Onsets, autocorrelation functions and spikes for direction based source separation

Leslie Smith and Dagmar Fraser

Department of Computing Science and Mathematics, University of Stirling, Stirling, UK. `Email: {lss, dsf}@cs.stir.ac.uk`

With thanks also to Steve Collins, University of Oxford

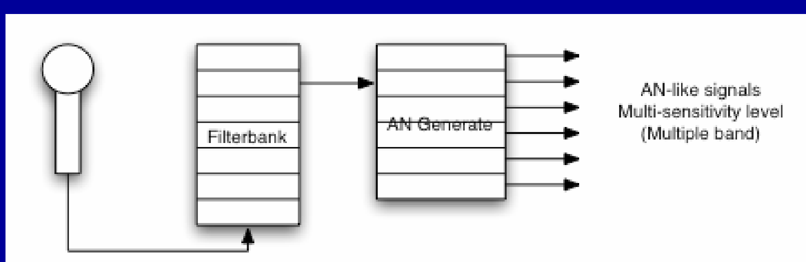
We acknowledge the support of the UK EPSRC for funding

Overview

- Pre-processing
- Onset finding
- Cross-correlation estimation of ITD
- Onset-based estimation of ITD
- Results and comparison

Setup

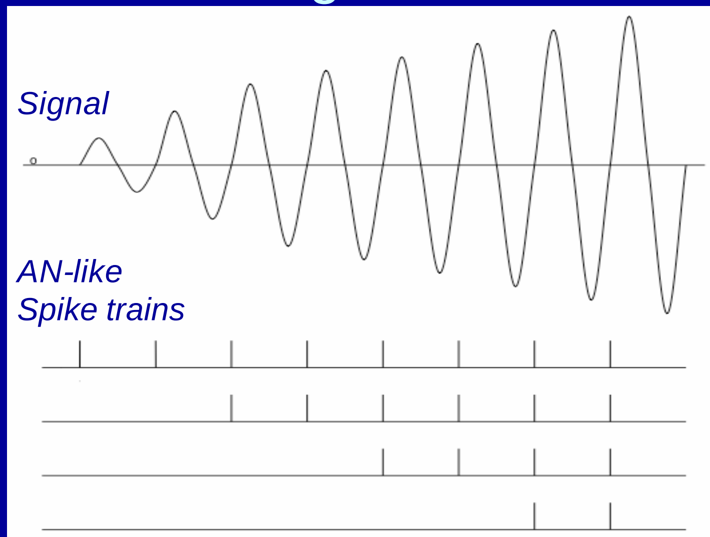
Initial processing



- *Each microphone signal is passed through a filterbank. Phase locked signals are generated by generating a pulse on positive-going zero crossings.*
- *Multiple spike trains per channel are generated.*

  Coding of dynamic range is by predicating spike generation on the pre-spike signal level.

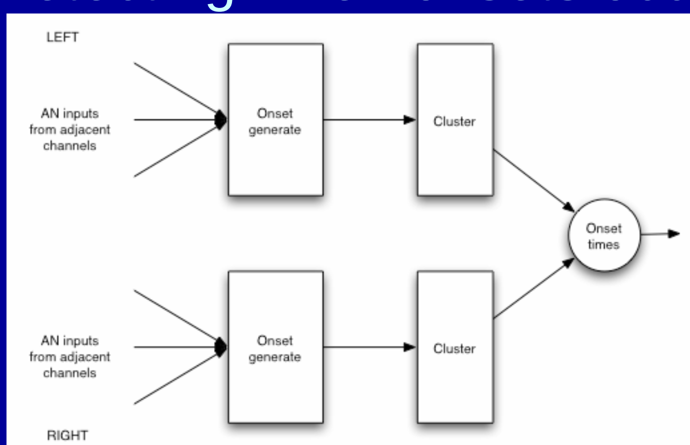
AN-like signals



Most sensitive

Least sensitive

Detecting when onsets occur



AN-like spikes are converged on to leaky integrate-and fire neurons through depressing spikes. These are clustered across time and frequency band to detect the intervals during which onsets occur.

