Auditory Perception Organization and Auditory Scene Analysis

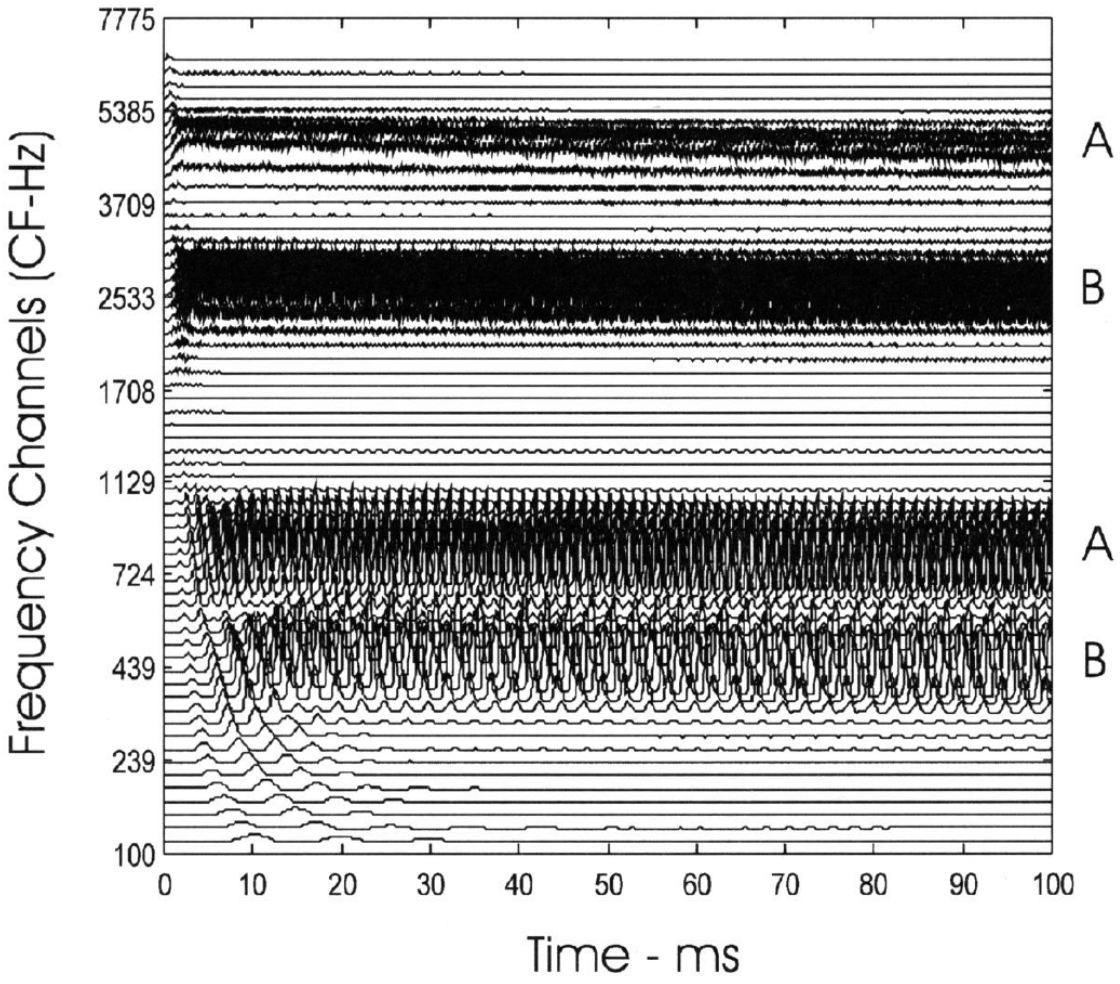
Wu Xihong

School of Artificial Intelligence

Need for sound segregation

- Ears receive mixture of sounds
- We hear each sound source as having its own appropriate timbre, pitch, location
- Stored information about sounds (eg acoustic/phonetic relations) probably concerns a single source
- Need to make single source properties (eg silence) explicit

Auditory stream analysis



Two sound sources A (450 and 2500 Hz) and B (725 and 5000 Hz).

How to analyze the mixture of sounds?

Mechanisms of segregation

- Primitive grouping mechanisms based on general heuristics such as harmonicity and onset-time - "***bottom-up***" / "pure audition"

- Schema-based mechanisms based on specific knowledge (general speech constraints?) - "***top-down***".

Segregation of simple musical sounds

- Successive segregation
  - Different frequency (or pitch)
  - Different spatial position
  - Different timbre

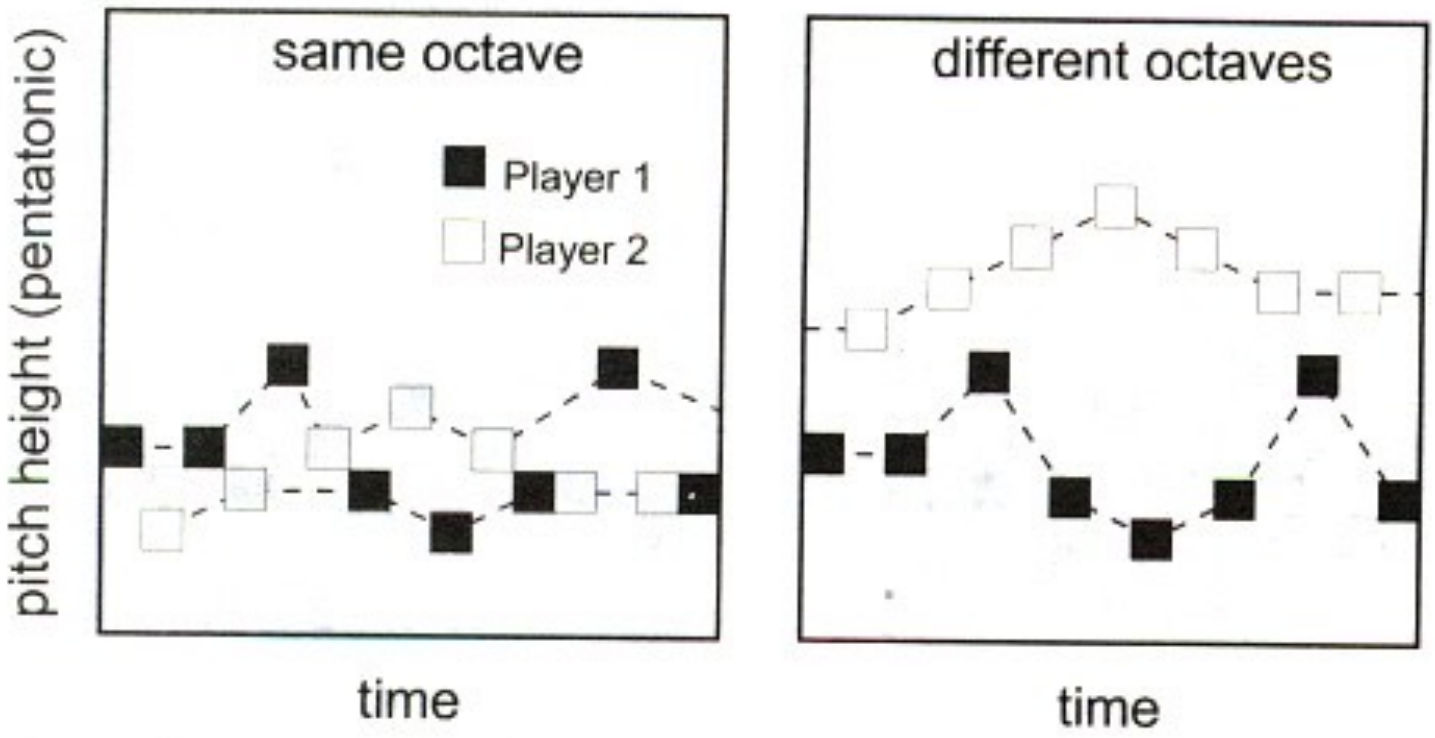
- Simultaneous segregation
  - Different onset-time
  - Irregular spacing in frequency
  - Location (rather unreliable)
  - Uncorrelated FM **not** used

Successive grouping by frequency

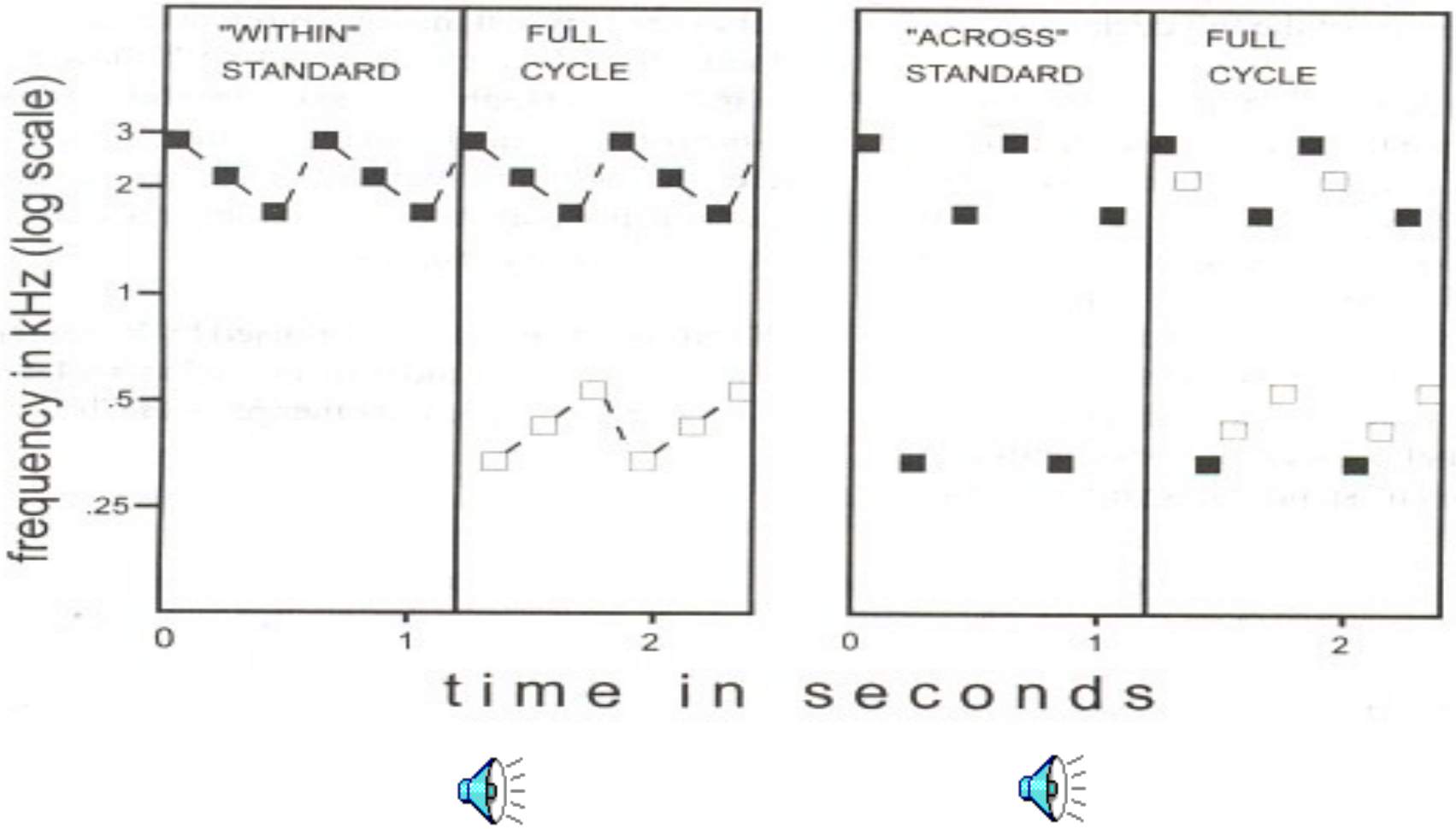
Streaming in African xylophone music

Successive grouping by frequency

Bugandan xylophone music: “Ssematimba ne Kikwabanga”

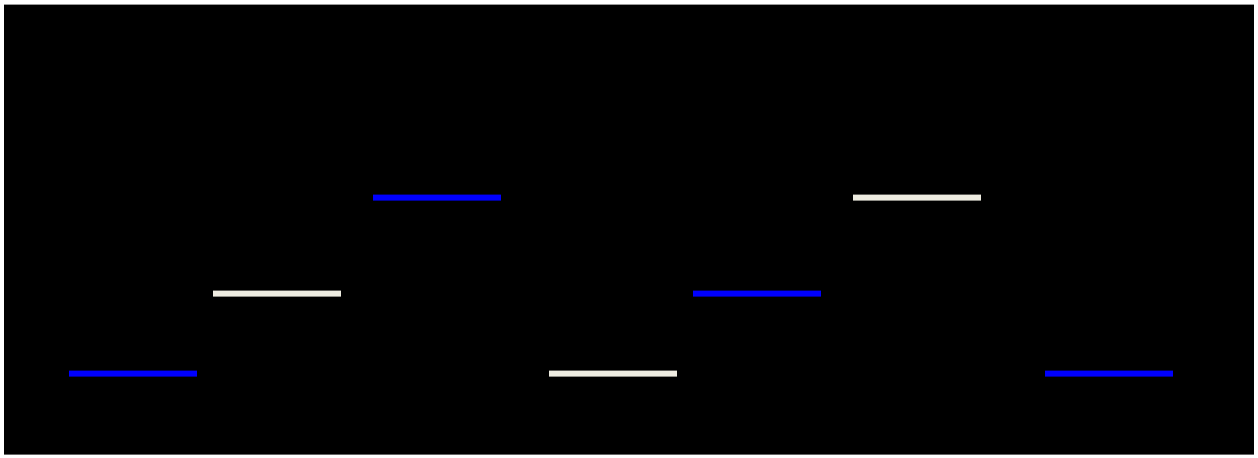


Successive grouping by frequency

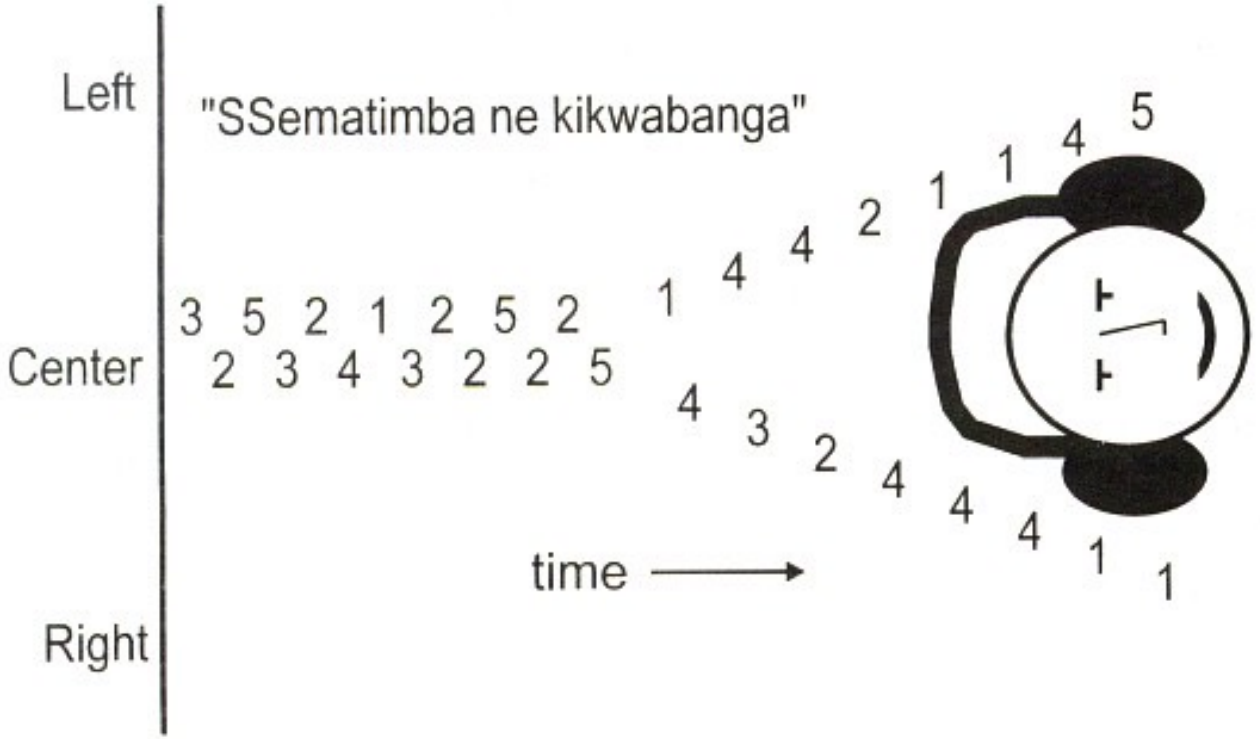


Successive grouping by timbre

Wessell illusion



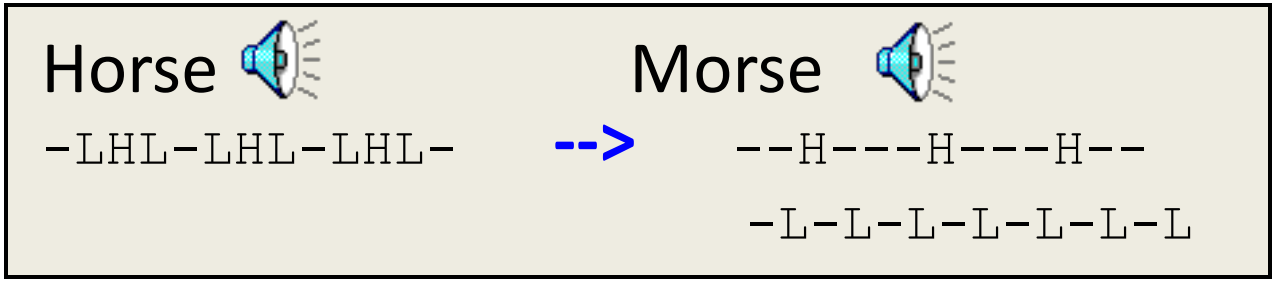
Successive groupingby spatial separation



