intelligibility of speech in noise after lossy audio data compression

Gaston Hilkhuysen & Mark Huckvale

department of Speech, Hearing and Phonetic Sciences
University College London (UK)

Centre for Law Enforcement Audio Research

overview

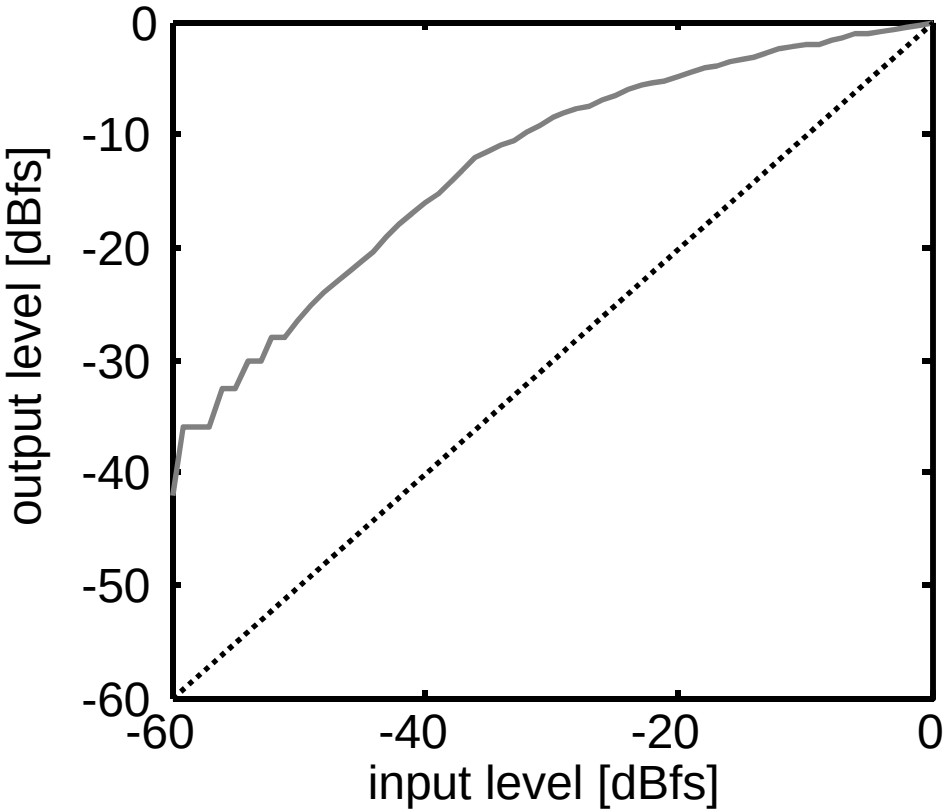
- compression strategies
- observed intelligibilities
- intelligibility metrics
- conclusions

G711 a-law

8 kHz sampling rate

12 bits compressed into 8 bits

=> 64 kbs CODEC



GSM 06.10

Regular Pulse Excitation - Long Term Prediction

8 kHz sampling; 20 m frames

- 160 samples

8 order LPC

32 bytes per frame

- =>13.2 kbs

MP3 / WMA

UCL

propriety patented algorithms

psycho-acoustically based

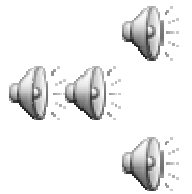
only transmit perceptually relevant

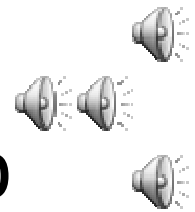
MP3 CBR, 16 kbs, 16 kHz, mono

MP3 CBR, 48 kbs, 44.1 kHz, mono

WMA 9.2 CBR, 16 kbs, 16 kHz, 16 bit mono

WMA 9.2 CBR, 48 kbs, 44.1 kHz, 16 bit mono





experimental design

CODEC

- PCM (256 kbs)
- G711 (64 kbs)
- GSM 06.10 (13 kbs)
- MP3 (48/16 kbs)
- WMA (48/16 kbs)

noise types

- car cabin noise
- babble

(5 SNRs per noise type)

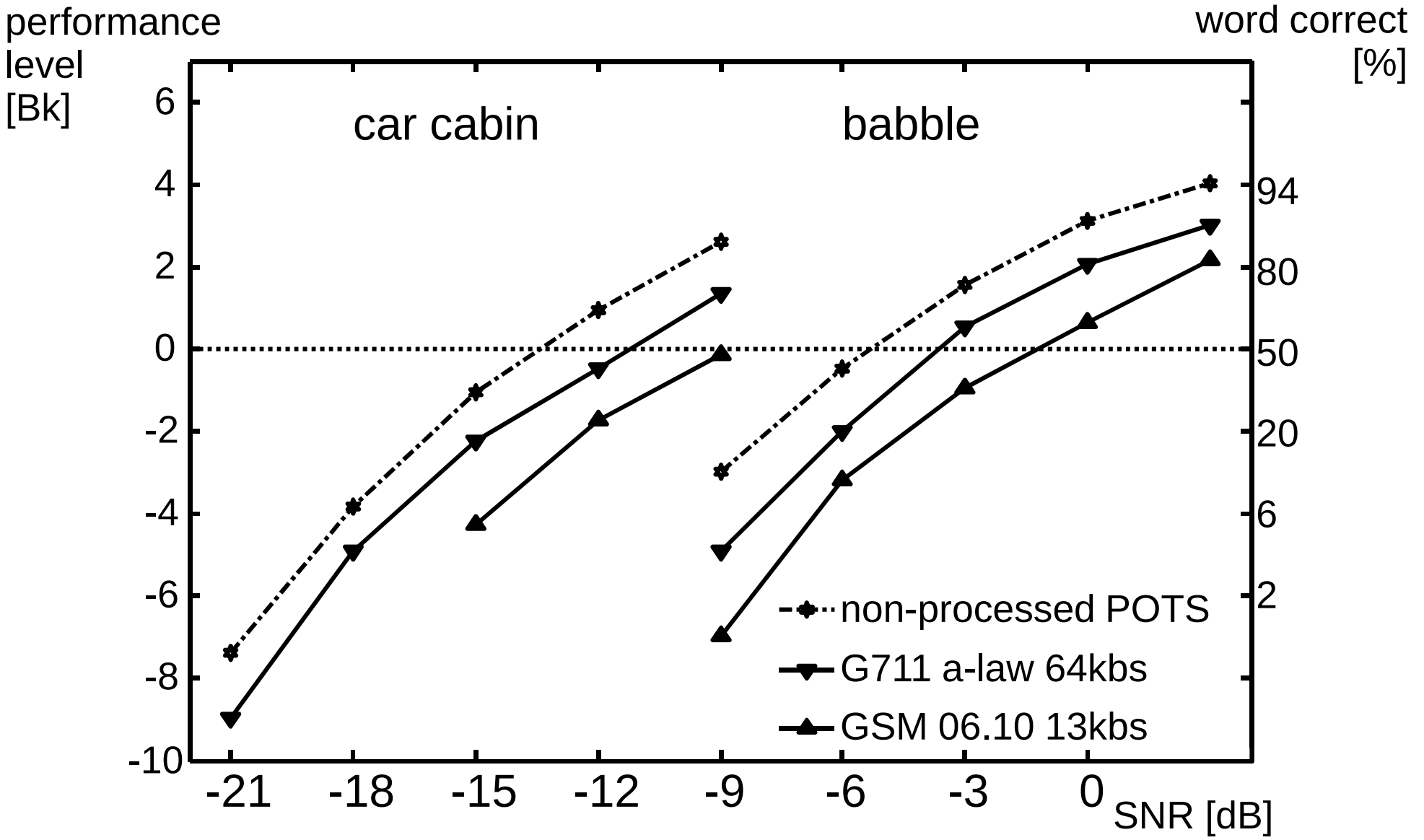
IEEE sentences

- 10 sentences per experimental condition
- 5 keywords per sentence
- 10 listeners per noise type

