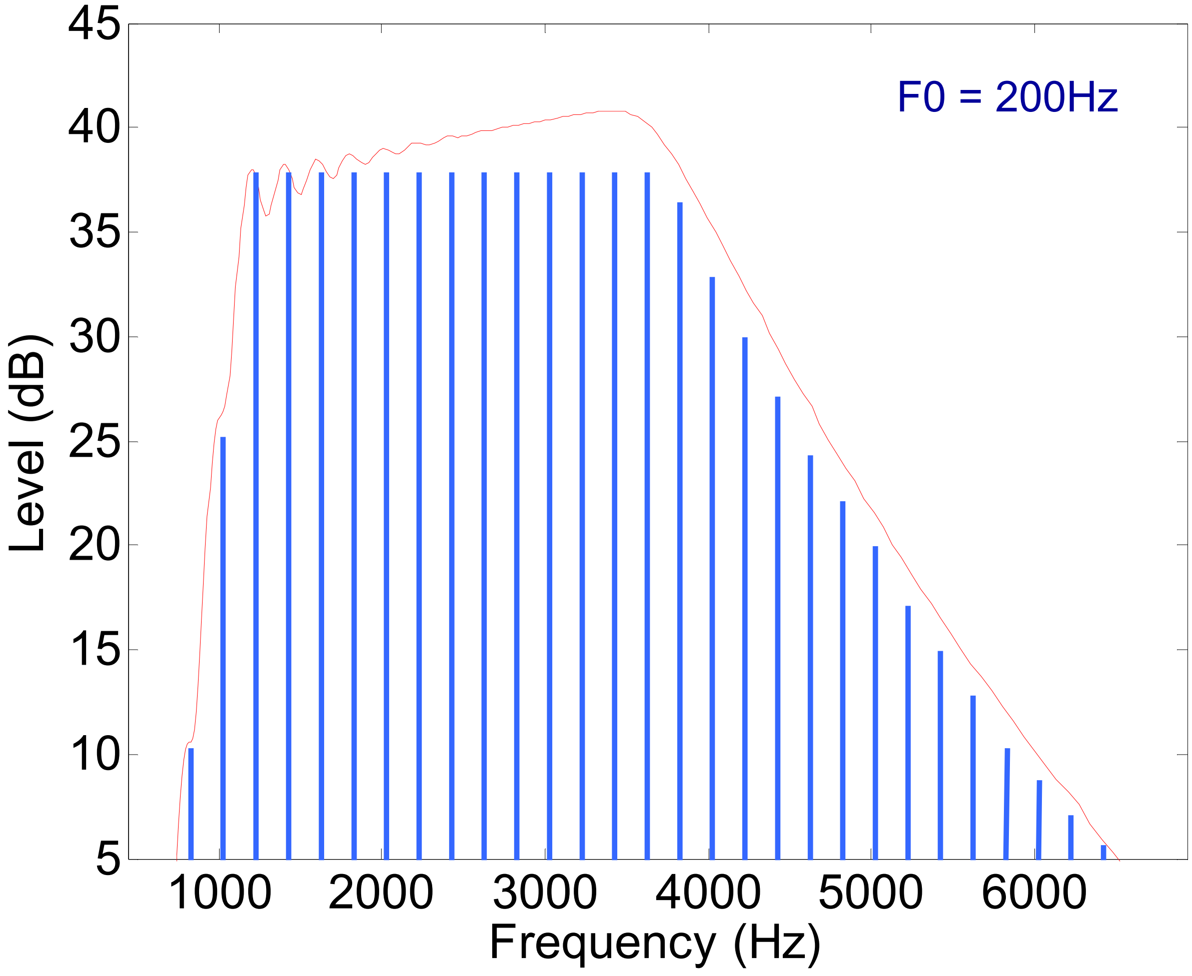
Streaming based on F0 differences



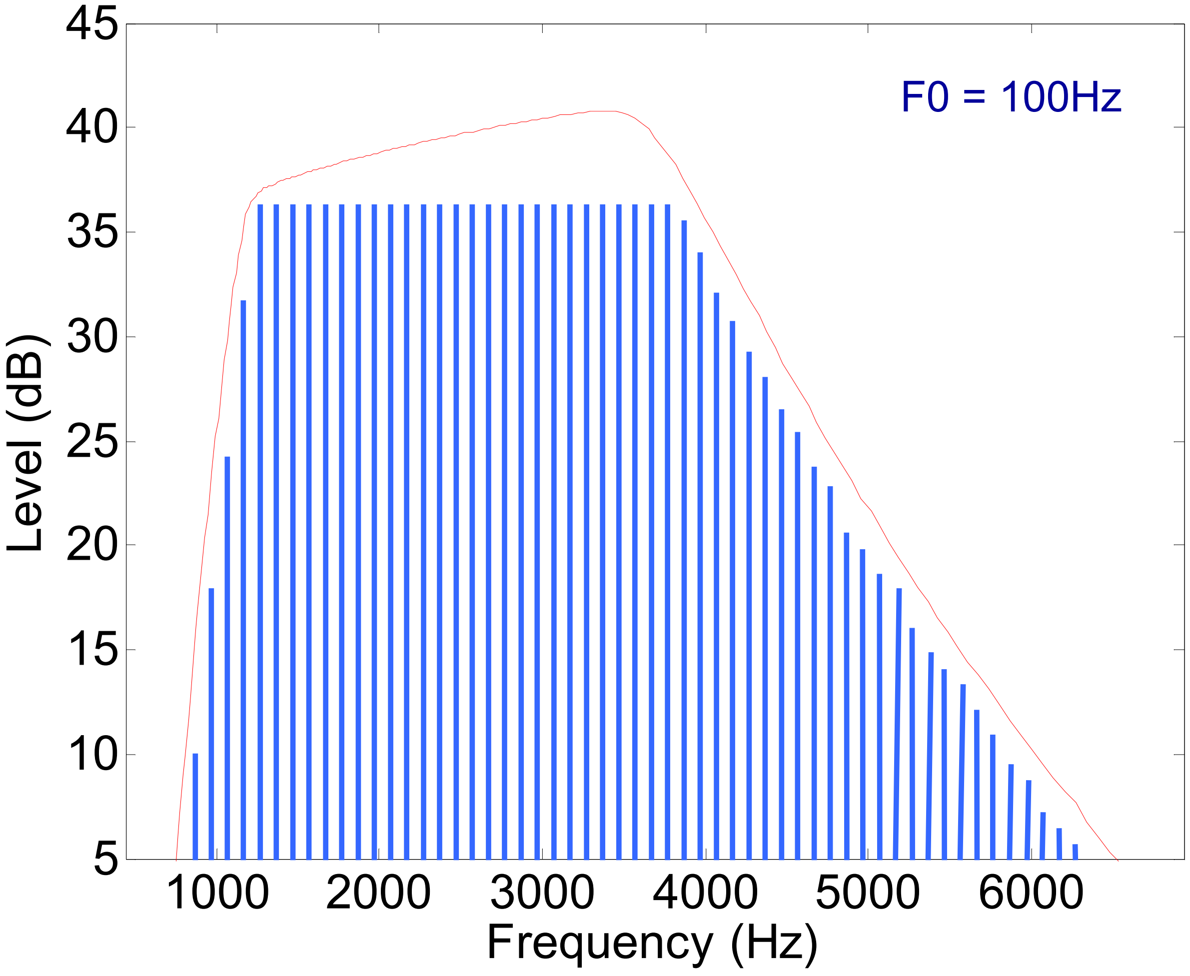
Auditory spectral excitation pattern evoked by bandpass-filtered harmonic complex