Some interesting points:

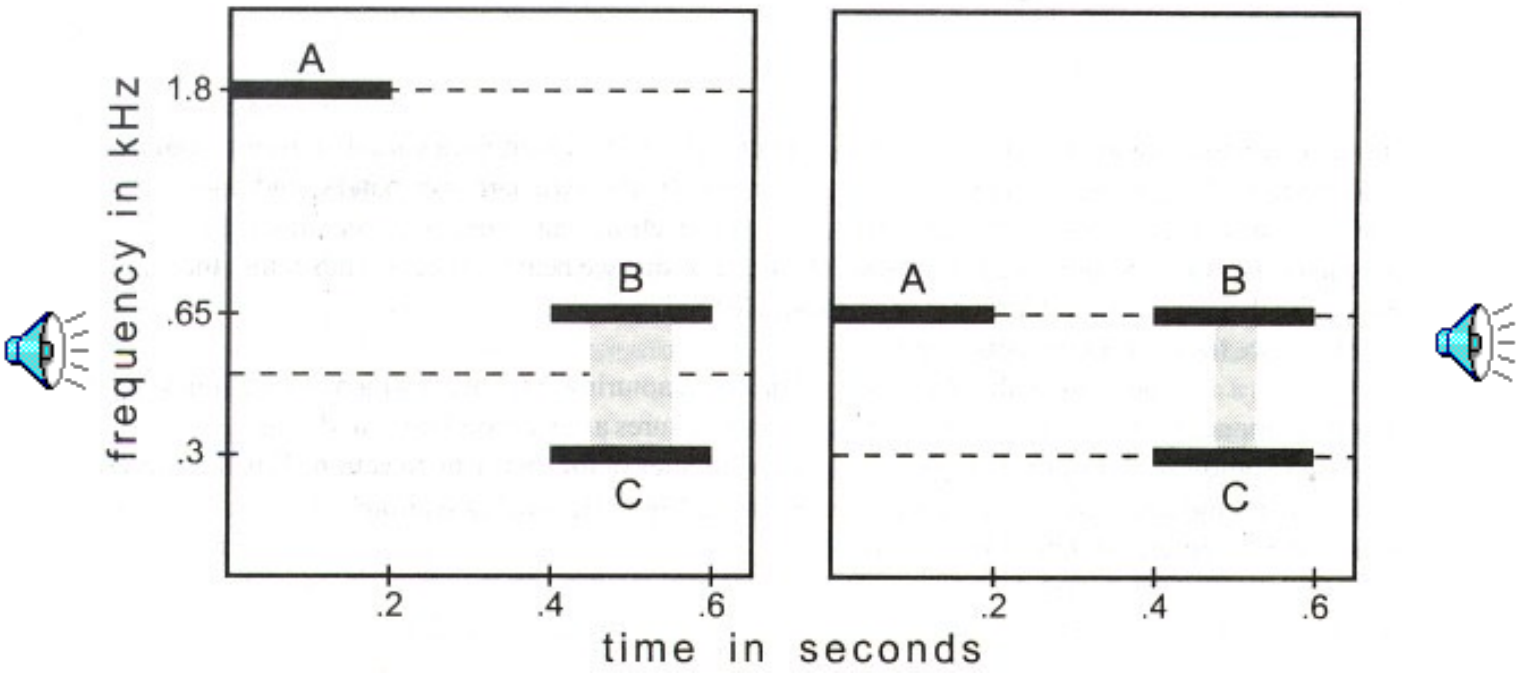
- Sequential streaming may require attention - rather than being a pre-attentive process.

Attention necessary for build-up of streaming (Carlyon et al, JEP:HPP 2000)



- Horse -> Morse takes a few seconds to segregate
- These have to be seconds spent attending to the tone stream
- Does this also apply to other types of segregation?

Capturing a component from a mixture by frequency proximity



A-B
A-BC

Freq separation of AB

Harmonicity & synchrony of BC

Disjoint Allocation?

Bregman & Pinker, 1978, Canad J Psychol

Simultaneous grouping

What is the timbre / pitch / location of a particular sound source ?

Important grouping cues

- continuity (or repetition) “Old + New”
- onset time
- harmonicity (or regularity of frequency spacing)

Bregman's Old + New principle

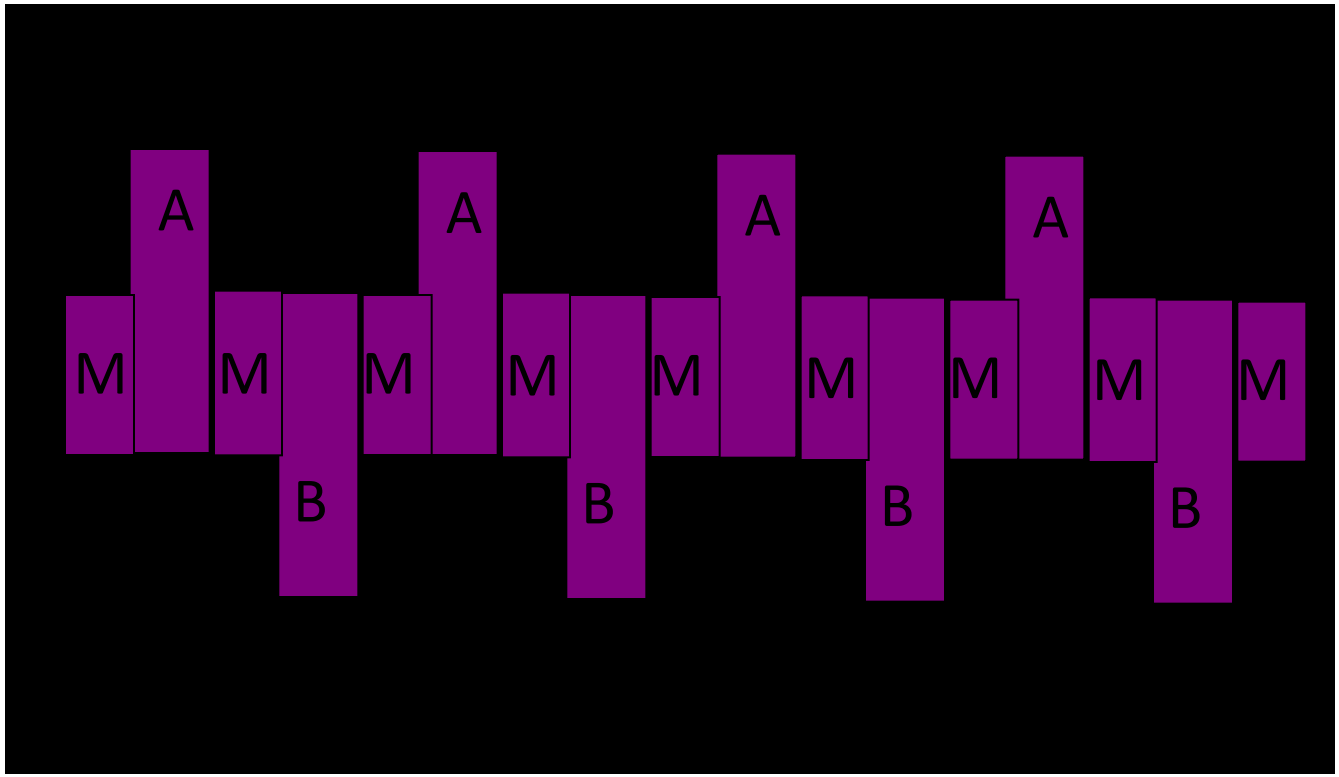
Stimulus: A followed by A+B

-> Percept of

A as continuous (or repeated)

with B added as separate percept

Old+New Heuristic

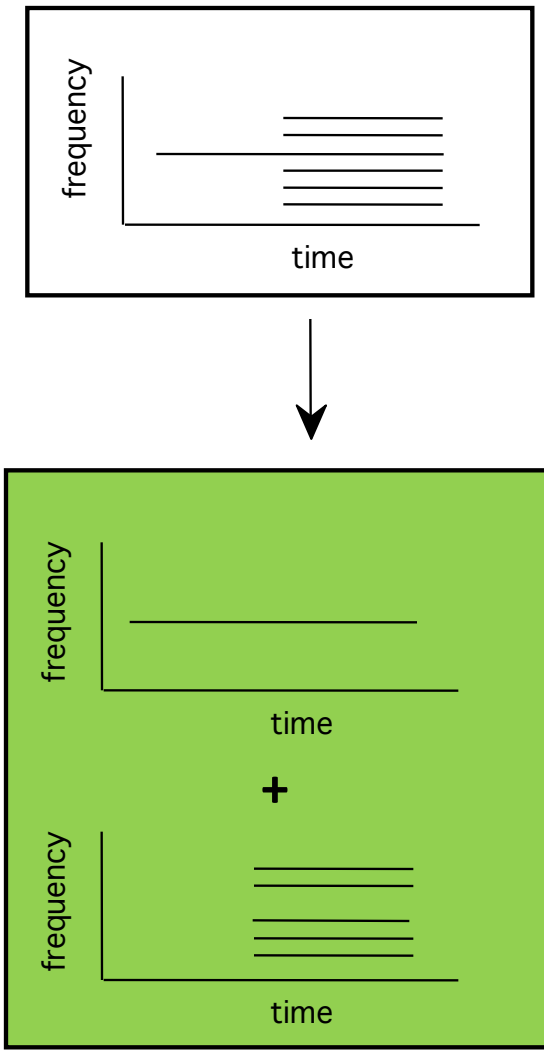


A
MAMB

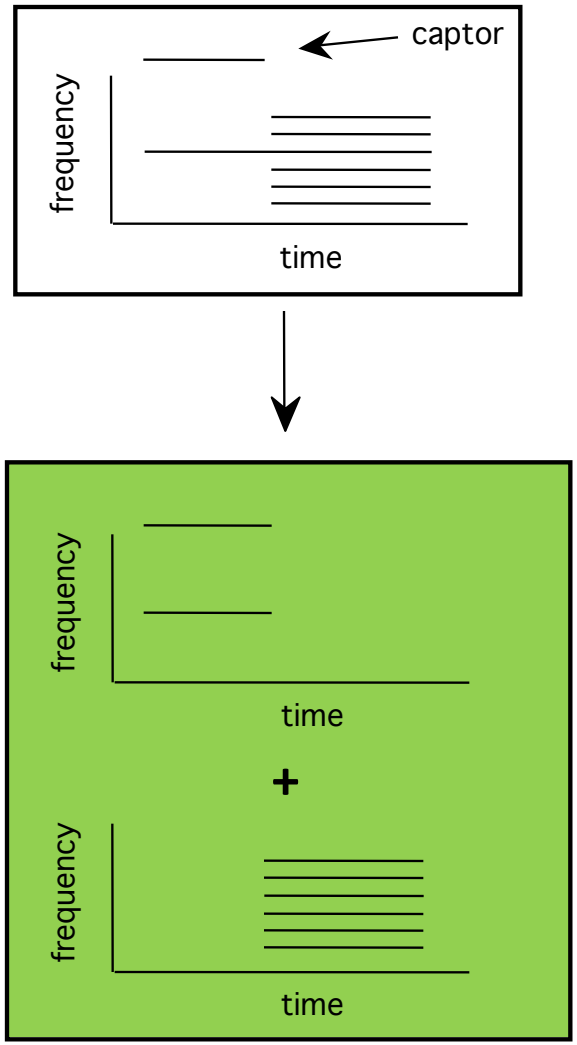
B
MAMB

Grouping & vowel quality

continuation removed from vowel

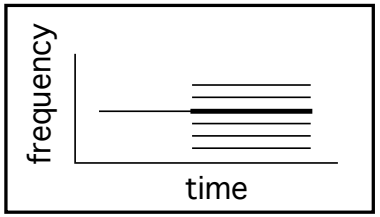


continuation not removed from vowel



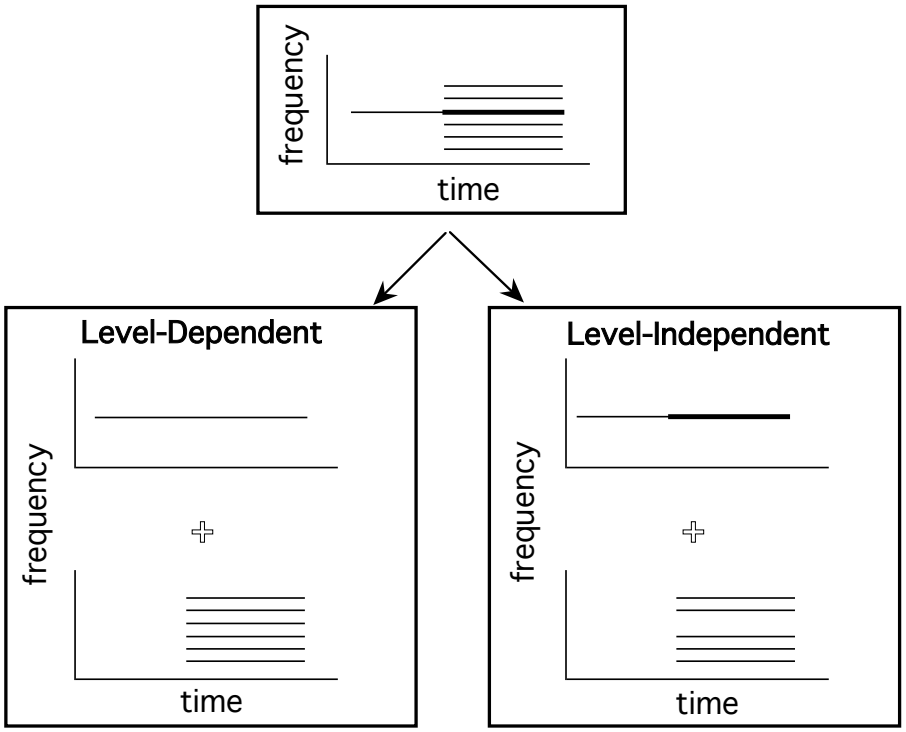
Onset-time: allocation is subtractive not exclusive

- Bregman's Old-plus-New heuristic



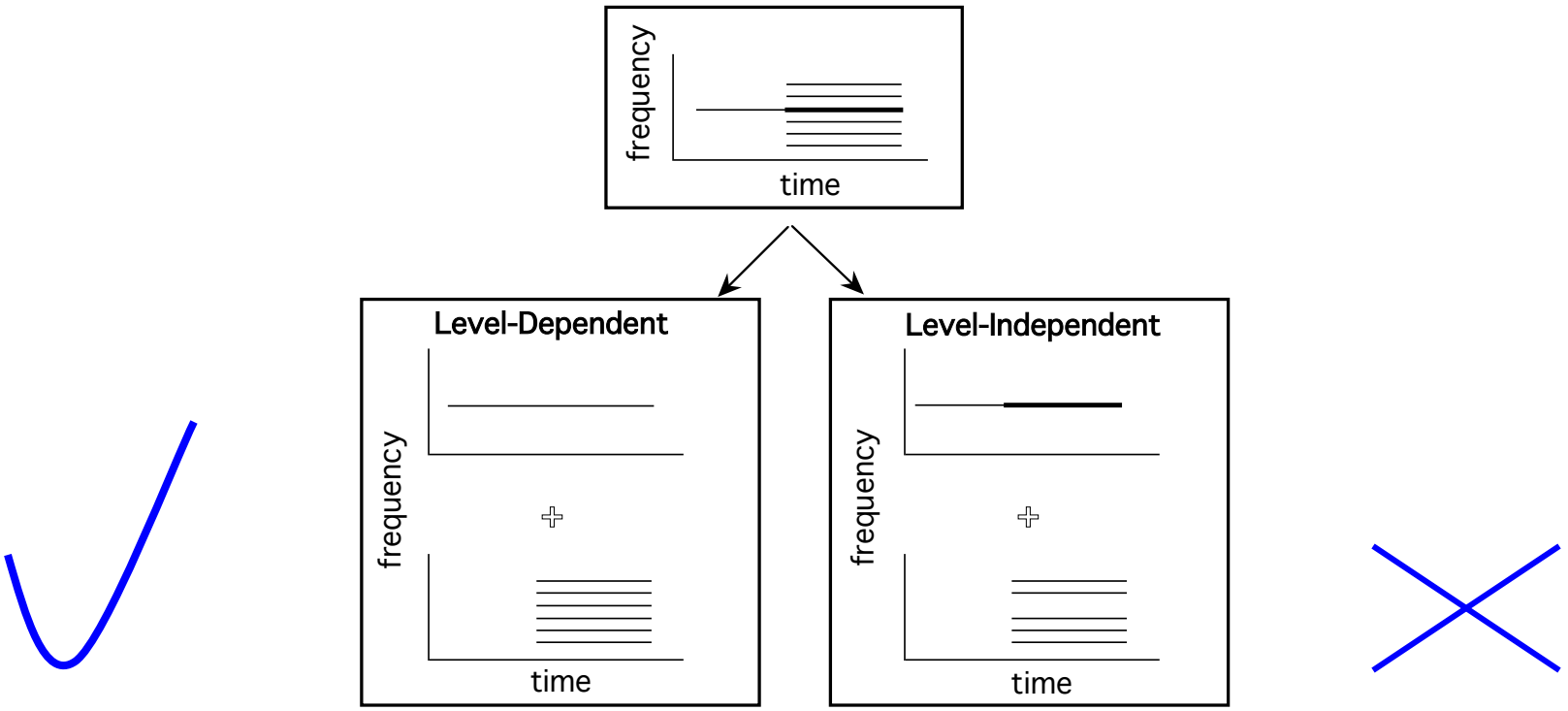
Onset-time: allocation is subtractive not exclusive

