Speech intelligibility prediction

Cassia Valentini

Enrich project meeting - October 2017

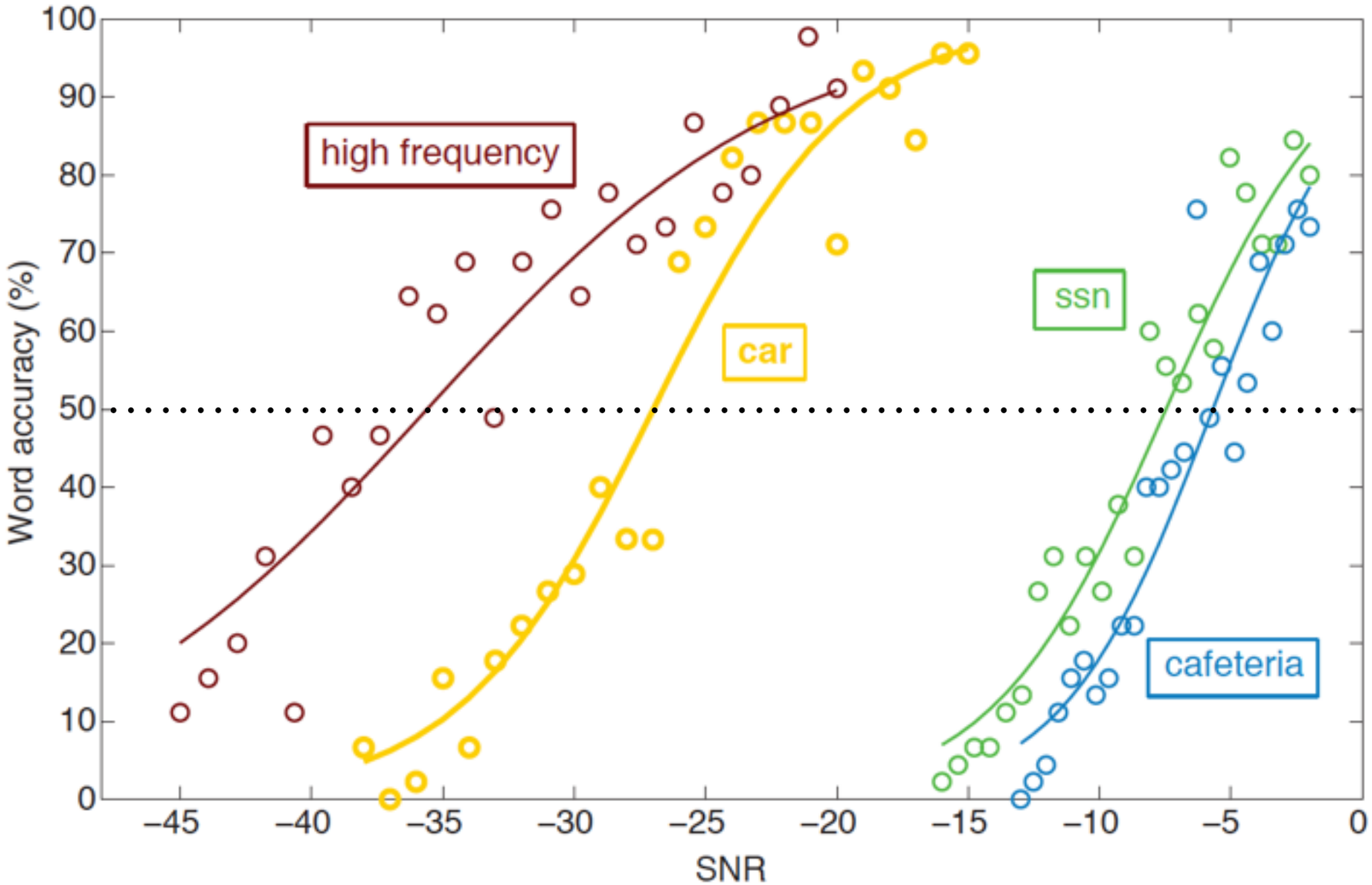
Outline

- Measures
- Evaluation
- Application

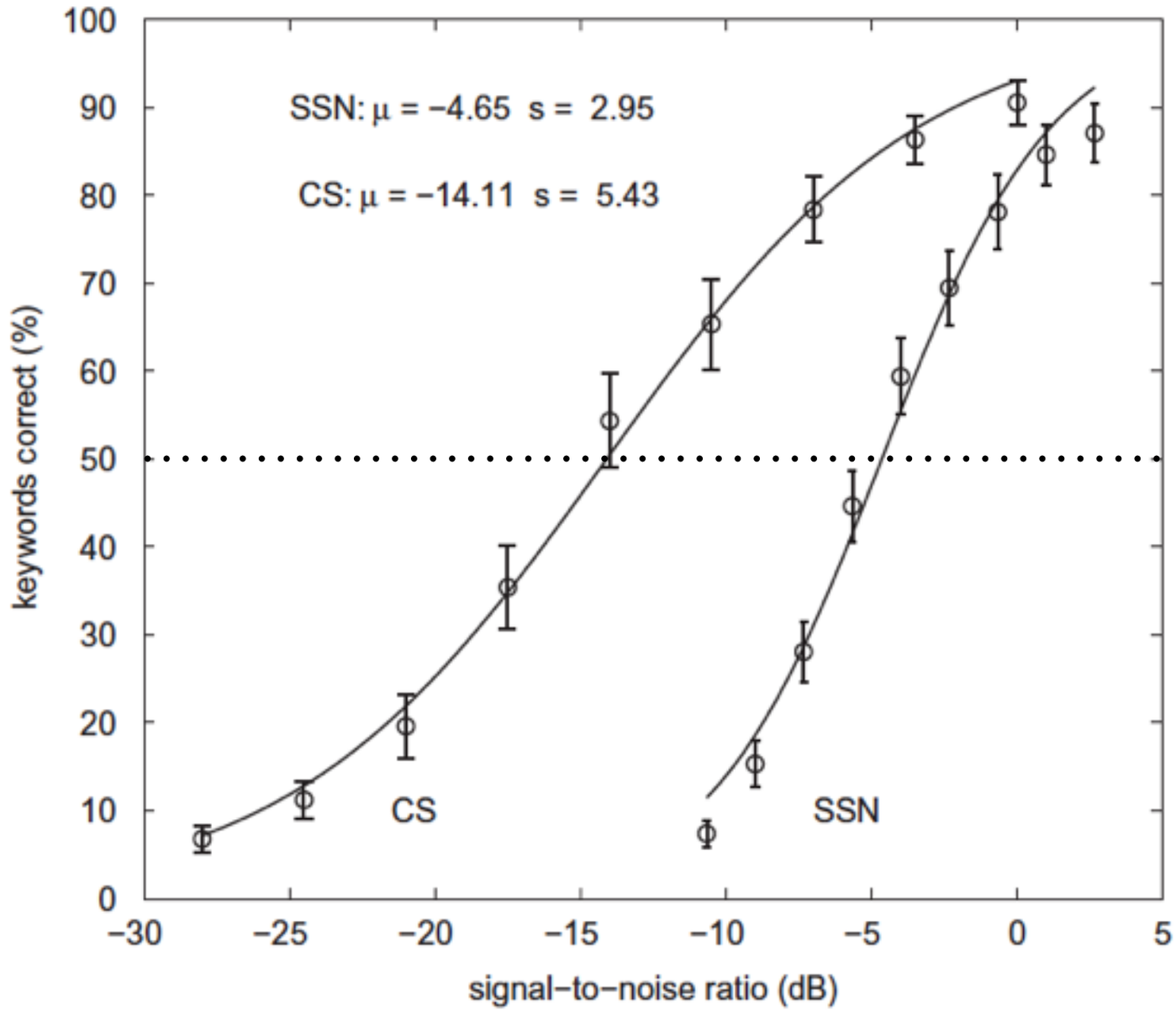
Why?

- Subjective evaluations are:
  - costly (time/money)
  - can only cover a limited amount of conditions (sentences/noises/speakers)
  - not easy to reproduce
- Objective scores are relatively easy to compute, can cover a wider range of conditions and are reproducible
- They can help the development of intelligibility enhancement techniques

Why?


high frequency
car
ssn
cafeteria
Word accuracy (%)
SNR
100
90
80
70
60
50
40
30
20
10
0
−45
−40
−35
−30
−25
−20
−15
−10
−5
0

Why?


SSN: μ = −4.65 s = 2.95
CS: μ = −14.11 s = 5.43
keywords correct (%)
signal−to−noise ratio (dB)
CS
SSN
100
90
80
70
60
50
40
30
20
10
0
−30
−25
−20
−15
−10
−5
0
5

Measures

Existing measures

- **CEP** - cepstral distortion (1976)
- **IS** - Itakura Saito distance (1976)
- **LLR** - log likelihood ration (1976)
- **LSD** - log spectral distortion (1976)
- **FWS** - frequency weighted SNR (1978)
- **WSS** - weighted spectral slope (1982)
- **AI** - articulation index (1962)
- **SII** - speech intelligibility index (1997)
- **ESII** - extended SII (2005)
- **CSII** - coherence based SII (2005)
- **STI** - speech transmission index (1980)
- **NCM** - normalised covariance measure (1996)
- **PESQ** - Perceptual evaluation of speech quality (2001)
- **GP** - glimpse distortion measure (2006)
- **DWGP** - distortion weighted GP (2014)
- **STOI** - short term objective intelligibility (2010)
- **CD** - Christiansen measure (2010)
- **SRMR** - speech-to-reverberation modulation energy ratio (2010)
- **SRMR-CI** - SRMR for cochlear implant (2013)
- **P.563** - Single-ended method for objective speech quality assessment in narrow-band telephony applications (2004)
- **NISA** - non intrusive speech intelligibility (2016)
- **BiDWGP** - binaural DWGP (2016)
- **BiSII** - binaural SII (1983)
- **BiSTI** - binaural STI (2008)
- **BiSMI** - binaural speech intelligibility model, based on SRMR (2013)

Classification

- Macroscopic / Microscopic
- Audibility-based / Distortion-based
- Intrusive / Non-intrusive
- Monaural / Binaural

Classification

- **Macroscopic** / Microscopic
- Audibility-based / Distortion-based
- Intrusive / Non-intrusive
- Monaural / Binaural

Classification

- **Macroscopic** / Microscopic
- **Audibility-based / Distortion-based**
- Intrusive / Non-intrusive
- Monaural / Binaural

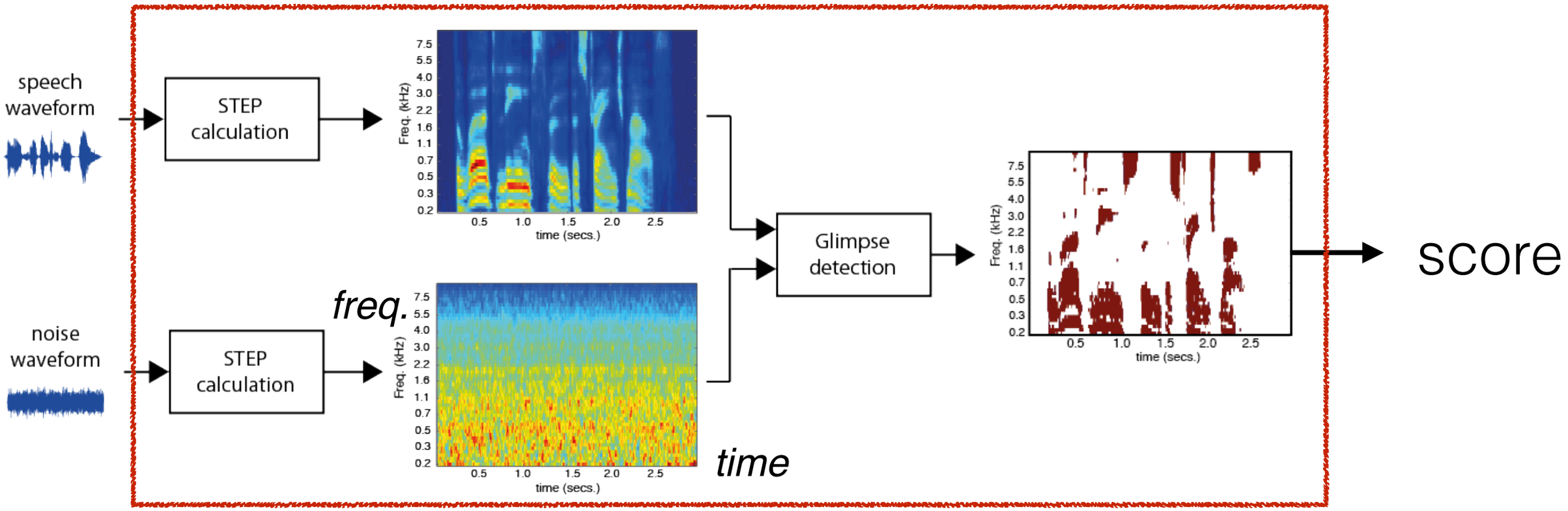
Audibility-based measures

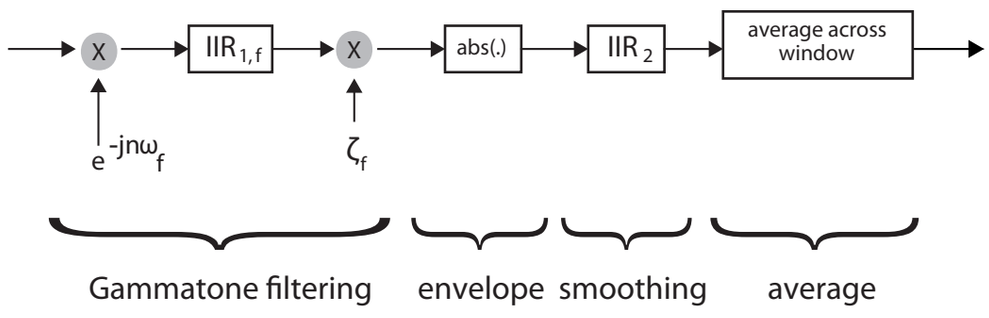
- Audibility measures calculate how audible speech is given an additive source of noise

- SNR-based measures are in this category

GP [Cooke, 2006]

Glimpse proportion measure





STEP calculation

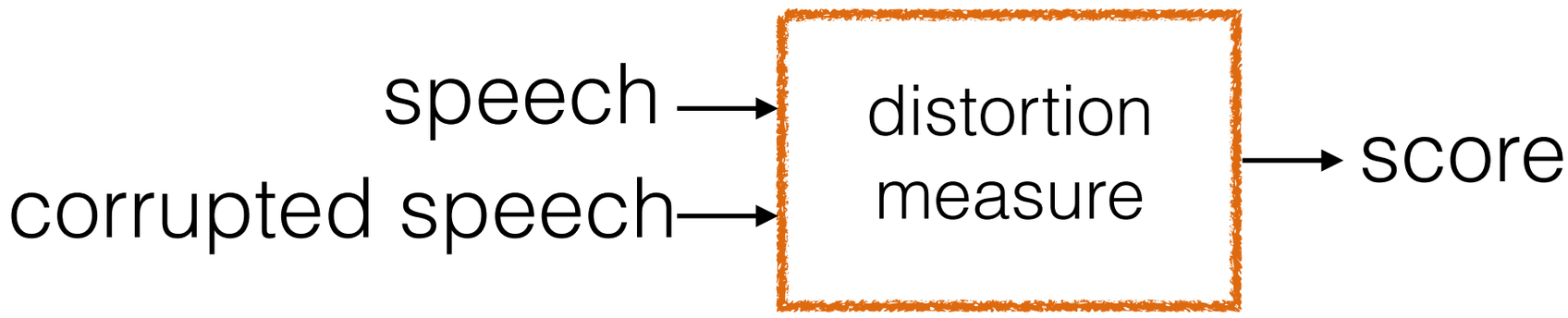
$$GP = \frac{100}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} \mathcal{H}(S_f(t) > (N_f(t) + \alpha))$$

SpectroTemporal Excitation Pattern

Distortion-based measures

- Distortion-based measures calculate how different distorted speech is given a speech reference

- Distortion: additive noise, reverberation, speech enhancement



- Correlation-based measures are in this category

STOI [Taal et al. 2011]

