Binaural processing

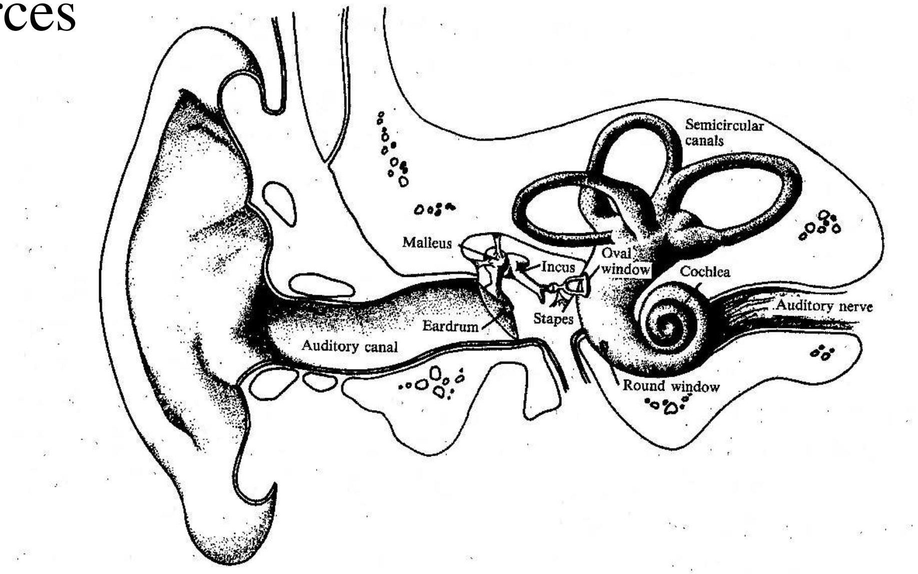
Sue Harding

Speech and Hearing Research Group,
University of Sheffield

Introduction

1. Binaural perception
   - Reasons for localising sound sources
   - Factors affecting localisation
   - Listener cues: interaural time and level differences, pinnae (outer ears)
   - Limitations of the auditory system
   - Effect of source and environment on localisation accuracy
   - Dealing with multiple sources

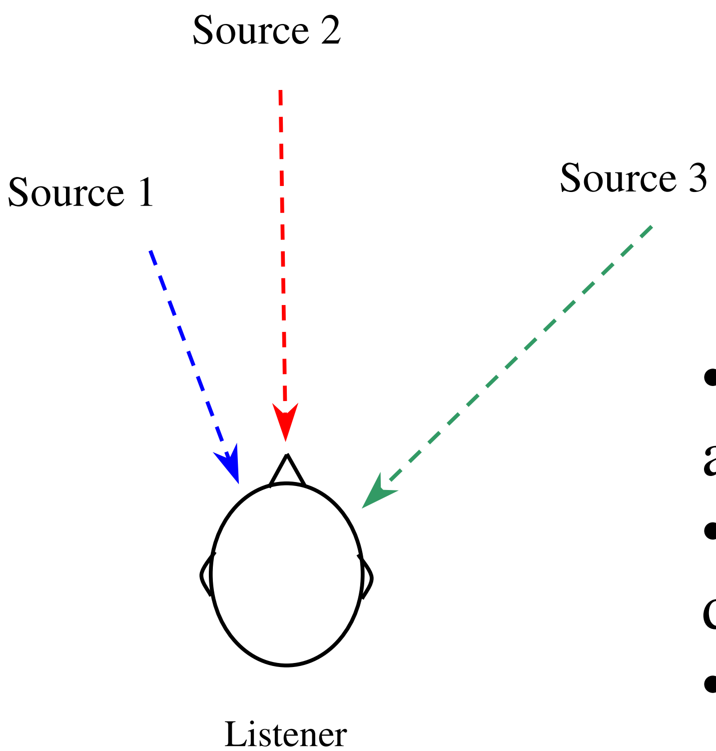
2. Models of binaural processing
   - Computational models



More information in Moore (1997); Brown & Wang (in press), Mackensen (2004)

Reasons for localising sound sources

- Nature of sound sources - how many, what are they
- Position of sound sources - movement towards us
- Information about the environment (obstacles)
- Improved communication (e.g. identifying a stream of speech)



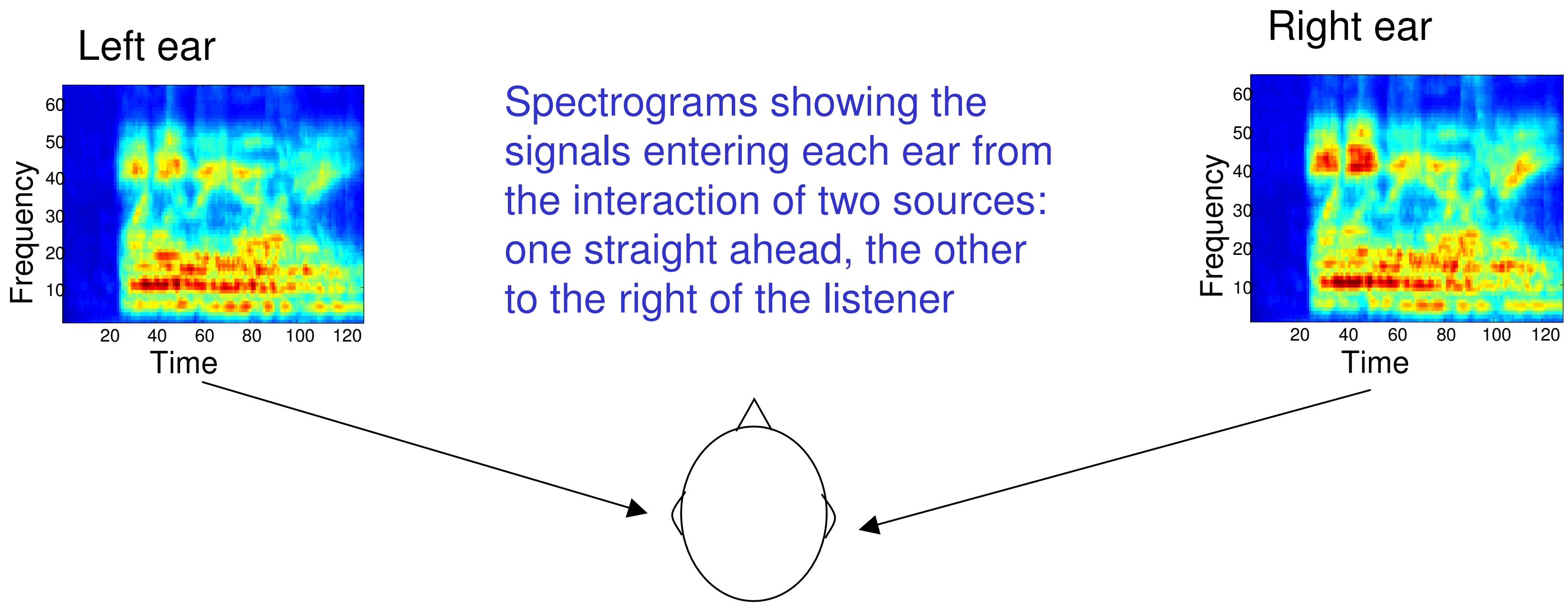
- Sounds may come from any direction and distance
- Multiple sound sources must be distinguished
- Ears are fixed on head (head may rotate or tilt)

Factors affecting localisation (1)

(based on Mackensen 2004)

1) Characteristics of sound sources

- Position of source(s) relative to listener (points on a sphere)
  - need to distinguish left/right, up/down, front/back
  - distance from listener
- Number of sources
- Spectral characteristics (frequency, bandwidth) of source
- Changes over time (spectral changes, duration, moving sources)



Spectrograms showing the signals entering each ear from the interaction of two sources: one straight ahead, the other to the right of the listener

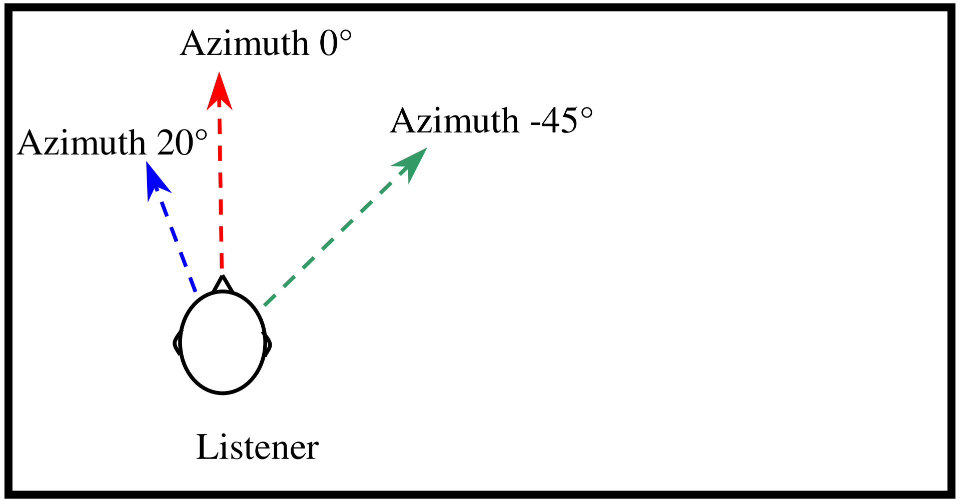
Factors affecting localisation (2)

2) Listening environment

- Surfaces of room or buildings (if any)
- Interactions between source(s) and environment

3) Characteristics of listener

- Monaural / binaural listening
- Pinnae (outer ears), head, torso
- Head movements
- Position of listener relative to surfaces and obstacles
- Non-acoustic information



Note: angle of source in the horizontal plane is defined by *azimuth,* and in the vertical plane by *elevation*

Cues available to the listener

Two major cues: use difference between input to each ear:

- interaural time differences (ITDs)
- interaural level differences (ILDs)

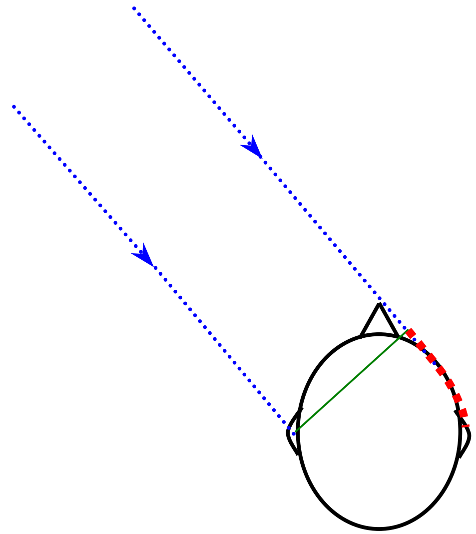
- particularly important in left/right distinction

ITD, ILD:

Left ear first, louder → source is on left

Right ear first, louder → source is on right

Sound source



Sound reaching ear further from source must travel around head;
it is delayed in time and is less intense than sound reaching ear nearer to source

Maximum ITD is approximately 690 μs for source close to one ear (minimum ITD is 0 μs)

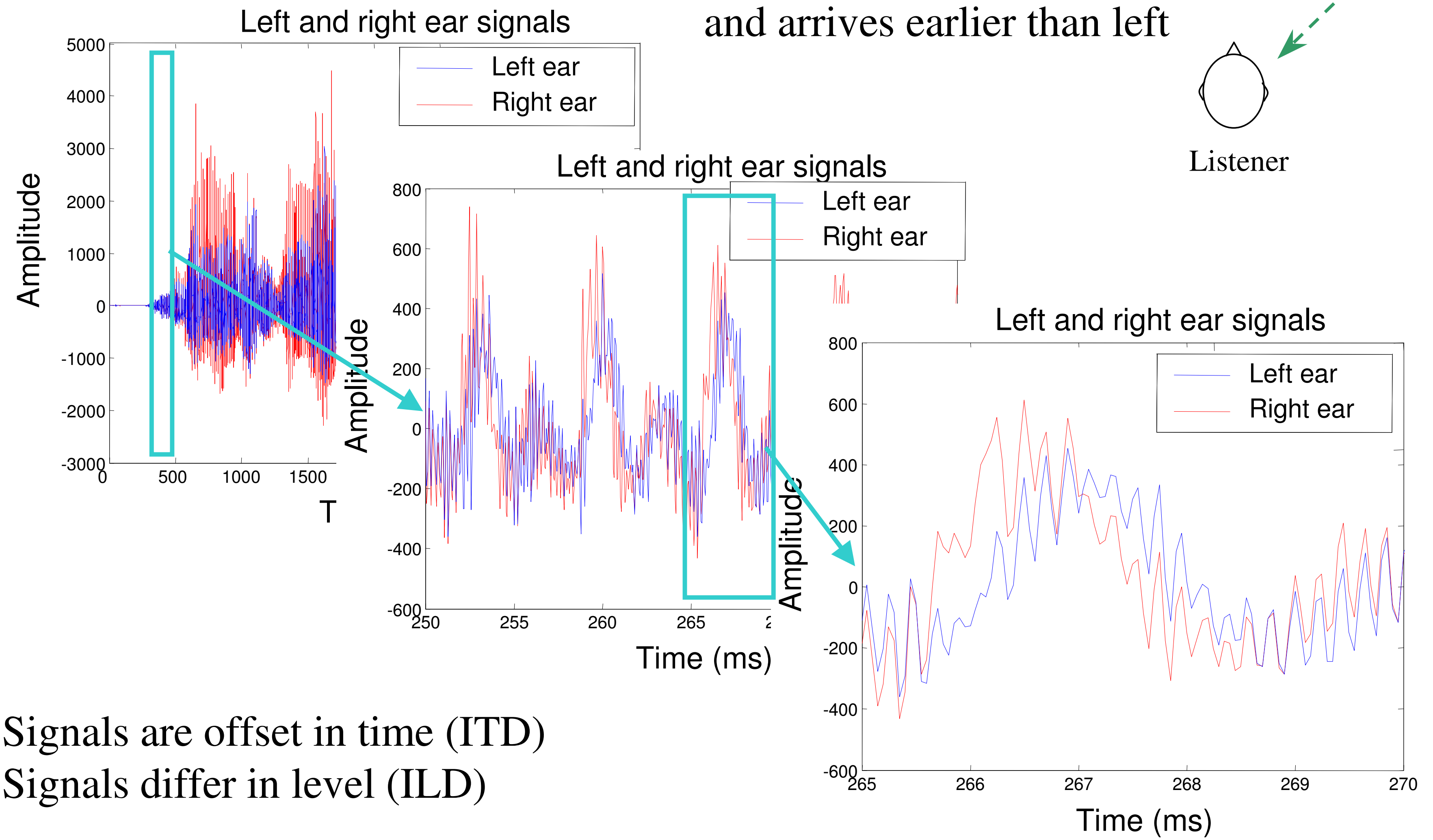
ILD is less effective at low frequencies (< 1500 Hz), as sound has long wavelength compared with head and can bend around head (no head shadow)

ITD is less effective at high frequencies (small wavelength; multiple cycles)

