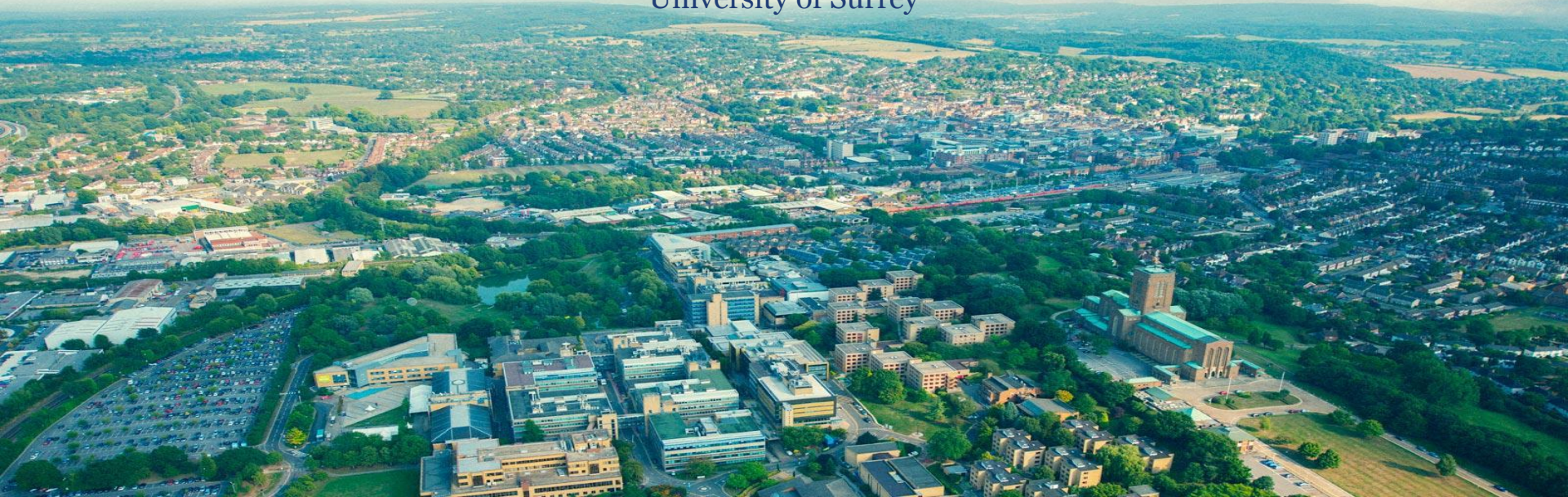
Evaluation of Musical Audio Source Separation: Objective and Subjective

Russell Mason, Ryan Chungeun Kim, Dominic Ward
University of Surrey



Musical audio source separation - aims

- Aim of musical audio source separation is to extract individual sources that make up a music mix
- Many applications, including:
  - soloing
  - karaoke
  - remixing
  - spatialisation
  - musical analysis
- Most of these are intended for humans to listen to
- Perceived quality is therefore essential

Musical audio source separation - our work

- MARuSS (Musical Audio Repurposing using Source Separation) project aims
  - How do we optimally separate musical sources from a mixture?
  - What are the perceived artefacts from imperfect separation?
  - How can we objectively measure the separation performance?
  - How does the separation quality affect the ability to remix the audio?

https://cvssp.github.io/maruss-website/

Background and context

- Evaluation is essential to the development of source separation algorithms
  - Subjective evaluation: human judgements
  - Objective evaluation: computational algorithms
- Machine learning is dependent on accurate evaluation
- Encourage:
  - better understanding of available methods;
  - better scientific methodology;
  - more careful consideration of details of results; and
  - consideration of perception.

Objective measurements

Objective performance evaluation

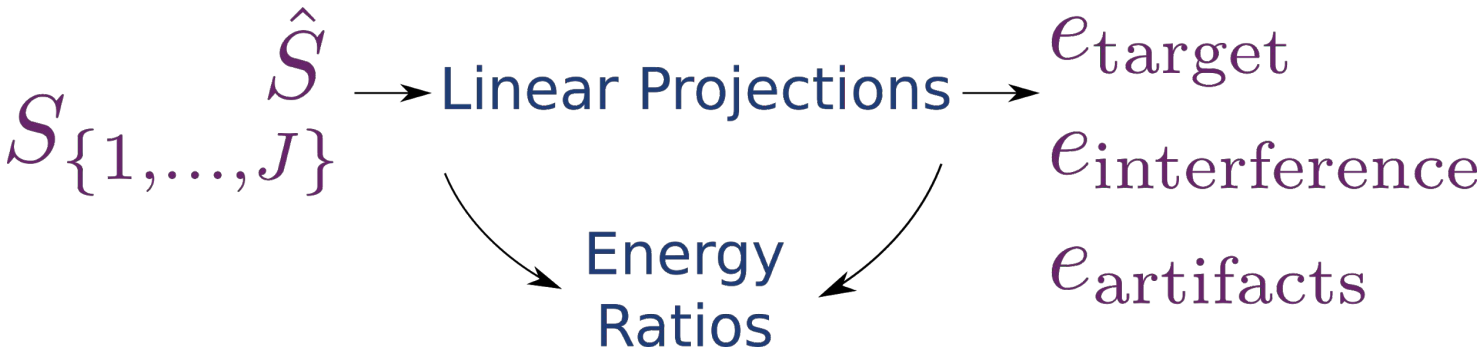
- Objective metrics allow for efficient and low-cost performance measurement