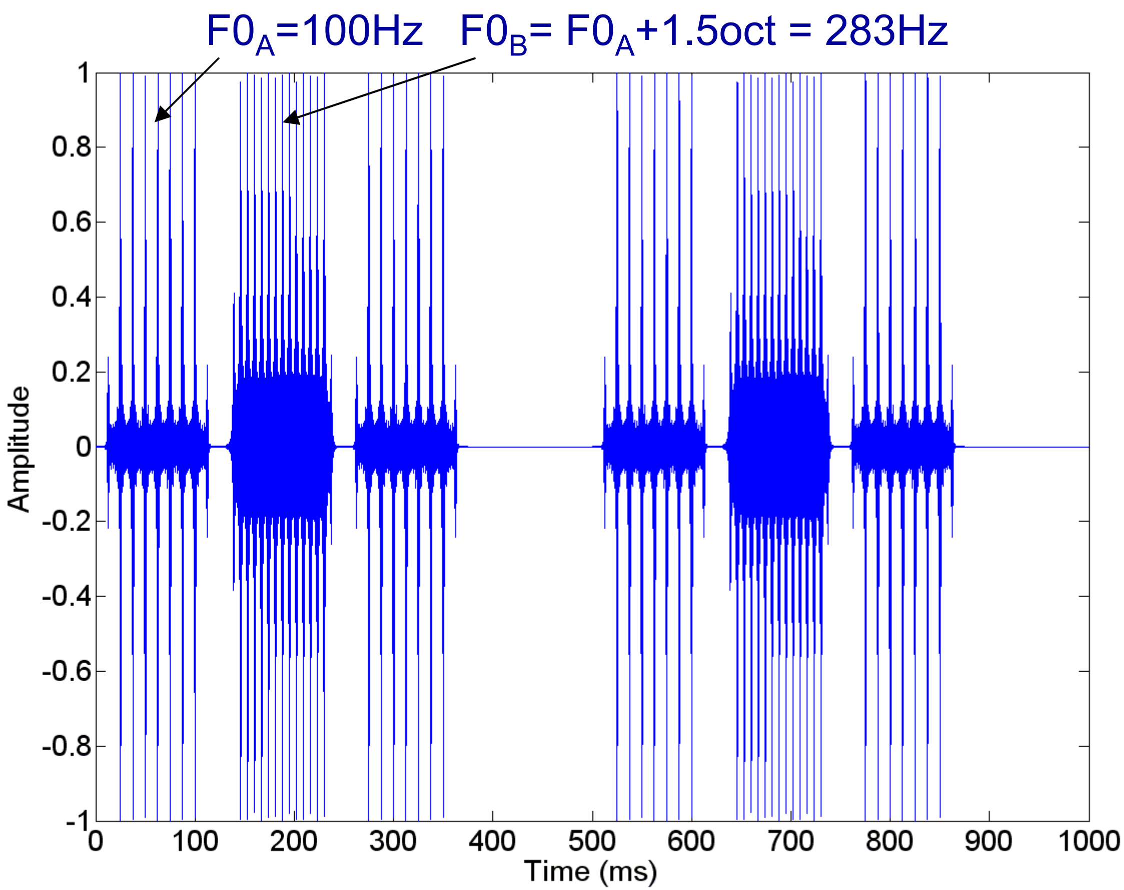


Auditory spectral excitation pattern evoked by bandpass-filtered harmonic complex



Auditory spectral excitation pattern evoked by bandpass-filtered harmonic complex





F0A=100Hz
F0B= F0A+1.5oct = 283Hz
Amplitude
Time (ms)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
100
200
300
400
500
600
700
800
900
1000