Interaural time and level differences

Comparison of signals entering each ear: source to the right of listener

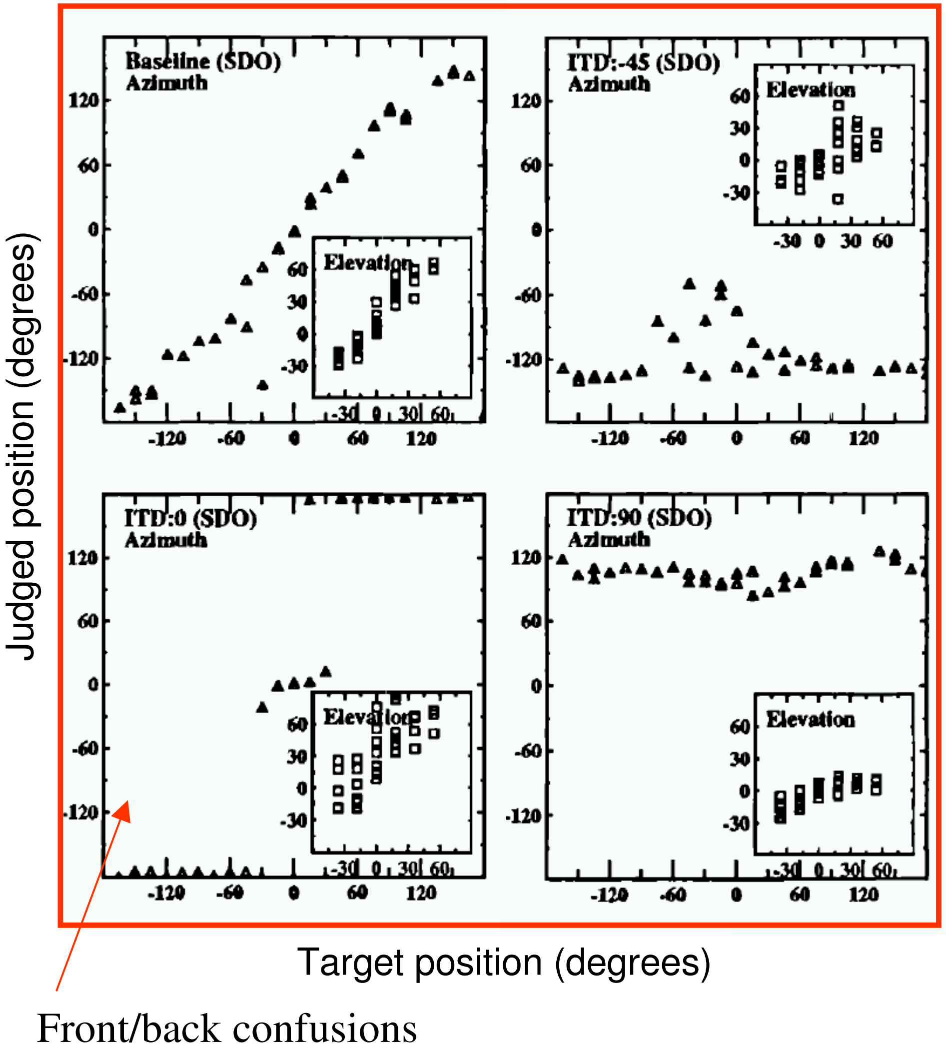
Right ear signal is louder and arrives earlier than left



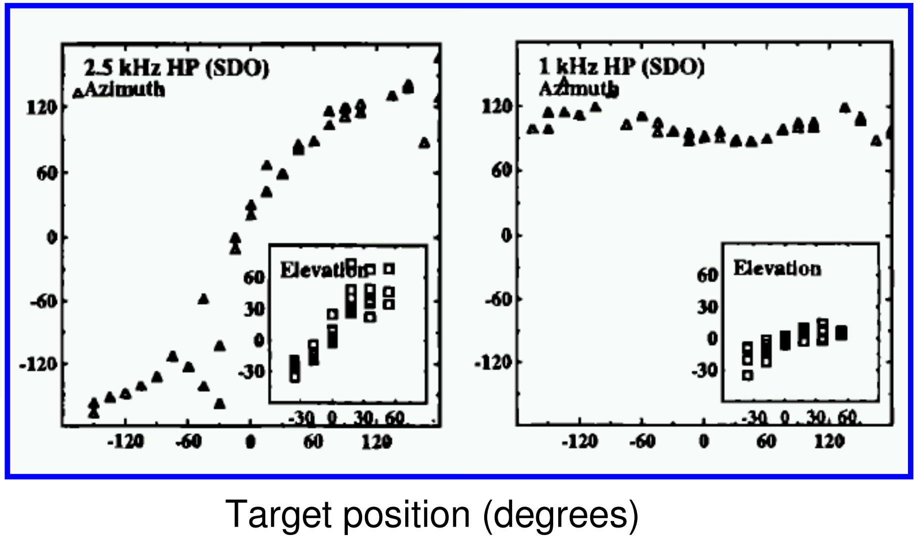
Signals are offset in time (ITD)
Signals differ in level (ILD)

ITD and ILD interaction

ITD dominates at low frequencies (Wightman & Kistler 1992)

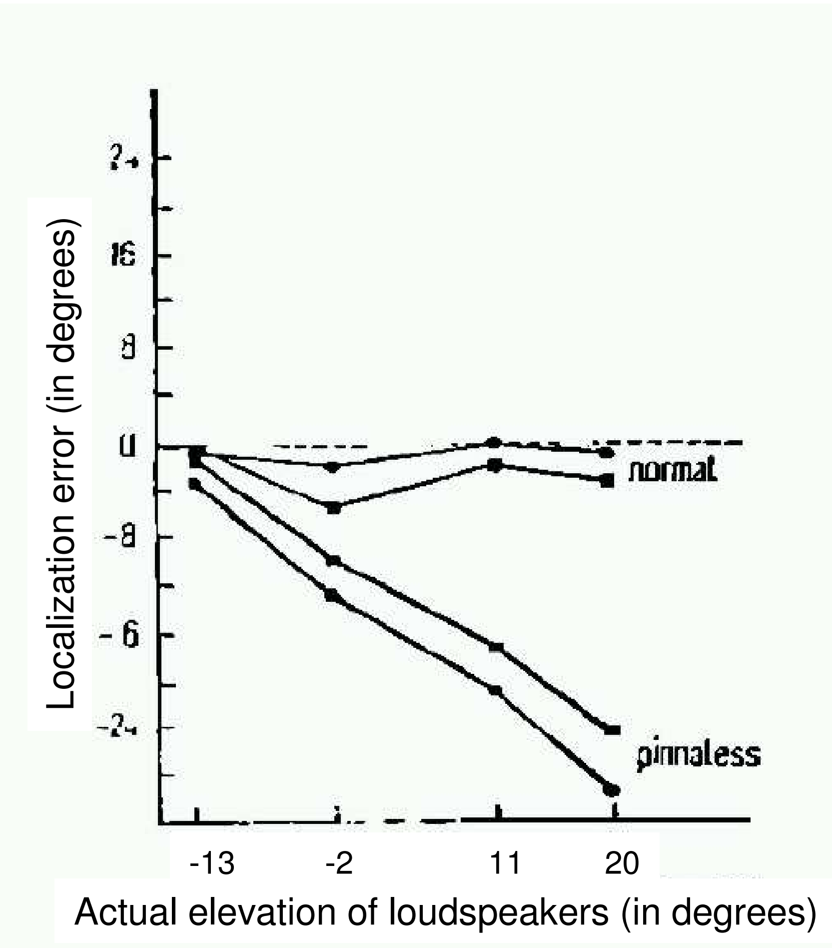


- Listeners presented with broadband noise at 36 spatial positions
- Phase of stimuli was manipulated to provide conflicting ITD cues (compared with ILD cues)
- ITD cues dominated
- High-pass filtering of stimuli (in ITD:90 condition) reduced effect of ITD



Pinnae (outer ears) - elevation

For source directly ahead of listener, same input arrives at each ear, whatever the elevation - additional information from pinna (outer ear) cues is used, e.g. Roffler & Butler (1967)

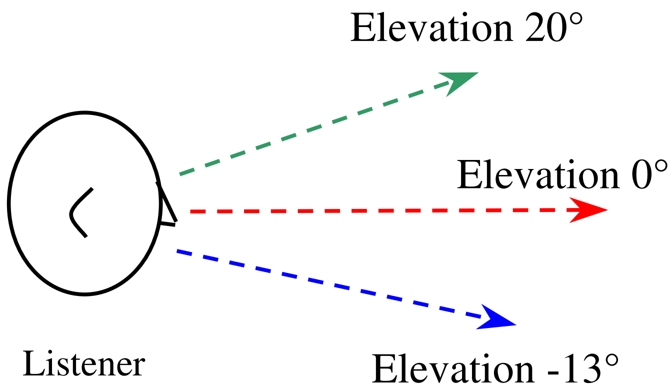


- Listeners pinnae were flattened and covered; no (minimal) head movements

- Two types of noise were presented (broadband or high frequency)

- Elevation of source was varied

- With pinnae flattened, most stimuli were judged to emanate from -13° elevation



Spectral modification due to pinnae

Pinnae, head and torso modify sound spectra depending on angle of incidence

Ratio of spectra of sound source and sound reaching eardrum gives **head-related transfer function (HRTF)** - shows frequency dependent peaks and troughs

Spectral modification cues are particularly important for distinguishing front/back and up/down, especially for sources directly ahead of the listener

Head-related transfer function (HRTF)

Head-related transfer function (HRTF):

- differs for each person
- varies with frequency and direction of source (e.g. Shaw 1974)

Listeners probably don't make use of HRTFs directly, but they can be used to simulate 3-D environments

Listeners can make use of the HRTF of other listeners for localisation, although only horizontal judgements are robust (front/back confusions are common) - Wenzel et al. (1993)

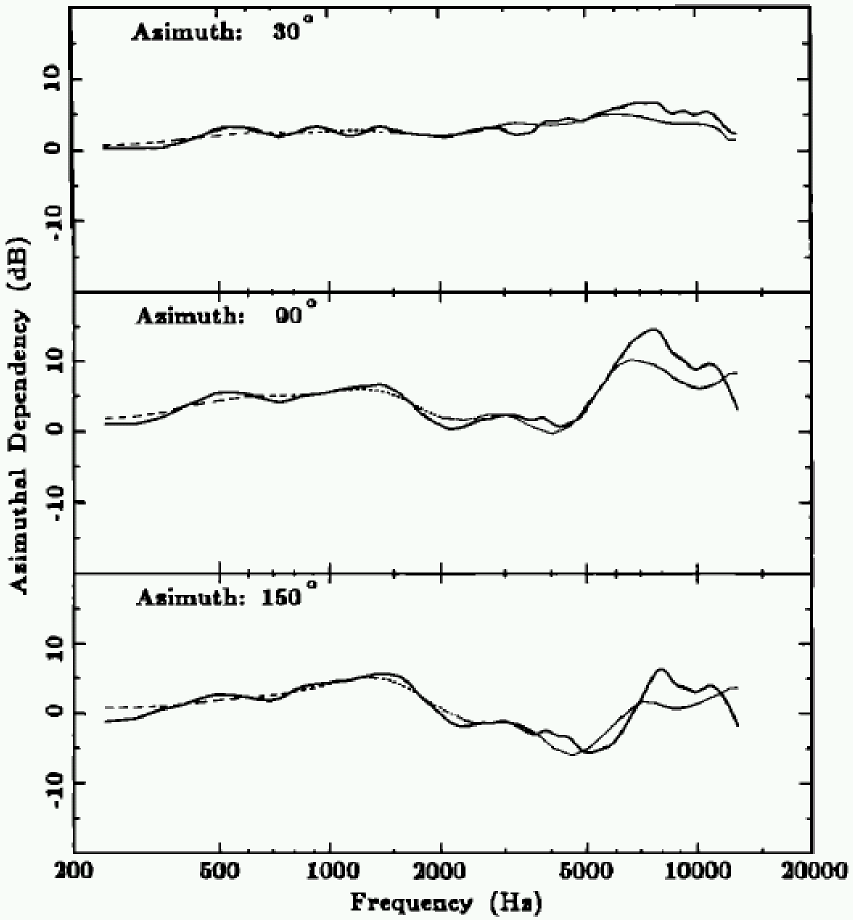


FIG. 7. The solid line in each panel represents an azimuthal dependency function (change in HRTF from 0 azimuth) averaged over ten subjects for a single source azimuth (0-deg elevation). The dotted line represents comparable data from Shaw (1974). Each panel shows the data from a different source azimuth.

Effect of source spectrum

Localisation accuracy affected by bandwidth and frequency of sources
e.g. Roffler & Butler (1967) measured elevation accuracy

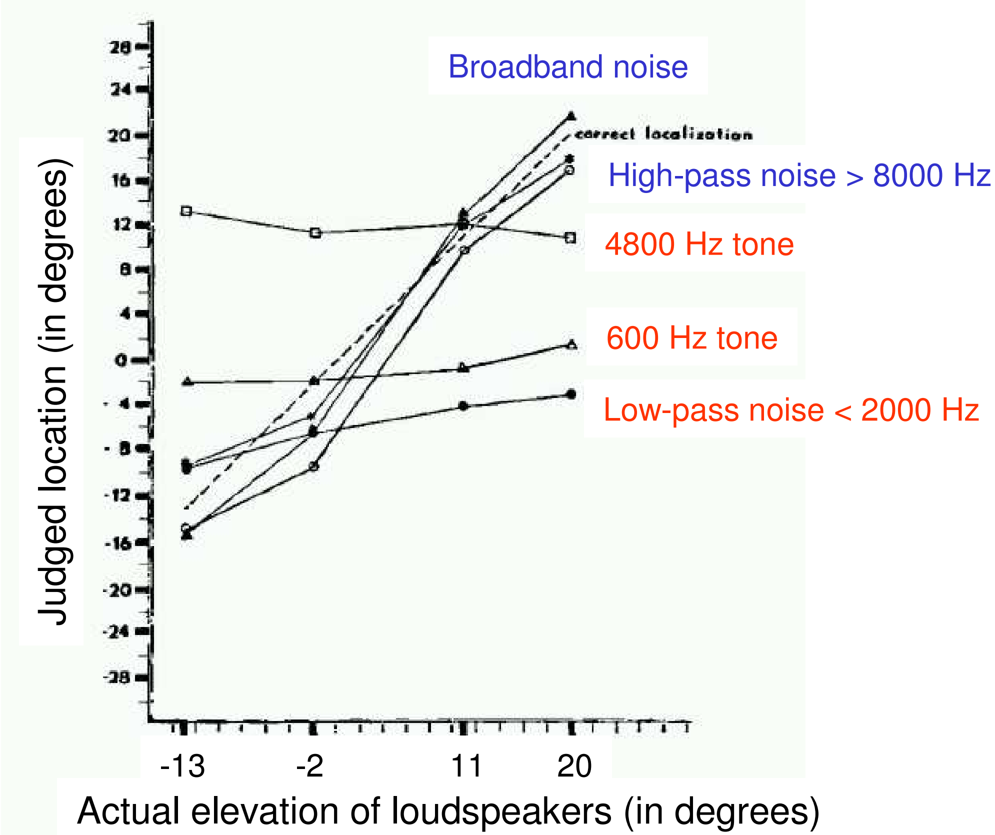


FIG. 2. Mean judged location at each loudspeaker position for simple and complex auditory stimuli. Open squares—4800 cps; open triangles—600 cps; closed circles—<2000-cps noise band; open circles—>8000-cps noise band; asterisks—>2000-cps noise band; closed triangles—broad-band noise.

- Azimuth 0°
- Loudspeakers at various elevation angles
- Pure tones, filtered or broadband noise
- No (minimal) head movements

- Perceived location was roughly constant for pure tones and low-pass noise
- High-pass or broadband noise could be localised

- Results due to effect of pinnae and head and wavelength of sound relative to these

Distance perception (1)

Distance perception is affected by:

• interaural level differences

- large ILDs indicate nearby source
- distance judgements are generally better when one ear is oriented towards source (Holt and Thurlow 1969)

• changes in spectrum and familiarity with sounds (e.g. Coleman 1962)

- high frequencies are attentuated due to absorbing properties of the air - comparisons of loudness and frequency spectrum are generally required

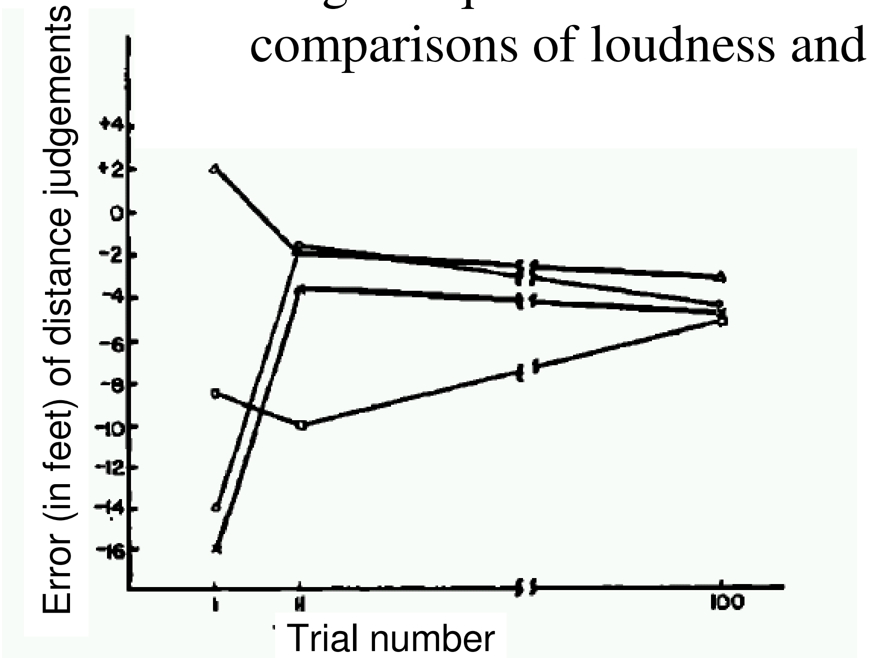


FIG. 1. Error of judged distance of sources of a previously unfamiliar sound as a function of trials and source distance.

