

BSS EVAL OR PEASS? PREDICTING THE PERCEPTION OF SINGING-VOICE SEPARATION

Dominic Ward, Hagen Wierstorf, Russell Mason, Emad M. Grais, Mark D. Plumbley

Audio Examples { bit.ly/2GutUKR }

Centre for Vision, Speech and Signal Processing | Institute of Sound Recording | University of Surrey, Guildford, UK



UNIVERSITY OF SURREY

EPSRC

Engineering and Physical Sciences Research Council

Objective Evaluation of Audio Source Separation

- Separating the singing-voice from music is a difficult task, however, deep-learning methods show significant improvements over traditional techniques such as NMF and ICA.
- Source separation introduces distortions and artifacts, which degrades the perceived sound quality.
- Trade-off between degree of separation and sound quality

How to evaluate separation performance?

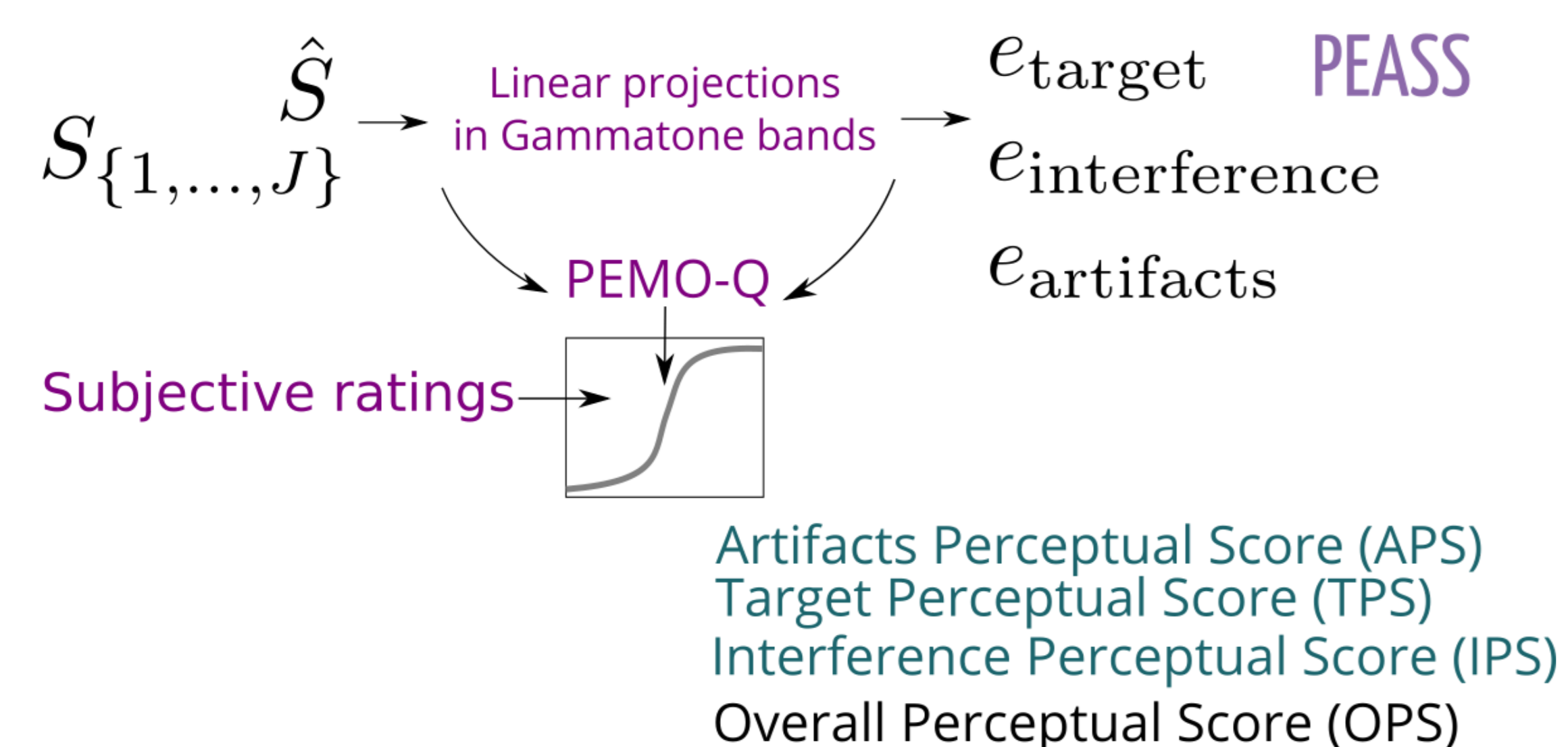
