STQ Workshop

Single-ended prediction of listening effort for smart speakers

Rainer Huber & Jan Rennies

Fraunhofer IDMT

Motivation

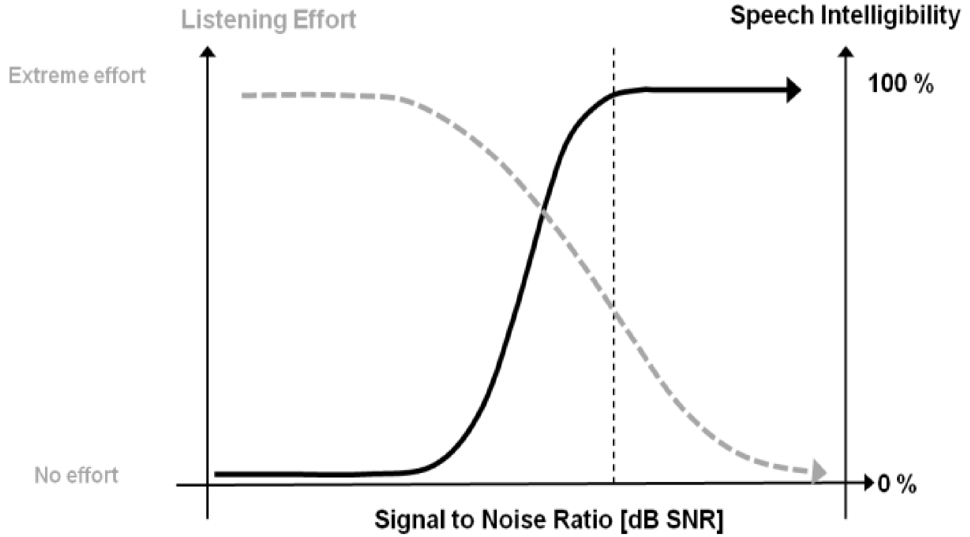
- Growing need for evaluating the speech output of smart speakers
- Current standardization specifications (ETSI TS 103 504) do not yet comprise assessments of quality, intelligibility, or listening effort of the (synthetic) speech output
- Recent studies have presented promising approaches of speech quality and naturalness of synthesized voices using single-ended („non-intrusive“) models (NISQA, NISQA-TTS; Mittag & Möller)

© Fraunhofer IDMT / Anika_Bödecker

- Our goal: Develop an instrumental, single-ended tool to measure listening effort for smart speaker speech output under realistic acoustic conditions

Why listening effort?

- Listening effort can still be affected by changes in noise levels at realistic SNRs, where speech intelligibility is already close to 100%
- Such conditions are often more representative of everyday-life listening conditions than very low SNRs (Smeds et al., 2015)



Klink et al. (2012)

