Modeling human speech recognition

Thomas Brand

Medizinische Physik, Universität Oldenburg

Cluster of Excellence “Hearing4All“

Why modeling human speech recognition?

- Evaluation of speech enhancement systems with listening experiments is time consuming and costly.
- How do humans solve the cocktail party problem?
- What are the consequences of hearing loss with respect to speech recognition?
- What is required to compensate for hearing loss?

Outline of this talk

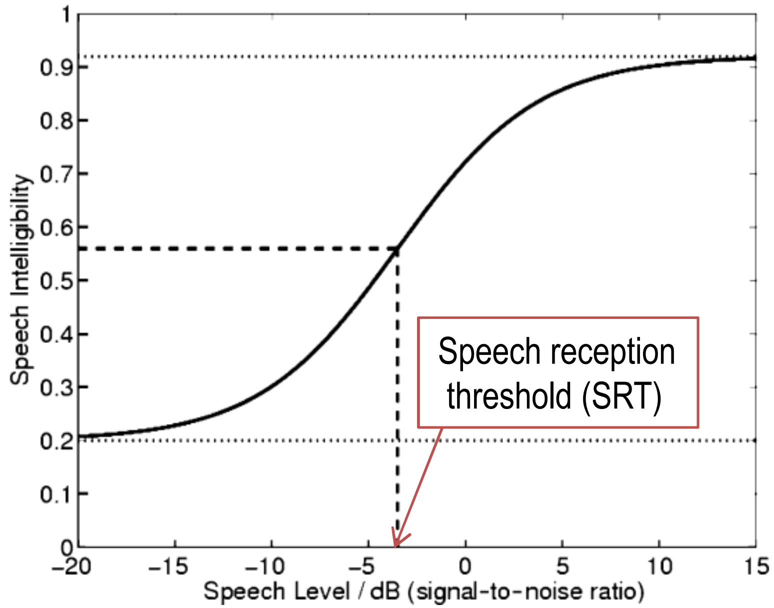
- Assessment of human speech recognition
  - speech intelligibility
  - masking
- Speech Intelligibility Prediction (SIP)
  - general model framework
  - example models
  - additional information
  - non-intrusive SIP
- Discussion

Assessment of human speech recognition

- Measures
  - speech quality
  - listening effort
  - intelligibility
  - …
- Speech materials
  - phonemes, single words
  - sentences, running speech
  - …
- Evaluation method
  - questionnaires
  - ratings
  - comparisons
  - recognition scores
  - …

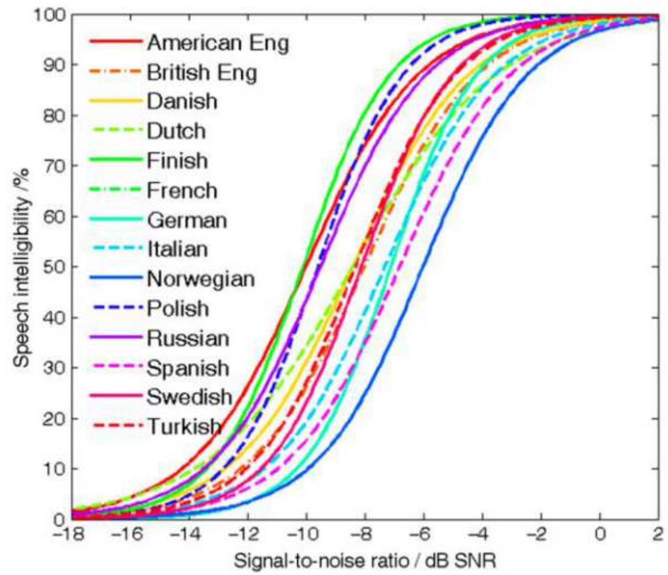
Speech Intelligibility (SI)

Intelligibility function



Matrix sentences

Name	Verb	Number	Adjective	Noun
Peter	got	three	large	desks.
Kathy	sees	nine	small	chairs.
Lucy	bought	five	old	shoes.
Alan	gives	eight	dark	toys.
Rachel	sold	four	thin	spoons.
Barry	likes	six	green	mugs.
Steven	has	two	cheap	ships.
Thomas	kept	ten	pink	rings.
Hannah	wins	twelve	red	tins.
Nina	wants	some	big	beds.

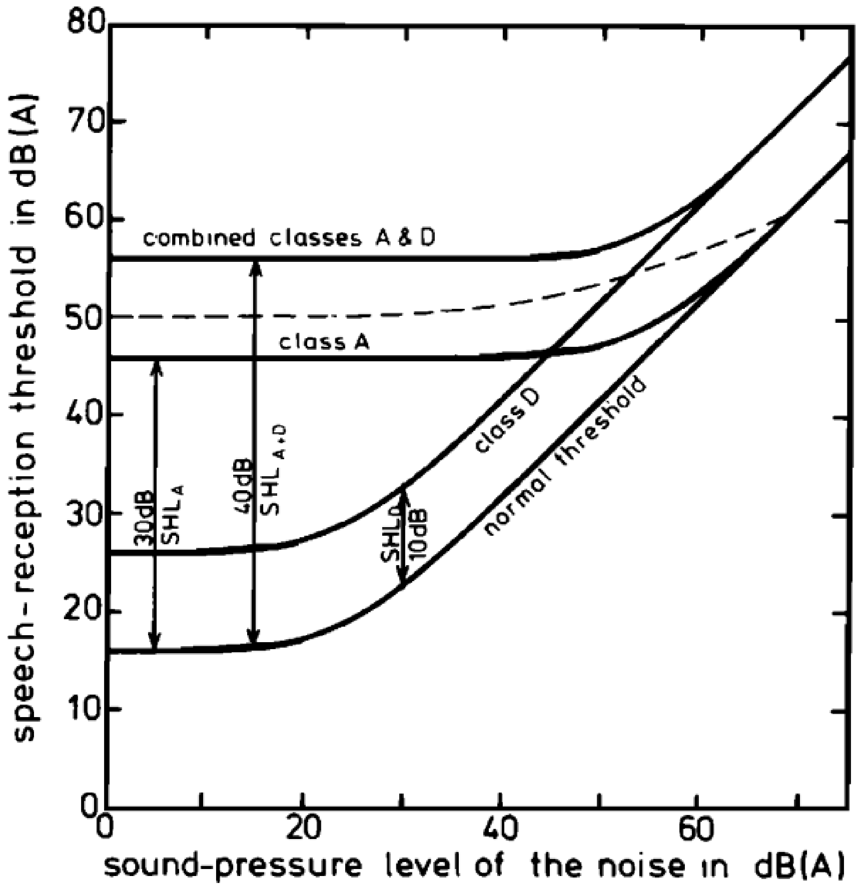


(Kollmeier et al, 2015, Hagermann 1982, Hochmuth et al 2012, Jansen et al 2012, Ozimek et al 2010, Wagener et al 1999, 2003, 2004, Warzybok et al 2011, Zokoll et al 2013, …)

Plomp Model

- simple phenomenological model
- fits very well to human data

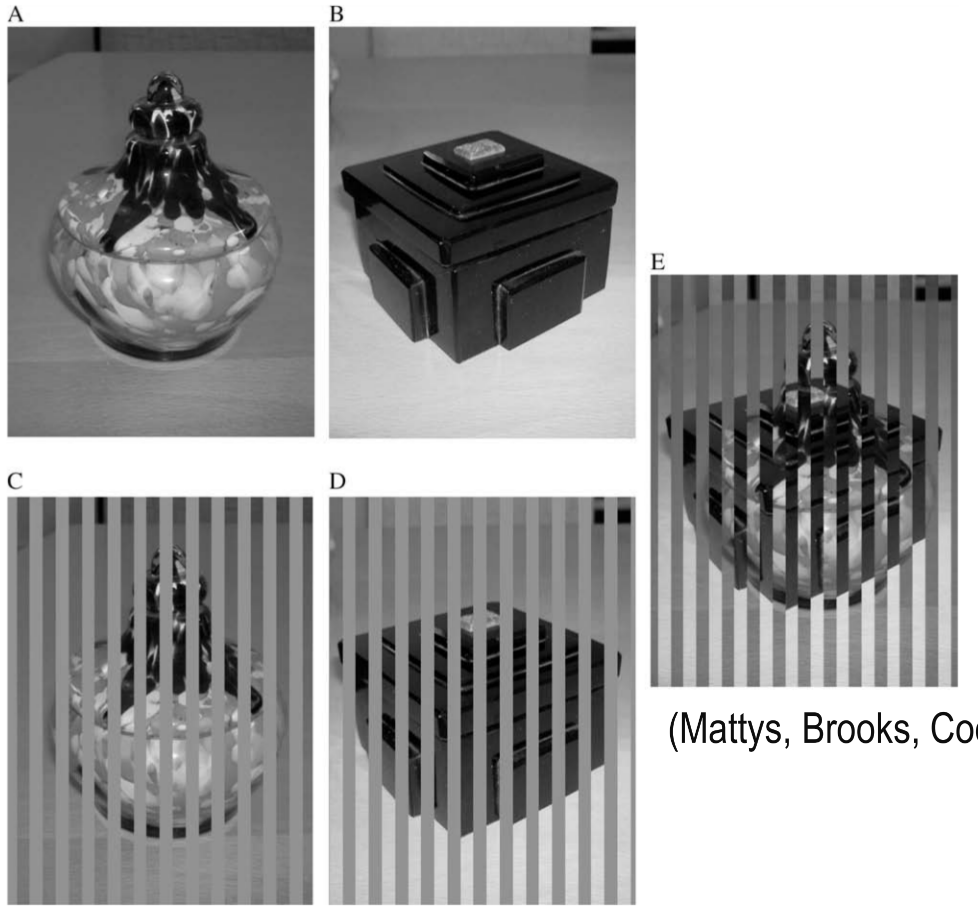
- SRT_{A+D} : speech reception threshold (SRT)
- L_n : sound pressure level of noise
- L_0 : SRT in quiet
- ΔL_{SN} : SRT in noise (in dB SNR)
- A : hearing loss of **class A (attenuation)**
- D : hearing loss of **class D (distortion)**



(Plomp, 1978)

$$SRT_{A+D} = 10 \cdot \lg(10^{(L_0+A+D)/10} + 10^{(L_n-\Delta L_{SN}+D)/10})$$

Masking



(Mattys, Brooks, Cooke, 2009)

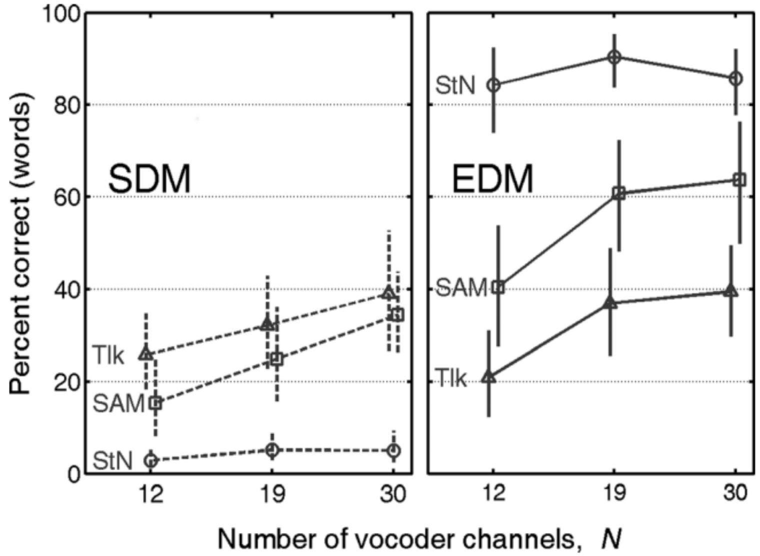
- ***Energetic masking:***

  Target energy is masked by interfering energy.

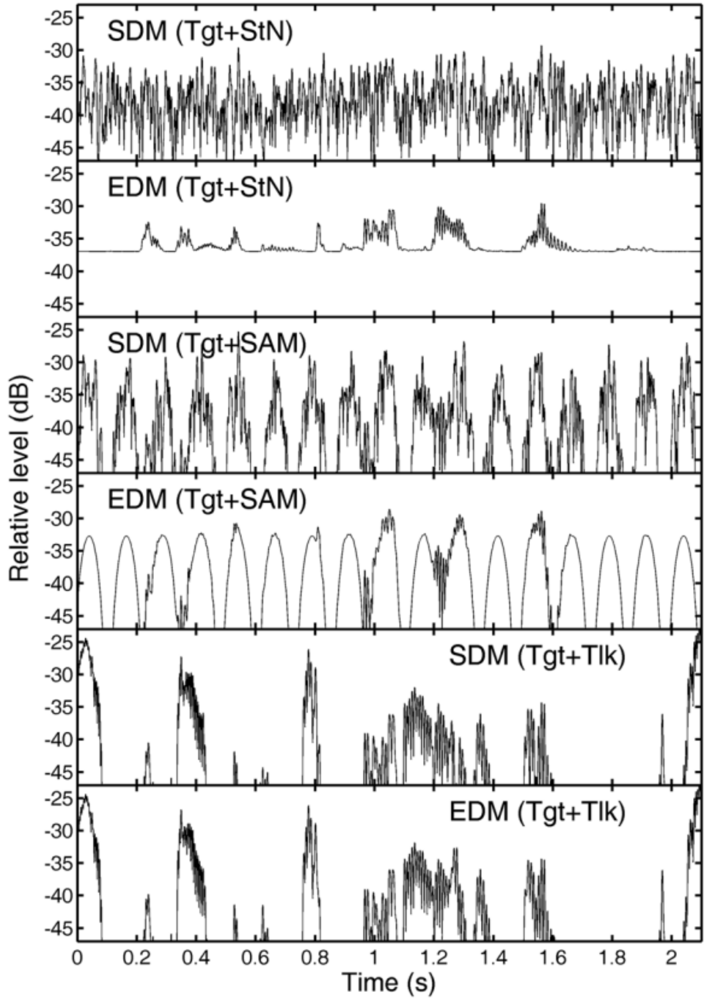
- ***Informational masking:***

  Target cannot be segregated from interferer.

