Projets et Applications Musicales

Sound Recording and Audio Source Separation

Azal LE BAGOUSSE, Robin WENDLING, Jiale KANG

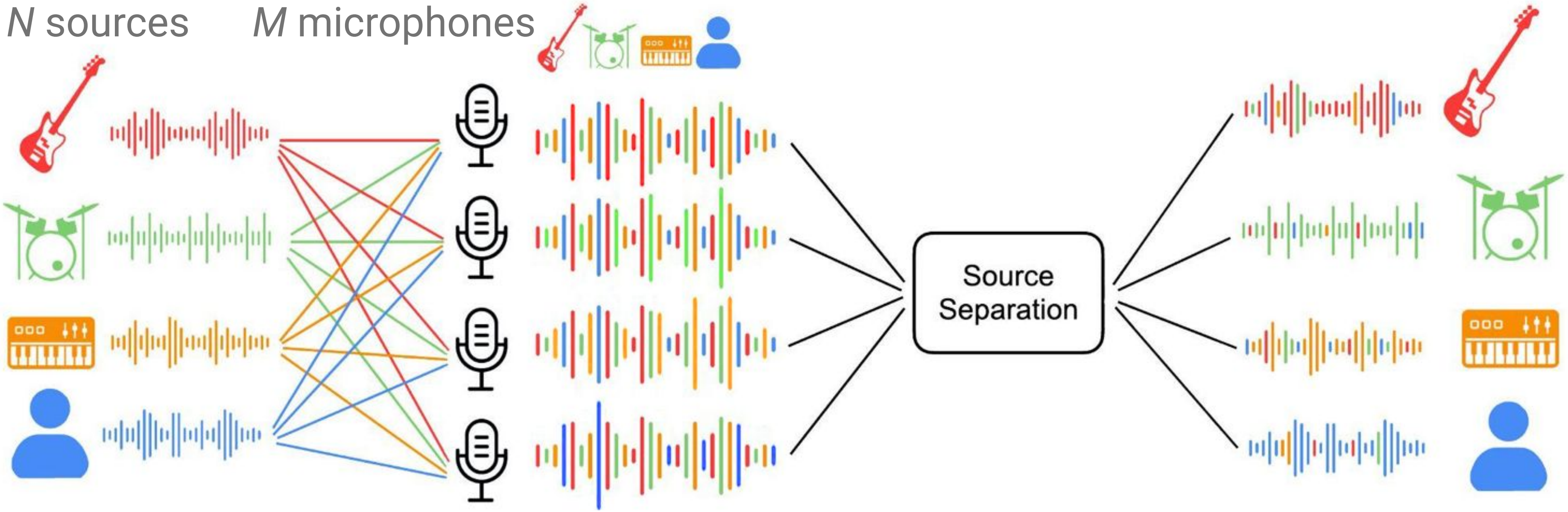
Supervisors: Benoit FABRE, Mathieu FONTAINE

M2 ATIAM 2024-2025

21, Feb, 2025

Problem statement

N sources *M* microphones



$\mathbf{X}_n = \{\mathbf{x}_{ftn}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T} \in \mathbb{C}^{F\times T\times M}$ $\{\hat{\mathbf{X}}_n\}_{n=1}^{N}$

Suppose mixture follows additive hypothesis $\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{x}_{ftn} \in \mathbb{C}^M$

State Of the Art

Sound recording

Microphone techniques :