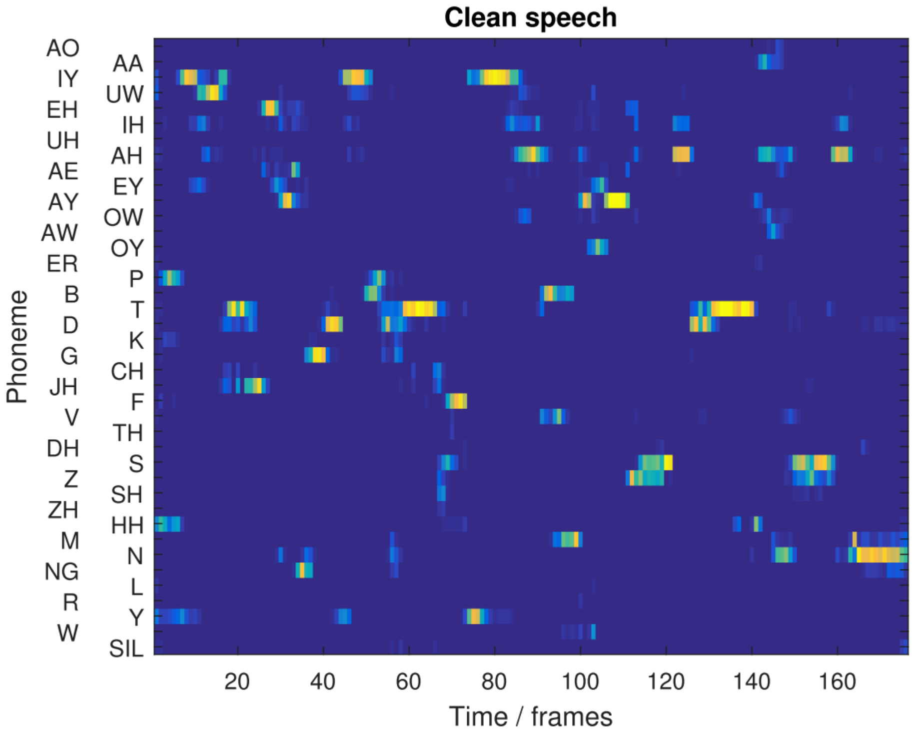
Approach

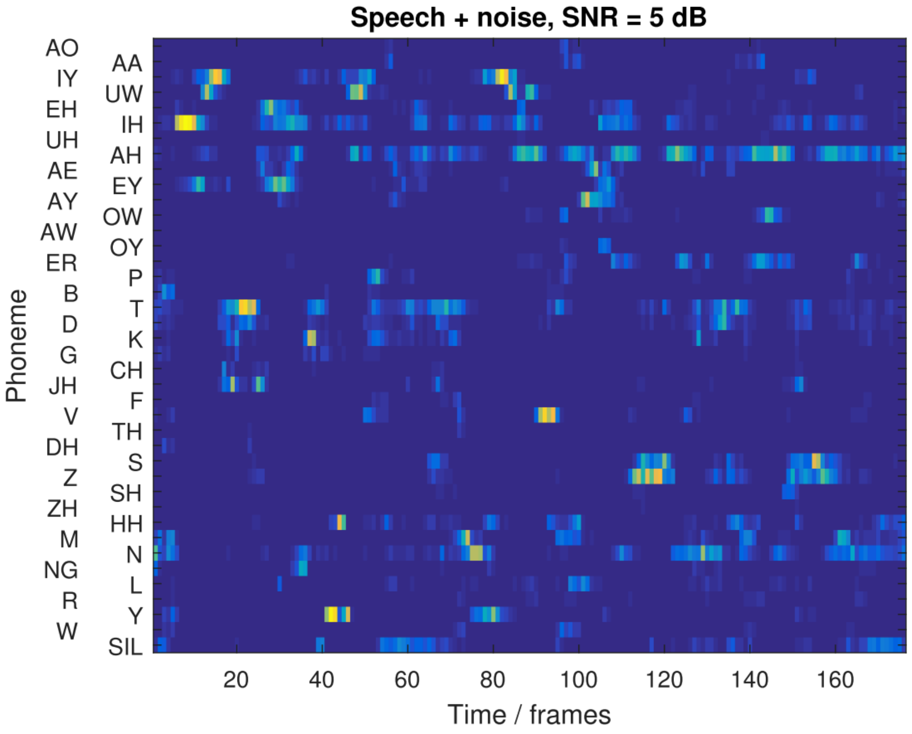
- Generate an audio database of simulated smart speaker voice output under realistic acoustic conditions
- Conduct listening tests to obtain ground truth date of subjectively perceived listening effort
- Validate and develop instrumental measures

Listening effort prediction from acoustic parameters

LEAP model (Huber et al., 2018a,b; Rennies et al., 2022)

- Employs a DNN-based automatic speech recognition engine, but does not evaluate the transcript of the voice recording, but instead an interim quantity, the so-called phoneme-posterior-probability (“posteriorgrams”)





Listening effort prediction from acoustic parameters

LEAP model (Huber et al., 2018a,b; Rennies et al., 2022)

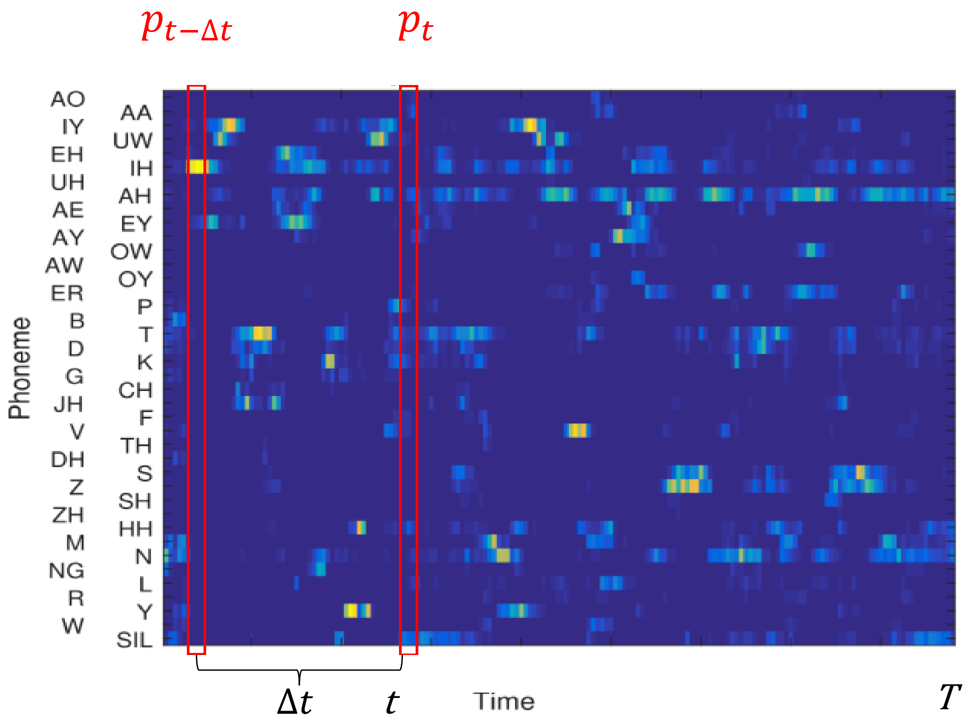
- Quantifies the degree of posteriorgram "smearing" by noise and/or other distortions by computing the „Mean Temporal Distance" („M-Measure"; Hermansky et al., 2013):