Masking and auditory filters



(Stone, Füllgrabe, Mackinnon, Moore, 2011)



- ***Energetic masking:***

  Target energy is masked by interfering energy.

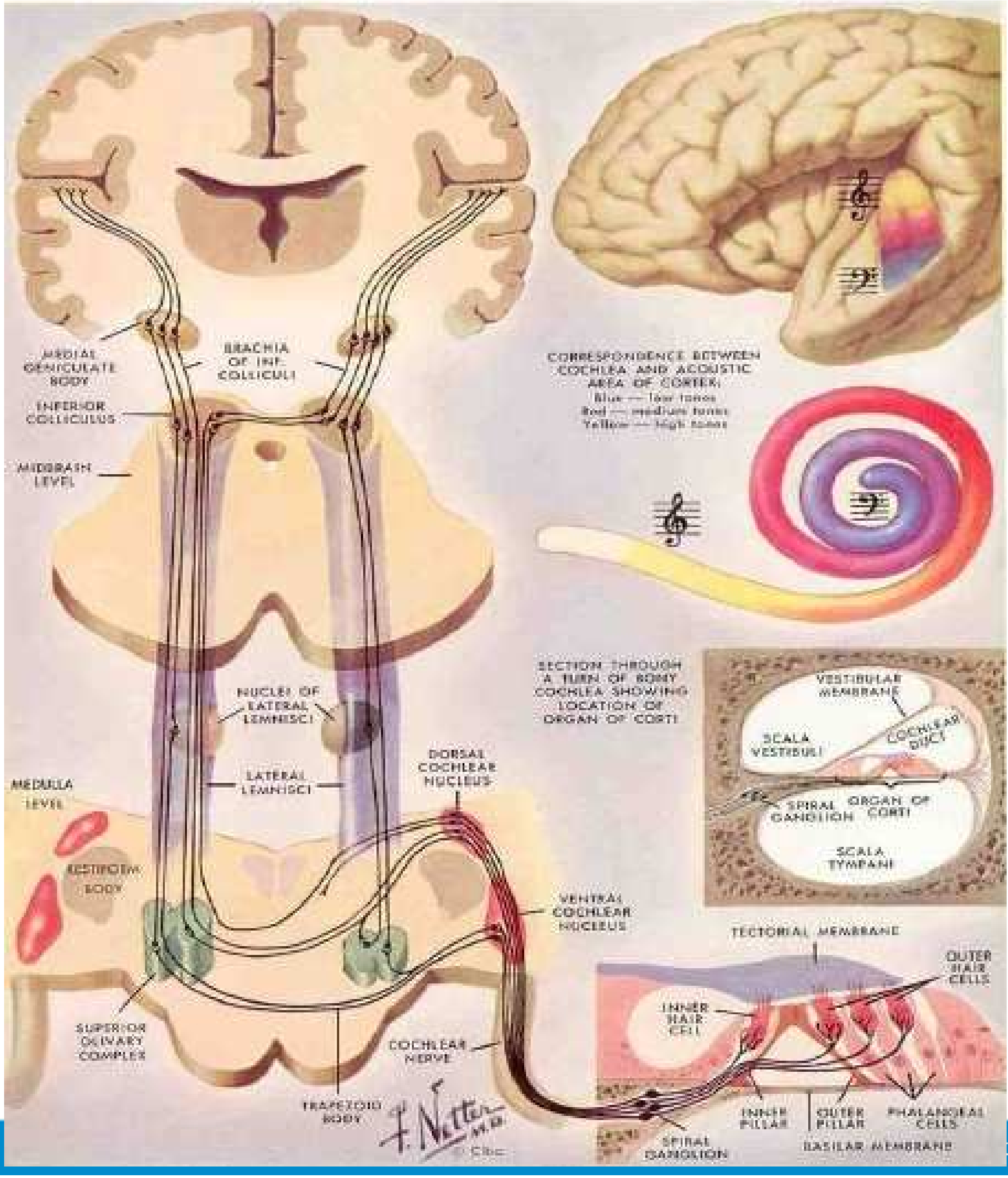
- ***Modulation masking:*** (Stone, Füllgrabe, Moore, 2012)

  Target modulations in auditory filters are masked by interfering modulations.

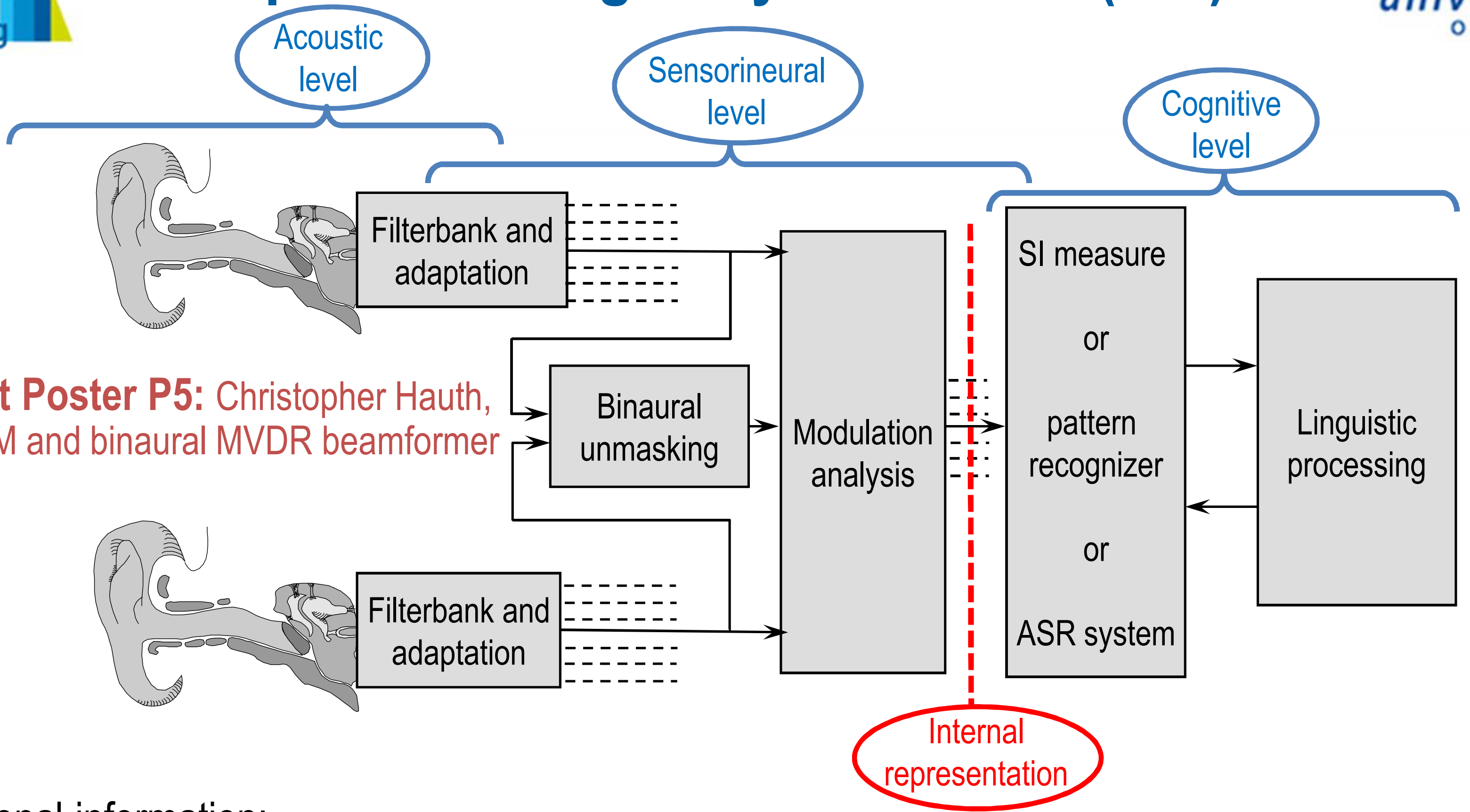
- ***Informational masking:***

  Target cannot be segregated from interferer.

Model of the auditory pathway



Effective model of auditory processing and Speech Intelligibility Prediction (SIP)

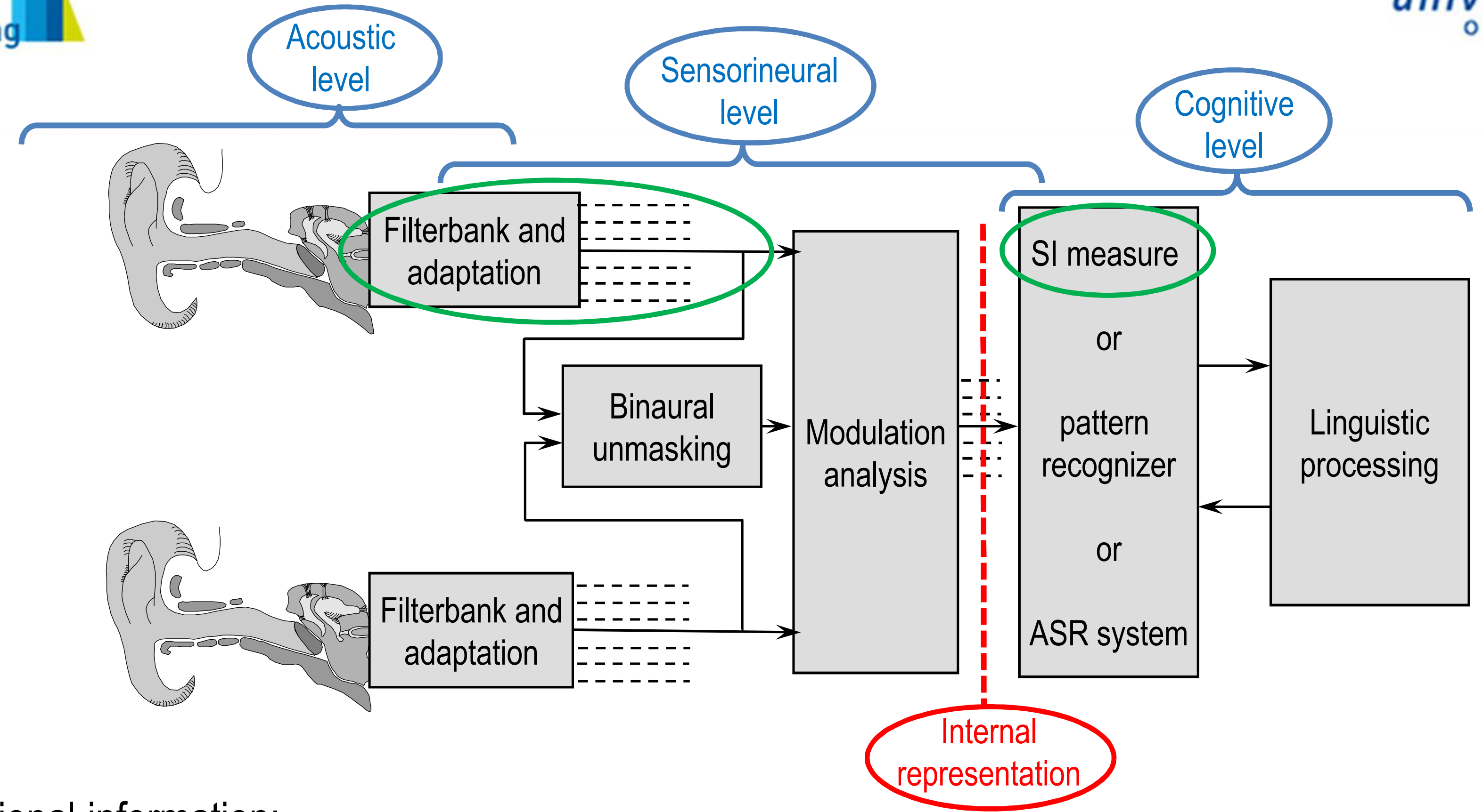


visit Poster P5: Christopher Hauth, BSIM and binaural MVDR beamformer

Additional information:

- Listener: hearing loss, linguistic abilities, …
- Signals: room acoustics, speech parameters, noise parameters, training material, …
- Linguistics: language, semantic context, …

Speech Intelligibility Prediction (SIP)



Additional information:

- Listener: hearing loss
- Signals: room acoustics, speech freq. spectrum, noise freq. spectrum
- Linguistics: semantic context