The onsets

- Are detected channel by channel
  - Selectivity depends on number and sharpness of filterbank channels
- Are detected with minimal latency
  - Uses a fast depressing synapse and leaky integrate-and-fire neuron (see IEEE TNNS November 2004)
- Are detected over a wide dynamic range
  - Due to the use of multiple sensitivity AN-like spike trains
  - And clustering them provides an indication of the duration of the onset itself

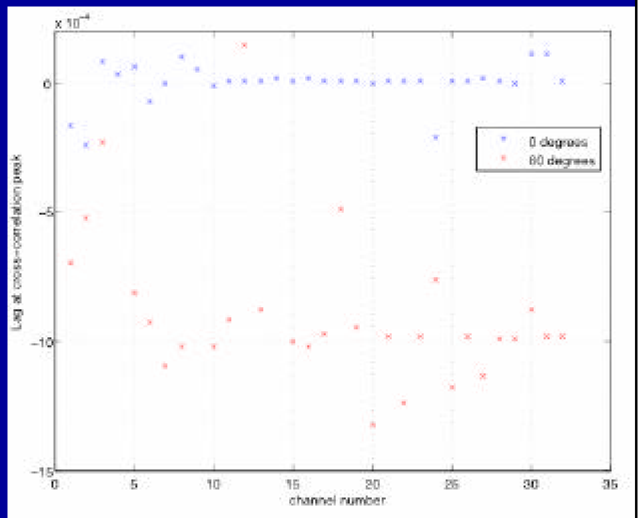
Calculating ITDs during onsets

- Two techniques

1. Cross-correlations: channel by channel during onset interval
2. Use spikes to find signal ITD directly

Cross-correlation ITD estimation

Find peak of each channel's x-correlation then histogram peaks with possible ITDs, and select largest bucket.

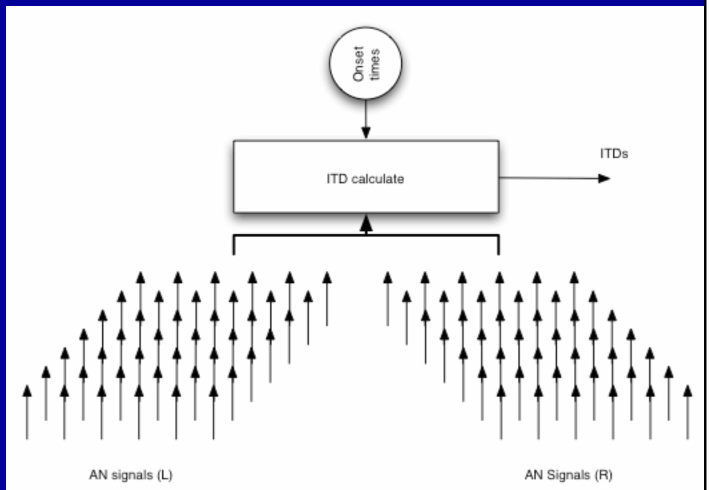


Spike based ITD estimation

- Which spikes to use?
  - Onset spikes from onset interval estimation
    - AN-like signals are converged to give reliable onset detection
    - But this reduces time accuracy because of different delays in different bands
    - Once we have converged the AN signals we cannot adjust for these differential delays.
  - Original AN-like spikes?

Calculating the ITDs using AN-like spikes

ITDs can be calculated from the AN pulses at times determined from the onsets intervals detected.



Why this might not be the best way (1)

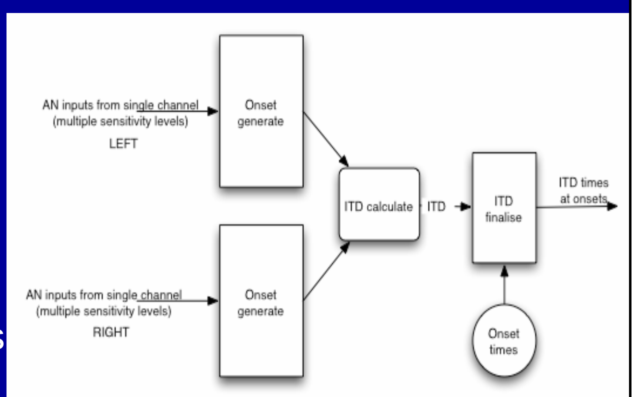
- Need to keep AN pulse times
  - (and there’s a lot of them)
- Each sensitivity and each channel has multiple AN zero-crossing times during an onset interval
  - Difficult to interpret for ITD when bandpass filter period < 2* ITD