$$M(\Delta t) = \frac{1}{T - \Delta t} \sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t),$$

with

$$D(x, y) = \sum_{i=1}^{N} x(i) \log\left(\frac{x(i)}{y(i)}\right) + \sum_{i=1}^{N} y(i) \log\left(\frac{y(i)}{x(i)}\right)$$

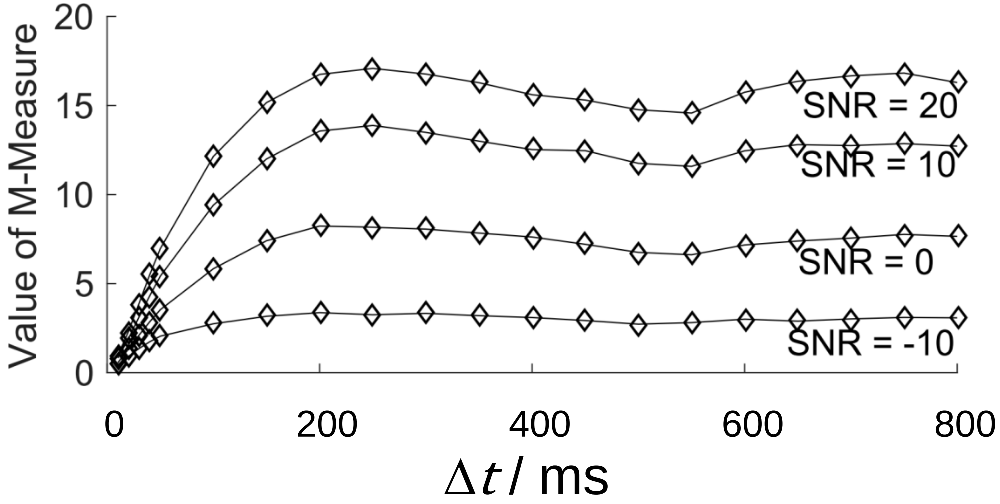
Kullback-Leibler divergence, aka KL „distance"



Listening effort prediction from acoustic parameters

LEAP model (Huber et al., 2018a,b; Rennies et al., 2022)

- Final predictor from obtained by averaging across multiple time-shifts
- Can be mapped onto scales as used in subjective listening tests



$$\bar{M} := \frac{1}{10}\sum_{n=1}^{10} M(300\text{ms} + n \cdot 50\text{ms})$$

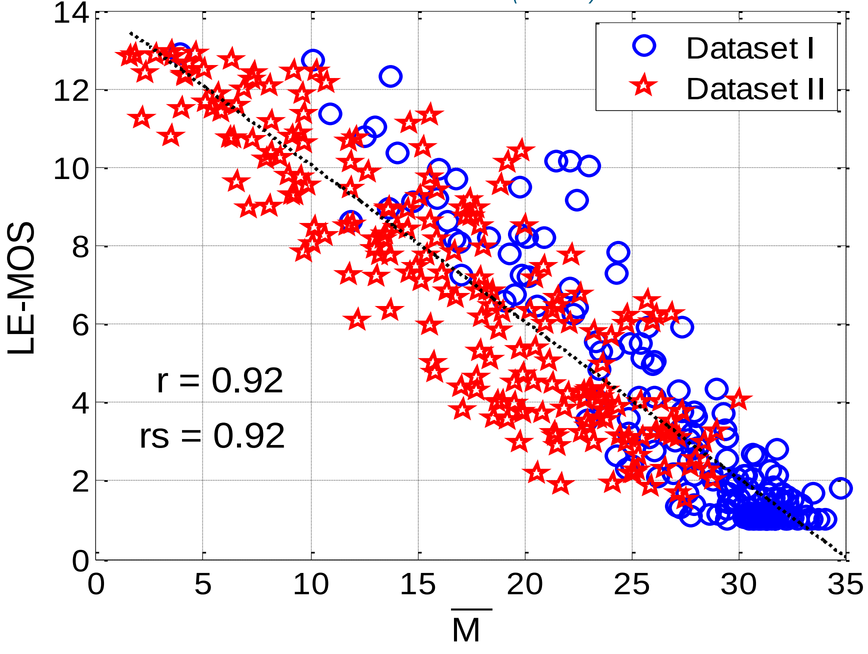
Listening effort prediction from acoustic parameters

