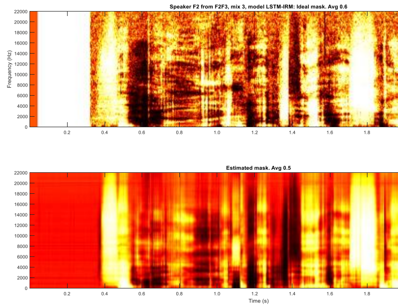
Gaurav Naithani
Tuomas Virtanen

Danish Sound Cluster 20221123

Deep neural networks for speaker separation for hearing impaired listeners

- but also some noise reduction

Lars Bramsløw



Speaker F2 from F2F3, mix 3, model LSTM-IRM: Ideal mask. Avg 0.6

Estimated mask. Avg 0.5

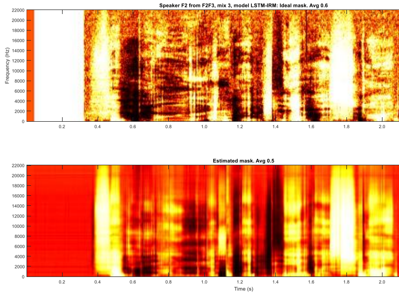
TAMPERE UNIVERSITY OF TECHNOLOGY

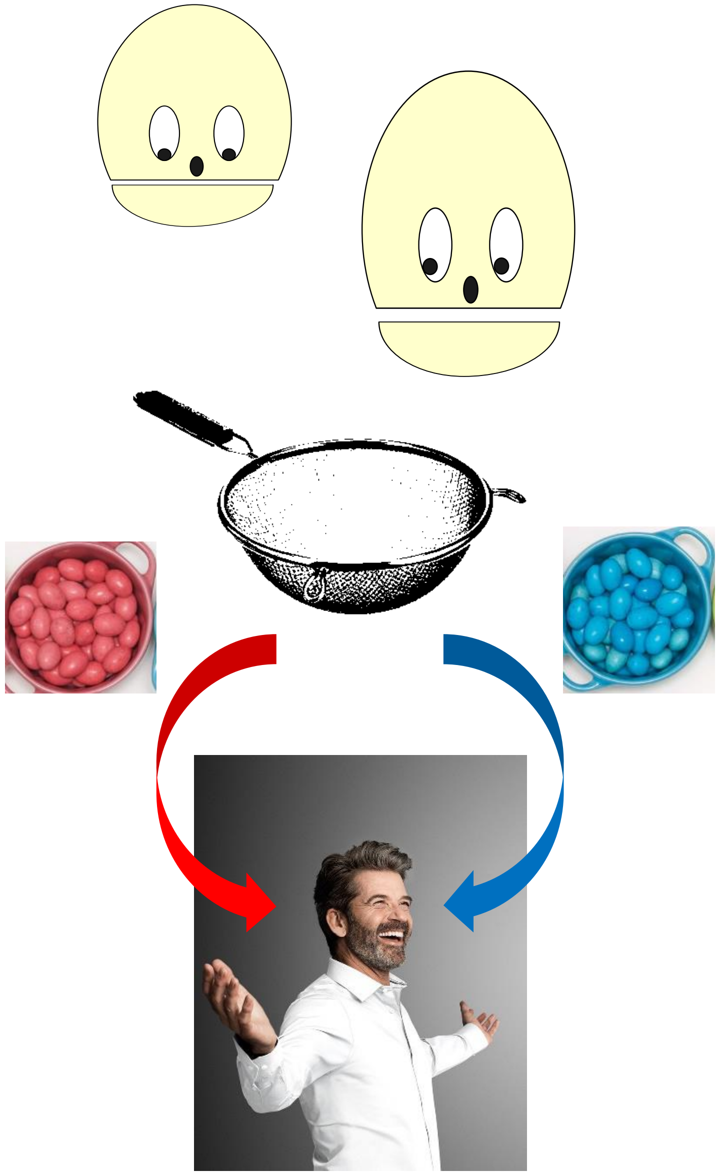
EriksholmResearchCentre
PART OF OTICON

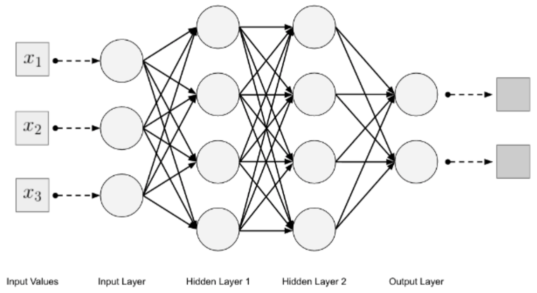
Overview

- Speaker separation: two competing voices
- Noise reduction: voice-in-noise
- Summary

Speaker separation (voice-on-voice)