STOI - short term objective intelligibility

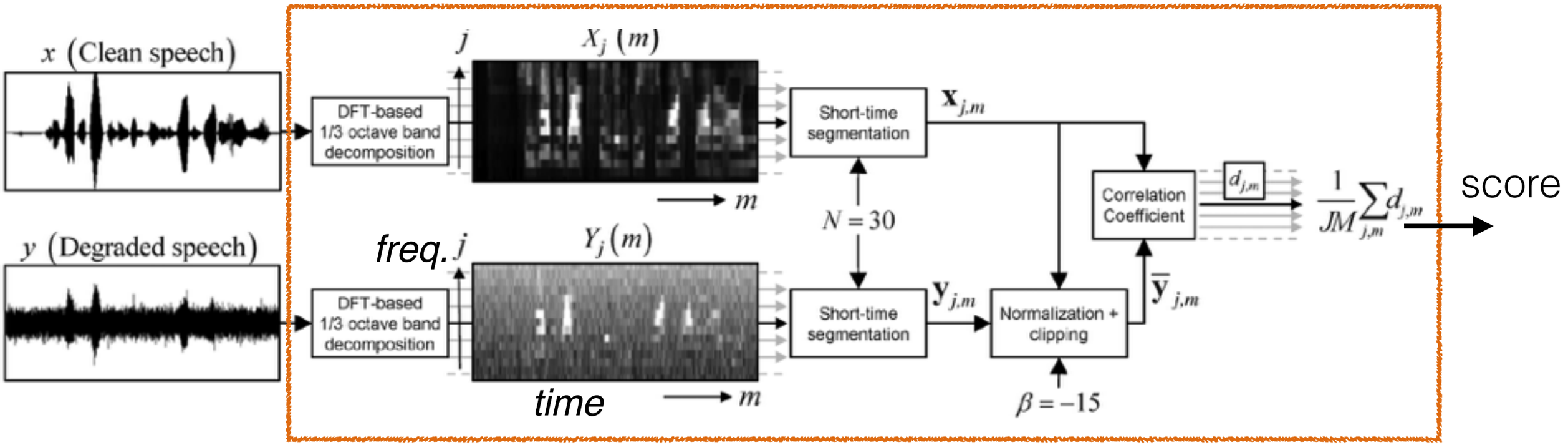


Figure taken from [Taal et al. 2011]

Classification

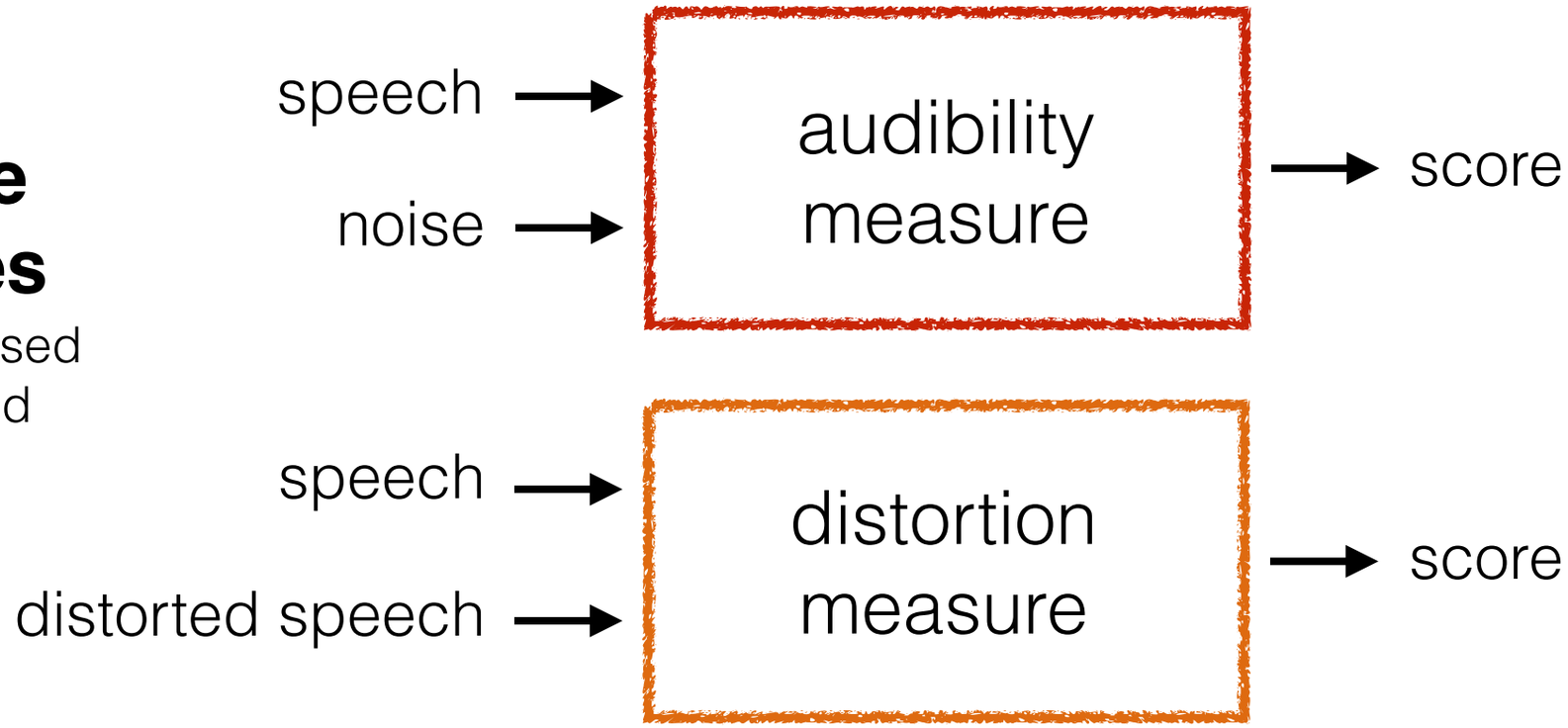
- **Macroscopic** / Microscopic
- **Audibility-based / Distortion-based**
- **Intrusive / Non-intrusive**
- Monaural / Binaural

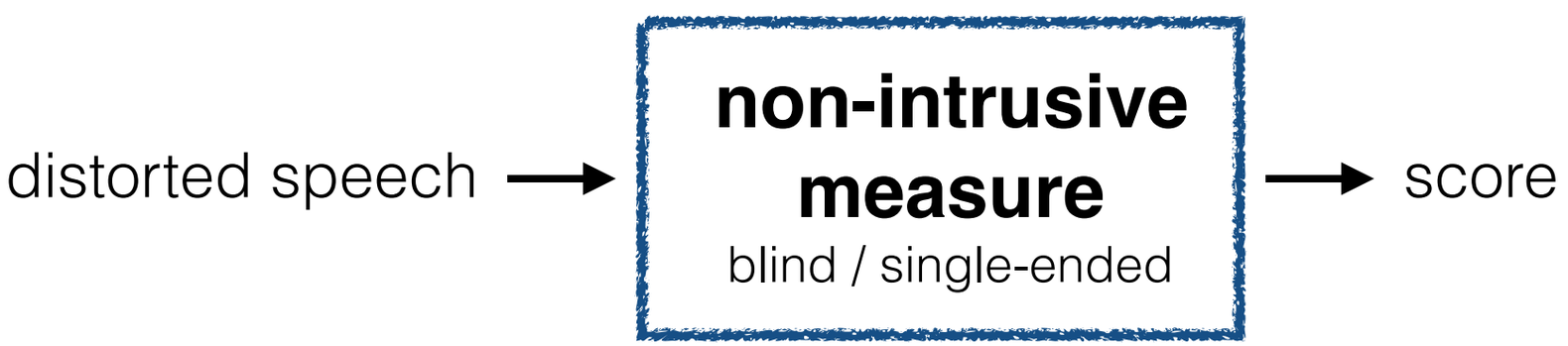
Intrusive/Non-intrusive

intrusive measures

comparison-based

double-ended





Classification

- **Macroscopic** / Microscopic
- **Audibility-based / Distortion-based**
- **Intrusive / Non-intrusive**
- **Monaural / Binaural**

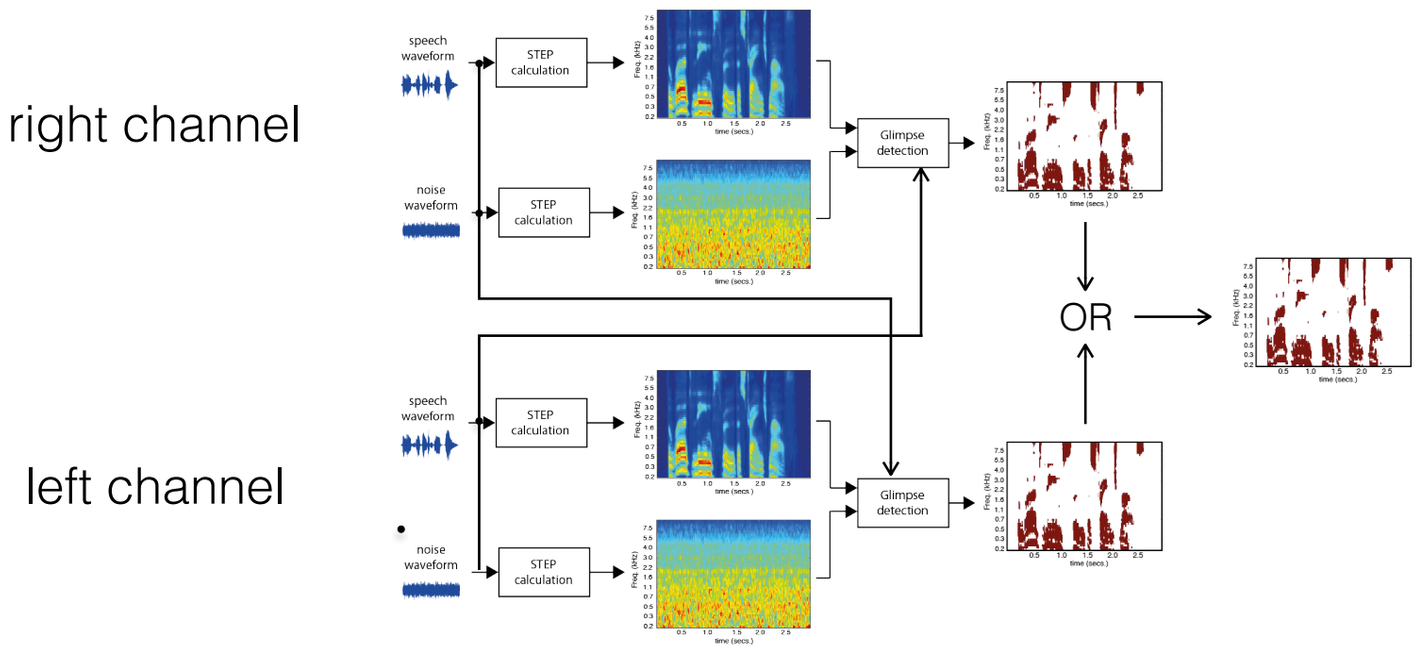
Monaural/Binaural

- Binaural measures attempt to model:
    - Better ear effect (head shadow effect)
        - how to choose from both ears?
          (select best overall score or on a frequency band basis)
- Binaural interaction (interaural time differences)
    - how to account for phase differences between speech and noise?

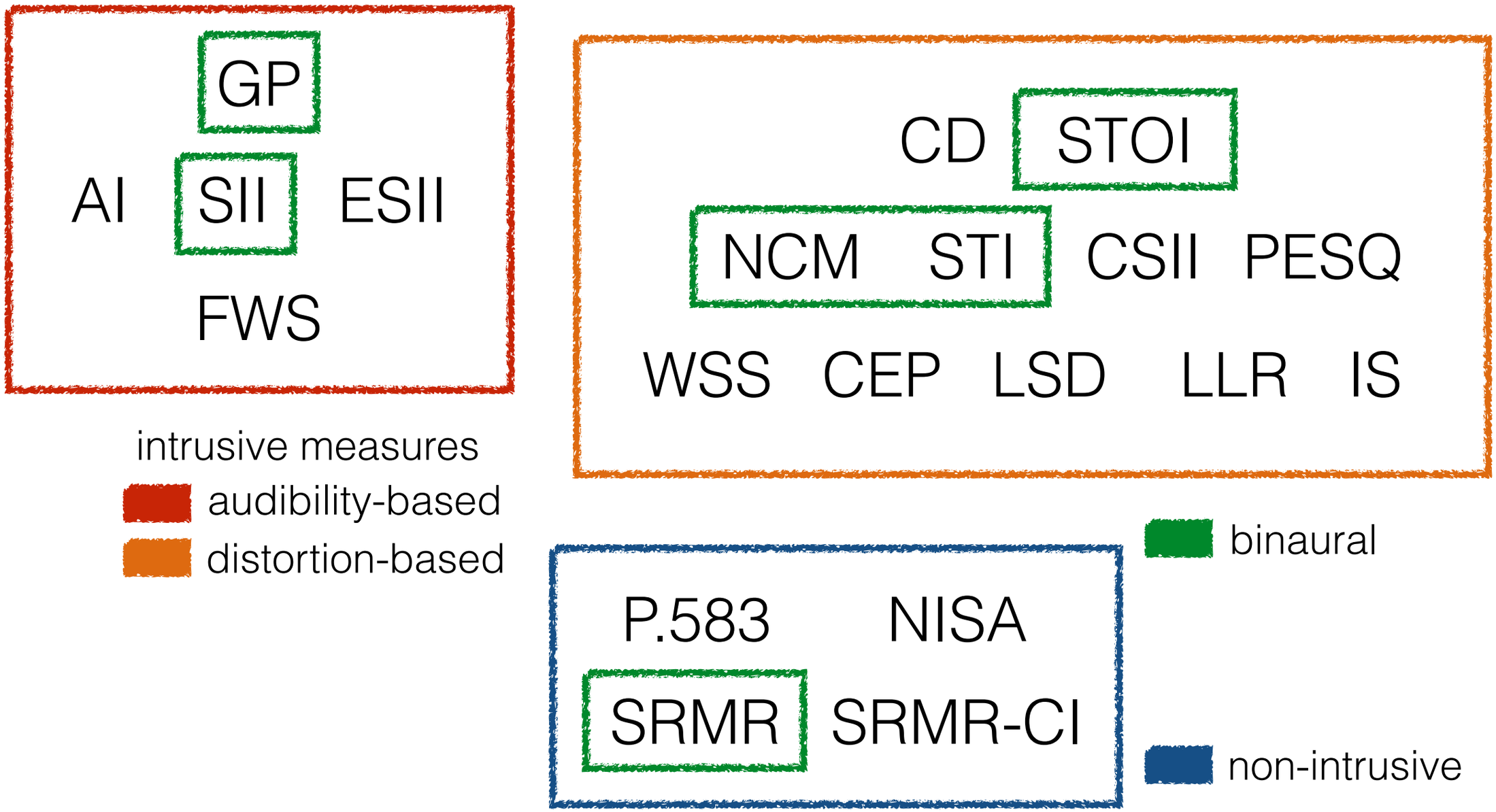
BiDWGP [Tang et al., 2016]

Binaural distortion weighted GP measure

- Better ear effect
  - a glimpse is counted when either or both ears detected it
- Binaural interaction
  - glimpse detection considers the binaural masking level difference (phase difference between speech and noise, noise signal coherence)



Existing Measures



Evaluation

Existing evaluation

- Additive noise and reverberation
  - Speech in noise [Chen et al., 2012; Taal et al., 2009]
  - Speech enhancement (noise suppression) [Taal et al., 2009]
  - Speech pre-enhancement (intelligibility enhancement) [Tang et al., 2016b]
  - Reverberation [Falk et al. 2010]
- Assistive listening devices [Falk et al. 2015]
- Synthetic speech in noise [Valentini et al., 2012a, Valentini et al., 2012b]

Evaluation procedure

- Objective scores should reflect subjective scores
- Subjective intelligibility score is derived from a listener's response to a task:
  - word accuracy at a sentence level
    - type of sentences: Harvard sentences, matrix sentences, SUS, SPIN…
- Compare scores at a condition
  (noise type, noise level, reverberation time, enhancement type)
  - scores are averaged across sentences (and listeners) belonging to the same condition

Evaluation procedure

- To linearise the relation between subjective and objective scores, a mapping function is applied to fit the output of each model to the mean listener scores

- Metrics
  - Pearson correlation
  - Kendall's tau (rank correlation)
  - Standard deviation of the error
- Scatter plots