- Bregman's Old-plus-New heuristic



Onset-time: allocation is subtractive not exclusive

- Bregman's Old-plus-New heuristic



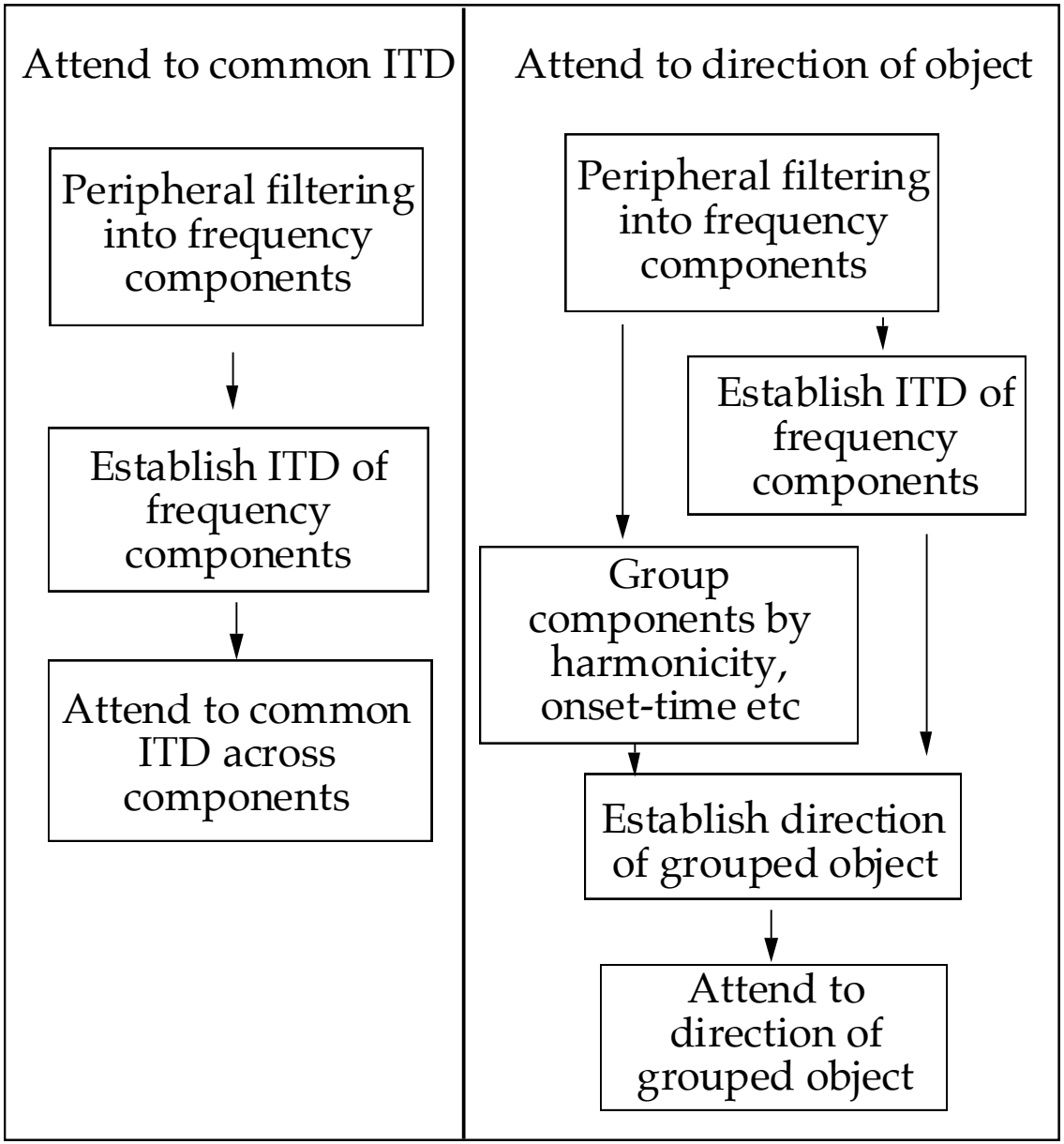
- Indicates importance of coding change.

Role of localisation cues

What role do localisation cues play in helping us to hear one voice in the presence of another ?

- Head shadow increases S/N at the nearer ear (Bronkhurst & Plomp, 1988).
  - … but this advantage is reduced if high frequencies inaudible (B & P, 1989)
- But do localisation cues also contribute to selectively grouping different sound sources?

Two models of attention



Some interesting points:

- Sequential streaming may require attention - rather than being a pre-attentive process.
- Parametric behaviour of grouping depends on what it is for.
- Not everything that is obvious on an auditory spectrogram can be used :
  - FM of Fo irrelevant for segregation (Carlyon, JASA 1991; Summerfield & Culling 1992)
- Although we can group sounds by ear, ITDs by themselves remarkably useless for simultaneous grouping.
  - Group first then localise grouped object.

The importance of Grouping

classify

group

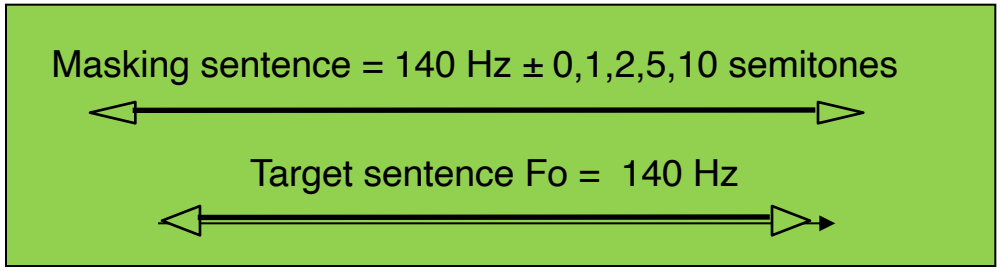
ΔFo between two sentences

(Bird & Darwin 1998; after Brokx & Nooteboom, 1982)

Two sentences (same talker)

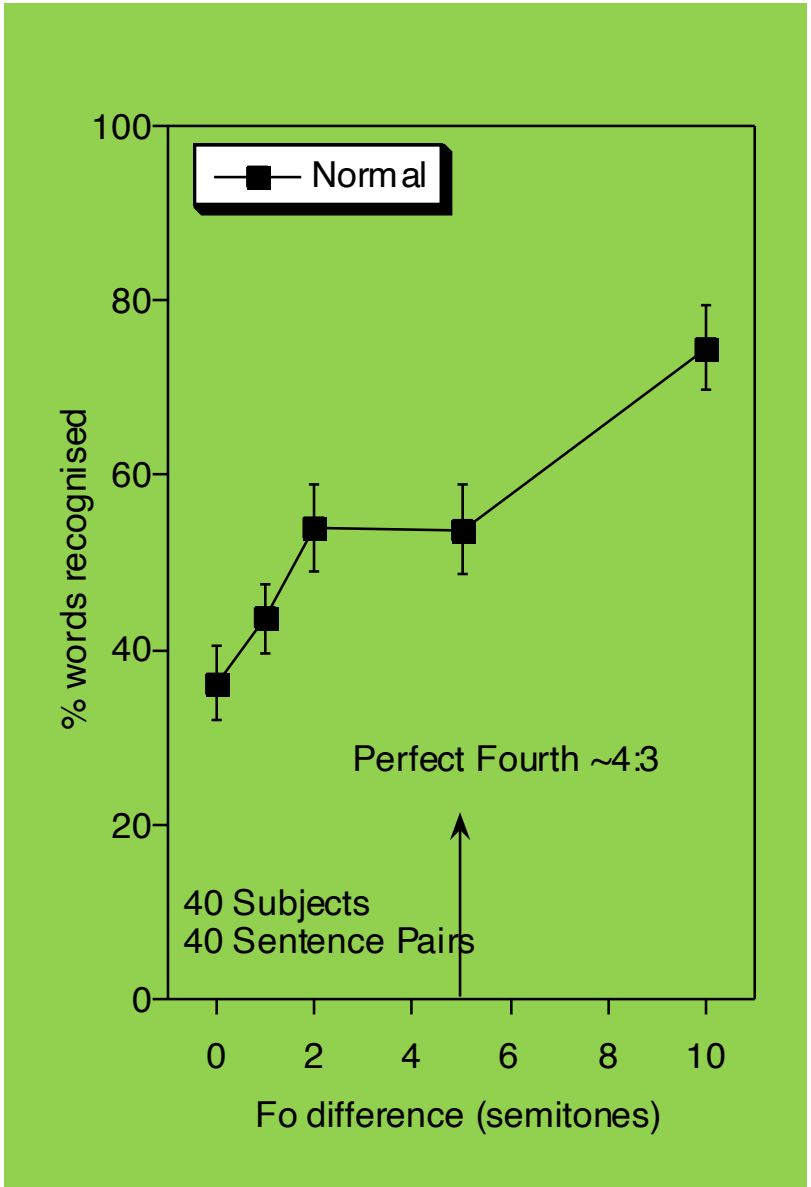
- only voiced consonants
- with very few stops

Thus maximising Fo effect

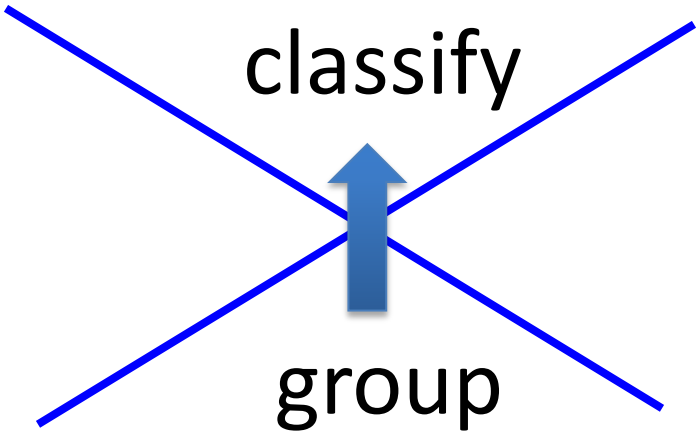


Task: write down target sentence

Replicates & extends Brokx & Nooteboom

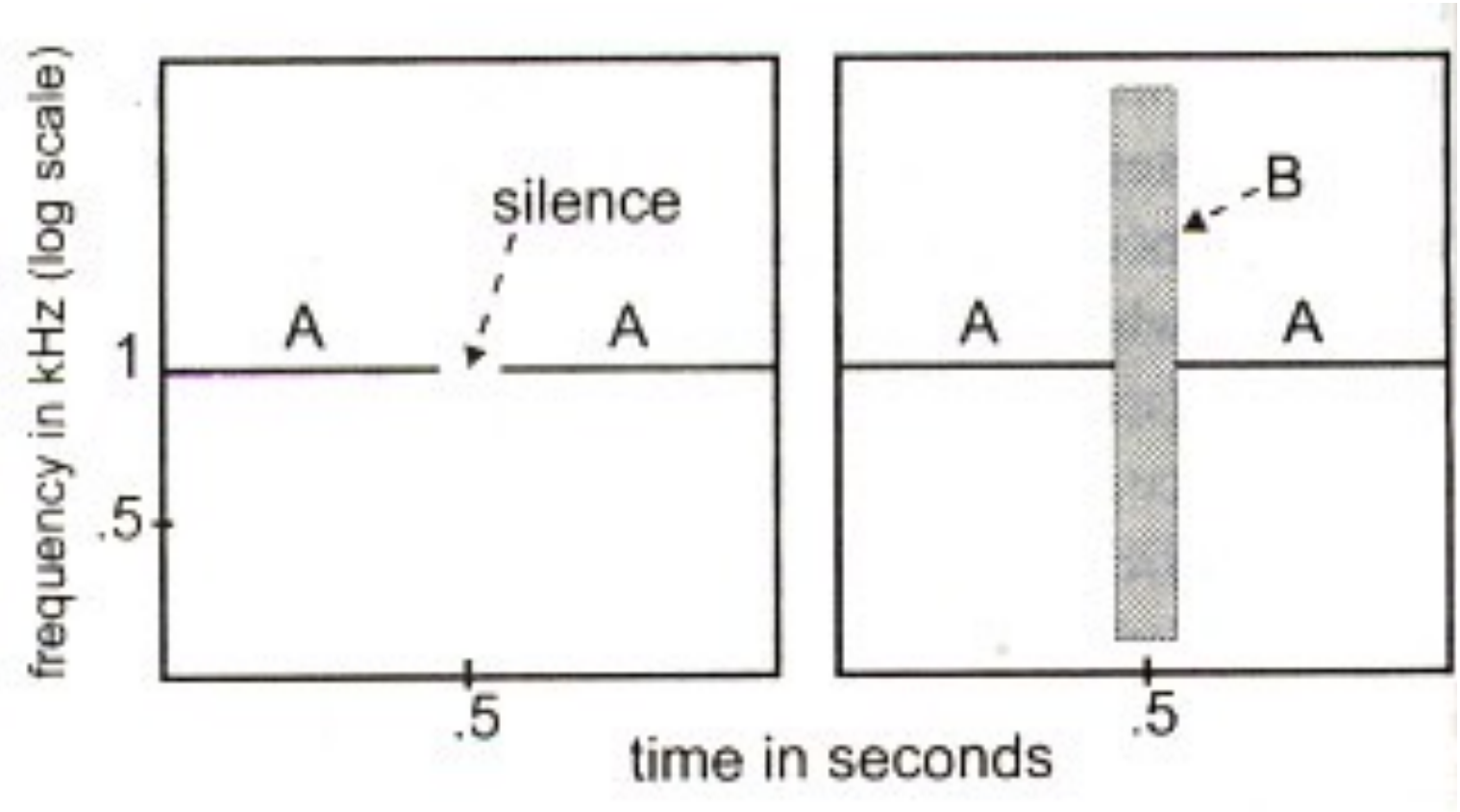


Grouping not absolute and independent of classification



Apparent continuity

If B would have masked if it HAD been there, then you don't notice that it is not there.



"栅栏效应"

Mechanisms of segregation

- Primitive grouping mechanisms based on general heuristics such as harmonicity and onset-time - "***bottom-up***" / "pure audition"

- Schema-based mechanisms based on specific knowledge (general speech constraints?) - "***top-down***.

Both approaches could be true!

Bregman long summary

Cues used by the ASA process

* The perceptual segregation of sounds in a sequence depends upon differences in their frequencies, pitches, timbres (spectral envelopes), center frequencies (of noise bands), amplitudes, and locations, and upon sudden changes of these variables. Segregation also increases as the duration of silence between sounds in the same frequency range gets longer.
* The perceptual fusion of simultaneous components to form single perceived sounds depends on their onset and offset synchrony, frequency separation, regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation, and parallel gliding of components. [Note to physicists: All these cases of fusion can be obtained at room temperature.]
* Different cues for stream segregation compete to control the grouping, and different cues have different strengths.
* Primitive grouping occurs even when the frequency and timing of the sequence is unpredictable.
* An increased biasing toward stream segregation builds up with longer exposure to sounds in the same frequency region.
* Stream segregation is context-dependent, involving the competition of alternative organizations,

Bregman long summary

Effects of ASA on perception

- A change in perceptual grouping can alter the perception of rhythms, melodic patterns, and overlap of sounds.
- Patterns of sounds whose members are distributed into more than one perceptual stream are much harder to perceive than those wholly contained within a single stream.
- Perceptual organization can affect perceived loudness and spatial location.
- The rules of ASA try to prevent the crossing of streams in frequency, whether the acoustic material is a sequence of discrete tones or continuously gliding tones.
- Known principles of ASA can predict the camouflage of melodies and rhythms when interfering sounds are interspersed or mixed with a to-be-recognized sequence of sounds.
- The apparent continuity of sounds through masking noise depends on ASA principles. Stimuli have included frequency glides, amplitude-varying tones, and narrow-band noises.
- A perceptual stream can alter another one by capturing some of its elements.