Coincident : MS Stereo

Near-coincident : ORTF, NOS

Spaced Microphone : Spaced omnis, spaced bidirectional…

Immersive : Optimized Cardioid Triangle (OCT), Ambisonic Microphones…

ORTF technique

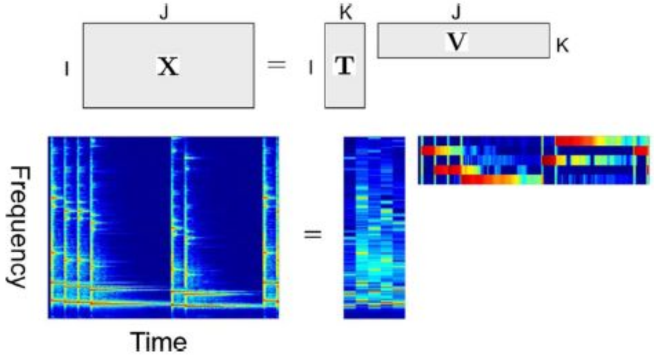
Ambisonic microphone

Techniques are selected based on what and where you want to record

Algorithms

- Independent Component Analysis (JADE/SOBI)
- Matrix factorisation methods (GaussMNMF, FastMNMF, ILRMA)
- Deep learning methods

DEMUCS Spleeter



Non-negative matrix factorisation principle
cr : Sawada 2013

See : **J.-F. Cardoso,** (1999)
See : **Hiroshi Sawada** et al, (2013)
See : **T. Sekiguchi,** (2020)
See : **Ono.** (2011)

Evaluation / Objective

Blind Source Separation Evaluation (BSS Eval), 2006

- Source-to-Distortion Ratio (SDR)
- Source-to-Interference Ratio (SIR)
- Sources-to-Artifacts Ratio (SAR)

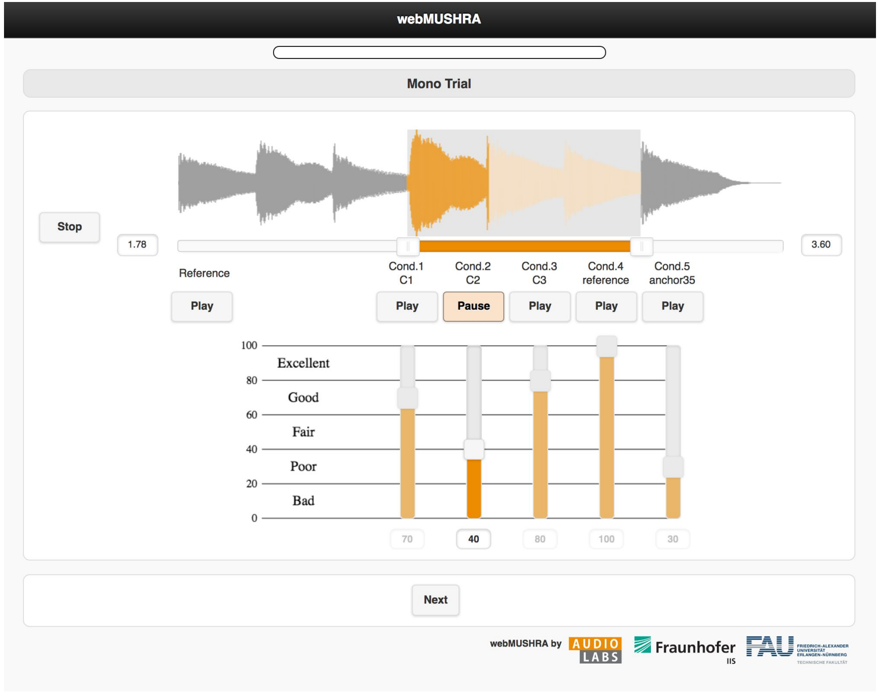
Perceptual Evaluation methods for Audio Source Separation (PEASS), 2011

