Separation of sound sources by humans and machines

Martin Cooke

Speech and Hearing Research
Department of Computer Science
University of Sheffield
http://www.dcs.shef.ac.uk/~martin

POP tutorial: Jan, 2006
For a longer (160 slides) version of this tutorial, see my NIPS tutorial available at

http://www.dcs.shef.ac.uk/~martin/nips.ppt
Part I:
The auditory scene analysis problem
The scope of auditory perception

Source: Compay Segundo “Ahora me da pena”
The scope of auditory perception

Now try to identify each instrument/voice as it comes in and follow it for a while

Source: Compay Segundo “Ahora me da pena”
Auditory perception answers these questions:

| What/who?       | Type of acoustic source eg talker, instrument, car engine,  
|                | *Eg for speech:* message content, talker identity, age, gender,  
|                | linguistic origin, mood, state of health, …  
| Where?         | Location: left, right, up down  
|                | Distance: promixity  
|                | Environment: bathroom, concert hall, open space?  
| How many?       | 1, 2, more  
| Transmission channel? | Telephone, radio, …  

POP tutorial: Jan, 2006
Issues facing everyday speech communication (and associated technology)

- **Distorting effect of channel**
  - none
  - telephone
  - low frequencies
  - high frequencies

- **additive noise from other sound sources**
  - car
  - pub

- **reverberation from late reflections**
  - none
  - moderate
  - strong

POP tutorial: Jan, 2006
Effects of additive noise on speech

Fourier amplitude and 
LPC smoothed spectra of 
the vowel in “head”

white noise or multi-talker 
babble mixed at 0 dB SNR

in white noise, spectral 
contrast reduced, 
harmonicity obscured

in babble, less reduction of 
spectral contrast, better 
preservation of harmonics

Source: Assmann & Summerfield (in press)
Effects of reverberation on speech

From Assmann & Summerfield (in press)

- Fills gaps associated with vocal tract closure
- Blurs onsets & offsets, reducing durational cues
- Extends noise bursts
- Flattens formant transitions in diphthongs & glides
- Removes evidence of amplitude modulation at pitch rate
- Preserves vowels

Source: synthetic data from Kalle Palomaki

POP tutorial: Jan, 2006
Effect of reverb on ASR performance

• AURORA task
• Missing data processing to handle reverberation
• Reverberation significantly reduces the binaural advantage

# Effects of noise on speakers

<table>
<thead>
<tr>
<th>Speech Adjustments</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>increase in vocal intensity (about 5 dB increase in speech for every 10 dB increase in noise level)</td>
<td>Dreher and O’Neill (1957)</td>
</tr>
<tr>
<td>decrease in speaking rate</td>
<td>Hanley and Steer (1949)</td>
</tr>
<tr>
<td>increase in average $F_0$</td>
<td>Summers et al. (1988)</td>
</tr>
<tr>
<td>increase in segment durations</td>
<td>Pisoni et al. (1985)</td>
</tr>
<tr>
<td>reduction in spectral tilt (boost in high frequency components)</td>
<td>Summers et al. (1988)</td>
</tr>
<tr>
<td>increase in F1 and F2 frequency (inconsistent across talkers)</td>
<td>Summers et al. (1988)</td>
</tr>
<tr>
<td></td>
<td>Junqua and Anglade (1990)</td>
</tr>
<tr>
<td></td>
<td>Young et al. (1993)</td>
</tr>
</tbody>
</table>

Table 5.2. Summary of changes in the acoustic properties of speech produced in background noise (Lombard speech) compared to speech produced in quiet.
Auditory scene analysis

**Key idea**

acoustic signals are littered with cues which allow our ears and brain to form separated perceptual representations (‘auditory streams’) for each individual source


Cooke (1991) PhD developed a system for *Computational Auditory Scene Analysis*
Acoustic sources and auditory streams

- Speech
- Music (Miles Davis)
- Babble
- Equal-energy mix

POP tutorial: Jan, 2006
Acoustic sources and auditory streams

A problem: sources overlap in the time and frequency domains
Part II:

Models of early processes in hearing
Preliminary: the spectro-temporal excitation pattern
Sound: from air to auditory nerve

1. Sound enters the outer ear as air pressure variations about atmospheric mean …

2. … and is converted to mechanical vibration of the oval window by the ossicles of the middle ear

3. … causing fluid vibrations in the incompressible cochlear liquids…

4. … giving rise to shearing movements between the basilar and tectorial membranes …

5. … which are detected by mechanical deflections of stereocilia of the inner hair cells …