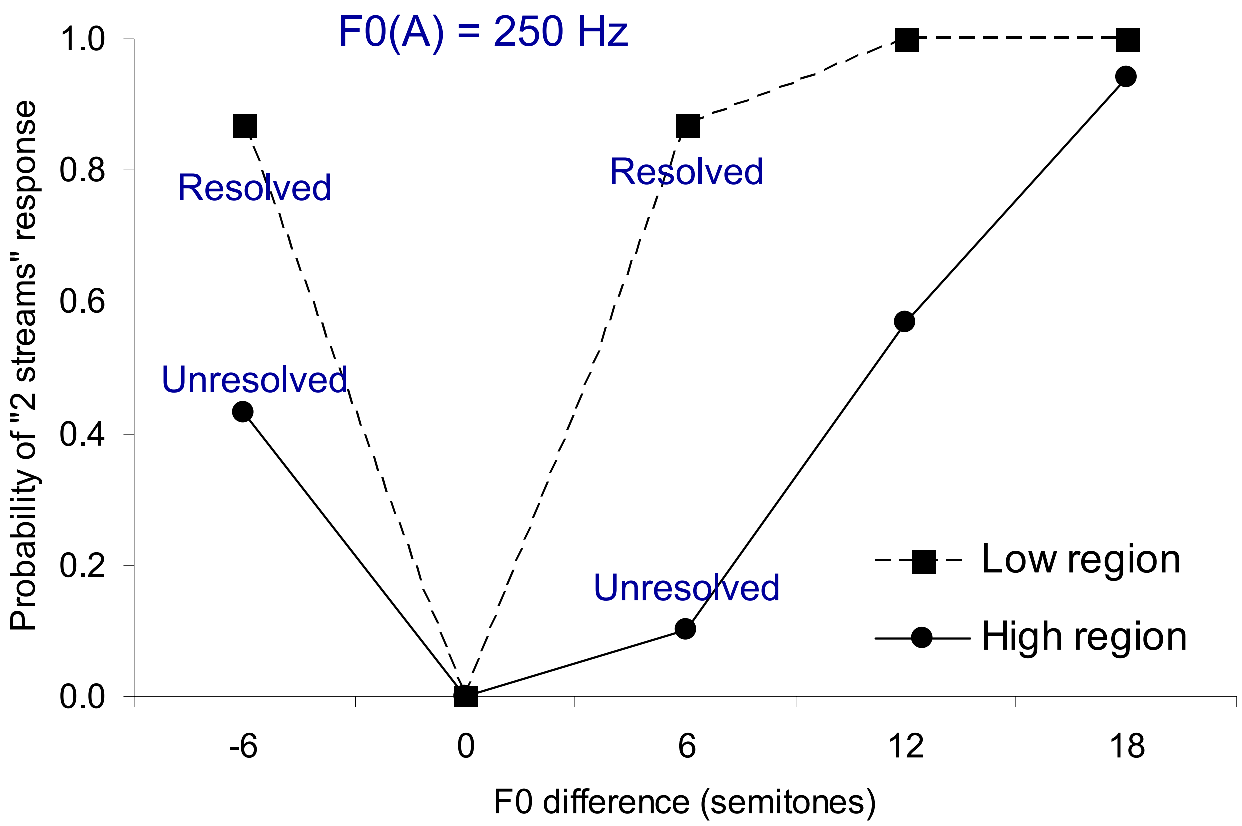
F0-based streaming with unresolved harmonics is possible...

Vliegen & Oxenham (1999); Vliegen, Moore, Oxenham (1999)

Grimault, Micheyl, Carlyon *et al.* (2000)

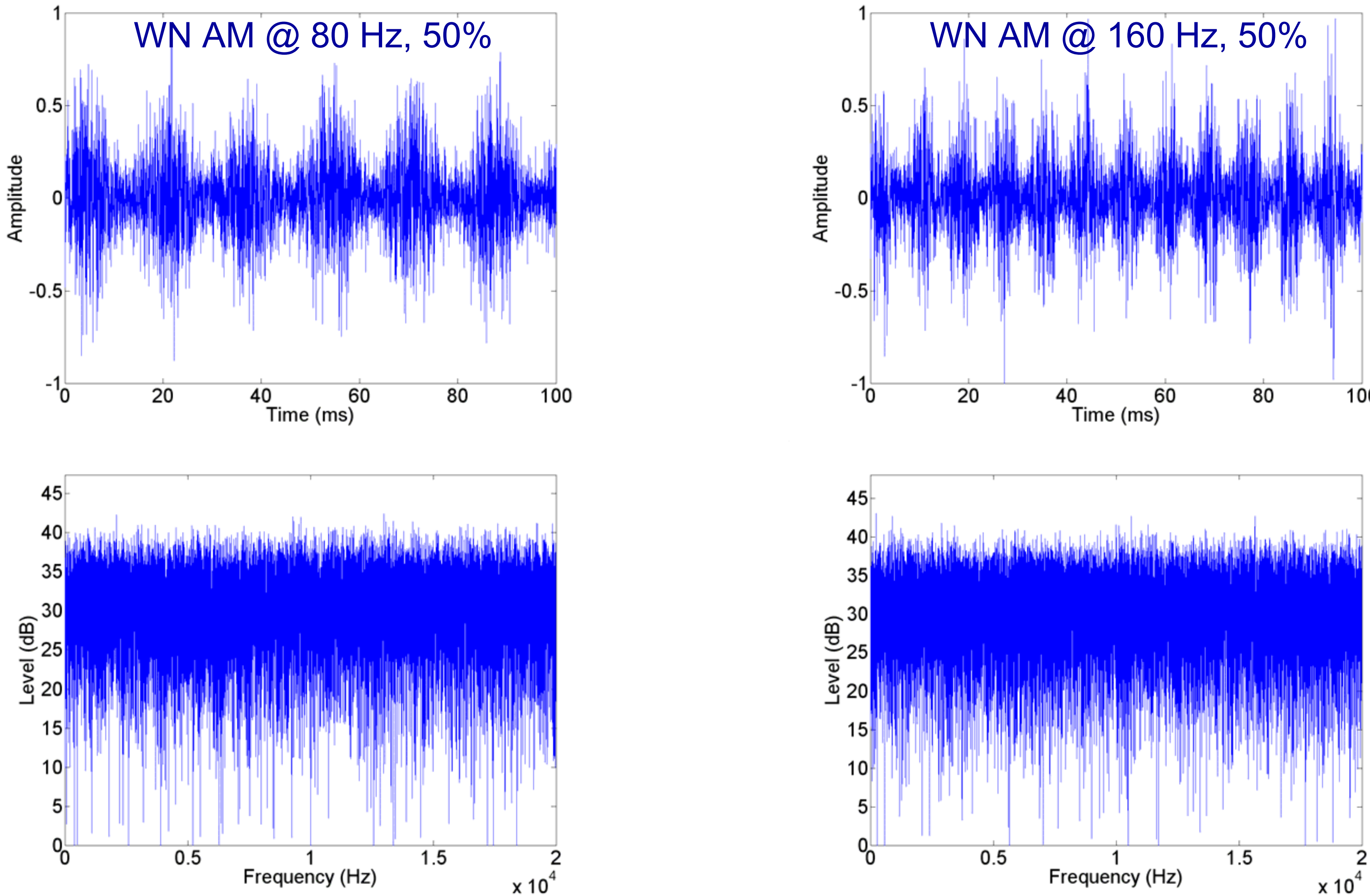
...but the effect is weaker than with resolved harmonics