- Separation evaluation criteria depends on source and intended application:
  - Speech intelligibility
  - Object-based audio for television
  - Upmixing and remixing music

- In many cases, especially for music, the final judgement is the human listener
  - Thus, ideally quantify separation quality using metrics that correlate with human judgements

Overview of BSS Eval

- BSS Eval: Blind Source Separation Evaluation
- Based on distortion decomposition between estimated source and target source



Sources-to-Artifacts Ratio (SAR)

Source Image-to-Spatial distortion Ratio (ISR)

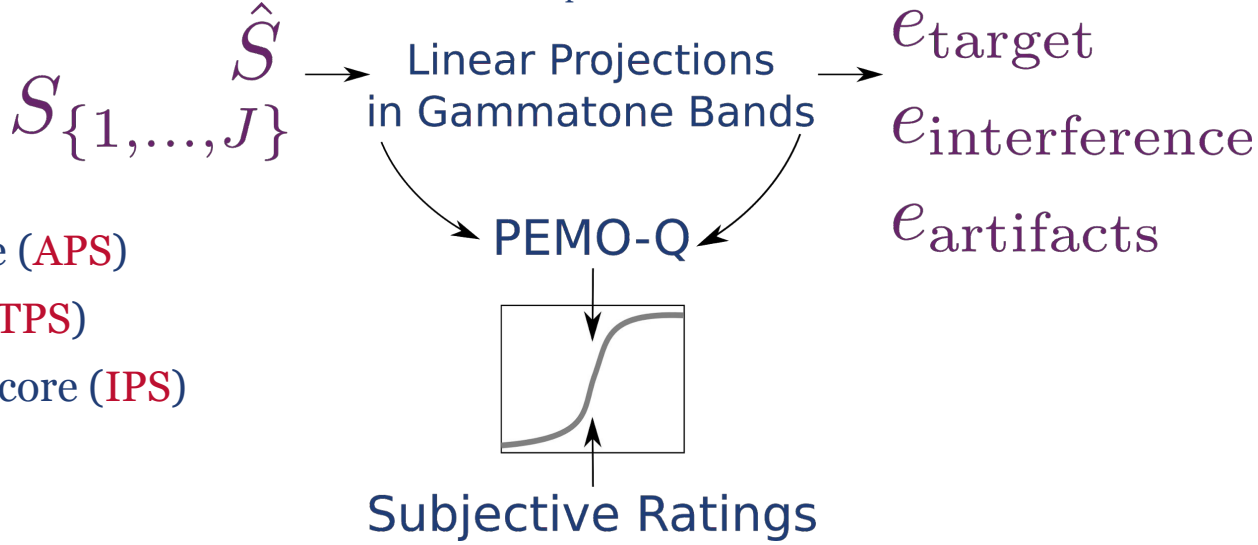
Source-to-Interference Ratio (SIR)

Source-to-Distortion Ratio (SDR)

Vincent et al. (2006; IEEE Trans)

Overview of PEASS

- PEASS: Perceptual Evaluation methods for Audio Source Separation
- Auditory motivated approach to error decomposition:
  - Gammatone analysis/synthesis filter bank
  - Perceptual Similarity Measure from PEMO-Q auditory model used to predict saliency of distortions
  - Nonlinear mapping is applied to combine the salience features onto final outputs



Artifacts-related Perceptual Score (APS)

Target-related Perceptual Score (TPS)

Interference-related Perceptual Score (IPS)

Overall Perceptual Score (OPS)

Emiya et al. (2012; IEEE Trans)

Case study: SiSEC 2016 MUS challenge

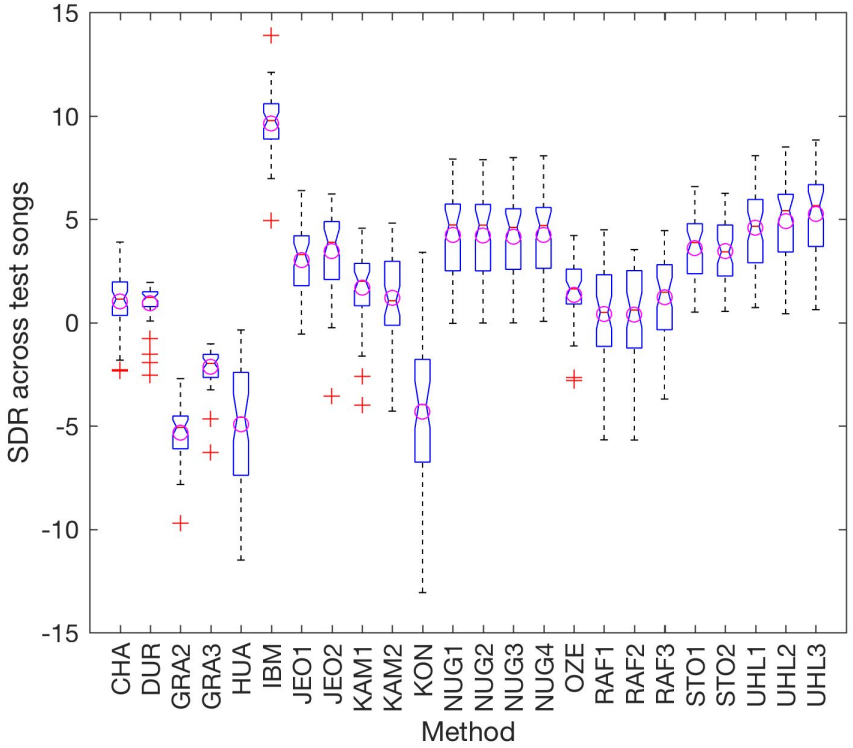
- Signal Separation Evaluation Campaign (see Antoine's talk on Wednesday)
  - Document the progress of the source separation community
  - Serve as a reference for a comparison of as many methods as possible
  - Provide training and evaluation data for the community