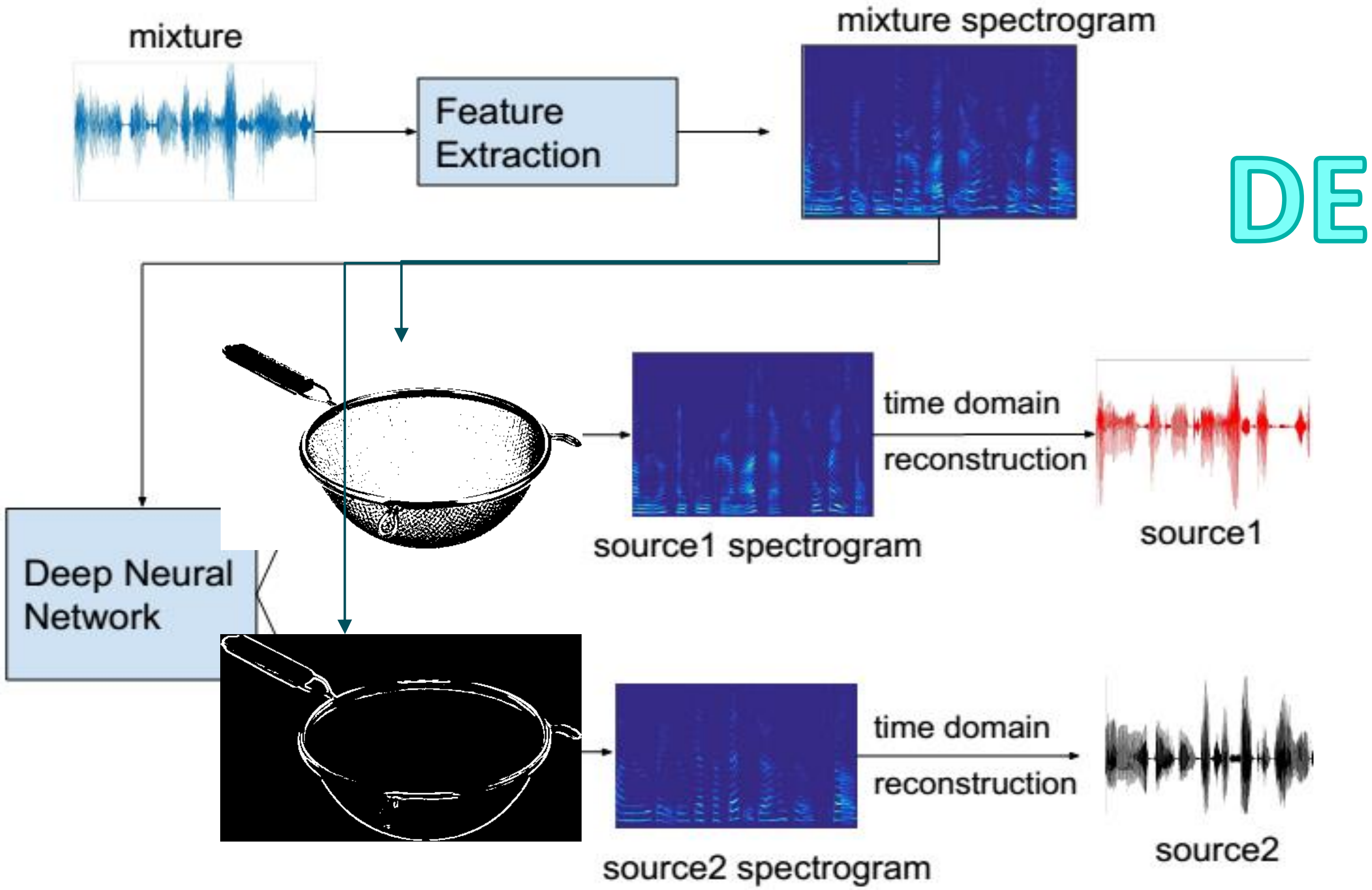






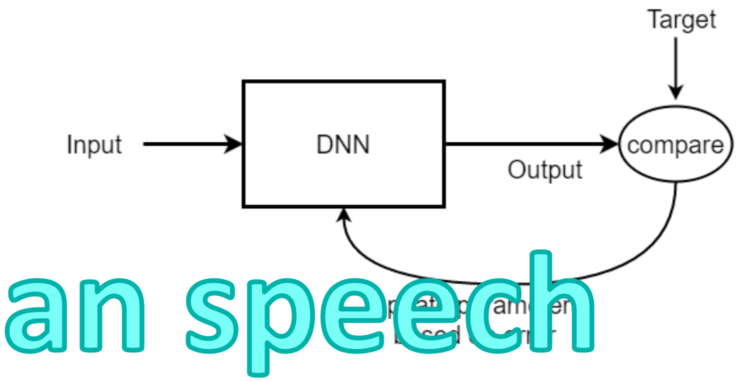
x_1
x_2
x_3
Input Values
Input Layer
Hidden Layer 1
Hidden Layer 2
Output Layer