Earlier validations using natural speech

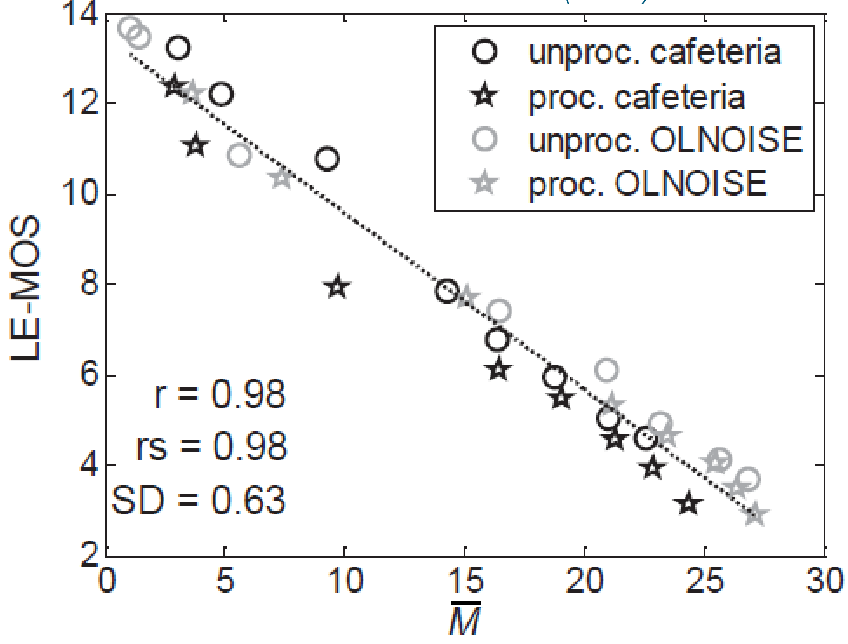
- High agreement between $\overline{M}$ and subjectively assessed listening effort of 450 TV audio clips (≈10s) with various backgrounds and SNRs

- Also high agreement between $\overline{M}$ and subjectively assessed listening effort for noisy speech processed by non-linear speech enhancement

Huber et al. (2020)



Huber et al. (2018)



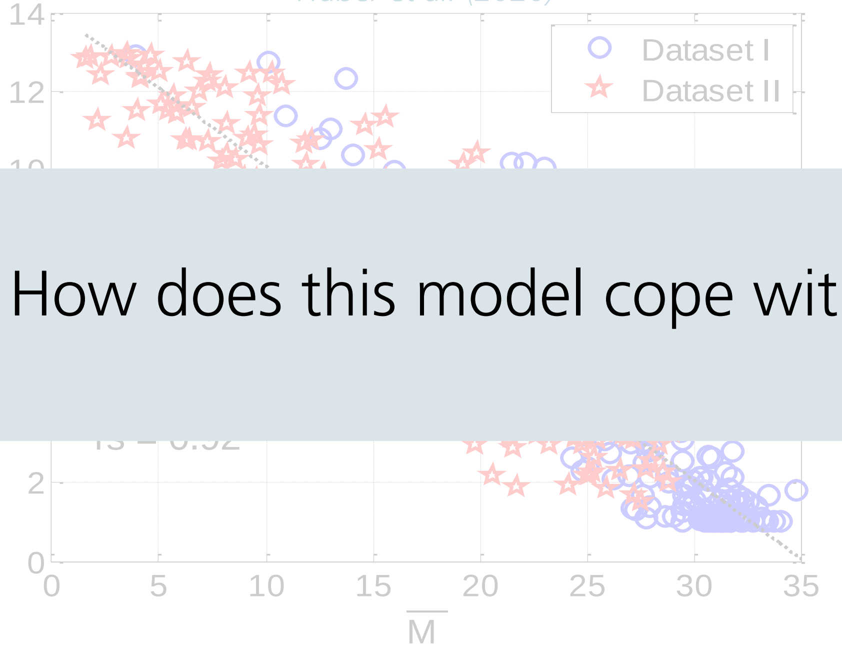
Fraunhofer IDMT

Listening effort prediction from acoustic parameters

Earlier validations using natural speech

- High agreement between $\overline{M}$ and subjectively assessed listening effort of 450 TV audio clips (≈10s) with various backgrounds and SNRs

Huber et al. (2020)



- Also high agreement between $\overline{M}$ and subjectively assessed listening effort for noisy speech processed by non-linear speech enhancement

Huber et al. (2018)

How does this model cope with synthetic speech in realistic listening conditions?

