On the Window-Disjoint-Orthogonality of Speech Sources in Reverberant Humanoid Scenarios

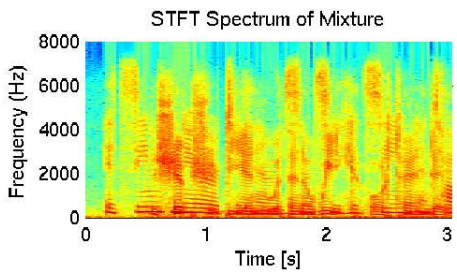
Sylvia Kümmel and Thorsten Herfet

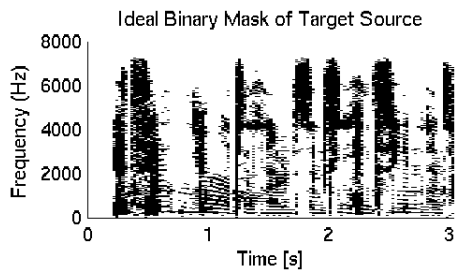
Telecommunications Lab, Saarland University, Germany

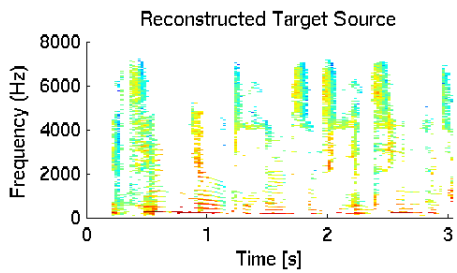
September 3, 2008

Introduction

- State-of-the-art Speech Source Separation algorithms are based on orthogonality of speech sources in the time-frequency (TF) domain
  - Speech signals are sparsely distributed in high-resolution TF representations
  - TF-spectra of different speech sources overlap only in few points → approximate orthogonality
  - TF-masks emphasize regions that are dominated by the target source and attenuate regions dominated by interfering sources







Introduction

- Orthogonality of speech sources in TF-domain has been investigated in detail for anechoic speech mixtures.
- Many practical applications require real reverberant conditions.
- Are source separation architectures based on the TF-orthogonality appropriate also in real world scenarios?
  - How do anechoic conditions influence the orthogonality?
  - How does a humanoid setup influence the orthogonality?
  - Is the ideal binary mask a valid final goal also in reverberant and humanoid setups?

Measuring Orthogonality (1)

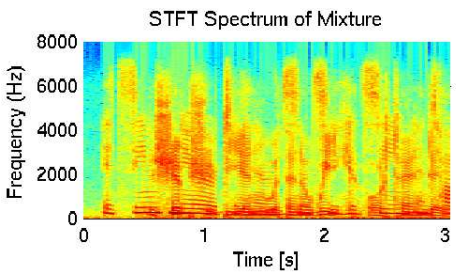
- Assume s_i is STFT spectrum of speech source i

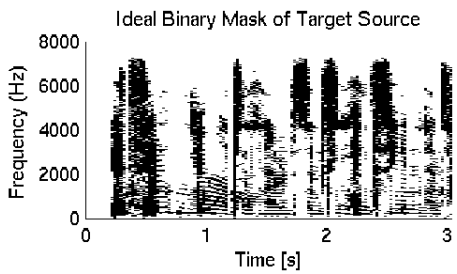
$$X(k,q) = \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} w_a(n) x(n+k) e^{-i2\pi \frac{qn}{N}}$$

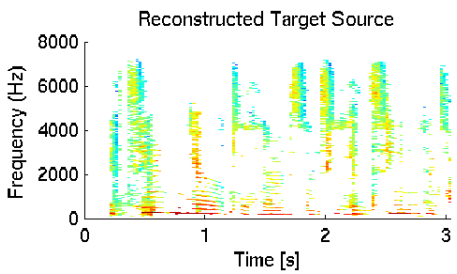
STFT paramters: samplingrate 44.1 kHz, window length 1024 samples, overlap 512 samples

- Ideal binary mask for target source s_i and interfering sources n_j

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x\ dB\ \forall j \\ 0 & \text{else} \end{cases}$$







Measuring Orthogonality (2)

- Preserved Signal Ratio (PSR)
  - How well does ideal mask preserve energy of target source?

$$PSR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||s_i(t,f)||^2}$$

- Signal to Interference Ratio (SIR)
  - How well does ideal mask suppress interfering sources?