- MUS(ic): Professionally-produced music recordings
  - DSD100: 100 professionally mixed songs of different musical styles and genres
  - 50 test set, 50 development set
  - Four stereo image sources: bass, drums, vocals, and other
  - Lots of interest: 24 participants (separation algorithms)
    - 10 blind; 14 supervised

SiSEC 2016 { http://sisec17.audiolabs-erlangen.de }
Liutkus et al. (2017; LVA/ICA)

Case study: SiSEC 2016 MUS challenge

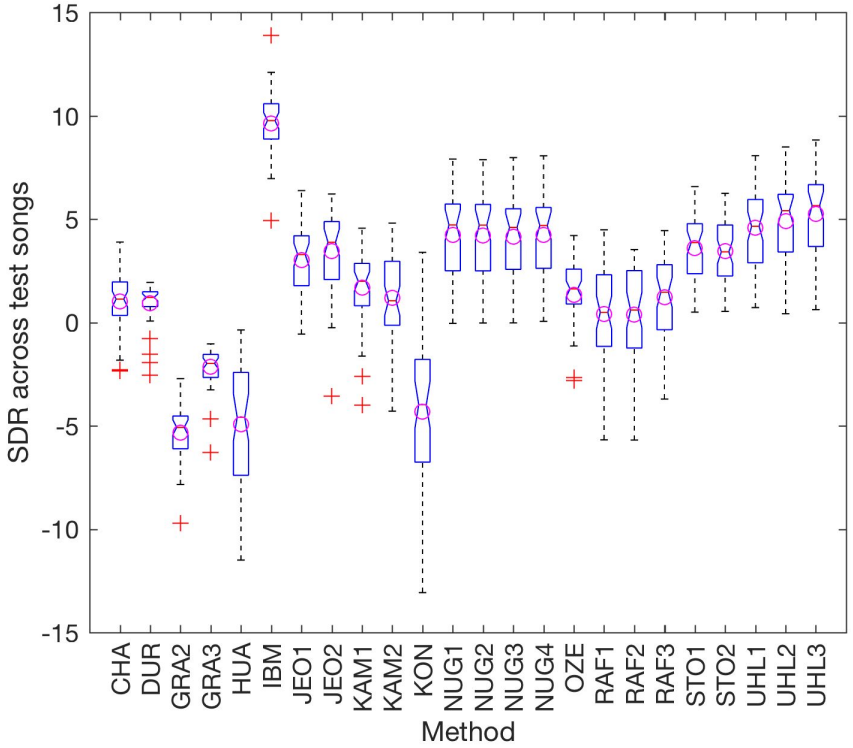
- Reporting solely mean results can be misleading
  - If the mean value is higher, is it definitely better?
  - Boxplots show a large amount of between-song variance
    - Reduced spread in error may be preferable
    - Are the observed differences in SDR systematic?



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

Case study: SiSEC 2016 MUS challenge

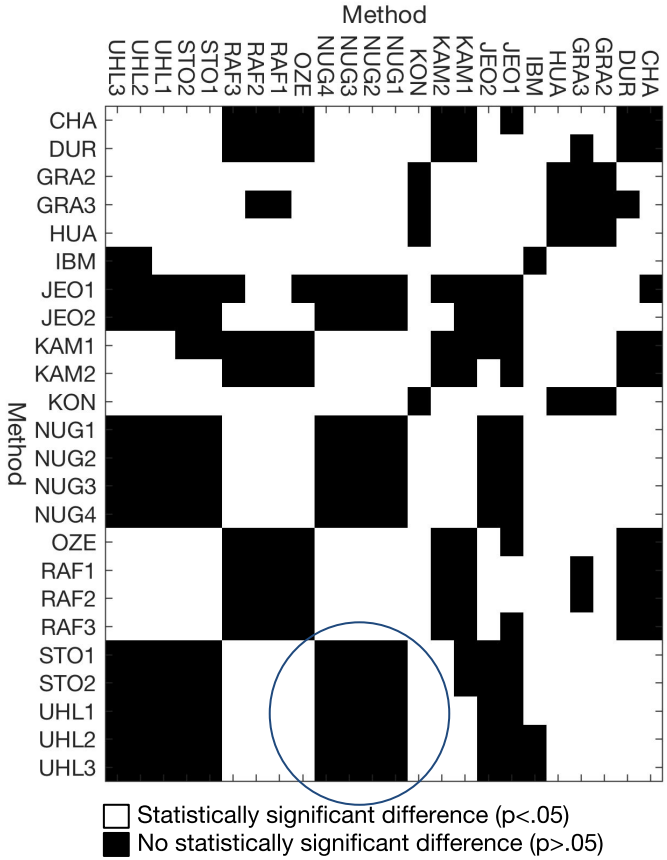
- One approach is to employ **null-hypothesis** statistical testing (NHST) in attempt to deal with influence of signal variability on our conclusions
  - Repeated-measures ANOVA:
    SDR ~ Method + (1 | Song)
  - Check model assumptions
  - Presence of an effect (p-value)
  - Carry out pairwise comparisons using an appropriate post-hoc test