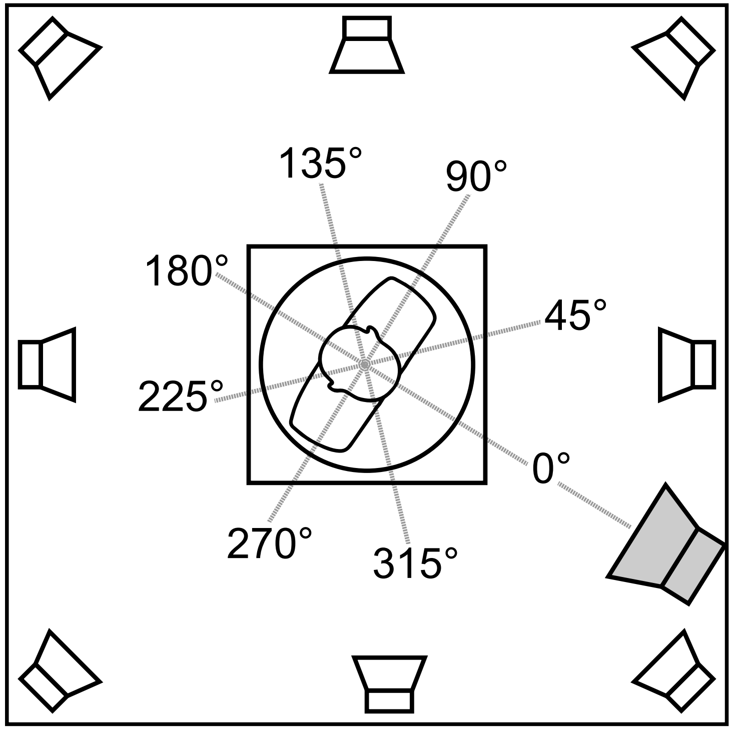
Subjective listening effort assessment

Methods

- Natural (standardized) speech stimuli from ETSI TS 103 281 and ITU-T Rec. P.501
- Synthetic speech stimuli
  - Exp 1: high-quality TTS systems, same sentences
  - Exp 2: TTS systems of different quality

- Standardized and combined reproduction of …
  - Noise → ETSI TS 103 224
  - Reverberation → ETSI TS 103 557

- Artificial head recordings with different simulated distances by project partner HEAD acoustics:
  1m (real), 3m (DRR ~ -10 dB), 10m (DRR ~ -20 dB), ∞ (only reverb)
- Separate recordings of direct sound, reverb, and noise for later mixing



Subjective listening effort assessment

Exp 1: Stimuli

Talkers:

- ITU-T P.501, female
- ITU-T P.501, male
- High-quality TTS, female
- High-quality TTS, male

RT_{60}:

- „medium“: 0,54s
- „high“: 1,2s
- „max“: 2,3s

Noise	Reverb	Distance/m	ΔSNR/dB
no	dry	1	
	max	3	
		10	
sink	medium	3	0
			-6
			-12
super market cashier	high	1	0
			-5
			-10
in train	medium	3	0
			-5
			-10
in bus	medium	1	0
			-5
			-10
office	high	3	0
			-12
train station	max	1	0
		3	0

Noise	Reverb	Speech	SNR / dB
in train	medium	TTS medium quality, female	-10
			-5
			0
			5
			10
OLNOISE	dry	OLSA	-8
			-3
			2
			7
			12

Overall 91 test signals of about 8-9s

Subjective listening effort assessment

Methode

- Assessment of subjectively perceived listening effort on 14-point categorical scale (Krüger et al., 2017)
- 18 normal-hearing listeners (31,8±8 years)
- Headphone presentation