Why this might not be the best way (2)

- There's a considerable amount of intensity difference between the signals.
  - particularly from head/ear based microphones but also from microphones on a panel due to different source distances, variation in microphone responses
- Need to combine AN-like pulses from different sensitivities
  - Have tried this! (ASA 2002,, Pittsburgh, 2003, Nashville!)
  - Difficult to get much accuracy

New technique: use two sets of onset spikes

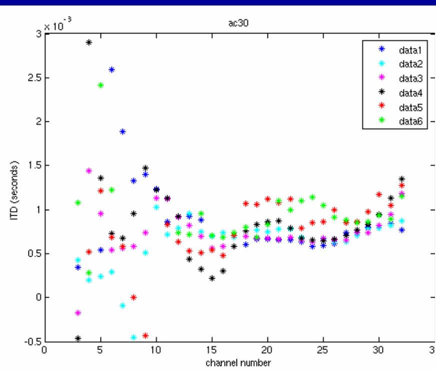
- First set of onset spikes is as before
  - Used to robustly find the onset intervals
- Second set has no convergence of AN-like signals
  - Maintains the accuracy of the timing



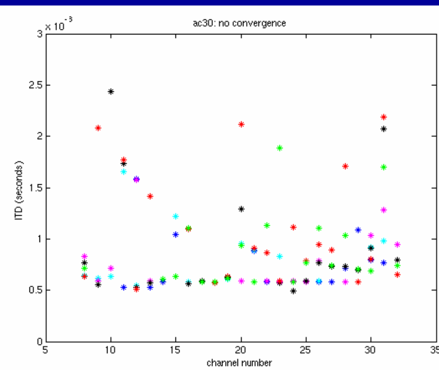
Advantages of 2-onset technique

- There are (far) fewer onsets than AN-like spikes
  - Easier to keep
  - Can use onset spikes from channels where period < 2 * ITD because of sparseness of spikes
  - Can easily estimate ITDs from onsets from different sensitivity levels.

Converged and unconverged onsets (1)



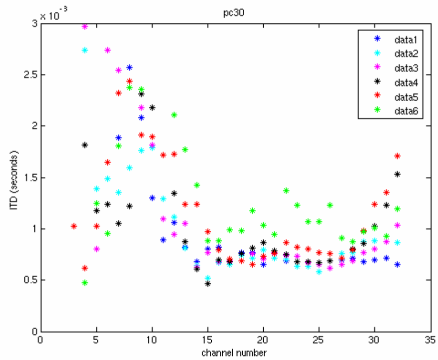
Converged onsets

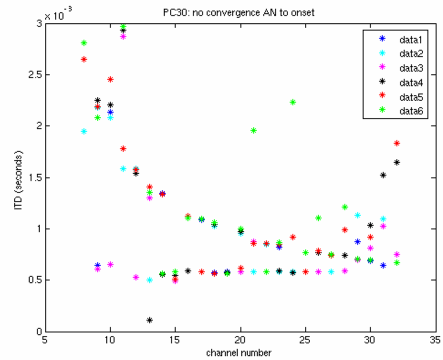


Non-converged onsets

Air-separated microphones, pink noise from 30 degrees. Different colours come from different sensitivity levels.

Converged and unconverged onsets (2)

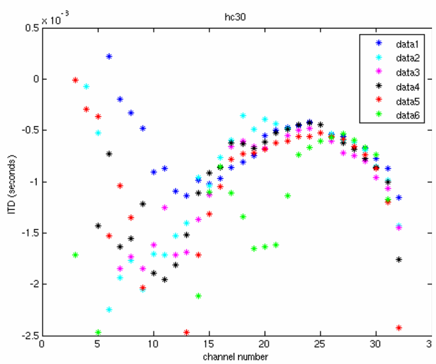


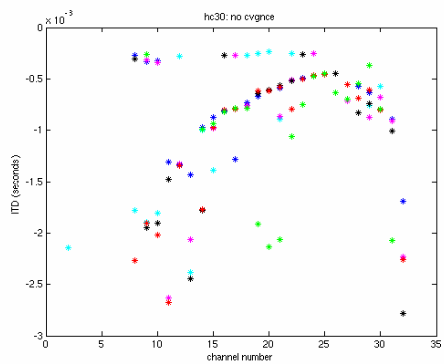


Converged onsets *Non-converged onsets*

Panel separated microphones, pink noise from 30 degrees

Converged and unconverged onsets (3)