6. … modulating the release of chemical neurotransmitter …

7. … which builds up and eventually produces an electrical impulse in an auditory nerve fibre
Typical model structure

Digitally recorded sound

Outer/middle ear transfer function

Filter

Filter

Filter

Filter

Filter

Filter

Filter

Filter

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Hair cell

Auditory nerve

POP tutorial: Jan, 2006
Cochlear filtering model

The *gammatone* function approximates physiologically-recorded impulse responses:

\[ g(t) = t^{n-1} \exp(-2\pi bt) \cos(2\pi f_0 t + \phi) \]

- \( n \) = filter order (4)
- \( b \) = bandwidth
- \( f_0 \) = centre frequency
- \( \phi \) = phase
Bank of gammatones

- Each position on the basilar membrane is simulated by a single gammatone filter with appropriate centre frequency and bandwidth.

- 32 filters are generally sufficient to cover the range 50-8 kHz.

- Note variation in bandwidth with frequency (unlike Fourier analysis).
Response to a pure tone

- Many channels respond, but those closest to tone frequency respond most strongly (*place coding*)

- The interval between successive peaks also encodes the tone frequency (*temporal coding*)

- Note propagation delay along the membrane model

POP tutorial: Jan, 2006
Example output for vowel “ah”

- Phase-locking to individual vowel harmonics at low frequencies, where the filter bandwidth is narrow enough to resolve them.
- Amplitude-modulated response at high frequencies, caused by beating (interaction of unresolved harmonics) in the wider auditory filters.
- Summed response across time gives an auditory spectrum.
Constructing the spectro-temporal excitation pattern
Part III:

Cues for separating sources
Illustration of potential cues in excitation patterns for speech

“… wash water all … ”

continuity

harmonicity

offset synchrony

common amplitude modulation

onset synchrony
Illustration of potential cues in music

- snare drum rhythm
- sax rhythm
- cadence & intervals

POP tutorial: Jan, 2006
Ugandan xylophone music

2 players alternate so their notes interleave (pentatonic scale)

- Notes occupy different pitch range
- Notes occupy same pitch range
- Notes have same pitch range but different timbre

Excitation patterns for same pitch

2nd player joins in

Source: Bregman & Ahad (1995); original demo by Wegner

POP tutorial: Jan, 2006
Cues for CASA?

Source properties
- event boundaries
- temporal modulations
- periodicity
- spatial location
- event sequence

Potential grouping cues
- across-freq synchrony of transients
- offsets
- onsets
- common across-freq envelope correlation
- common FM
- fine-structure periodicity
- harmonicity
- AM at F0
- ITD
- IID
- spectral
- good continuation
- similarity

POP tutorial: Jan, 2006
Periodicity cues in auditory nerve

- Common amplitude modulation
- Harmonicity
- Fine structure periodicities (common intervals)

nerve firing probabilities (model)
Neural autocoincidence

Licklider (1951)
Autocorrelogram

- Short-term autocorrelation of the output of each channel of the auditory periphery model
- Low frequency channels respond to individual harmonics, showing AC peaks at the period of the closest harmonic and at multiples of that period
- A summary autocorrelogram is formed by summing responses across frequency
- Peaks from harmonics of the same F0 show constructive interference

Autocorrelogram (ACG) & summary ACG of a double vowel, showing F0s
Visualising grouping cues

‘Pitch’ spectrograms
- Energy (value)
- Pitch (hue)
- Pitch strength (saturation)

A: pitch & strength from location & height of dominant ACG peak
B: ... from dominant amplitude modulation freq & depth
Neural cross-coincidence

Jeffress (1948)
Cross-correlogram

- One frame (30 ms) of a mixture of spatialised male and female speech, located at -20 and +20 degrees azimuth
- Ideally, a CC should show a spines at delays corresponding to the ITDs of each sound source
- summary CC emphasises such delays, reducing problems due to false peaks
- remaining problem:
  - Multiple peaks at high frequencies where wavelengths are shorter than ear separation; effect is to limit the number of sources localisable to 2 (humans=6)