• loudspeakers at various distances (approx. 3 m to 9 m)

• stimuli presented in random order

• listeners could not judge the distance of a sound on first hearing, but their judgements improved on subsequent trials

Distance perception (2)

Distance perception is affected by:

- sound level and expectations (e.g. Gardner 1969)
- environment & reverberation (covered next)

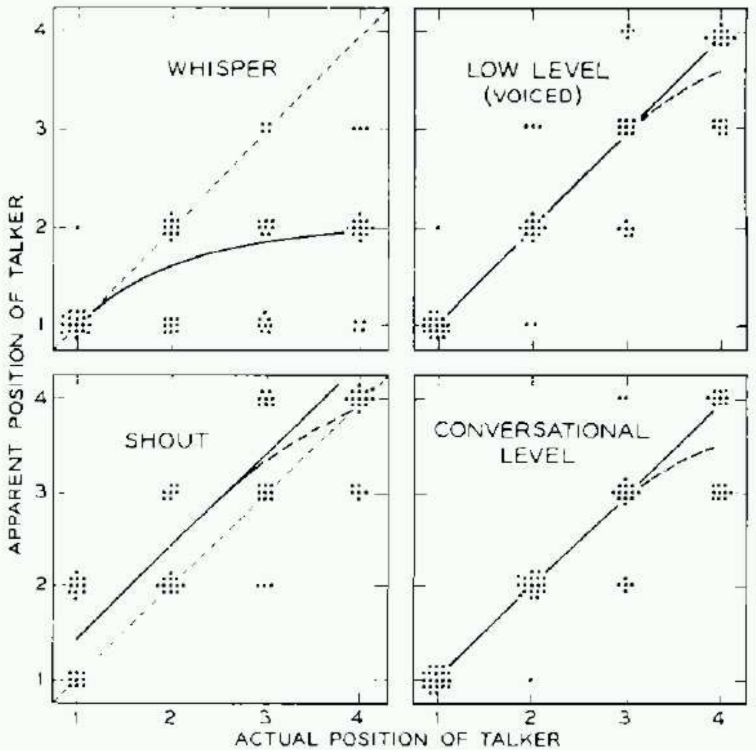


FIG. 7. Comparison of actual and apparent locations of 0° sources of live speech.

- 4 loudspeakers at azimuths 0°
- distance 3, 10, 20, 30 feet (approx. 1 m to 10 m)
- anechoic conditions

- perceived distance determined by level
- but whispered speech always assumed to be nearby

Note also that localisation of unfamiliar sound sources is poorer than that of familiar sources (Plenge 1972) - but familiarisation occurs within a few seconds

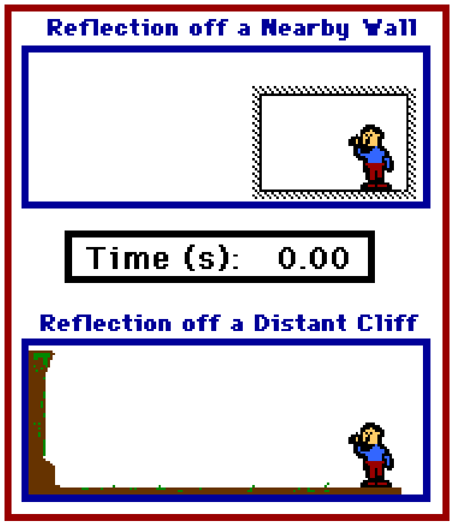
Effect of listening environment

Sounds are affected by the environment:

- room surface, buildings
- other nearby objects

and the position of the listener and sound source relative to these

In reverberant environment, reflected sound reaches ears after delay



Longer delays (> ~40 ms for complex sounds) are heard as distinct echoes

(audibility of echoes is also influenced by other factors, e.g. changes in spectrum and direction of incidence)

Shorter delays are fused

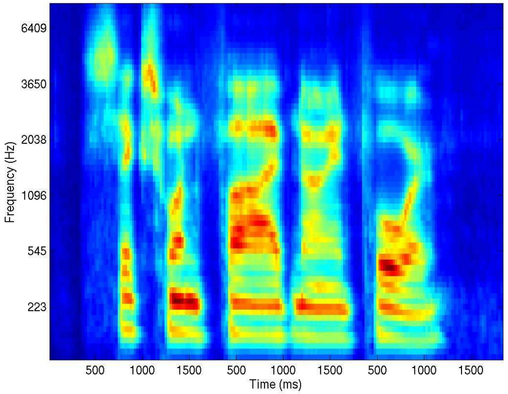
http://gbs.glenbrook.k12.il.us/Academics/gbssci/phys/mmedia/waves/er.html

Reverberation

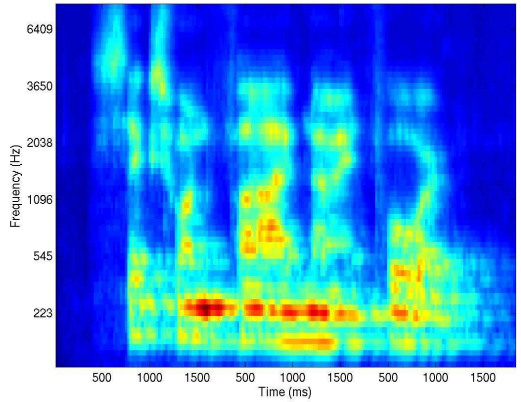
Surfaces can be characterised by the reverberation time T_{60} (the time taken for the sound level to decay by 60 dB after the sound source is turned off)

Examples (simulated reverberation using **roomsim** software):

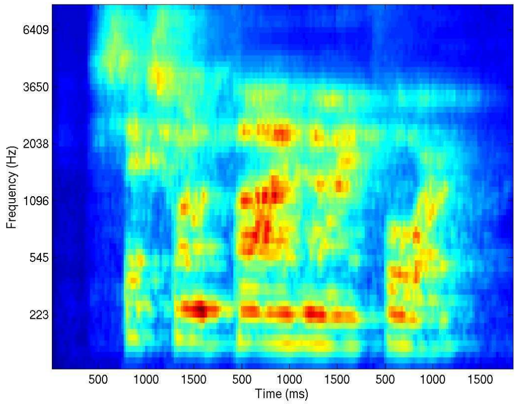
- anechoic (no reverberation)



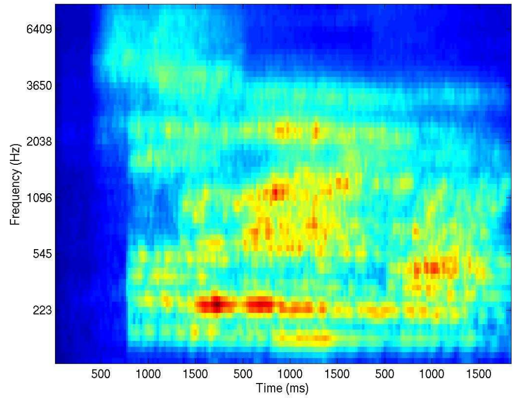
- acoustic plaster (T_{60} = 0.34 s)



- platform floor wooden (T_{60} = 0.51 s)



- glazed wall (T_{60} = 8.70 s)



For a given surface, reverberation time also varies according to frequency

Reverberation effects

Distance perception:

- in anechoic environment, affected by sound level
- in reverberant conditions, independent of sound level - reflections used (Nielsen 1992)
- ratio of direct to reflected sound can be used to judge distance (Mershon and Bowers 1979)

Reverberation also affects determination of azimuth and elevation of source:

- 'precedence effect' - if two sounds are fused, location is determined principally by first sound (also affected by duration, intensity, consistency between sounds)
- increased reverberation can decrease localisation accuracy, especially for low frequencies

Limitations of auditory system - MAA

Minimum audible angle (MAA): defines resolution of auditory system (ability of listeners to detect a shift in direction from reference direction)

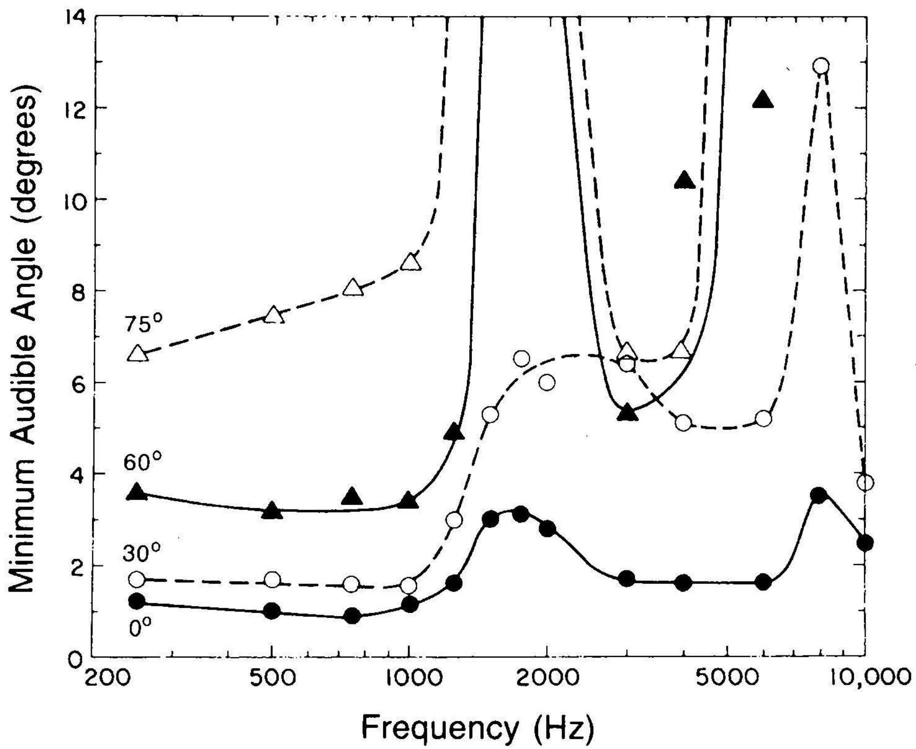


FIG. 6.5 The minimum audible angle (MAA) for sinusoidal signals, plotted as a function of frequency; each curve shows results for a different reference direction. Data from Mills (1958, 1972).

Frequency and azimuth dependent (Mills 1958,1972):

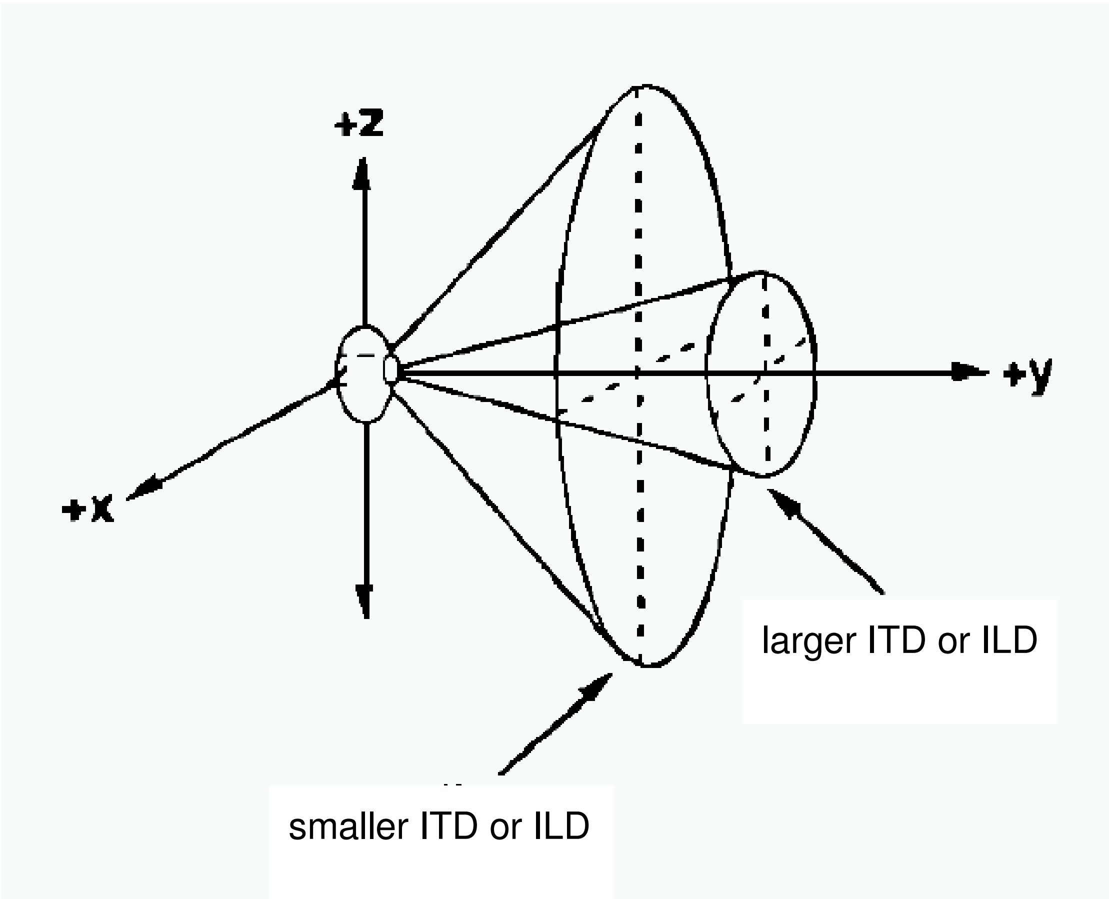
- around 1° for source at 0° azimuth and frequencies below 1000 Hz

- worsens for sources at larger azimuths and higher frequencies, i.e. listeners cannot detect change in direction for large angles

Cone of confusion

Multiple points in space have same ITD and ILD

For a spherical head and ignoring pinnae, the surface (centred on interaural axis) on which these points lie is known as the ‘cone of confusion’ (strictly a hyperboloid)



Head movements

Head movements may resolve ambiguities: comparing changes in perceived location with changes in head position e.g. Mackensen (2004)

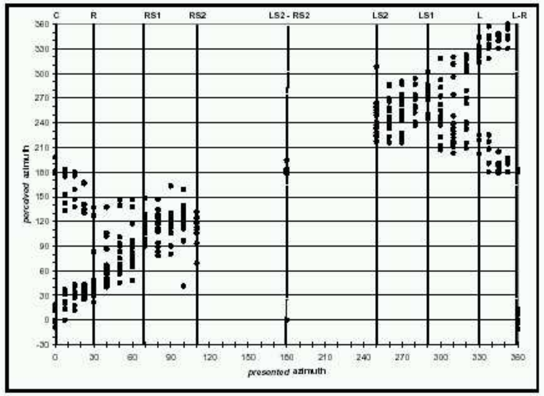


Fig. 4.7: Localization using a fixed dummy head without head tracking under natural listening conditions (IRT studio). A lot of front-back-inversions occur in the region between the first surround loudspeakers LS1 and RS1.