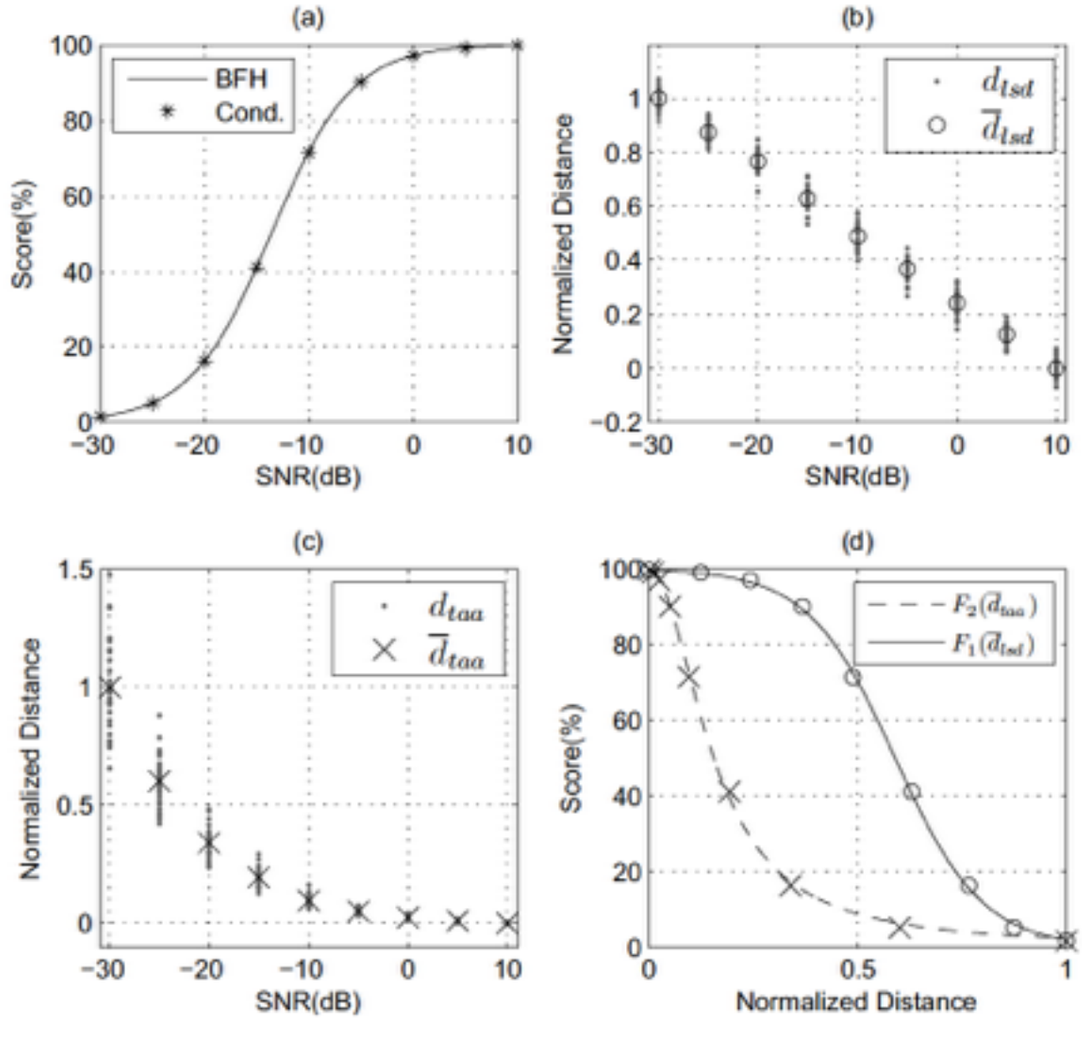


Figure from [Taal et al. 2009]

Evaluation example

Text-to-speech intelligibility in noise

- Which measures best predict the impact of noise on the intelligibility of an HMM-based text-to-speech voice

- Intrusive measures are better than non intrusive ones

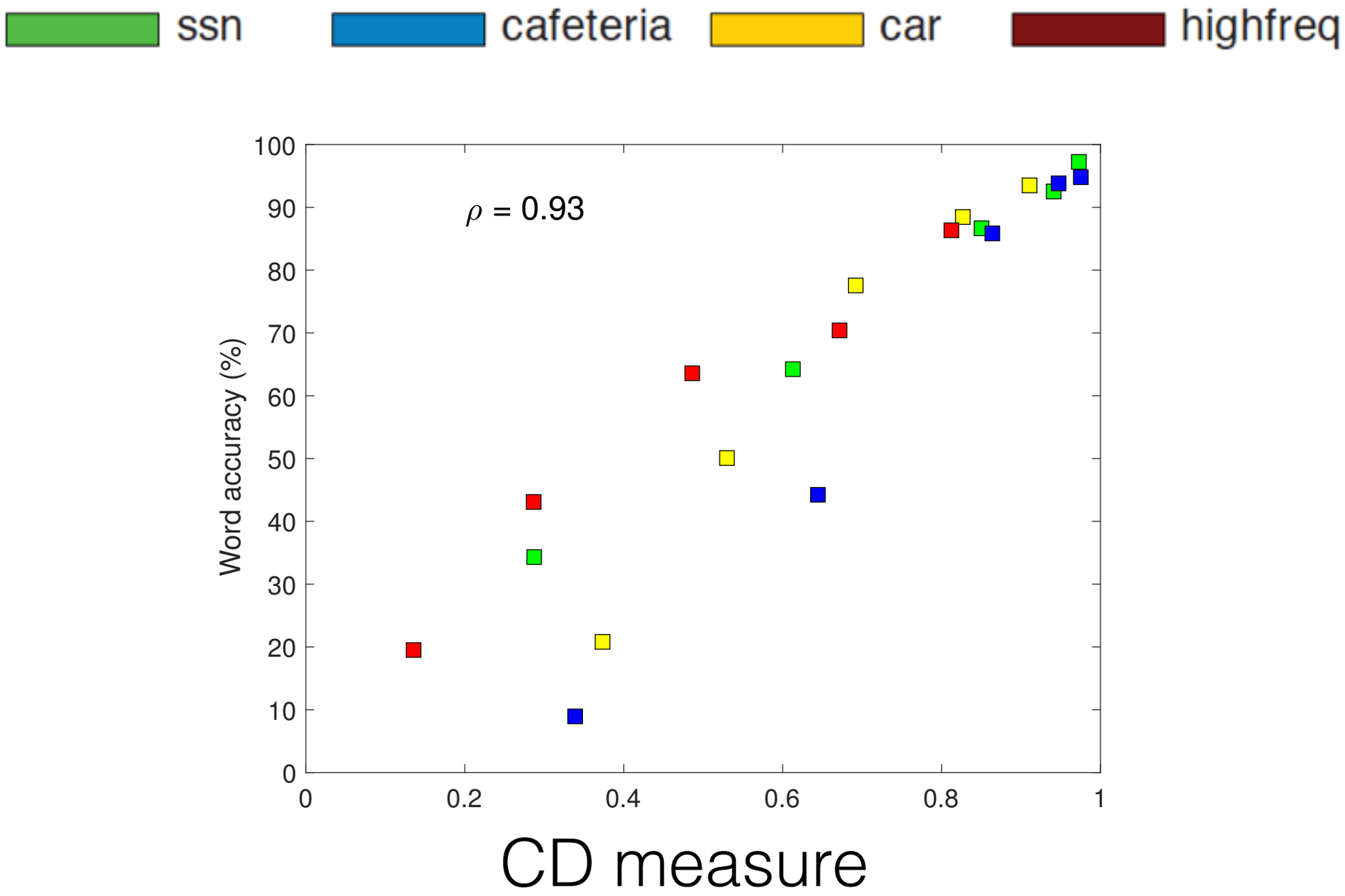
- Assumption: synthetic speech is as intelligible as natural speech in quiet

- Use clean synthetic speech as reference

- Monaural intrusive measures

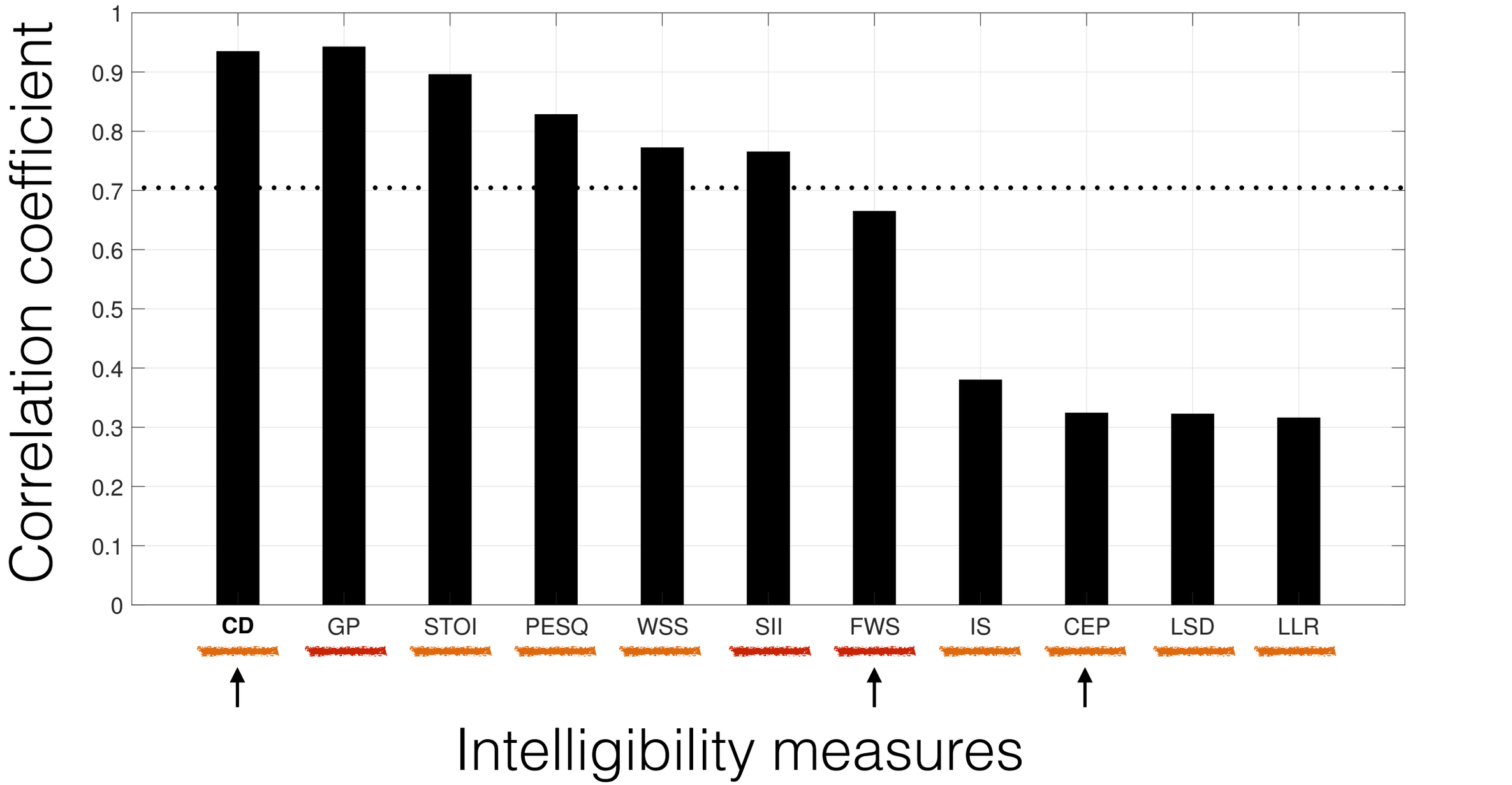
- Listening experiment 1:

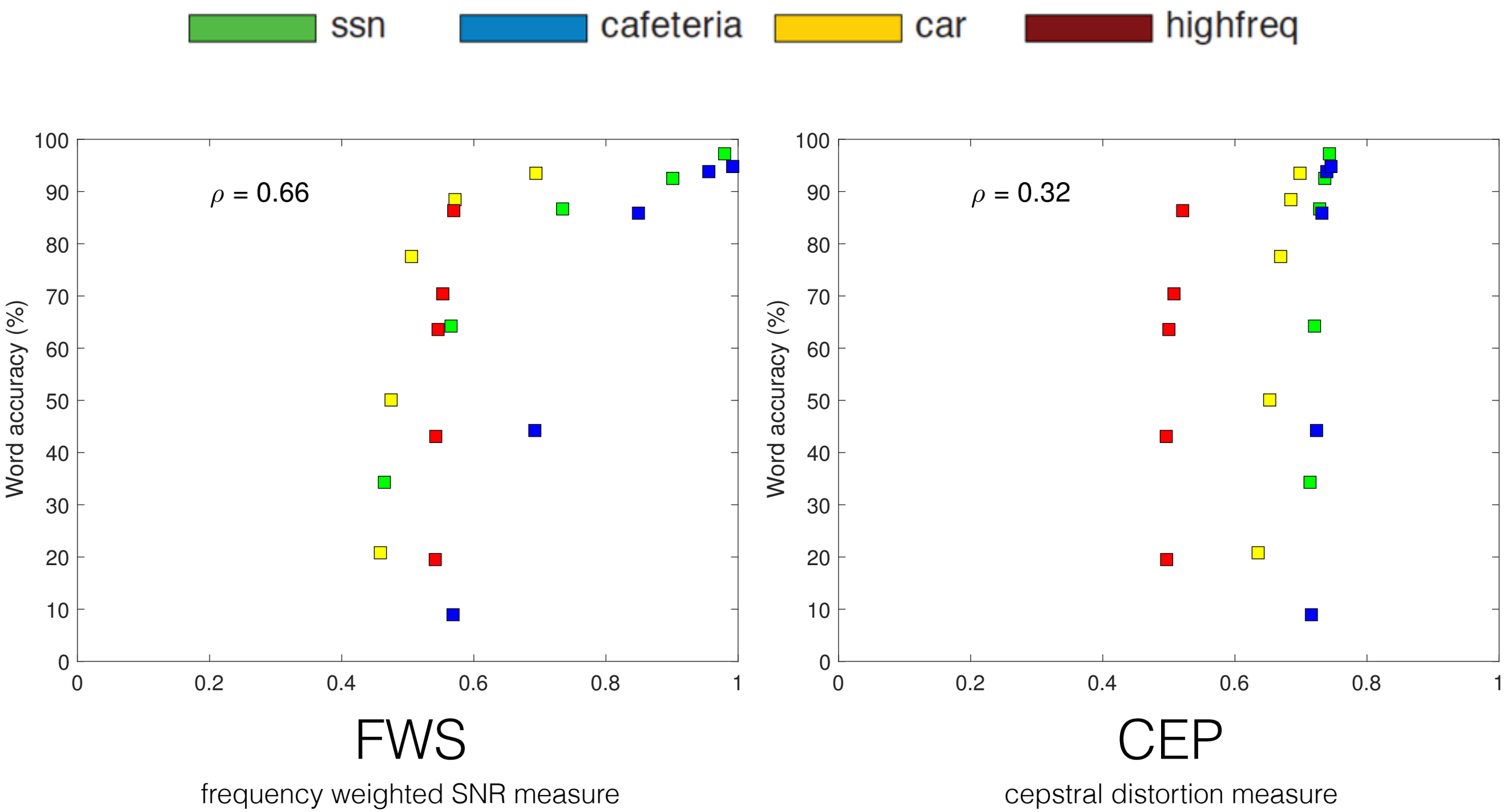
  - 4 noises at 5 SNR levels


ssn
cafeteria
car
highfreq
ρ = 0.93
Word accuracy (%)
100
90
80
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
CD measure