Separating two voices with low delay

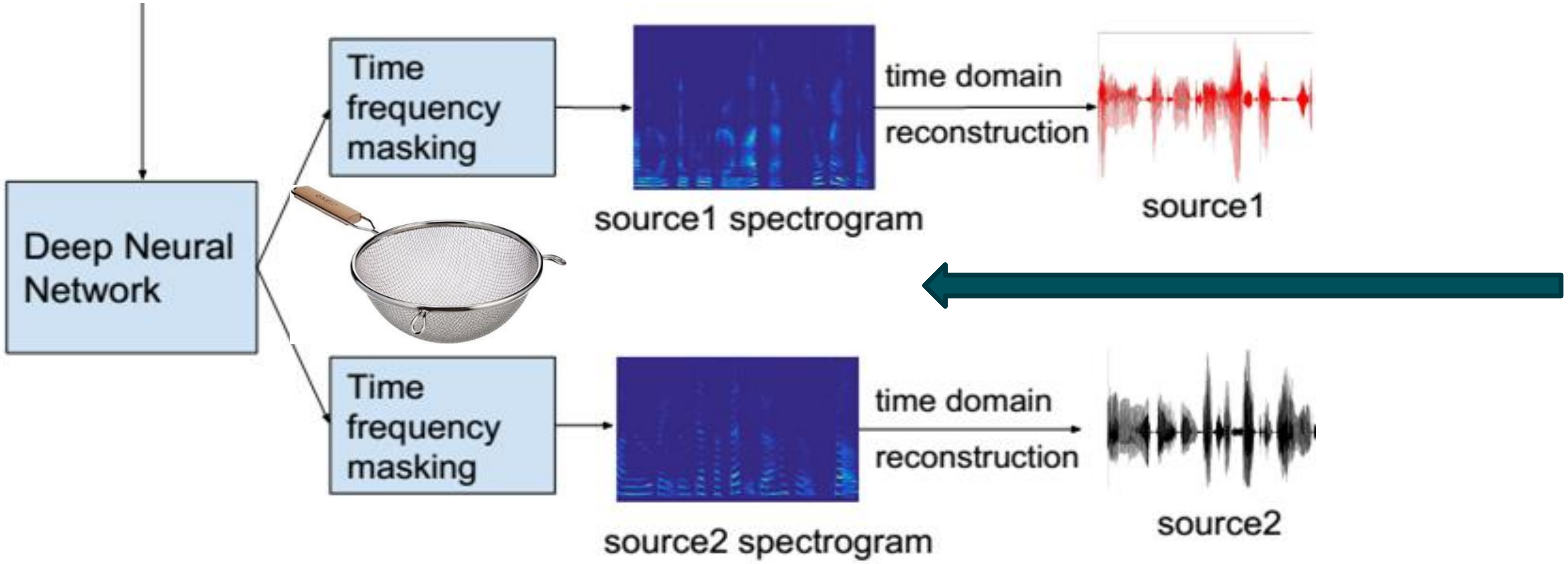


DELAY 8 ms

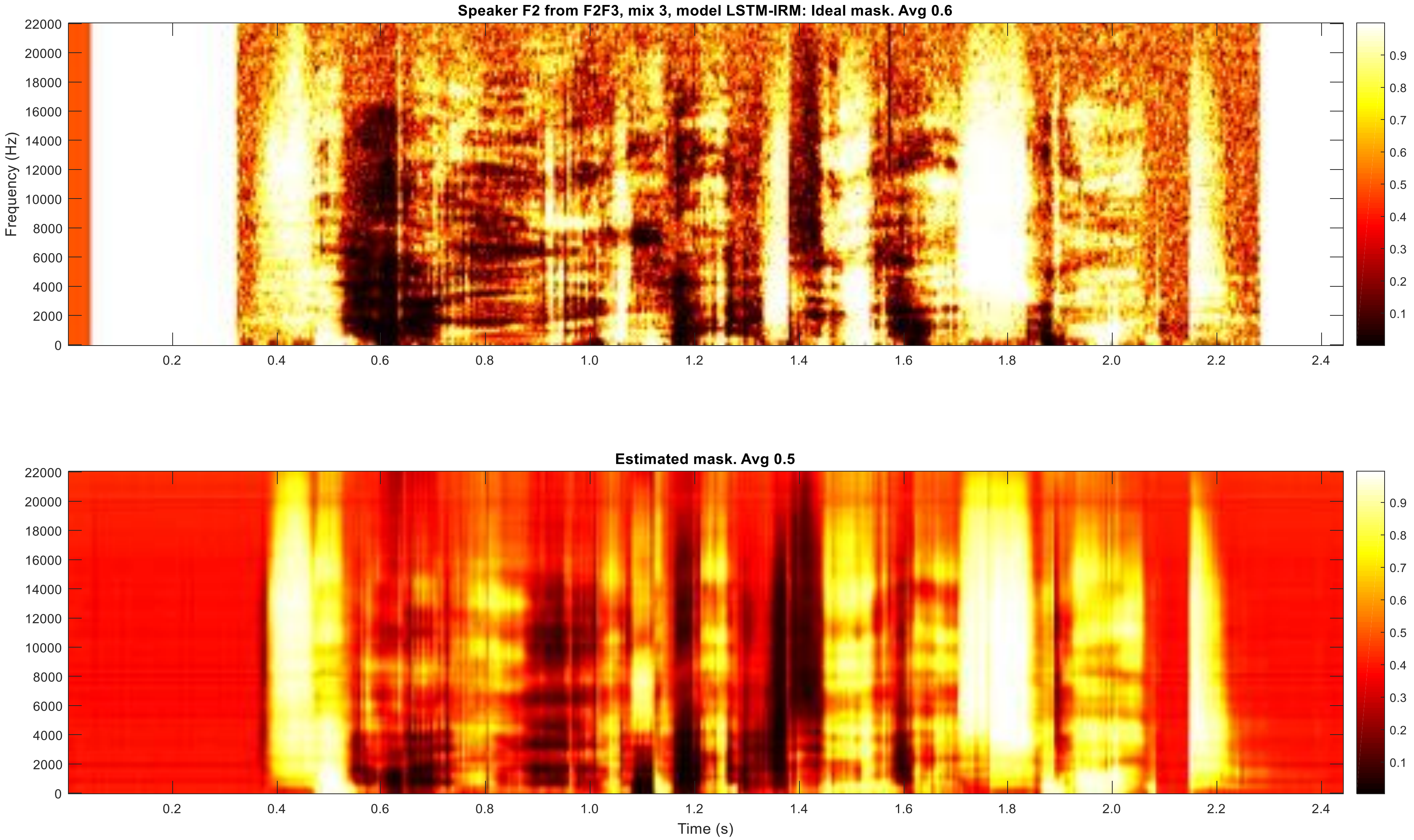
Training the DNN with the truth



3 min of clean speech



Ideal and estimated ratio mask



Competing voices separated

Derhjemme spiser vi ikke kød

Statuen har ikke noget hoved

- Example: pairs of sentences from the Danish Hearing In Noise Test
- Voices known in training
- Lots of glimpsing possible

M1 + M2

~M1

~M2

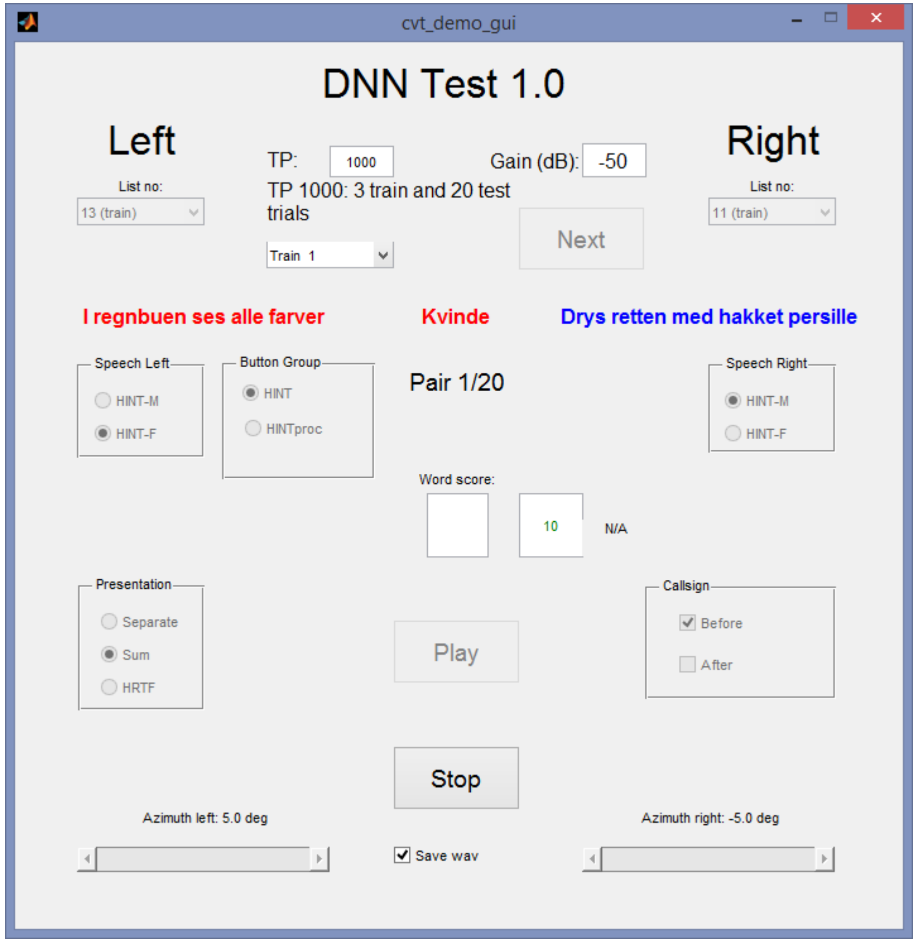
Danish HINT material

- Overall: 13 lists of 20 sentences each
- Talkers: 3 male, 3 female (originally 1 male)
  - Combined in male-male, female-female and male-female pairs.
  - The DNN is speaker-specific, trained per speaker pair
- DNN validation: 1 list
- Training material: 4 lists
- Listening test: 8 lists

Speech intelligibility benefit: Competing voices test

Statuen har ikke noget hoved

Derhjemme spiser vi ikke kød



cvt_demo_gui

DNN Test 1.0

Left

List no: 13 (train)

TP: 1000

Gain (dB): -50

TP 1000: 3 train and 20 test trials

Train 1

Next

Right

List no: 11 (train)