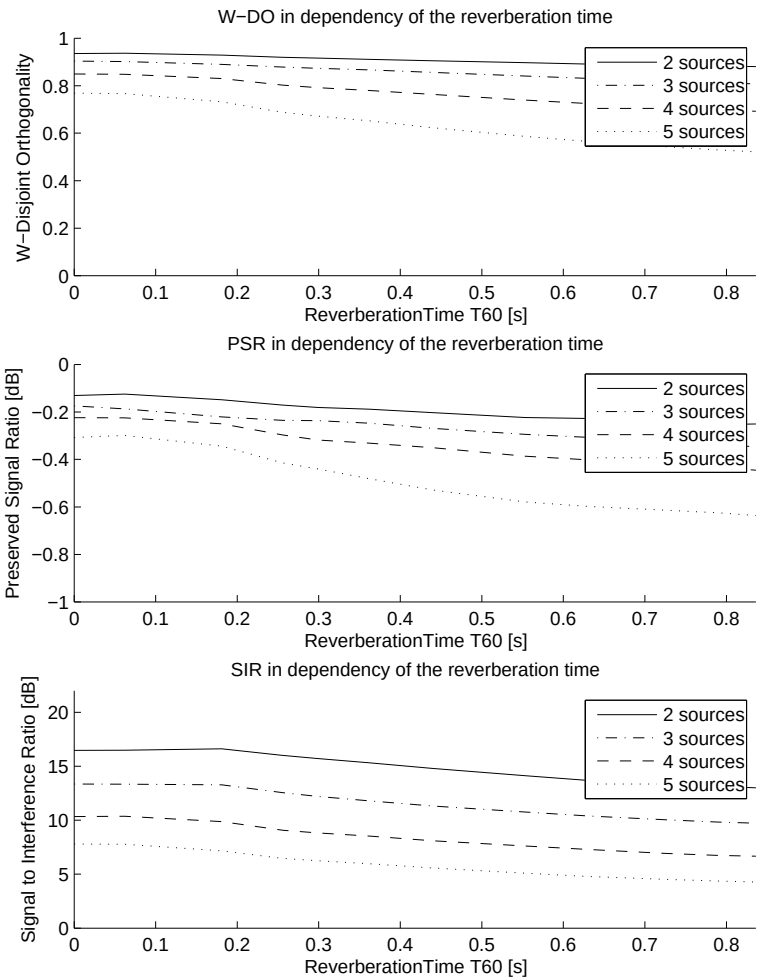
$$SIR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||\Omega_i(t,f)\sum_{j\neq i} s_j(t,f)||^2}$$

- Window-Disjoint Orthogonality (WDO)
  - Combined and normalized measurement of PSR and SIR

$$WDO = PSR - PSR/SIR$$

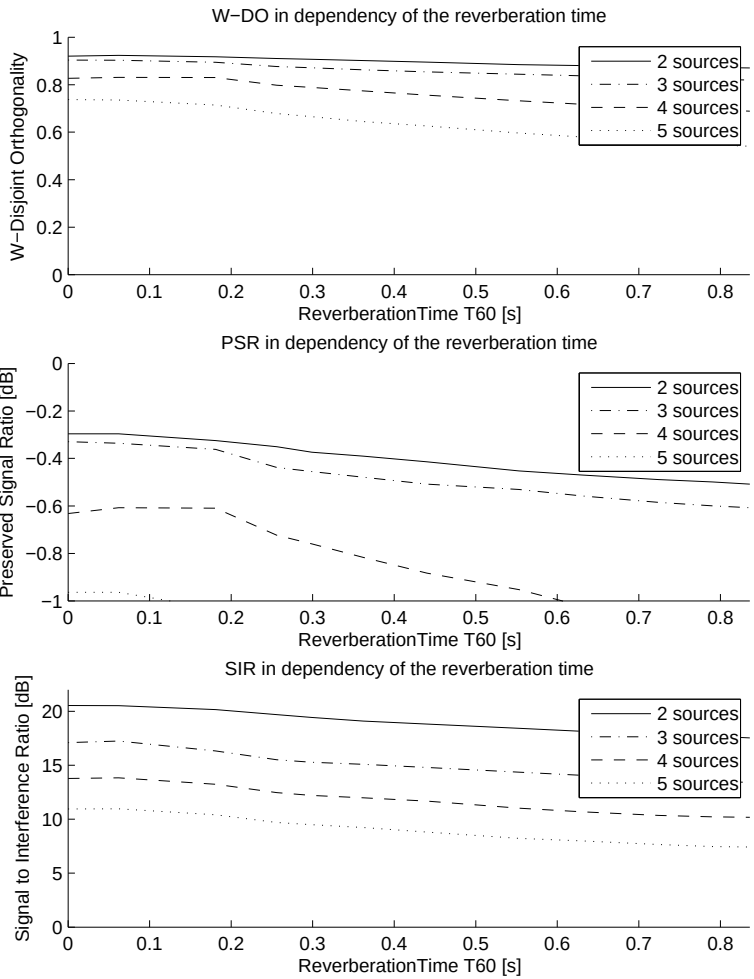
  - WDO = 1 $\rightarrow$ perfect orthogonality
  - WDO $\rightarrow -\infty \rightarrow$ no orthogonality