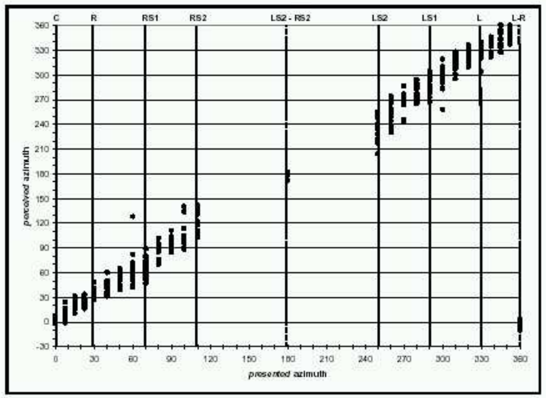


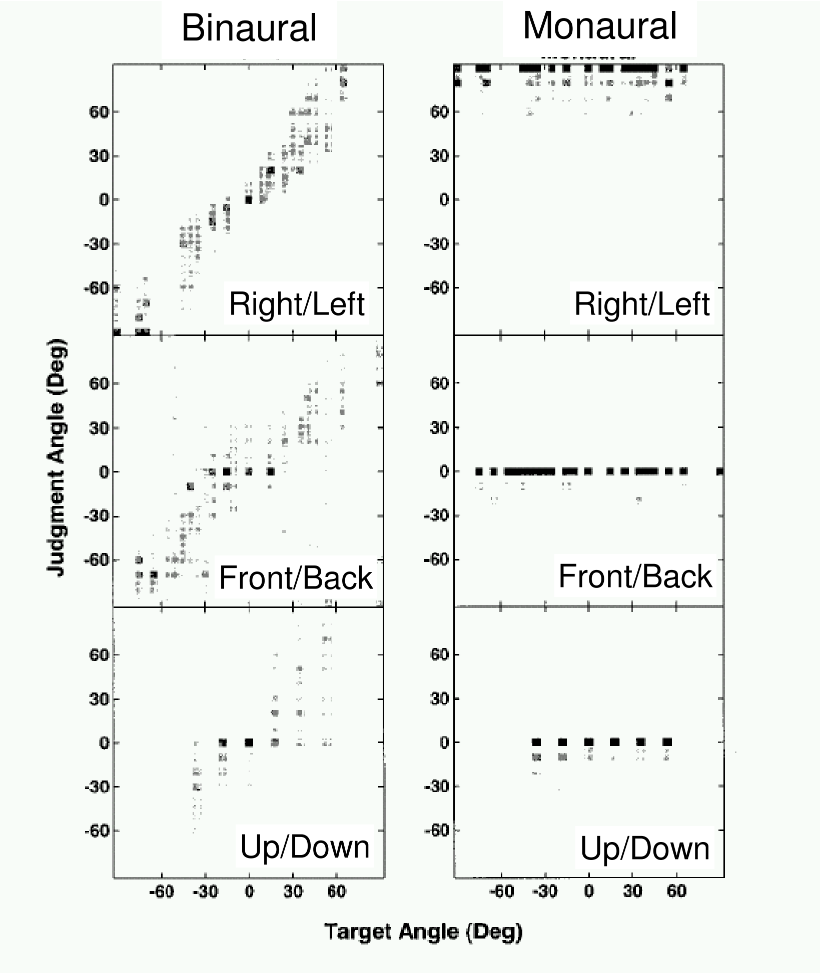
Fig. 4.8: Localization using a dummy head *with* head tracking in the studio (natural listening conditions). The various front-back-inversions of the fixed case do not exist anymore, and the localization resembles that of natural hearing.

Increased front-back confusions when listeners are unable to use head movements (top)

Front-back confusions do not arise when head movements are allowed (bottom)

Monaural localisation

Localisation ability is severely disrupted if only one ear is stimulated, although listeners can make use of minimal information in the other ear (Wightman and Kistler 1997)



- Broadband stimuli based on listener's own HRTF presented via headphones
- Level difference of 70 dB between right and left ears
- Azimuth and elevation of source was varied
- Listeners could judge position in binaural but not monaural condition

(Note: earlier experiments suggested monaural localisation was possible, but this was probably due to low level input in other ear)

Moving sources

Moving source exhibits spectral and other changes

Only slow changes can be followed (Perrott & Musicant 1977)

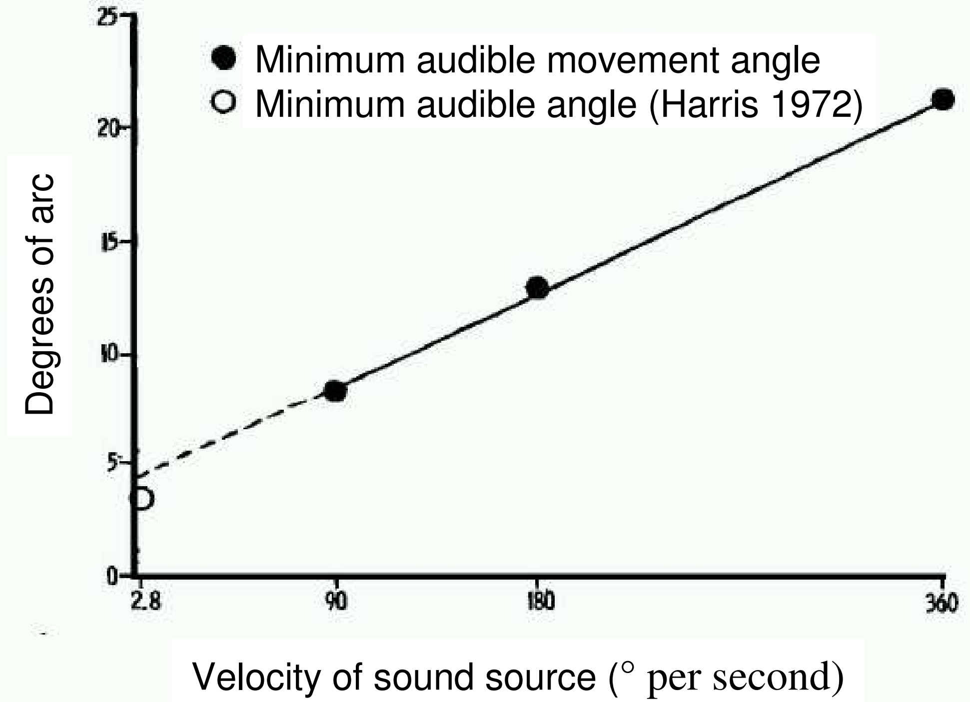


FIG. 1. Minimum audible movement angle thresholds were determined using a least-squares analysis for a 75% correct criteria. Angles (in degrees of arc) are plotted as a function of the velocity of the moving sound source. The dotted line represents an extension to Harris' MAA (1972).

- 500 Hz sine wave presented through moving or stationary loudspeaker
- listeners were asked whether sound was from moving or fixed speaker

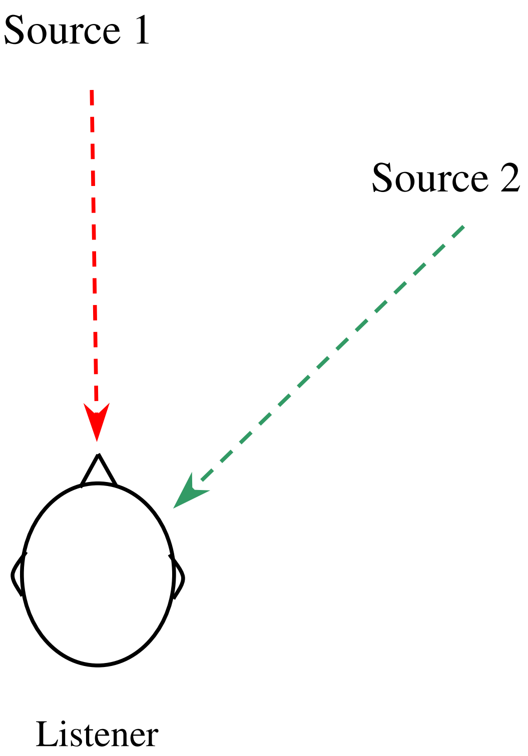
- detectable angle depends on speed of movement of source, up to ~21° for source moving at 360°/s

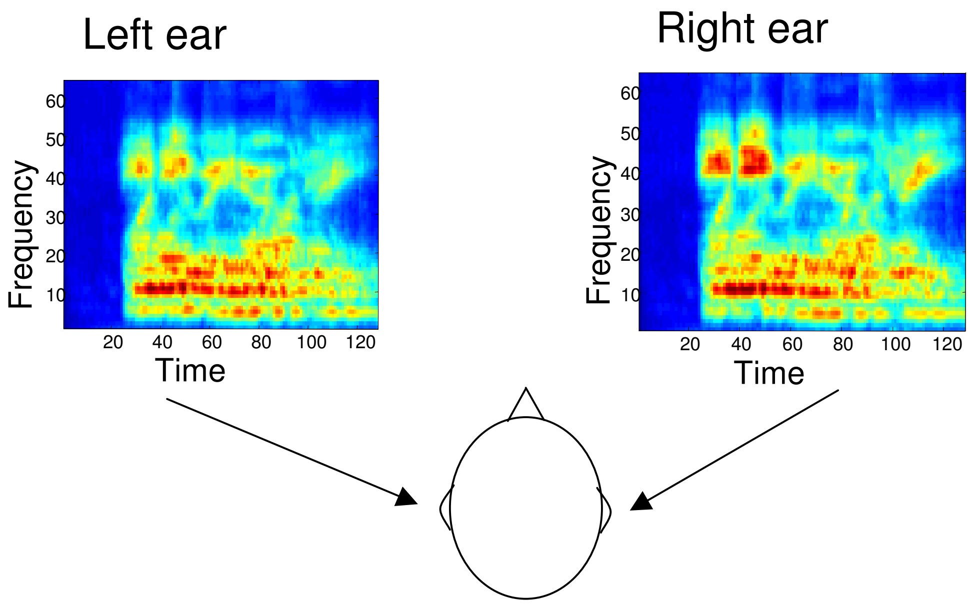
Dealing with multiple sources

Rarely have only one source present

How easy is it to segregate multiple sources?

- Where is each source?
- Which parts of signal were produced by each source?





Localisation in noise

Presence of background noise reduces localisation accuracy of click-train stimuli, especially front/back distinction (Good & Gilkey 1996)

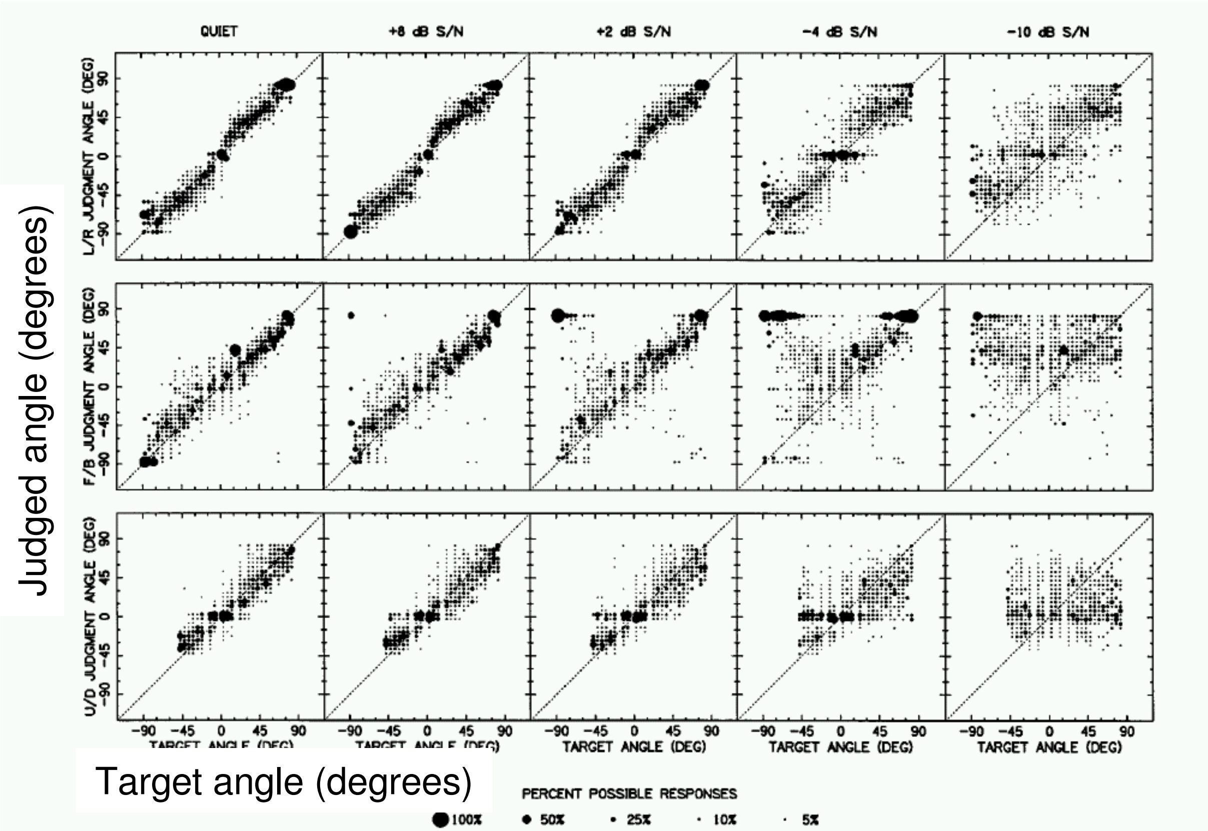


FIG. 1. Performance in the left/right (L/R), front/back (F/B), and up/down (U/D) dimensions is shown for one subject (MG) at each of five signal-to-noise ratios (relative to the same-speaker detection threshold). In each panel, the judgment angle is plotted as a function of the target angle. Targets are grouped into 5°-wide target-angle bins. The size of each symbol represents the percentage of the total number of judgments in each target-angle bin that fall within each 5°-wide judgment-angle bin. The masker, if present, was always located at 0° L/R, 90° F/B, and 0° U/D.

Top row: left/right judgements; middle: front/back; bottom: up/down

- Stimuli (broadband click-trains) presented at 239 sptial locations (azimuth 0 - 360°, elevation -45° - 90°)
- Stimuli masked by broadband noise at 0° azimuth and elevation (directly in front of listener)
- Signal-to-noise ration (SNR) ranged from +14 to -13 dB
- Results show target angle v. judged angle for each SNR
- Localisation accuracy decreased with increasing SNR

Localisation of multiple talkers

Localisation accuracy is not affected by the number of competing talkers, as long as both ears are used (Hawley et al 1999)

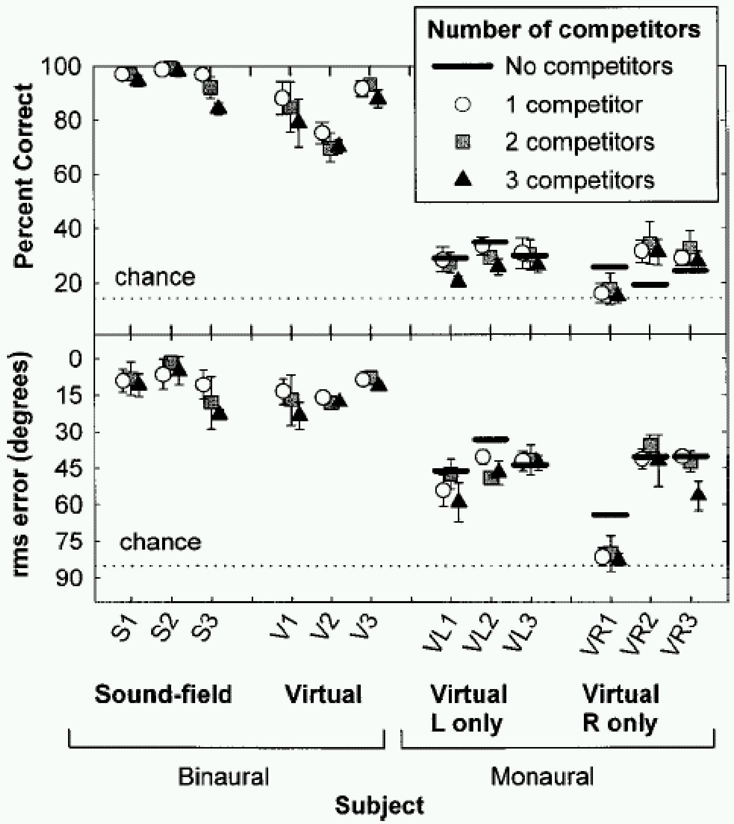


FIG. 6. Results for localization experiments for individual listeners. The percent correct and the root-mean-square (rms) error are plotted for each competing sentence configuration. The symbol and shading denoting the number of competing sentences (white for one, gray for two, black for three). Dark lines denote performance for localization in quiet for monaural listeners.