I regnbuen ses alle farver

Kvinde

Drys retten med hakket persille

Speech Left
○ HINT-M
● HINT-F

Button Group
● HINT
○ HINTproc

Pair 1/20

Speech Right
● HINT-M
○ HINT-F

Word score: 10 N/A

Presentation
○ Separate
● Sum
○ HRTF

Play

Callsign
☑ Before
□ After

Stop

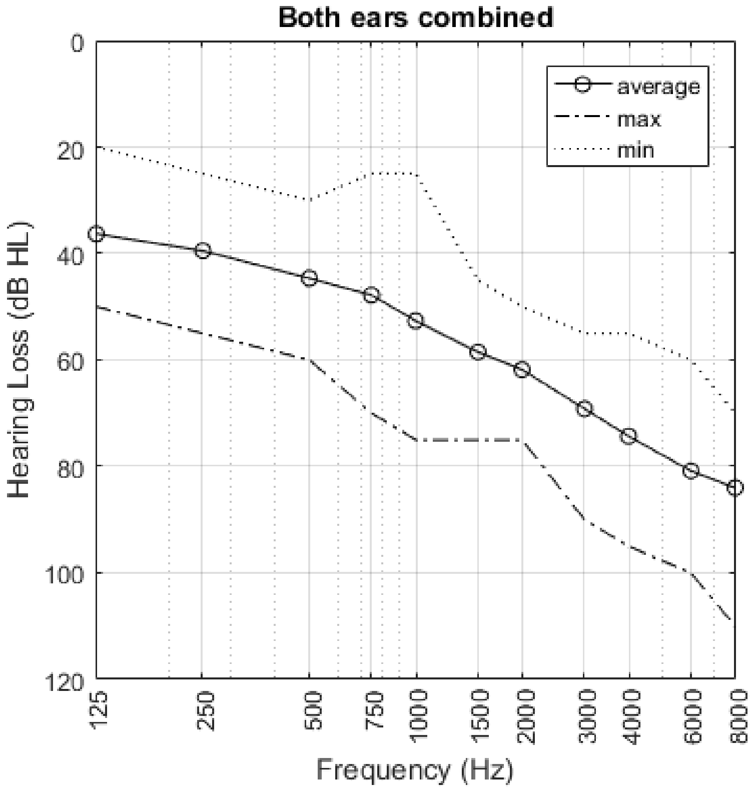
Azimuth left: 5.0 deg

☑ Save wav

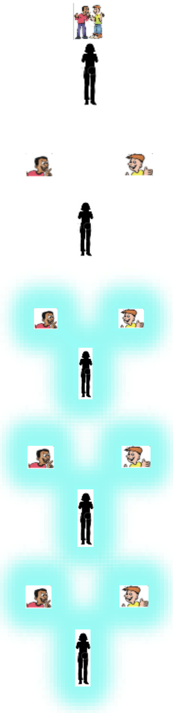
Azimuth right: -5.0 deg

- Pairs of sentences from the Danish HINT (Hearing In Noise Test).
- Cueing by
  - First word = Single target
  - Last word = Dual targets: Competing Voices (CVT)
- Hearing loss compensated individually (NAL-R)