Grimault, Micheyl, Carlyon *et al.* (2000)



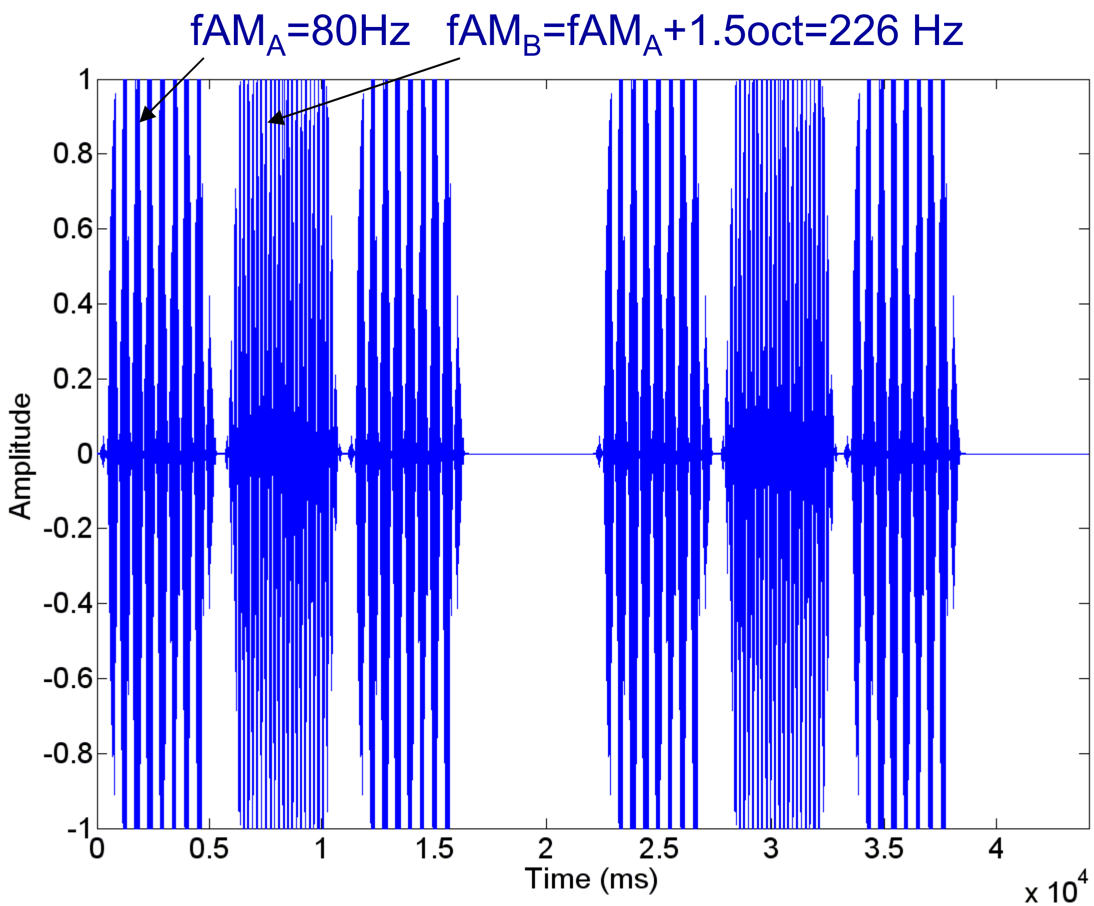
AM-rate-based streaming

Grimault, Bacon, Micheyl (2002)