- Stimuli (sentences) presented at 7 spatial locations (azimuth -90° - +90°) through loudspeakers or headphones

- Stimuli masked by 1 to 3 competing sentences (same talker) at azimuth separation ranging from 0° to 180°

- All sentences had same level

- Localisation accuracy was good when both ears used; poor when only one used

Contrast with Good & Gilkey (who found presence of background noise reduces localisation accuracy) : may be due to stimuli and/or conditions (SNR)

Intelligibility of multiple talkers

Proximity of talkers has more effect than number of talkers, on both intelligibility and localisation accuracy

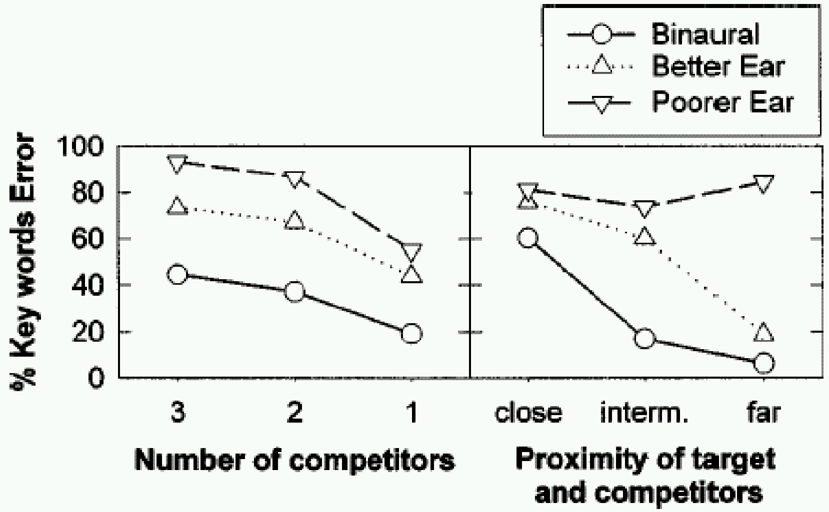


FIG. 4. Error rates in virtual-listening speech intelligibility experiments as a function of the number of competitors and the proximity of the competitors to the target location for binaural (circle), better monaural ear (up-pointing triangle) and poorer monaural ear (down-pointing triangle).

- Stimuli (sentences) presented at 7 spatial locations (azimuth -90° - +90°) through loudspeakers or headphones

- Stimuli masked by 1 to 3 competing sentences (same talker) at azimuth separation ranging from 0° to 180°

- All sentences had same level

- Speech intelligibility was affected by proximity of competing speech

Later results (Hawley et al 2004) suggest better intelligibility for speech masked by speech than speech masked by noise

Multiple sources - BMLD

Listening with two ears can reduce threshold of audibility

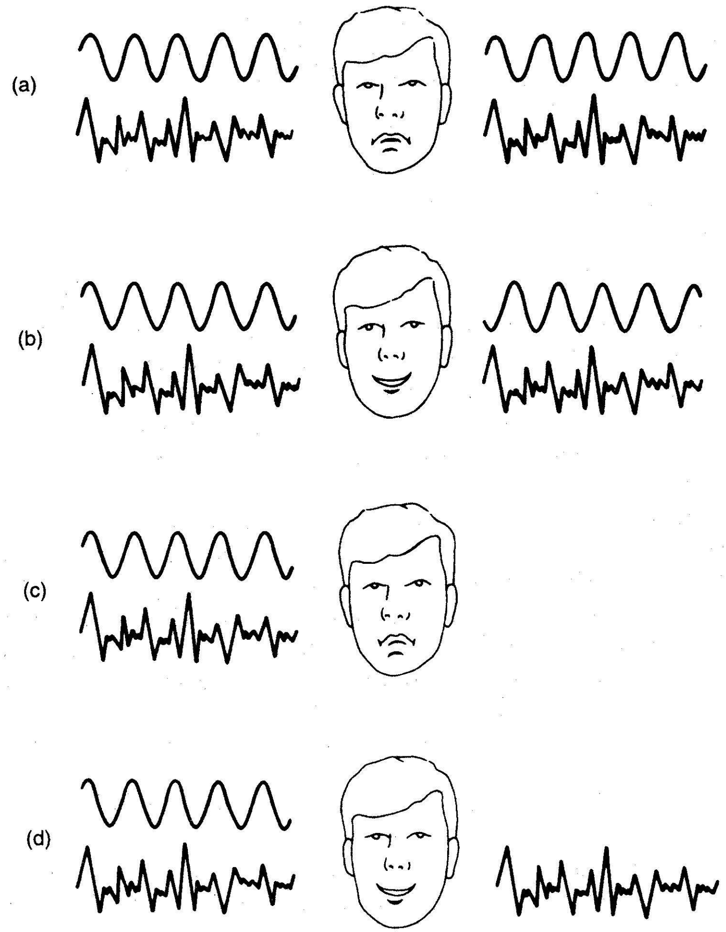


FIG. 6.9 Illustration of two situations in which binaural masking level differences (MLDs) occur. In conditions (a) and (c) detectability is poor, while conditions (b) and (d), where the interaural relations of the signal and masker are different, detectability is good (hence the smiling faces).

- If a tone is just masked by a broadband noise when presented to both ears, then if the phase of the tone is changed by 180° it becomes audible

- If the noise is increased to just mask the tone again, difference is 'binaural masking level difference' (BMLD)

- If noise and just-masked tone are fed to one ear only, then noise alone is fed to other ear, tone becomes audible

- If tone is then added to second ear, it becomes inaudible

Segregation and grouping

Segregation of sources (using frequency, time cues) should help interpretation of environment
- basis of Auditory Scene Analysis (Bregman 1990)
- features with similar properties should be grouped together

Idea is supported by some experiments, e.g.:

- increased intelligibility if talkers are at different locations (Hawley et al 1999, 2004)
- decreased intelligibility if speech is alternated from one ear to the other, depending on rate of switching (Cherry & Taylor 1954)

but evidence also exists that ear of presentation doesn't always segregate, i.e. cues for segregation can be overridden, e.g.:

- speech sound split between two ears is fused into a whole (Broadbent 1955; Broadbent & Ladefoged 1957)
- duplex perception: partial speech sound in one ear plus non-speech chirp in another fuses into complete speech sound plus segregated chirp

2. Computational models

Early models (coincidence, equalisation-cancellation)

Later developments

Computational source localisation

Problems and suggested solutions

Early models of binaural processing

Two classical models:

a) Jeffress (1948) coincidence-based model

- coincidences in neural firings from each ear for corresponding frequency bands are identified using a delay mechanism
- ITD sensitive units

b) 'Equalisation-cancellation' model

Kock (1950), developed by Durlach (1963)

- designed to model binaural masking level differences (BMLD)
- signal in one ear is transformed so that one component (the 'masker') matches that in the other ear; then one signal is subtracted from the other

Later models of binaural processing

Jeffress (1948) coincidence-based model, adapted by Colburn (1973), plus later developments

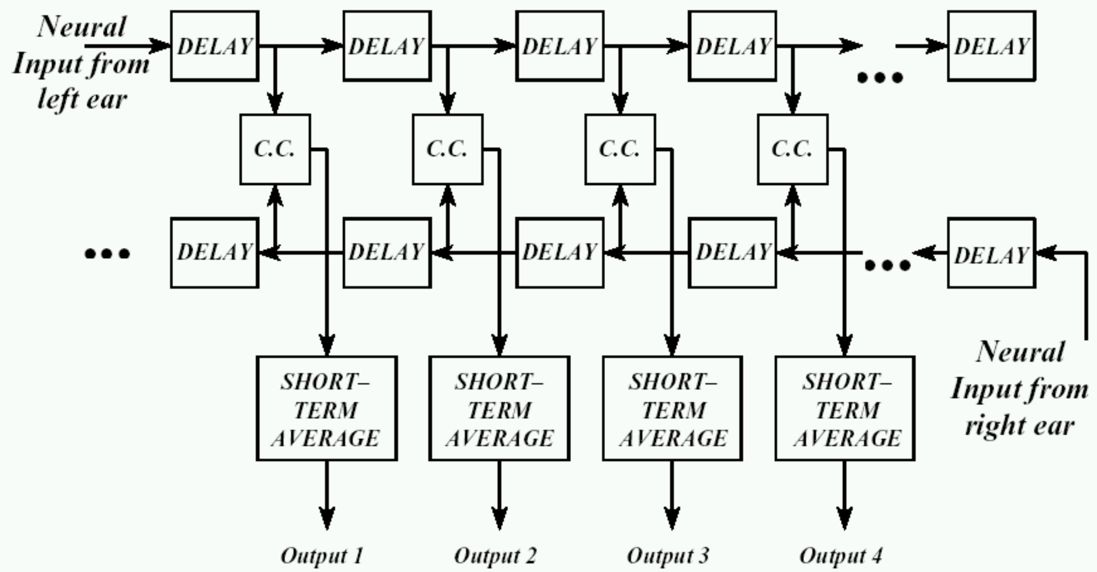


Figure 2. Schematic representation of the Jeffress place mechanism. The blocks labelled *C.C.* record coincidences of neural activity from the two ears (after the delays are incurred).

Figure from Stern & Trahiotis (1995)

Other extensions exist and incude additional features such as inhibition, HRTF adaptation, ILD weighting e.g. Stern, Colburn & Trahiotis; Blauert, Lindemann and colleagues

Includes model of auditory nerve activity

Implemented as cross-correlation between the neural responses to stimuli (early models used the stimuli directly)

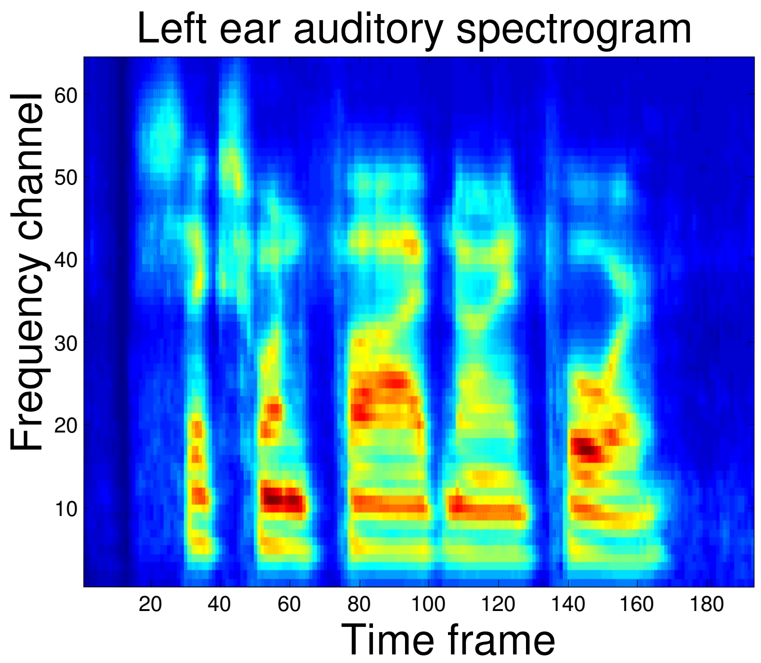
Considered as a generalisation of the EC model - interaural delays perform equalisation role

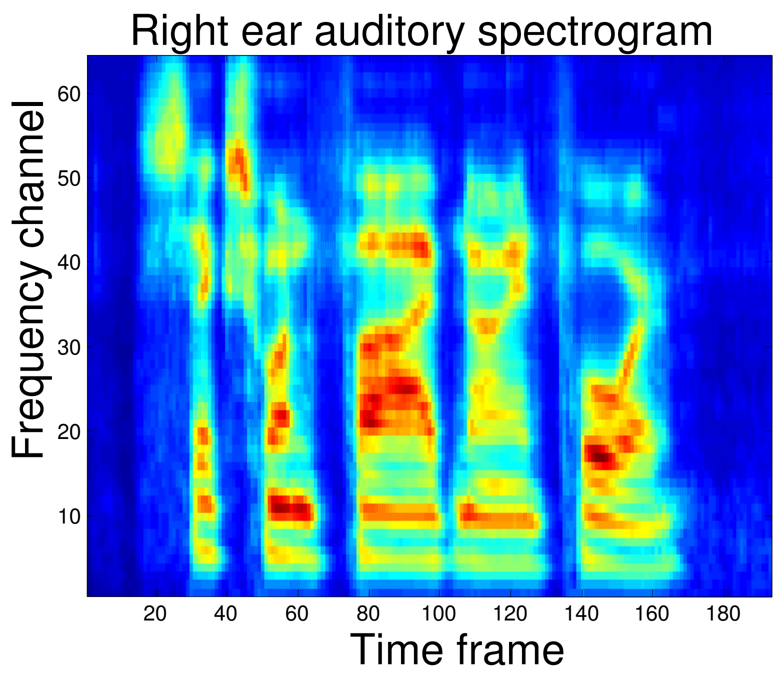
Computational source localisation

Typical steps in processing:

- Monaural processing of signal entering each ear, using auditory filterbank within moving analysis window (typically 20 ms, shifted by 10 ms)

  (Note: use input to ear, i.e. stimuli processed using HRTF)
- ITD: cross-correlation between left and right ear BM activity
- ILD: ratio of left and right ear envelope



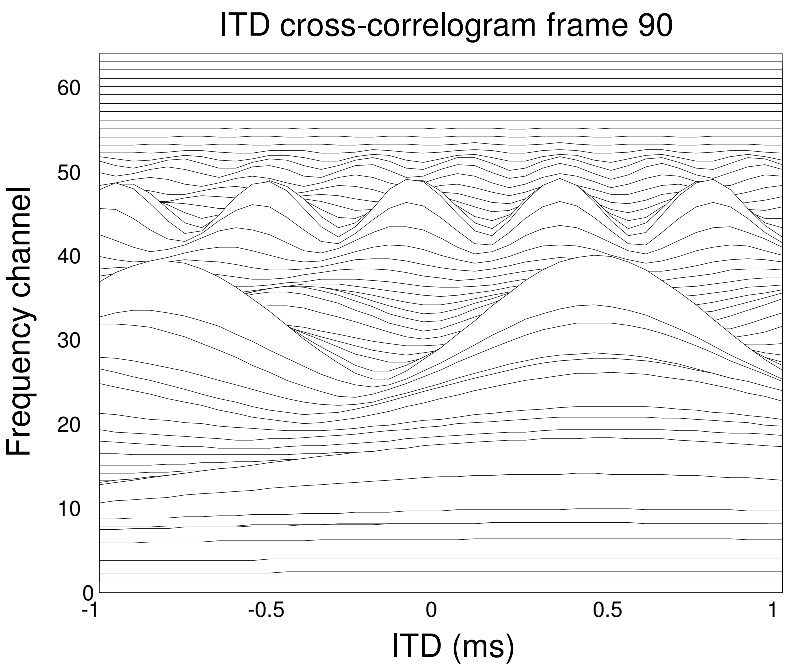


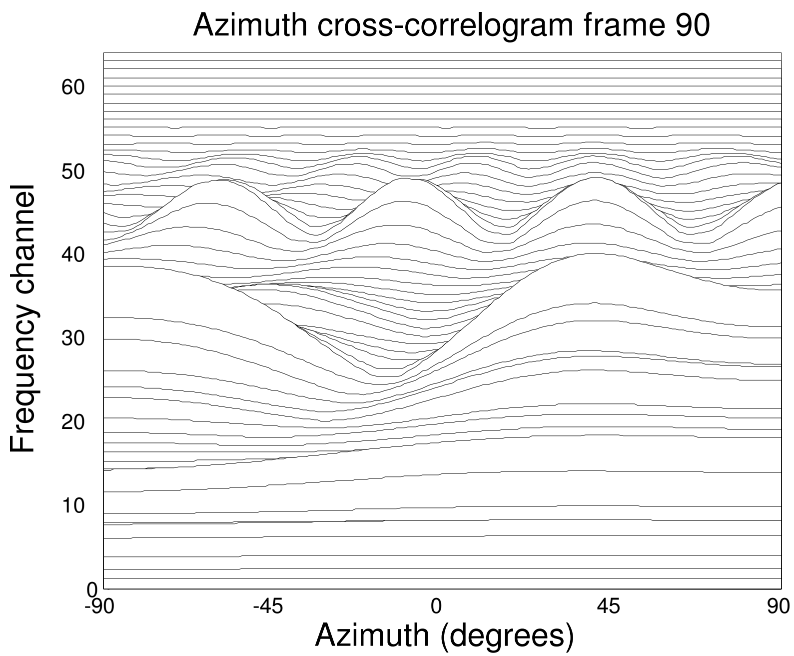
Cross-correlation for a single source

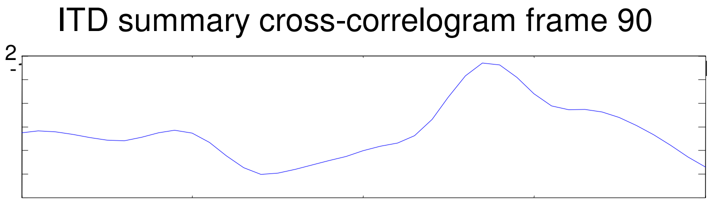
Cross-correlogram example for a ***single*** source at azimuth 40 in ***anechoic*** conditions (time frame 90)