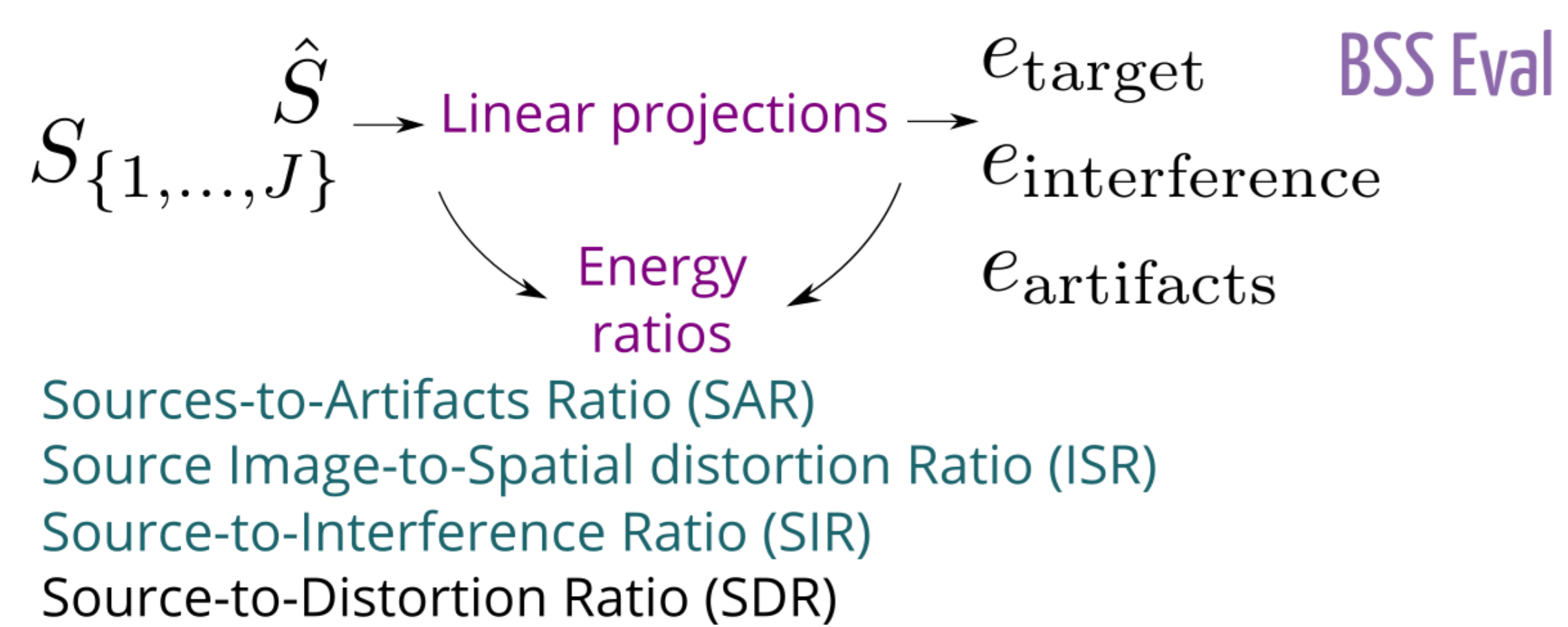
Few researchers conduct listening assessments, but instead resort to objective toolkits:

- BSS Eval**¹: Blind Source Separation Evaluation toolbox
- PEASS**²: Perceptual Evaluation methods for Audio Source Separation

Both approaches based on distortion decomposition between estimated source \hat{S} and target source S :

$$\hat{S} - S = e_{\text{target}} + e_{\text{interference}} + e_{\text{artifacts}}$$

Error components estimated through least-squares projections of estimated and true sources.



¹ Vincent et al. (2006) { [10.1109/tasl.2011.2109381](https://doi.org/10.1109/tasl.2011.2109381) }

² Emiya et al. (2012) { [10.1109/tasl.2011.2109381](https://doi.org/10.1109/tasl.2011.2109381) }

Subjective Listening Assessment

Can these toolkits be used to predict the perception of singing-voices extracted by modern source separation systems?