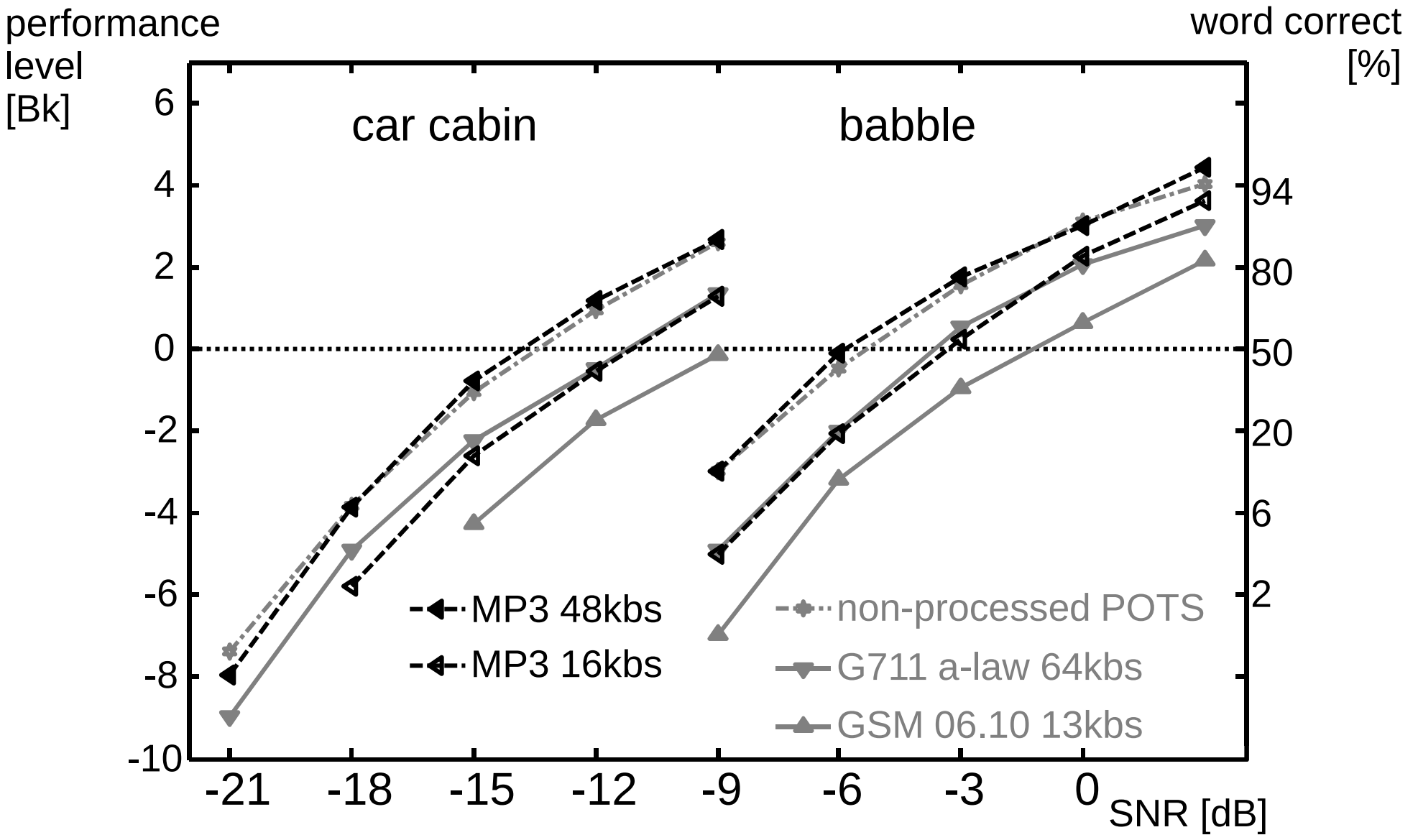
500 observations per experimental condition