## Contribution of other factors to the perception of everyday speech in masking noise

Technique: how much extra masking (in dBs) can listeners tolerate to reach a criterion level of intelligibility? Each extra dB provides 5-10% increase in intelligibility.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 dB or more linguistic entropy (low vs high predictable sentences)</td>
<td>Up to 5 dB</td>
</tr>
<tr>
<td>Up to 5 dB intensity differences</td>
<td>Up to 8 dB</td>
</tr>
<tr>
<td>Up to 8 dB single competing speaker vs many talkers noise</td>
<td>Up to 8 dB</td>
</tr>
<tr>
<td>Up to 8 dB binaural cues</td>
<td></td>
</tr>
<tr>
<td>1 dB location cues in reverberant environments</td>
<td></td>
</tr>
<tr>
<td>3 dB binaural cues to location</td>
<td></td>
</tr>
<tr>
<td>4 dB improved SNR at closest ear</td>
<td></td>
</tr>
<tr>
<td>Up to 15 dB visual cues eg lipreading</td>
<td></td>
</tr>
</tbody>
</table>


POP tutorial: Jan, 2006
The case of two radios

Suppose the task is to monitor two radio channels simultaneously, monaurally. At what level should you set the volumes of each channel?

*Worst strategy:* same volume

*Best strategy:* around 9 dB apart (improvement of 5 dB SRT; Brungart, 2001)
Vision influences what we hear …

McGurk effect
[video demo]

Source: McGurk & McDonald (1976) "Hearing lips and seeing voices", Nature
... and sound influences what we see

Sound-induced visual rabbit
[DEMO]

Source: http://neuro.caltech.edu/~kamitani/audiovisualRabbit/
Kamitani, Y & Shimojo, S (2001)
Part IV:

Summary of computational approaches to source separation
I: Hard-core primitive auditory scene analysis

**Organisational cues in target speech**

**Principle:** A sound mixture decomposed at the auditory periphery can be reassembled into its constituent sources by the application of grouping principles such as harmonicity, onset synchrony, continuity, etc.


**Issues**
- How to combine cues
- Grouping is not all-or-nothing
- Different thresholds for different tasks (Darwin)
- No really successful model of sequential grouping

POP tutorial: Jan, 2006
## 10 years of progress in primitive computational auditory scene analysis

<table>
<thead>
<tr>
<th>Original mix</th>
<th>Automatic separation systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech + telephone</td>
<td>Cooke (1991)</td>
</tr>
<tr>
<td>2 talkers (m/m)</td>
<td>Wang &amp; Brown (1999)</td>
</tr>
<tr>
<td>2 talkers (m/f)</td>
<td>Hu &amp; Wang (2002)</td>
</tr>
</tbody>
</table>
II: Full primitive auditory scene analysis

Organisational cues in target speech
Organisational cues in background

Principles
(i) grouping cues in the background can help unmask the target speech
(ii) unexpected energy while tracking one source can reveal the presence of another source (Bregman’s old+new principle)
(iii) the residue left after extracting one or more sources can be processed to reveal further sources

Status: perceptual evidence for the power of background periodicity in helping identify the foreground

Models
(i) Cancellation models of double vowel perception (Lea, 1992, de Cheveigné, 1993++)
(ii) Residue models (eg Nakatani et al, 1998)
III: Speech is special

Organisational cues in target speech

**Principle:** speech identification processes have privileged access to the mixture signal and take what they need for classification

“Speech is beyond the reach of Gestalt grouping principles” (Remez et al, 1994)

Models: could actually work in practice but yet to be demonstrated computationally

Issues

- Listeners have difficult identifying speech mixtures when potential cues for organisation are degraded (cocktail party sine-wave speech)

POP tutorial: Jan, 2006
IV: Hard-core model-based explanation

Organisational cues in target speech

Organisational cues in background

Models for target speech

Models for background

Principle: all energy in the mixture can be explained by an appropriate combination of prior models for all sources present at any moment.

Models
- HMM decomposition (Varga & Moore, 1990)
- Parallel Model Decomposition (Gales & Young, 1993)
- MaxVQ (Roweis, 2001)

Issues
- Need to know how many sources are present at each time
- Need models for all possible sources
- Computationally complex for N > 2, and too complex in practice for N = 2 if the background source is non-trivial

POP tutorial: Jan, 2006
V: Full Auditory Scene Analysis account

- Organisational cues in target speech
- Organisational cues in background
- Models for target speech
- Models for background

**Principle:** source separation and identification requires the action of both innate, primitive, grouping principles and learned schemas

**Champions:** Bregman; application to speech (Darwin)

**Models:** to some extent, the systems of Weintraub (1985) and Ellis (1996) applied bottom-up and top-down influences

**Issues**
- Very few CASA systems have exploited models for the speech target
- Level(s) at which primitive and schema processes could be integrated/conflicts resolved is not clear
VI: Energetic masking

- Organisational cues in target speech
- Organisational cues in background
- Models for target speech
- Models for background
- Energetic masking