Bregman long summary

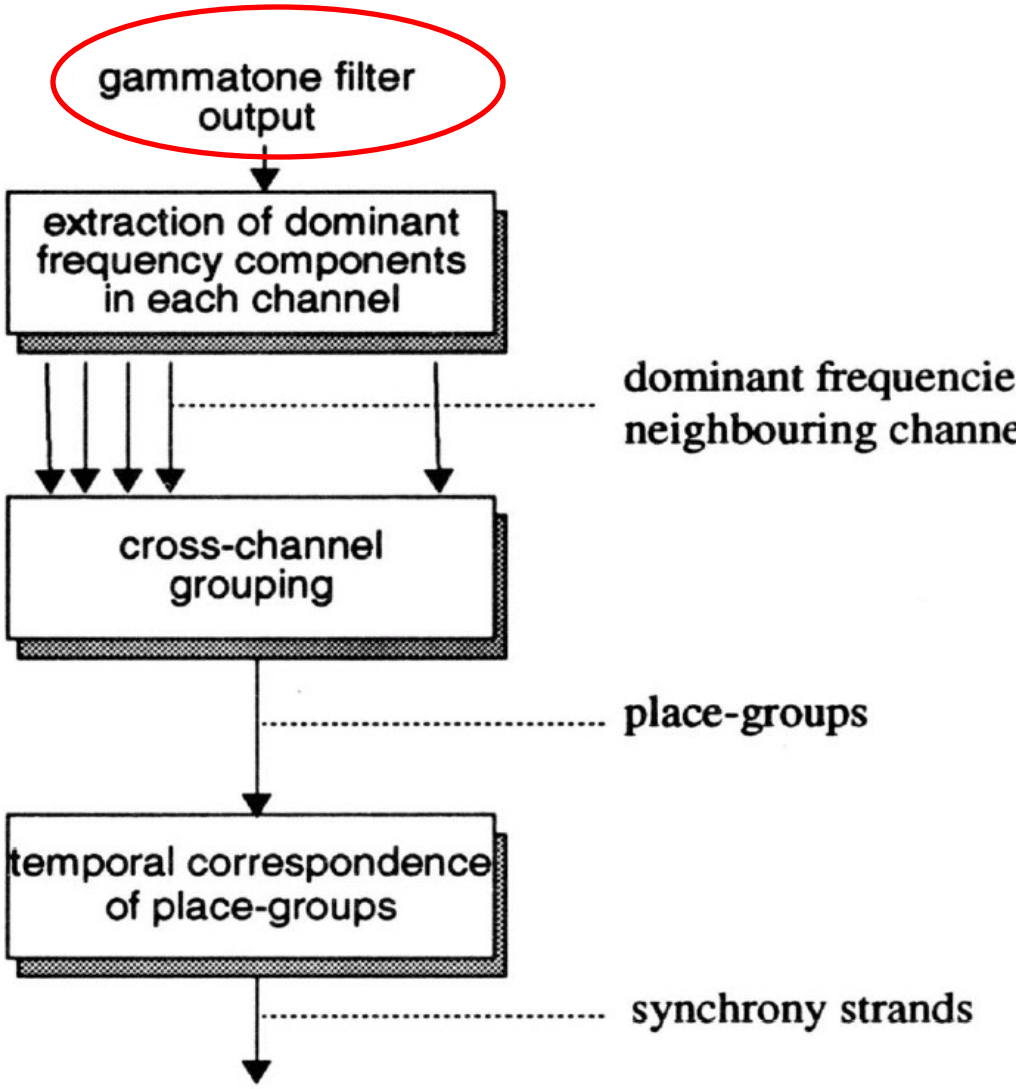
- The apparent spatial position of a sound can be altered if some of its energy becomes grouped with other sounds,
- Comodulation masking release (CMR) does not make the presence of the target more discriminable by simply altering the timbre of the target-masker mixture. It actually increases the subjective experience that the target is present.
- Sequential capturing can affect the perception of speech, specifically the integration of perceptually isolated components in speech-sound identification.
- The segregation of vowels increases when they have different pitches and different pitch transitions. We have looked at synthetic vowels that do or do not have harmonic relations between frequency components,
- ASA principles help explain the construction of music, e.g., rules of voice leading.
- ASA principles are used intuitively by composers to control dissonance in polyphonic music.
- The segregation of streams of visual apparent motion works in exactly the same way as auditory stream segregation.

Computational Auditory Scene Analysis

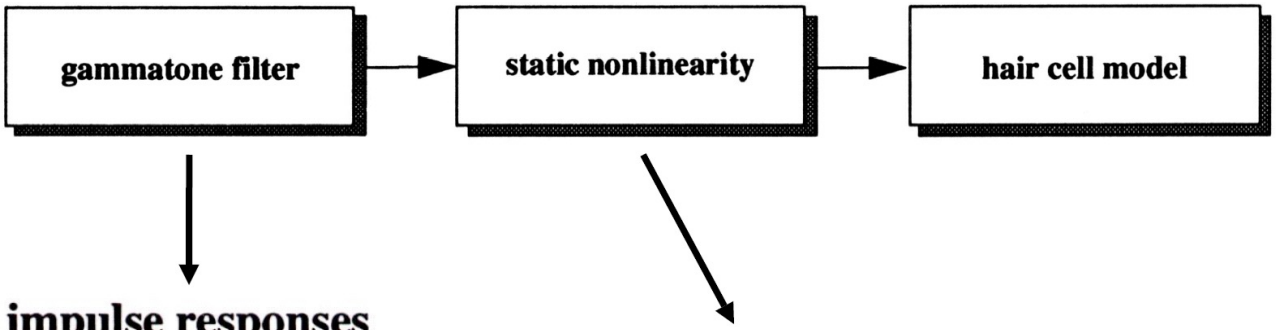
Computational problems in ASA

- segment or separate the sounds
- factor out background noise
- factor out head and pinnae filtering
- group sound elements

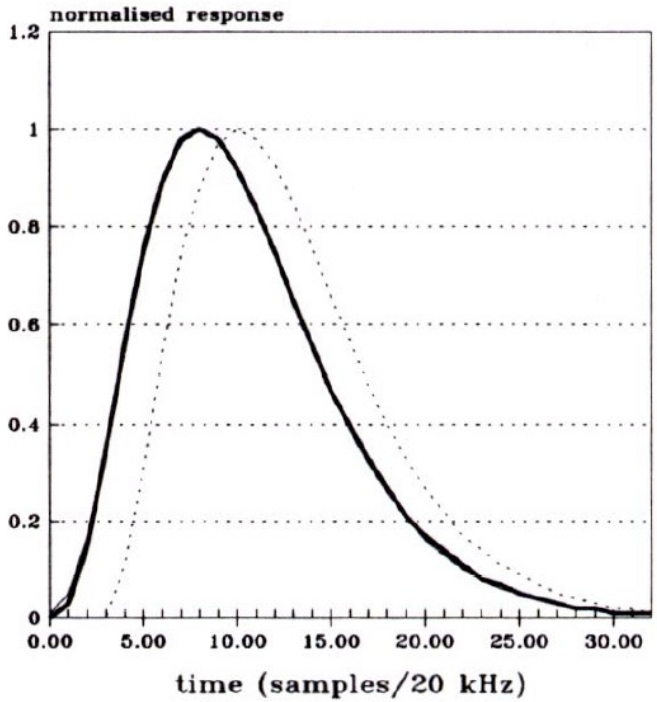
Framework of Computational Auditory Scene Analysis

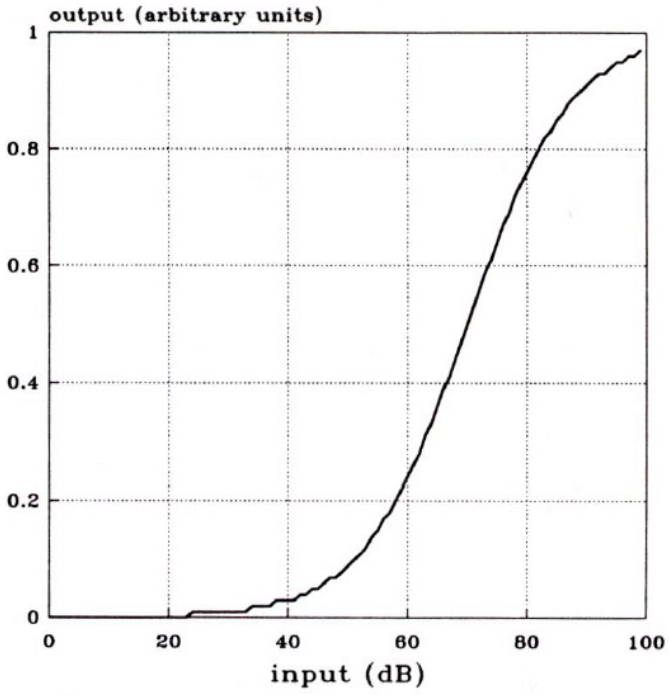


Auditory periphery model

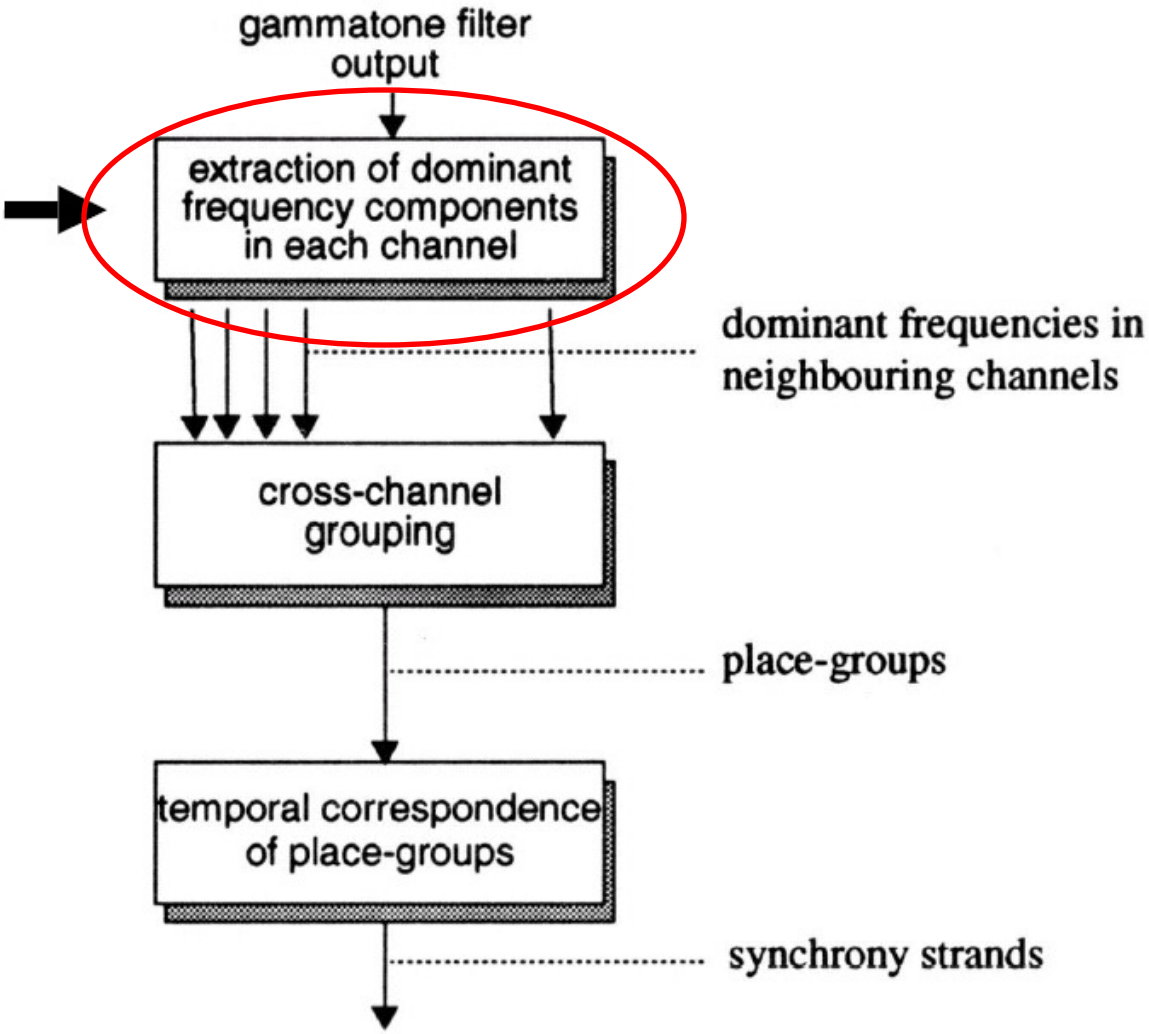


impulse responses





Next stage: extraction of dominant frequency



Dominant frequency estimation

Want a pure frequency representation in order to apply grouping rules.

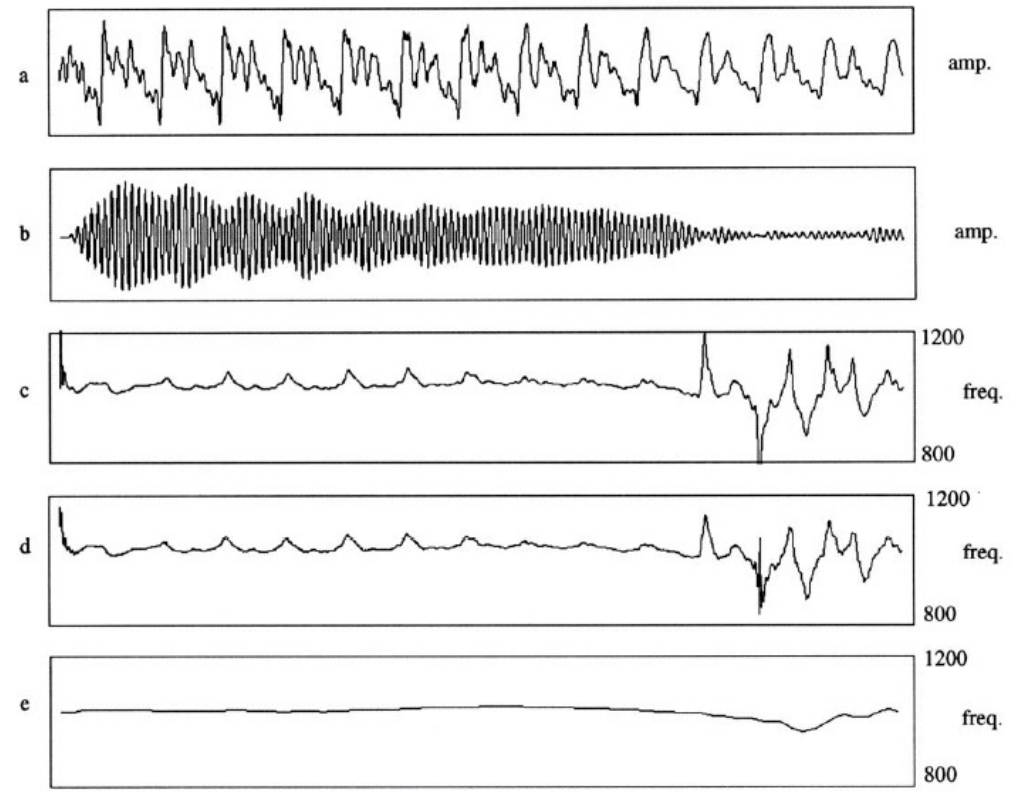


Figure 3.2 Dominant frequency estimation: *a*: waveform for the vowel /ae/; *b*: auditory filter output (CF: 1 kHz); *c*: $\nu(t)$ by analytic signal method; *d*: instantaneous frequency by linear prediction analysis; *e*: $\hat{\nu}(t)$.