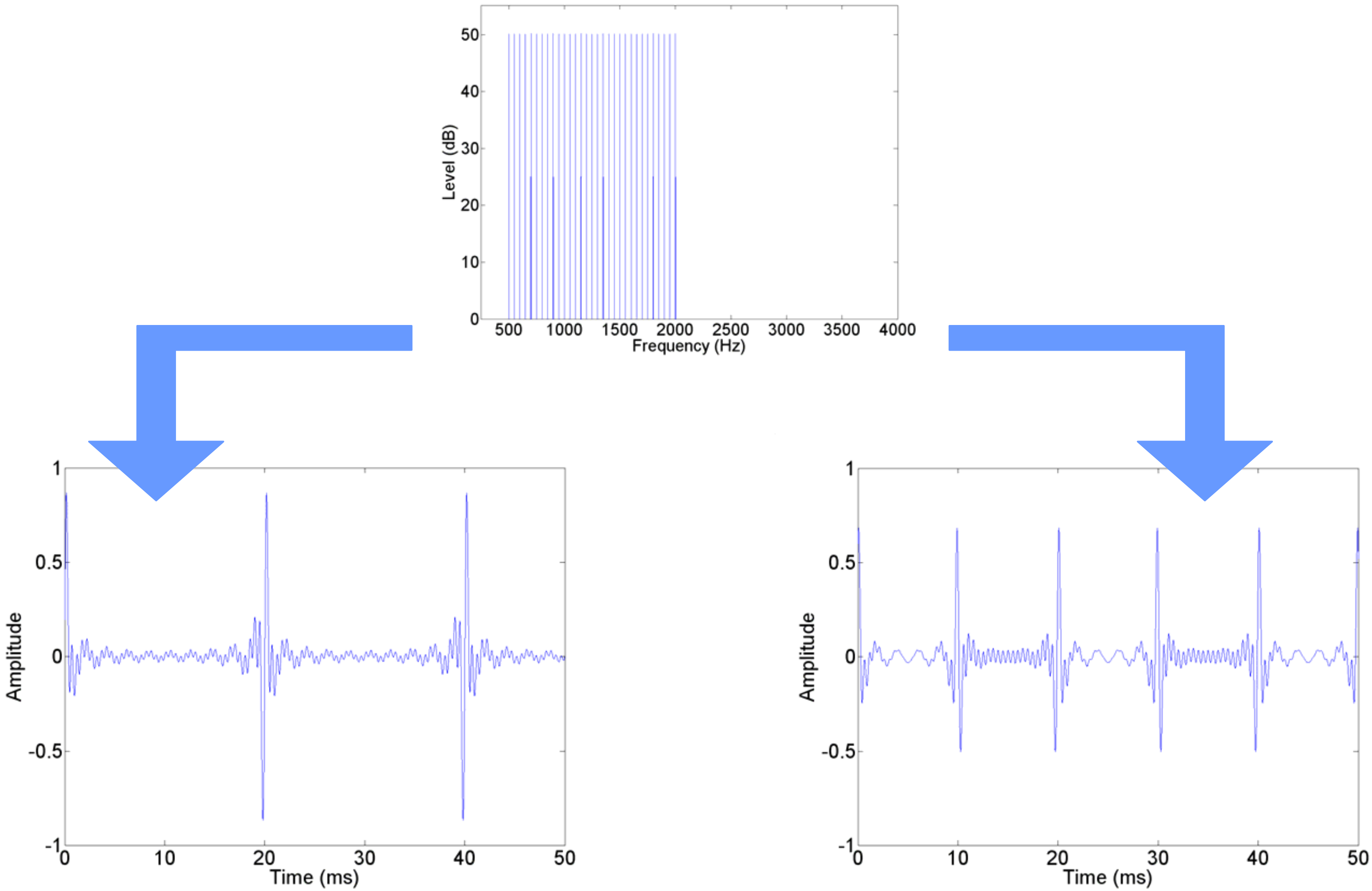
AM-rate-based streaming

Grimault, Bacon, Micheyl (2002)



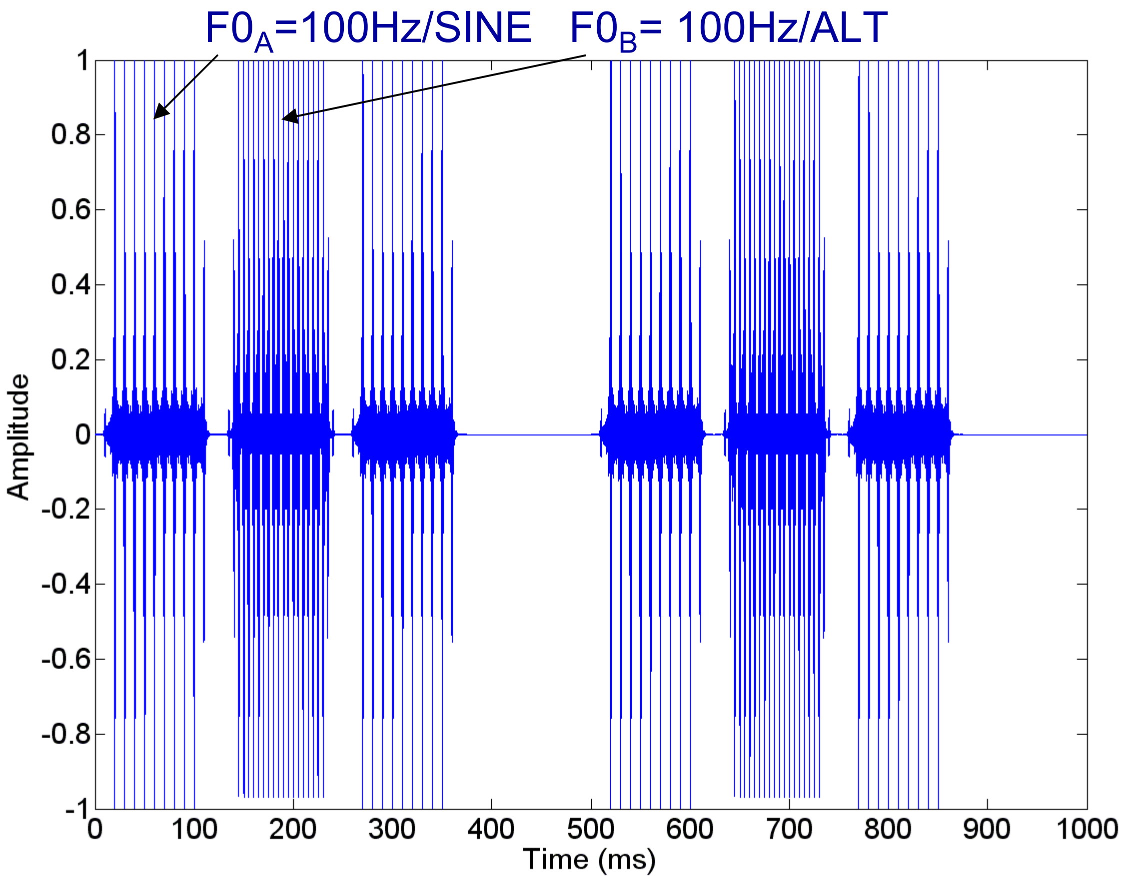
Phase-based streaming

Roberts, Glasberg, Moore (2002)