**Principle:** the intelligibility of speech in a mixture is largely determined by peripheral masking

**Models:** articulation index (French & Steinberg, 1947; Kryter, 1962); Speech Intelligibility Index (ANSI S3.5, 1997); Speech Transmission Index (Steeneken & Houtgast, 1980; 1999); Speech Recognition Sensitivity (Musch & Buus, 2001); Spectro-Temporal Modulation Index (Elhilali, Chi & Shamma, 2003)

**Issues**
- Detection of the unmasked portions
- AI, STI etc are macroscopic models of intelligibility
VII: Linguistic masking of speech by speech

**Principle:** the intelligibility of speech in a mixture is determined not only by audibility but by the degree to which the background and foreground can be confused

‘Perceptual masking’ (Carhart et al, 1969)

**Recent studies:** Brungart et al (2001+); Freyman et al (2001+)

**Models:** None, but a prototype model of energetic and informational masking was presented by Barker & Cooke at the Hanse meeting based on competition within a speech decoder

**Issues:**
- Informational masking is too much of a catch-all term; factors other than foreground/background confusions may have a role over and above energetic masking eg distractors
VIII: Stationarity

Principle: stationary backgrounds are easily compensated

Models: lots – spectral subtraction (Boll), minimum statistics (Martin, 1993), histogram partitioning (Hirsch & Ehrlicher, 1995)

Issues
- While this is a bad approximation to everyday backgrounds, many models/algorithms embody this constraint implicitly or otherwise
- Must be used in conjunction with other processes
- Not clear to what extent listeners exploit stationarity (perhaps implicitly via enhancement of dynamics)
IX: Independence

**Principle:** exploit statistical independence of sources (Comon, 1994)

**Models:** Bell & Sejnowski (1995); Lee et al (1997); Smaragdis (2003)

**Issues**

- Reverberant energy correlated with direct energy
- Listeners manage with 1 or 2 sensors regardless of the number of sources
- Debate over whether “the cocktail party problem is beyond scope of ICA”

“One of the original motivations for ICA research was the cocktail-party problem [...] blind separation of audio signals is, however, much more difficult than one might expect [...] due to these complications, it may be that prior information, independence and nongaussianity of the source signals are not enough” (Hyvarininen et al, 2001, *Independent Component Analysis*)
X: Sparsity and redundancy

Principles
(i) spectro-temporal modulations of speech (and possibly the background too) allow relatively clear but sparse views of the target;
(ii) redundancy of speech makes identification possible in spite of missing information.


Issues
• detection and integration of sparse information in speech
The glimpsing hypothesis: listeners separate sources by exploiting brief regions where the target source is dominant.

Recall the overlap problem

Time domain

Frequency domain

Mean energy spectra

-80 -70 -60 -50 -40 -30 -20

80 200 500 1000 2000 5000 20000

frequency

log energy

speech
Miles
babble

POP tutorial: Jan, 2006
A single frequency slice
Sparse information in mixtures

Energy within 3 dB of value in mix

speech

music

babble

remainder

Energy within 3 dB of other source

speech/music

speech/babble

music/babble

speech/music/babble
An explanation of the ‘dominance effect’

1. As pointed out many times (e.g. Varga & Moore, 1991), the energy in dBs of a mixture is nearly equal to the dB energy of the most intense source in the mixture.

   \[ \log(x+y) \approx \max(\log(x), \log(y)) \]

   This approximation is at its worst when the constituents are equally intense.

2. Two or more modulated sources rarely inject similar energies in the same frequency region at the same time.

POP tutorial: Jan, 2006
Listening to sparse information

- Talker 1
- Talker 2
- Mix@0dB
- One or other talker dominant

With added noise

POP tutorial: Jan, 2006
Sparse-sampling of music

Green = speech shaped regions
### Summary of possible ingredients for computational source separation

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisational cues in target source</td>
<td>Auditory scene analysis</td>
</tr>
<tr>
<td>Organisational cues in background</td>
<td>Model-based separation</td>
</tr>
<tr>
<td>Prior models for target</td>
<td>Signal processing/robust ASR</td>
</tr>
<tr>
<td>Models for background</td>
<td>Statistics, information theory, machine learning</td>
</tr>
<tr>
<td>Energetic masking</td>
<td></td>
</tr>
<tr>
<td>Informational masking</td>
<td></td>
</tr>
<tr>
<td>Stationarity of background</td>
<td></td>
</tr>
<tr>
<td>Source independence</td>
<td></td>
</tr>
<tr>
<td>Sparsity and redundancy</td>
<td></td>
</tr>
</tbody>
</table>

POP tutorial: Jan, 2006