no effort

very little effort

little effort

moderate effort

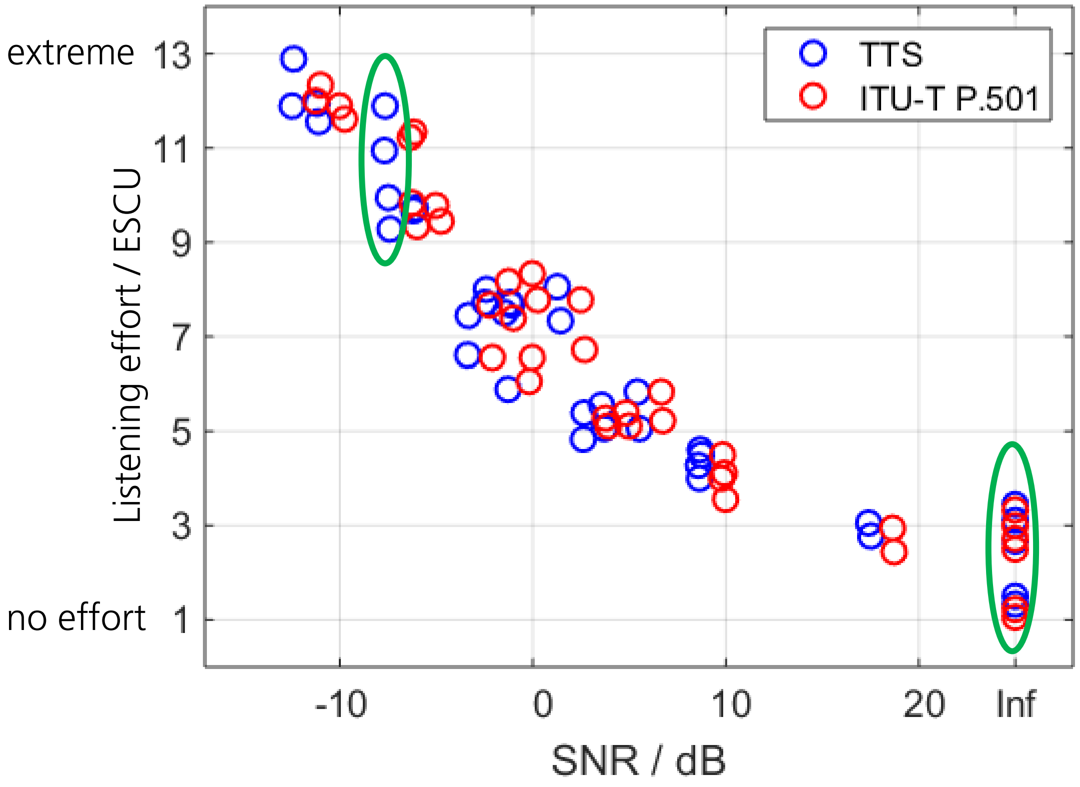
considerable effort

very high effort

extreme effort

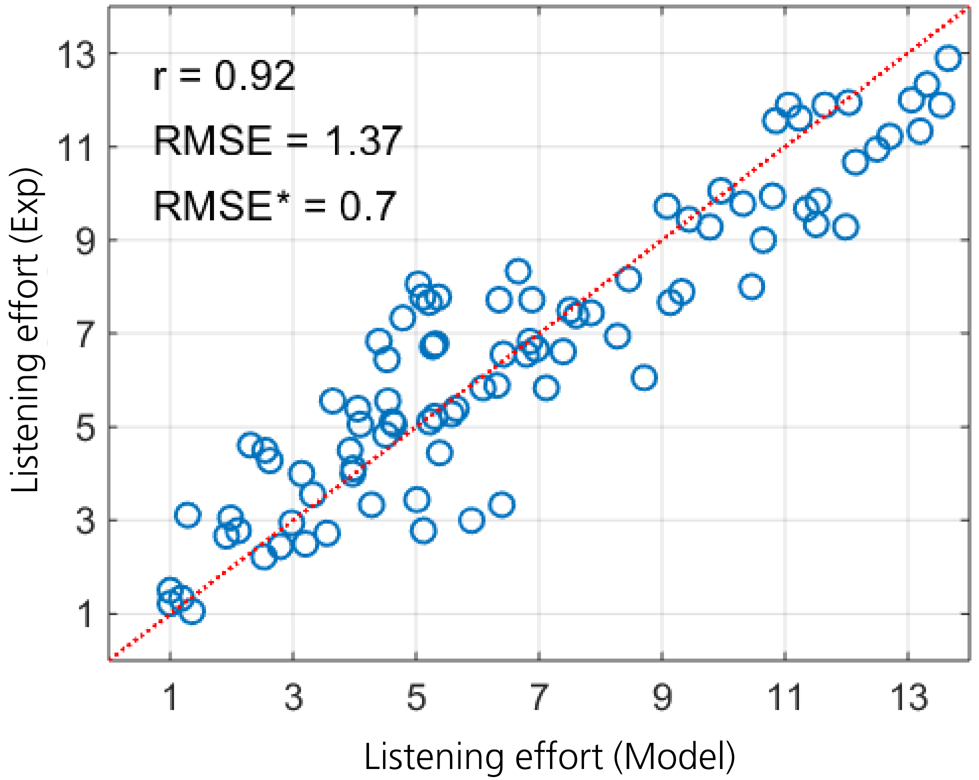
Subjective listening effort assessment
Exp 1: Results

- Subjects made use of entire rating scale
- No apparent difference between natural (ITU-T P.501) and synthetic (TTS) talkers
- Different noise types and reverb produce different (mean) listening effort ratings at the same SNR