Phase-based streaming

Roberts, Glasberg, Moore (2002)



Conclusion:

The formation of auditory streams is determined primarily by peripheral frequency selectivity,

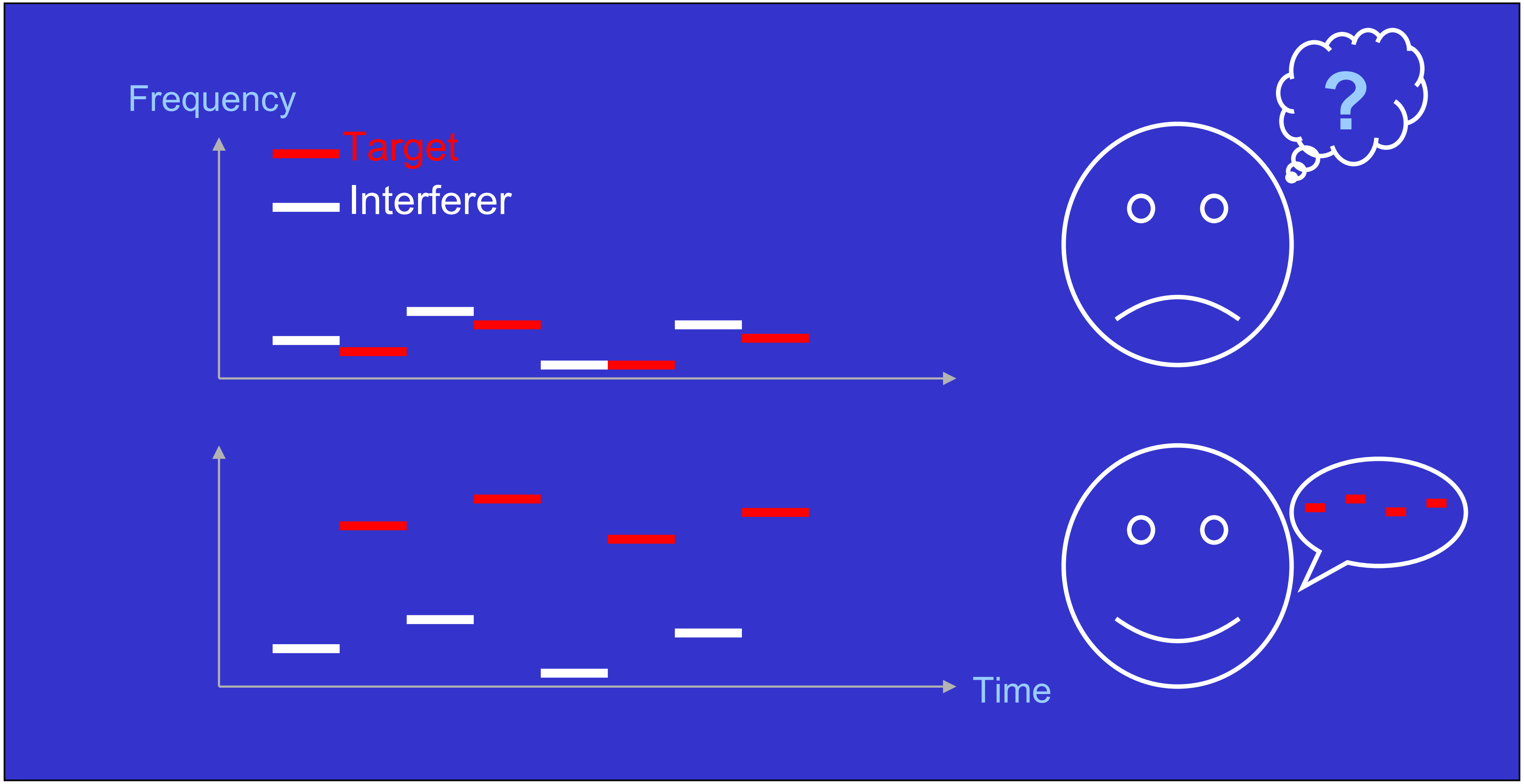
but some streaming may be produced even by sounds that excite the same peripheral channels

Does streaming influence other aspects of auditory perception?

Stream segregation can help...