Results

intrusive measures
audibility-based
distortion-based

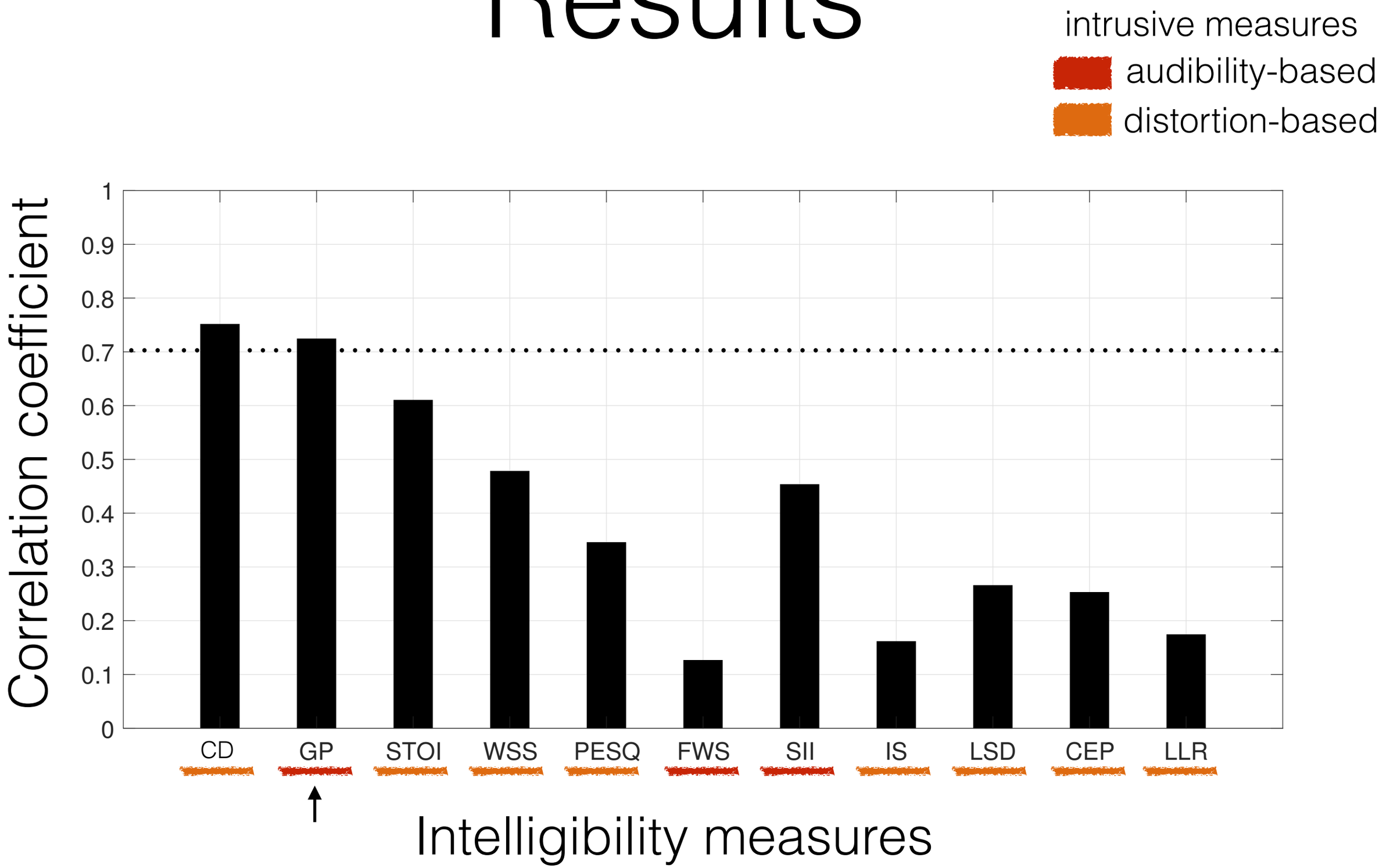



ssn
cafeteria
car
highfreq
Word accuracy (%)
ρ = 0.66
FWS
frequency weighted SNR measure
ρ = 0.32
CEP
cepstral distortion measure

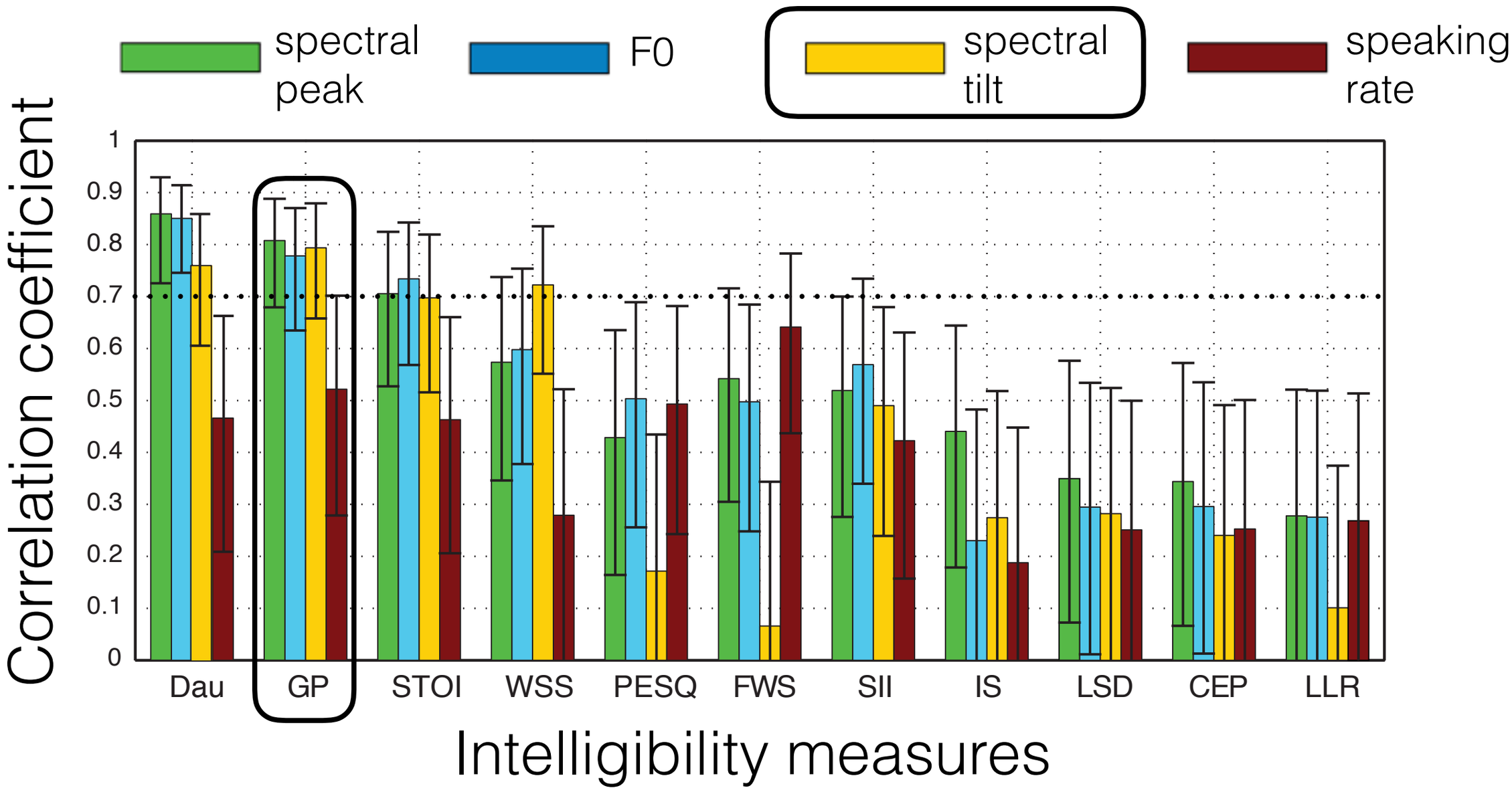
Evaluation with a text-to-speech voice in noise

- Intelligibility in noise of an HMM-based text-to-speech voice that has been enhanced:
    - spectral peak enhancement
    - mean F0 increase
    - line spectral pair (LSP) shift (spectral tilt changes)
    - speaking rate changes
- Listening experiment 2:
    - 4 noises at 4 SNR levels
    - 4 speech modifications at 2 or 3 levels

Results

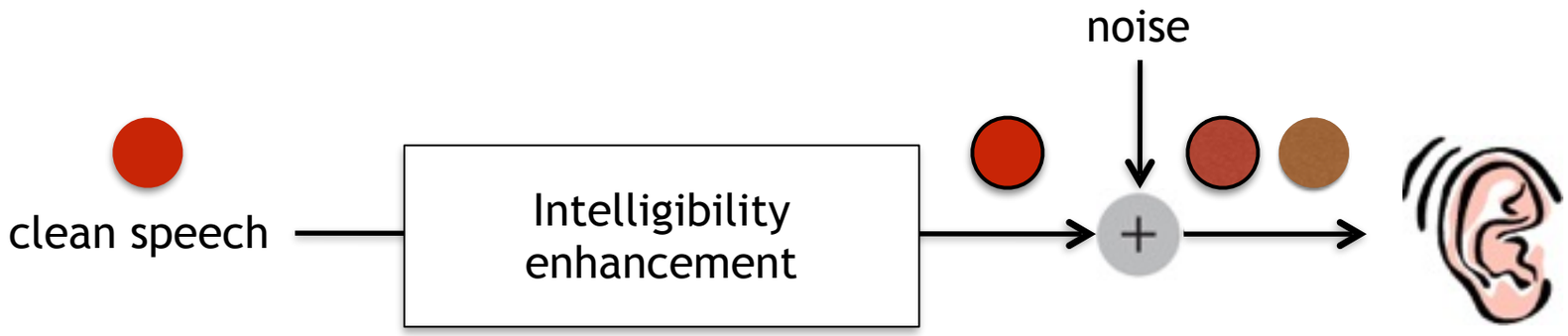

intrusive measures
audibility-based
distortion-based
Correlation coefficient
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CD
GP
STOI
WSS
PESQ
FWS
SII
IS
LSD
CEP
LLR
Intelligibility measures

Results with different modification types



Application

Speech pre-enhancement

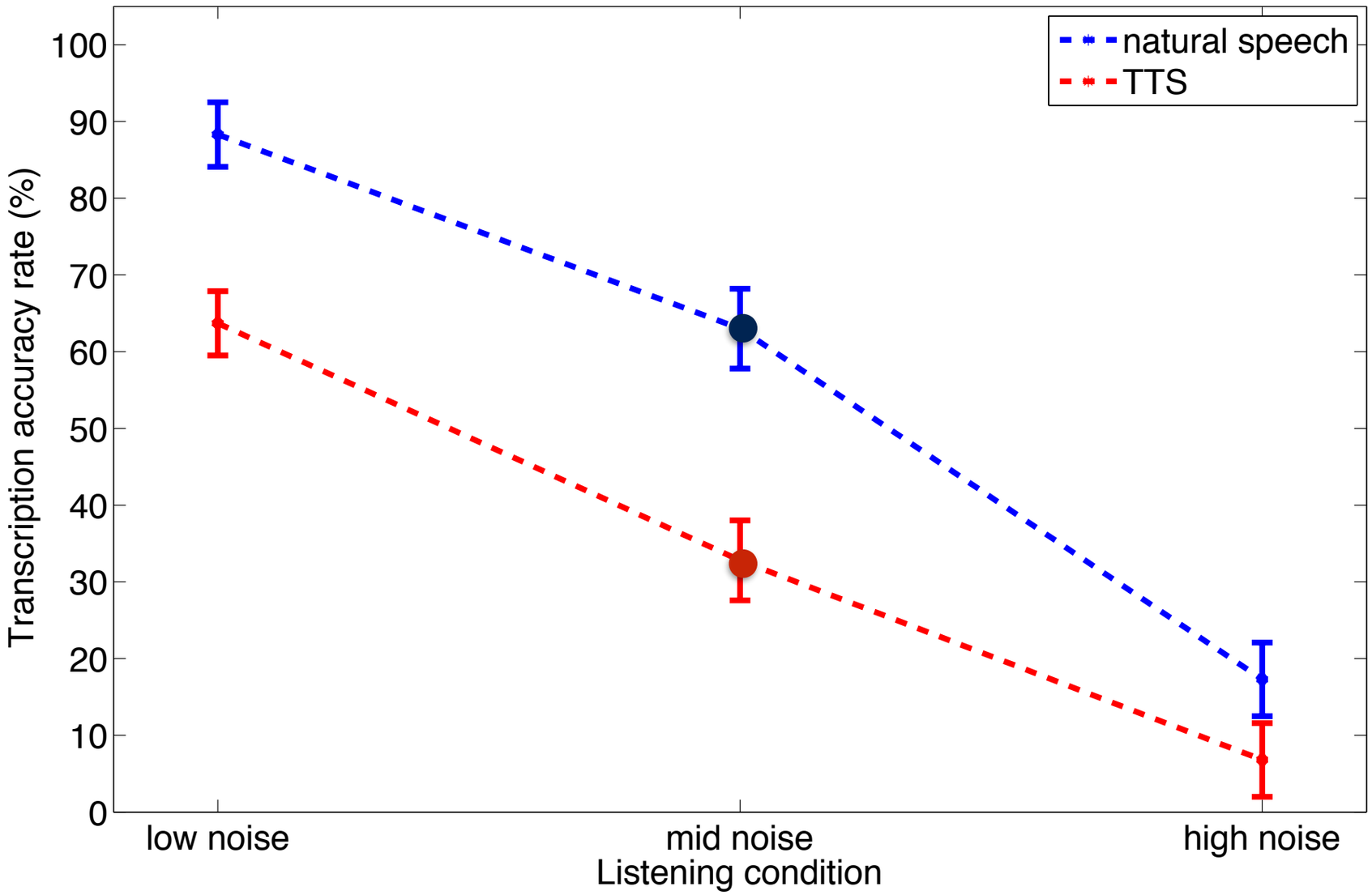


- A measure can be used to control pre-enhancement parameters
  - in order to max intelligibility [Sauert et al. 2009, Tang et al. 2012, Valentini et al. 2012, Taal et al. 2012, Aubanel et al. 2013]
  - according to intelligibility levels [Schepker et al. 2013]

Speech pre-enhancement

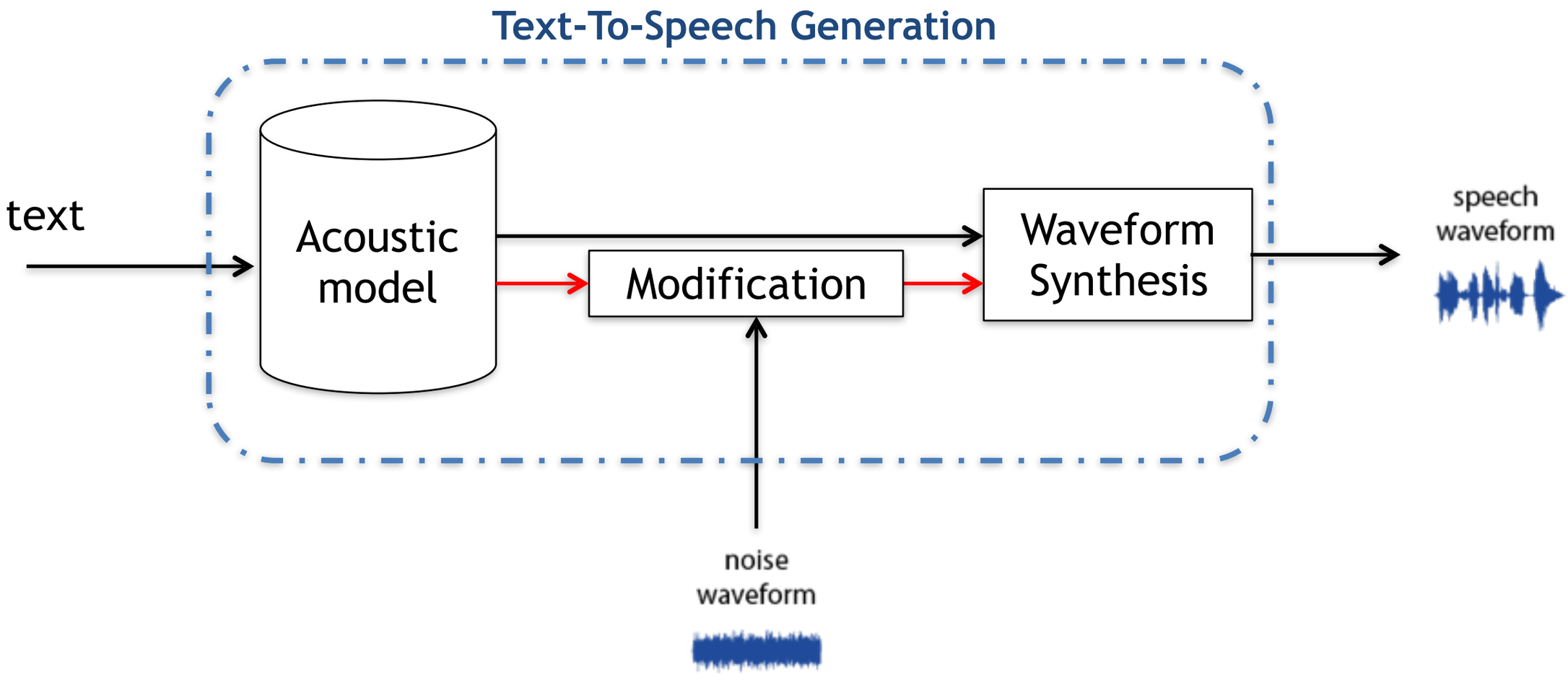
- Define:
  - Objective measure to use (see evaluation results)
    - GP, SII, STOI
  - What should be changed (spectrum, F0, speaking rate)
  - Constraints
    - overall SNR, overall loudness
    - range of modification
  - Optimisation routine
    - genetic algorithms, gradient descent, greedy search, dynamic programming