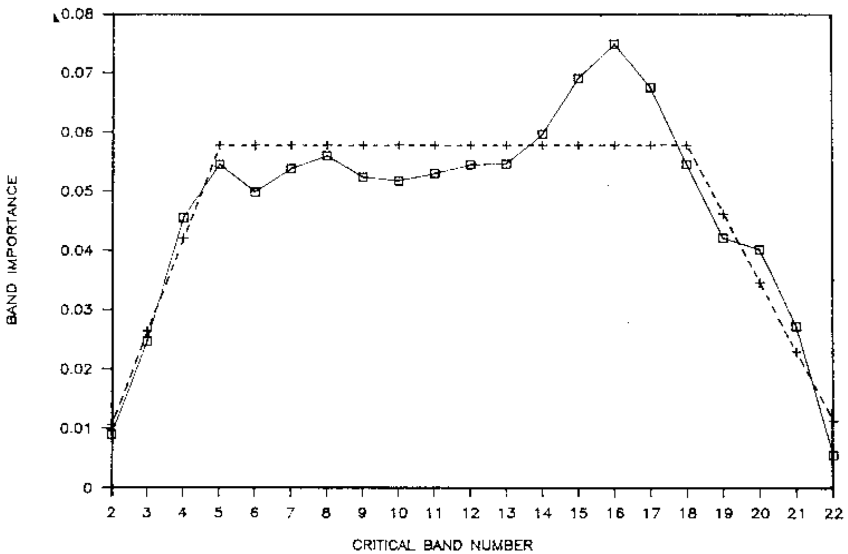
Articulation Index (AI)
Speech Intelligibility Index (SII)

$$(1-s)=\prod_i (1-s_i)$$ Fletcher-Steward independent channel model

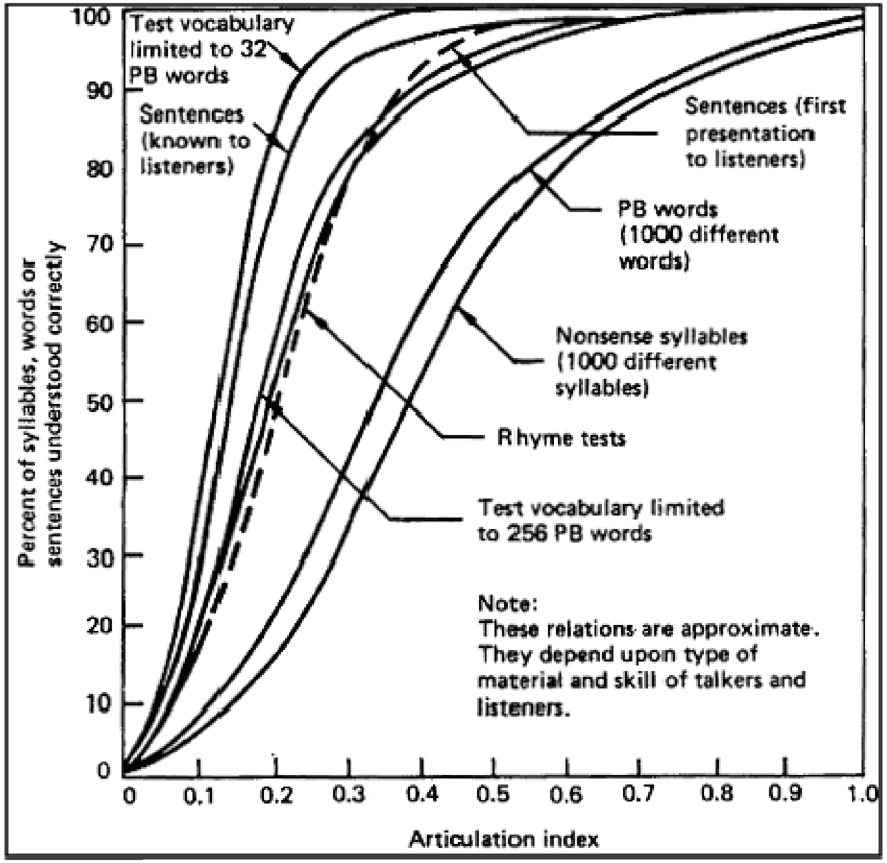
$$\log(1-s)=\sum_i \log(1-s_i)$$

$$\mathrm{AI}=\sum_i \mathrm{AI}_i \quad , \qquad \text{with } \mathrm{AI}=-k\log(1-s)$$

$$\mathrm{AI}=\sum_i \frac{W_i(\mathrm{SNR}_i+15)}{30}$$



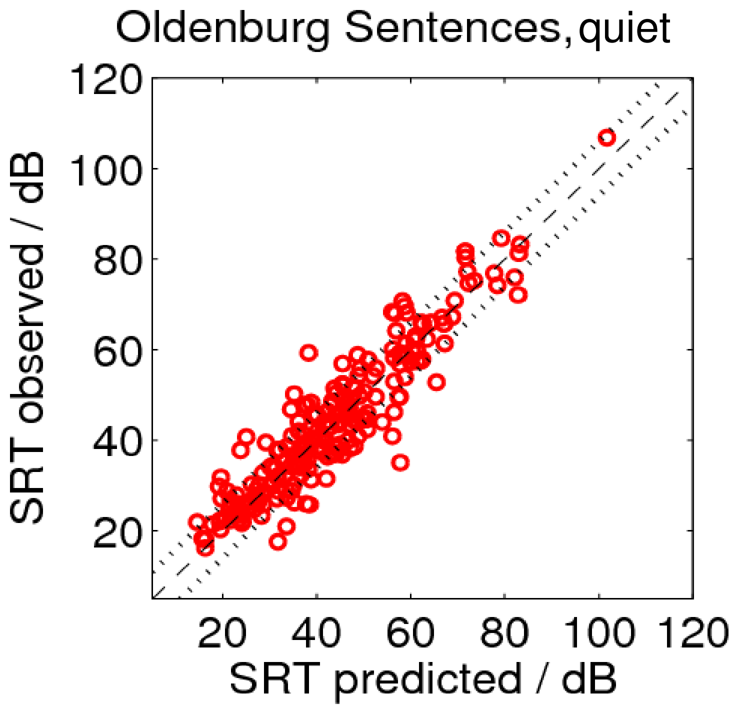
(Pavlovic, 1987)

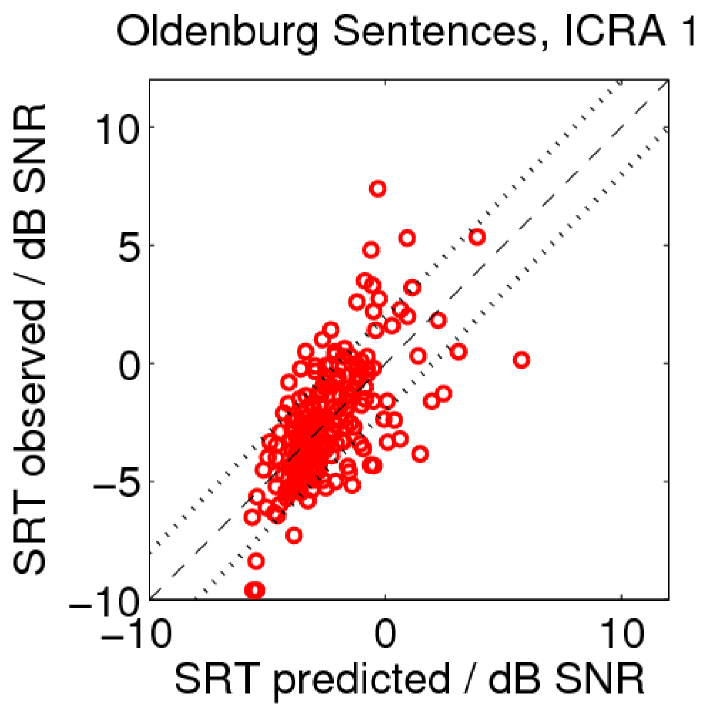


(ANSI S3.5-1969)

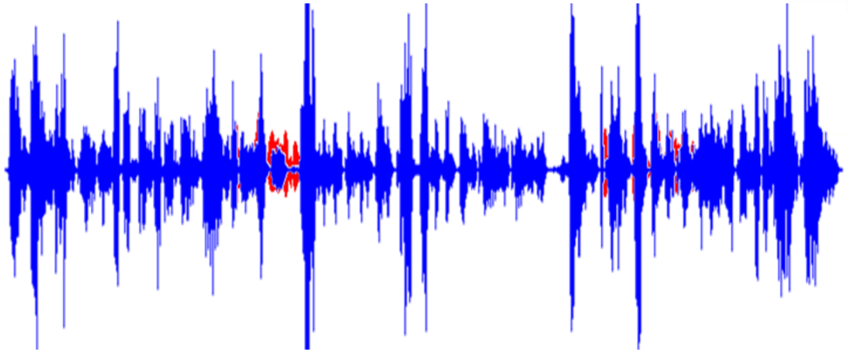
Articulation Index (AI)
Speech Intelligibility Index (SII)

- predicts SRTs in quiet very well
- combined effect of hearing loss and noise is difficult to predict

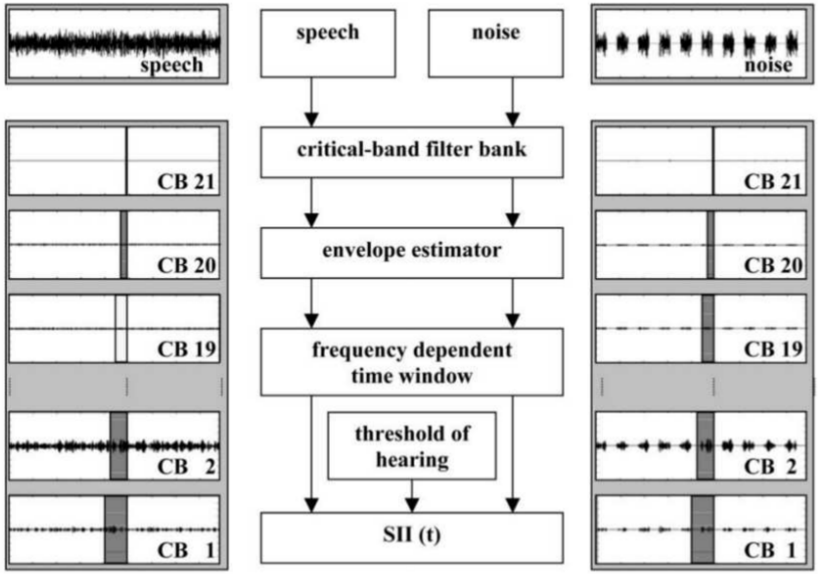




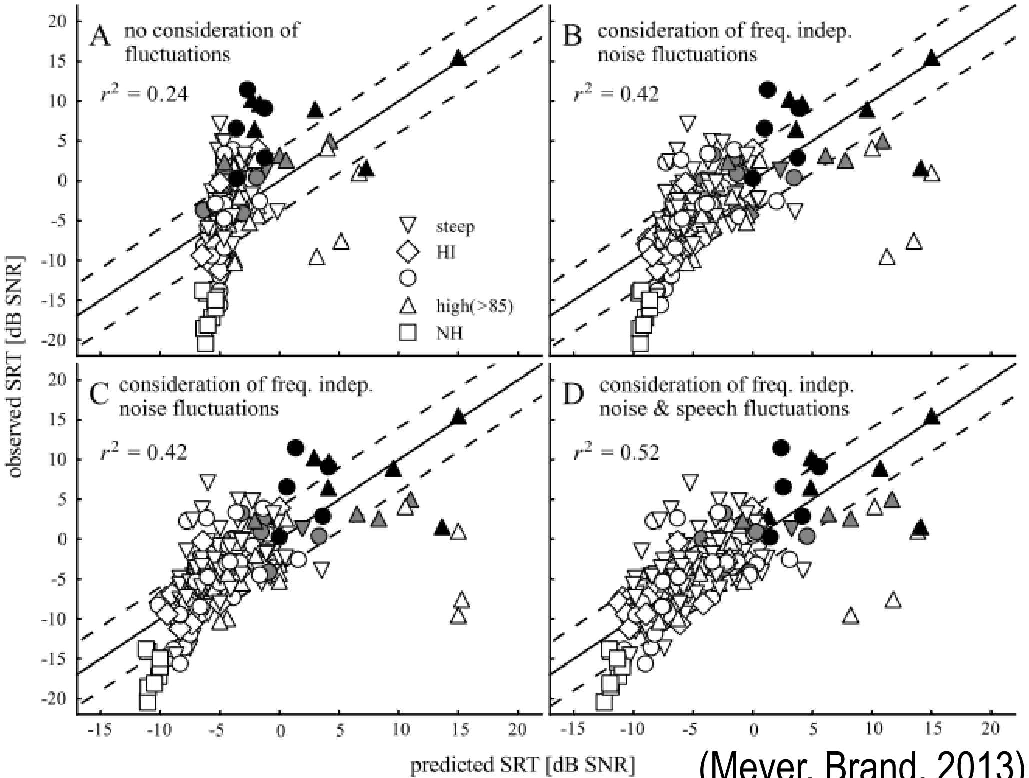
Extended SII (ESII)
short-term SII (stSII)



Temporal context:
"listening in the gaps"

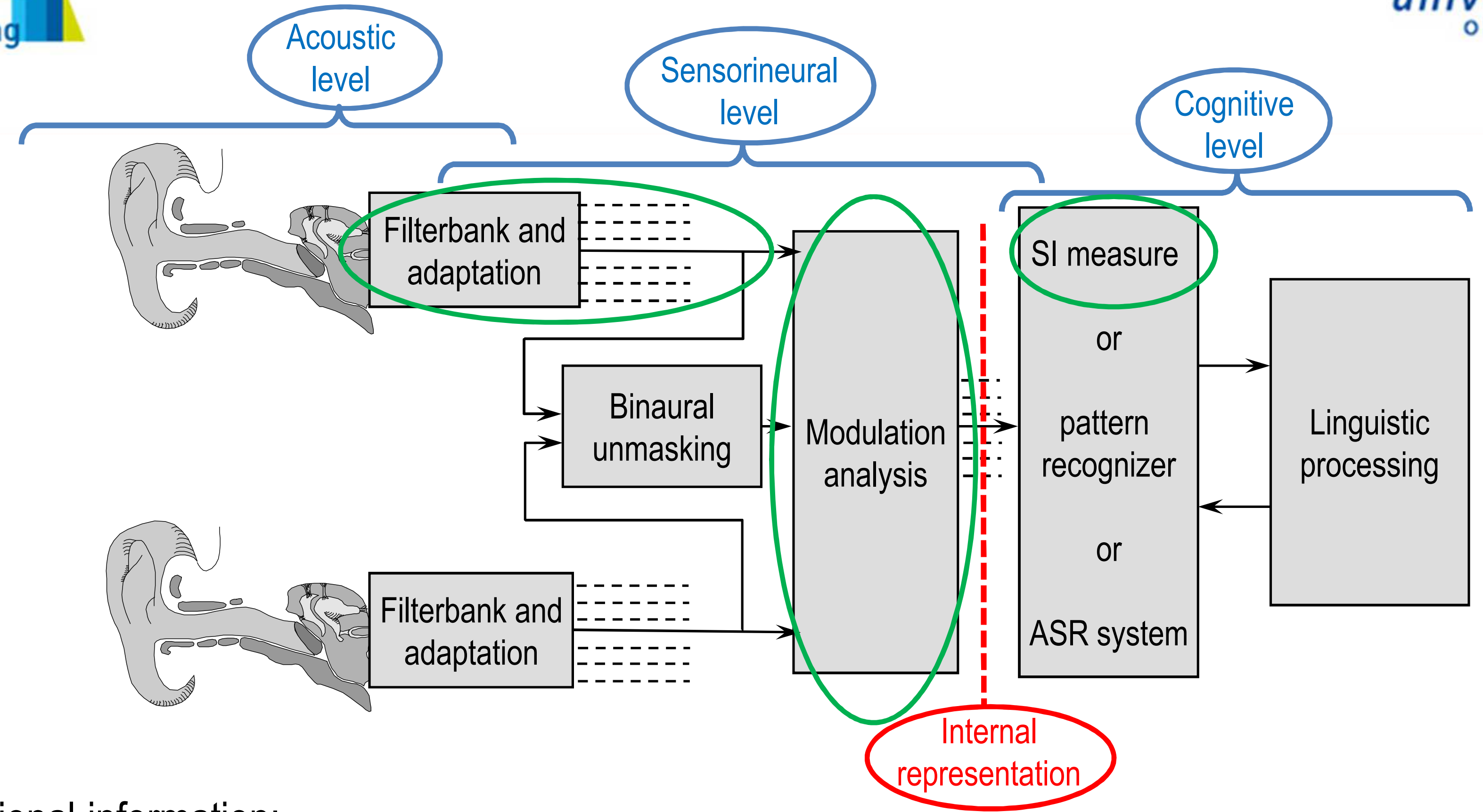


(Rhebergen, Versfeld, 2005)
(Rhebergen, Versfeld, Dreschler, 2006)