- Perceptually motivated assessment
- Integrate auditory models

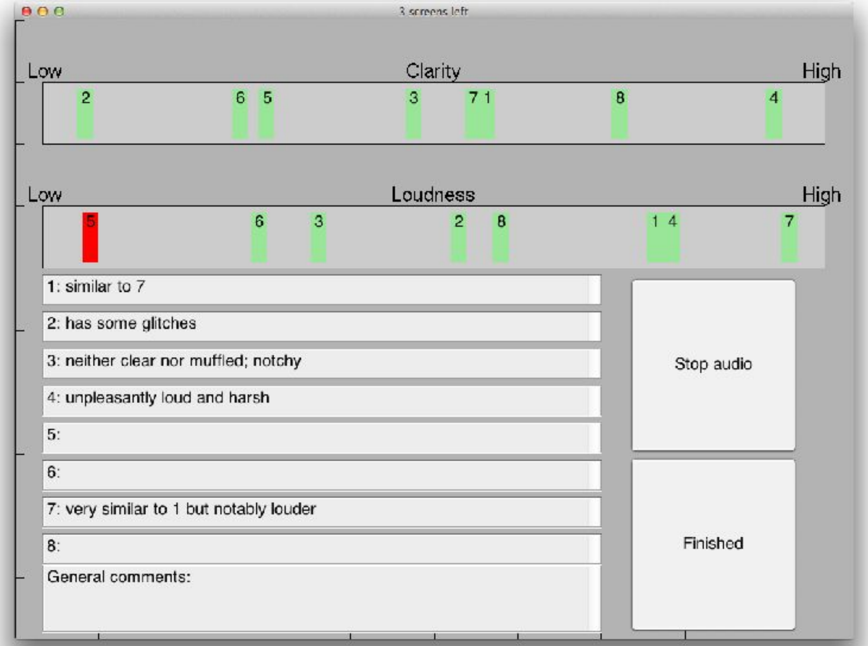
Fréchet Audio Distance (FAD), 2019

- Need large database of studio recorded audio
- A distance

Evaluation / Subjective



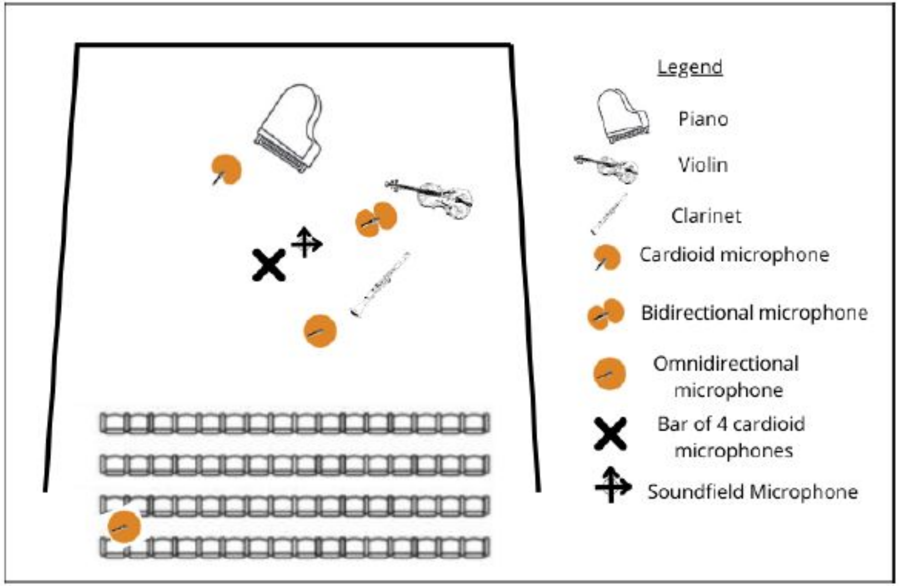
Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), 1996