a) waveform for the vowel /ae/

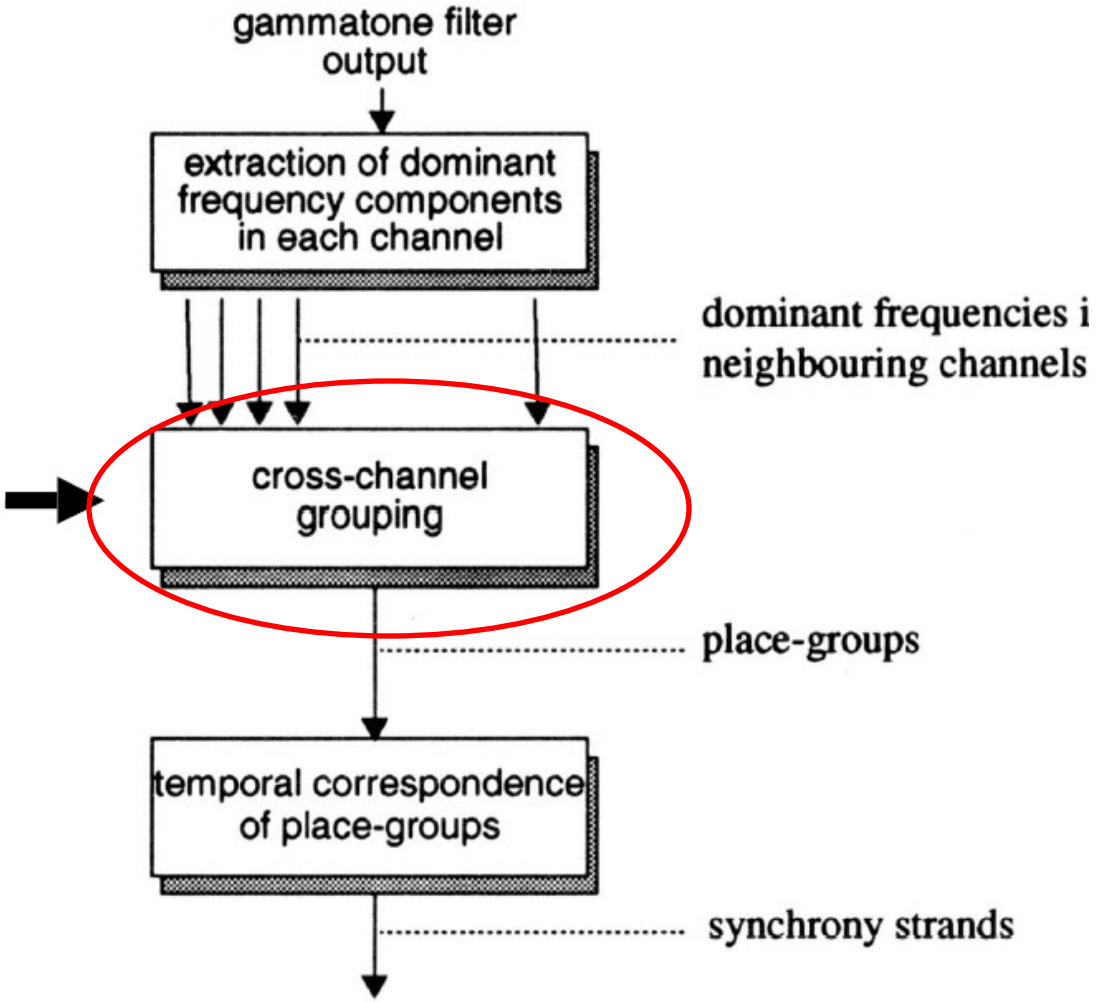
b) auditory filter output for 1kHz center frequency

c) instantaneous frequency from Hilbert transform of gammatone filter

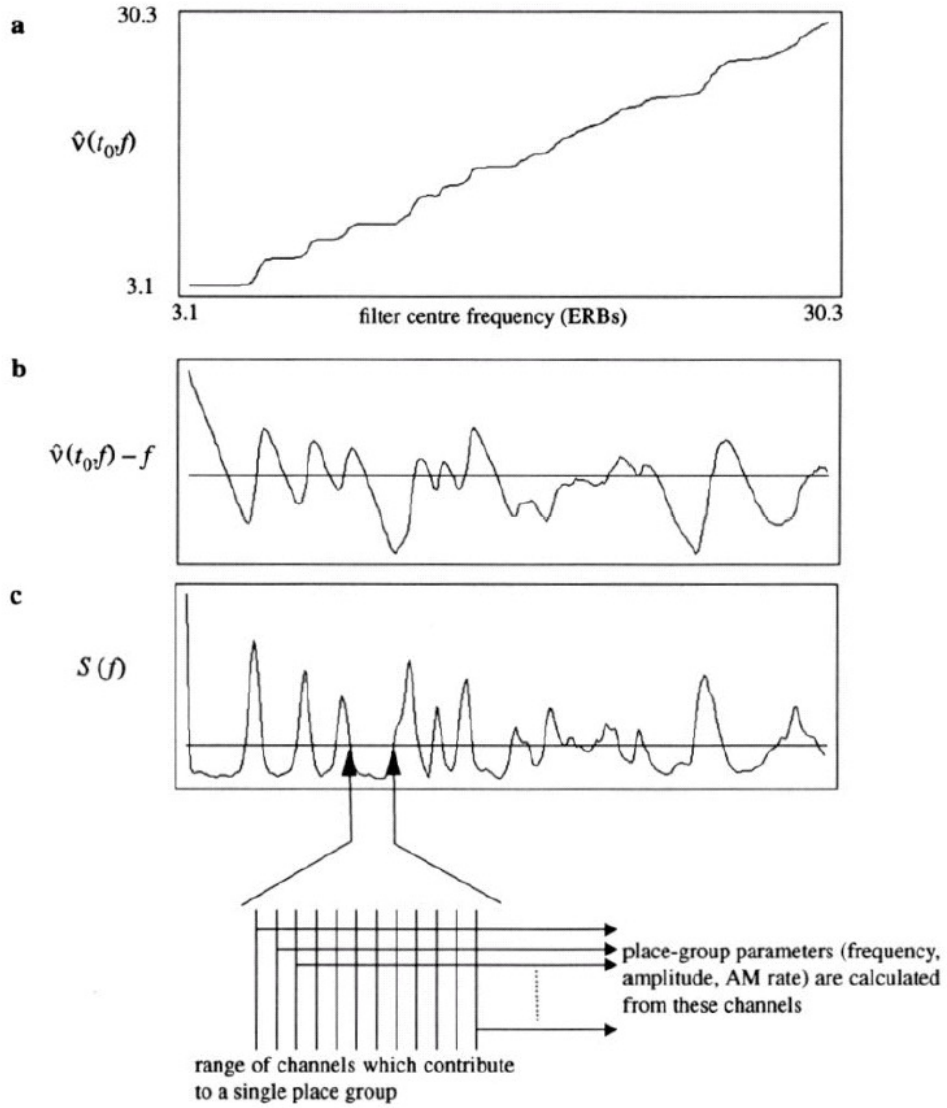
d) an alternative method of estimating instantaneous frequency

e) smoothed instantaneous frequency median filtering over 10ms, plus linear smoothing also allows for data reduction: samples are collapsed over 1ms window

Next stage: grouping frequency channels



Calculation of place groups



What frequencies belong to the same sound?

• Goal of this stage is to locate and characterize intervals along filterbankwith synchronous activity.

(a) Notice that center frequency is not distributed evenly due the median filtering in the frequency estimation stage.

(b) The instantaneous frequency varies around the estimated frequency.
To group all channels centered around the "dominant frequency", i.e. grouping by spectral location.

(c) *S(f)* is a smoothed frequency derivative estimate. Channels at the minimum are grouped together.

Plot of place groups for a speech waveform

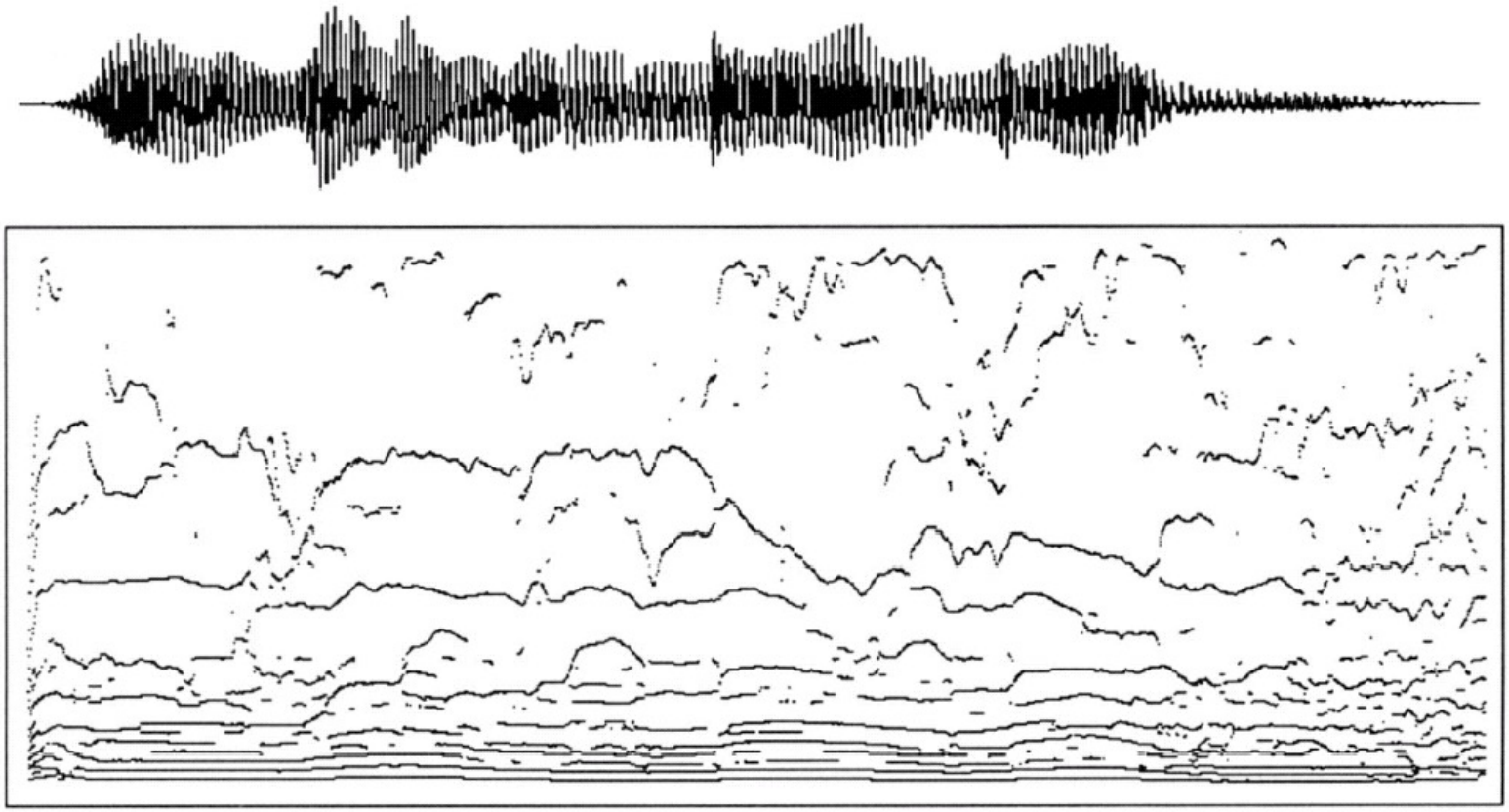
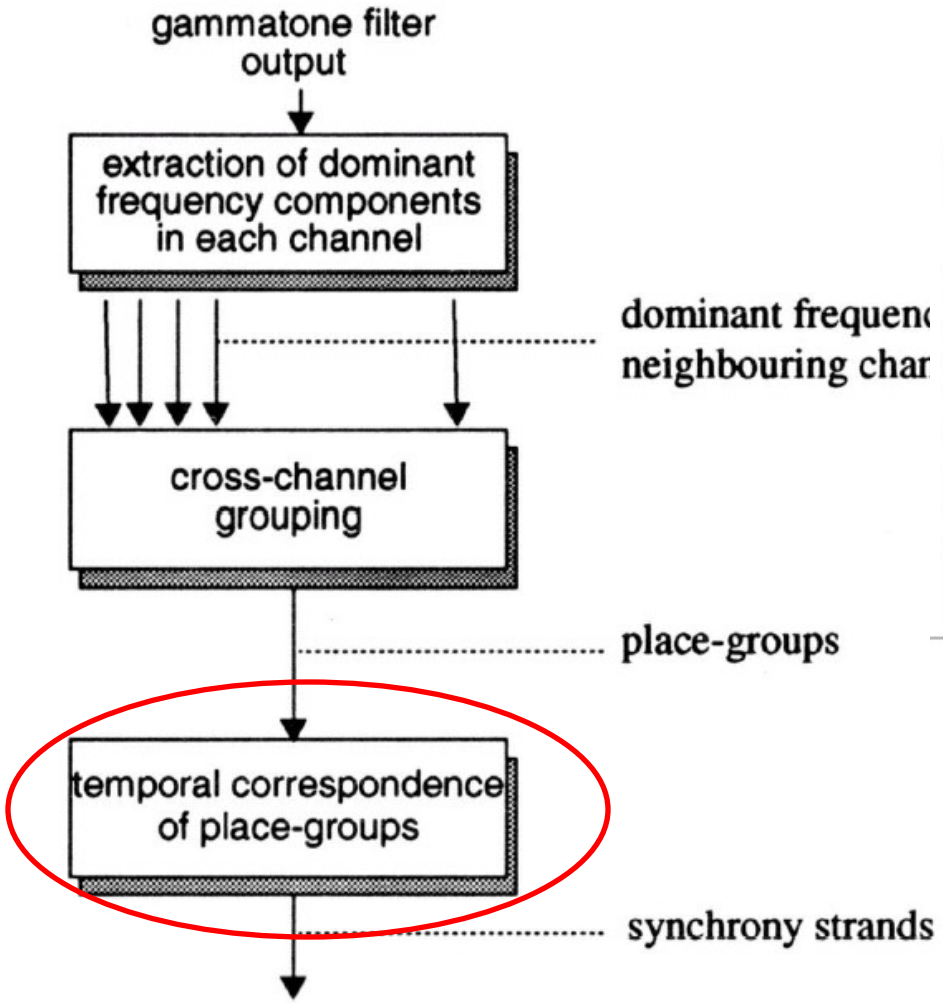


Figure 3.4 Place-groups for the utterance whose waveform is shown. Frequency axis is linear in Hz.

Synchrony strands



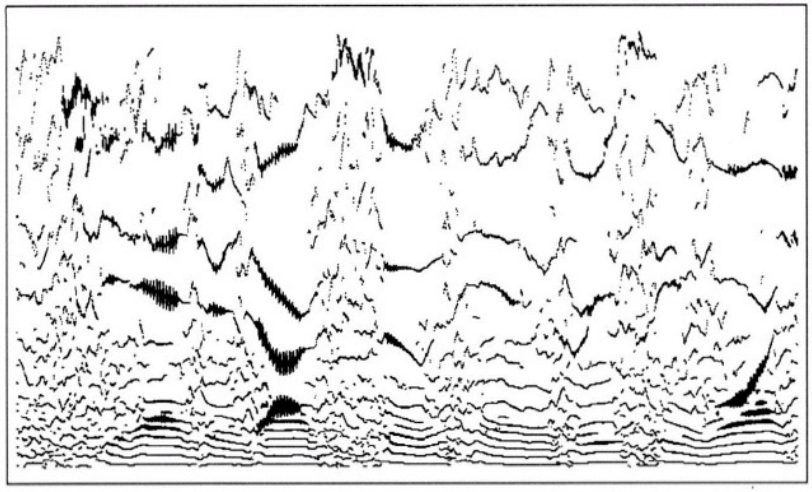


Figure 3.7 Male speech (2.5 s duration).