- Need more evidence to address suitability of BSS Eval
- Few studies have investigated generalisation of PEASS

Methodology

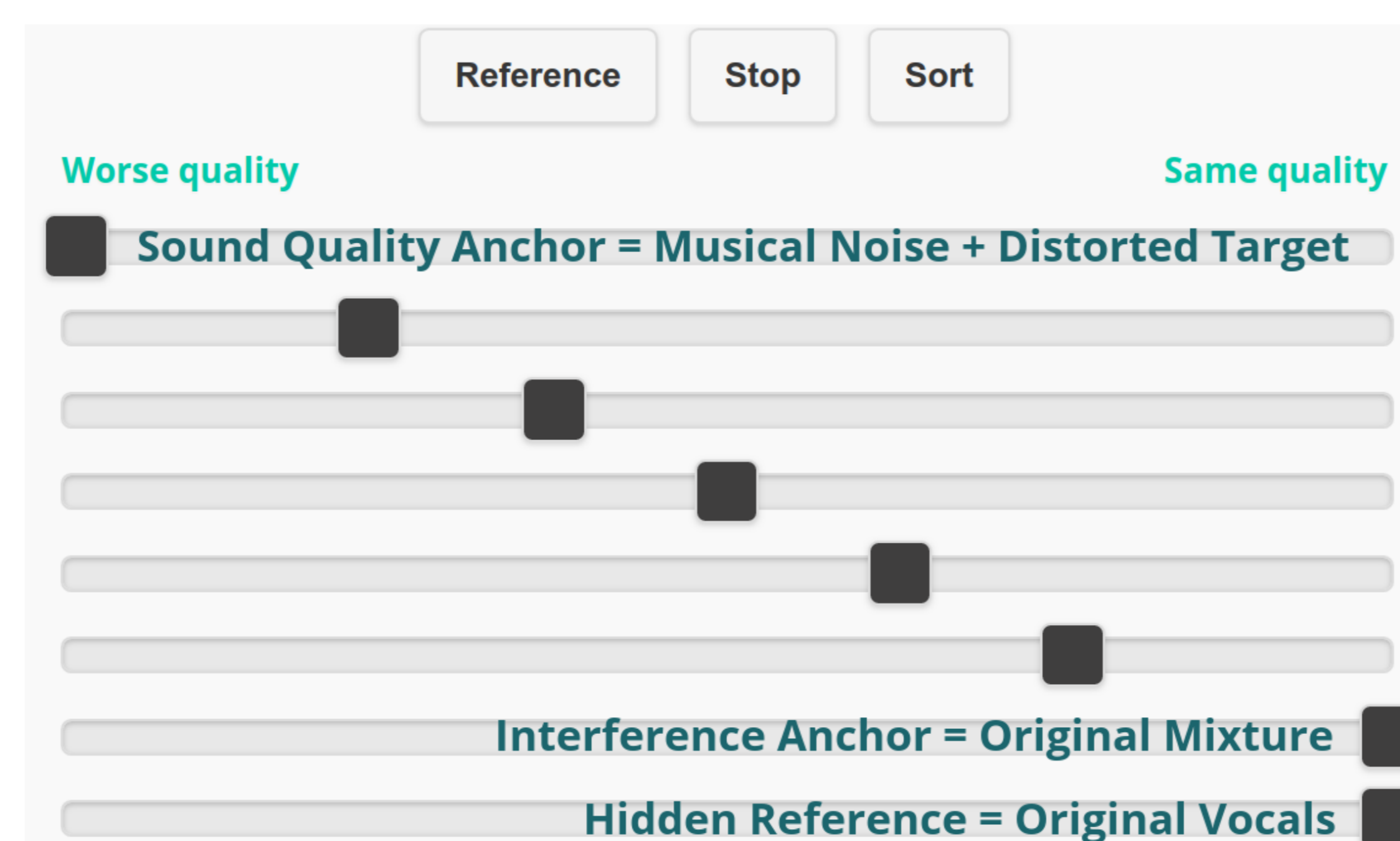
Task 1: Sound Quality

Sound quality relates to the amount of artifacts or distortions that you can perceive. These can be heard as tone-like additions, abrupt changes in loudness, or missing parts of the audio.