Converged onsets *Non-converged onsets*

Head separated microphones, pink noise from 30 degrees

Extracting the ITD from the spikes

- We use the peak of a simple histogram to make an estimate of the ITD
  - Project times onto Y (ITD) axis in (e.g.) 50 μsecond buckets.
  - Estimate ITD as centre of bucket with largest number of ITDs
- We can attempt to optimise this using

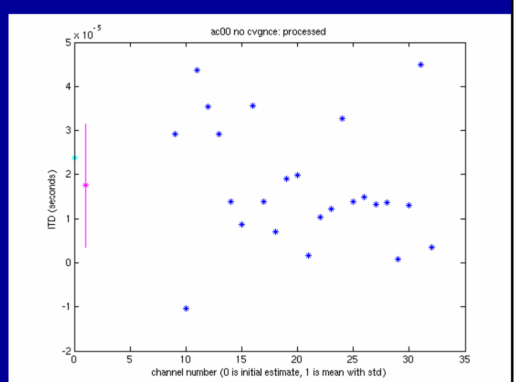
$$\mathrm{ITD}_{\mathrm{estimate2}} = \mathrm{ITD}_{\mathrm{estimate1}} \pm N * 1/f_c \quad \textit{(1)}$$

to improve the estimate
  - But this may be trickier than it looks

Example

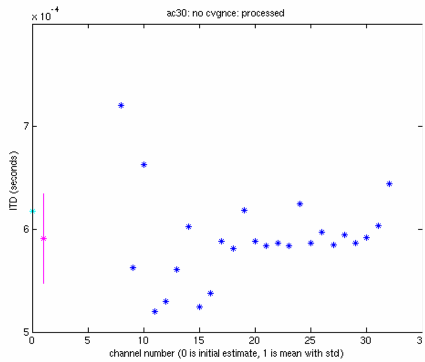
Pale blue dot shows original estimate (bucket size=50mS).

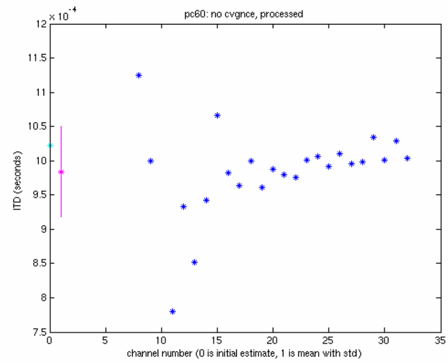
Dark blue dots show final estimate for each channel using a simple one-jump gradient descent technique which minimises the sum squared error of (1) while choosing N.



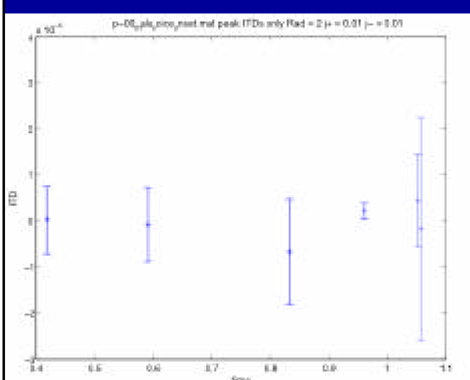
Pink dot is mean of final channel estimates, line is standard deviation.