70 experimental conditions in total

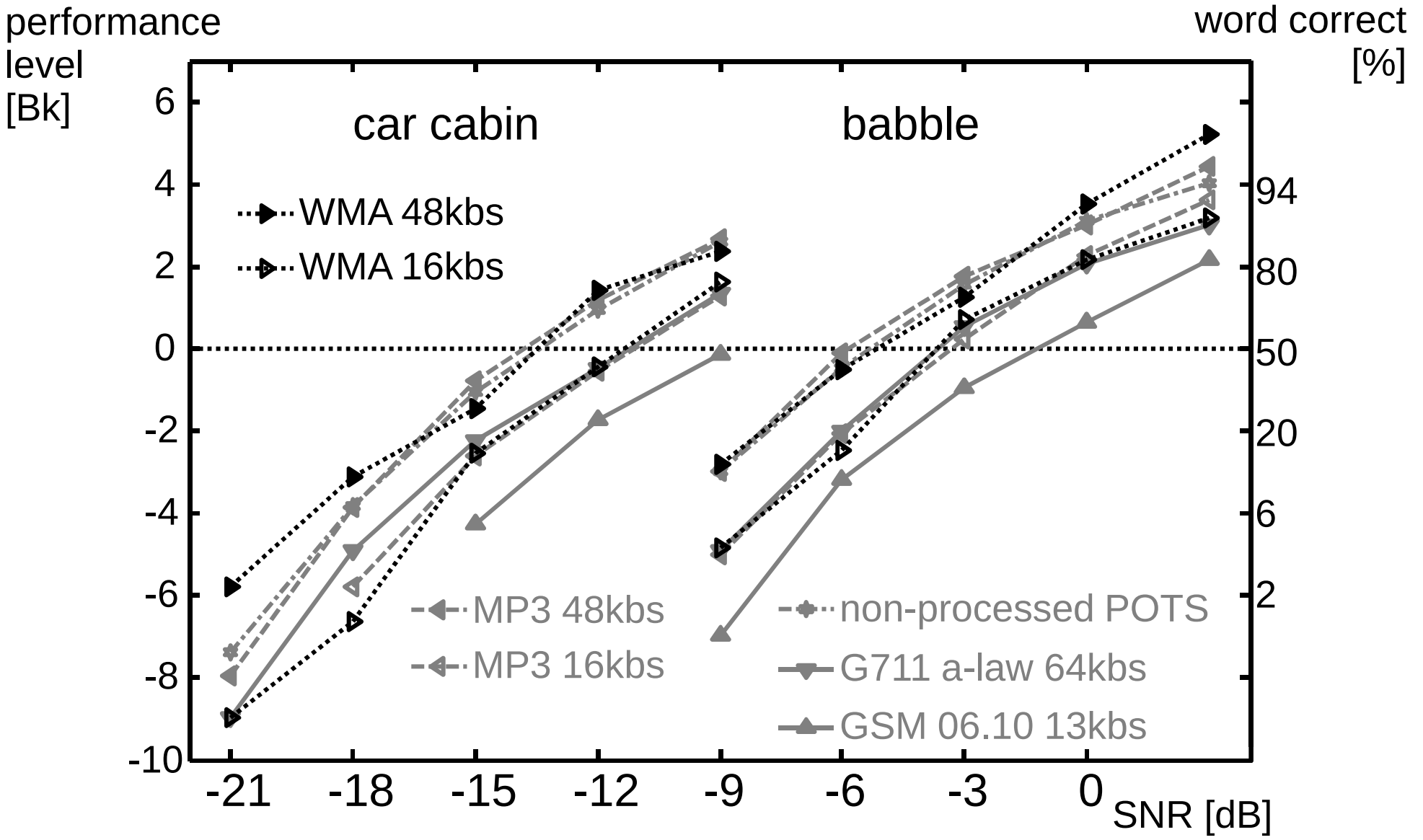
observations



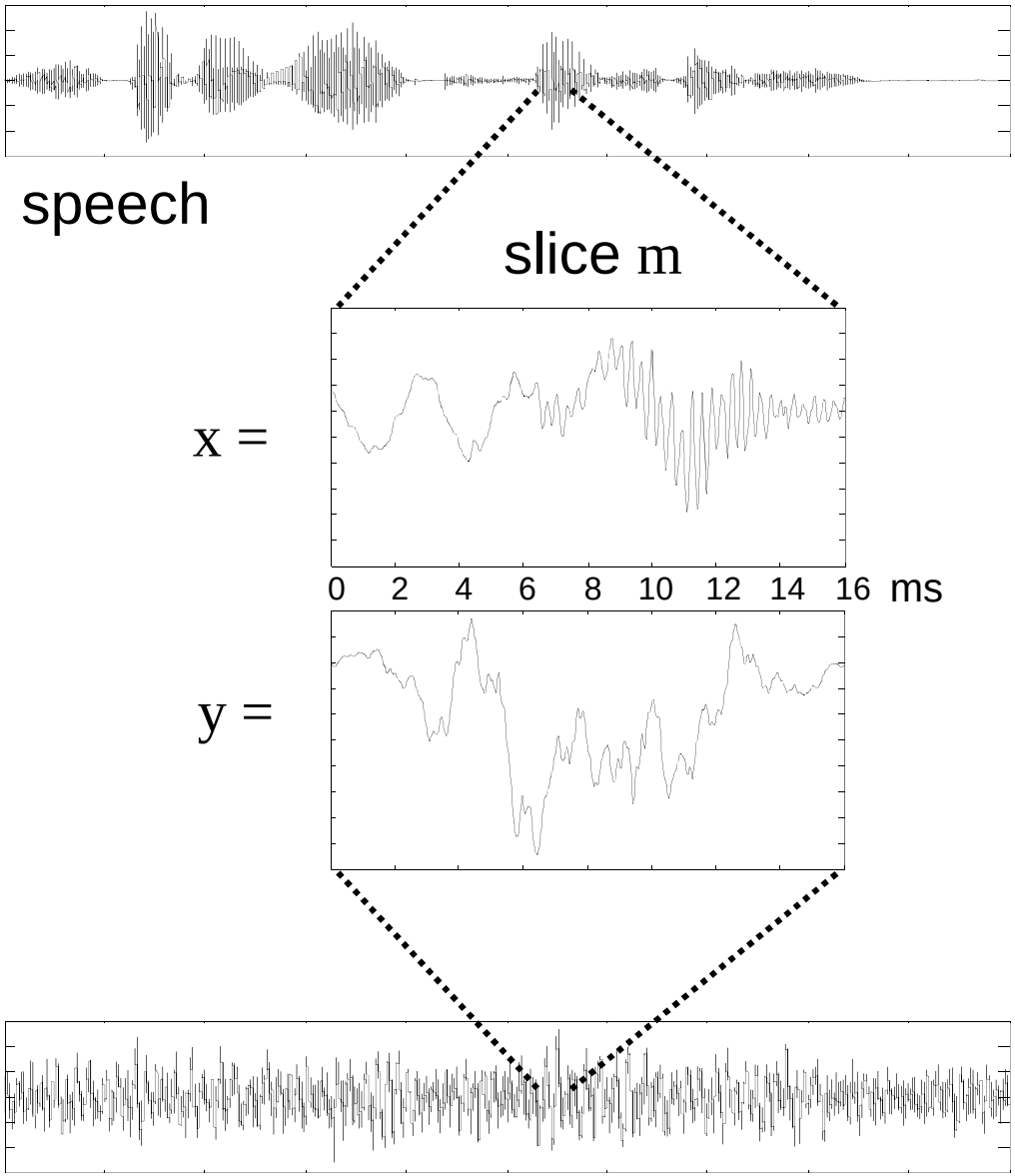
observations



observations



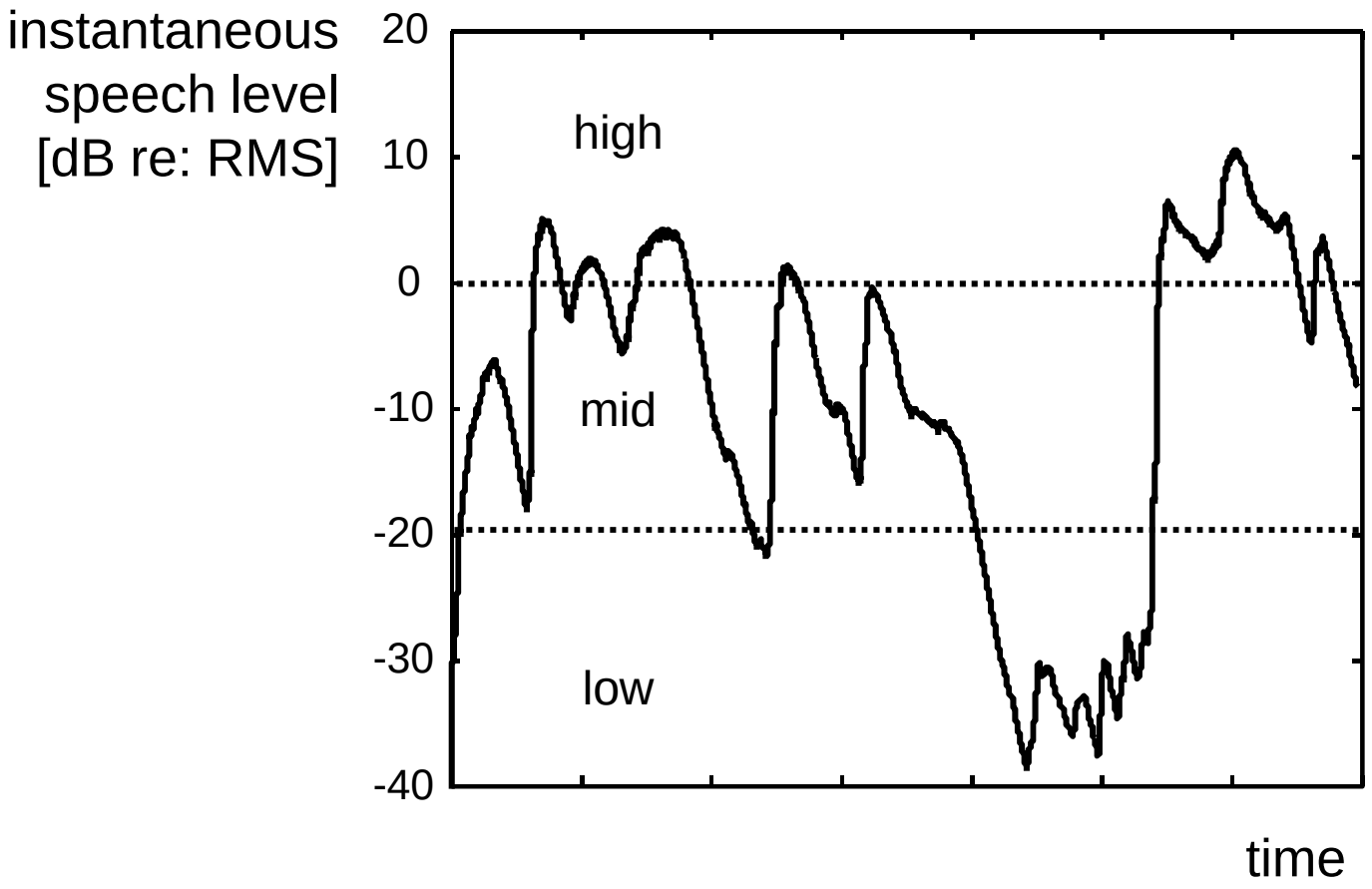
coherence SII