(Meyer, Brand, 2013)

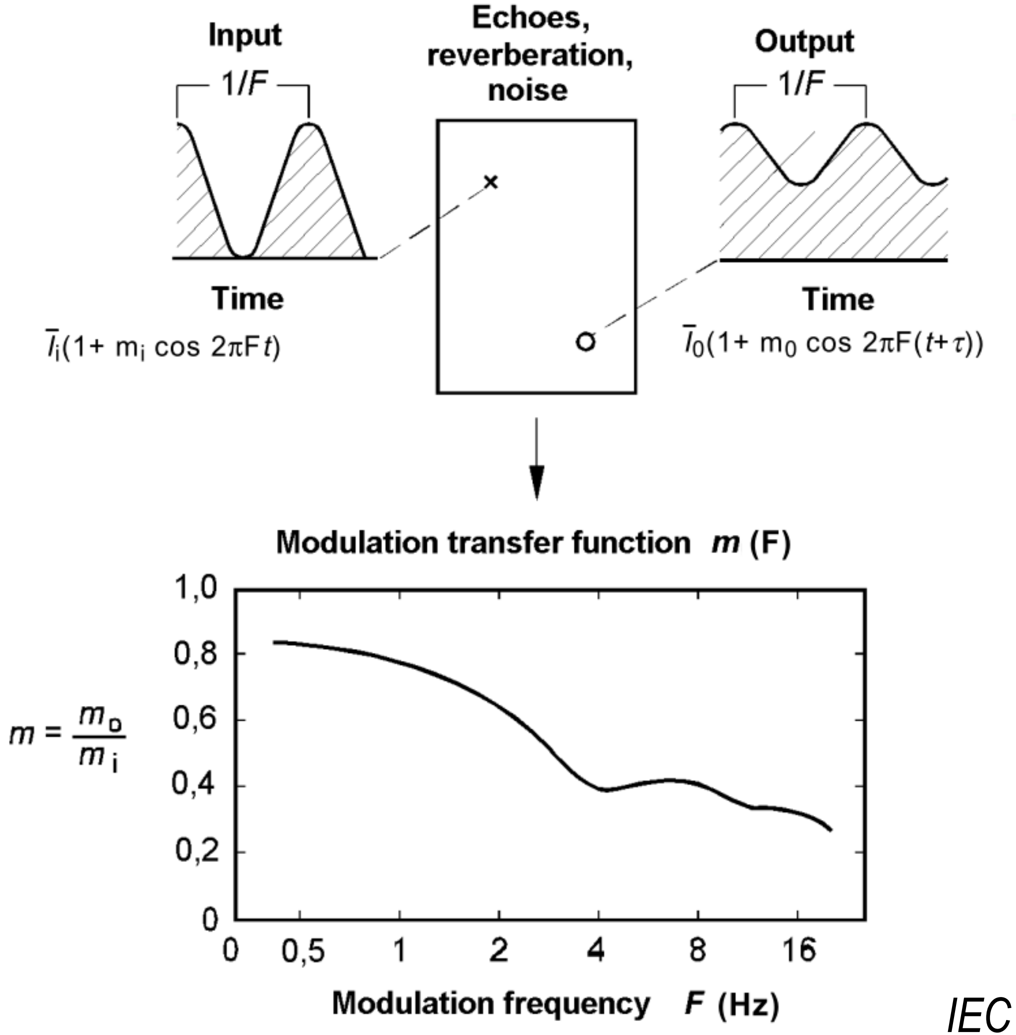
Speech Intelligibility Prediction (SIP)



Additional information:

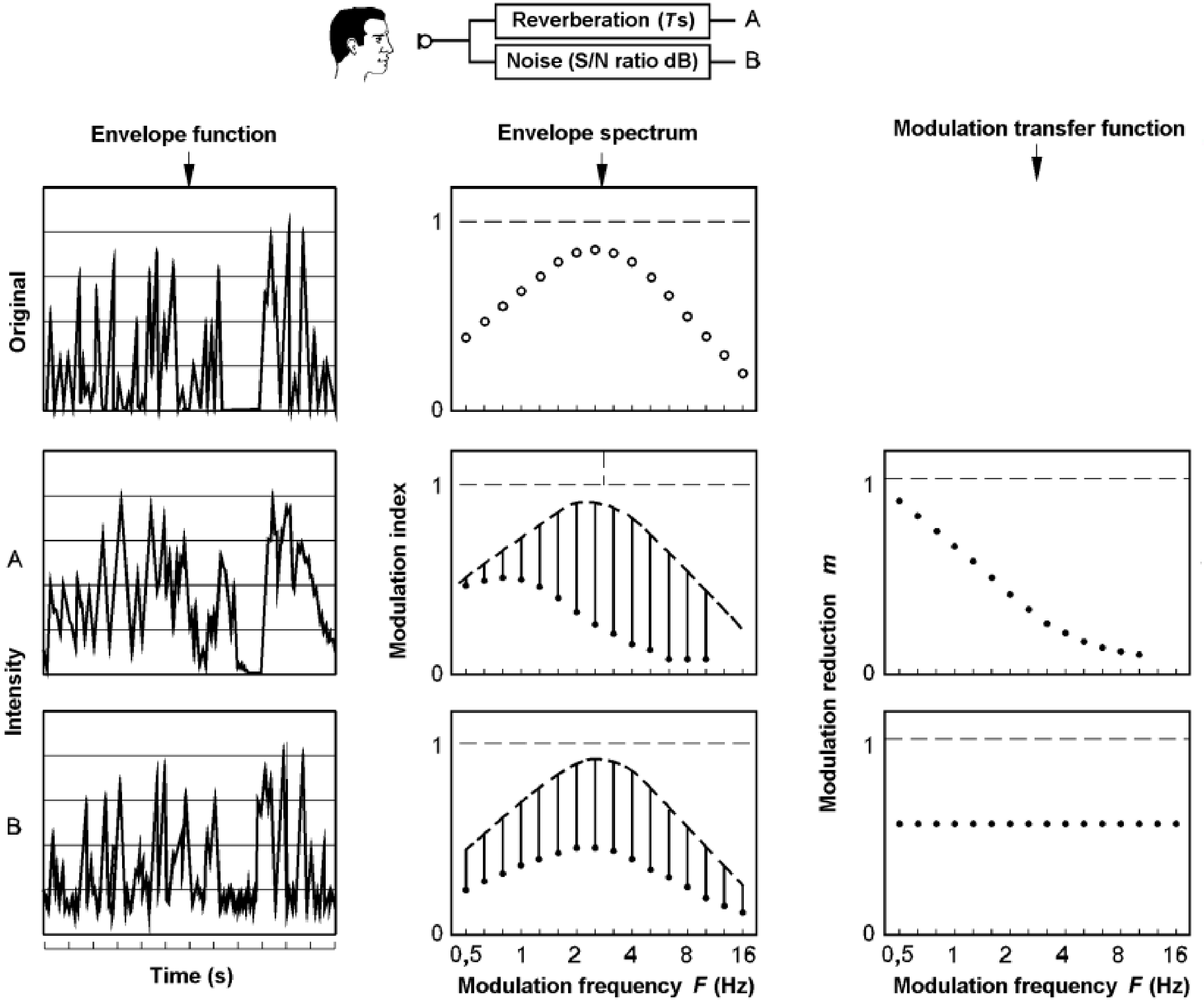
- Listener: hearing loss
- Signals: room acoustics, speech(+ noise), noise(+ speech)
- Linguistics: semantic context (reference curve)

Speech Transmission Index (STI)



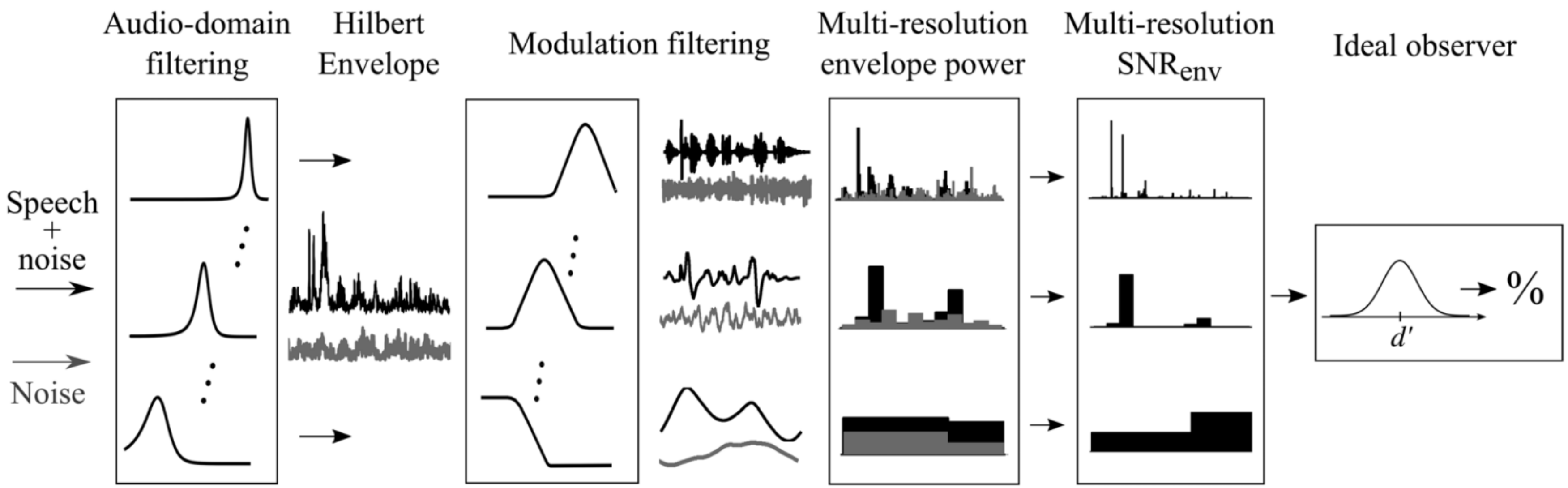
IEC 60268-16

Speech Transmission Index (STI)



IEC 60268-16

Multi resolution speech-based envelope power spectrum model (mr-sEPSM)



(Jørgensen, Ewert, Dau, 2013)

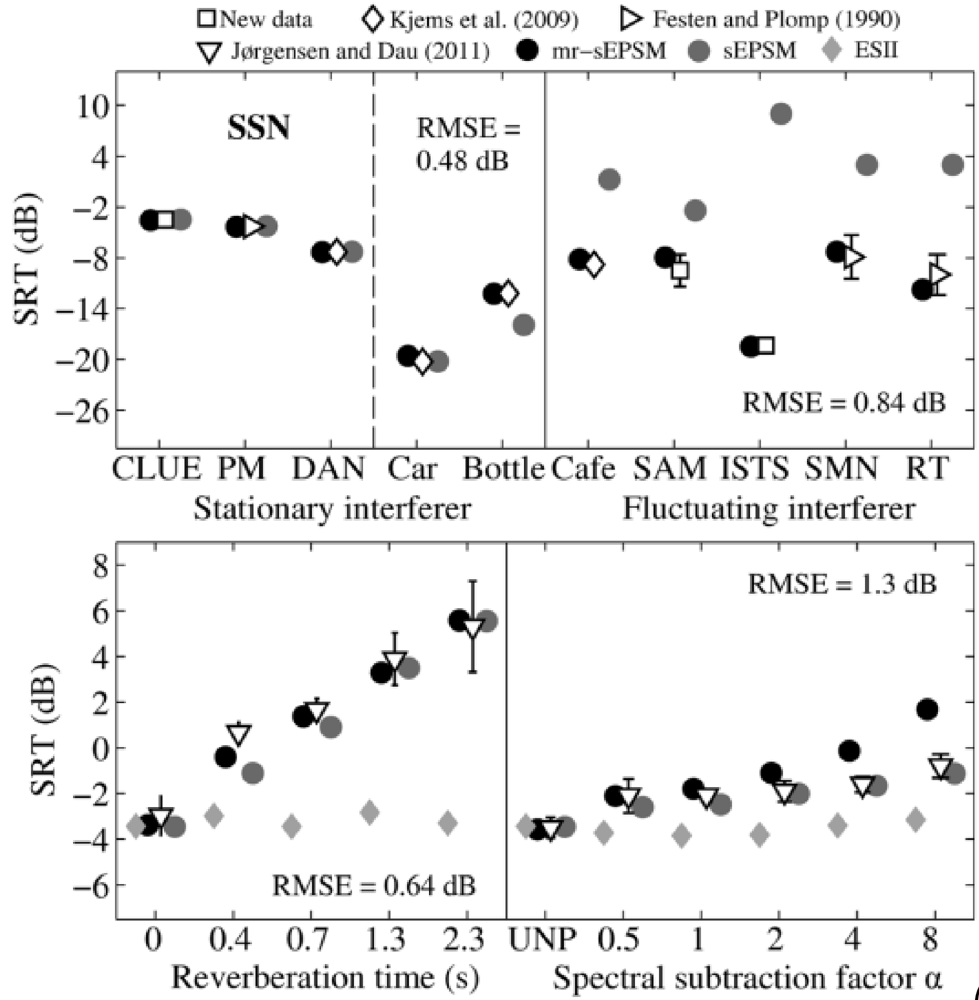
$$\mathrm{SNR}_{\mathrm{env}}(p) = \left[\sum_{n=1}^{9} \mathrm{SNR}_{\mathrm{env}}^{2}(p,n)\right]^{1/2}$$

$$d' = k(\mathrm{SNR}_{\mathrm{env}})^{q}$$

$$\mathrm{SNR}_{\mathrm{env}} = \left[\sum_{p=1}^{22} \mathrm{SNR}_{\mathrm{env}}^{2}(p)\right]^{1/2}$$

$$P_{\mathrm{correct}}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right)$$

Multi resolution speech-based envelope power spectrum model (mr-sEPSM)



(Jørgensen, Ewert, Dau, 2013)

Speech-to-Reverberation Modulation energy Ratio (SRMR) (Falk, Zheng, Chan, 2010)

- developed as objective measure for quality and intelligibility of (de)reverberated speech
- *non-intrusive*


www.spartanavenue.com
SPARTANS
5
ELITE

Speech-to-Reverberation Modulation energy Ratio (SRMR) (Falk, Zheng, Chan, 2010)

- Developed as objective measure for quality and intelligibility of (de)reverberated speech
- *non-intrusive:* no need of a reference signal

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4} \sum_{j=1}^{23} \sum_{m=1}^{M} \varepsilon_{j,k}(m)}{\sum_{k=5}^{8} \sum_{j=1}^{23} \sum_{m=1}^{M} \varepsilon_{j,k}(m)}$$

(Falk, Zheng, Chan, 2010)
(Santos, Senoussaoui, Falk, 2014)
(Senoussaoui, Santos, Falk, 2015)

Speech-to-Reverberation Modulation energy Ratio (SRMR)

Table 1. Performance results for the SRMR-based and benchmark metrics.