Application example



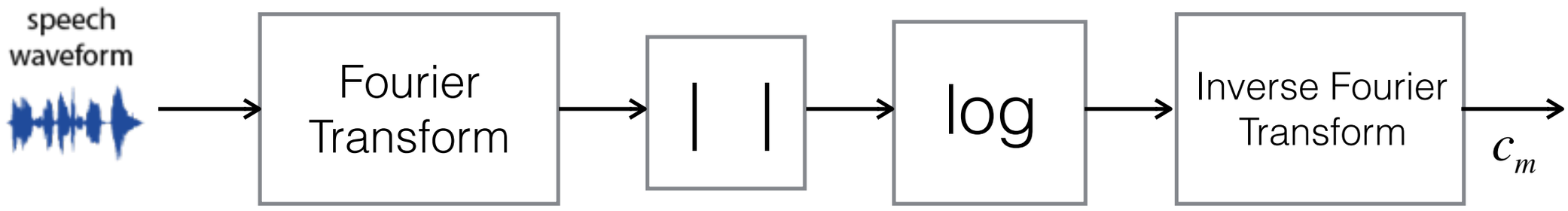
Hurricane Challenge results
for speech-shaped noise

Solution


Text-To-Speech Generation
text
Acoustic model
Modification
Waveform Synthesis
speech waveform
noise waveform

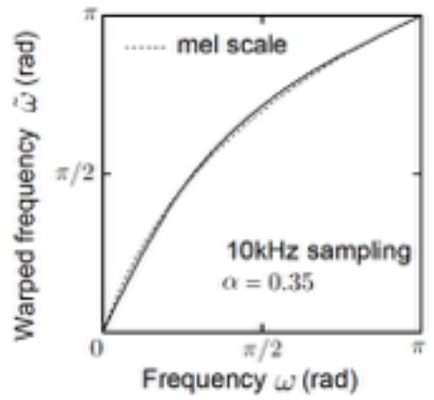
Spectrum

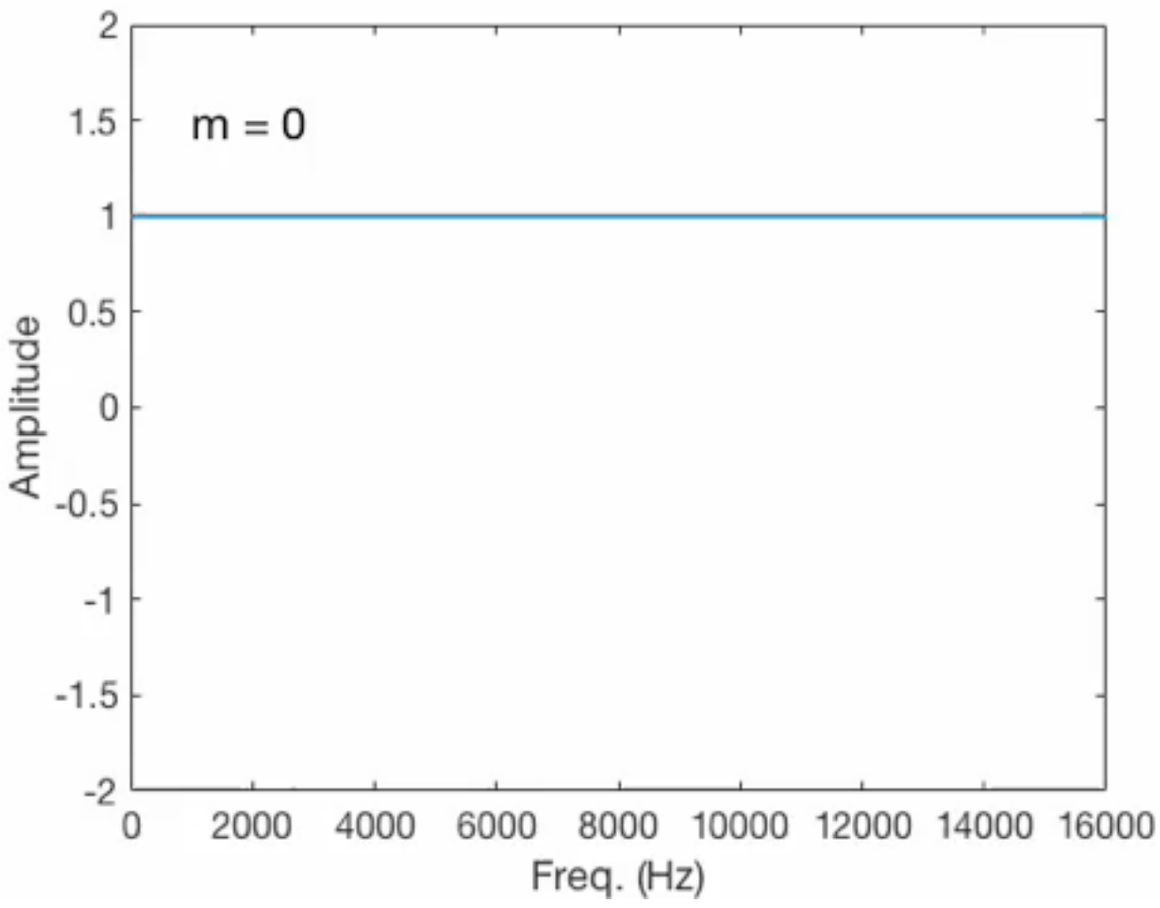
- The log spectrum envelope $\log|H(w_k)|$ can be parametrised via cepstral coefficients:



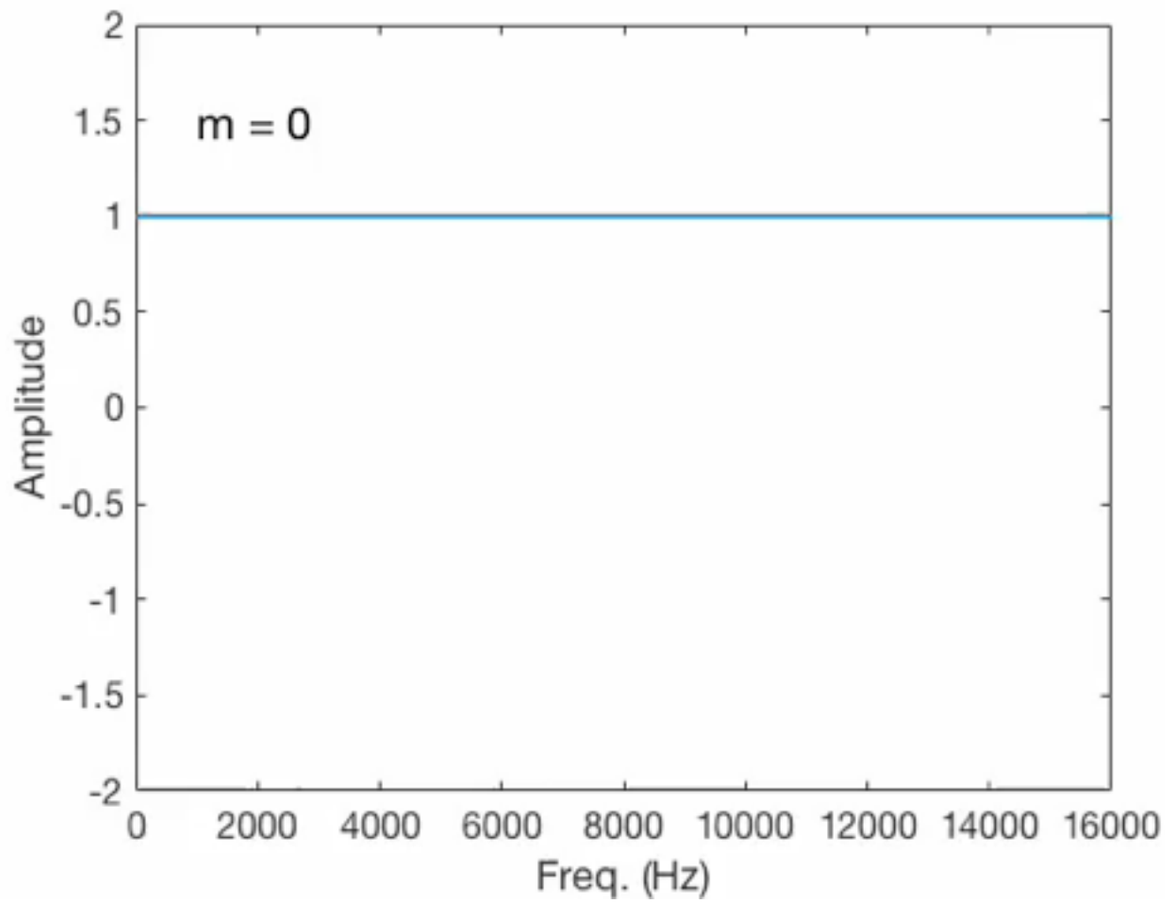
$$\log|H(w_k)| = \sum_{m=0}^{M} c_m \cos(w_k)$$