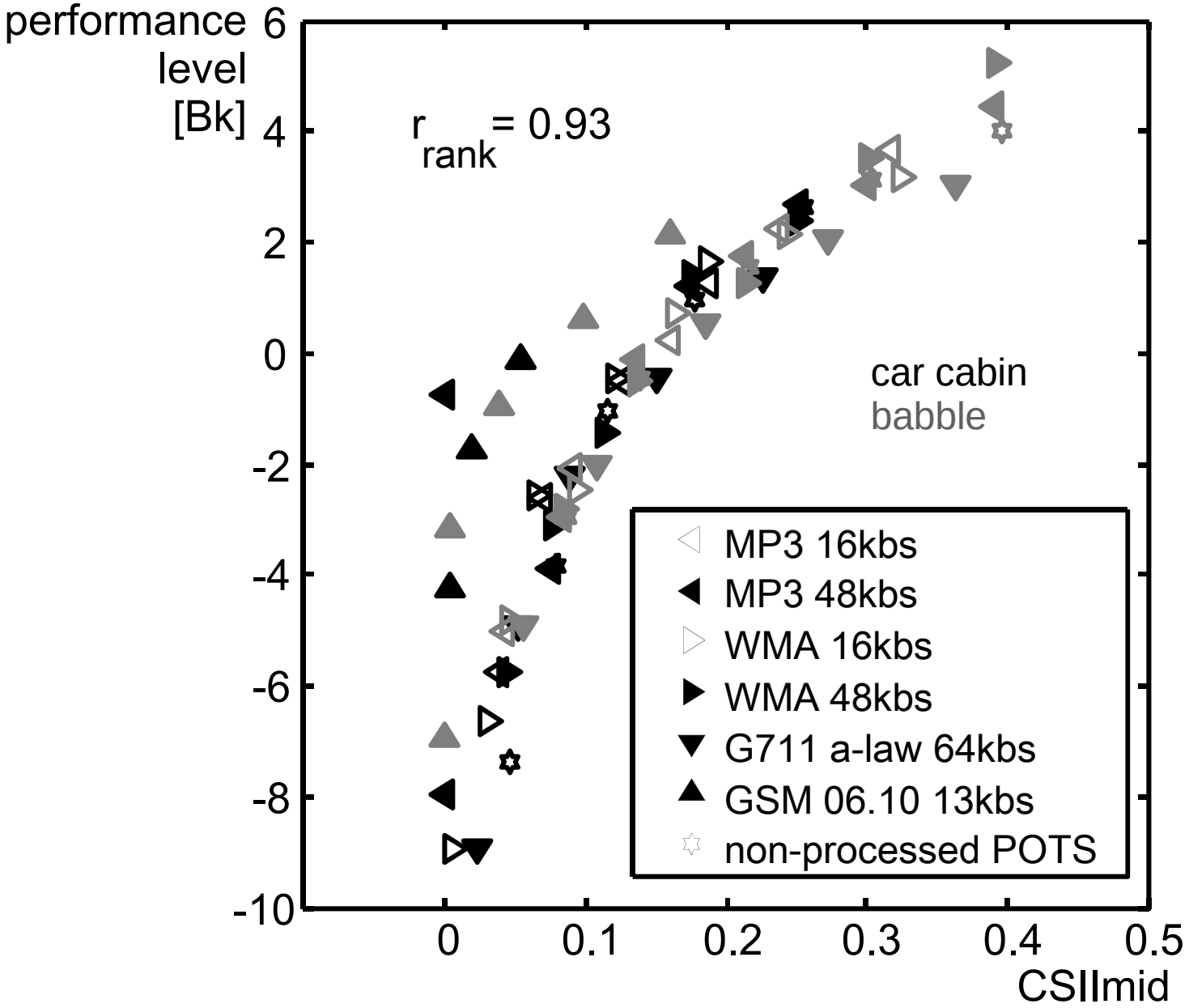
$$|\gamma(k)|^2 = \frac{\left|\sum_m X_m(k) Y_m^*(k)\right|^2}{\sum_m |X_m(k)|^2 \sum_m |Y_m(k)|^2}$$

$$SDR(j) = \frac{\sum_k W_j(k) |\gamma(k)|^2 S_{yy}(k)}{\sum_k W_j(k)[1 - |\gamma(k)|^2] S_{yy}(k)}$$