15 hearing-impaired listeners



Processing

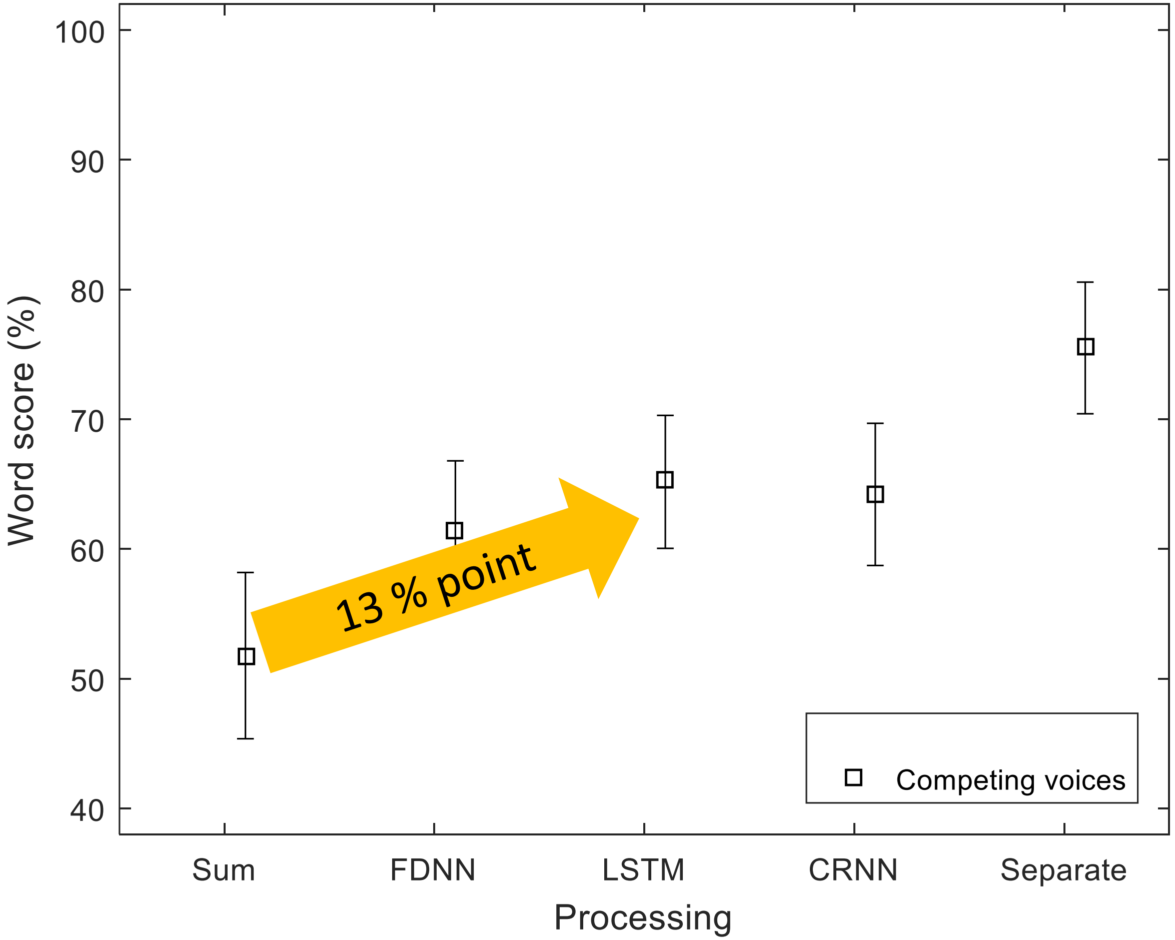


1. Sum (unprocessed)
2. Separate (ideal)
3. Feed Forward DNN (FDNN)
4. Long-Term Short-Term Memory Neural Net (LSTM)
5. Convolutional Recurrent Neural Net (CRNN)