WDO under reverberant conditions



- Simulated room impulses with specified T_{60} and known ground truth signals
- WDO decreases
- Room impulses smear energy in time and frequency
  $\rightarrow$ sparseness decreases
- SIR decrease at $T_{60} = 0.4$s $\approx$3dB
- Low SIR gains (17 dB (2 sources) to 8 dB (5 sources))
  - Sources exhibit approximate orthogonality
  - But also overlap in many parts

How to enhance low SIR? Divide Separation Process



- Estimate coarse binary masks with bins that exhibit large orthogonality
  → high SIR
- Use ideal mask with higher threshold

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x\ dB\ \forall j \\ 0 & \text{else} \end{cases}$$

- For 6 dB mask, SIR increase of 5 dB
  For 9 dB mask, SIR increase of 8 dB
  → few interfering energy
- But decrease of PSR
  → low speech quality
  → Refill mask: Keep SIR, incr PSR

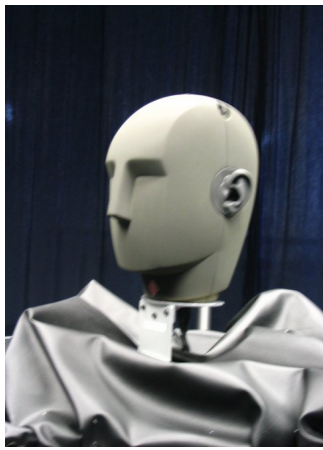
How to enhance low SIR?

Enhanced Separation Process

1. Estimate coarse binary masks with bins that exhibit large orthogonality → high SIR
   - Unique spatial position
   - Harmonic analysis
2. Refill masks by cognitive models (keep SIR, increase PSR)
   - Refill missing harmonics
   - Consistent On/Offset over time/frequency
   - Step 1+2 → ideal binary mask in optimal case
   - Not necessarily best speech quality
3. Postprocessing algorithms
   - Eliminate sharp peaks in spectrum
   - Use models of human speech production to shape spectrum

WDO under humanoid conditions

- Humanoid Conditions
  - Pinnae and outer ear structures filter signals
  - HRTF affects TF-spectrum → affects orthogonality
  - Sources are spatially separated in auditory scene
  - Two ears available → which ear is better?