Comparison of subjective and predicted listening effort

- Mapping of M-Measure → listening effort scale taken from earlier studies, not adapted to current data
- Very high agreement between model predictions and mean subjective ratings
- So far, LEAP does not comprise an explicit binaural processing stage, binaural effects simplified by „better ear listening“

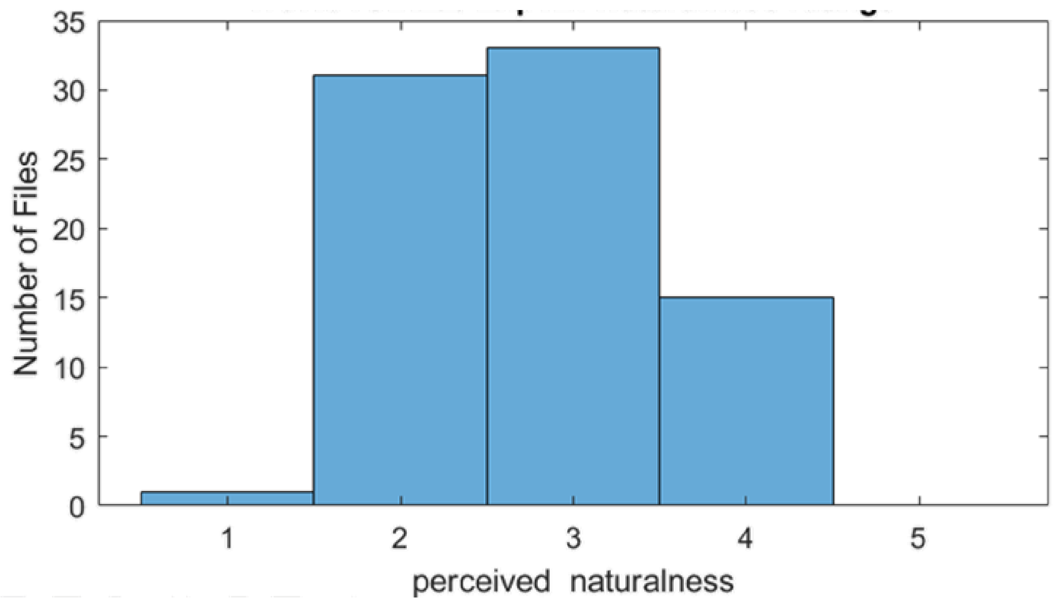


Subjective listening effort assessment

Exp 2: Methods

- Larger variety in TTS quality, from very unnatural to very natural
- 20 normal-hearing listeners (21-30 years)
- Different sentences uttered by 20 different artificial talkers:

  Anna, Birgit_low, Conrad, Dieter_high, Dieter_normal, Google_basic_A_pitch, Google_basic_B_pitch, Google_basic_E_norm, Google_Basic_E_speed_mod, Google_Basic_E_speed_pitch_mod, Google_WaveNet_E_normal, Google_WaveNet_F_speed_pitch_mod, Hans, Hedda, iSpeech_female, iSpeech_male, Petra, Siri_female, Siri_male, Vicki