$$\log\left|H(w_k)\right| = \sum_{m=0}^{M} c_m \cos(w_k)$$



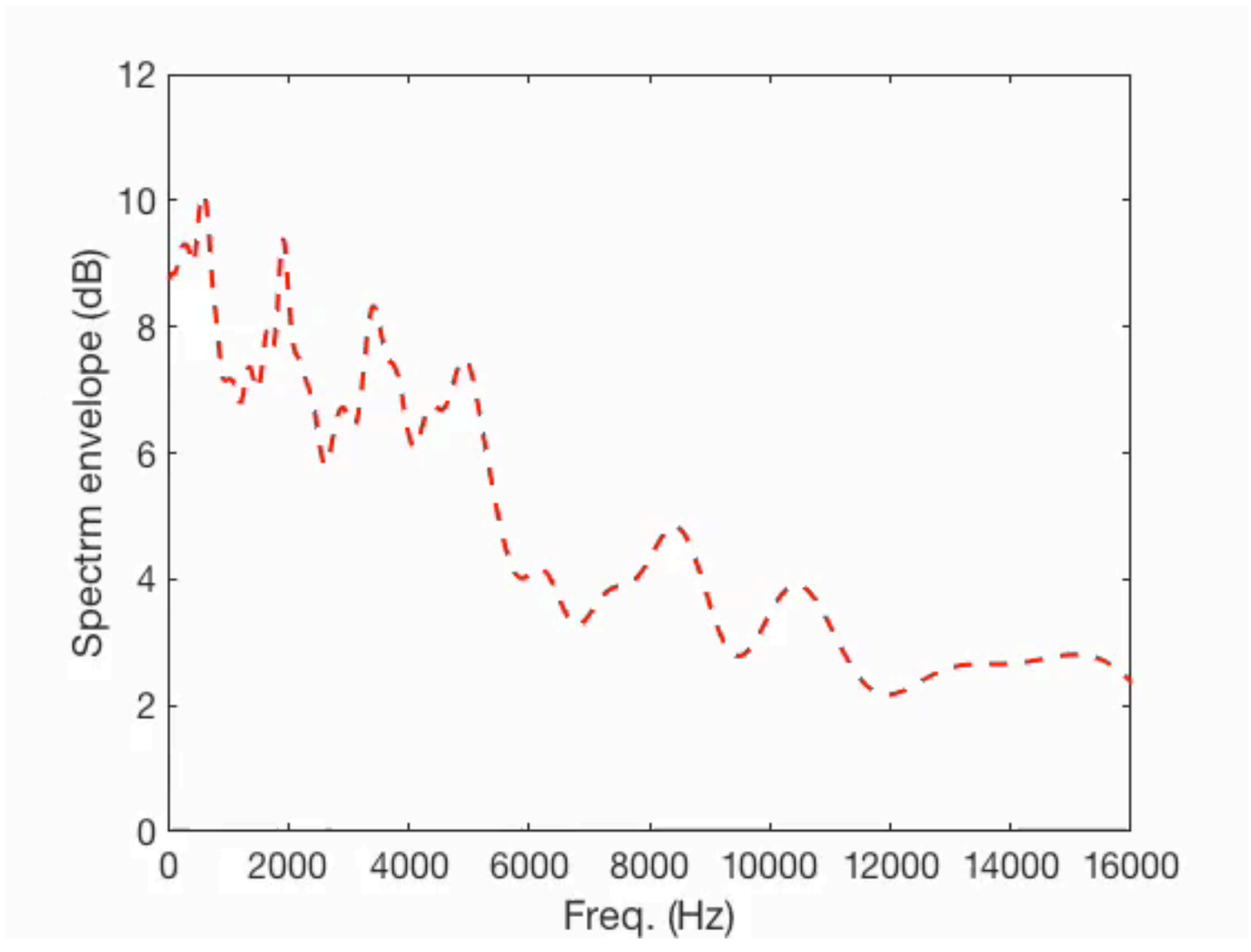


Cepstral coefficients

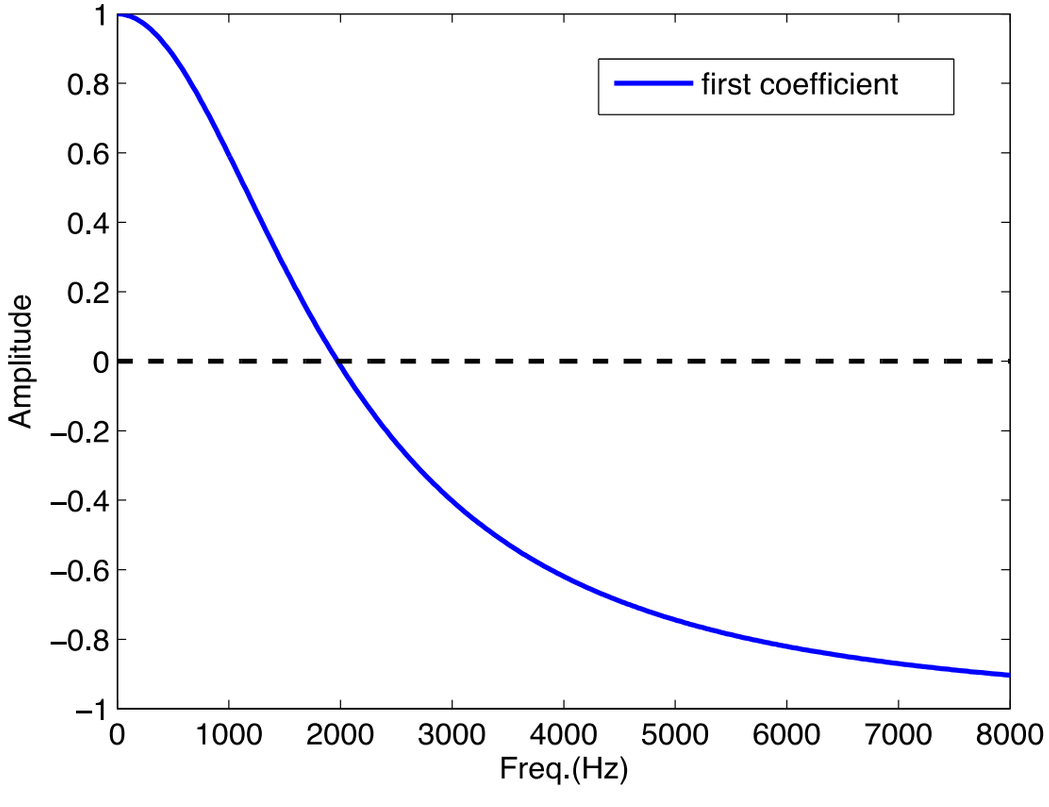


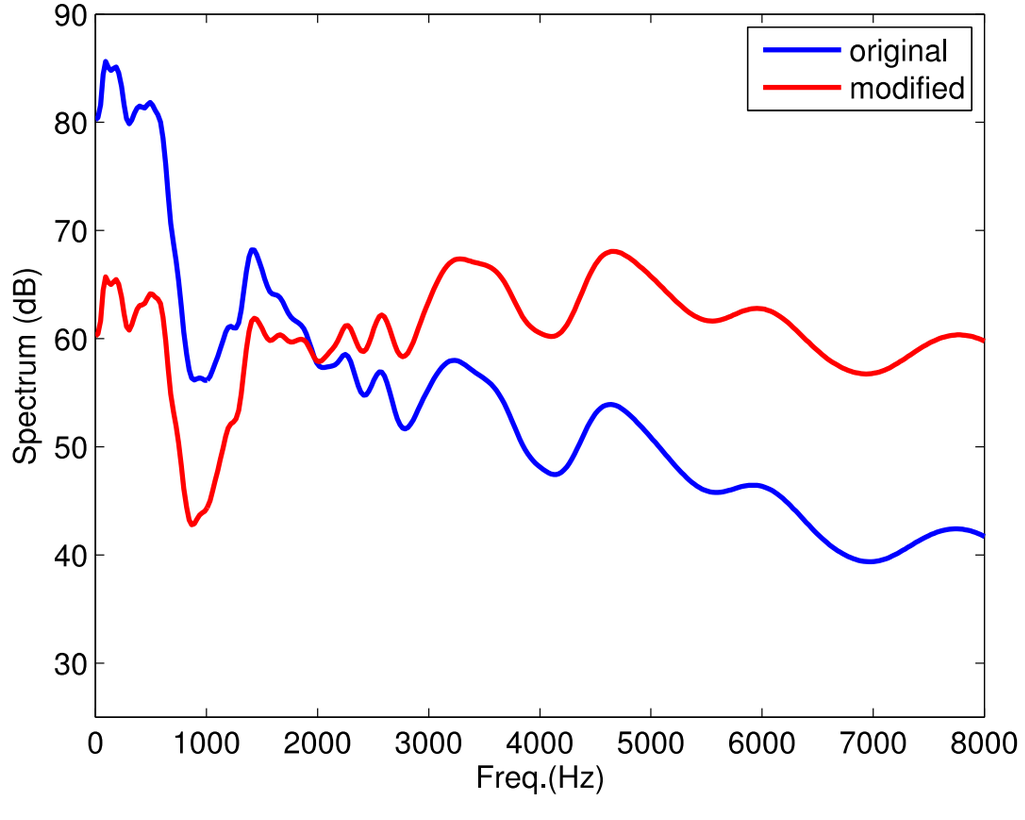
Mel cepstral coefficients

$$\log\left|H(w_k)\right| = \sum_{m=0}^{M} c_m \cos(w_k)$$

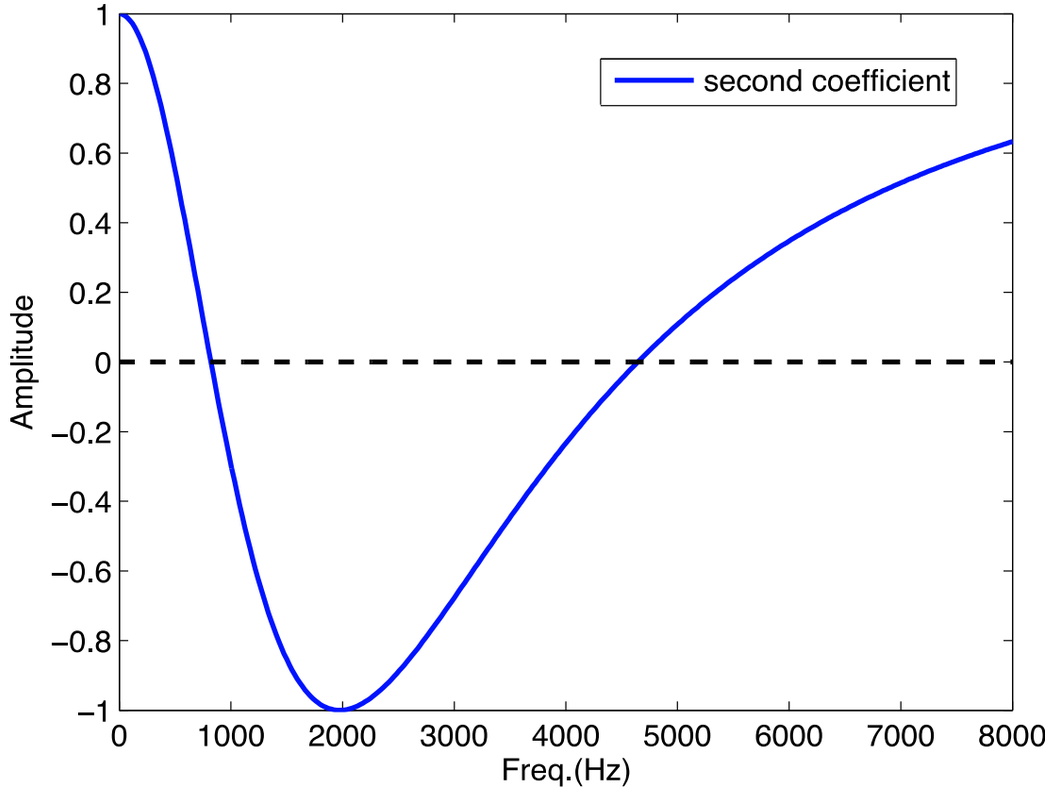


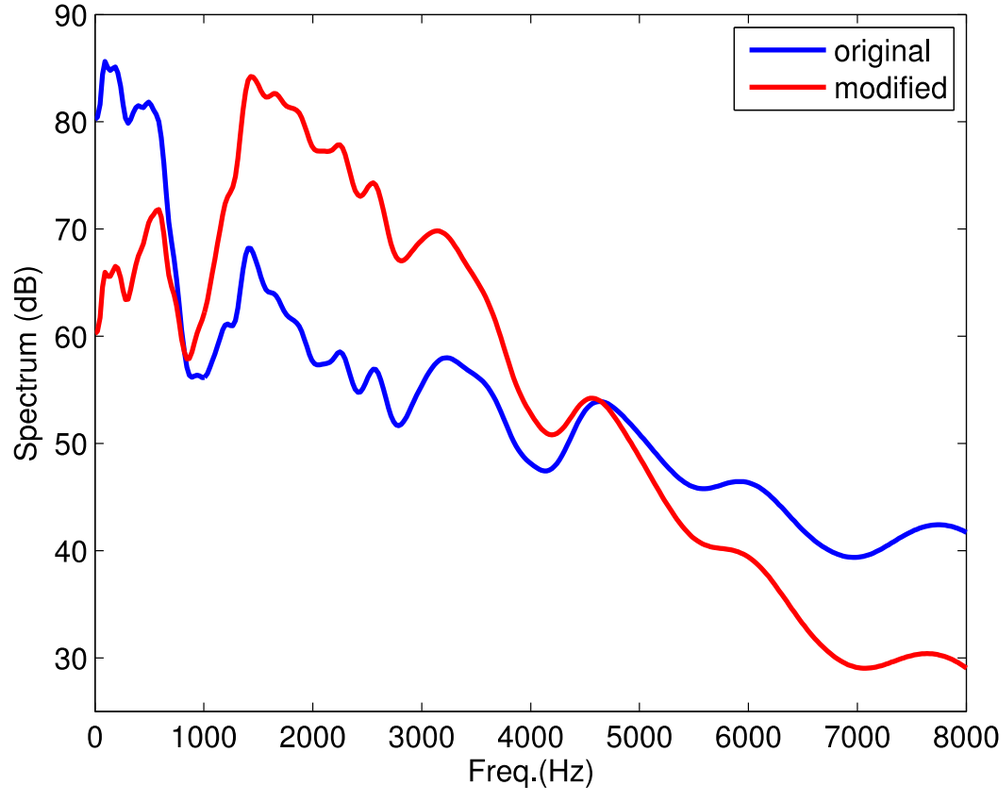
$$\log\left|H(w_k)\right| = \sum_{m=0}^{M} c_m \cos(w_k)$$



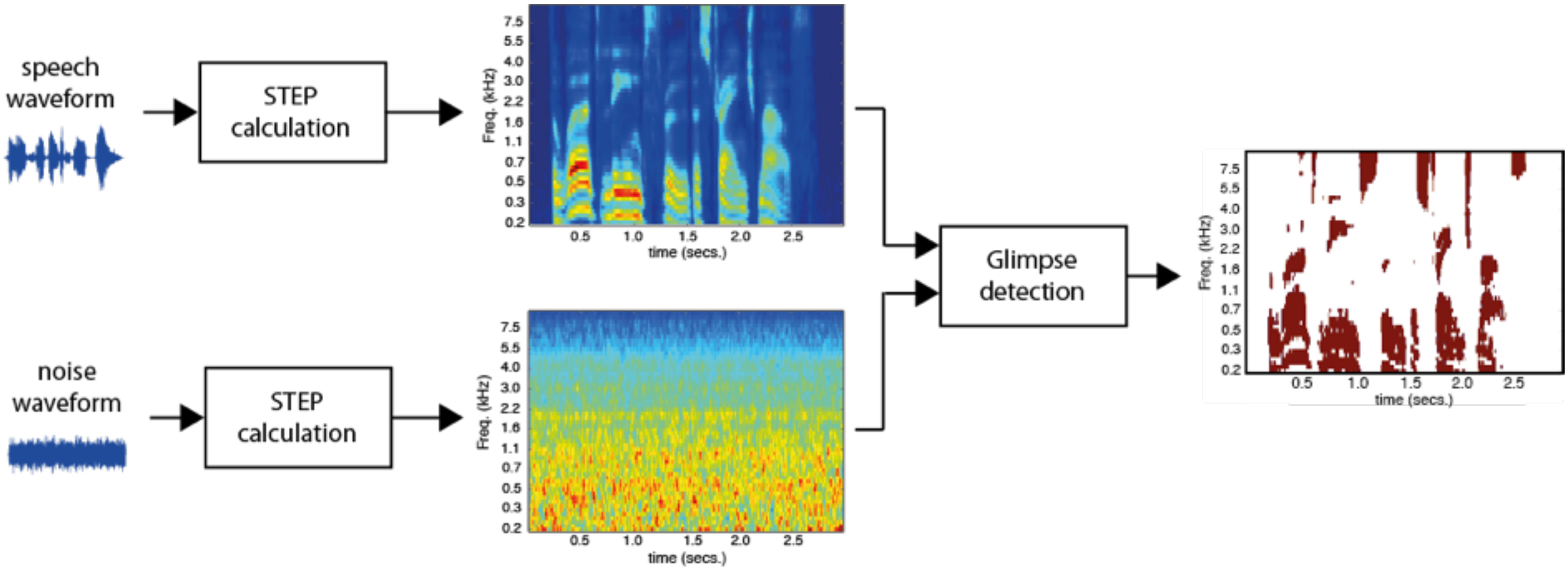


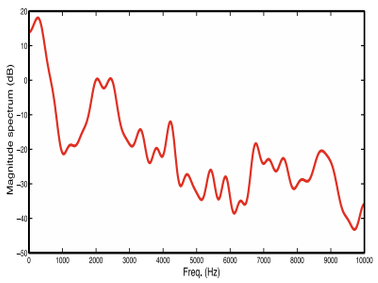
$$\log\left|H(w_k)\right| = \sum_{m=0}^{M} c_m \cos(w_k)$$