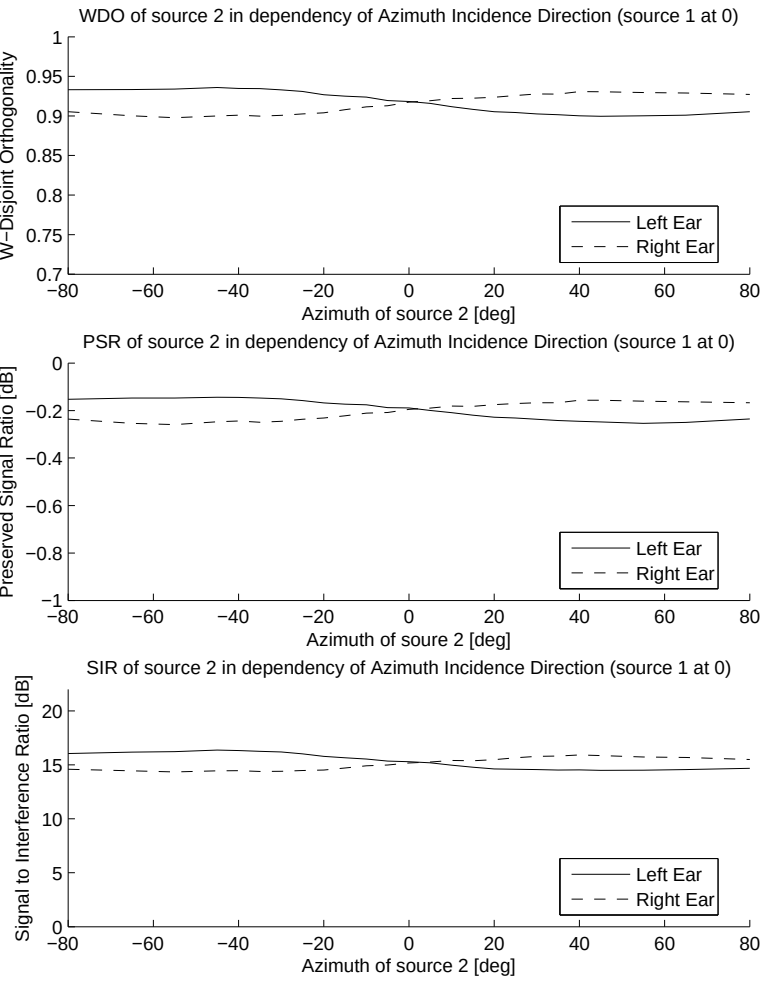


- How does orthogonality change under humanoid conditions?
  - SIR equal to anechoic case for sources with large incidence angle difference ($> 50°$) if ear closer to source is chosen
  - SIR for nearby sources (difference $< 10°$) drops by approximately 3 dB
  - Max. SIR is 17 dB → 3dB decrease influences speech quality

WDO under humanoid conditions

Src 1 is fixed at $0°$,

Src 2 moves from $-80°$ to $80°$

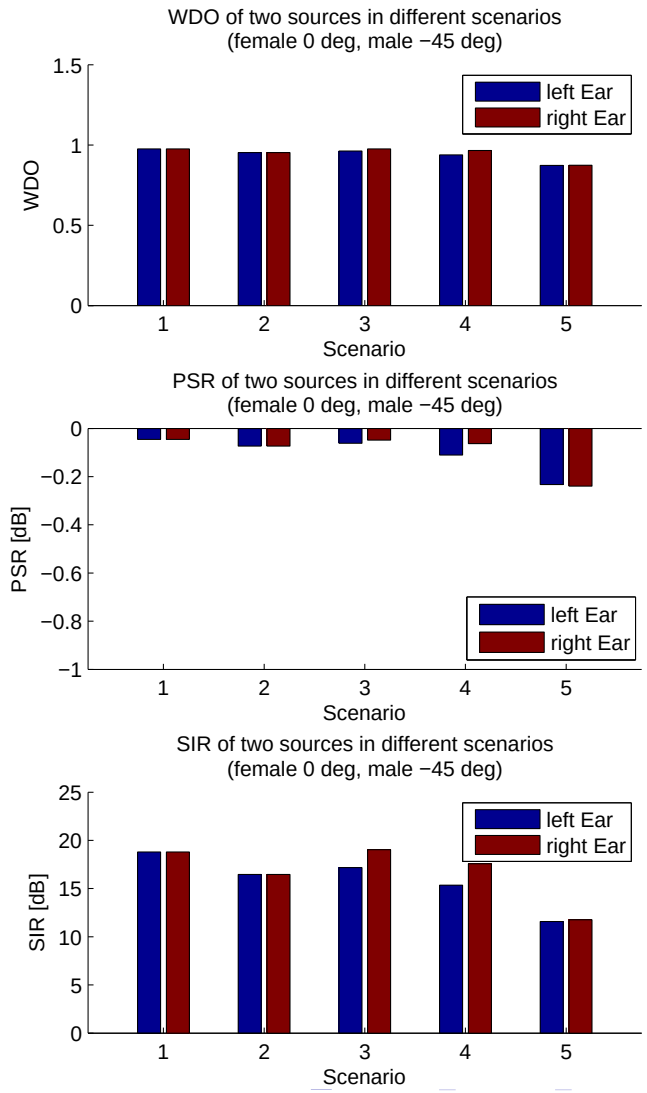


- Best values are obtained if sources are far away from each other
  - For right ear WDO is best if src 2 is positioned in the right hemisphere
  - source is nearer to right ear $\rightarrow$ src 2 is louder than src 1
  - src 1 is attenuated by head shadow
- If source positions are known
  - Automatically choose better ear
  - Use ear with higher expected SIR for demixing
  - Move head to optimal position

WDO under real reverberant humanoid conditions

- S1 Anechoic case (0-dB mask)
- S2 Simulated reverberant case $T_{60} = 0.4\ s$ (0-dB mask)
- S3 Simulated HRTF filtering (0-dB mask)
- S4 Simulated reverberant HRTF filtering (0-dB mask)
- S5 Real recordings of a human dummy head in a normal office room with $T_{60} = 0.4\ s$ (0-dB mask).

- WDO of real recordings lower than in simulated cases
  - SIR about 5 dB lower than anechoic



Conclusions

- Speech signals are not completely orthogonal
  - Echoes and HRTF filtering introduce further overlap between different speech sources in the STFT domain.
  - Overall the SIR is decreased by up to 6 dB.
- Restrict mask estimation to spectral parts that exhibit high orthogonality
  - High SIR (more than 9 dB SIR gain)
  - Low PSR (to low for human or automatic listener)
  - Refill mask by appropriate cognitive model (increase PSR)
  - Ideal adaption to receiver (human or ASR)