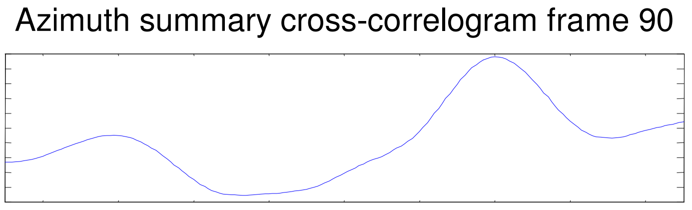
Highest peak in each frequency channel indicates ITD and therefore position of source: convert ITD to azimuth (e.g. using empirical data)

Can sum over all channels and/or over time



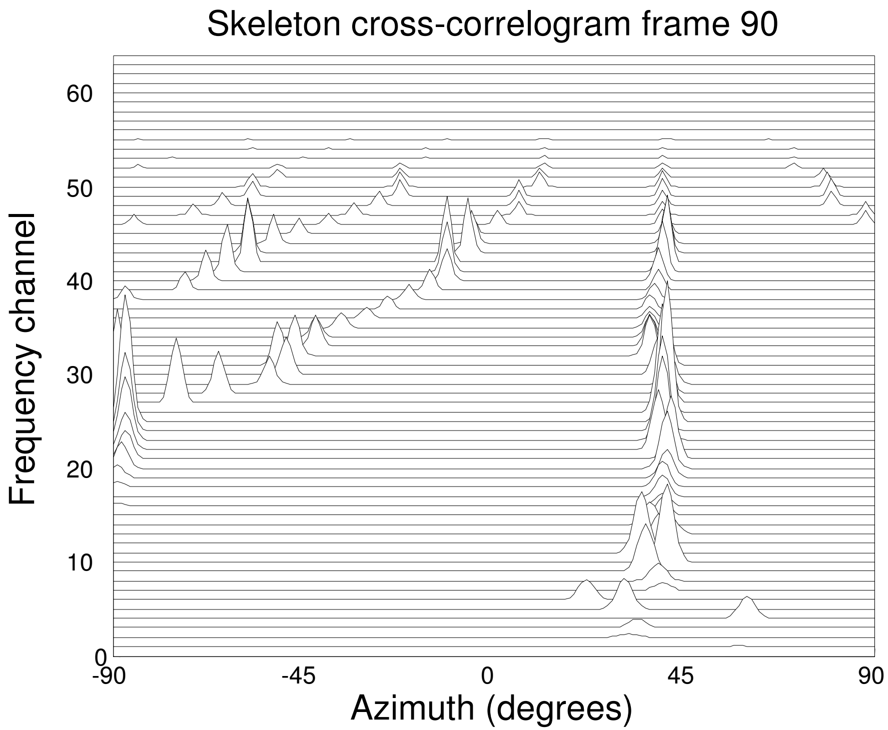


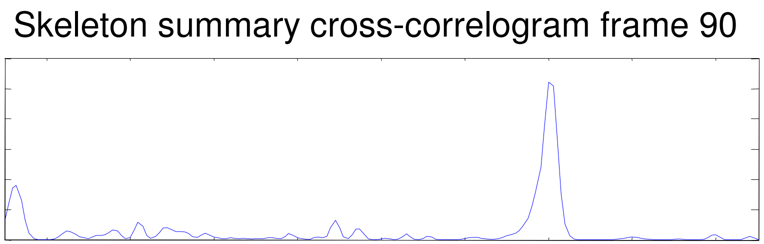




Skeleton cross-correlogram

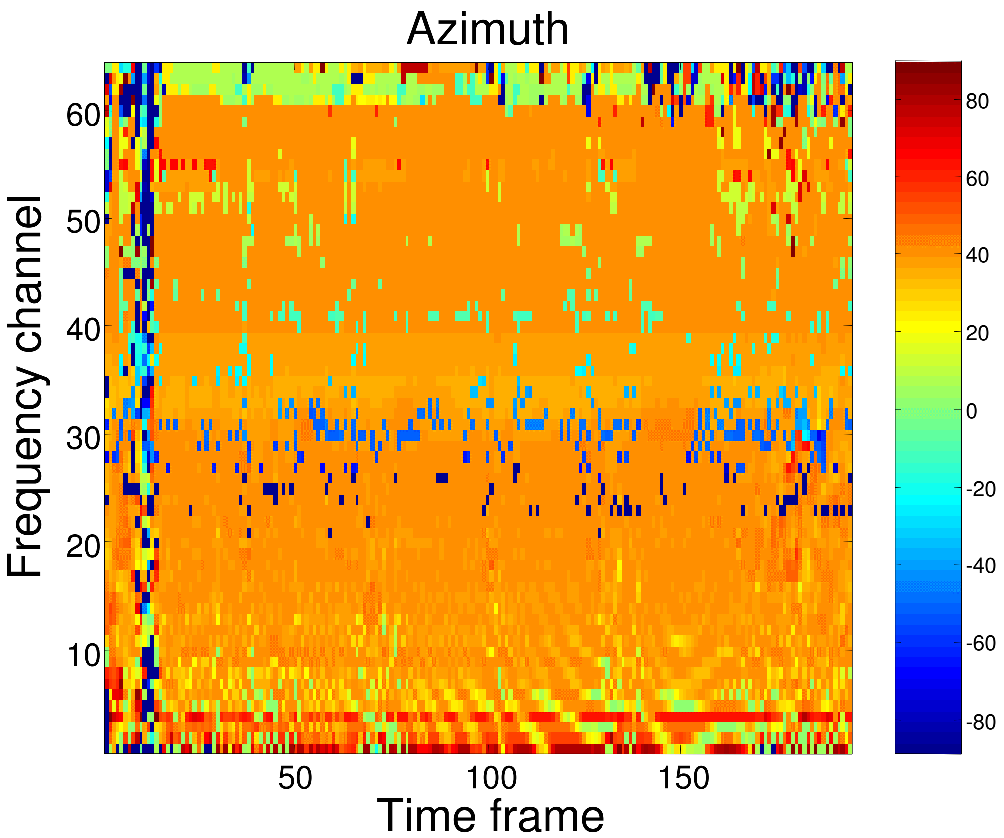
Peaks in cross-correlogram are broad - resolution can be improved by sharpening peaks to produce ‘skeleton’ cross-correlogram – maximum peak is reduced to an impulse and convolved with a Gaussian





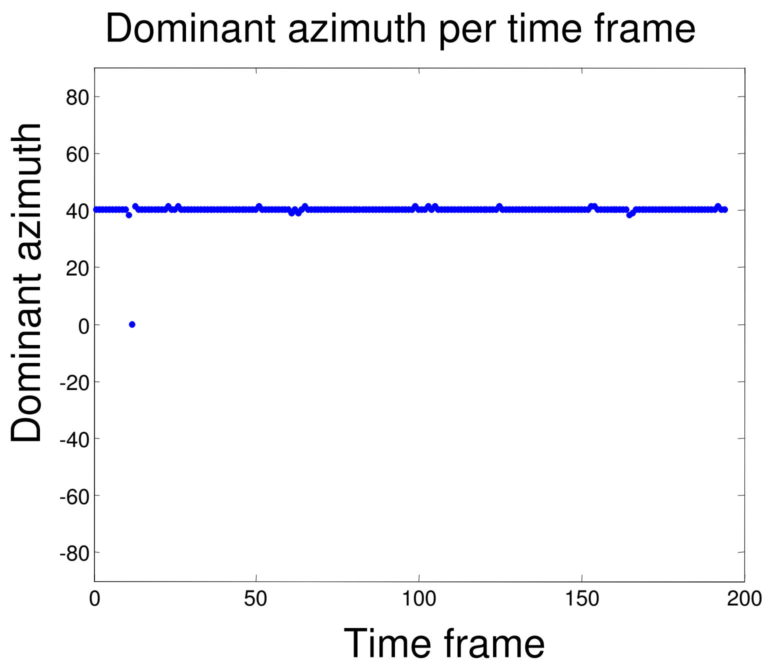
Localising time-frequency elements (1)

Highest peak in each frequency channel indicates azimuth of dominant source in that channel – but not always accurate, even for a single anechoic source



Colour bar shows azimuth value (orange corresponds to azimuth 40 degrees, i.e. actual azimuth of source)

Highest peak in summary cross-correlogram gives dominant azimuth per time frame

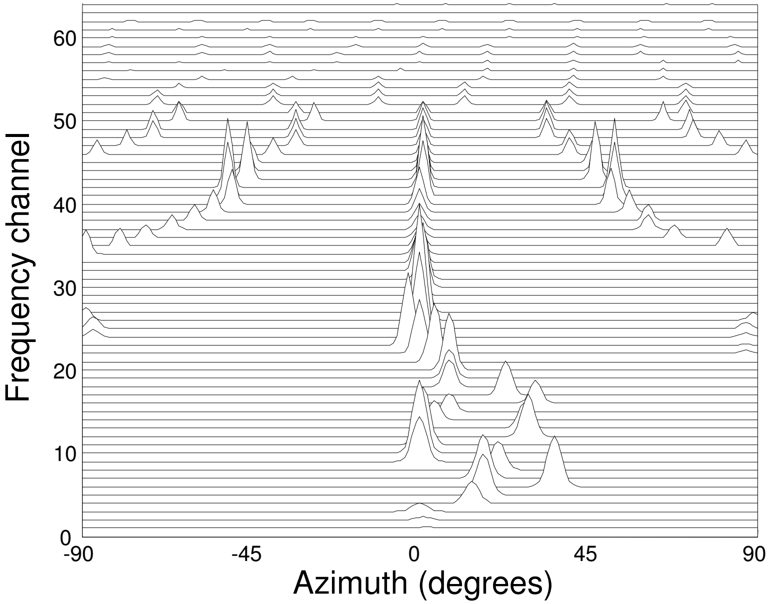


Cross-correlation for multiple sources

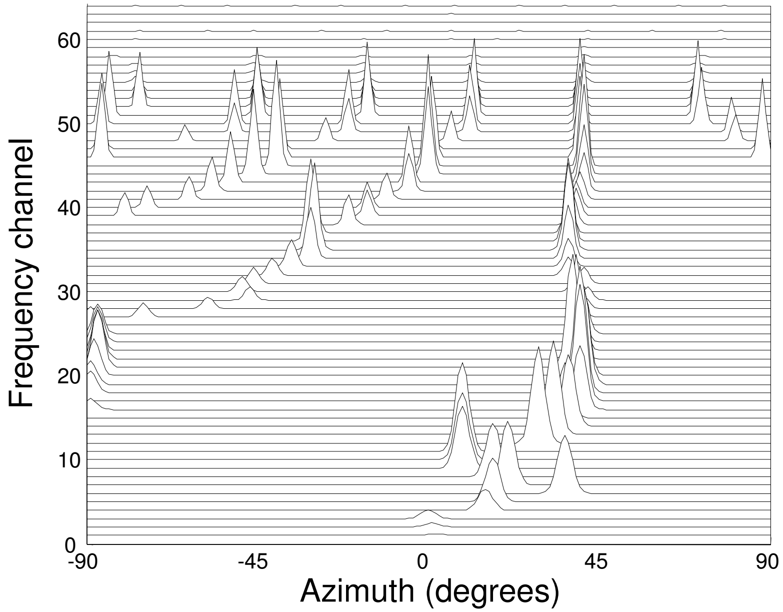
Cross-correlogram example for ***two sources***, one at azimuth 0, one at azimuth 40, in ***anechoic*** conditions (time frames 90 and 105)

Dominant source differs in different time frames

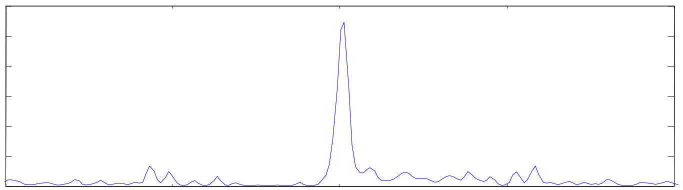
Skeleton cross-correlogram frame 90



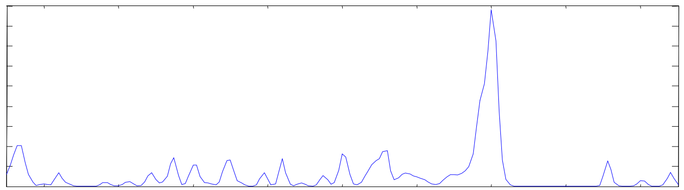
Skeleton cross-correlogram frame 105



Skeleton summary cross-correlogram frame 90

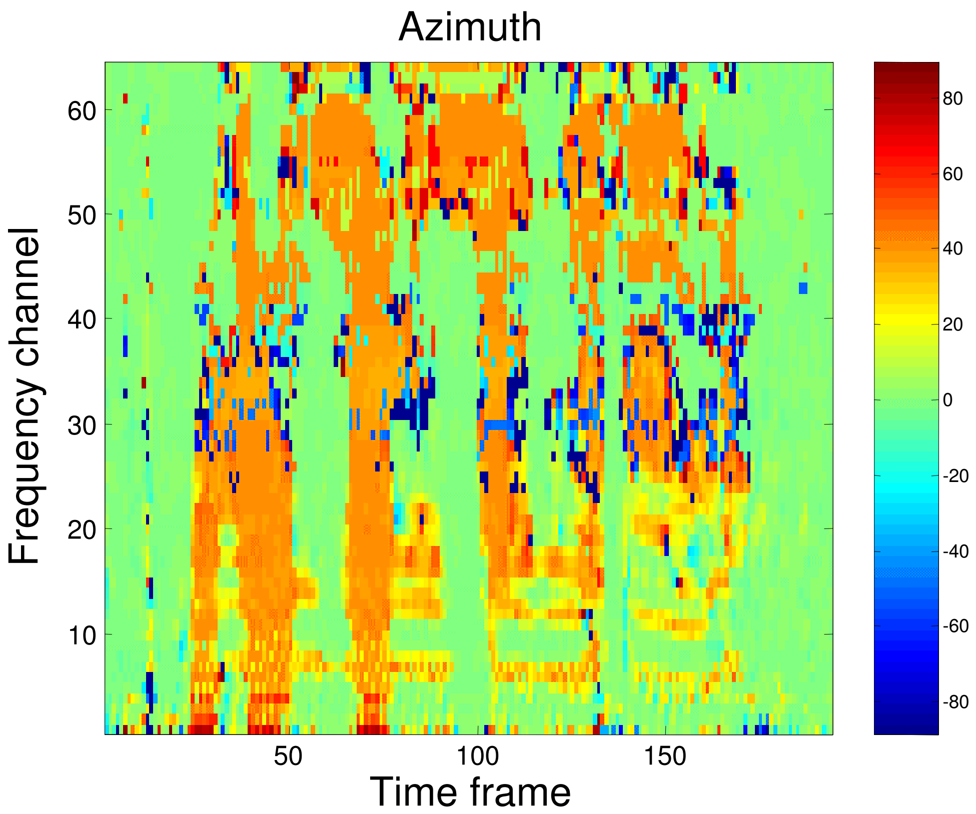


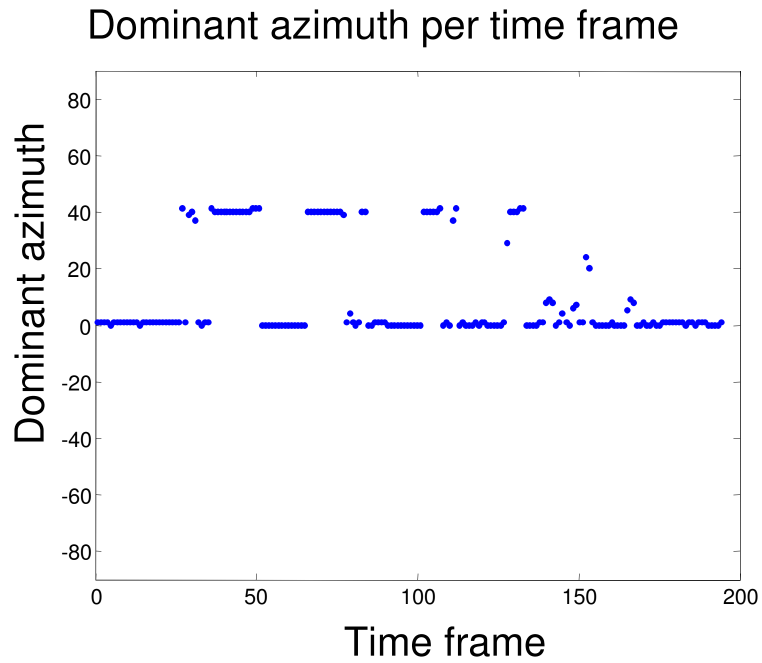
Skeleton summary cross-correlogram frame 105