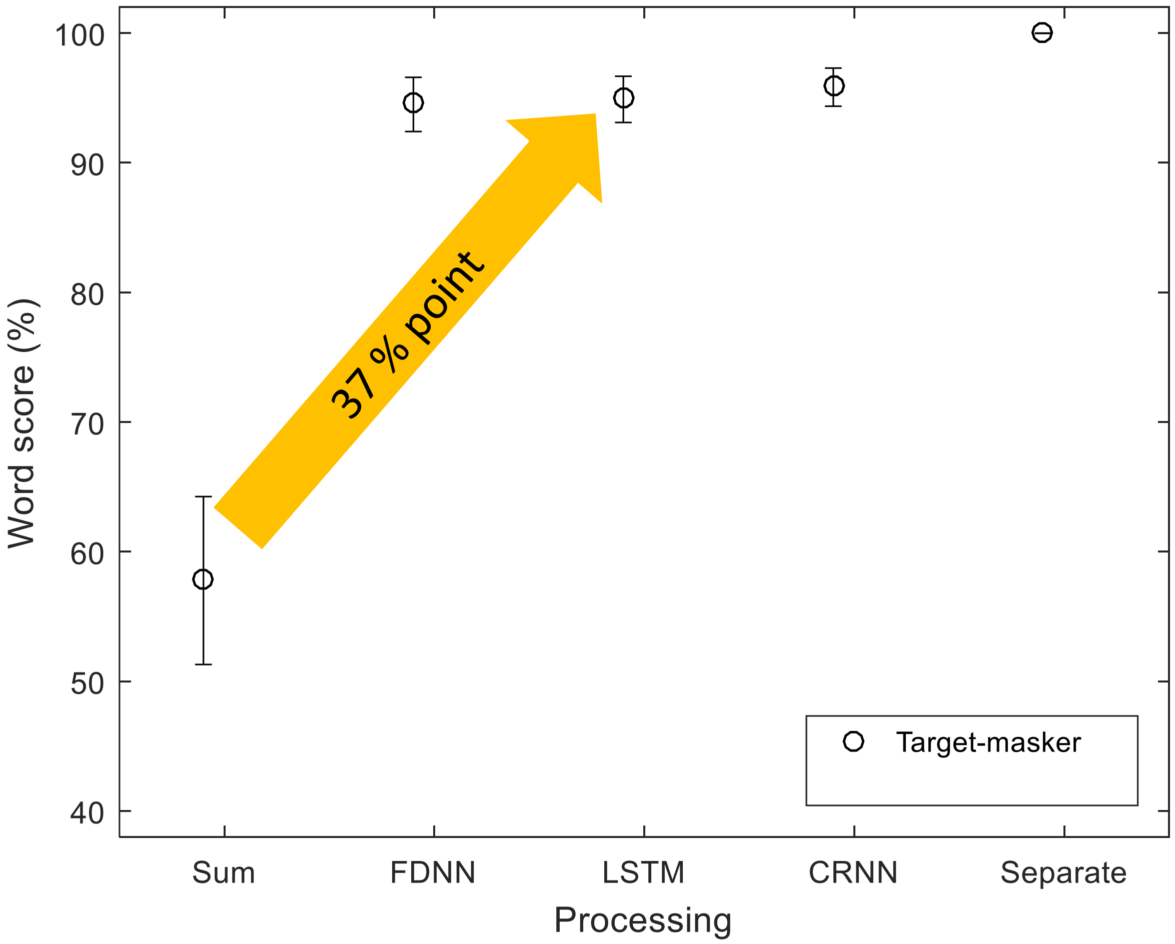
Roughly 3.5 mio weights

Naithani et al, CHAT 2017, Stockholm
Naithani et al, WASPAA 2017, Mohonk

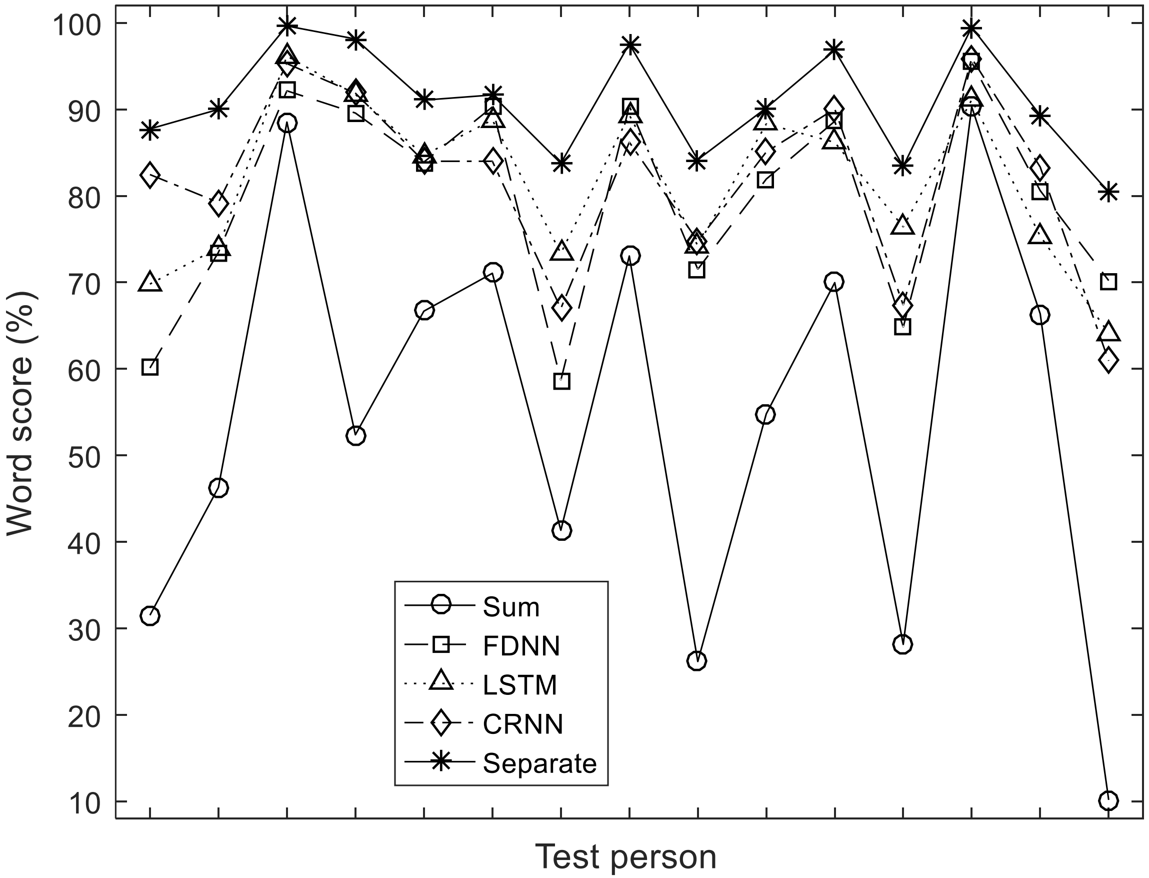
Speech segregation results (competing voices)



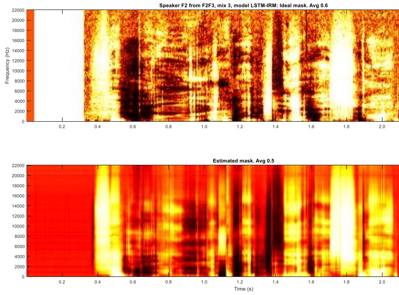
Speech separation results (single target)



Different benefit for different people



Noise reduction



Voice in noise

- Known and unknown voices in known noise
  - more common scenario
  - evaluate generalization ability
- New DNN+mask candidates

Named DNN conditions

1. Sum (= input)
2. FDNN known voice
3. LSTM known voice
4. LSTM unknown voice
5. LSTM unknown voice + multi resolution mask
6. LSTM unknown voice + phase sensitive mask
7. Ideal ratio mask

Maximum 20 dB attenuation (except 7.)

Roughly 3.5 mio weights

Test stimuli

- Danish HINT sentences
  - M1-M6, F1-F6 (12 talkers)
  - 200 – 260 sentences ~ 6 min

- Target talkers:
  - M1, M2, F1, F3
  - Speaker dependent: train on these (test other sentences)
  - Speaker independent: do not train on these (test all sentences)

- Noise from the 'ICRA natural sound library'
  - P1: Party noise
    - train at -3..+3 dB SNR, test at +0 dB
  - S1: Shopping center noise
    - train at -3..+3 dB, test at +0 dB.