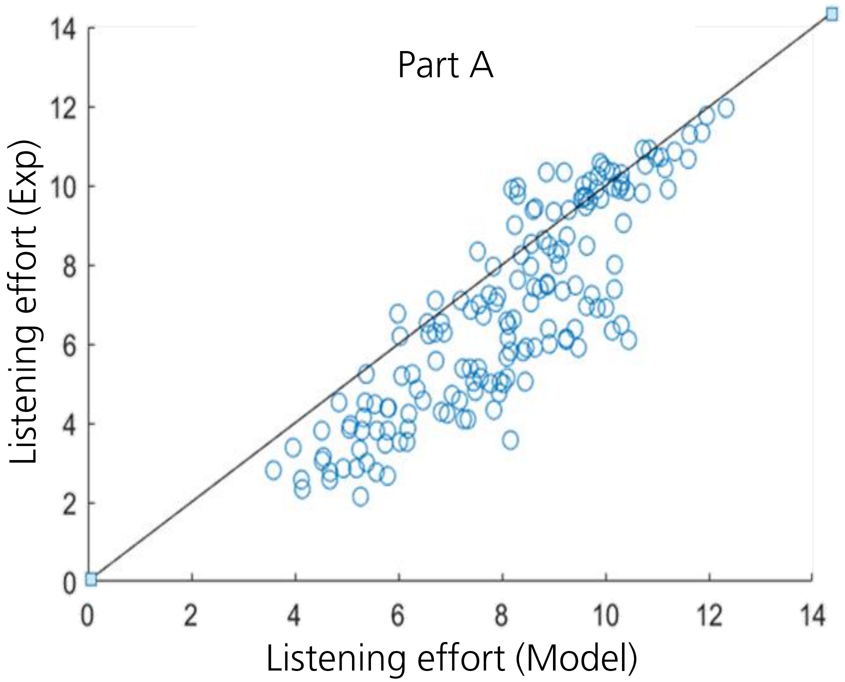


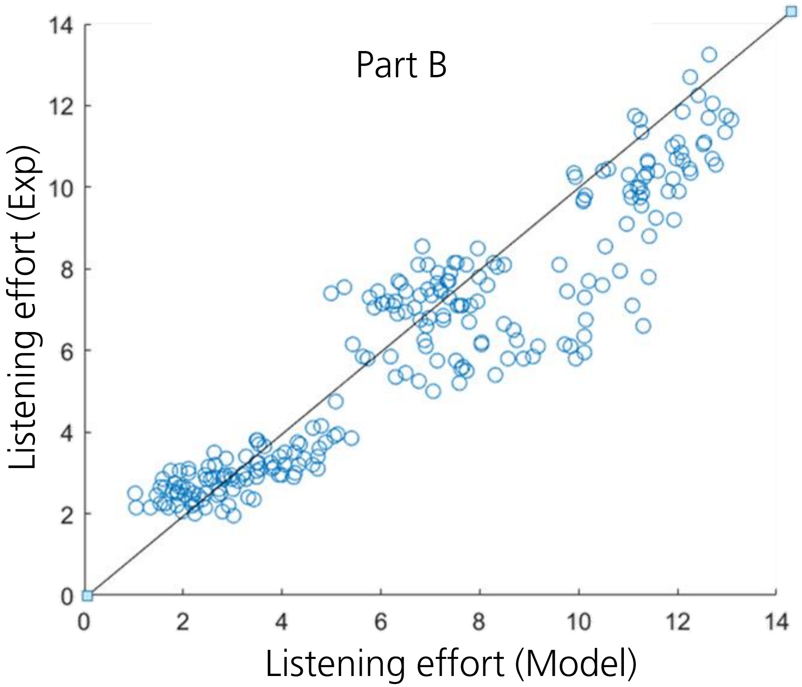
- Different noise types, different SNRs
  - Part A: sink, office
  - Part B: train, sink, train, cafeteria, metal grinder, different lateral positions relative to target speech

Subjective listening effort assessment

Exp 2: Results

- Good general agreement between model and experiment in both parts, slight overestimation of listening effort on average
- „Better-ear“ model seems sufficient also for strongly lateralized noise sources





Conclusions

- Prediction model based on ASR technology procudes accurate listening effort predictions for a variety of listening conditions
  - No adaptation of mapping function to new data
  - No strong differences between natural speech and high-quality synthetic speech
  - Very low-quality TTS likely requires other assessment methods
  - Additional binaural processing stage probably not required / additional complexity not justified
- Promising approach as single-ended assessment tool for smart speaker voice output under realistic acoustic conditions including noise and reverb

Thank you very much!

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Funded by the German Federal Ministry for Economic Affairs and Climate Action
Project BASS, FKZ: KK5048502GR0

jan.rennies-hochmuth@idmt.fraunhofer.de

Fraunhofer
IDMT

Referemces

Hermansky, H. Variani, E., & Peddinti, V. (2013). Mean temporal distance: predicting ASR error from temporal properties of speech signal. In Proc. IEEE Conf. Acoust. Speech, Signal Process. (ICASSP), 7423–7426, DOI: 10.1109/ICASSP.2013.6639105.

Huber, R., Baumgartner, H., Krishnan, V.N., Goetze, S., & Rennies, J. (2020). Single-ended Prediction of Listening Effort for English Speech. Fortschritte der Akustik – DAGA 2020, Hanover, Germany.

Huber, R., Krüger, M., & Meyer B.T. (2018a). Single-ended prediction of listening effort using deep neural networks, Hear. Res., vol. 359. pp. 40-49.

Huber, R., Pusch, A., Moritz, N., Rennies, J., Schepker, H., & Meyer, B.T. (2018b). Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system. Speech Communication; 13th ITG-Symposium, Oldenburg, Germany, 2018, 86–90.

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017a). Development of an adaptive scaling method for subjective listening effort. Journal of the Acoustical Society of America 141, 4680–4693, DOI: 10.1121/1.4986938.

Referemces

Mittag, G. & Möller, S. (2021). Deep Learning Based Assessment of Synthetic Speech Naturalness. https://doi.org/10.48550/arXiv.2104.11673

Rennies, J., Röttges, S., Huber, R., Hauth, C.F. & Brand, T. (2022). A joint framework for blind prediction of binaural speech intelligibility and perceived listening effort. Hearing Research 426. https://doi.org/10.1016/j.heares.2022.108598

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. Journal of the American Academy of Audiology 26, 183–196. DOI: 10.3766/jaaa.26.2.7.

