Source Segregation

Chris Darwin
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 (e.g., acoustic/phonetic relations) probably concerns a single source
- Need to make single source properties (e.g., silence) explicit
Making properties explicit

- Single-source properties not explicit in input signal
- eg silence (Darwin & Bethel-Fox, JEP:HPP 1977)

NB experience of yodelling may alter your susceptibility to this effect
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

Bugandan xylophone music: “Ssematimba ne Kikwabanga”

Track 7

Track 8
Streaming occurs for sounds

– with same auditory excitation pattern, but different periodicities


– with Huggins pitch sounds that are only defined binaurally

Carlyon & Akeroyd
Huggins pitch

"a faint tone"

Frequency

Time

Noise

$\Delta \phi$

Interaural phase difference

0

2$\pi$

500 Hz

Frequency
Successive grouping by frequency

[Graph showing frequency in kHz (log scale) over time in seconds for "WITHIN" and "ACROSS" conditions, comparing standard and full cycle.]
Successive grouping by timbre

Wessell illusion
Successive grouping by spatial separation
Sach & Bailey - rhythm unmasking by ITD or spatial position?

ITD sufficient but, sequential segregation by spatial position rather than by ITD alone.
## Build-up of segregation

<table>
<thead>
<tr>
<th>Horse</th>
<th>Morse</th>
</tr>
</thead>
<tbody>
<tr>
<td>-LHL-LHL-LHL-</td>
<td>--H---H---H--</td>
</tr>
<tr>
<td></td>
<td>-L-L-L-L-L-L-L</td>
</tr>
</tbody>
</table>

- Segregation takes a few seconds to build up.
- Then between-stream temporal / rhythmic judgments are very difficult
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?
Continuity and streaming

Discontinuous frequency changes produce more streaming than do continuous changes

Bregman CD track 12

iTunes
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
Rhythmic masking release

Simultaneous onset prevents a component from forming part of a sequential stream

Bregman CD Track 22

iTunes
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
Percept
Rate of onset and continuity

Rapid increases in level lead to Old+New
Gradual just heard as increase

Bregman CD track 32
iTunes
Grouping & vowel quality

[Diagram with axes labeled 'frequency' and 'time']
Grouping & vowel quality (2)

continuation removed from vowel

continuation not removed from vowel
Onset-time: allocation is subtractive not exclusive

- Bregman’s Old-plus-New heuristic

- Indicates importance of coding change.
Asynchrony & vowel quality

![Graph showing the relationship between onset asynchrony and F1 boundary, with 8 subjects and no 500 Hz component.]

- Onset Asynchrony $T$ (ms)
- F1 boundary (Hz)
- No 500 Hz component
- 8 subjects
- $T$ = 90 ms
Mistuning & pitch

Mean pitch shift (Hz)

% Mistuning of 4th Harmonic

8 subjects

90 ms
Onset asynchronous & pitch

Mean pitch shift (Hz) vs. Onset Asynchrony T (ms)

- ±3% mistuning
- 8 subjects

Graph shows data for vowel and complex sounds with 90 ms onset asynchrony.
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.
Effectiveness of a parameter on grouping depends on the task. Eg

- 10-ms onset time allows a harmonic to be heard out
- 40-ms onset-time needed to remove from vowel quality
- >100-ms needed to remove it from pitch.
Minimum onset needed for:

Harmonic in vowel to be heard out: c. 10 ms

Harmonic to be removed from vowel: 40 ms

Harmonic to be removed from pitch: 200 ms
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.
Continuity & grouping

Harmonic

1. Pulsing complex

Enharmonic

1. Pulsing high tone
2. Steady low tone

Group tones; then decide on continuity.
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)
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

Masking sentence = 140 Hz ± 0,1,2,5,10 semitones
Target sentence Fo = 140 Hz

Task: write down target sentence

Replicates & extends Brokx & Nooteboom
McAdams FM in sung vowels

Bregman demo 24
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?
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.
Separating two simultaneous sound sources

- Noise bands played to different ears group by ear, but...

- Noise bands differing in ITD do not group by ear
Segregation by ear but not by ITD
(Culling & Summerfield 1995)

Task - what vowel is on your left? (“ee”)
Two models of attention

Attend to common ITD

Peripheral filtering into frequency components

Establish ITD of frequency components

Attend to common ITD across components

Attend to direction of object

Peripheral filtering into frequency components

Establish ITD of frequency components

Group components by harmonicity, onset-time etc

Establish direction of grouped object

Attend to direction of grouped object
Phase Ambiguity

500 Hz: period = 2 ms

R leads by 1.5 ms
L leads by 0.5 ms

cross-correlation peaks at +0.5 ms and -1.5 ms
auditory system weighted to one closest to zero

500-Hz pure tone leading in Right ear by 1.5 ms
Heard on Left side
Disambiguating phase-ambiguity

- Narrowband noise at 500 Hz with ITD of 1.5 ms (3/4 cycle) heard at lagging side.
- Increasing noise bandwidth changes location to the leading side.

Explained by across-frequency consistency of ITD.

(Jeffress, Trahiotis & Stern)
Resolving phase ambiguity

500 Hz: period = 2ms
L lags by 1.5 ms or L leads by 0.5 ms?

300 Hz: period = 3.3ms
L lags by 1.5 ms or L leads by 1.8 ms?

Cross-correlation peaks for noise delayed in one ear by 1.5 ms
Segregation by onset-time

ITD: \( \pm 1.5 \text{ ms} \) (3/4 cycle at 500 Hz)
Segregated tone changes location

![Graph showing segregated tone changes location.