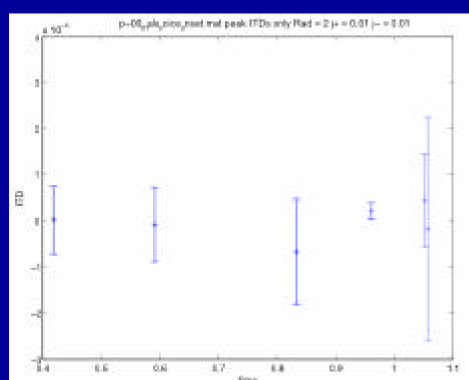
ITD estimate from 2-onset technique





Applying to speech sounds



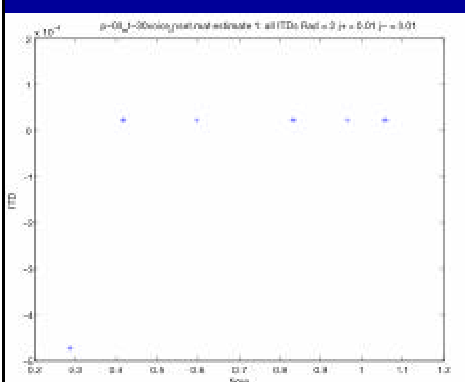


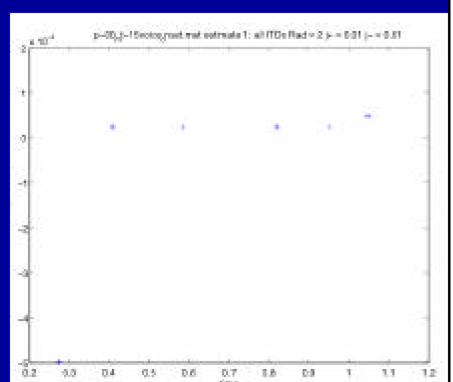
1 second of speech, straight ahead. Left shows all ITDs used at Onset times, right shows only ITDs at peak intensities.

What’s wrong

- The results are disappointing, compared with the results for the pink noise
- Problem: estimating f_c and N in equation 1.
  - Setting f_c to centre of band is ok if energy is equally distributed
    - As in noise
  - Or if signal peaks near f_c
  - Neither is true of speech
- May be better to stick with original estimate!

Results from original estimate

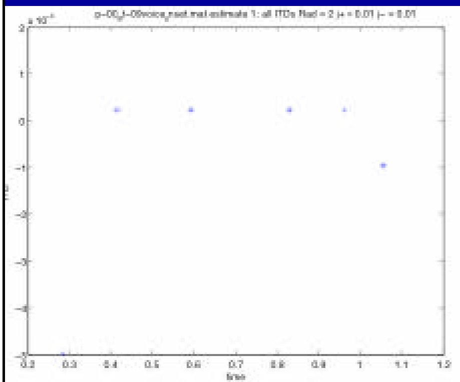


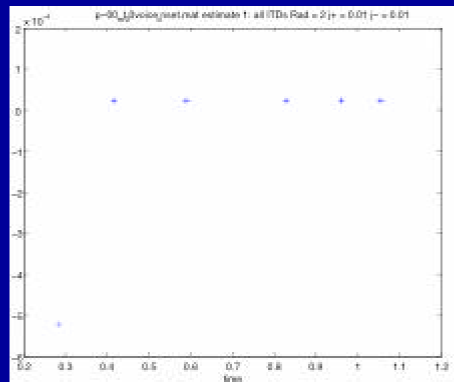


30dBSNR *15dBSNR*

Speech with 1Khz tone in background. Speech is at 0 degrees, Noise is at 30 degrees