Metric	ρ_p	ρ_{sp}	ρ_{sig}	RSD%	RMSE
SRMR-based metrics					
SRMR	0.68	0.86	0.78	0.22	15.45
$\mathrm{SRMR}_{\mathrm{norm}}$	0.77	0.93	0.92	0.09	9.48
Benchmark metrics					
POLQA	0.68	0.94	0.94	0.09	7.81
NCM	0.57	0.72	0.53	0.15	22.98
CSII	0.51	0.71	0.46	0.26	23.80
STOI	0.44	0.77	0.36	0.08	23.24
PESQ	0.64	0.90	0.92	0.08	10.05
oPESQ	0.89	0.88	0.92	0.09	10.12
ANIQUE+	0.81	0.88	0.91	0.32	11.68
ModA	0.81	0.86	0.86	0.15	15.95
P.563	0.38	0.33	0.34	0.24	28.14

(Santos, Senoussaoui, Falk, 2014)

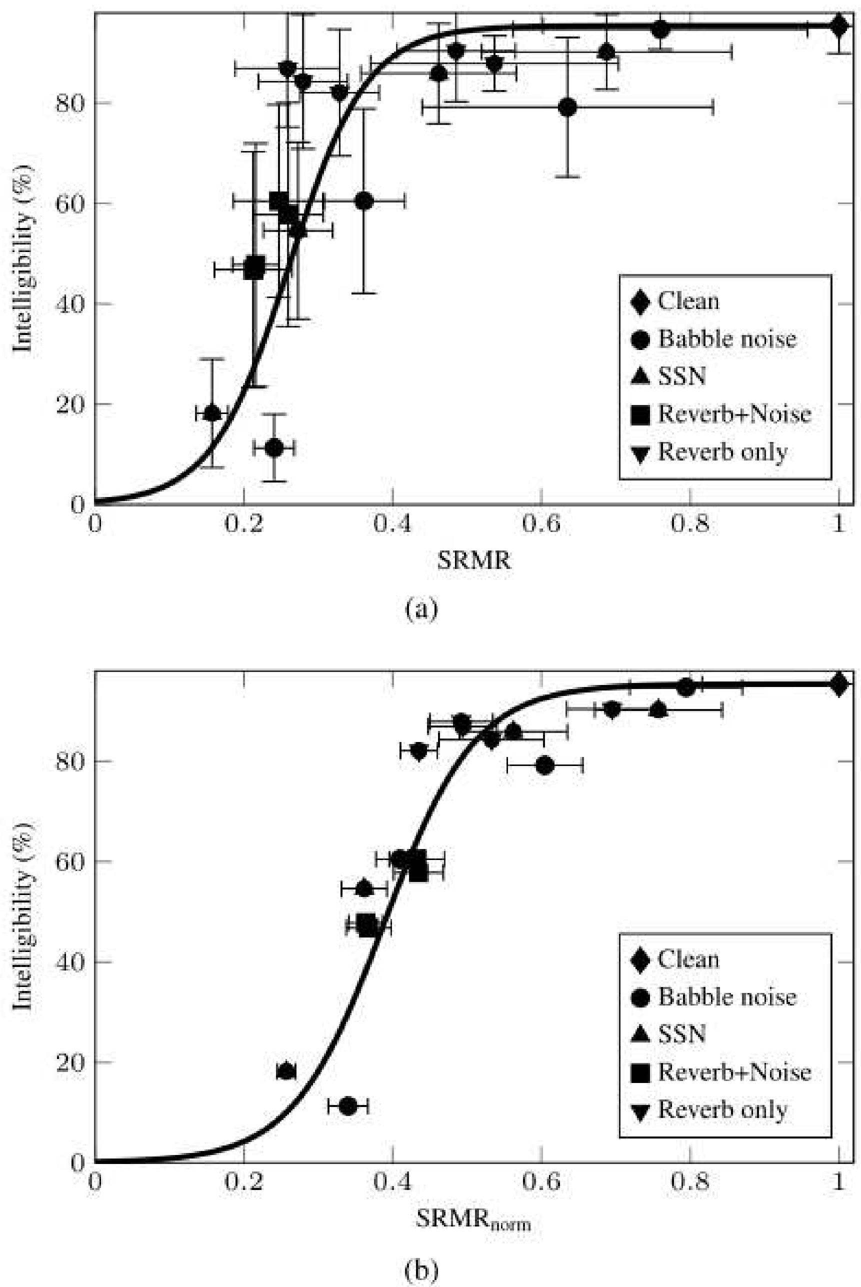
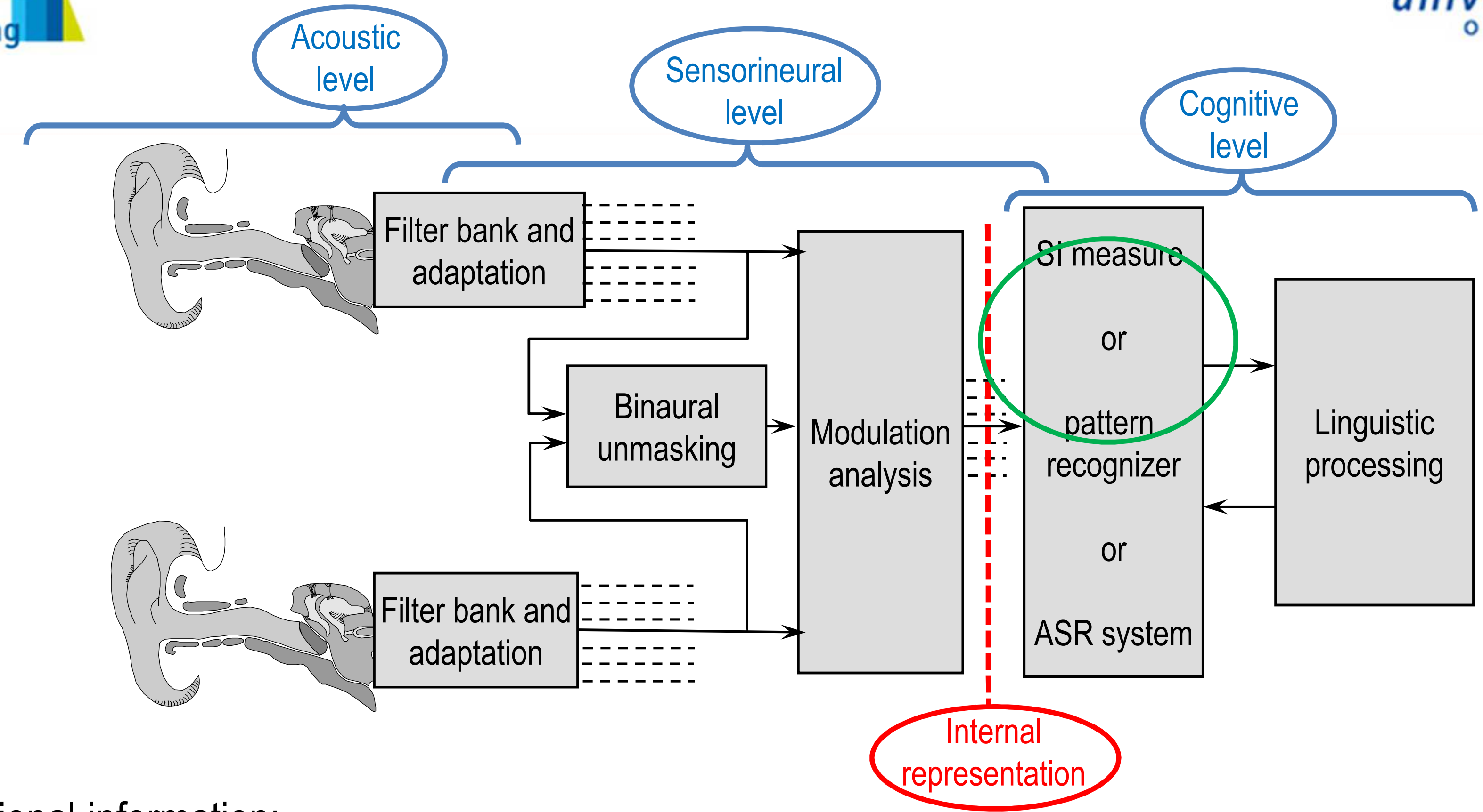


Fig. 3. Scatterplots for SRMR (a) and $\mathrm{SRMR}_{\mathrm{norm}}$ (b).

Speech Intelligibility Prediction (SIP)

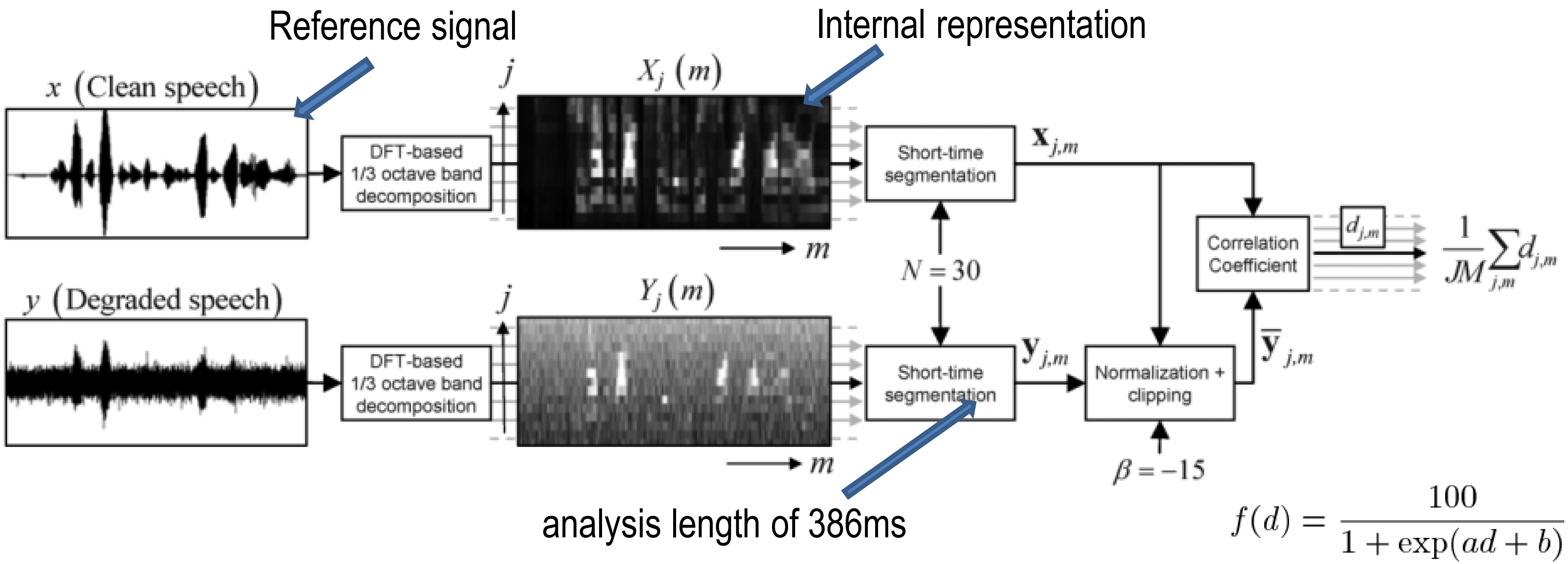


Additional information:

- Listener: hearing loss, linguistic abilities
- Signals: crosscorrelation between clean speech and degraded speech
- Linguistics: semantic context, …