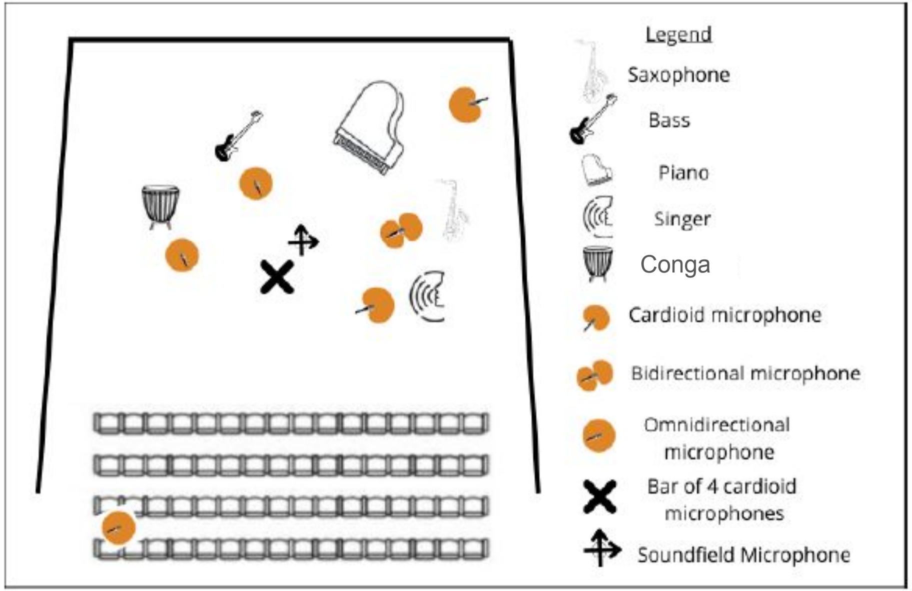
Audio Perceptual Evaluation (APE), 2014

Methods

Recording session



Configuration of the recording of Schubert trio



Configuration of the recording of the jazz quartet

Recording session

Table: Microphone details for instruments and mixture microphones

Instrument	Microphone Model	Directivity
Clarinet (Classical)	AT4050	Cardioid
Violin (Classical)	AT4050	Cardioid
Piano (Classical)	DPA 4007	Omnidirectional
Conga (Jazz)	DPA 4007	Omnidirectional
Bass (Jazz)	DPA 4007	Omnidirectional
Piano (Jazz)	AT4050	Cardioid
Alto Sax (Jazz)	AT4050	Bidirectional
Mixture Microphones		
4 microphones bar	2 Schoeps MK4 (center), 2 DPA 4011 (ext)	
Additionnal microphone	Soundfield	
Room Microphone		
Omnidirectional	DPA 4007	

Auxiliary microphone (sax)

Bar holding the 4 cardioid microphones for mixture recording

Soundfield microphone

Gaussian MNMF

- NMF extension to multichannel audio
- Local gaussian model

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\left(0,\, \lambda_{ftn}\, \mathbf{G}_{nf}\right) \qquad \lambda_{ftn} = \sum_{k=1}^{K} w_{nkf}\, h_{nkt}$$

- Power spectral density λ_{ftn}
- Spatial covariance matrix $\mathbf{G}_{nf}$
- Multichannel Wiener filter

See : **Hiroshi Sawada** et al, (2013)

FastMNMF2

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\left(0,\, \lambda_{ftn}\, \mathbf{G}_{nf}\right)$$

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf}\, h_{nkt}$$

- Joint diagonalization for covariance matrices = common diagonalizer Qf for each *freq* bin

$$\mathbf{G}_{nf} = \mathbf{Q}_f^{-1}\, \mathrm{Diag}(\tilde{\mathbf{g}}_n)\, \mathbf{Q}_f^{-H}, \quad \forall n,$$

- Frequency-invariant spatial parameters

$$\tilde{\mathbf{g}}_n = [g_{n1}, g_{n2}, \ldots, g_{nM}]$$

ILRMA

$$\mathbf{x}_{ftn} \sim \mathcal{N}_{\mathbb{C}}\left(0,\, \lambda_{ftn}\, \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}}\right),$$

$$\lambda_{ftn} = \sum_{k=1}^{K} w_{nkf}\, h_{nkt}$$

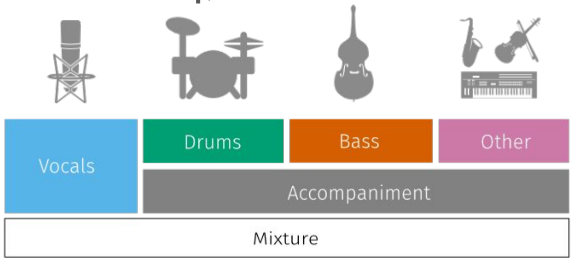
- Rank-1 spatial covariance model
- Controlled/less reverberant spaces