Results from original estimate cont'd



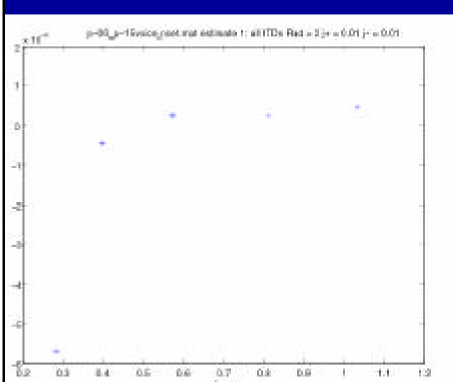


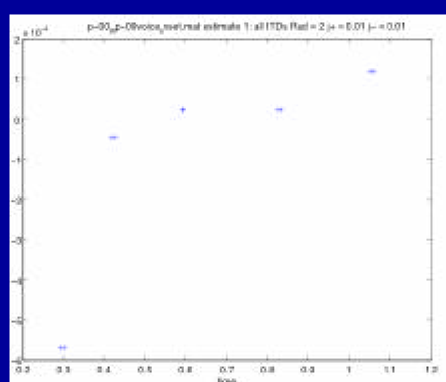
9dBSNR

0dBSNR

Speech with 1Khz tone in background

Results from original estimate cont'd



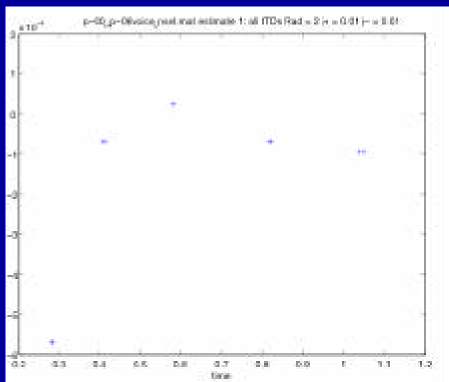


15dB SNR

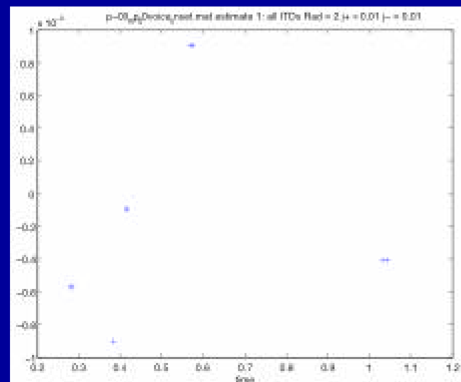
9dBSNR

Speech with pink noise in background

Results from original estimate cont'd



6dB SNR

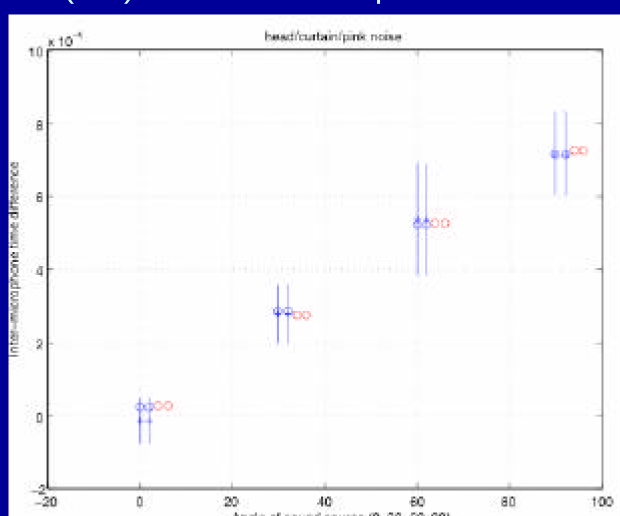


0dBSNR

Speech with pink noise in background

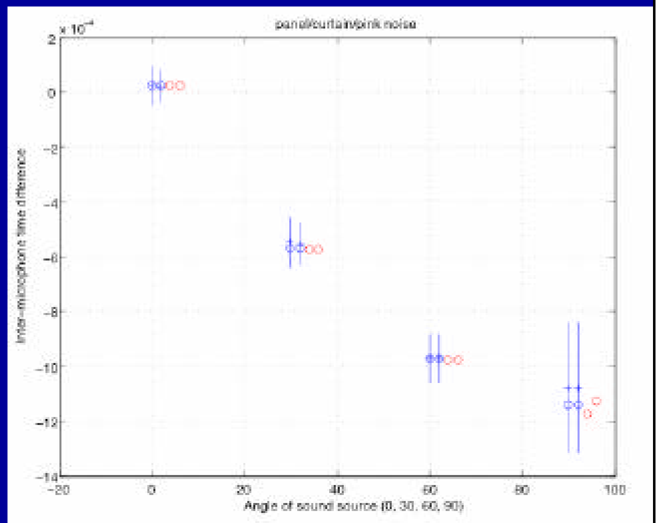
Comparing cross-correlation and spike based ITD estimation: head (ear) mounted microphones

- Sound is short pink noise pulse
- Red circle is cross-correlation ITD estimate,
- Blue circle is 1st estimate for onset based ITD calculation, blue cross and line is the adjusted estimate

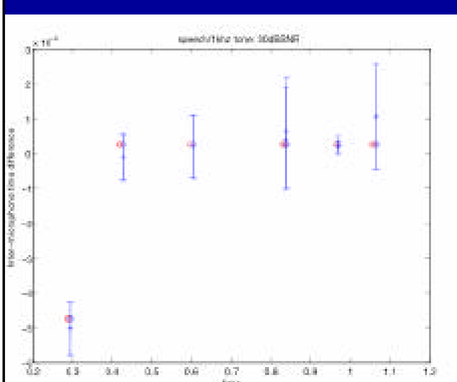


Comparing cross-correlation and spike based ITD estimation: panel mounted microphones