Case study: SiSEC 2016 MUS challenge

- Simpson et al. (2016; EUSIPCO):
  - Model assumptions were not met, so a non-parametric Friedman test was used to test main effect of method
  - Post-hoc test carried out using Wilcoxon signed-rank test with Bonferroni correction to control for false positive ($p<0.05$)
  - Insufficient evidence against the null (no difference) between the two best performing DNN systems

    (UHL and NUG)

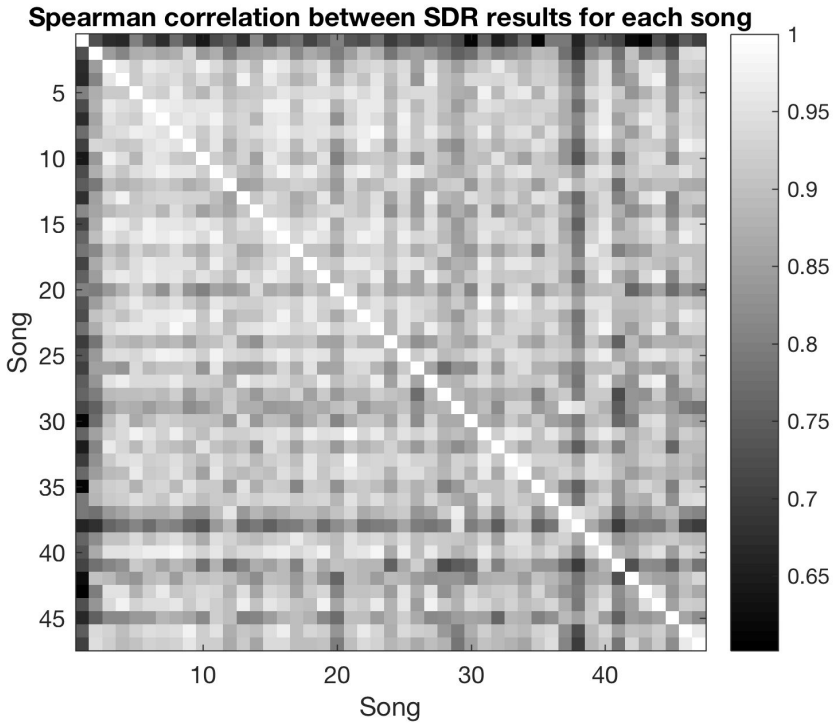


Case study: SiSEC 2016 MUS challenge

- Don't lose your head:
  - P-value = probability of observing a difference as large as the one you have, **if the null-hypothesis were true**
  - Just because $p < 0.05$, doesn't mean you can accept the alternative hypothesis (we infer that method A is 'better' than method B)
- Parameter estimation:
  - In most cases we want to know "how much", taking into account uncertainty due to sampling error
  - Consider supplementing your effect with a Bayesian credible interval to report the uncertainty of the summary statistic
  - **Jasp-stats.org:** Conduct classical analyses as well as Bayesian analyses (Wagenmakers et al. 2018; Psychon Bull Rev)

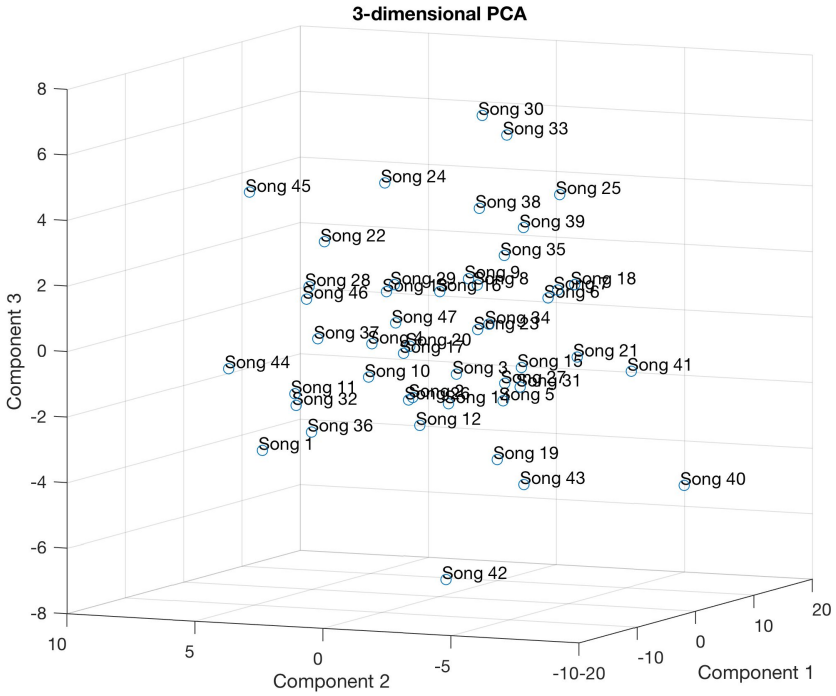
Case study: SiSEC 2016 MUS challenge

- Results highly dependent on the signals used for testing
  - Scores and rankings can change greatly by the selection of songs used
  - Can examine similarity of results or rankings using correlation analysis