Deep learning models

DEMUCS

- Both time and frequency domain
- U-Net, Transformer
- 4-layer Encoder * 2 (T / F)
- 4-layer Decoder * 2 (T / F)

MUSDB18-hq, **10** hrs music tracks



Spleeter (commercial)

- U-Net
- 6-layer Encoder
- 6-layer Decoder

vocals / drums / bass / piano / others

Bean (private dataset), **79** hrs music

mainly pop and rock music

Evaluation / BSS Eval

Suppose :

acceptable deformation of the true source
interference from other undesired source
noise from source
noise from algorithm

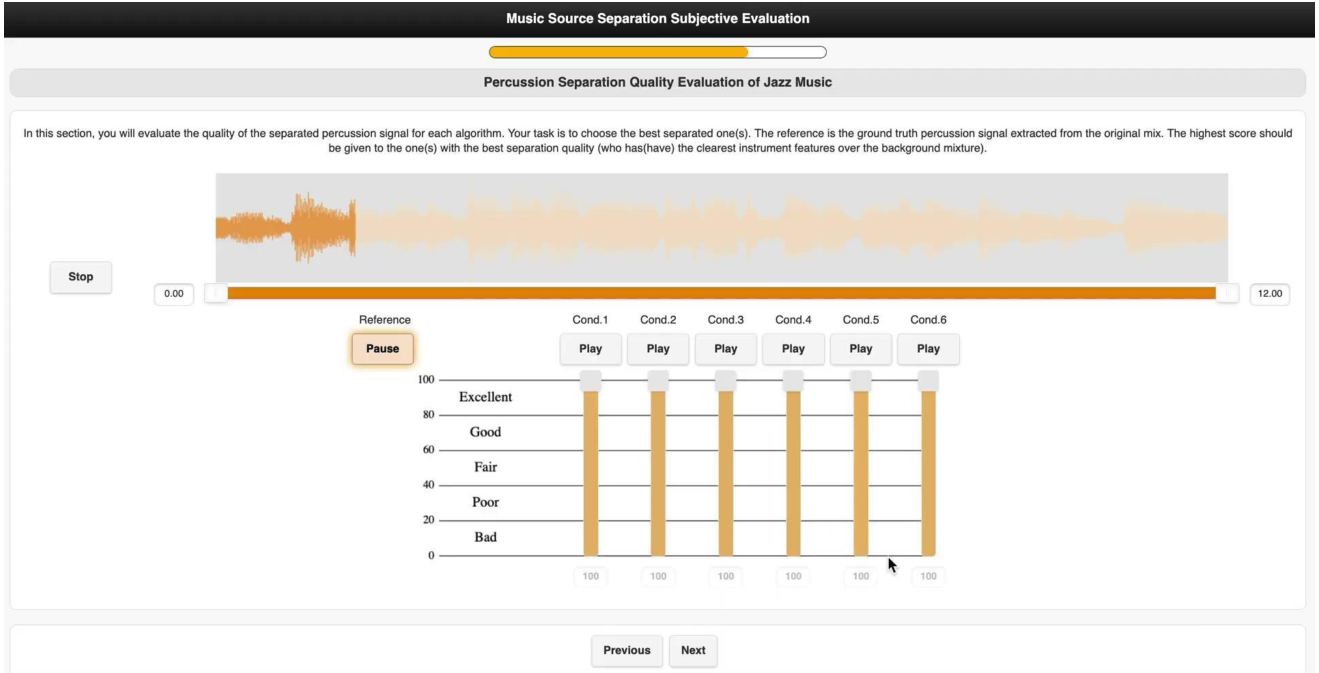
observed signal

Then :

SDR = overall quality of the separated source
SIR = suppression of undesired sources in the separated signal
SAR = amount of additional noise introduced by the separation process