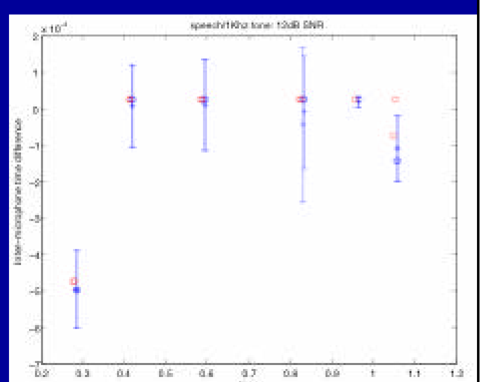
- Sound is short pink noise pulse
- Red circle is cross-correlation ITD estimate,
- Blue circle is 1st estimate for onset based ITD calculation, blue cross and line is the adjusted estimate



Comparison cont'd: speech in sine wave noise.



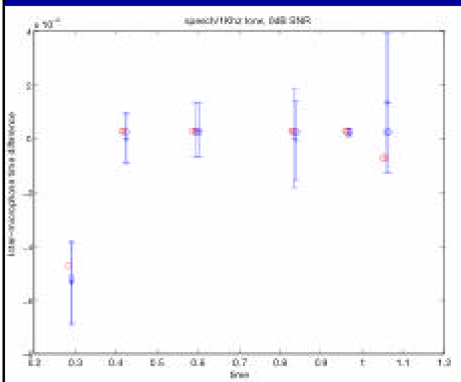
30 dB SNR

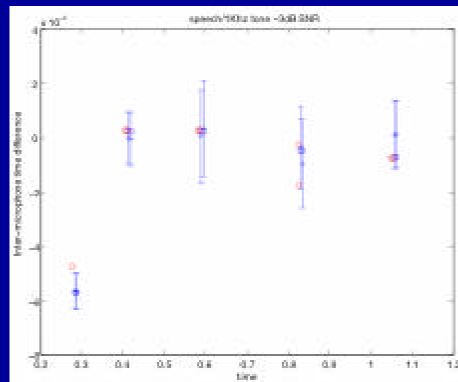


12 dB SNR

Speech (0 degrees) with 1Khz background noise (30 degrees) Microphones panel mounted.

Comparison cont’d: speech in sine wave noise.

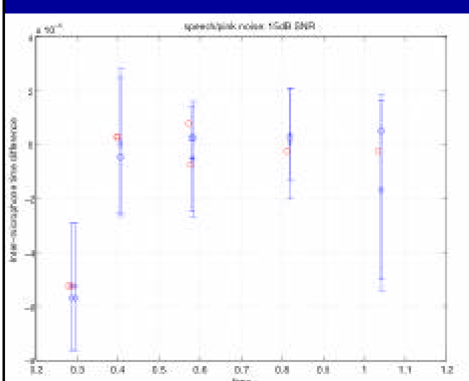


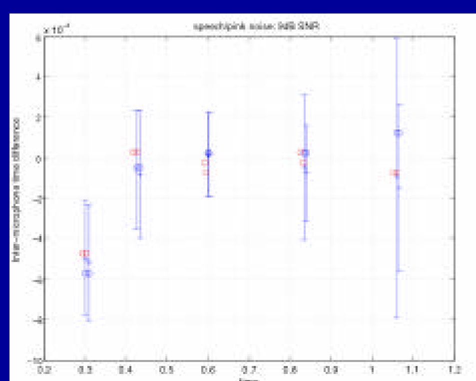


0 dB SNR *-3 dB SNR*

Speech (0 degrees) with 1Khz background noise (30 degrees)

Comparison cont’d: speech in pink noise.