From Mel cepstral coefficients to the GP measure

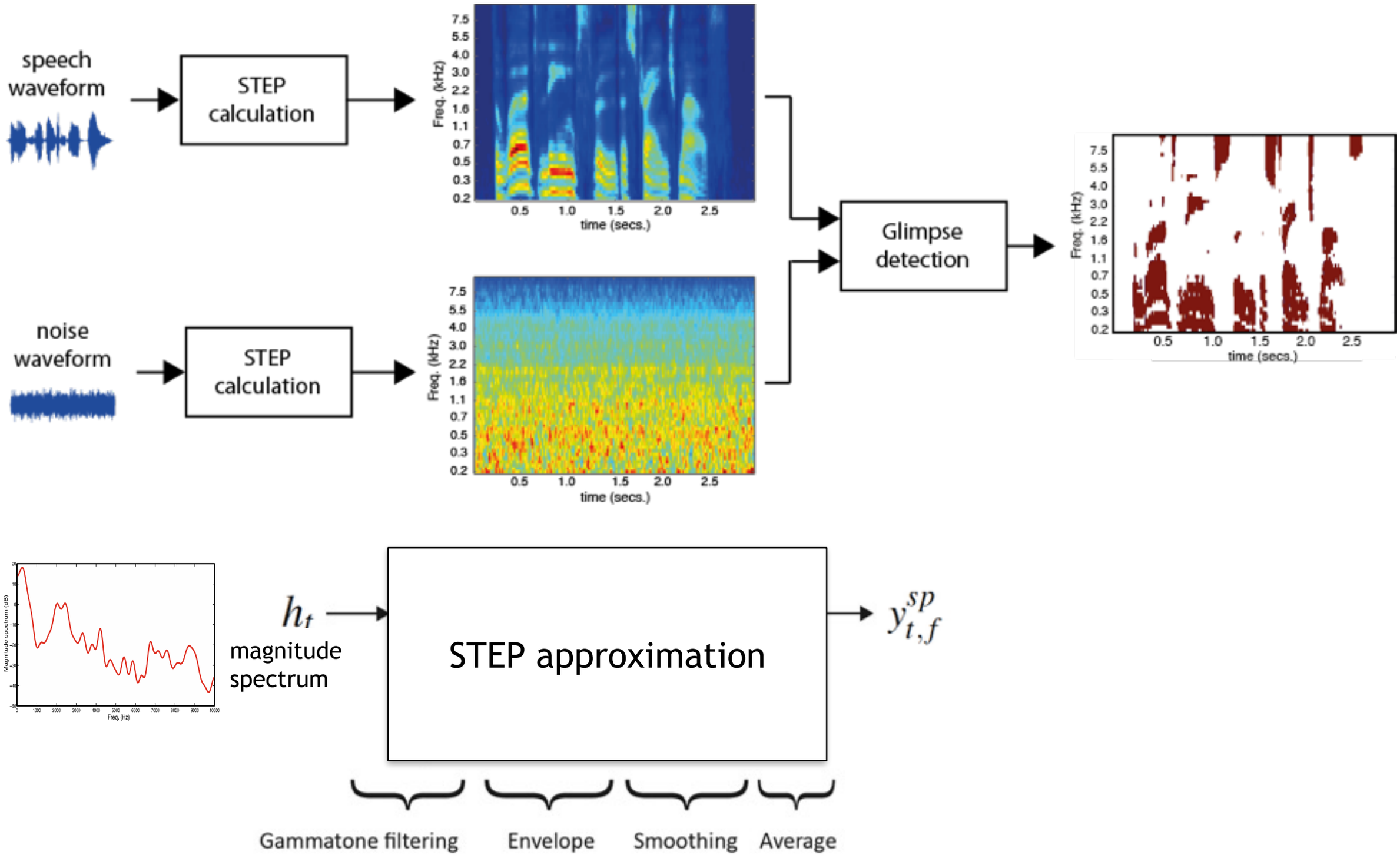




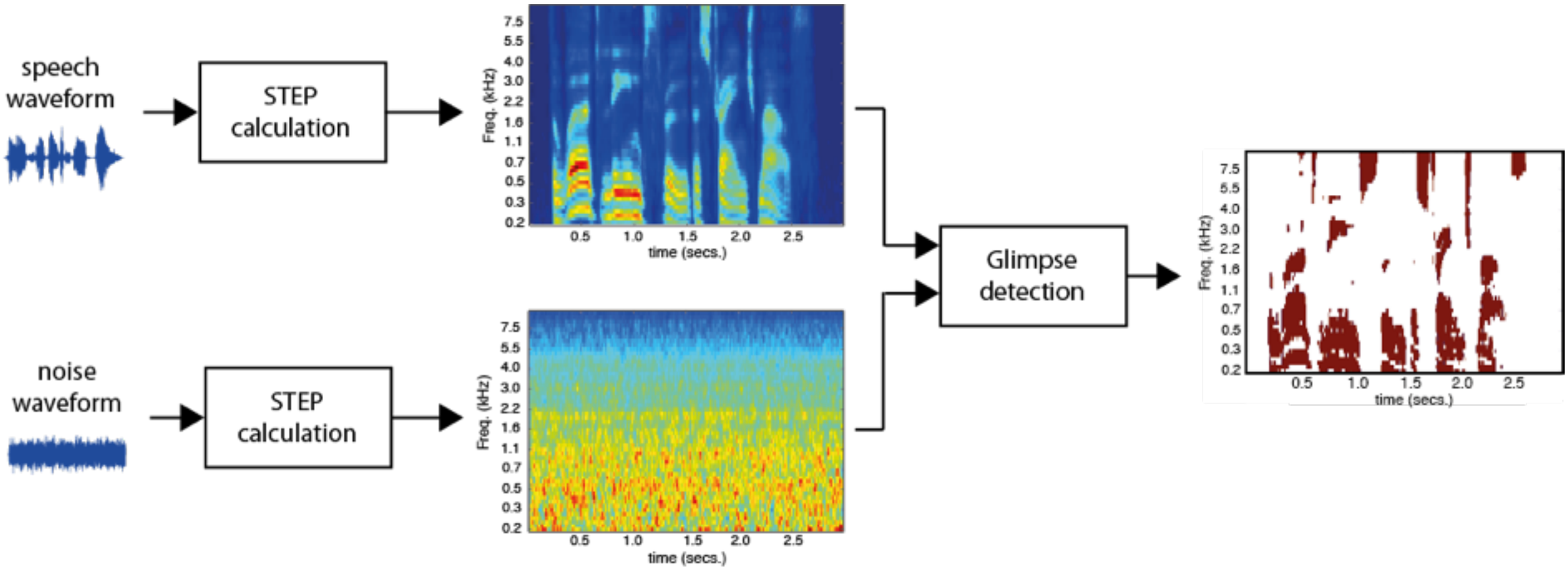
h_t

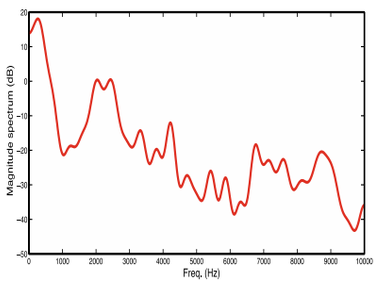
magnitude spectrum

From Mel cepstral coefficients to the GP measure

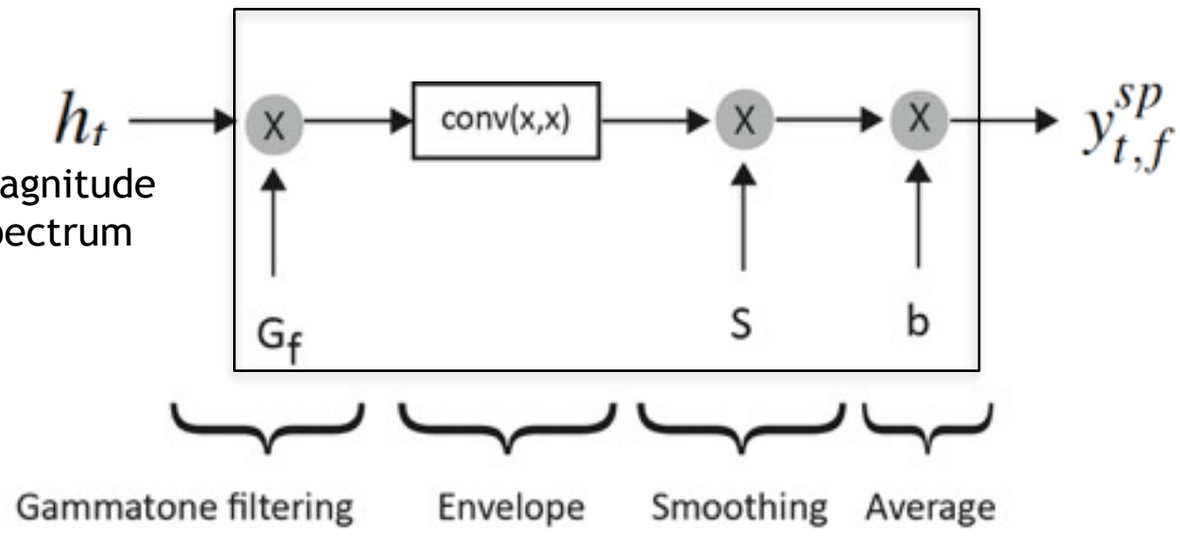


From Mel cepstral coefficients to the GP measure

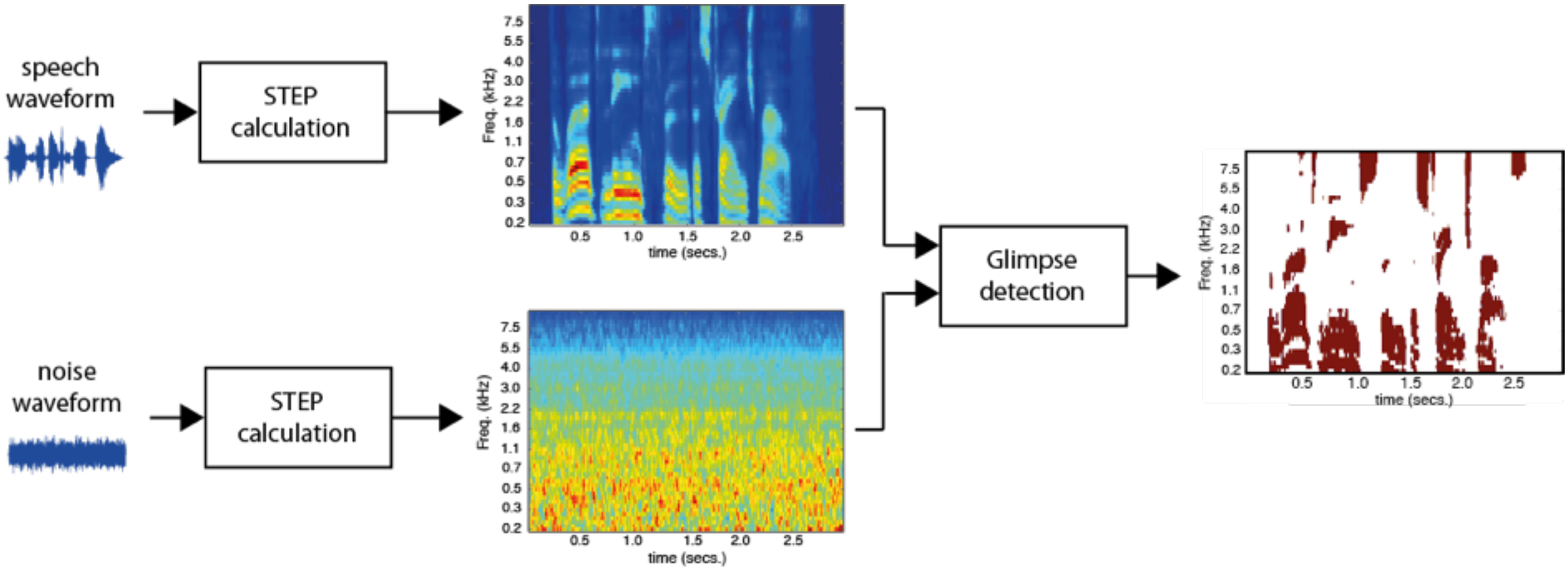


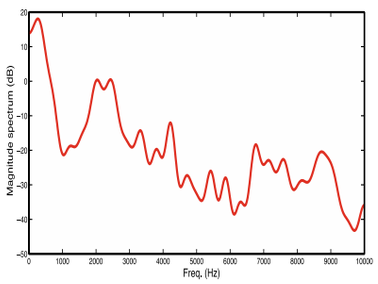


magnitude spectrum



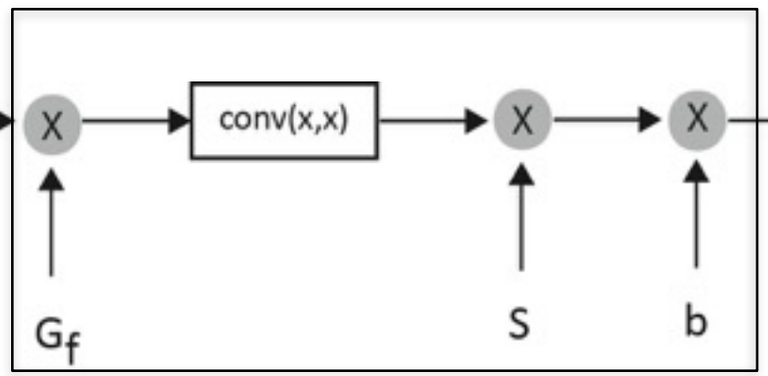
From Mel cepstral coefficients to the GP measure





h_t

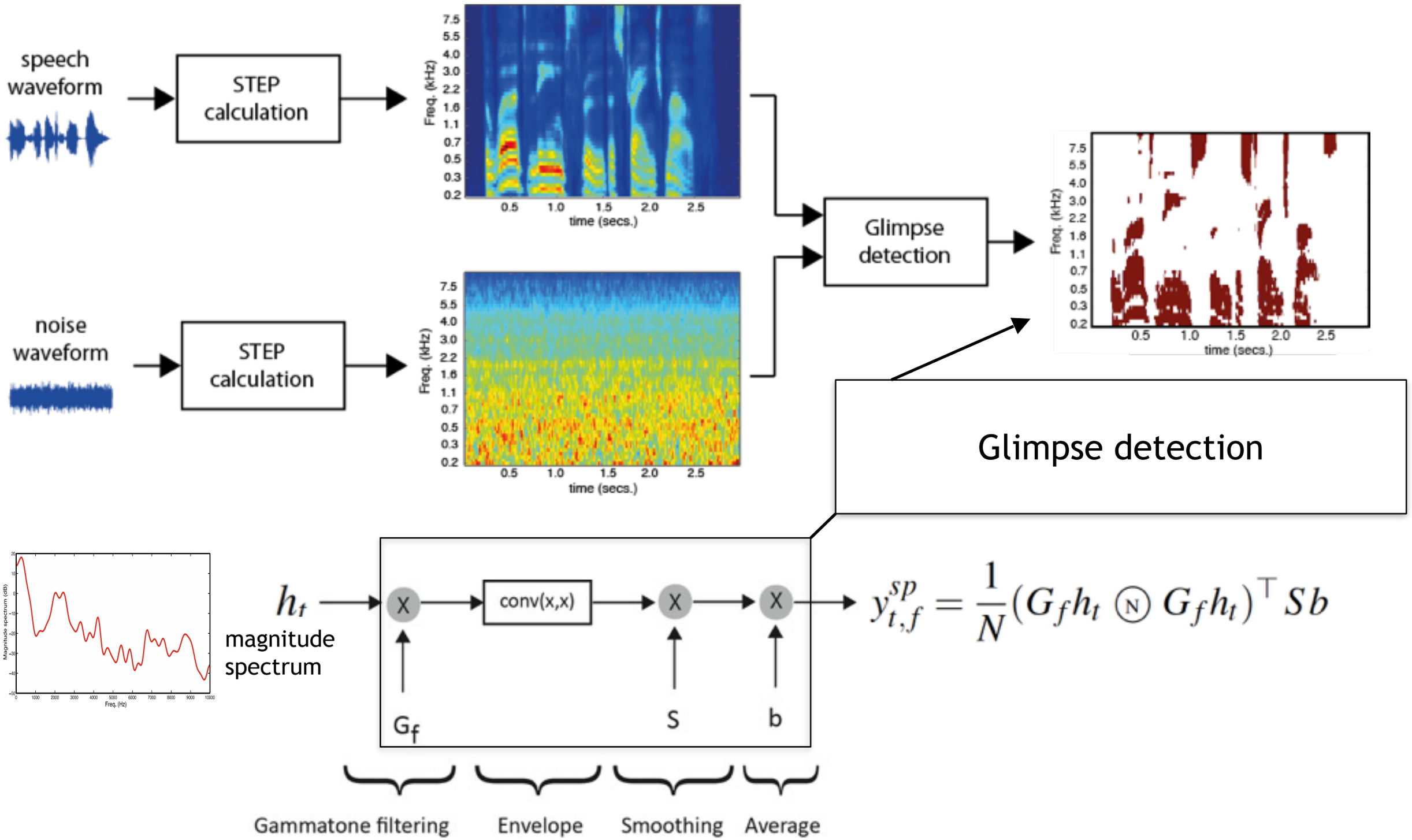
magnitude spectrum



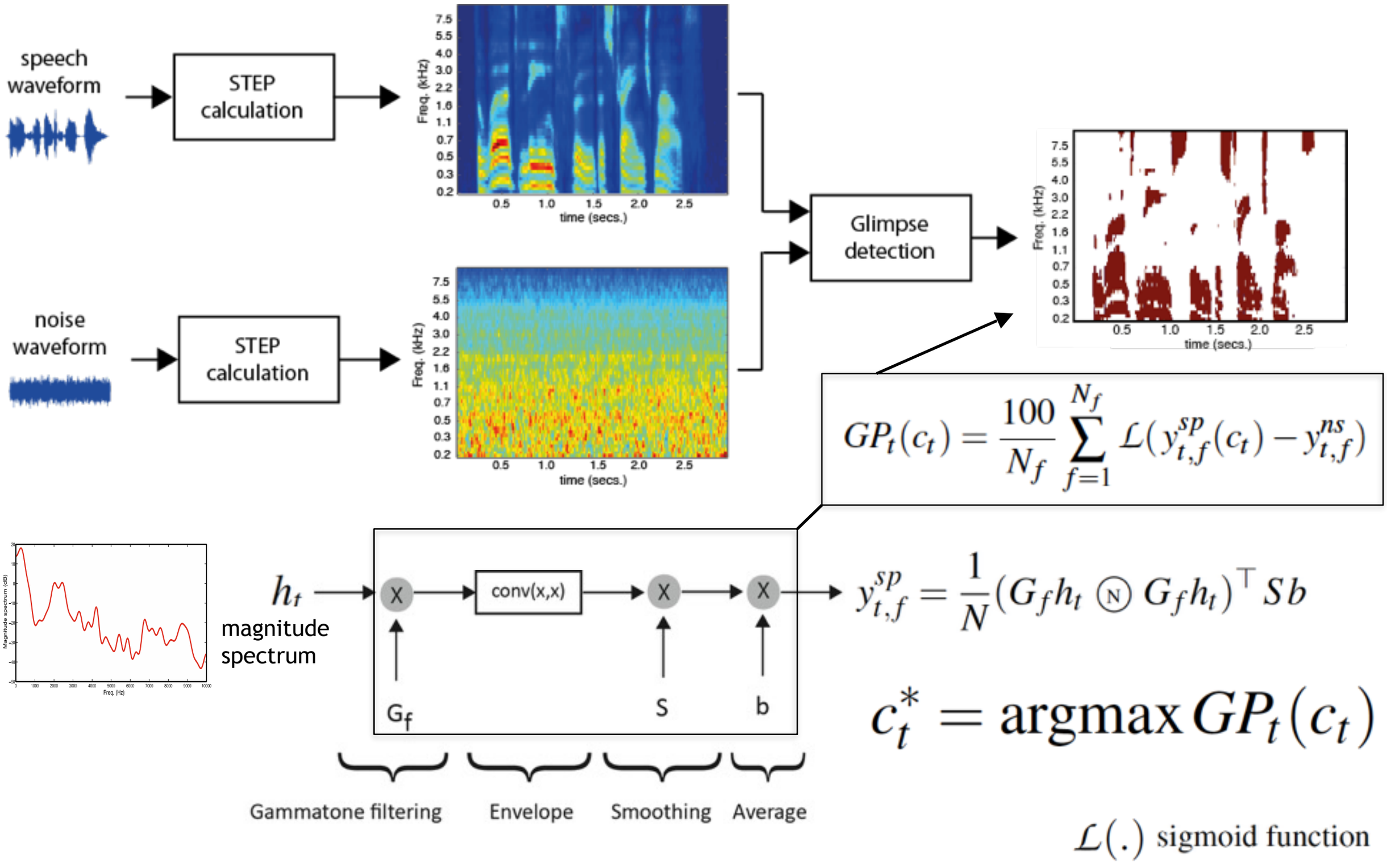
$$y_{t,f}^{sp} = \frac{1}{N}(G_f h_t \circledN G_f h_t)^\top S b$$