Task 2: Interference

Interference describes the loudness of the instruments compared to the loudness of the vocals. For example, 'strong interference' indicates a strong contribution from other instruments, whereas 'no interference' means that you can only hear the vocals. Interference does not include artifacts or distortions that you may perceive.

- 24 Listeners performed a MUSHRA-style experiment
- 16 songs, using *singing-voice* as the target source
- Listeners compared 5 algorithms selected pseudorandomly from 21 systems for each song³
- Sound quality anchor and interference anchor included



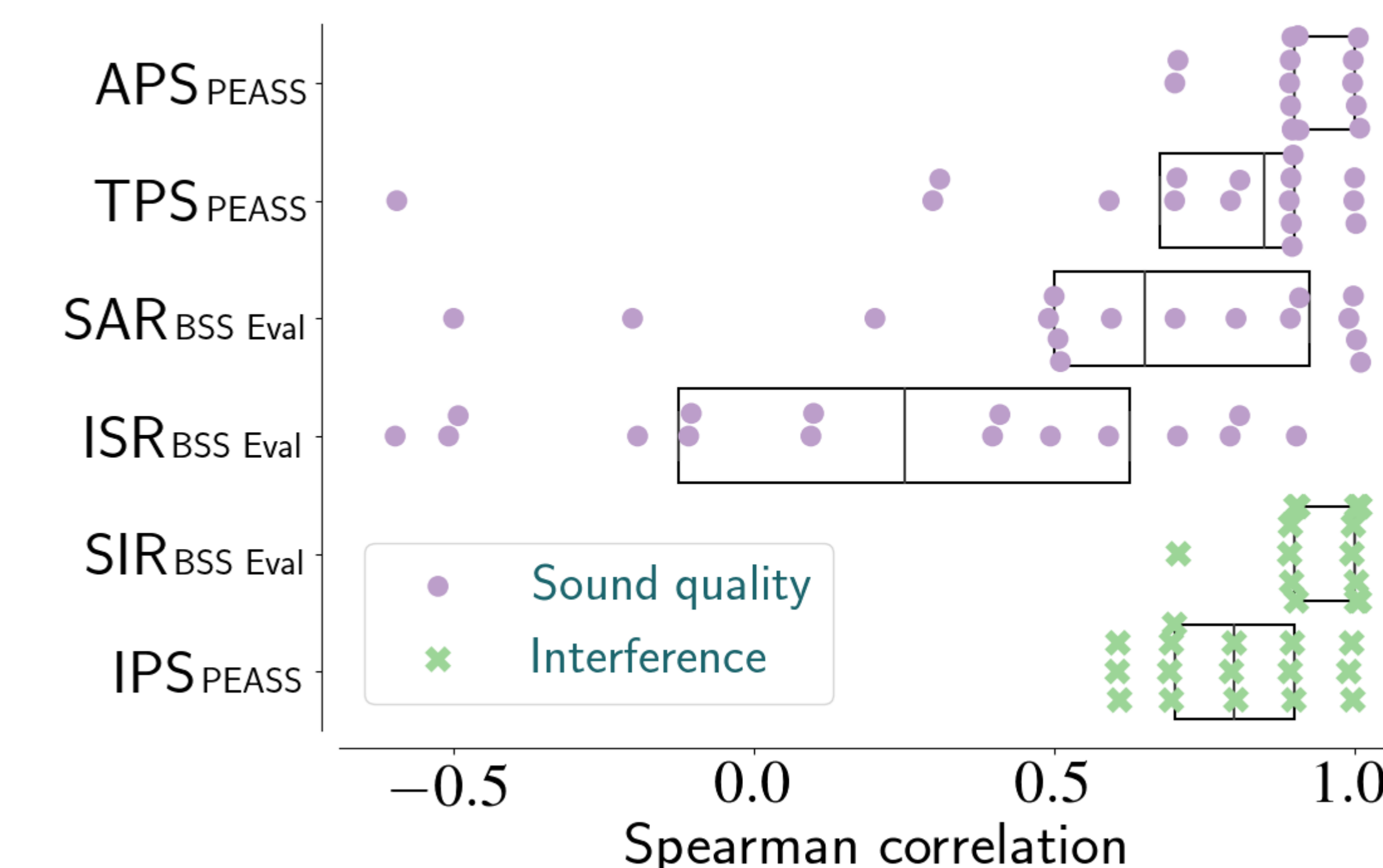
Interface for Task 1. Examples at { bit.ly/2GutUKR }