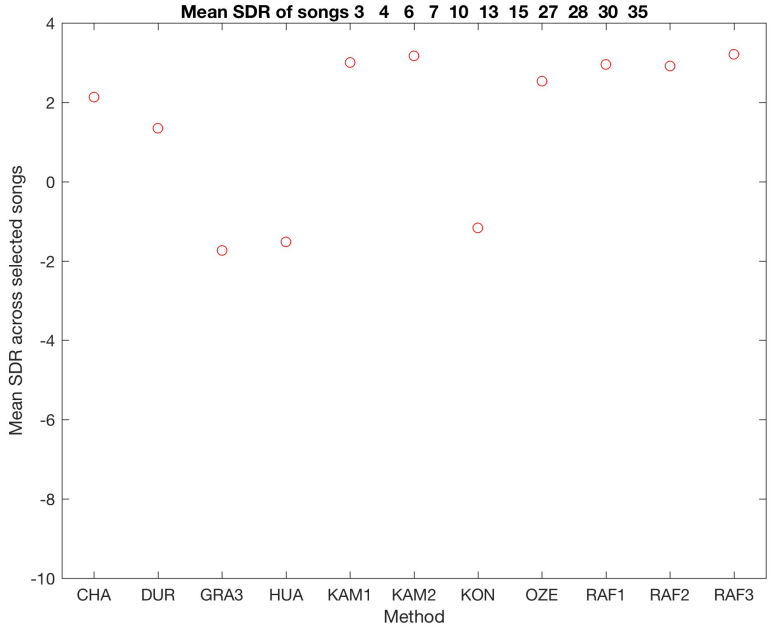
Case study: SiSEC 2016 MUS challenge

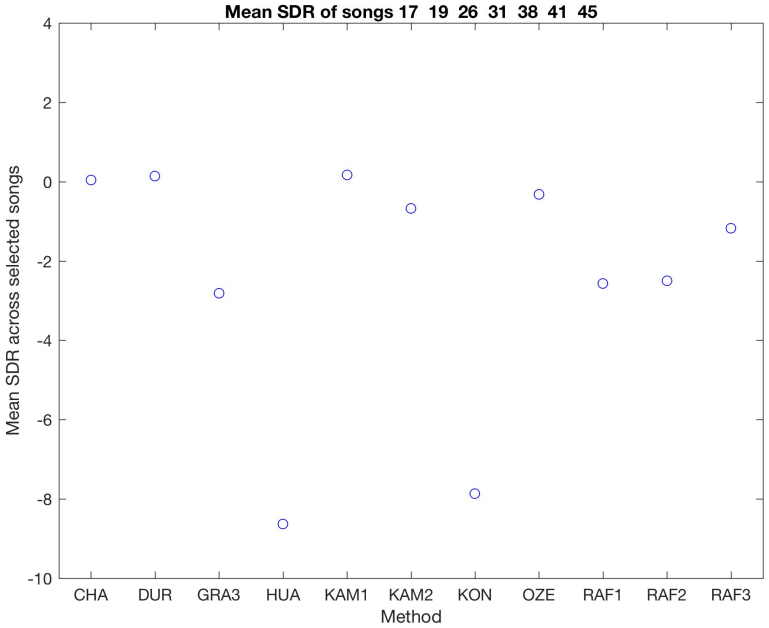
- Results highly dependent on the signals used for testing
  - Scores and rankings can change greatly by the selection of songs used
  - Can examine similarity of results or rankings using correlation analysis
  - Principal component analysis can show more information about the structure of differences between results for each song



Case study: SiSEC 2016 MUS challenge

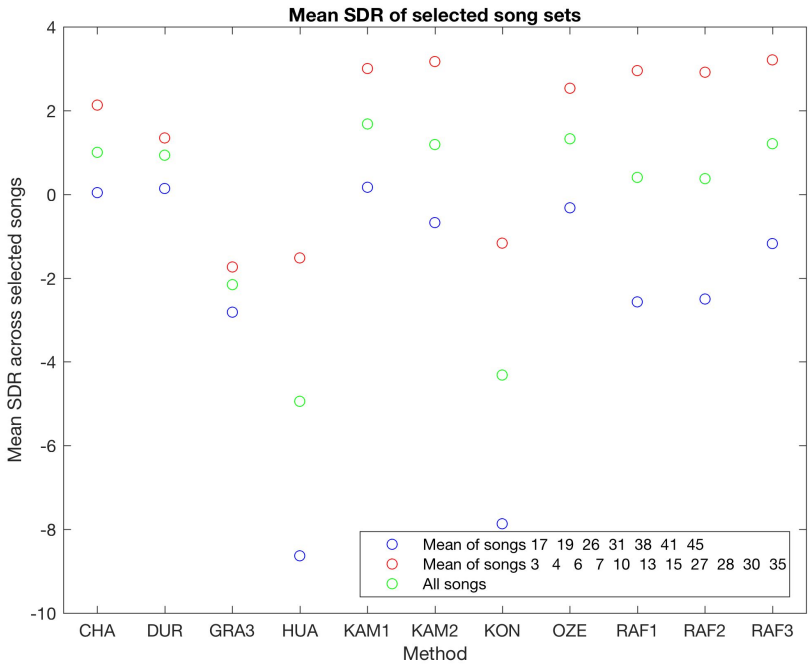
- Results highly dependent on the signals used for testing
  - Scores and rankings can change greatly by the selection of songs used
  - Example differences between song sets