Evaluation / MUSHRA

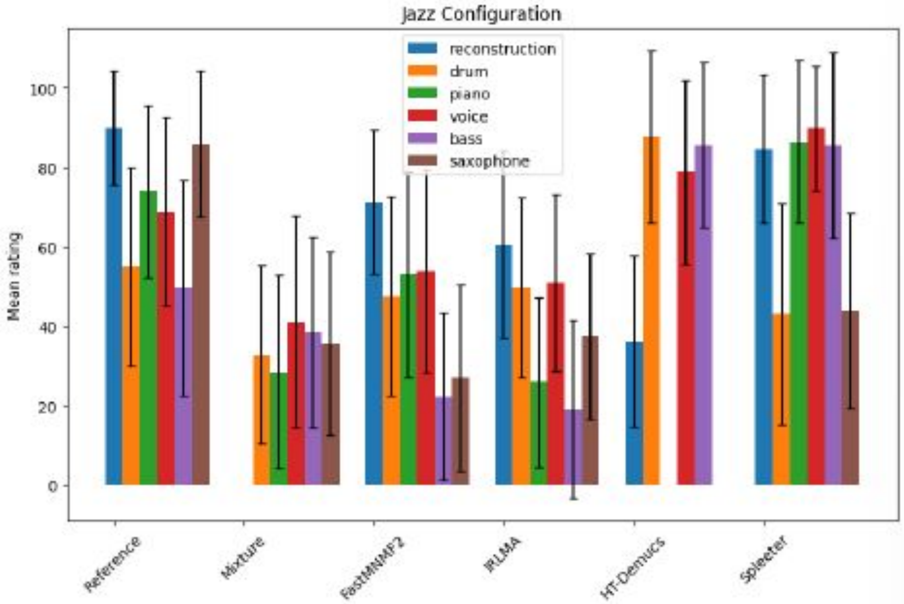
MUSHRA test example (how to use)