coherence SII



CSII performance function

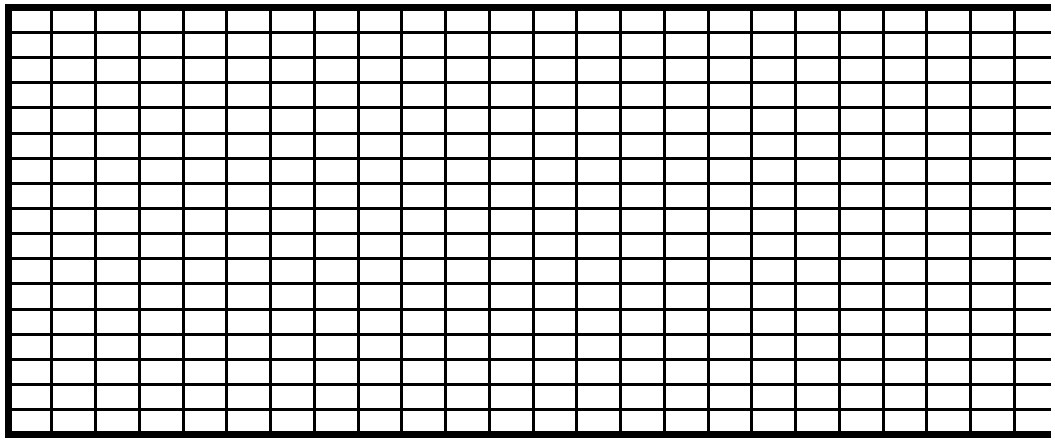


STOI

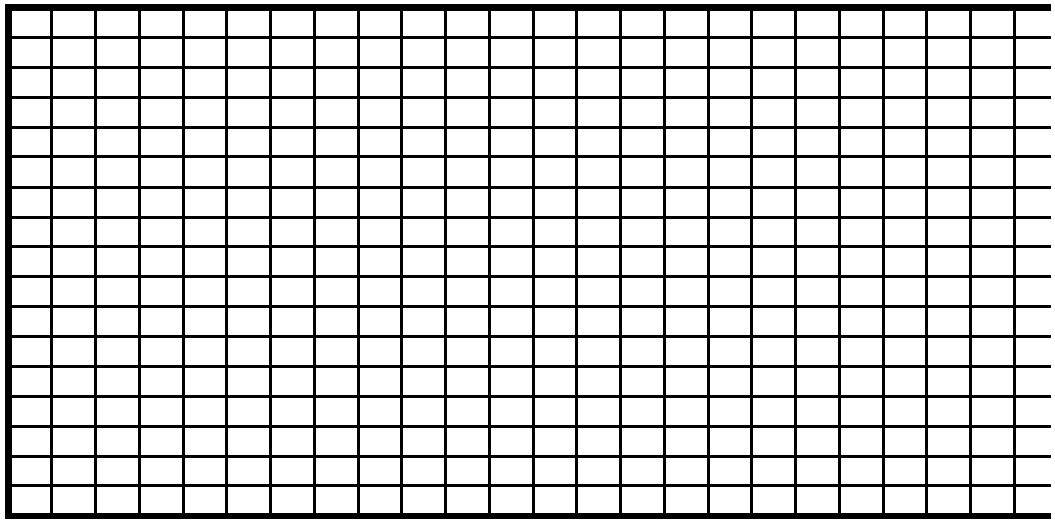
spectrogram pixels

- 1/3 octave x 13 ms

speech



speech + noise



STOI

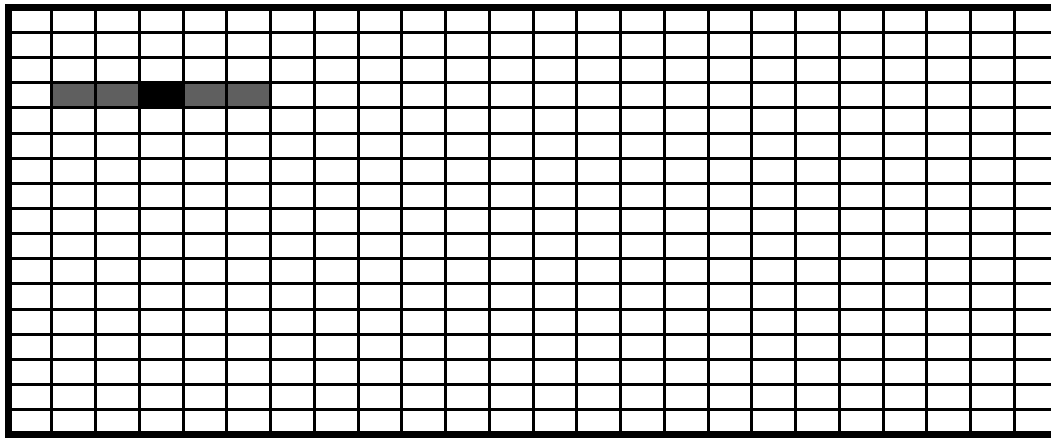
spectrogram pixels

- 1/3 octave x 13 ms

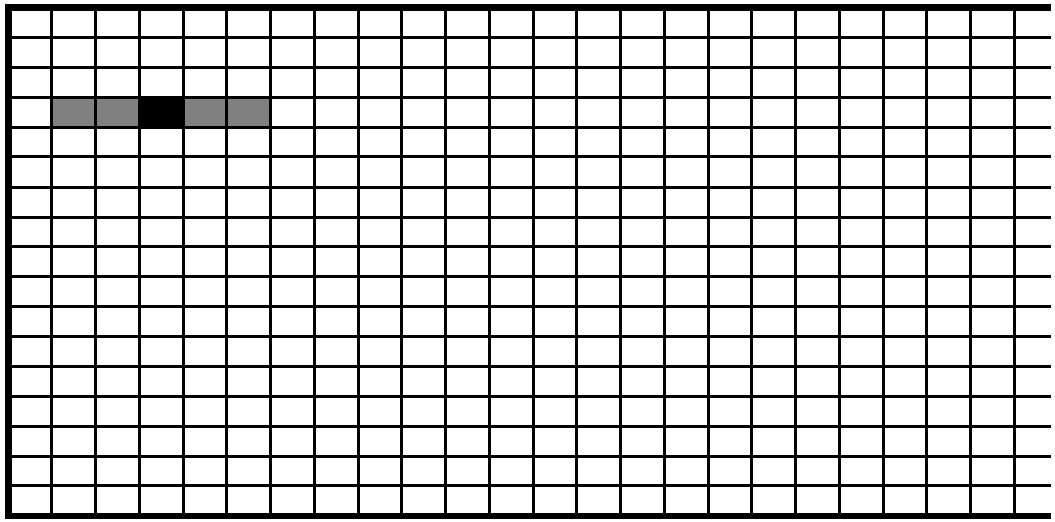
equalize energies in corresponding 400 ms ranges

limit pixel difference to range <-15..15> dB

speech



speech + noise



STOI

spectrogram pixels

- 1/3 octave x 13 ms

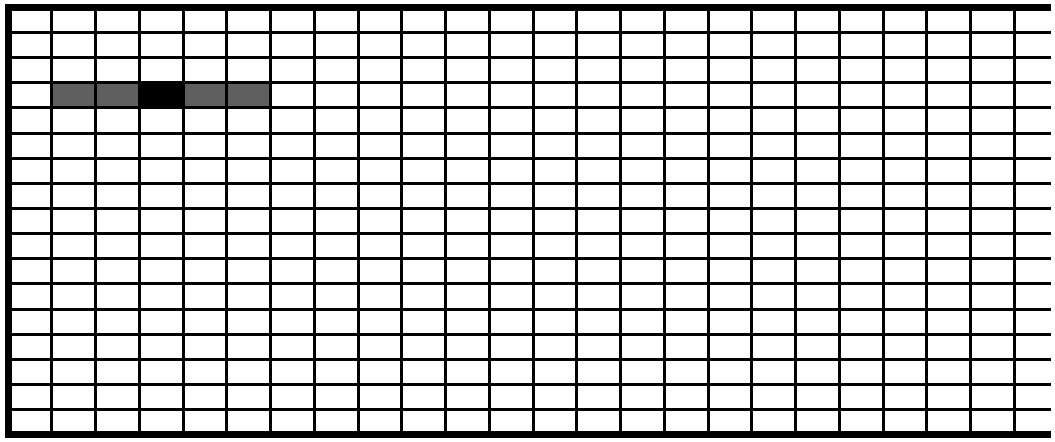
equalize energies in corresponding 400 ms ranges