Localising time-frequency elements (2)

Anechoic sources can be distinguished if sufficiently well separated in space, but some inaccuracies arise





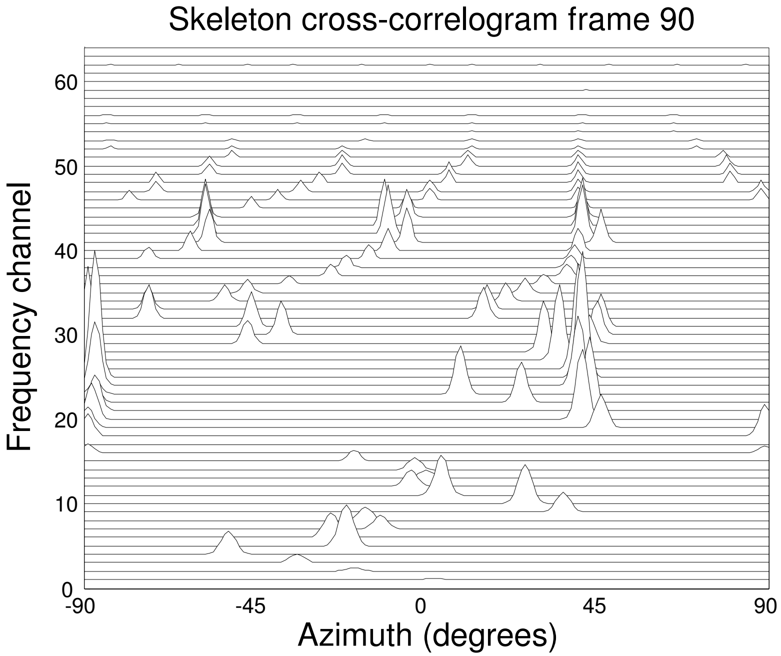
Colour indicates azimuth value (orange corresponds to azimuth 40 degrees, green to 0 degrees)

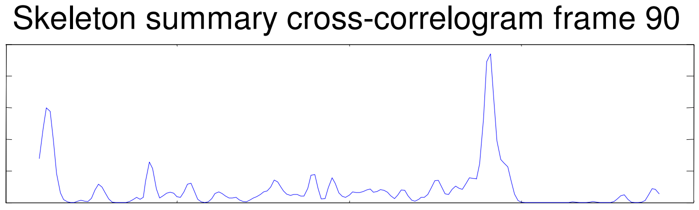
Cross-correlation in reverberation

Cross-correlogram example for a ***single source*** at azimuth 40 in ***reverberant*** conditions (time frame 90)

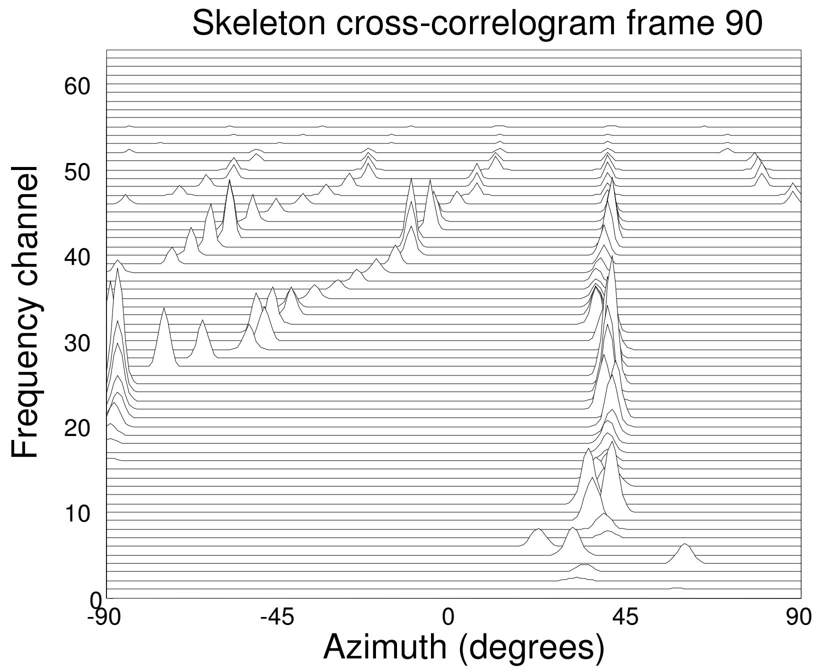
Additional peaks appear

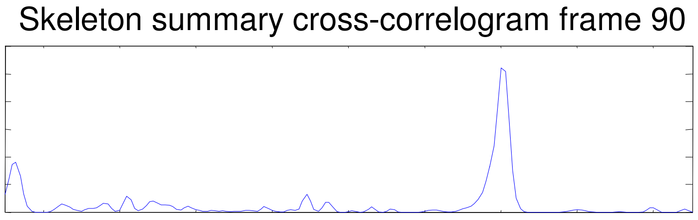
Reverberant





Anechoic (for comparison)

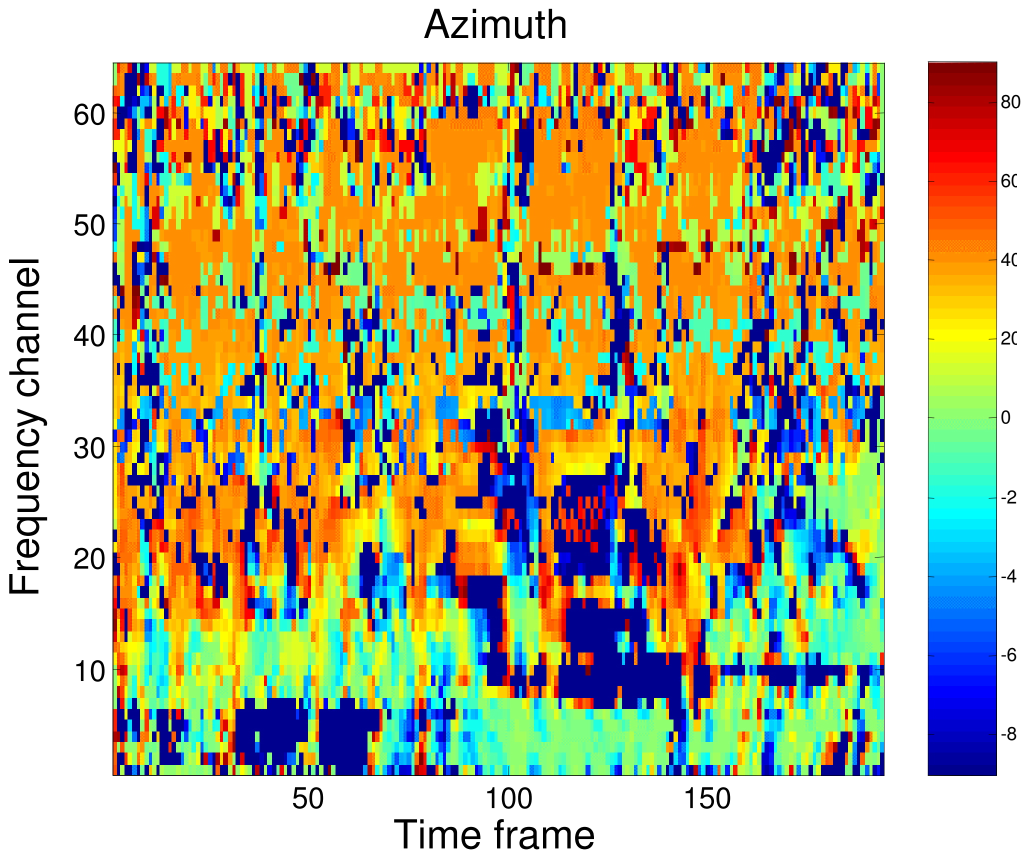


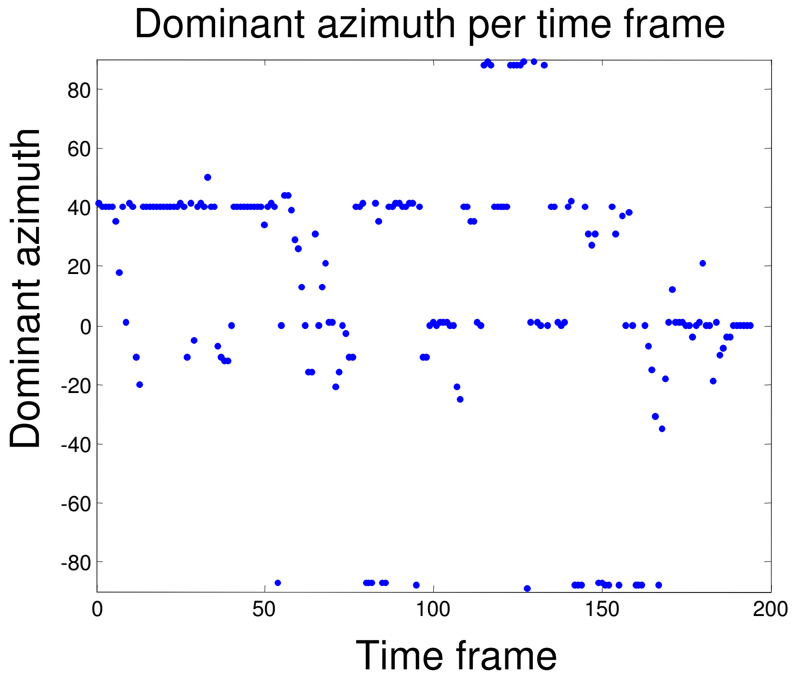


Localising time-frequency elements (3)

Localisation accuracy deteriorates in reverberant conditions

Note example is for a ***single*** source at azimuth 40



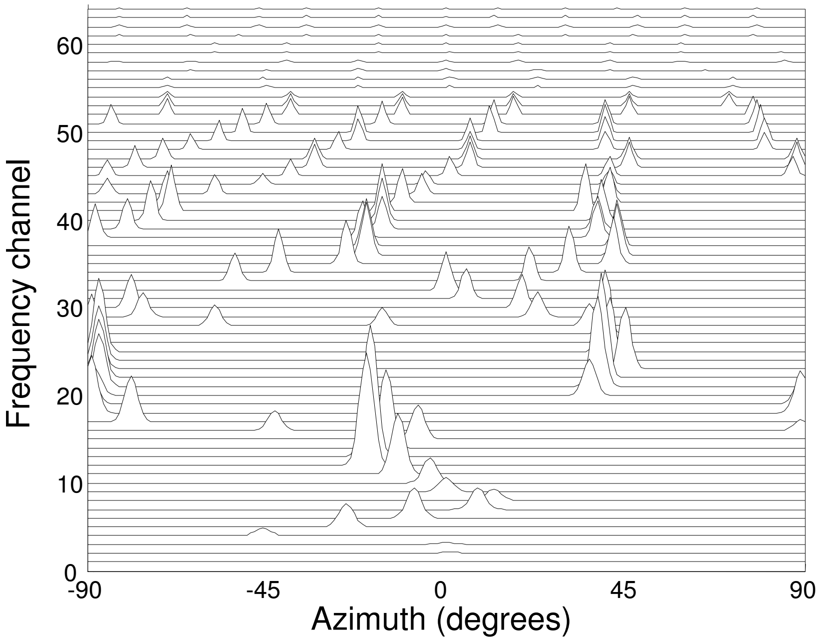


Colour indicates azimuth value (orange corresponds to azimuth 40 degrees)

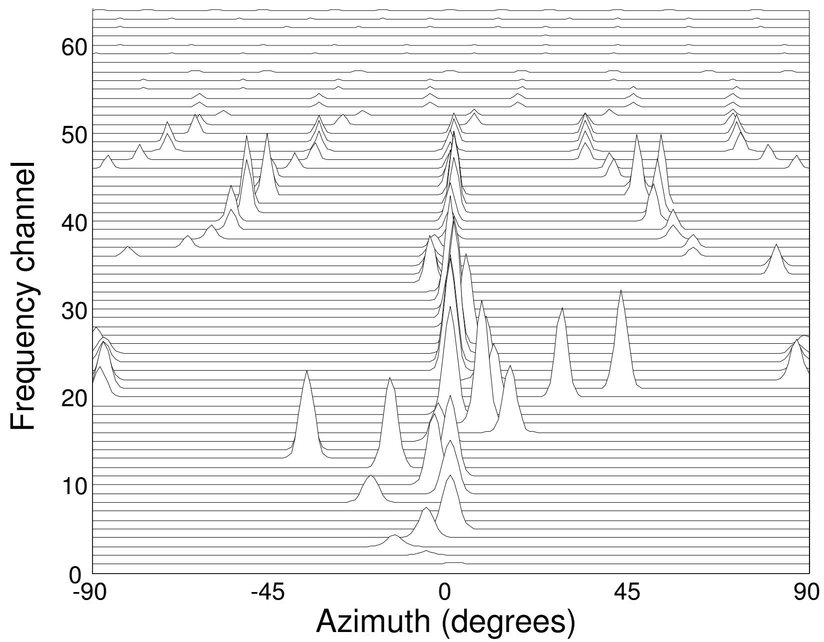
Multiple sources with reverberation

Cross-correlogram example for ***two sources***, one at azimuth 0, one at azimuth 40 in ***reverberant*** conditions (time frames 40 and 90)