Improved recognition of intervleaved melodies

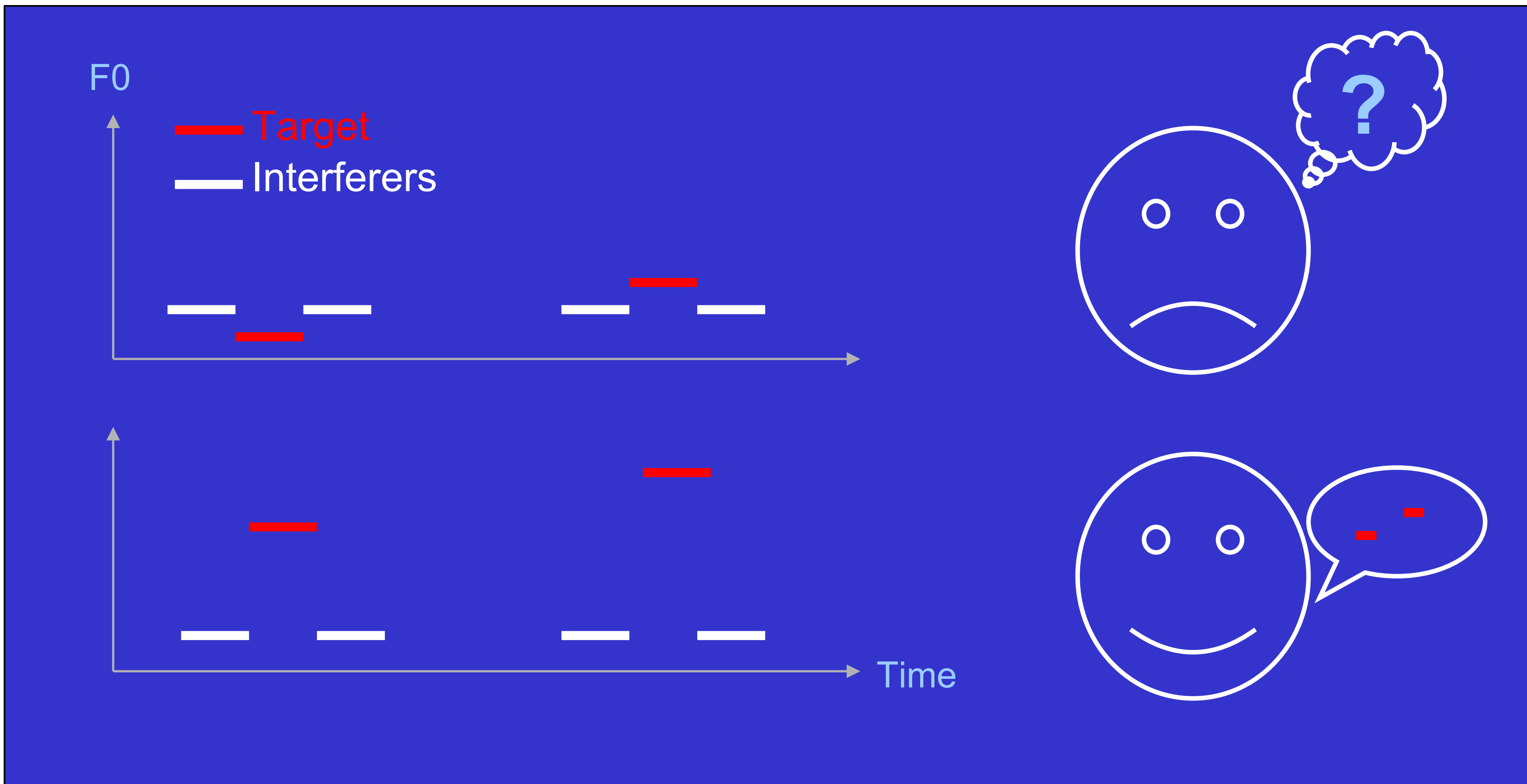
Dowling (1973), Dowling et al. (1987),
Hartmann & Johnson (1991), Vliegen & Oxenham (1999),
Iverson (1995), Cusack & Roberts (2000), Bey & McAdams (2002)



Stream segregation can help...

Improved (pitch) discrimination of target tones separated by extraneous tones

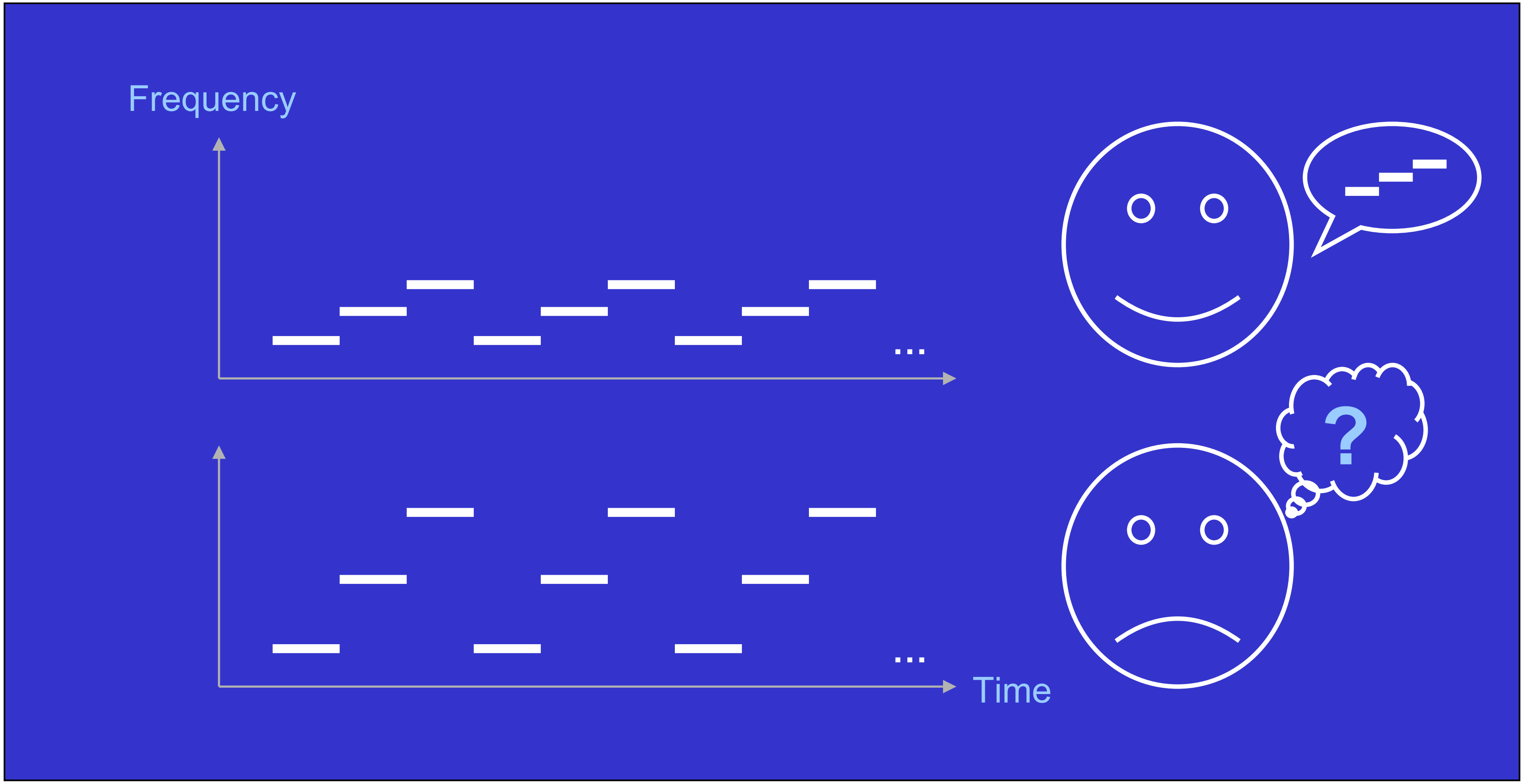
Jones, Macken, Harries (1997)
Micheyl & Carlyon (1998)
Gockel, Carlyon, & Micheyl (1999)



Stream segregation can harm...