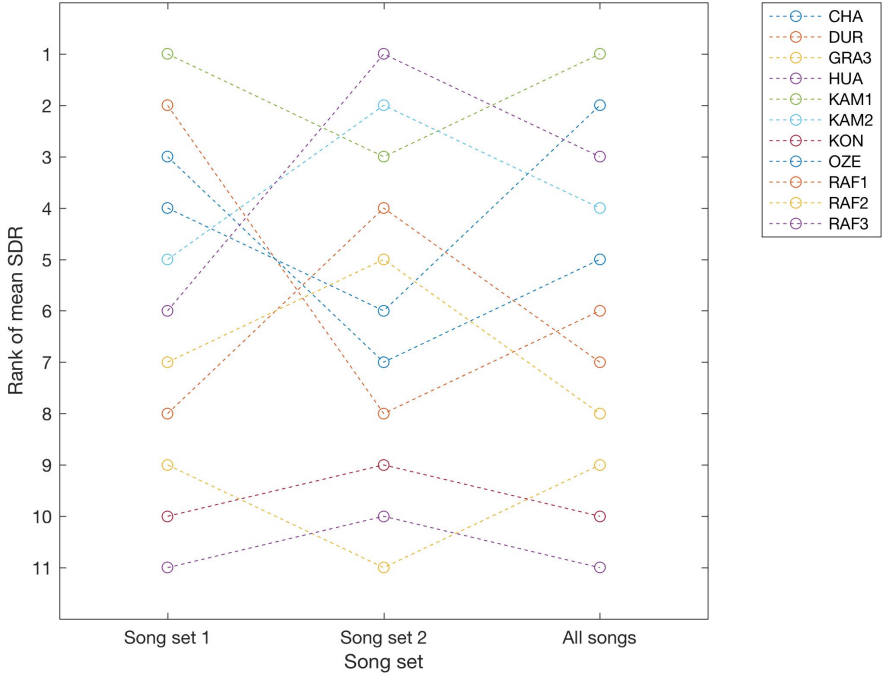
Case study: SiSEC 2016 MUS challenge

- Results highly dependent on the signals used for testing
  - Scores and rankings can change greatly by the selection of songs used
  - Example differences between song sets



Case study: SiSEC 2016 MUS challenge

- Results highly dependent on the signals used for testing
  - Scores and rankings can change greatly by the selection of songs used
  - Example differences between song sets



Objective measurements summary

- Currently two objective performance evaluation toolkits (BSS Eval, PEASS)
  - Both based on error decomposition
  - Multi-criteria: artifacts, interference, target distortion, overall distortion
  - BSS Eval based on signal to noise ratios
  - PEASS designed to improve correlation with human judgements
- Analysis
  - Look beyond differences in SDR
  - Undertake repeated-measures analysis to determine presence of effects
  - Present confidence/credible intervals to gauge effect size and relate to practical importance
  - Examine results for each song to determine genre- / source-dependent effects

Subjective perception

Importance of perception

- Physical measures don't always reflect how sound is perceived
- For example:
  - 13 dB signal to noise ratio (SNR) can result in poor audio quality
  - 13dB SNR can be imperceptible if the noise is located temporally and spectrally close to another signal
- What ultimately matters for most musical source separation applications is what a listener perceives

Original track

Track with 13dB SNR white noise

Track with 13dB SNR perceptually shaped noise

Subjective attributes of separated audio

- Common subjective attributes based on BSS Eval metrics (SDR, ISR, SIR, SAR)
  - SDR (Source-to-Distortion Ratio) → Global quality
  - ISR (Source Image-to-Spatial distortion Ratio) → Preservation/distortion of target source
  - SIR (Source-to-Interference Ratio) → Suppression of other sources (interferences)
  - SAR (Source-to-Artefact Ratio) → Presence of additional artificial noise (artefacts, musical noise)

- Recent studies show these are not perceptually independent
  - Examples:

Confusions in perception of BSS Eval-motivated attributes

- Artefacts vs target distortion? (Emiya et al. 2011; IEEE)

Original mixture

Voice with added artefacts

Voice with elements removed (distortion), no intention of additional component

  - Distortion (physical loss) confused with artefacts (additional artificial noise)

- Interference vs artefacts (artificial musical noise)? (Ward et al. 2018; ICASSP)

Original voice

Voice with interfering sources (=other tracks)

Voice with elements removed (distortion) & added artefacts, no intention of interfering sources

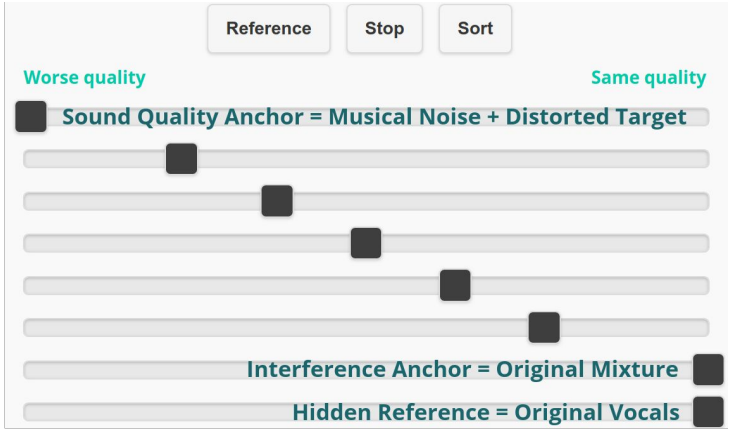
  - "They were uncomfortable assigning 'no interference' to artificial musical noise."
  - Source of "interference" – other sources? Or musical noises?