- Onset Asynchrony (ms) on the x-axis.
- Pointer IID (dB) on the y-axis.
- Two sets of data points: Pure and Complex.
- Pure: Solid black squares for right (R) and solid black circles for left (L).
- Complex: Open black squares for right (R) and open black circles for left (L).
]
Segregation by mistuning

![Frequency vs. Duration Graph]

- **In tune**
  - Frequency (Hz)
    - 800
    - 600
    - 400
    - 200
  - Duration (ms)
    - 0
    - 400

- **Mistuned**
  - Frequency (Hz)
    - 800 (highlighted)
  - Duration (ms)
    - 0
    - 80
    - 400
Mistuned tone changes location

![Graph showing mistuning effects](image)
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.”
Hierarchy of sound sources?

Orchestra
1° Violin section
   Leader
   Chord
   Lowest note
   Attack
2° violins…

Corresponding hierarchy of constraints?
Is speech a single sound source?

Multiple sources of sound:
- Vocal folds vibrating
- Aspiration
- Frication
- Burst explosion
- Clicks

Nama: Baboon's arse
Tuvan throat music
Tuvan throat music
Mechanisms of grouping / segregation

• Primitive grouping mechanisms based on general heuristics such as harmonicity and onset-time - “bottom-up” / “pure audition”
  – Evidence: Fo-diffs on simultaneous speech

• Schema-based mechanisms based on specific knowledge (general speech constraints?) - “top-down” / “segregation by recognition”
  – Evidence: sine-wave speech
Sine-wave speech: one is OK...
(Bailey et al., Haskins SR 1977; Remez et al., Science 1981)
... but how about two?

Barker & Cooke, Speech Comm 1999

Onset-time provides only bottom-up cue
Both approaches could be true

• Bottom-up processes constrain alternatives considered by top-down processes
e.g. cafeteria model (Darwin, QJEP 1981)

Evidence:
Onset-time segregates a harmonic from a vowel, even if it produces a “worse” vowel
(Darwin, JASA 1984)
Low-level cues for separating a mixture of two sounds such as speech

Look for:

• harmonic series
• sounds starting at the same time
Plan

• How does ΔFo help in separating sound sources?
  – within vs across-formant grouping

• Effect of localisation cues on grouping & attention
  – Grouping by ear & by ITD
  – Maintaining attention to sound source (ITD, prosody, VT length)
PAT-generated sentence
“What did you say before that?”

when Fo the same -125 Hz
(either natural or monotone),

- listeners heard:
  - one voice only 16/18
  - in one place 18/18

when Fo different -125 /135
(monotone),

- listeners heard:
  - two voices 15/18
  - in two places 12/18

But as B & L admit ...
p 216 describes experiment (suggested by Arnold).

- Speech fuses

- but polyphonic music sounds weird since different notes are heard at different ears
Conclusion

Common Fo integrates

- broadband frequency regions of a single voice
- coming simultaneously to different ears
- into a single voice heard in one position.

But what about Fo’s ability to separate different voices? *(original B & L question)*
• Cutting (1976) /da/ F1 + F2 on same Fo to diff't ears, only 60% “one-item” responses

• Listening to Broadbent & Ladefoged-type sentences gives me a very clear impression of two different things on the two ears.

• Does common Fo help to integrate formants?
ΔFo improves identification

double vowels

sentences

% correct

semitones

Assmann & Summerfield 200ms
Brokx & Notteboom
Mechanisms of $\Delta$Fo improvement

- A. Across formant grouping by Fo
- B. Better definition of individual formants - especially where harmonics resolved

- B more important than A for double vowels (Culling & Darwin, JASA 1993).

- Also true for sentences?
ΔFo between two sentences
(Bird & Darwin 1998; after Brokx & Nooteboom, 1982)

Two sentences (same talker)
• only voiced consonants
• (with very few stops)

Masking sentence = 140 Hz ± 0,1,2,5,10 semitones
Target sentence Fo = 140 Hz

Task: write down target sentence

Replicates & extends Brokx & Nooteboom
Chimeric sentences
(Bird & Darwin, Grantham 1998)

Fo below 800 Hz

0

100-100

1

100-106

2

100-112

5

100-133

10 semitones

100-178

Fo above 800 Hz

100-00

800.00

60.5 dB

0.5 dB
ΔFo only in low or high freq. regions

- all the action is in the low freq region (<800 Hz)
Segregating Fo-chimeric sentences

- inappropriate pairing of Fo only detrimental at/above 4 semitones

- so across-formant grouping only important at/above 4 semitones

- 40 Subjects
- 40 Sentence Pairs

Fo difference (semitones)

% words recognised

- Normal
- Fo-swapped
- Same Fo in High Pass
- Same Fo in Low Pass
Summary of $\Delta$Fo effects in separating competing voices

- Intelligibility increased by small $\Delta$Fo in F1 region...
- … but not by $\Delta$Fo in higher freq. region.
- Across-formant consistency of Fo only important at larger $\Delta$Fo
Hi / Low complementarity

Intelligibility of competing voices increased in:

- Low frequencies: Fo differences allow better estimates of F1
- High frequencies: spatial separation head-shadow -> better S/N ratio (Bronkhorst & Plomp, 1988). But may not be audible?
Harmonicity or regular spacing?

Similar results for harmonic and for linearly frequency-shifted complexes.

Roberts and Brunstrom: Perceptual coherence of complex tones (2001)
J. Acoust. Soc. Am. 110
Auditory grouping and ICA / BSS

• Do grouping principles work because they provide some degree of statistical independence in a time-frequency space?

• If so, why do the parametric values vary with the task?
Speech
music
Speech music
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.

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.
* 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.