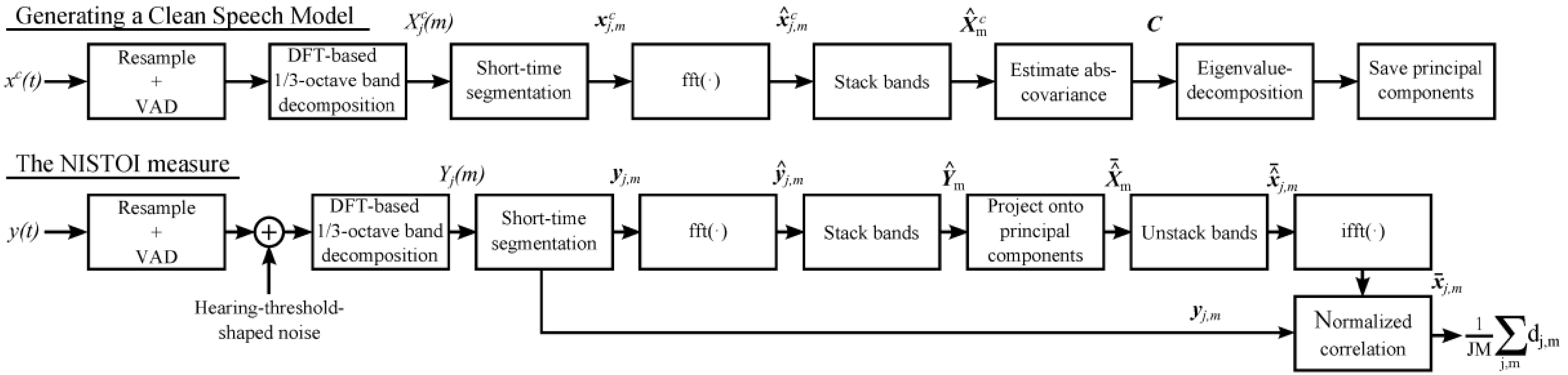
Short-time objective intelligibility measure (STOI) (Taal, Hendriks, Heusdens, Jensen, 2011)

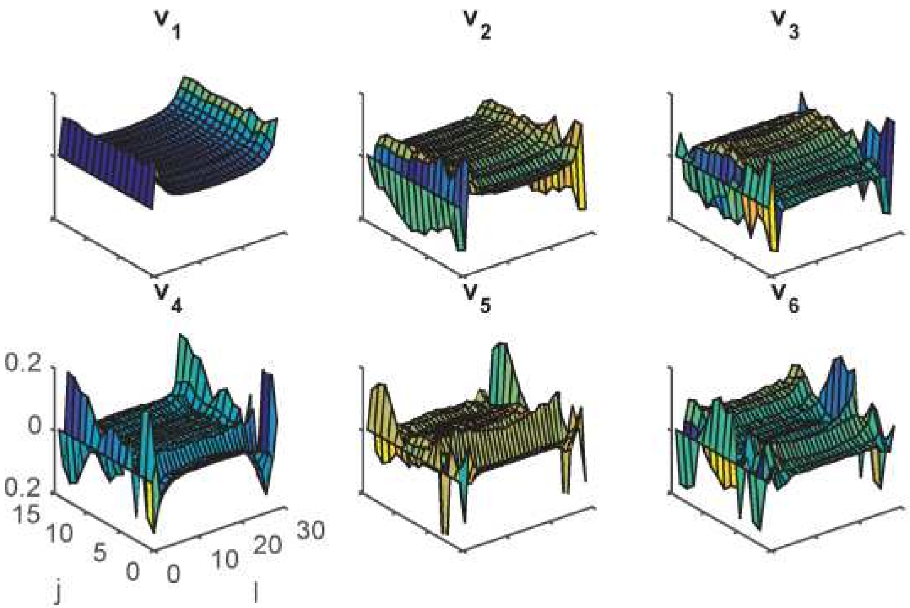
- Purpose: Prediction of the benefit of noise-reduction algorithms

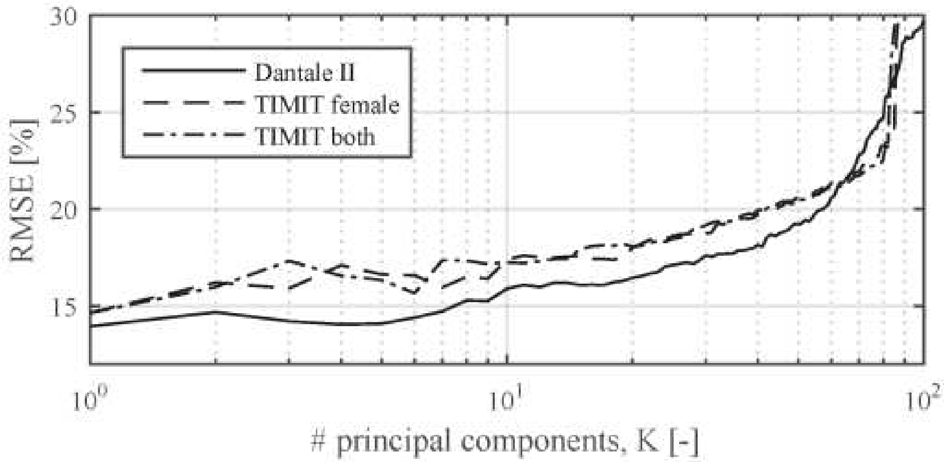


- Good prediction of benefit due to noise reduction (e.g. ideal binary masks, Ephraim-Malah), because the disadvantageous effect of distortion on speech is taken into account.

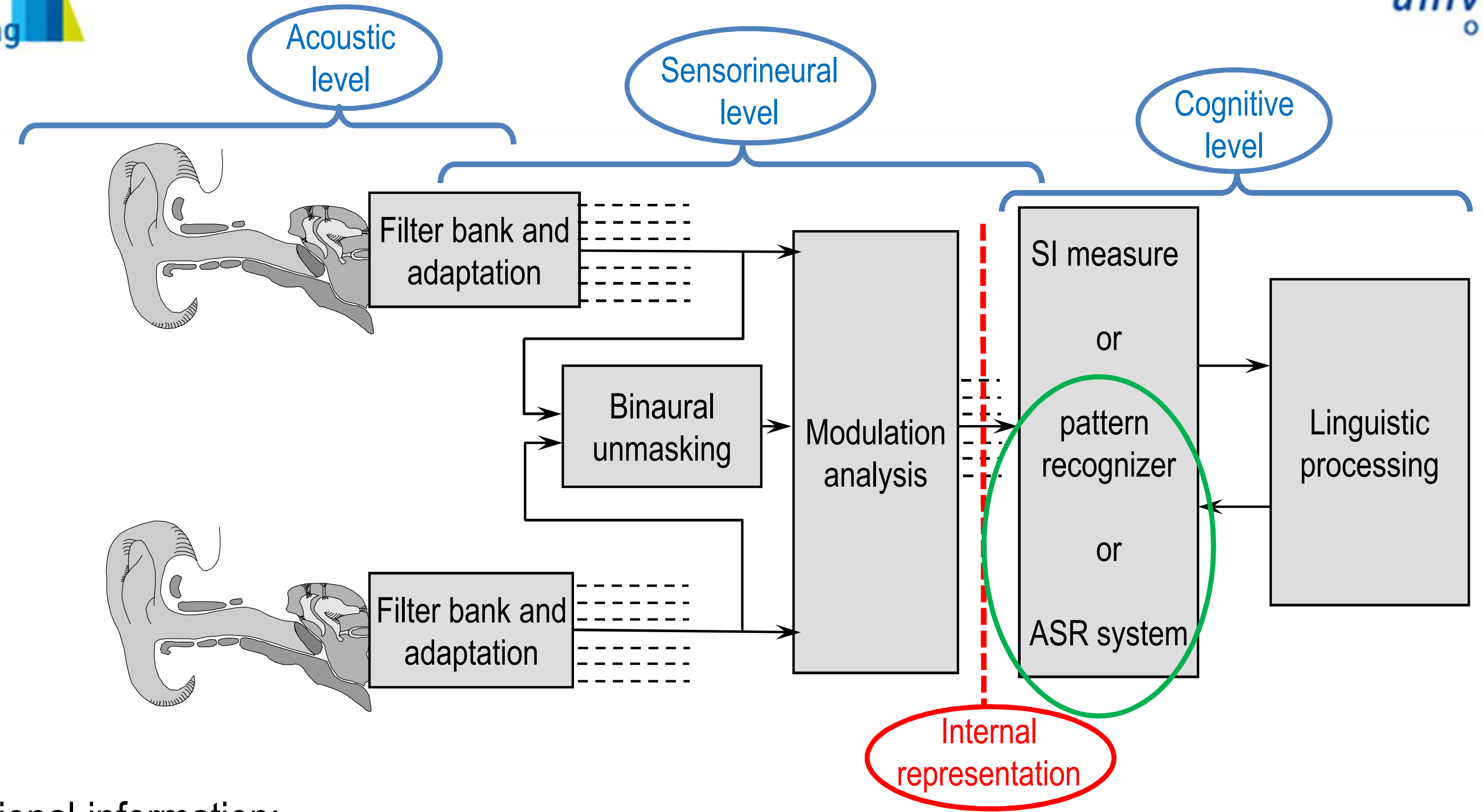
ESTOI (Jensen, Taal 2016)
NI-STOI (Andersen, de Haan, Tan, Jensen, 2017)







Speech Intelligibility Prediction (SIP)

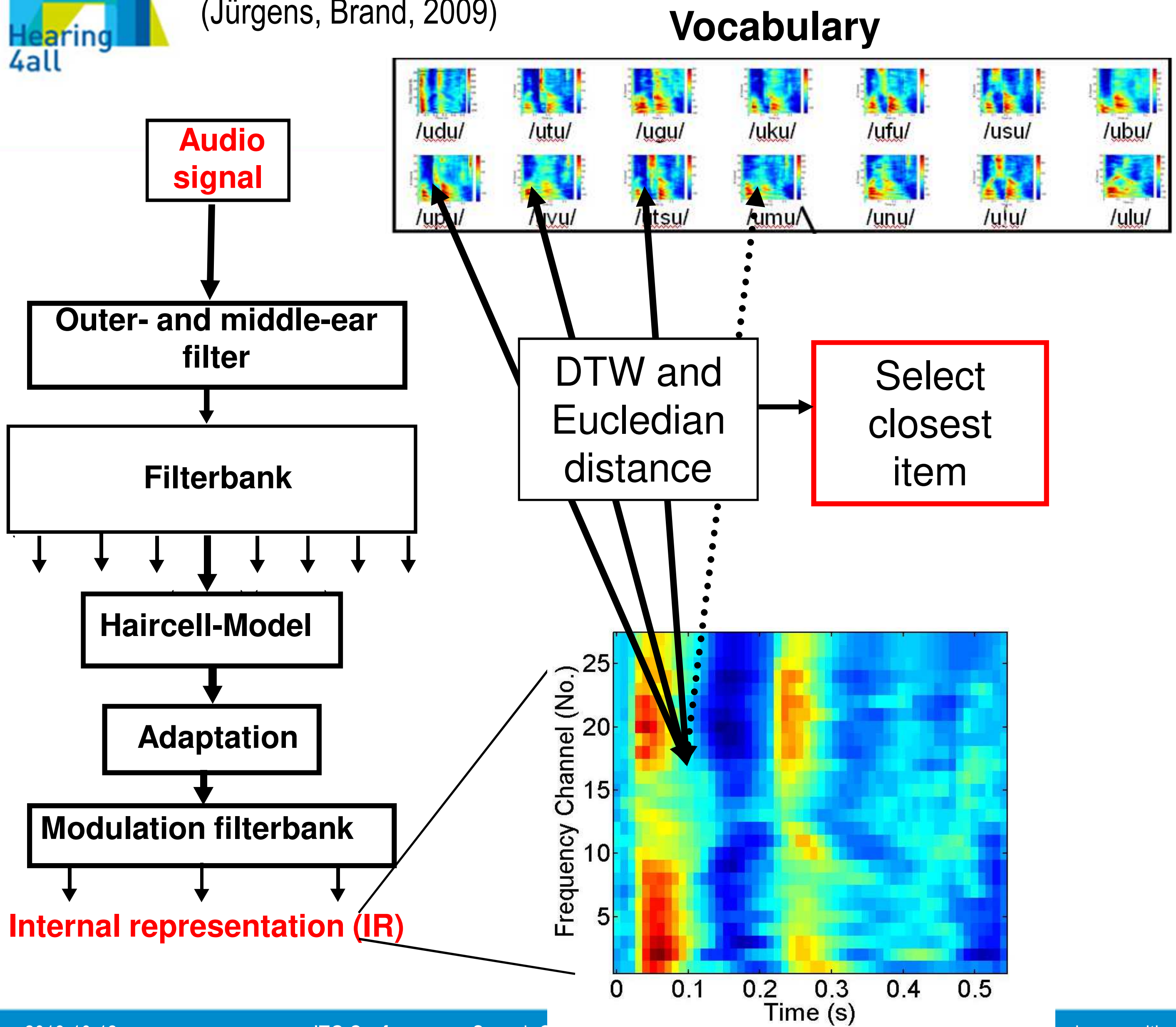


Additional information:

- Listener: hearing loss
- Signals: vocabulary or training material
- Linguistics: language model

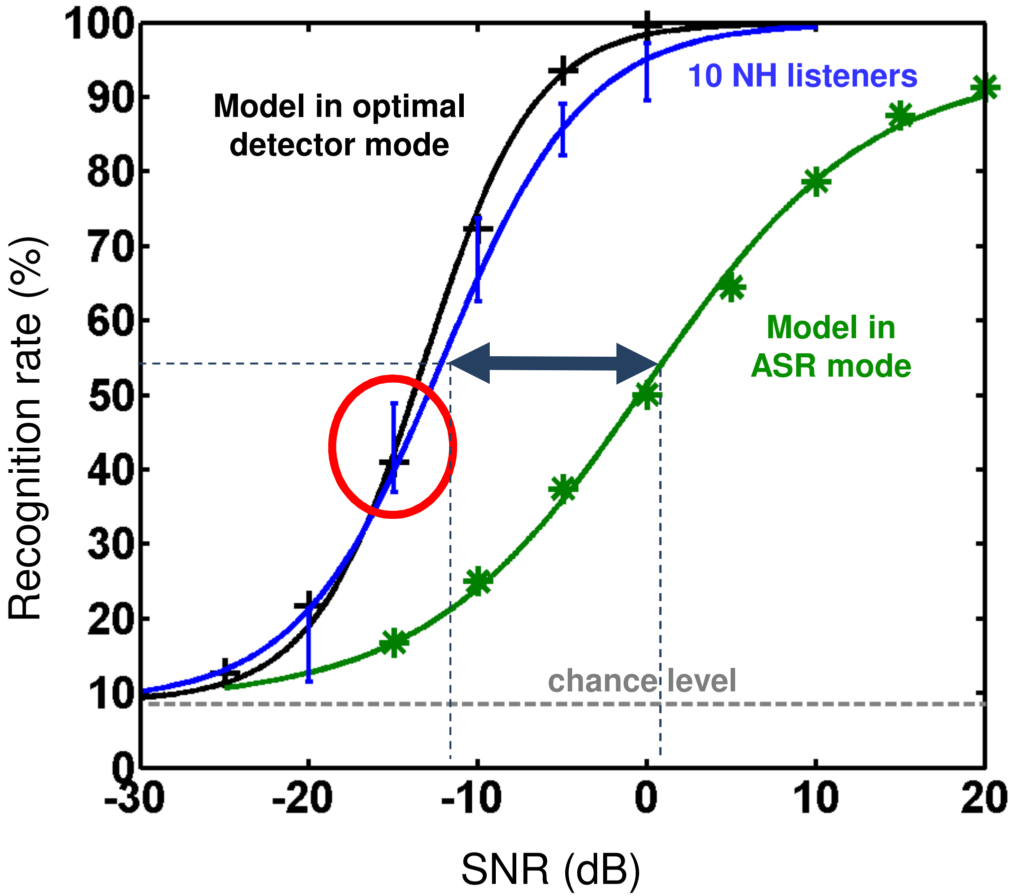
Auditory model and pattern recognizer

(Jürgens, Brand, 2009)