Attempts toward “the right questions”

- Introduction of alternative questions
  - Simpson et al. (2017; LVA-ICA): similarity of vocal (↔SAR) / similarity of “*loudness balance*” (↔SIR)
  - Ward et al. (2018; ICASSP): more general scale “*sound quality*” (artefacts, distortions) / interference explained as (vocals to accompaniment loudness balance)
  - Cartwright et al. (2018; ICASSP): new scale “*lack of distortions to the target source*”
  - More controlled tests (e.g., separate manipulation of “artefacts” in stimuli) could help to tackle the confusion issues
- Identifying perceptual (independent) dimensions from scratch
  - e.g., Cano et al. (2018; ICASSP)
    - Attribute elicitation → grading
    - 2 dimensions found sufficient: one correlated with “interference” and one with “artifact” / “target distortion”

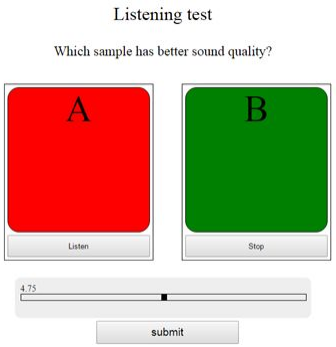
How to undertake subjective evaluations of separated audio

- Multi-stimulus evaluation
  - Ranks and scores are directly collected (saves time)
  - E.g., “Please rate the samples in terms of attribute XXX”
  - Listeners can compare multiple systems-under-test on the same stimulus
  - Often contains hidden external reference and anchors
  - Needs more concentration
  - Method used by Emiya et al. (2011; IEEE), Cano et al. (2016; EUSIPCO), Cartwright et al. (2016; ICASSP), Ward et al. (2018; ICASSP)

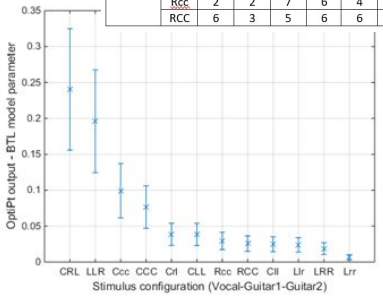


How to undertake subjective evaluations of separated audio

- Pairwise comparison
  - Intuitive in overall preference / quality investigation, including “no reference” situations
  - e.g., “Which sample do you prefer?”
  - Indirect method, so requires an estimation stage, to map preference orderings to latent scores on an interval scale
  - Time consuming - many comparisons needed
  - Method used by Cartwright et al. (2018; ICASSP)



		Number of losses											
		Ccc	CCC	Cll	CLL	Crl	CRL	Llr	LLR	Lrr	LRR	Rcc	RCC
Number of wins	Ccc	0	7	9	8	9	6	11	4	11	10	10	6
	CCC	5	0	8	7	9	4	9	3	12	8	10	9
	Cll	3	4	0	6	3	1	5	2	9	6	5	7
	CLL	4	5	6	0	5	3	6	2	11	10	6	6
	Crl	3	3	9	7	0	1	9	1	10	7	8	6
	CRL	6	8	11	9	11	0	11	10	12	11	10	12
	Llr	1	3	7	6	3	1	0	1	10	7	6	5
	LLR	8	9	10	10	11	2	11	0	12	12	11	11
	Lrr	1	0	3	1	2	0	2	0	0	2	2	6
	LRR	2	4	6	2	5	1	5	0	10	0	4	4
	Rcc	2	2	7	6	4	2	6	1	10	8	0	8
	RCC	6	3	5	6	6	0	7	1	6	8	4	0



Subjective perception summary

- Quality attributes in the context of source separation:
  - BSS eval-inspired quality descriptors initially introduced
  - Not perceptually independent, needs further investigations
- Various evaluation methods/interfaces
  - Multi-stimulus, pairwise comparison(, perceptual mapping/elicitation)
  - Choose one that suits purpose / design factors (questions to ask, # of stimuli, test duration, results of interest, etc.)

Relating objective and subjective results

Relating objective and subjective results

- For musical audio source separation applications, perceived quality is usually essential
- Subjective evaluation is time-consuming
- Objective evaluation would be quicker and more reliable, as long as it accurately matches perception

Limitations of current objective measures

- BSS eval:
  - Based on energy ratios expressed in decibels
  - Known that signal to noise ratios are not perceptually optimal
- PEASS:
  - Developed to predict aspects of perception
  - Are the attributes the most appropriate?
  - Are the measured results sufficiently accurate for a wide range of songs and separation algorithms?
- Recent studies show inconclusive performances in predicting the subjective data
  - Gupta et al. (2015; WASPAA)
  - Cano et al. (2016; EUSIPCO)
  - Cartwright et al. (2016; ICASSP)
  - Simpson et al. (2017; LVA-ICA)

Limitations of current objective measures

Are we rating the right thing?