limit pixel difference to range <-15..15> dB

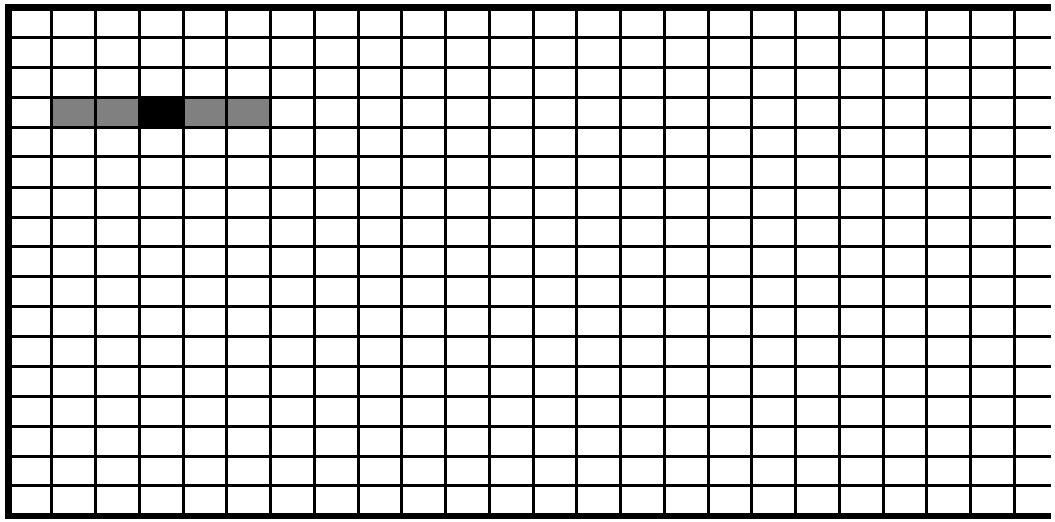
correlate regions

average correlations

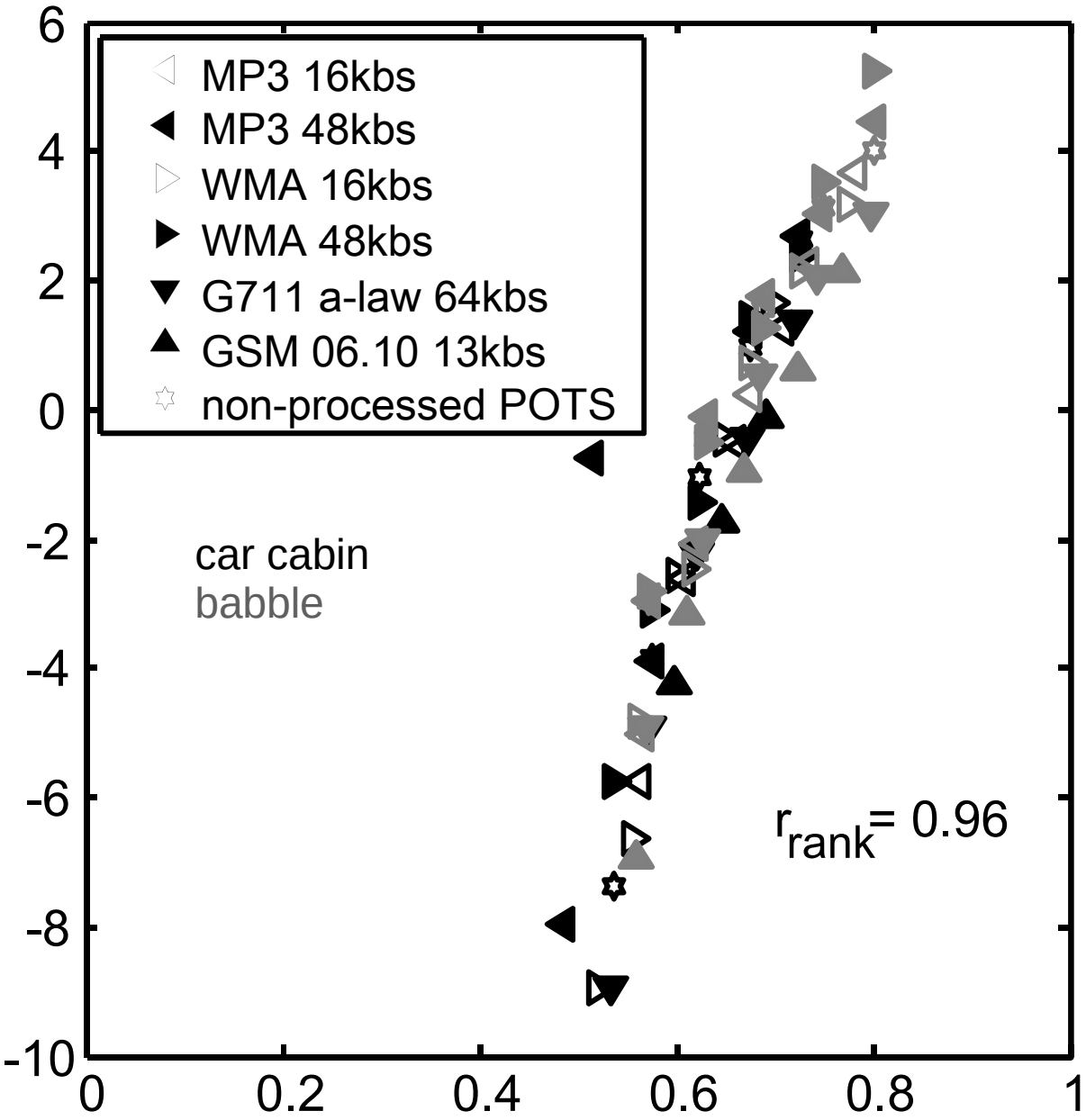
speech



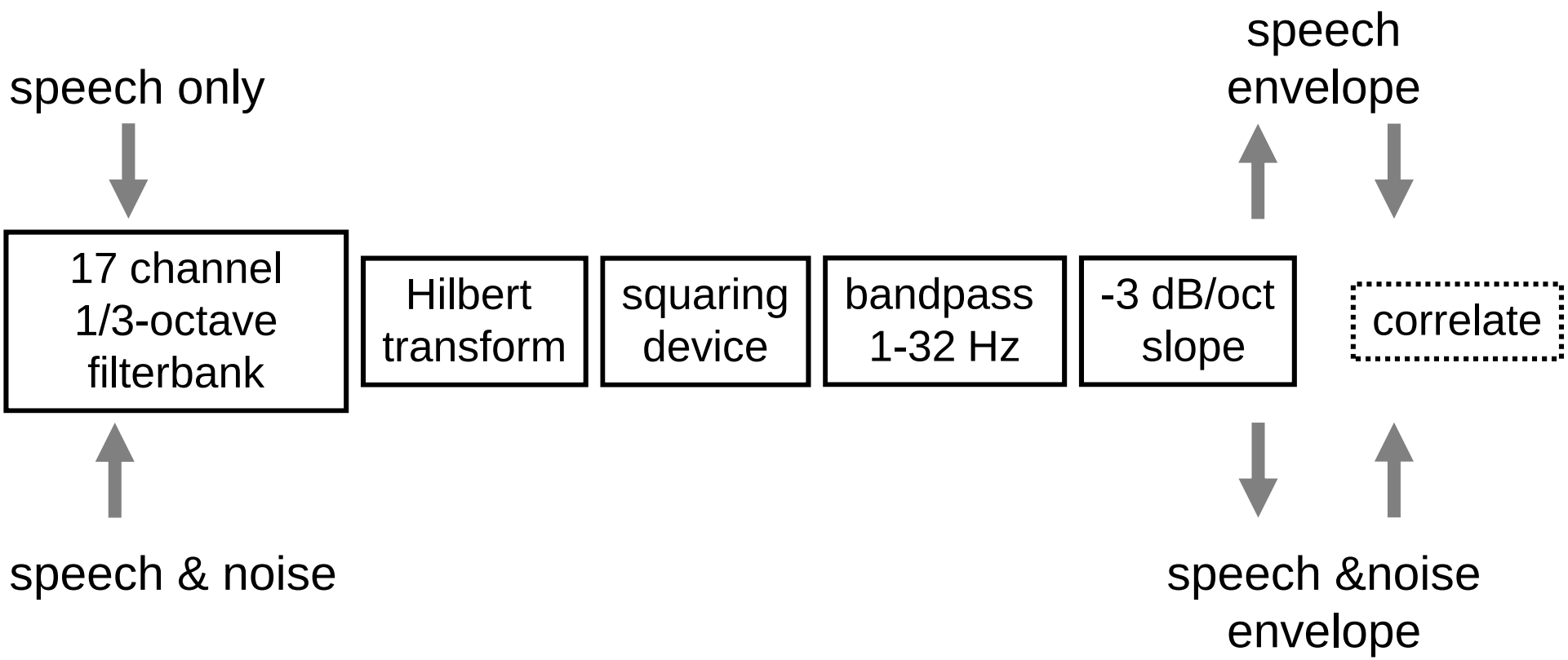
speech + noise



STOI performance functions

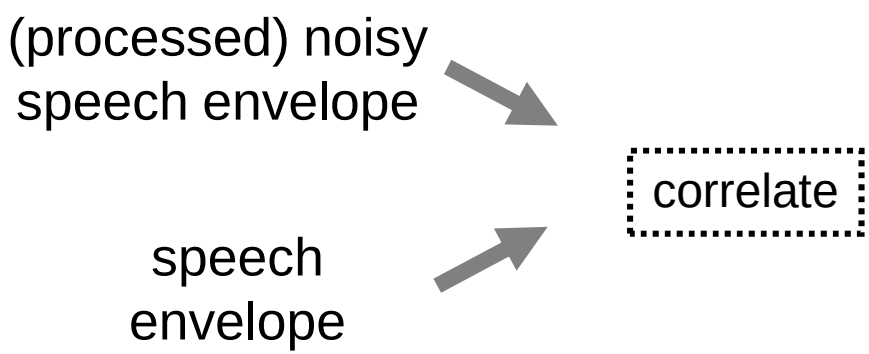


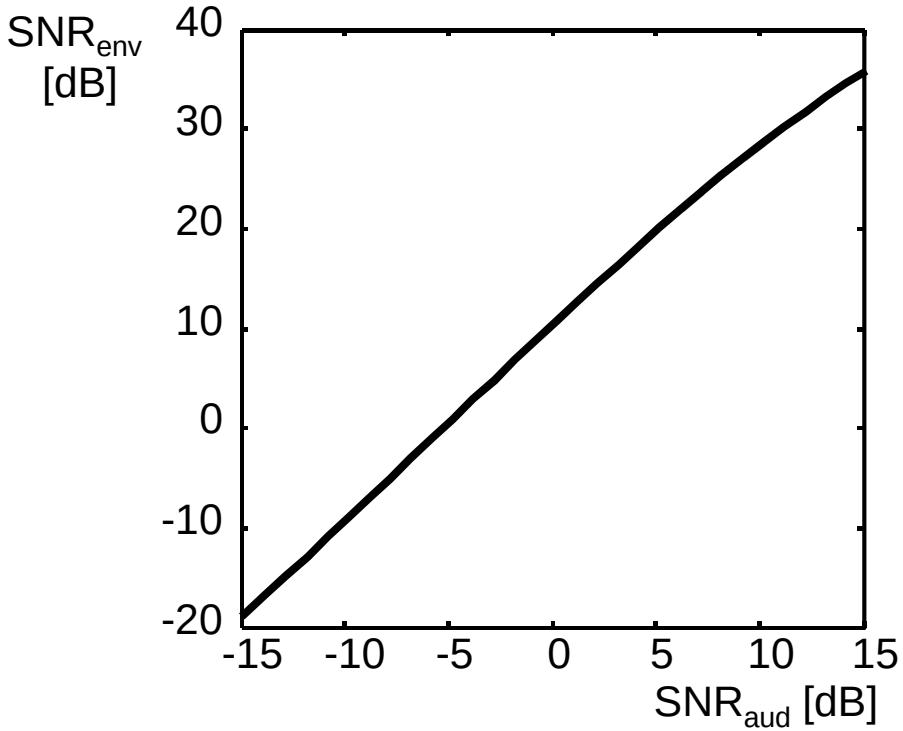
SNR in the envelope domain