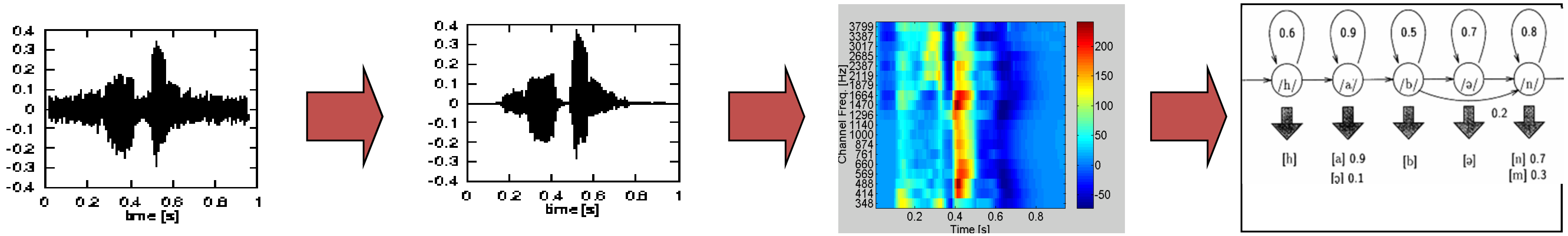


Auditory model and pattern recognizer

(Jürgens, Brand, 2009)

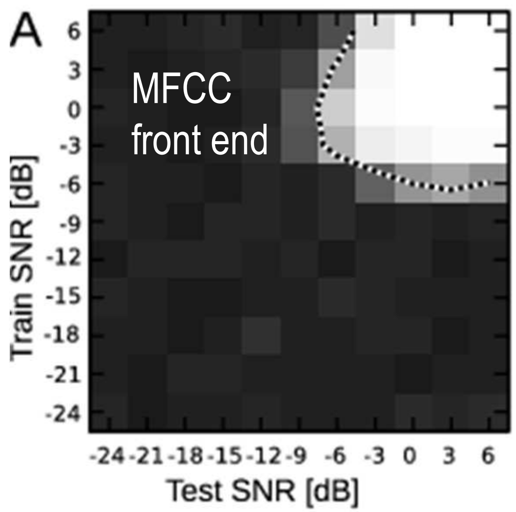


Framework for Auditory Discrimination Experiment simulation (FADE) (Schaedler et al. …)



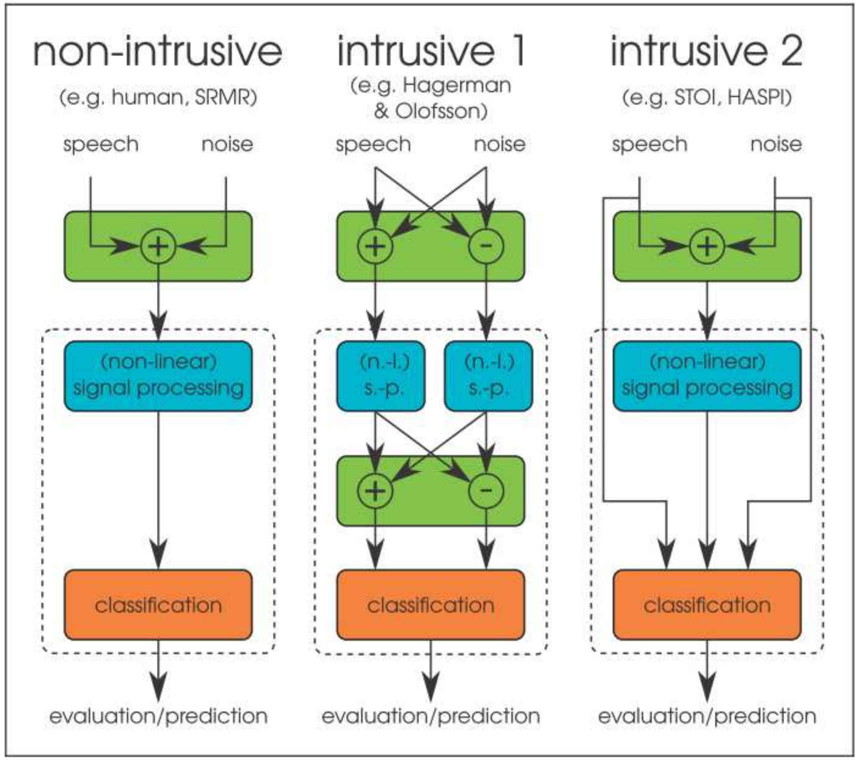
Time signal Preprocessing Feature extraction Classification (HMM)

- Applicable as well to other discrimination experiments
- System trained at different SNRs with matrix sentence set
- Select training SNR with lowest SRT prediction
- **visit Poster P6:** David Hülsmeier: FADE with everyday sentences

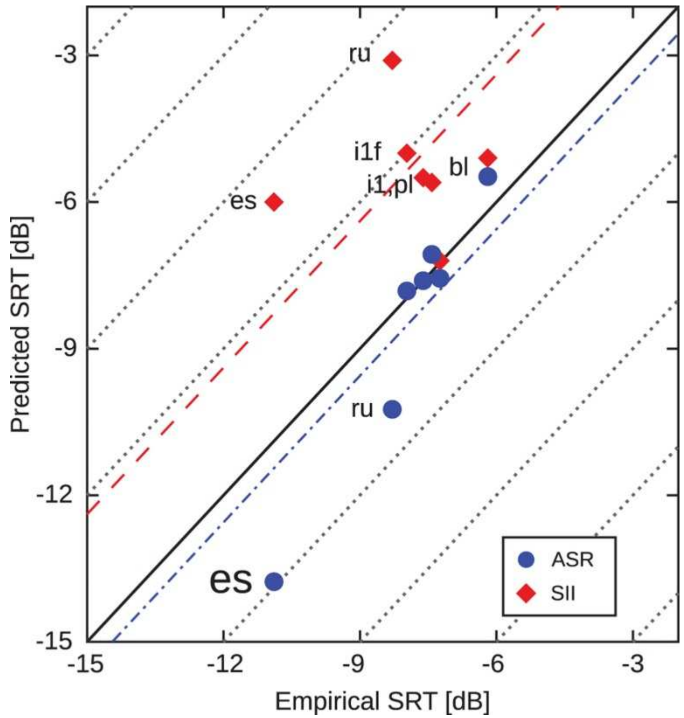


(Schaedler, Warzybok, Ewert, Kollmeier, 2016)

Framework for Auditory Discrimination Experiment simulation (FADE)

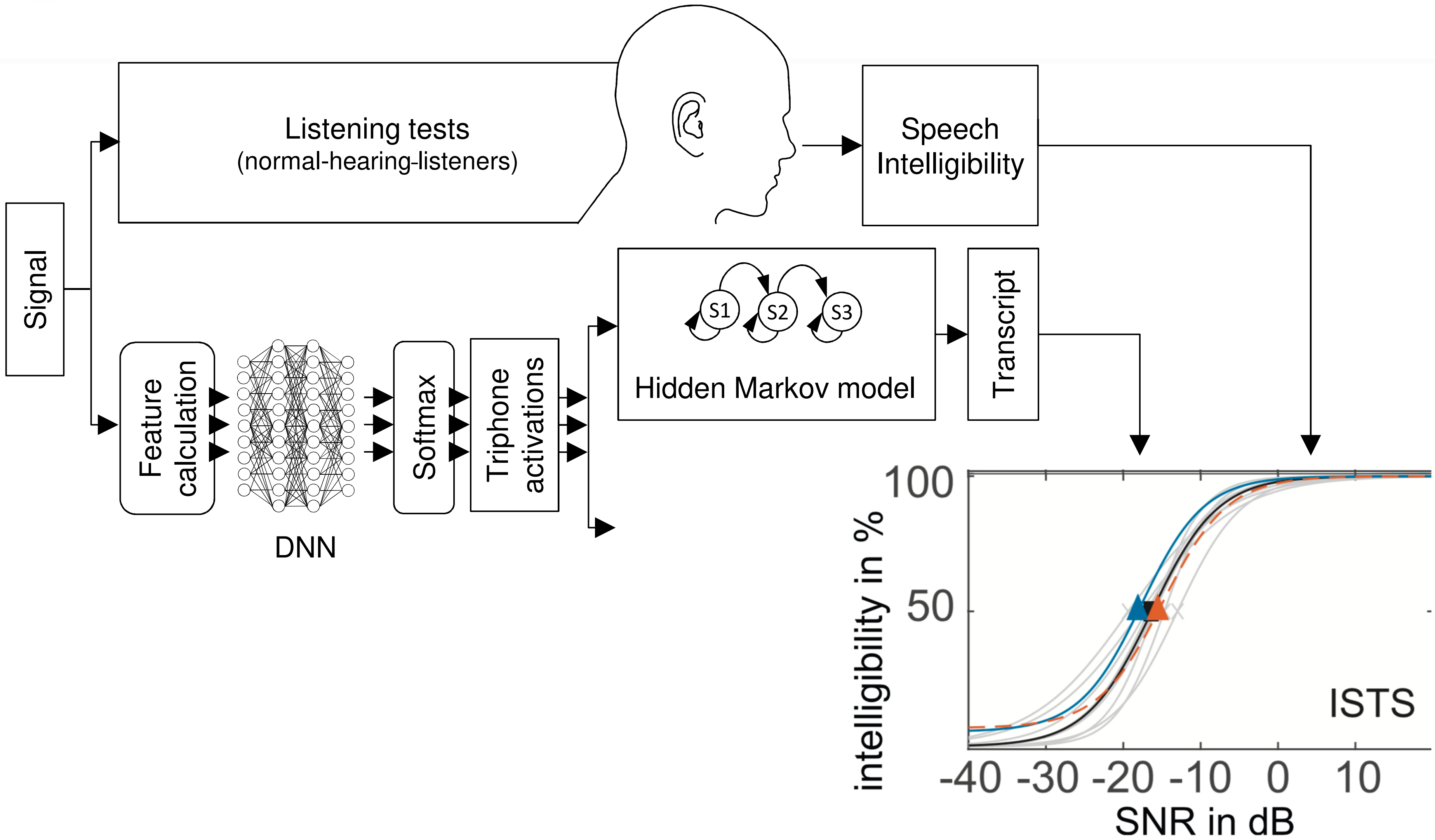


(Schädler, Warzybok, Kollmeier, 2018)

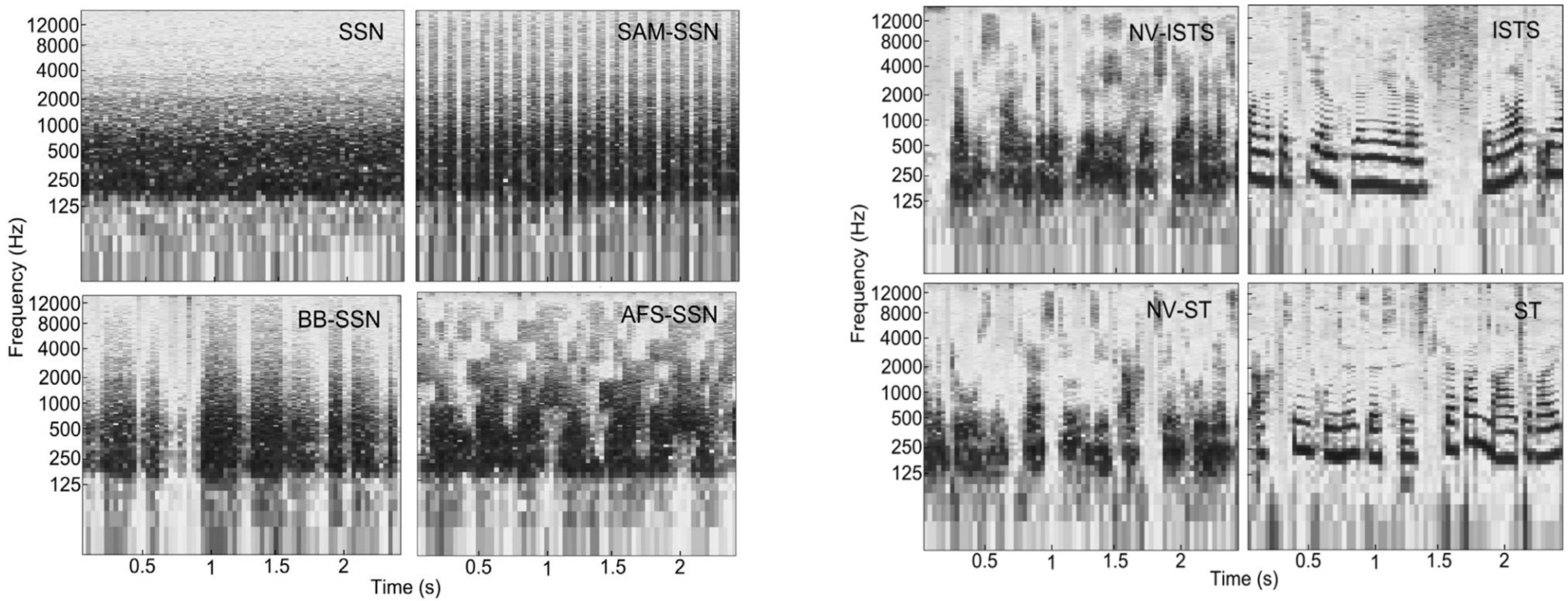


(Schädler, Warzybok, Hochmuth, Kollmeier, 2015)