Results & Discussion

Comparison between algorithms

3 days / 53 participants / ⅓ Female, ⅔ Male / 19-62 yo



Statistics for Jazz Configuration in MUSHRA evaluation, normalized from 0 to 100

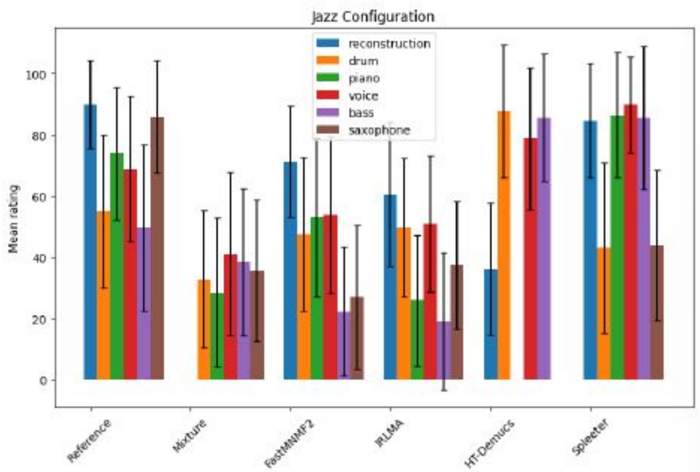
Jazz piano separation results

FMNMF2

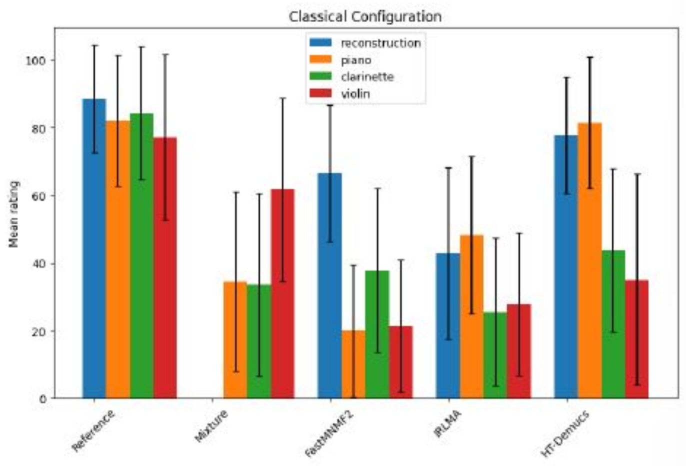
ILRMA

Spleeter

Influence of genres and input



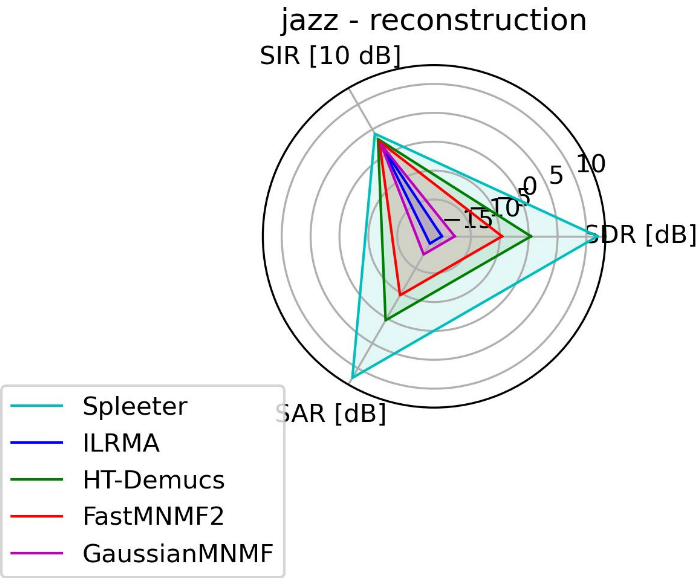
Statistics for Jazz Configuration in MUSHRA evaluation, normalized from 0 to 100



Statistics for Classical Configuration in MUSHRA evaluation, normalized from 0 to 100

- Inputs : Auxiliary Microphones / 4-cardioid-bar / Soundfield recordings
  Best results = stack of auxiliary microphones
  [MUSHRA tracks input = 4-cardioid + room microphone]

Objective vs. Subjective evaluation



Polar graph of the objective metrics

Table 3: Objective metric results for the jazz reconstruction

Algorithm/Metric	SDR (dB)	SIR (dB)	SAR (dB)
GaussMNMF	-9.6	22.1	-9.5
FastMNMF2	-4.6	26.0	-4.6
ILRMA	-15.0	16.3	-14.9
Demucs	0.4	30.8	0.4
Spleeter	11.9	41.3	11.9

See : Rumbold et al. (2024) and Emiya et al. (2011)

Conclusion

Conclusion

Key factors affecting performance:

- **Playing techniques**: avoiding homorhythmic, electronic instruments;
- **Recording conditions**: microphone placement and instruments spacing influence leakage;
- **Reverb effects**: highly reverberant environment introduce longer decay.

Optimizing strategies:

- Auxiliary and directional microphones -> pre-isolate sources;
- Dictionary: map spectrogram to basis matrix -> MNMF based algorithms;
- Training on common stem dataset -> deep learning models;

Jazz configuration at CRR 93

Thank you !

Classical configuration at CRR 93

Project website with all results:

https://kjle.github.io/PAM-Music-Source-Separation/

References

- E. Vincent, R. Gribonval, and C. Fevotte. “Performance measurement in blind audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing 14.4 (2006)*, pp. 1462–1469.
- Michael Schoeffler et al. “webMUSHRA — A Comprehensive Framework for Web-based Listening Tests”. In: *Journal of Open Research Software (Feb. 2018)*.
- Romain Hennequin et al. “Spleeter: a fast and efficient music source separation tool with pre-trained models”. In: *Journal of Open Source Software 5.50 (2020)*, p. 2154.
- A. Défossez et al. Demucs: Deep Extractor for Music Sources. 2019. arXiv: 1911 . 13254 [cs.SD].
- Kouhei Sekiguchi et al. “Fast Multichannel Nonnegative Matrix Factorization With Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020)*, pp. 2610–2625.
- Hiroshi Sawada et al. “Multichannel Extensions of Nonnegative Matrix Factorization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing 21.4 (2013)*, pp. 750–762.
- A. Ozerov and C. Févotte. “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing 18.3 (2010)*, pp. 550–563.
- George T ZANETAKIS Erika RUMBOLD and Bryan PARDO. “CORRELATIONS BETWEEN OBJECTIVE AND SUBJECTIVE EVALUATIONS OF MUSIC SOURCE SEPARATION”. In: *SMC 2024 Book (2024)*, pp. 538–545. doi: 10 . 5281 / zenodo . 13918961.
- Niklas HARLANDER Valentin EMIYA Emmanuel VINCENT and Volker HOHMANN. “Subjective and objective quality assessment of audio source separation”. In: *IEEE Transactions on Audio, Speech and Language Processing 19 (2011)*, pp. 2046–2057. doi: 10.1109/TASL.2011.2109381.

Appendices

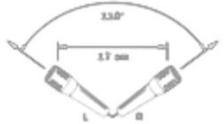
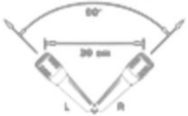
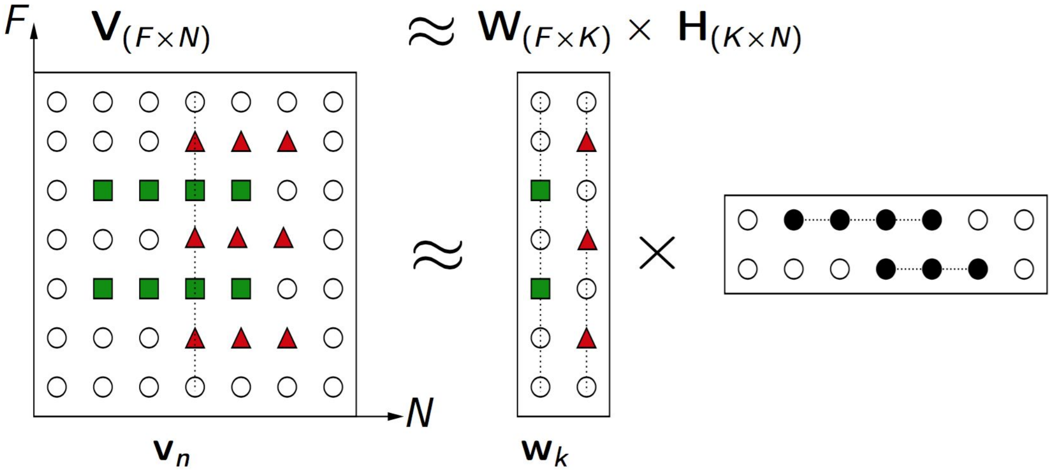
STEREO PICKUP SYSTEMS	MICROPHONE TYPES	MICROPHONE POSITIONS	
X-Y	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 135° SPACING: COINCIDENT	
ORTF (FRENCH BROADCASTING ORGANIZATION)	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 110° SPACING: NEAR-COINCIDENT (7 IN.)	
NOS (DUTCH BROADCASTING FOUNDATION)	2 - CARDIOID	AXES OF MAXIMUM RESPONSE AT 90° SPACING: NEAR-COINCIDENT (12 IN.)	
MS (MID-SIDE)	1 - CARDIOID 1 - BIDIRECTIONAL	CARDIOID FORWARD-POINTED; BIDIRECTIONAL SIDE-POINTED; SPACING: COINCIDENT	
SPACED	2 - CARDIOID OR 2 - OMNIDIRECTIONAL	ANGLE AS DESIRED SPACING: 3-10 FT.	

Figure A1 : Different kinds of microphone techniques for sound recording

Appendices

Figure A2 : NMF method



spectrogram ≅ basis × activation