Gammatone filtering | Envelope | Smoothing | Average

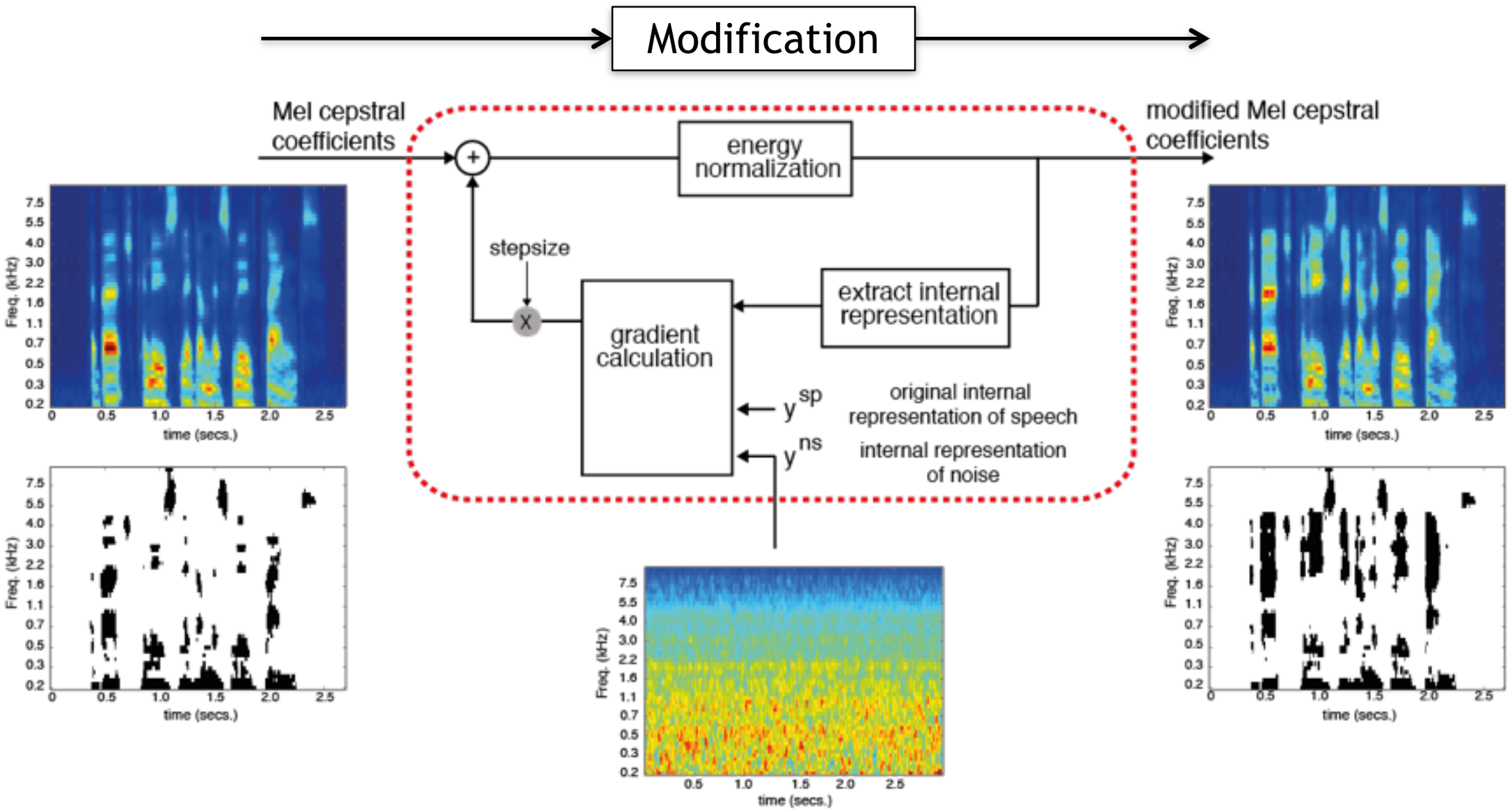
From Mel cepstral coefficients to the GP measure



From Mel cepstral coefficients to the GP measure



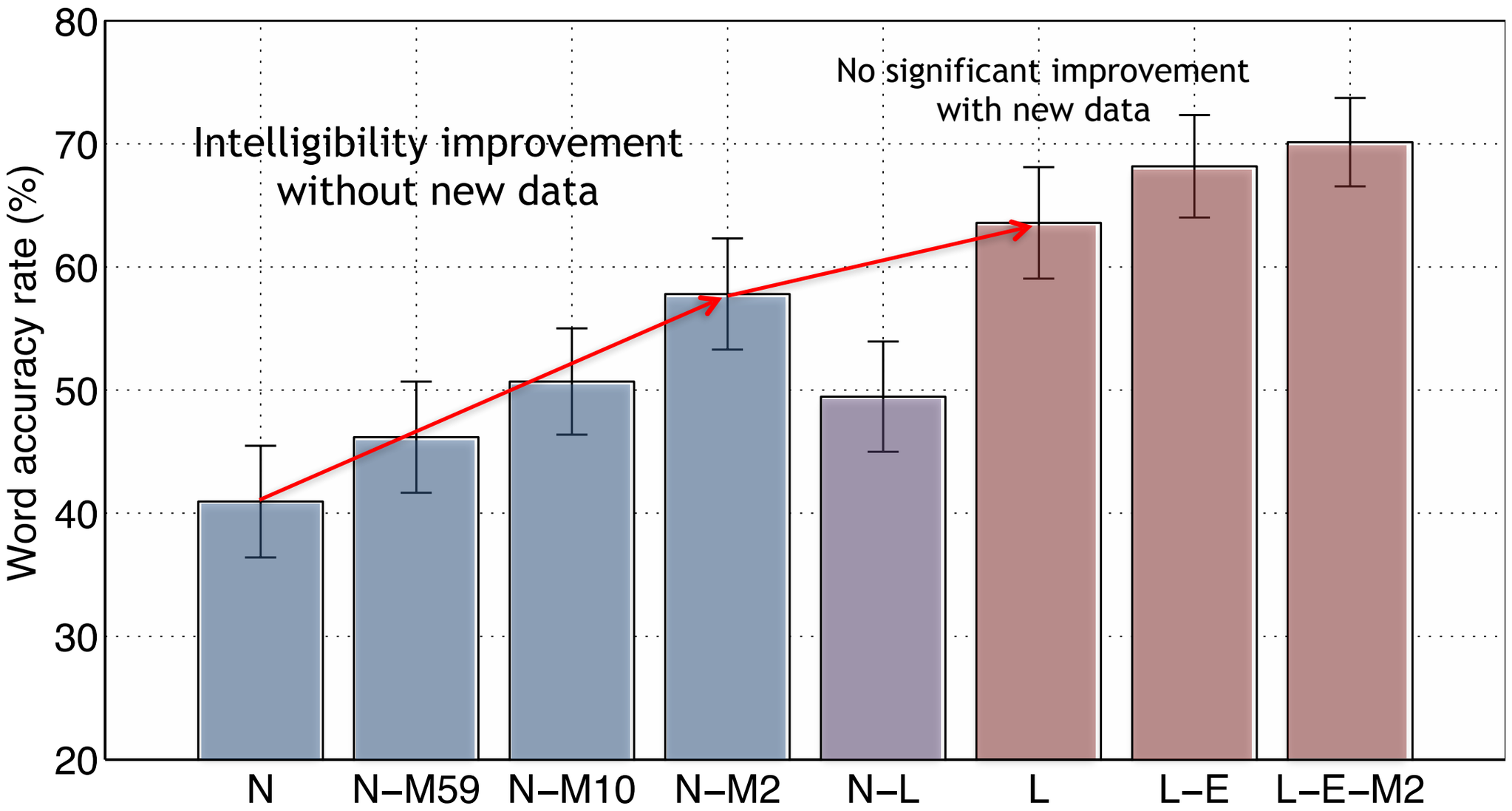
Mel cepstral modifications based on the GP



[Valentini et al., 2012]

[Valentini et al., 2012]

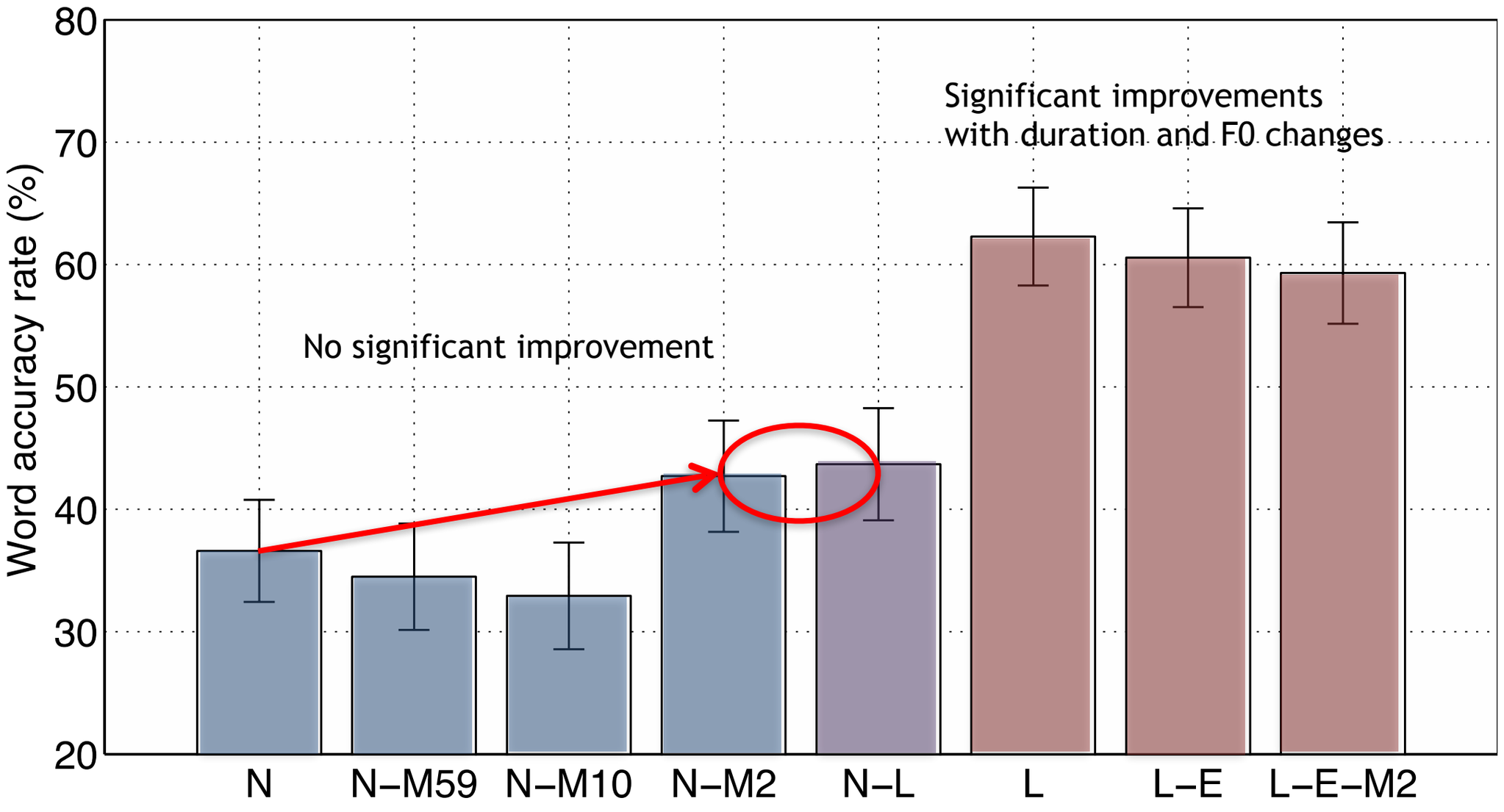
Intelligibility scores in speech-shaped noise



"The birch canoe slid on the smooth planks."

[Valentini et al., 2012]

Intelligibility scores in competing speaker



“The birch canoe slid on the smooth planks.” MALE SPEAKER

Conclusions

- Measures, evaluation and application
- Ultimately, listening experiments provide the gold standard (?)
- Measures tend to be proposed in a context
- In other contexts they might not work very well
- Know their limitations but exploit their advantages!
- Identify which condition the measures work reasonably well

Questions?

References

Intrusive measures

[Gray et al 1976] Distance measures for speech processing (LSD, CEP, LLR, IS)

[Tribolet et al. 1978] A study of complexity and quality of speech waveform coders (FWS)

[Klatt et al. 1982] Prediction of perceived phonetic distance from critical-band spectra: a first step (WSS)

[ANSI S3.5-1997] Methods for Calculation of the Speech Intelligibility Index (SII)

[IEC 2011] Objective rating of speech intelligibility by speech transmission index (STI)

[Kates et al 2005] Coherence and the speech intelligibility index (CSII)

[Rix et al 2001] Perceptual evaluation of speech quality (PESQ)

[Holube et al. 1996] Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model (NCM)

[Christiansen et al. 2010] Prediction of speech intelligibility based on an auditory preprocessing model (CD)

[Cooke et al. 2006] A glimpsing model of speech perception in noise (GP)

[Tang 2014] Speech intelligibility enhancement and glimpse-based intelligibility models for known noise conditions (DWGP)

[Taal et al. 2010] A short-time objective intelligibility measure for time-frequency weighted noisy speech (STOI)

[Zureck 1993] Acoustical Factors Affecting Hearing Aid Performance (BiSII)

[Wijngaarden et al. 2008] Binaural intelligibility prediction based on the speech transmission index

[Tang et al. 2016] Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms (BiDWGP)

Non intrusive measures

[ITU-T Rec. 2004] - The ITU-T Standard for single-ended speech quality assessment (P-563)

[Falk et al. 2010] A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech (SRMR)

[Falk et al. 2011] QUANTIFYING PERTURBATIONS IN TEMPORAL DYNAMICS FOR AUTOMATED ASSESSMENT OF SPASTIC DYSARTHRIC SPEECH INTELLIGIBILITY

[Santos et al. 2013] Objective speech intelligibility prediction for cochlear implant users in complex listening environments (SRMR-CI)

[Cosentino et al. 2013] A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals (BiSIM)

References

Evaluation

[Hu et al. 2007] Evaluation of objective measures for speech enhancement

[Taal et al. 2009] An Evaluation of objective quality measures for speech intelligibility prediction

[Chen et al. 2012] Predicting the intelligibility of vocoded speech

[Tang et al. 2016b] Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech

[Falk et al 2015] Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices

[VB et al., 2012] Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise

Application

[Schepker et al. 2013] Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression

[Taal et al. 2012] A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure

[Valentini et al., 2012] Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise

[Sauert et al. 2009] Near end listening enhancement optimised with respect to speech intelligibility index

[Aubanel et al. 2013] Information-preserving temporal reallocation of speech in the presence of fluctuating maskers

PhD theses

[Tang 2014] Speech intelligibility enhancement and glimpse-based intelligibility models for known noise conditions

[Taal 2013] Prediction and Optimization of Speech Intelligibility in Adverse Conditions

[Valentini-Botinhao 2013] Intelligibility enhancement of synthetic speech in noise

[Christiansen 2012] Listening in adverse conditions: Masking release and effects of hearing loss

[Al Dabel 2016] Intelligibility model optimisation approaches for speech pre-enhancement

Code for measures

- SII: http://www.sii.to
- eSII, DWGP: http://listening-talker.org/legacy.html
- WSS, LLR, IS, CEP, fwSRN: https://uk.mathworks.com/support/books/book48837.html
- PESQ and composite measure: http://ecs.utdallas.edu/loizou/speech/software.htm
- STOI: http://siplab.tudelft.nl/users/cees-taal
- SRMR and SRMR-CI: https://github.com/MuSAELab/SRMRToolbox
- BiSTI, BiSII, BiNCM, BiDWGP: https://dx.doi.org/10.17866/rd.salford.3172921
- BiSTOI: http://kom.aau.dk/project/Intelligibility/