Detrimental effect on temporal order identification

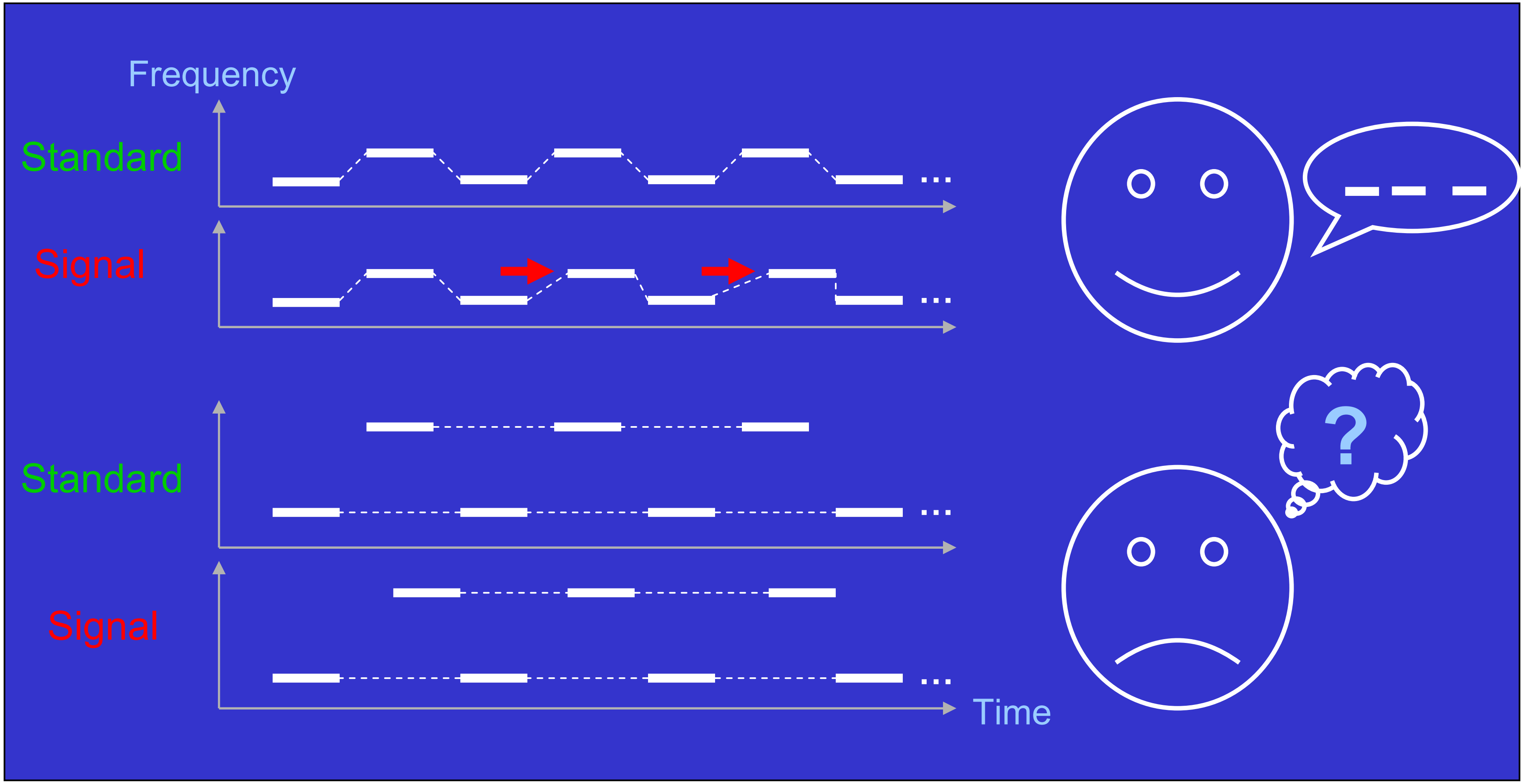
Bregman & Campbell (1971)



Stream segregation can harm...

Loss of fine temporal relationships

Brochard, Drake, Botte, & McAdams (1999)
Cusack & Roberts (2000)
Roberts, Glasberg, & Moore (2003)



References

Books, reviews on ASA:

- Darwin CJ & Carlyon RP (1995) Auditory grouping. In: Hearing (Ed. BJ Moore), Acad. Press, NY
- Bregman (1990) Auditory scene analysis. MIT Press, Cambridge MA.

Misc:

- Darwin CJ, Ciocca V. (1992) Grouping in pitch perception: effects of onset asynchrony and ear of presentation of a mistuned component. J Acoust Soc Am. 91, 3381-3390.
- Darwin CJ, Gardner RB. (1986) Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. J Acoust Soc Am. 79, 838-845.

On the neural mechanisms streaming:

- Fishman YI et al. (2001) Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hear Res. 151, 167-187.

Computer models of sound segregation:

- Cariani PA (2001) Neural timing nets. Neural Netw. 14, 737-753
- de Cheveigne A, et al. (1995) Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement. J Acoust Soc Am. 97, 3736-3748.
- Assmann PF, Summerfield Q. (1990) Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. J. Acoust. Soc. Am. 88, 680-697.