Implementation of auditory grouping principles

The algorithm implements the following grouping principles:

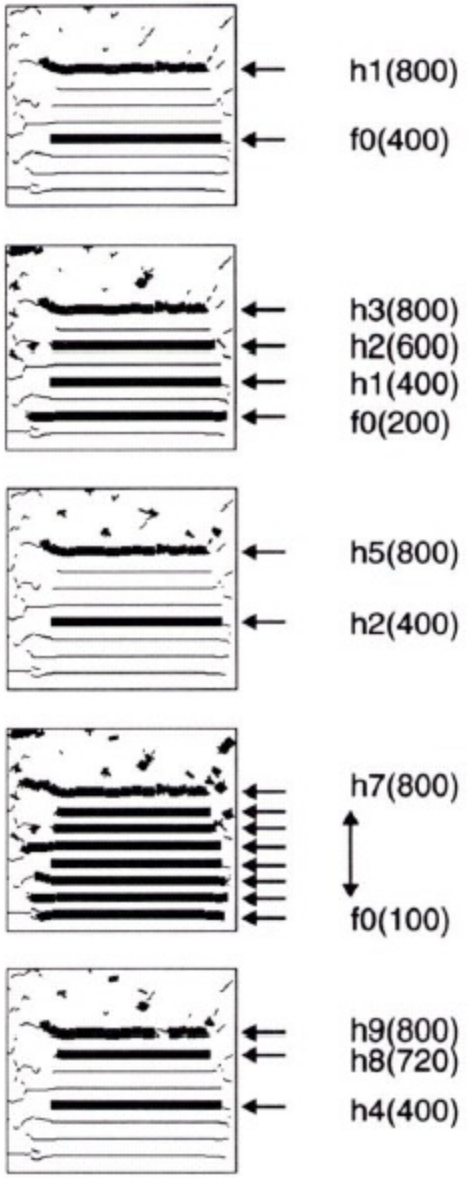
1. Harmonicity

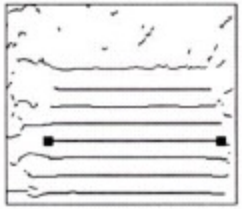
2. Common amplitude modulation

3. Common frequency movement

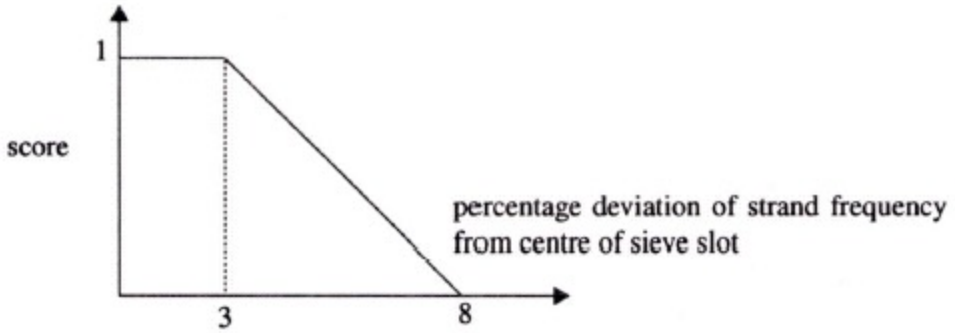
1. Harmonic grouping

Harmonic groups discovered by various synchrony strands



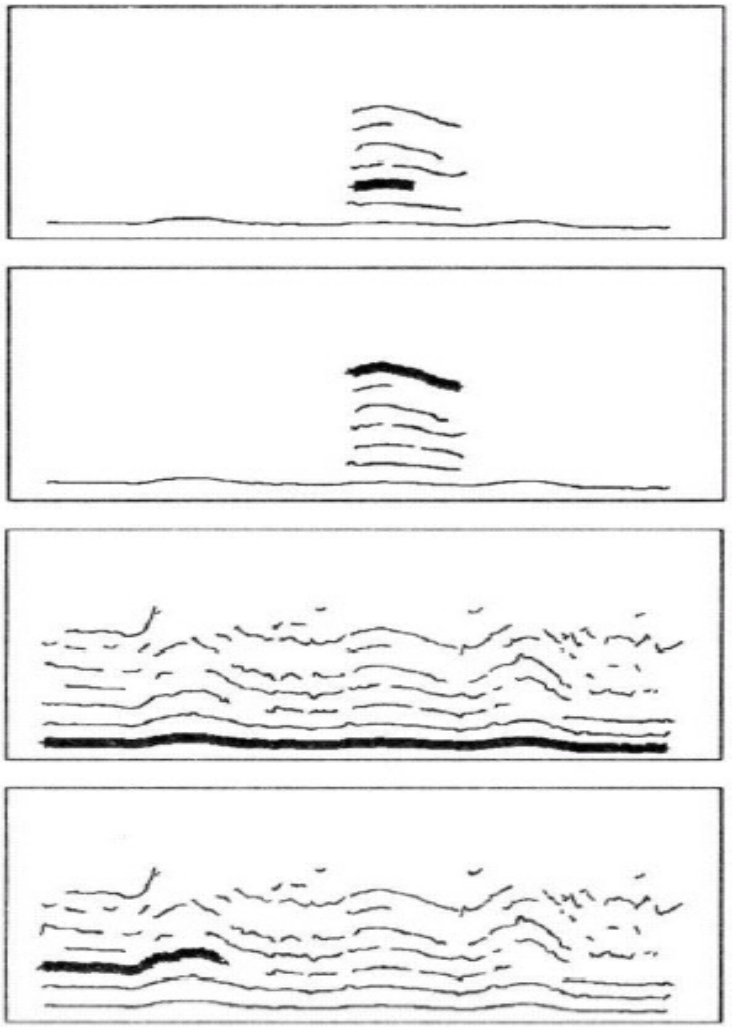


Scoring function used in assessing how well strands fit sieve slots



Harmonic constraint propagation for a voiced utterance

1. Seed (highlighted) attracts several supporters to form a harmonic group
2. A new focus is chosen, but recruits few new supporters to group.
3. New focus (f0 itself) successfully attracts virtually all the harmonically related strands in the utterance.
4. Process halts when no temporal extension to the group is possible.



2. Amplitude modulation constraint propagation

1. Seed selection (thick line)

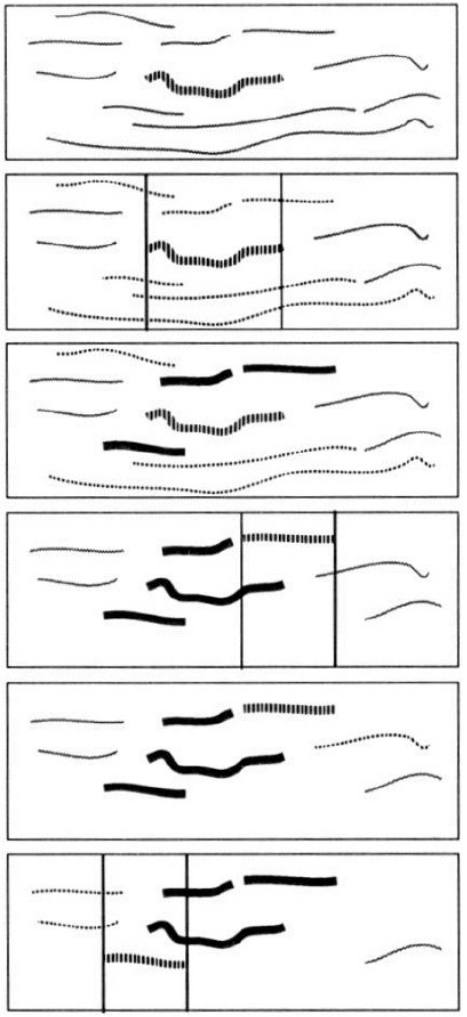
2. Choose ‘simultaneous set’ i.e. those strands which overlap in time with the Seed

3. Apply grouping constraint, e.g. suppose the black strands share a common rate of amplitude modulation with the seed

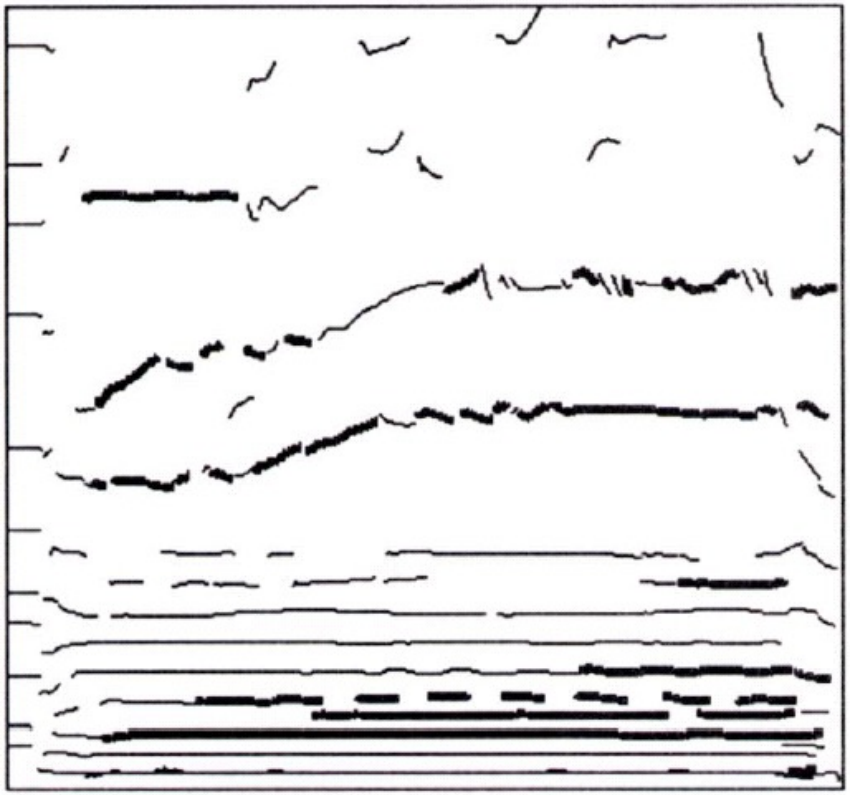
4. Sequential phase: choose strand to extend the temporal basis of the group

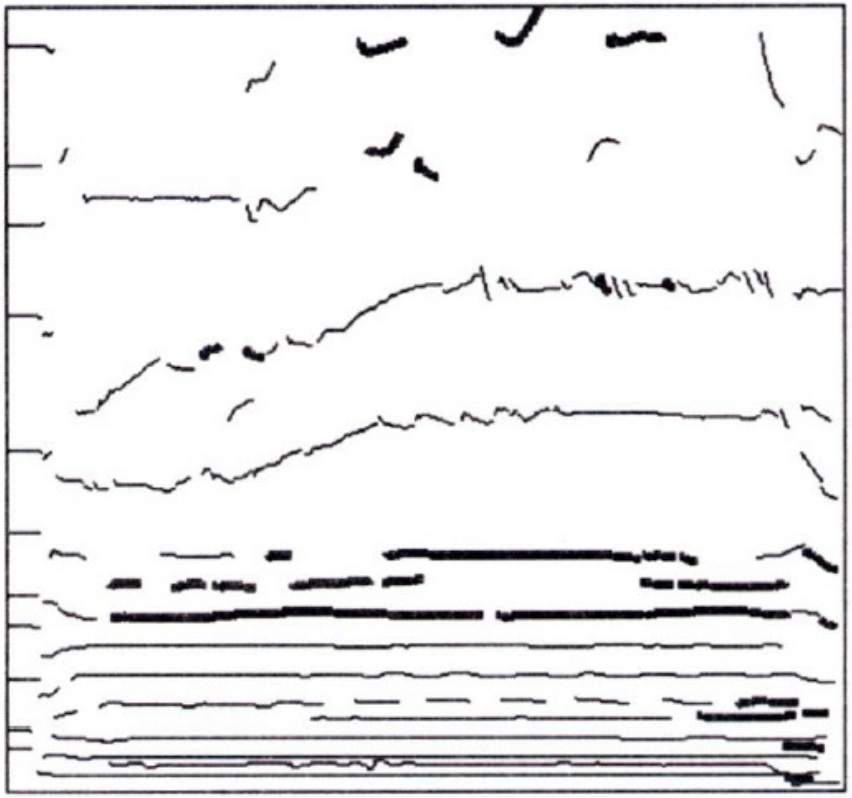
5. Back to simultaneous phase: consider for similarity strands that overlap with new seed

6. Seeds may be selected from any in the group which extended in time

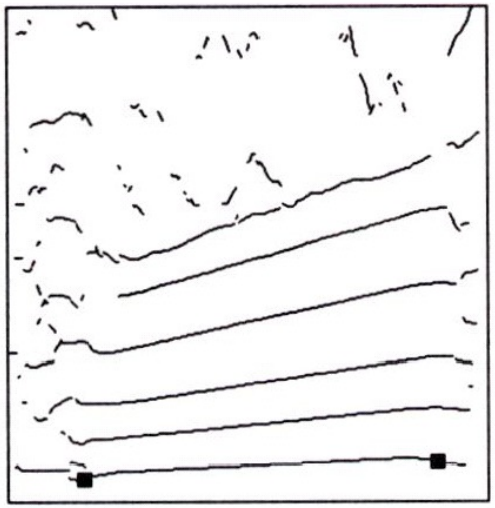


Grouping by common amplitude modulation

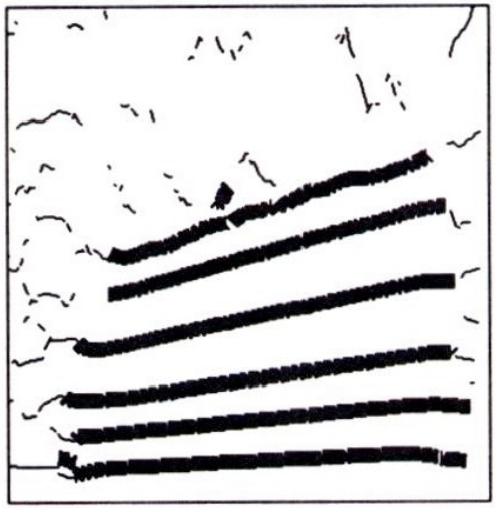




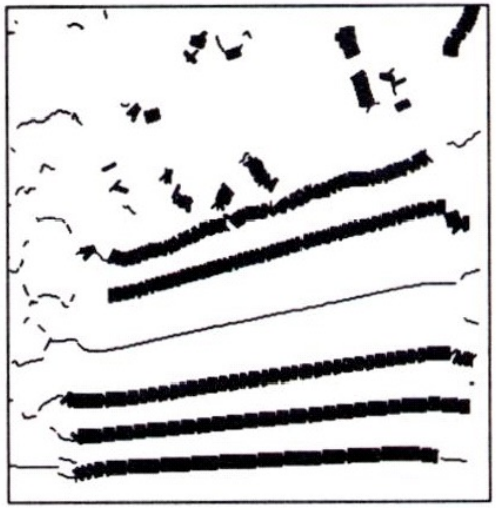
3. Grouping by common frequency movement



common FM grouping



harmonicity grouping



Harmonic grouping of frequency modulated sounds

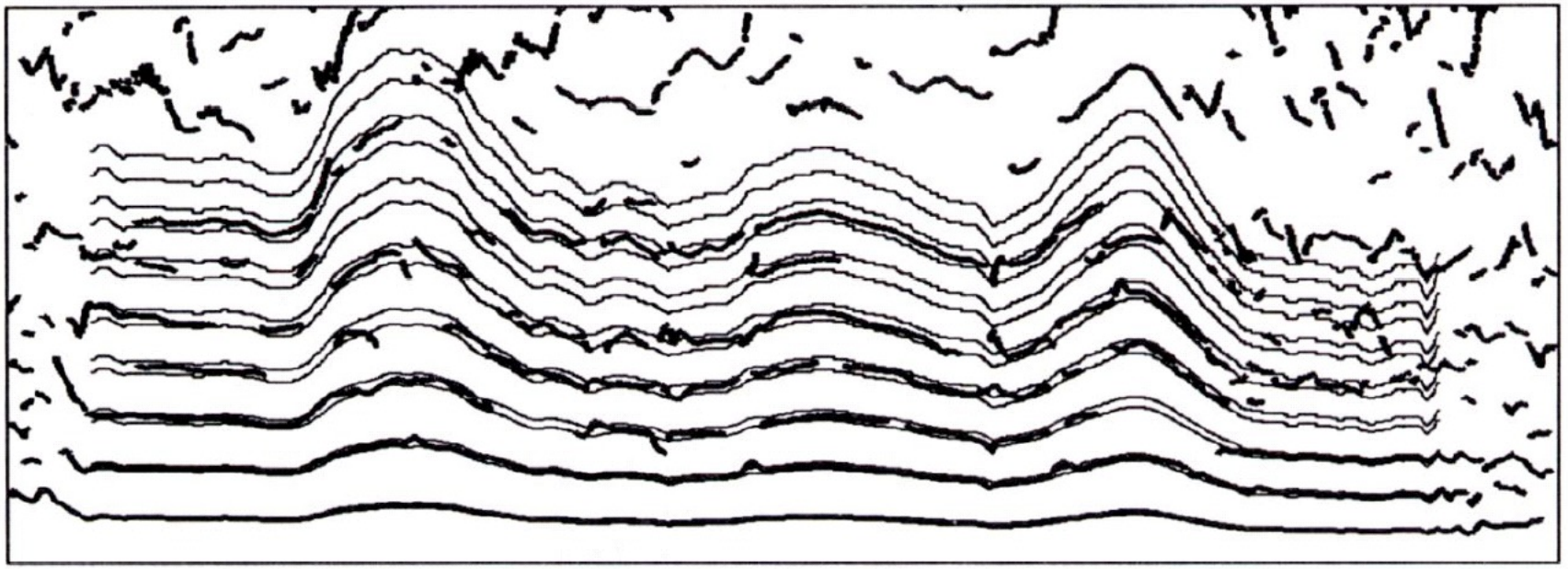
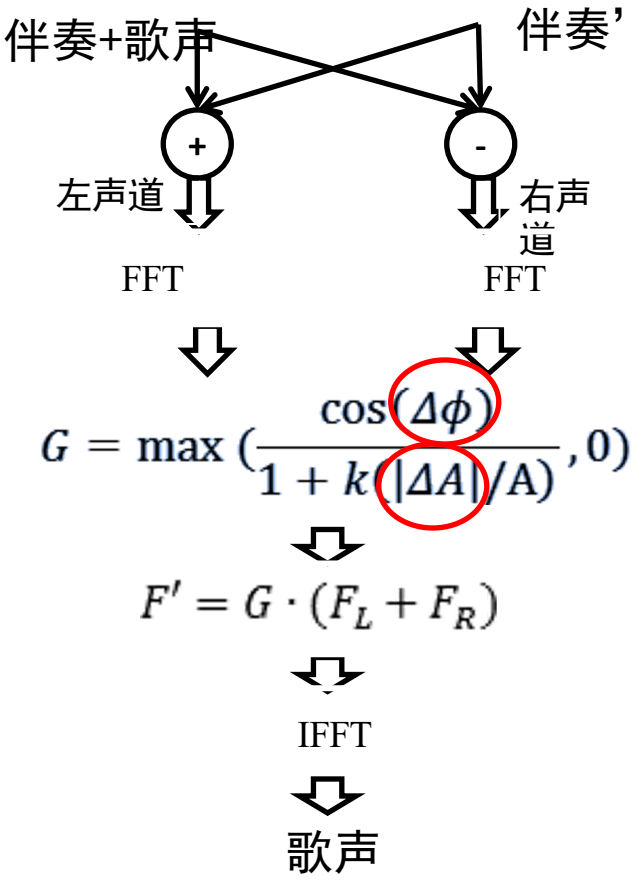
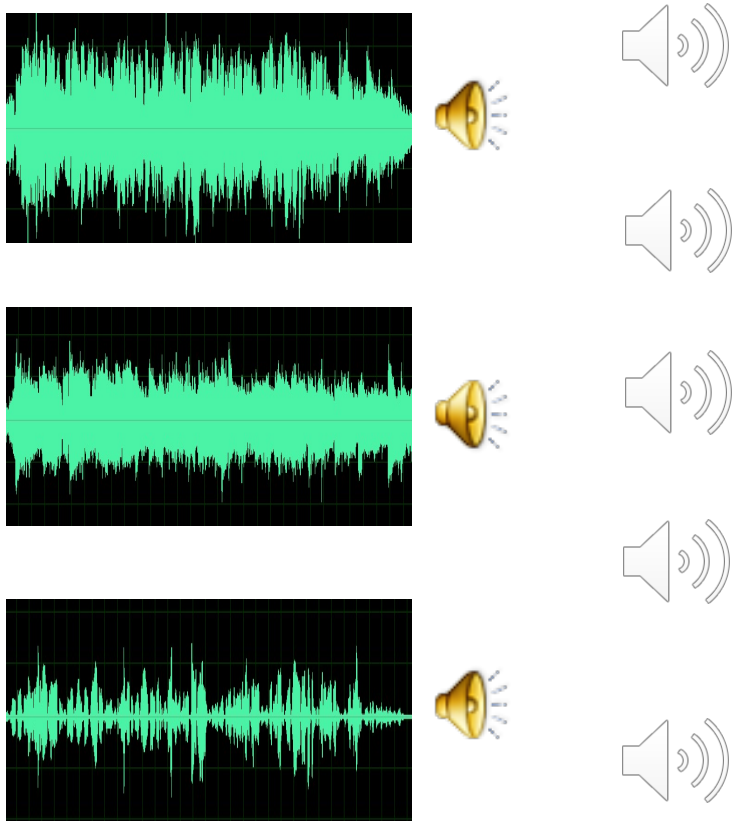