$$SNR_{env} = 10\log_{10}\left(\frac{r^2}{1-r^2}\right)$$

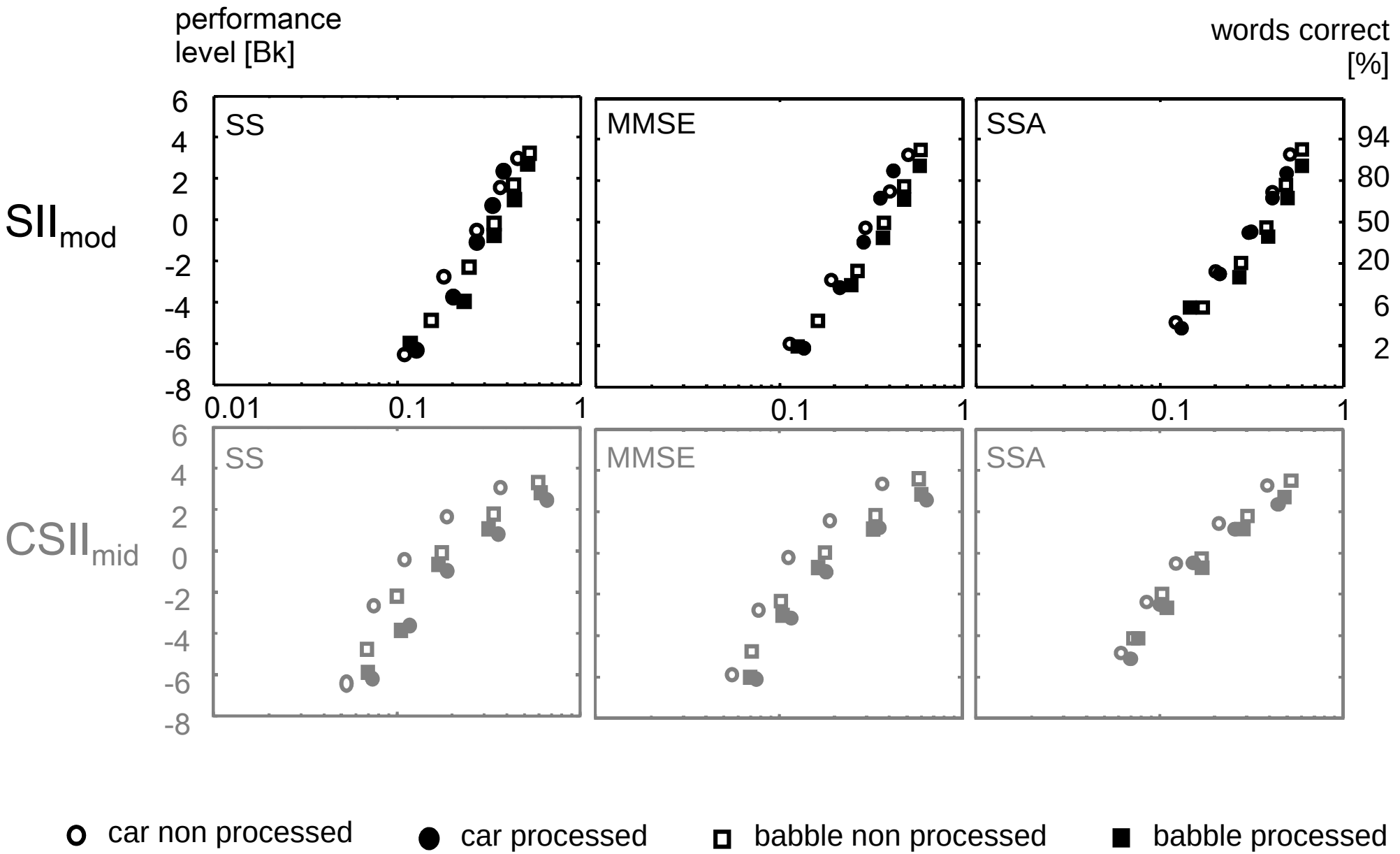
SNR in envelope domain



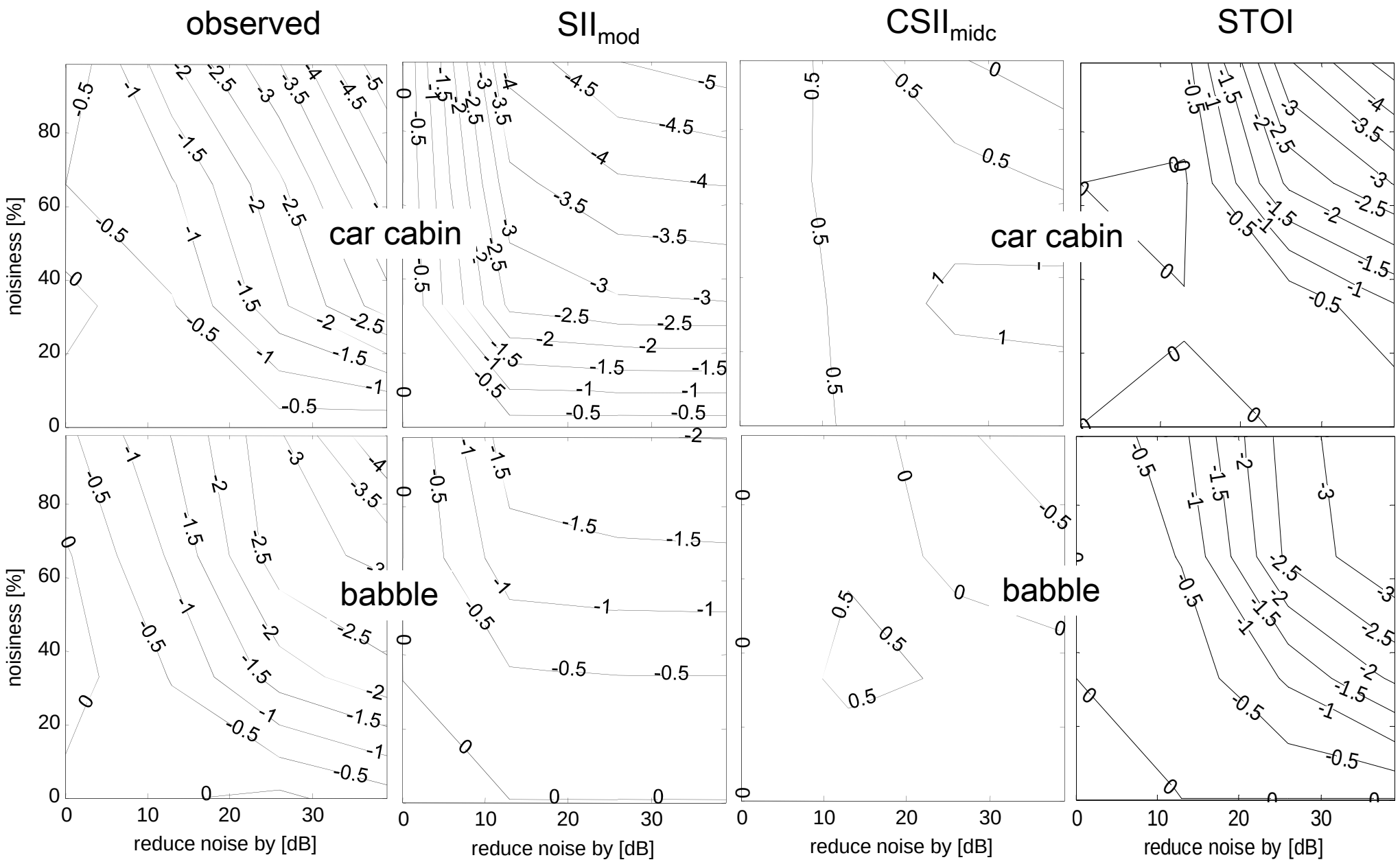


$$SNR_{env} = 10\log_{10}(\frac{r^2}{1-r^2})$$

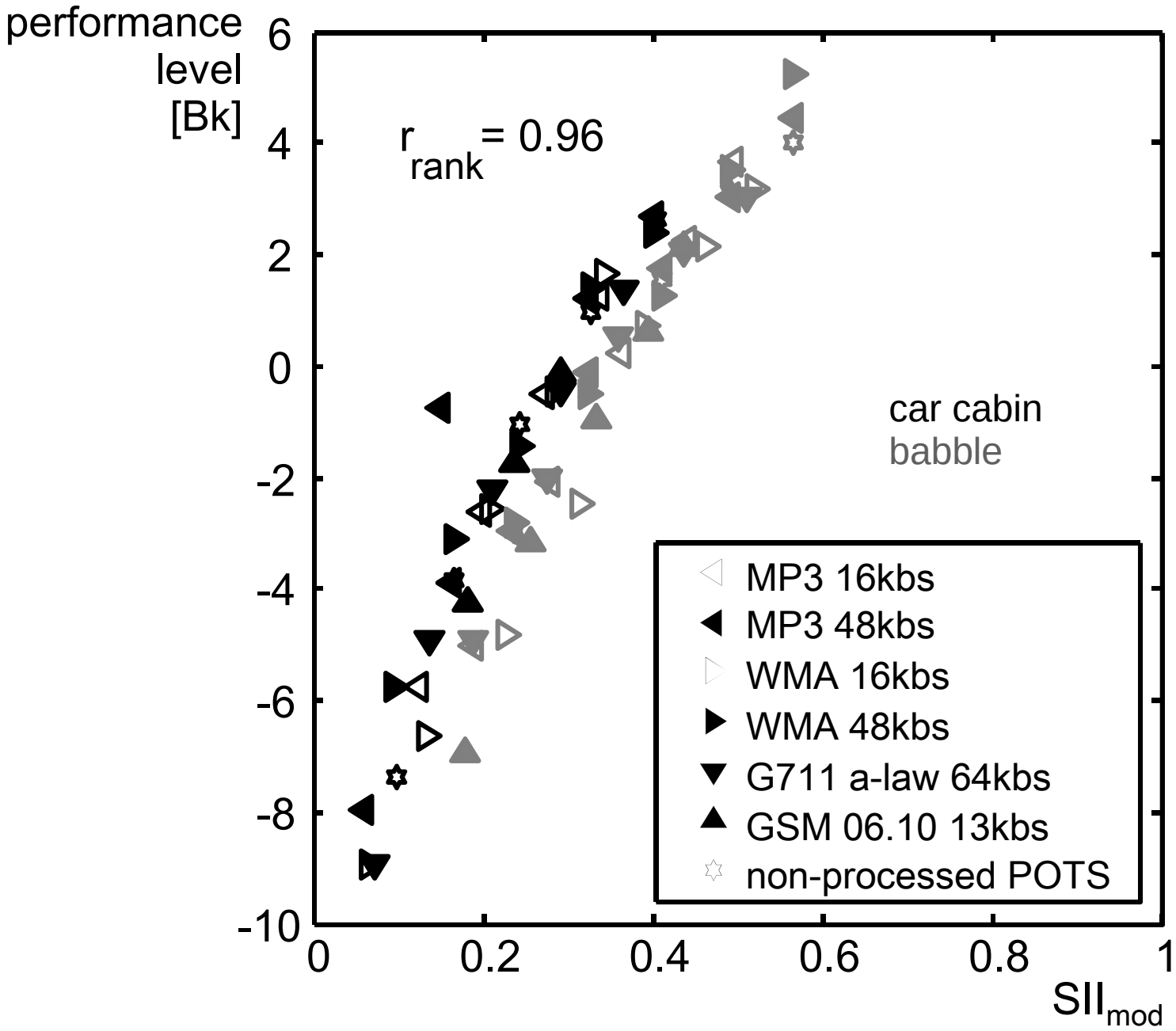
modelling NR effects



noise reduction parameter optimization



SII_{mod} performance functions



conclusions

CODECs alter intelligibility of noisy speech

- negatively/positively(!)

effects can be predicted

- STOI/SII_{mod}

SII_{mod} may be improved by improving SII predictions for speech in babble

- modulations deteriorate intelligibility
- dip listening imporves intelligibility

Special thanks to **CLEAR**

UCL

Mike
Brooks

Dushyant
Sharma

Nikolai
Gaubitch

Mark
Wibrow

Patrick
Naylor

Mark
Huckvale