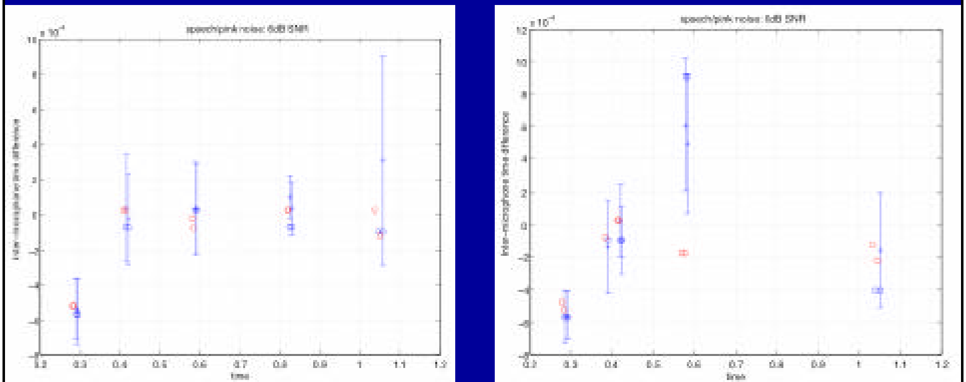




15dB SNR *9dB SNR*

Speech (0 degrees) with pink noise at 30 degrees

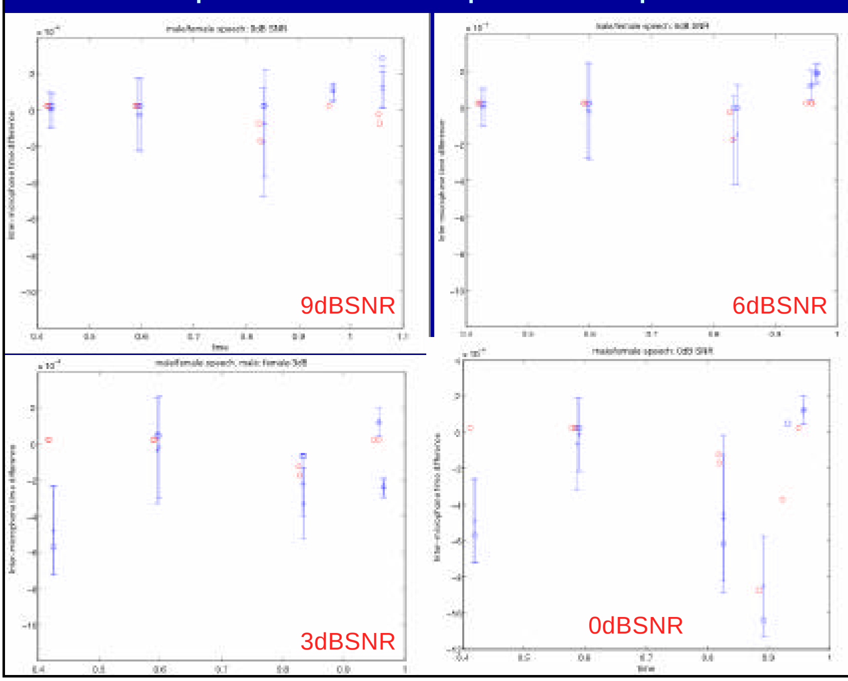
Comparison cont'd: speech in pink noise.



6dB SNR *0dB SNR*

Speech (0 degrees) with pink noise at 30 degrees

Comparison cont'd: speech in speech noise.



Discussion (1)

Which method of ITD estimation is better?

- Cross-correlation uses the ‘whole” signal, but the onset system uses just the zero-crossing
  - But for many sounds there is virtually no difference.
- Both are relatively accurate, compared to non-onset based methods when there is more than one sound source
  - Cross-corrrelation seems marginally better in high noise levels: but difference is very small.
- Onset based technique may be less computationally intensive: but this depends on the implementation technology.
  - Refinement of the histogram-based estimate seems not to help

Discussion (2)

What are the sources of error?

- Initial sound quantisation (96 KS/sec)
  - About 10μseconds
- Histogramming quantisation (50 μsecond buckets)
- Effect of IID on bandpassed signal
  - Large signals in adjacent bands which are of different sizes at each microphone affect bandpassed signal phase, moving of the zero-crossing time
- Onsets from different sources which occur at almost the same time will interfere with each other
  - We rely on actual onsets being sparse.

Discussion (3)

- Ways forward
  - How to join onsets across time?
    - By location? But people can separate speech/music from a mono radio
    - By the signal characteristic at onset? E.g. by estimating vocal tract length for speech, or the characteristic of the attack for musical instruments
  - Can we use the characteristics of the onsetting of the signal to help identify phonemes?
    - See poster 4ASCx
  - How about resynthesising sounds/speech just from the onsets?
    - Working on this!

End of talk

- Thank you for your attention.