Figure 5.2 One of a series of temporally-extended harmonic sieves generated from a seed strand. Thin lines represent the +/-3% boundaries of sieve channels. Some strands fall wholly or partly into the sieve, whilst others do not. Strands may contribute to more than one sieve channel.

歌声分离和提取


伴奏+歌声
伴奏’
+
-
左声道
右声道
FFT
FFT
$G = \max\left(\frac{\cos(\Delta\phi)}{1 + k(|\Delta A|/A)}, 0\right)$
$F' = G \cdot (F_L + F_R)$
IFFT
歌声



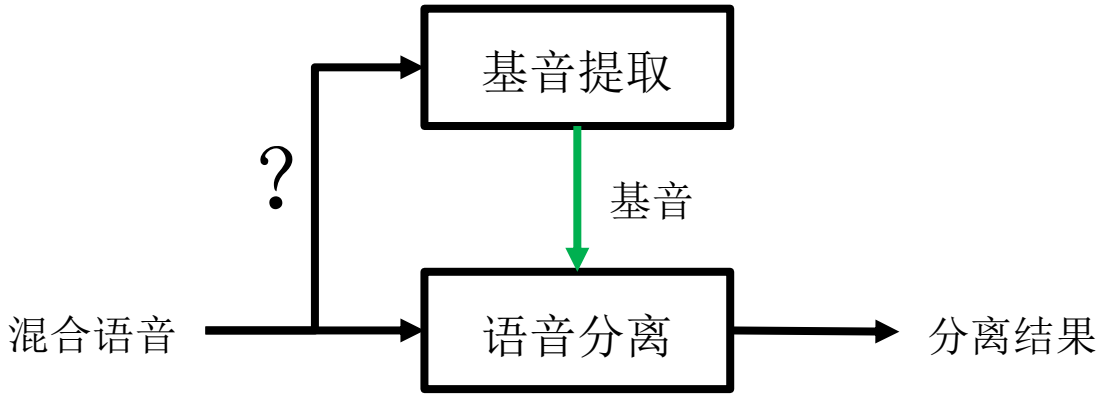
基频提取与语音分离的关系

基频提取存在的问题

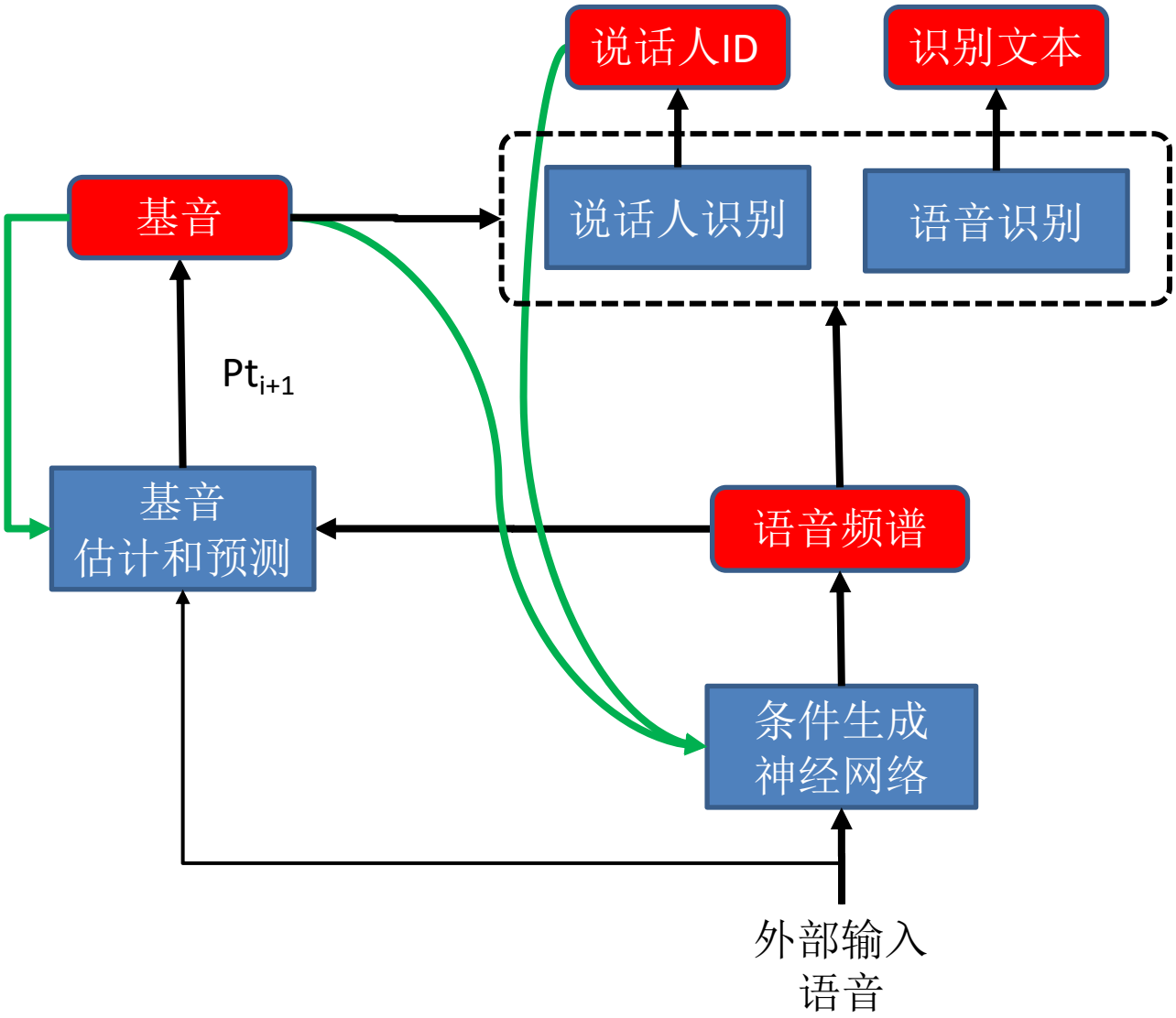
- 信号处理的方法
  - 相关法、倒谱法、…
  - 局部单帧信息
- 深度神经网络
  - 可利用上下文信息
  - 混合语音的基频提取困难

语音分离问题

- 时频掩膜方法：质量较差
- 生成神经网络方法
- 基频驱动的条件式生成方法
  - ✓ 基频轨迹连续性
  - ✓ 置换问题和说话人数目



迭代式“基频预测-语音分离”

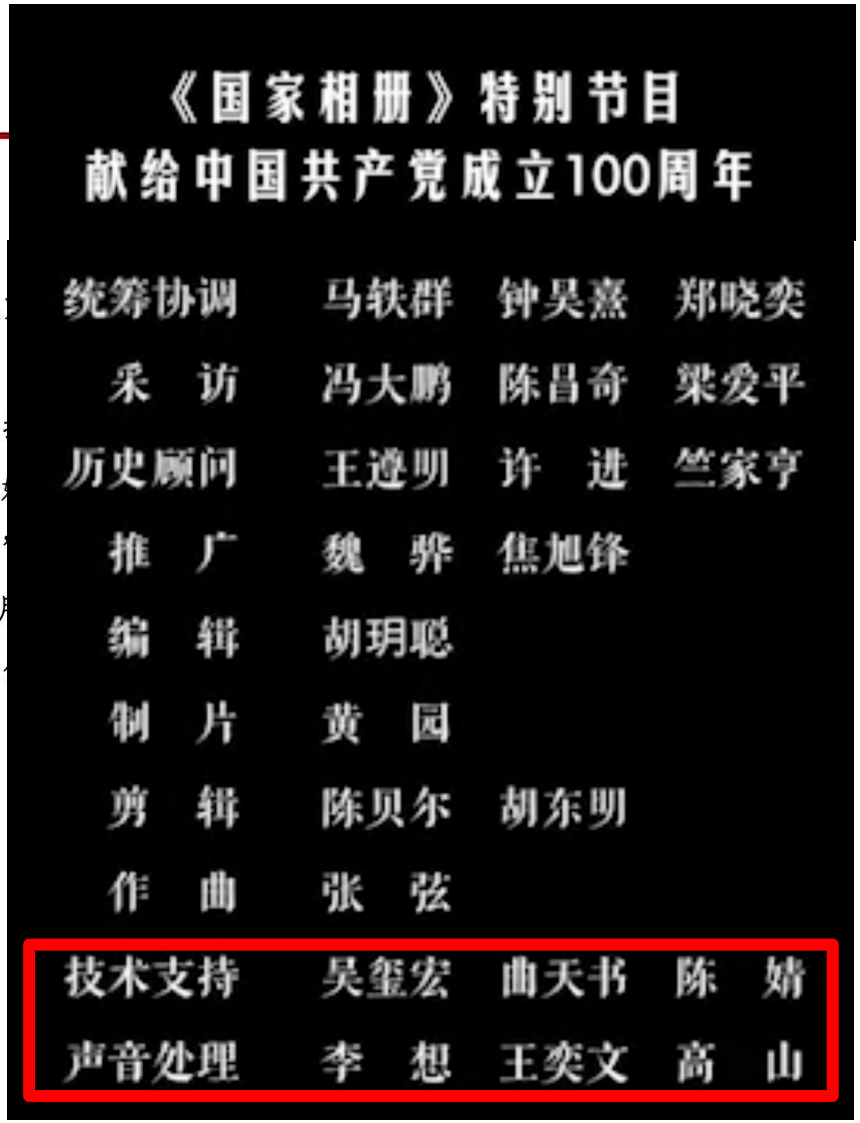


性能评价

与其他方法在语音分离性能上的比较

Method	2-speakers			3-speakers		
	SDRi(dB)	PESQ	STOI(%)	SDRi(dB)	PESQ	STOI(%)
uPIT	9.4	2.6	87.7	4.7	2.1	79.2
DPCL	9.4	2.7	88.4	7.1	2.2	82.1
TasNet	11.1	2.9	90.7	9.6	2.4	84.6
Conv-TasNet	**15.6**	3.2	93.9	13.1	2.6	88.2
Pitch-cGAN (mix phase)	10.4	2.7	89.3	8.9	2.2	84.9
Pitch-cGAN (end2end)	15.3	3.4	94.3	13.4	2.8	89.3
Pitch-cGAN (update pitch)	15.5	**3.4**	**94.6**	**13.6**	**2.9**	**90.1**

应用效果


《国家相册》特别节目
献给中国共产党成立100周年
统筹协调 马轶群 钟昊熹 郑晓奕
采 访 冯大鹏 陈昌奇 梁爱平
历史顾问 王遵明 许 进 竺家亨
推 广 魏 骅 焦旭锋
编 辑 胡玥聪
制 片 黄 园
剪 辑 陈贝尔 胡东明
作 曲 张 弦
技术支持 吴玺宏 曲天书 陈 婧
声音处理 李 想 王奕文 高 山

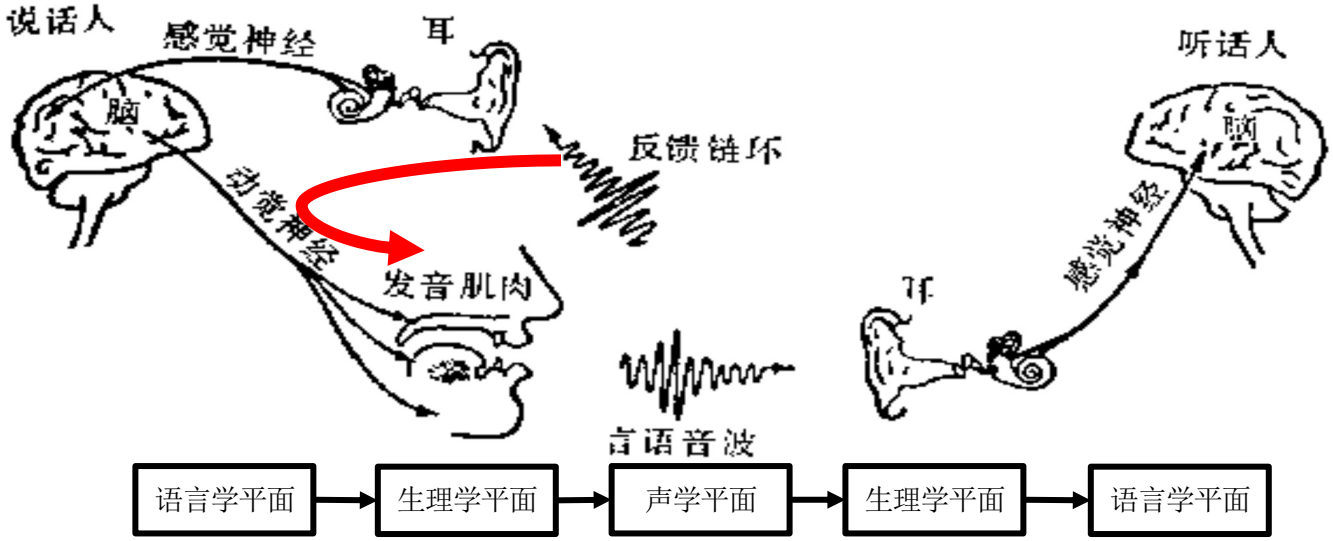
Is speech a single sound source ?