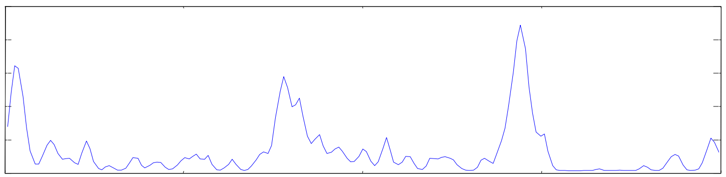
Skeleton cross-correlogram frame 40



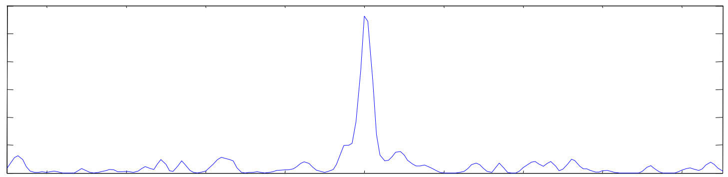
Skeleton cross-correlogram frame 90



Skeleton summary cross-correlogram frame 40



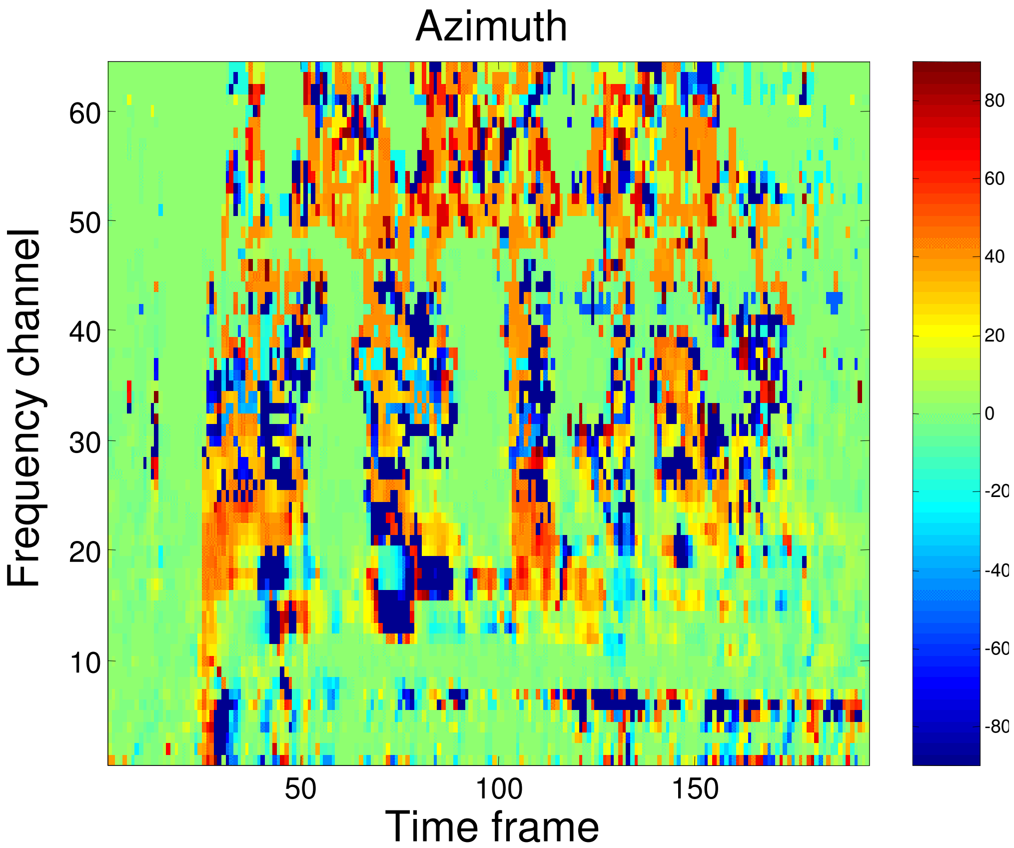
Skeleton summary cross-correlogram frame 90

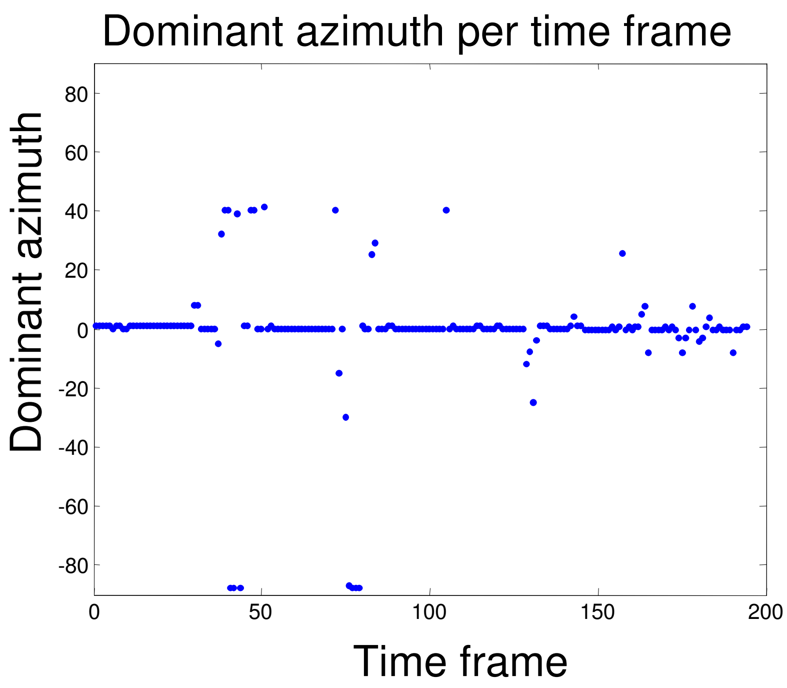


Localising time-frequency elements (4)

Localisation accuracy is poor compared with anechoic conditions

Source at azimuth 0 dominates (symmetry of room and source-listener aids localisation); other source is poorly localised



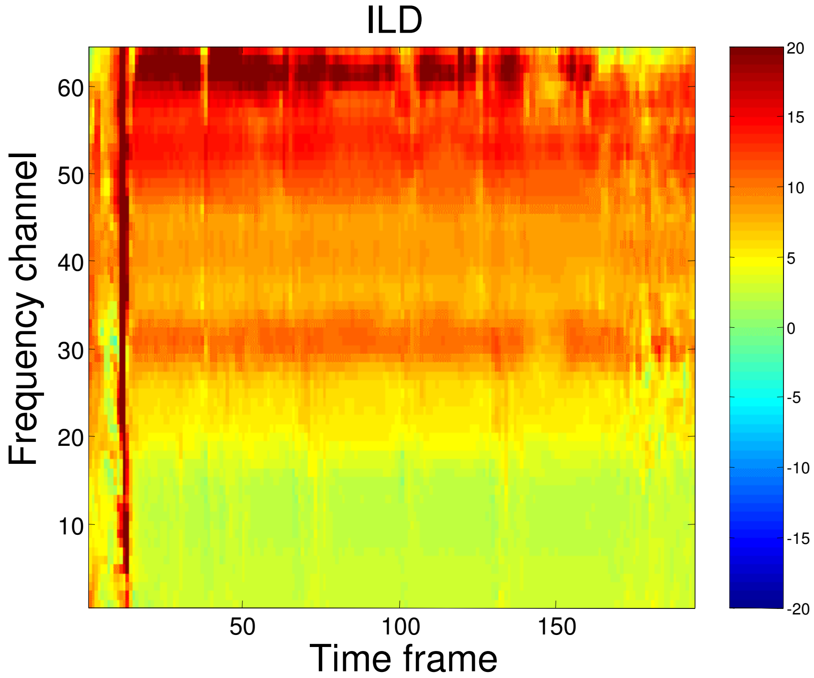


Colour indicates azimuth value (orange corresponds to azimuth 40 degrees; green to 0 degrees)

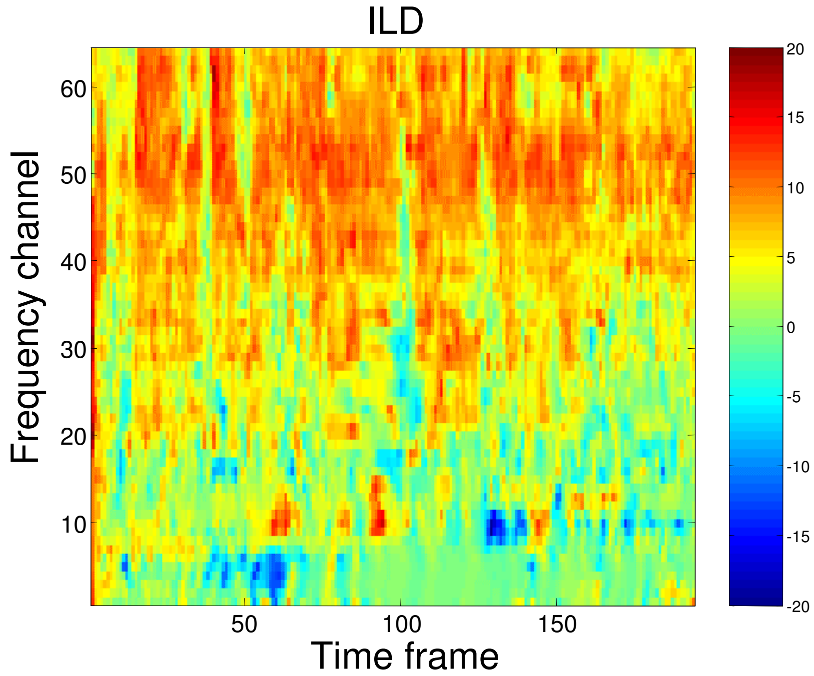
Interaural level difference (1)

ILD cue is less reliable in reverberant conditions
ILD is stronger at higher frequencies

Single source, azimuth 40, anechoic



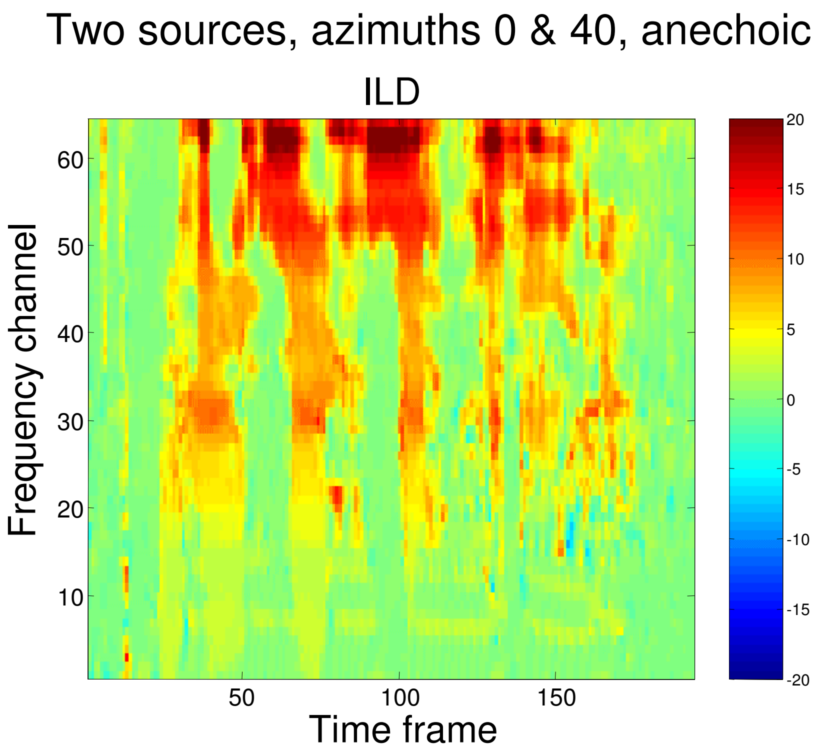
Single source, azimuth 40, reverberant

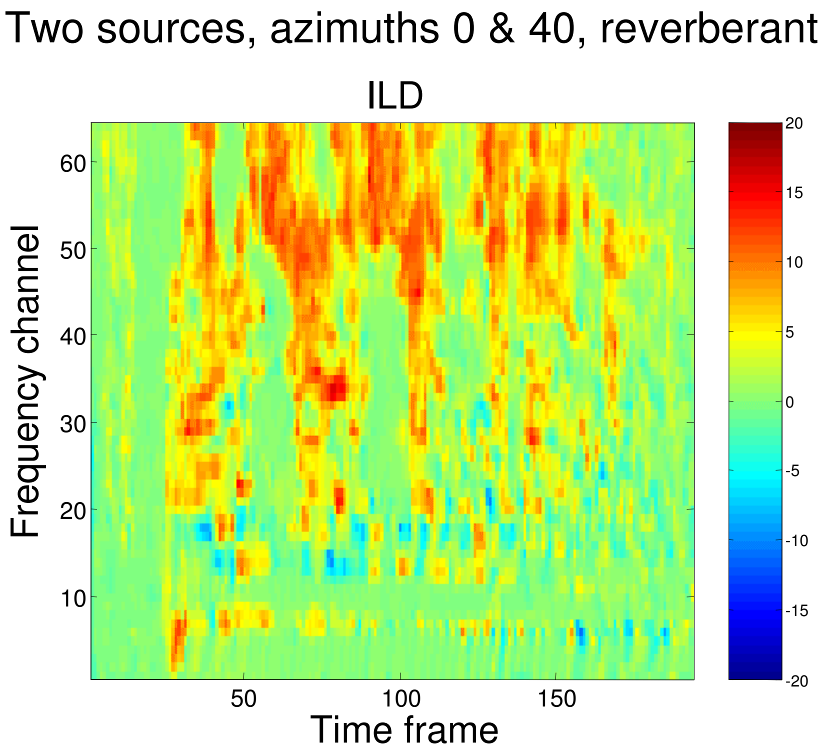


Interaural level difference (2)

ILD cue is less reliable in reverberant conditions

ILD is stronger for source on one side of head





Localisation problems

Problems with cross-correlograms:

a) multiple peaks at high frequencies
b) interactions between sources – incorrect, broad or reduced peaks
c) reverberation effects
d) moving sources

Suggested solutions:

a) Sum cross-correlogram across frequency channels (Lyon 1983)
b) Convert from ITD to azimuth, using supervised training or empirical data (Bodden 1993)
c) Weight frequency bands according to their importance (Bodden 1993)
d) Track peaks over time; measuring amplitude changes (Bodden 1993)
e) Sharpen cross-correlation peaks – skeleton cross-correlogram (Palomaki et al 2004)
f) Subtract (stationary) background cross-correlogram (Braasch 2002)
g) Ignore low-amplitude peaks in cross-correlogram – use ‘interaural coherence’ (Faller & Merimaa 2004)
h) Use template matching (‘stencil’) to identify muliple peaks (Liu et al 2000)
i) Track moving sources using hidden Markov models (Roman & Wang 2003)

Summary

Binaural sound localisation uses cues:

- interaural time difference (ITD)
- interaural level difference (ILD)
- pinna cues

ITD dominates, but cues interact in complex ways (not fully understood)

Cues are affected by:

- nature of source: position, frequency, bandwidth, movement, interactions between sources
- listening environment: proximity and type of surfaces and other obstacles (reverberation)
- listener characteristics: pinnae, head movements, position relative to surfaces & obstacles

- Computational models use HRTFs, cross-correlation and level differences
- Processing of multiple sound sources and reverberation is particularly problematic

References (1)

Bodden (1993), Acta Acustica 1:43-55
Braasch (2002), Acustica/Acta Acustica 88:956-969
Bregman (1990). Auditory scene analysis, MIT Press
Broadbent (1955), Q. J. Expt. Psych. 7:46-47
Broadbent & Ladefoged (1957), JASA 29:708-710
Brown & Wang, Eds. (in press)
Cherry & Taylor (1954), JASA 26:554-559
Colburn (1973), JASA 54:1458-1470
Coleman (1962), JASA 34:345-346
Durlach (1963), JASA 35(8):1206-1218
Faller & Merimaa (2004), JASA 116(5):3075-3089
Gardner (1969), JASA 45(1):47-53
Good & Gilkey (1996), JASA 99:1108-1117
Hawley et al. (1999), JASA 105:3436-3448
Hawley et al. (2004), JASA 116(2):1057-1065
Holt & Thurlow (1969), JASA 46(6):1584-1585
Jeffress (1948), J. Comp. Physiol. Psych. 41:35-39
Kock (1950), JASA 22:801-804
Liu et al. (2000), JASA 108(4):1888-1905
Lyon (1983), Proc. ICASSP 1148-1151

References (2)

Mackensen (2004), PhD thesis
Mershon & Bowers (1979), Perception 8:311-322
Mills (1958), JASA 30:237-246
Mills (1972), in Tobias (Ed.), Foundations of modern auditory theory Vol. II, Academic Press
Moore (1997), An introduction to the psychology of hearing, MIT press
Nielsen (1992), AES Convention, Vienna
Palomaki et al. (2004), Speech Comm. 43(4):361-378
Perrott & Musican (1977), JASA 62:1463-1466
Plenge (1972), Acustica 26:213-221
Roffler & Butler (1967), JASA 46(3):1255-1259
Roman & Wang (2003), Proc. ICASSP V:149-152
Shaw (1974), JASA 39:465-470
Wenzel et al (1993), JASA 94:111-123
Wightman & Kistler (1992), JASA 91(3):1648-1661
Wightman & Kistler (1997), in Gilkey & Anderson (Eds.)., Binaural and spatial hearing in real and virtual environments, Erlbaum