https://icra-audiology.org/

Voice-on-noise test

Statuen har ikke noget hoved

- Sentences from the Danish HINT
- 0 dB SNR

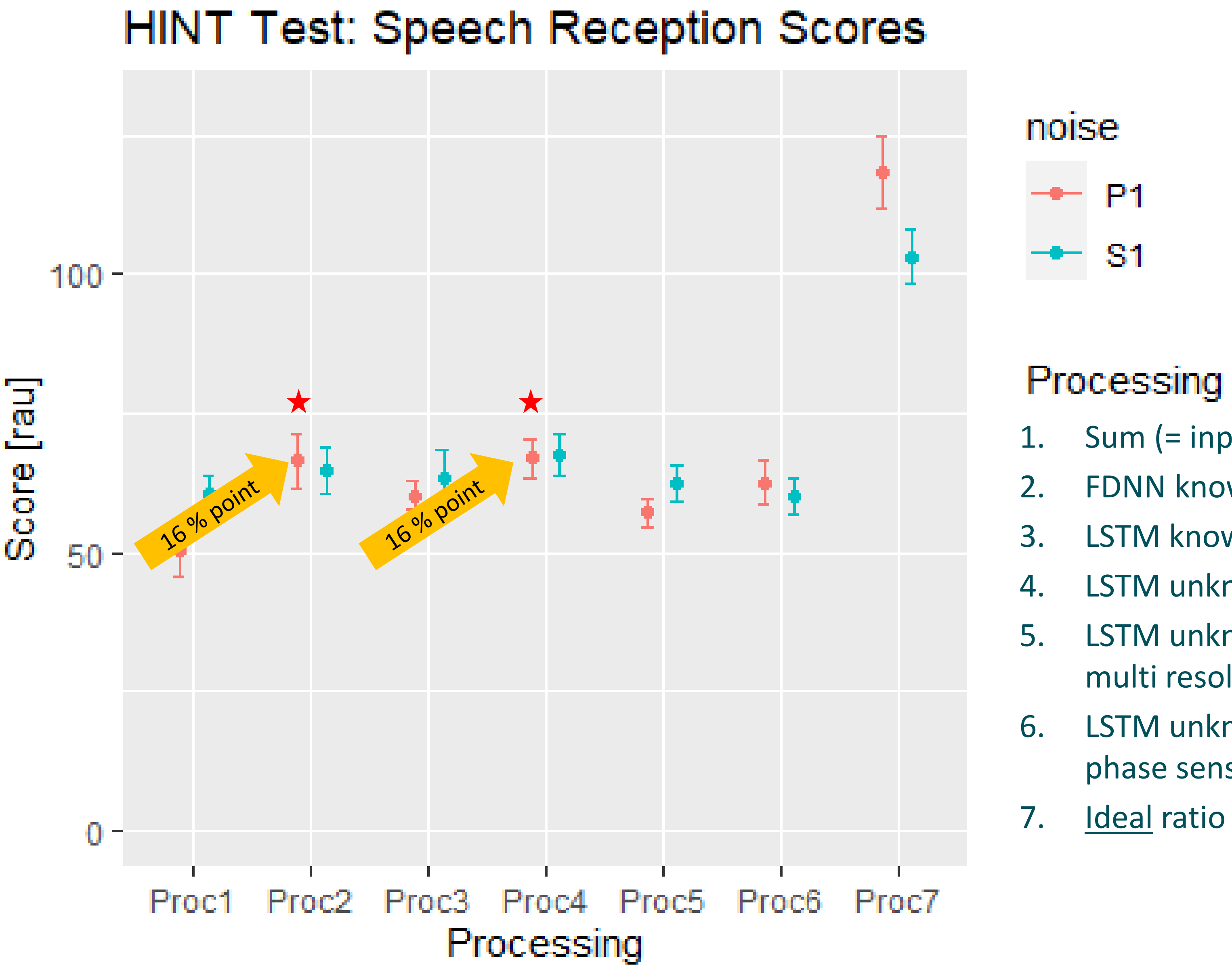
M1 + P1

~M1

M1 + S1

~M1

Results: post hoc



Processing

1. Sum (= input)
2. FDNN known voice
3. LSTM known voice
4. LSTM unknown voice
5. LSTM unknown voice + multi resolution mask
6. LSTM unknown voice + phase sensitive mask
7. Ideal ratio mask

Conclusion, speaker separation

- Competing voices test: Relevant, significant effect = 13% point.
  The user has all the information!

- Target-masker test: Large effect = 37% point
  The user must chose!

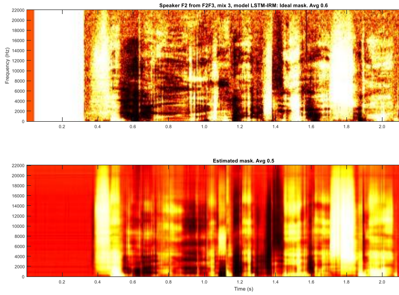
- All DNN modes (topologies) give the same improvement.

Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., and Virtanen, T. (2018). “Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm,” J. Acoust. Soc. Am., 144, 172–185. doi:10.1121/1.5045322

Conclusion, noise reduction

- Party noise: ~16 %-point (1.5 dB)
  - known voice FDNN
  - unknown voice LSTM
- Shopping centre: no benefit
  - Less modulated = less glimpses
- Unknown noise is a challenge!

Thank you!


Speaker F2 from F2F3, mix 3, model LSTM-IRM: Ideal mask. Avg 0.6
Frequency (Hz)
22000
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Estimated mask. Avg 0.5
Time (s)