Multiple sources of sound:

- Vocal folds vibrating
- Aspiration
- Frication
- Burst explosion
- Clicks

言语链

- 言语链中人的听觉系统和发声系统构成一个闭合链，为一个子链。



（摘自《言语链：说和听的科学》P.B.邓斯、E.N.平森著，曹剑芬等译，中国社会科学出版社）

具身认知

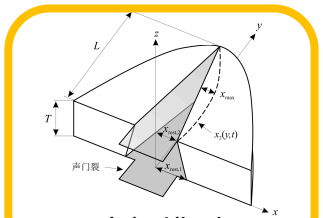
将语音的听觉表示转换成发声的肌肉控制，发声器官的姿态成为语音的具有物理意义的表示

具身学习

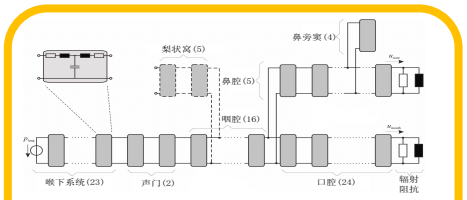
协同物理的发声过程与听觉的逆过程，实现物理系统约束下的自监督的学习

Models of Speech Production

- VocalTractLab (VTL) models (Birkholz, 2013)

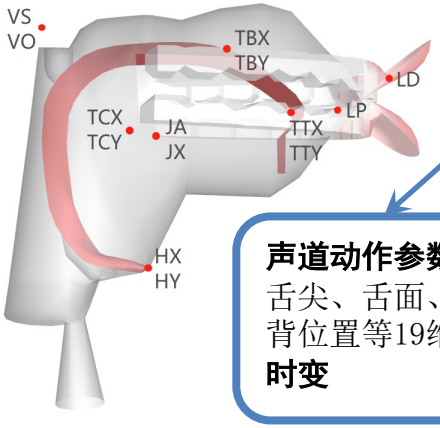


声门模型



传输线声学模型

声管几何模型

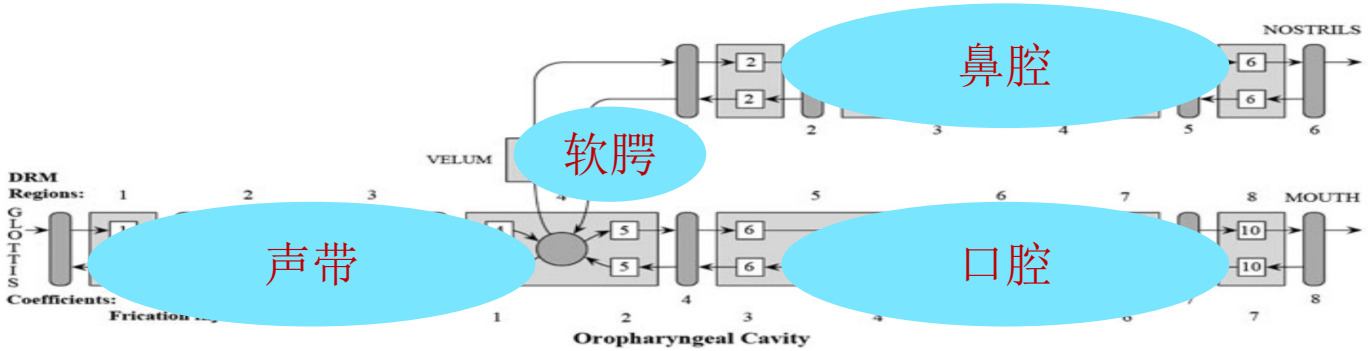


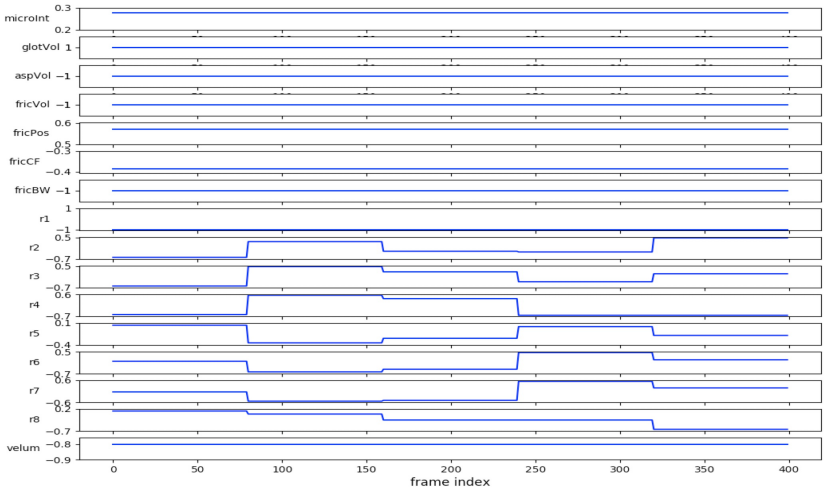
声门动作参数：
基频、气压、送气强度等11维；
时变

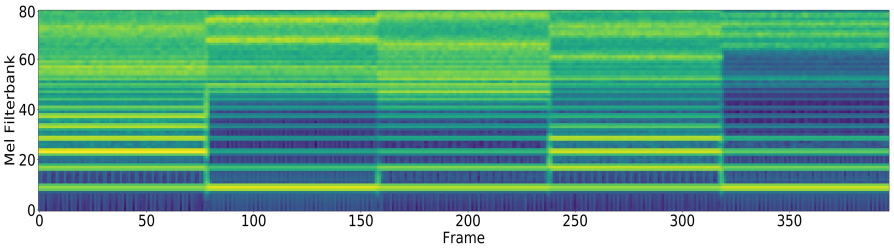
声道动作参数：
舌尖、舌面、舌背位置等19维；
时变

声道结构参数：
唇宽、咽长、上下臼齿高度等13维；**非时变**

声门-声道发声滤波器TRM[Hill, 2017]*模型

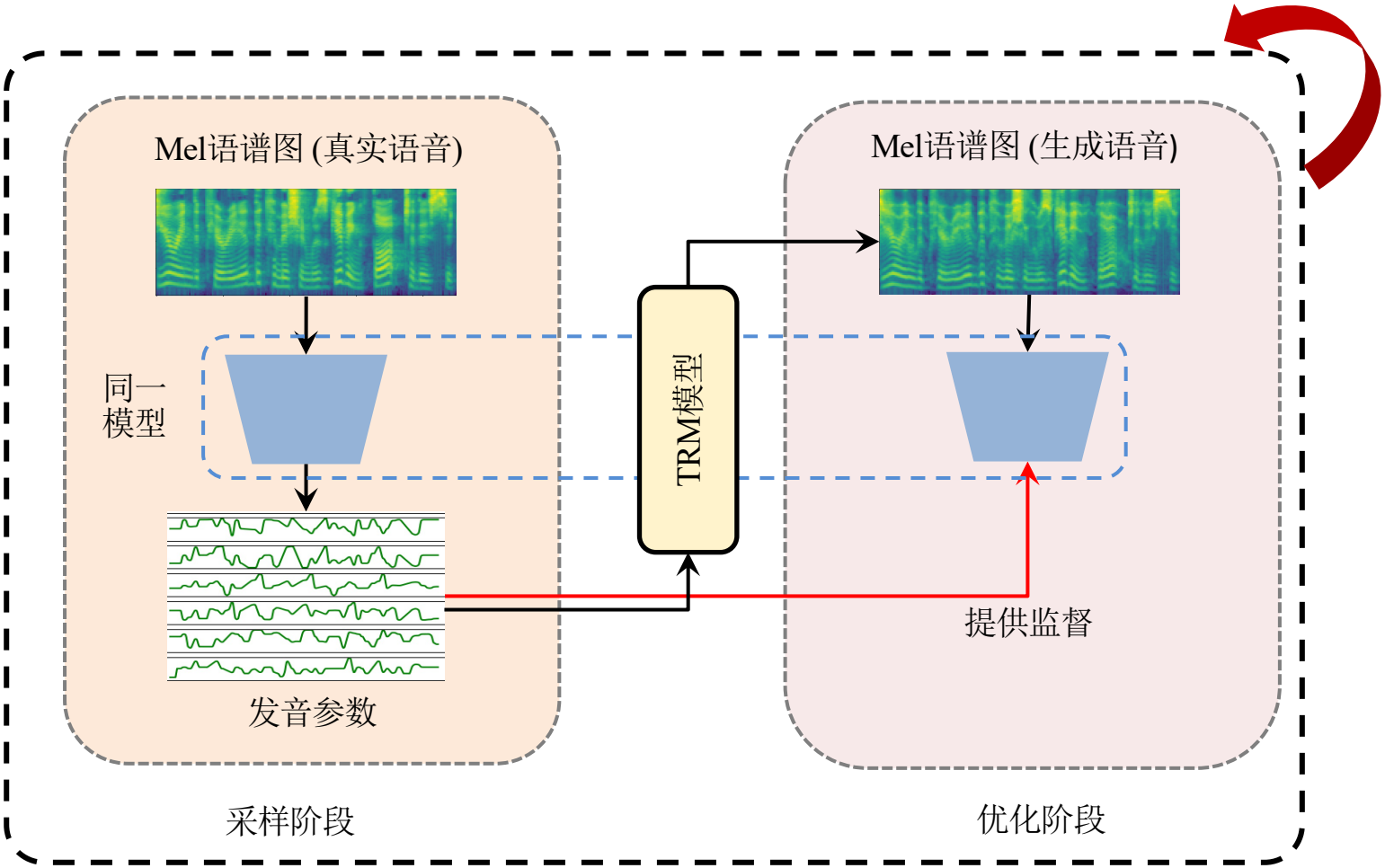






* Tube Resonance Model

具身自监督学习框架

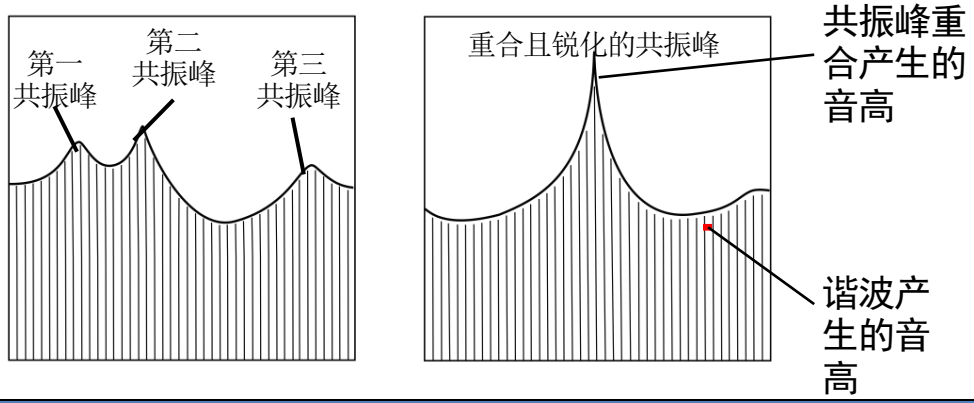


演示：呼麦及其产生机制

- 呼麦
  - 蒙古族、图瓦人的一种民族音乐艺术形式
  - 单个歌者同时产生两个音高

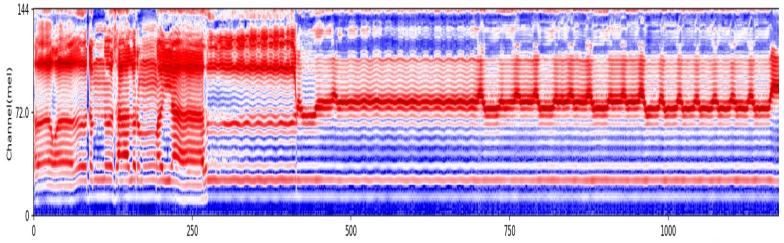
■ **产生机制：**

➢ 一种运用泛音的歌唱方式，第一、第二共振峰重合

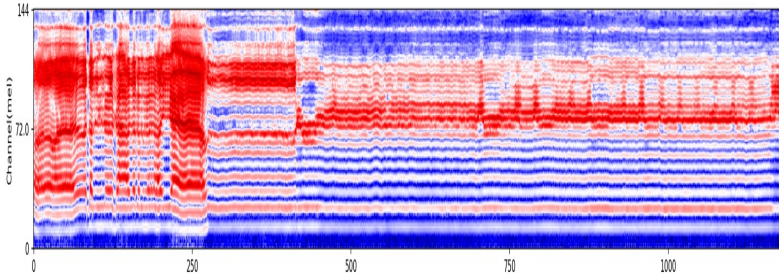


问题：发音器官采取何种姿态可使得前两个共振峰重合？

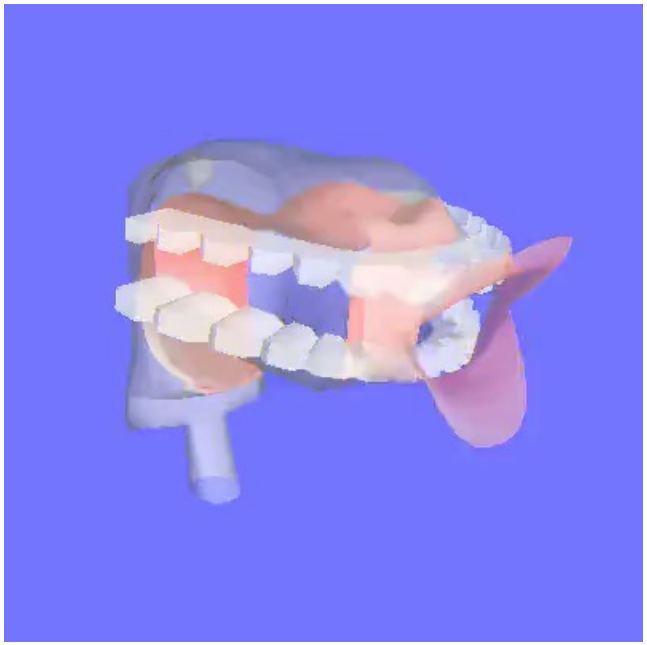
演示：呼麦反演



（a）实录呼麦的部分语谱

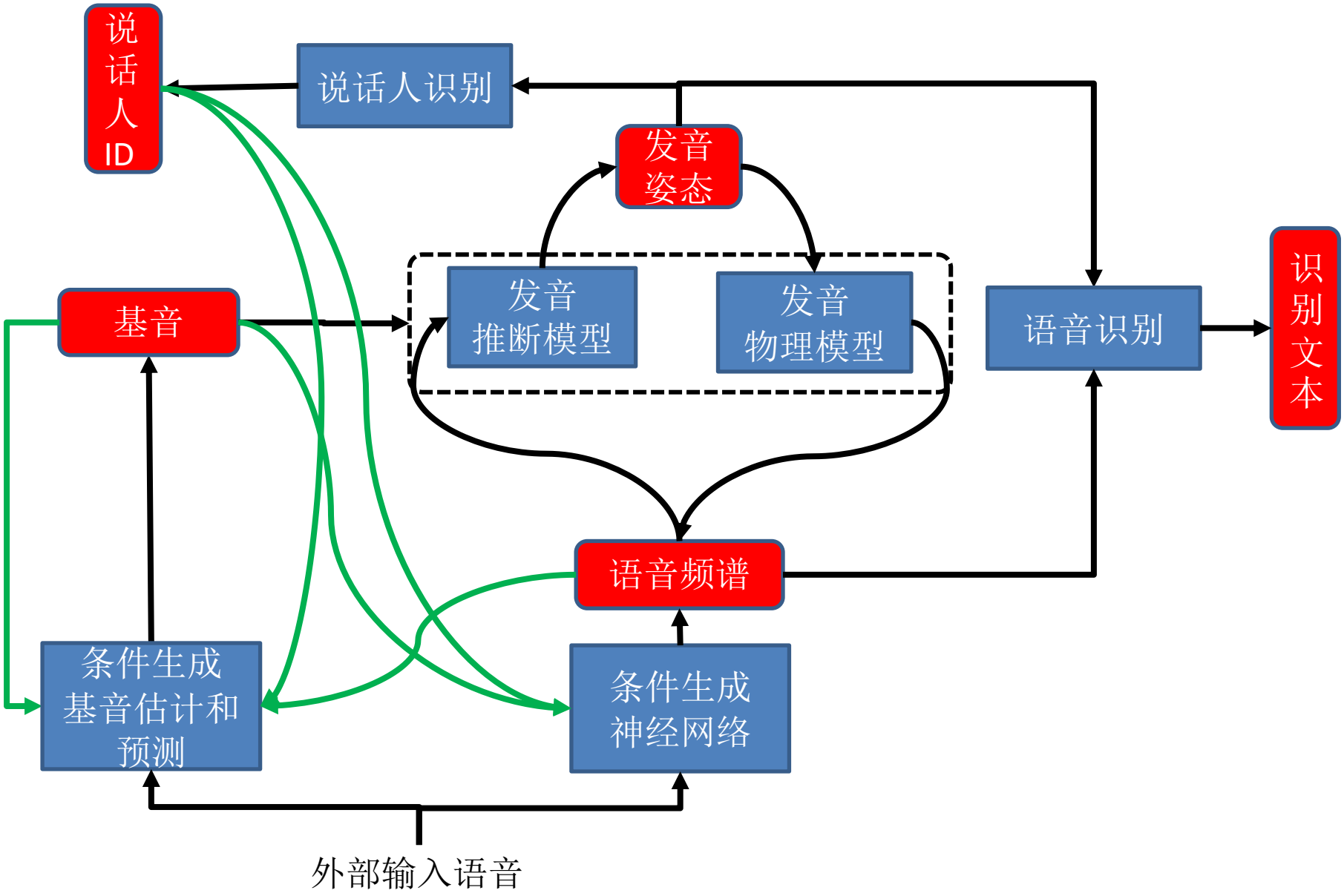


（b）重构呼麦的部分语谱



重构的呼麦与估计的发音动作

迭代式“基频、音色预测的语音分离”



Q&A?