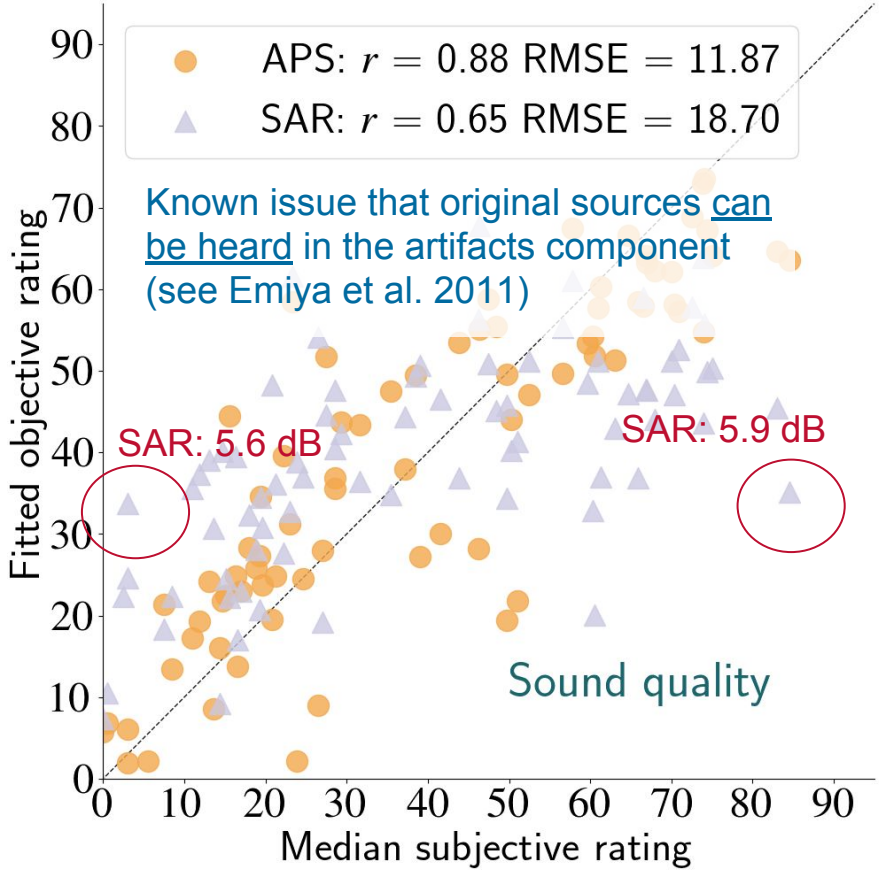
Are we measuring the right thing?

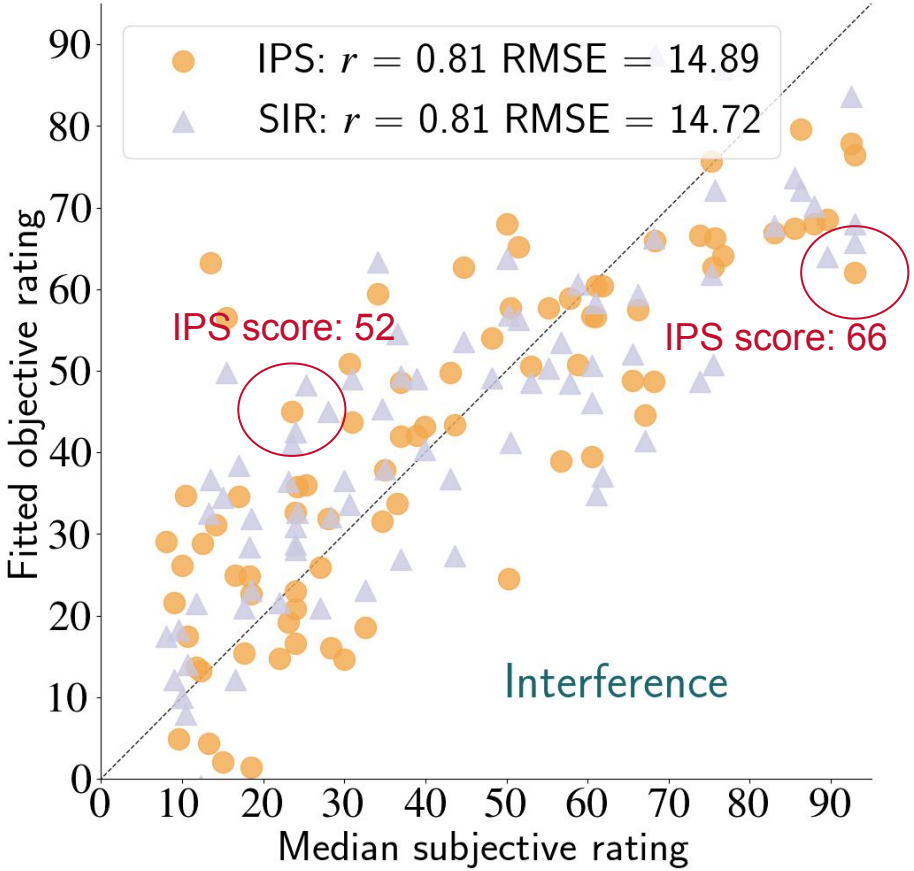
Are we correctly relating these to each other?

Examples of BSS eval and PEASS successes and failures

- Ward et al (ICASSP; 2018) performed a Multi-Stimulus Evaluation to further assess predictive capability of BSS Eval and PEASS
- 24 listeners were asked to judge Sound Quality and Interference of vocals separated from 16 pop and rock songs by a range of algorithms
- General findings:
  - Sound Quality (artifacts + subtractive distortions)
    - APS generally more consistent than SAR for both within- and across-song correlations
  - Interference (other sources)
    - SIR and IPS comparable in performance

Examples of BSS eval and PEASS successes and failures





Areas for future development

- We need:
  - to determine the most important perceptual attributes
  - to refine and develop more accurate and consistent objective metrics
  - collect more subjective data for purposes of fitting, training and validation
- Application-based objective metrics:
  - soloing – ✓
  - karaoke – ✓
  - remixing – ✗
  - spatialisation – ✗

Summary

Summary

- Evaluation is critical to the future development of source separation algorithms
- Objective evaluation algorithms need to properly reflect the most important parameters
- Subjective evaluation is slower and more time consuming but is the gold standard for some applications
- We still need to refine the attributes that participants rate
- With further development, objective evaluation may be able to accurately and consistently predict the results of subjective evaluation
- In all cases we need to use good scientific method:
  - consistent use of carefully selected signals;
  - detailed statistical analysis to tell the full story (not just means); and
  - examination of the full details of the results (to help further development).