³ SiSEC 2016 { <http://sisec17.audiolabs-erlangen.de> }

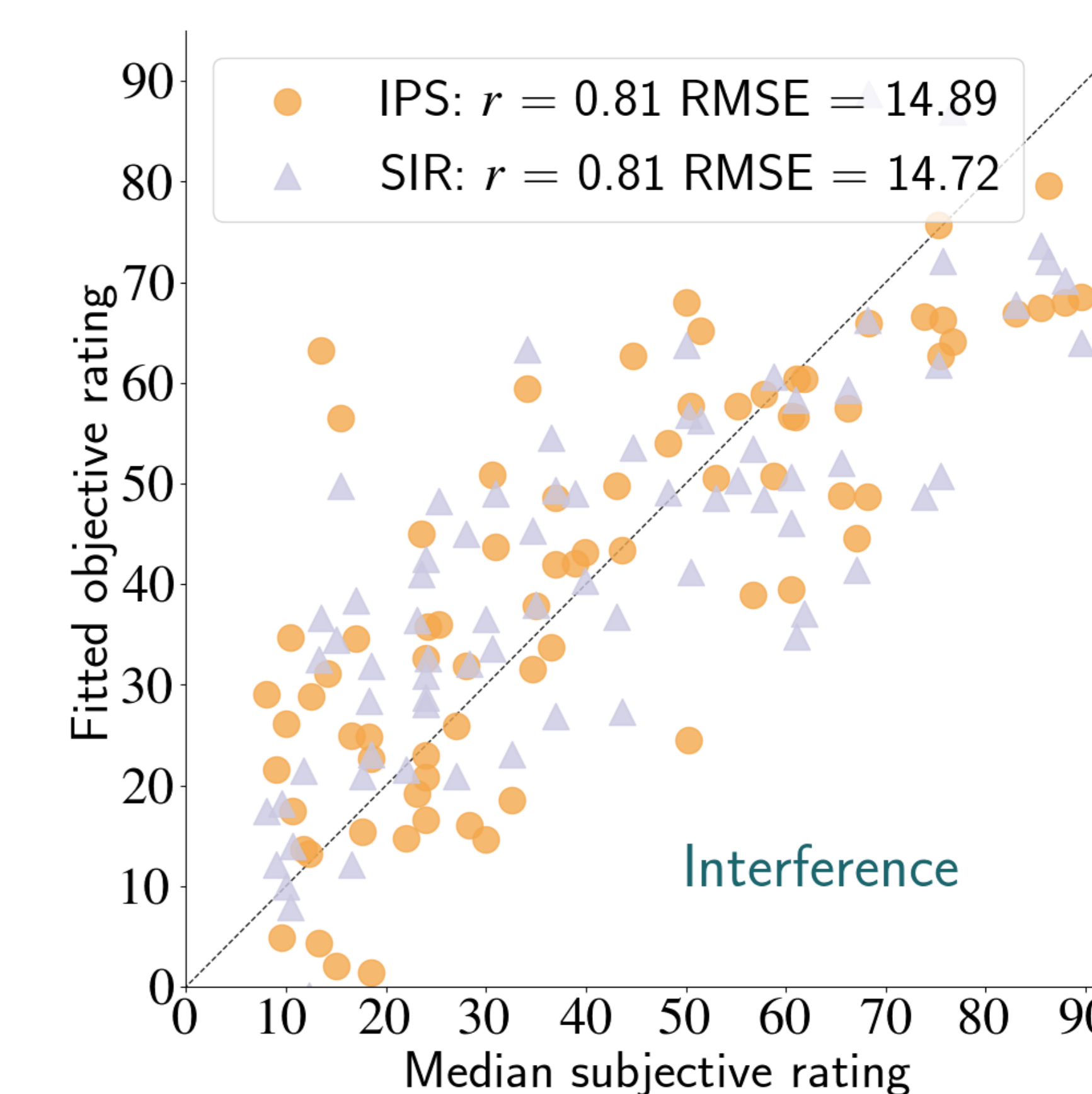
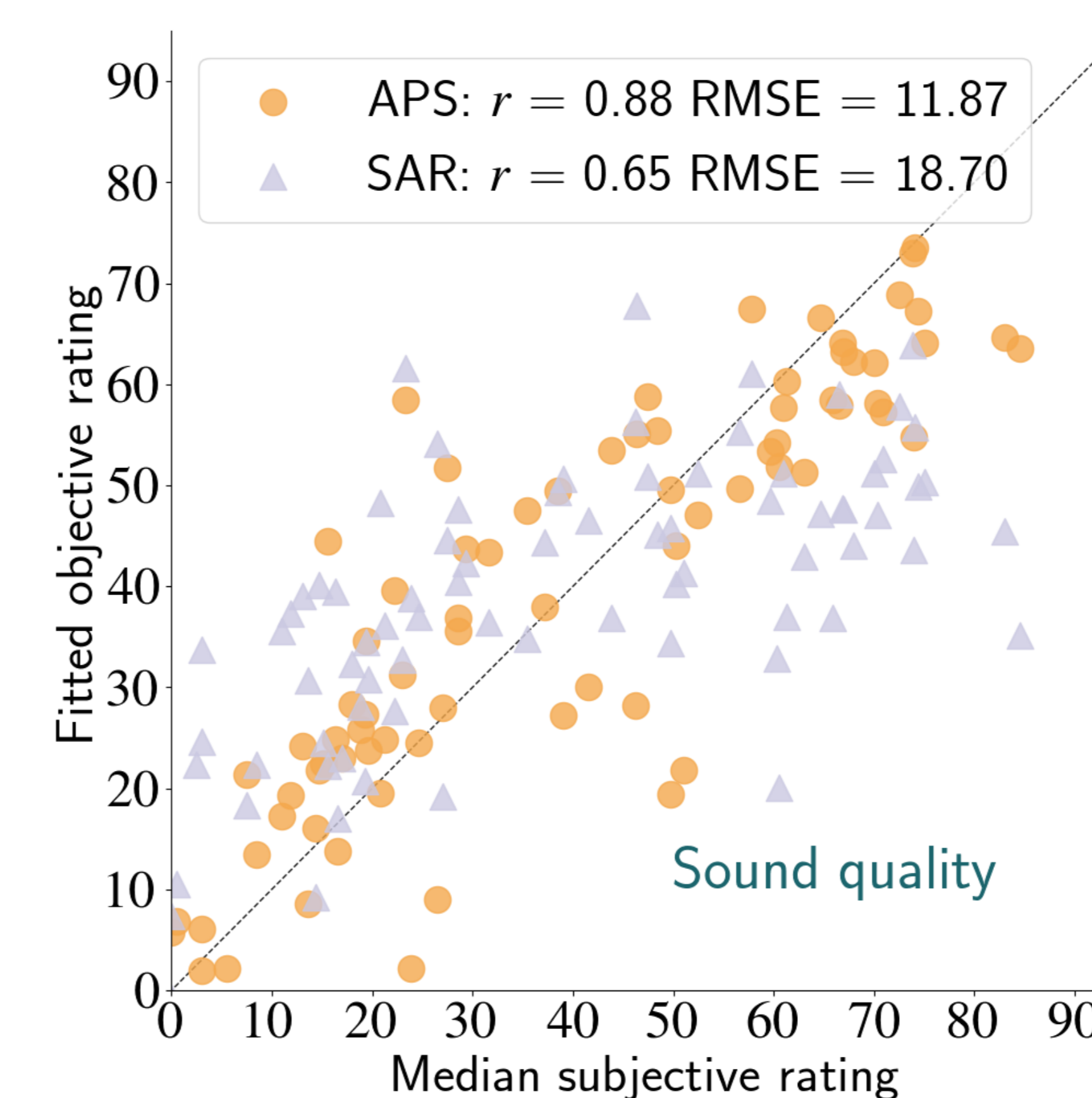
Results

Within-song Spearman correlations

- Spearman correlation between objective measures and medians of subjective ratings
- Assesses rank agreement of 5 separation algorithms per song
- 16 within-song correlations per metric



Linear-Fitted Objective Measures vs Subjective Medians



Conclusions and Reflections

- Important to reinforce attribute definitions with audio examples
- APS of the PEASS toolkit showed the strongest predictive ability
- IPS (PEASS) and ISR (BSS Eval) were comparable in performance
- Metrics far from perfect (large RMSE) when considering 100-point scale
- Remapping of features necessary to better predict the perceptual scales used here
- Next time, emphasise **overall sound quality** as some listeners focused only on voice
- We are currently running **similarity** experiments for assessing SDR and OPS