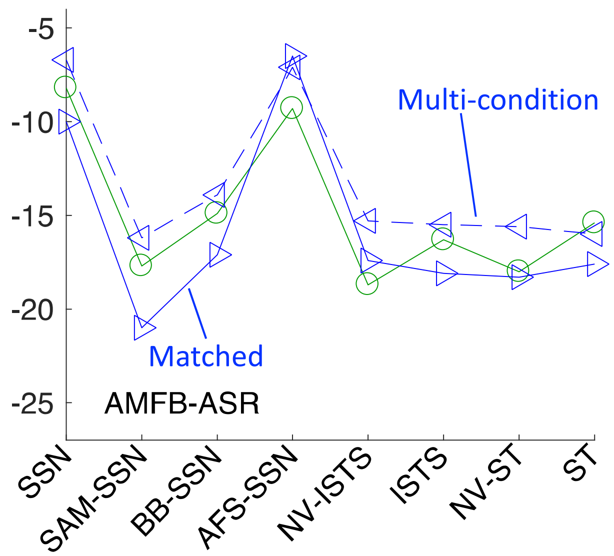
Using deep learning for prediction of speech intelligibility (Spille, Ewert, Kollmeier, Meyer, 2018)



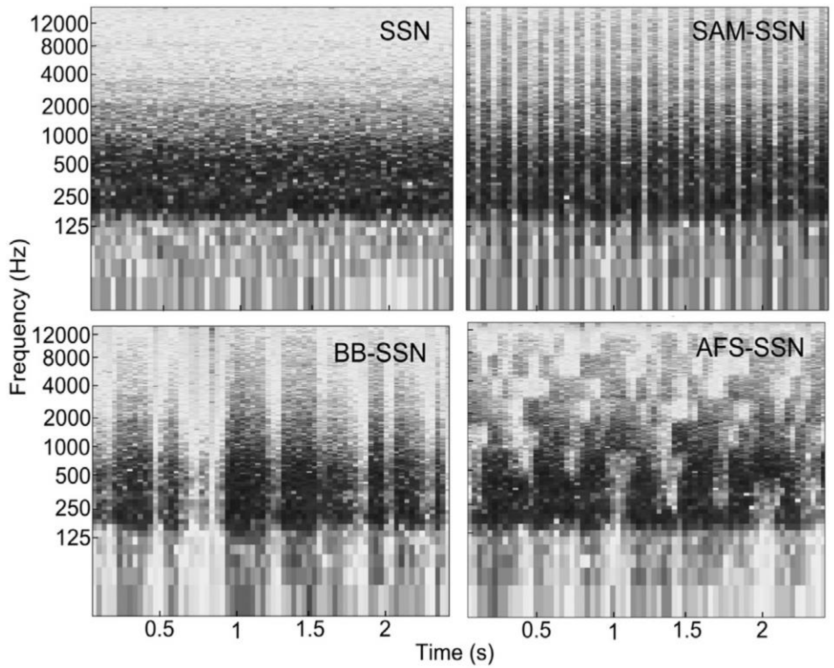
Evaluation with modulated and speech-like interferers (Schubotz, Brand, Kollmeier, Ewert, 2016)

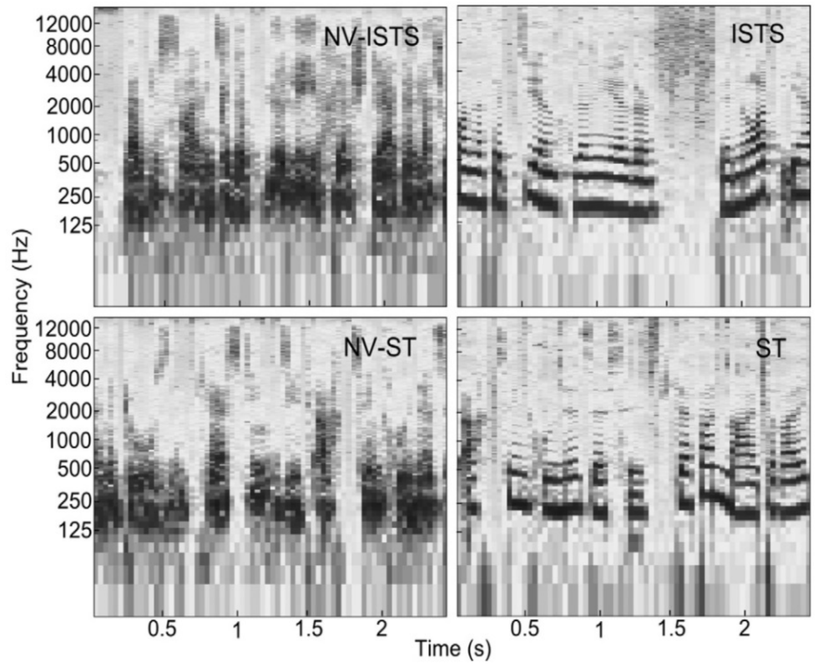


(Spille, Ewert, Kollmeier, Meyer, 2018)



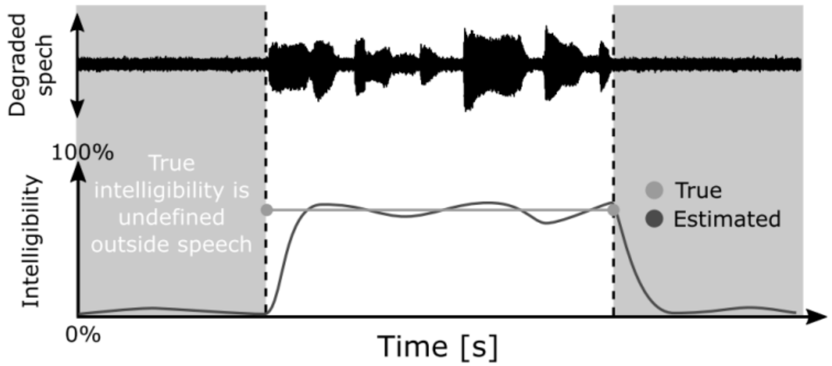
(Schubotz, Brand, Kollmeier, Ewert, 2016)



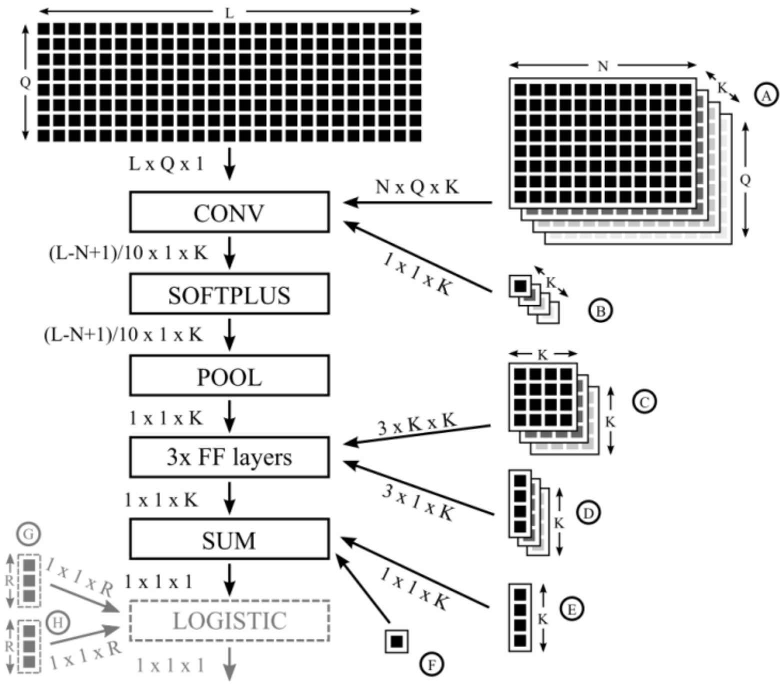


Non-intrusive SIP using Convolutional Neural Networks (CNN)

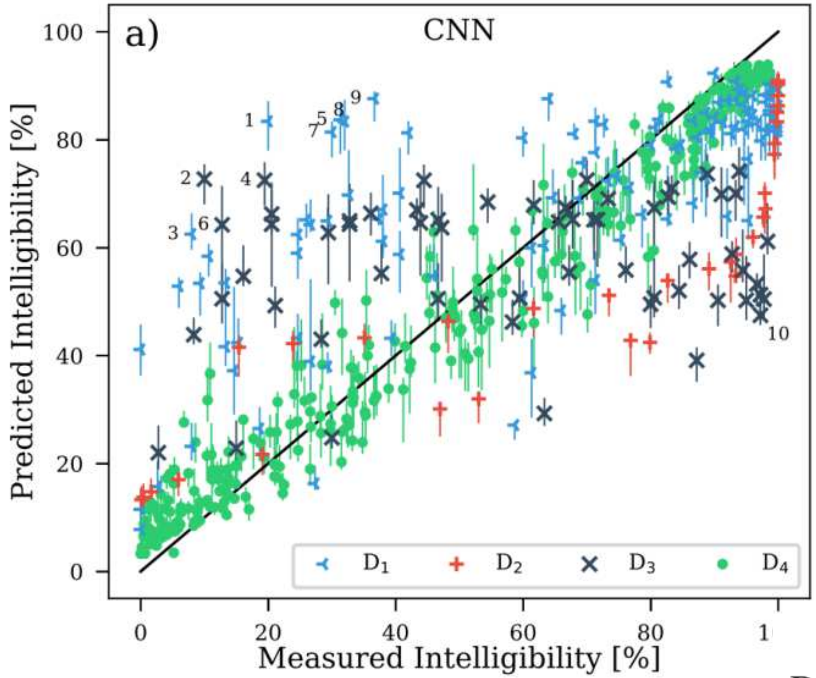
Hypothesis: SIP is a simple problem compared to ASR and can be solved by comparatively smaller neural networks and less training material.



(Andersen, de Haan, Tan, Jensen, 2018)



Non-intrusive SIP using Convolutional Neural Networks (CNN)



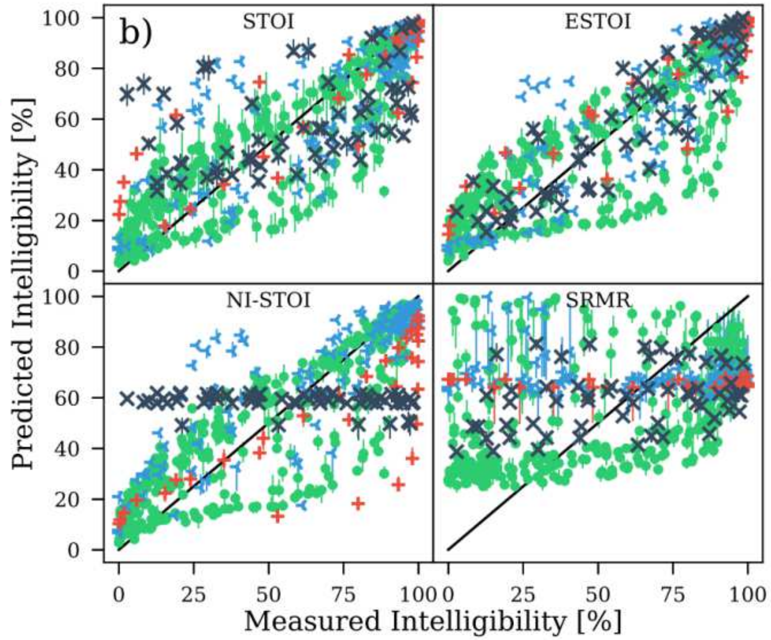


Fig. 4. The median of predictions for the proposed CNN, and for four other SIP algorithms, plotted against corresponding measured intelligibility. The error bars show the 25th and 75th percentiles of predictions. Colors/symbols indicate which dataset each condition belongs to. For the proposed method, the ten conditions with the largest absolute prediction errors are numbered in descending order. Descriptions of these are given in Table II.

(Andersen, de Haan, Tan, Jensen, 2018)

TABLE III
PERFORMANCE METRICS FOR THE FIVE SIP ALGORITHMS.

SIP algorithm	RMSE	Kendall's Tau
CNN	17.69 pp	0.667
STOI	18.94 pp	0.658
ESTOI	17.11 pp	0.692
NI-STOI	19.90 pp	0.629
SRMR	32.77 pp	0.281

Non-intrusive SIP using Convolutional Neural Networks (CNN)

TABLE V
PREDICTED INTELLIGIBILITY FOR VARIOUS NOISE RECORDINGS FROM THE NOISE-X DATABASE [68]. PREDICTIONS WERE MADE USING THE LOGISTIC FUNCTION ASSOCIATED WITH DATASET D_4.

Recording	Median prediction
White noise	1.2%
Pink noise	1.2%
HF channel	1.3%
Jet cockpit 1	1.3%
Jet cockpit 2	1.4%
Car interior	1.8%
F16 cockpit	1.8%
Destroyer engine room	2.0%
Factory floor 2	2.8%
Tank noise	2.9%
Military vehicle	3.3%
Destroyer operations room	7.5%
Factory floor 1	22.3%
Machine gun	38.6%
Speech babble	41.4%

(Andersen, de Haan, Tan, Jensen, 2018)

Conclusions

- Simple intrusive SIP models explain a lot.
- Relatively simple non-intrusive models are available.
- ASR models are applicable to many situations.
- Evaluation measurements are still strongly recommended.
- Challenges: speech maskers, …
- **Visit poster P5** (Christopher Hauth: Performance prediction of binaural MVDR beamformer using BSIM)
- **Visit poster P6** (David Hülsmeier: Extension of FADE for everyday sentences)

Many thanks for your attention!