Perceptual Evaluation of Source Separation: Current Issues with Listening Test Design and Repurposing

Ryan Chungeun Kim

Centre for Vision, Speech and Signal Processing (CVSSP)

University of Surrey, U. K.

Contents

- The MARuSS project at Surrey: overview
- Key issues in perceptual evaluations and listening test design
- Source separation and repurposing
- Summary

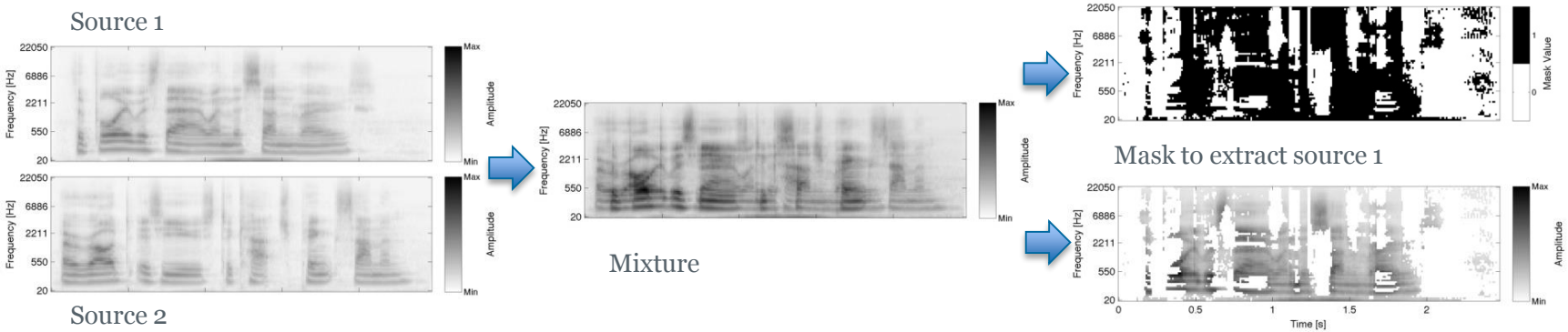
The MARuSS project: overview

<Musical Audio Repurposing using Source Separation>

- Separation of legacy-format music mix (e.g. stereo) with consideration of “repurposing”: remixing or upmixing
- Development of evaluation techniques to judge the outcome
- https://cvssp.github.io/maruss-website/

The MARuSS project: overview

<Musical Audio Repurposing using Source Separation>



(from Institute of Sound Recording (IoSR) blog, http://iosr.surrey.ac.uk/blog/2013-10-23.php)

- Emerge of machine learning techniques to estimate masks to apply
- Design and application of novel deep learning techniques / structural arrangements

The MARuSS project: overview

Works within the MARuSS project

- 2015: “Deep Karaoke”
  - Use of a DNN to estimate the binary mask towards vocal extraction
- 2016: Use of a set of DNNs
  - Different types of time-freq masks (binary, ratio, etc.) are estimated
  - Combinations are used for the final output

The MARuSS project: overview

Works within the MARuSS project

- 2017: Multi-stage separation
  - Stage 1: the separated sources are considered as mixtures for the input of the second stage
  - Stage 2: the separated sources are further enhanced (separately or jointly) to deliver the final estimates
- 2017: Use of other deep learning techniques
  - Convolutional Denoising Autoencoders

The MARuSS project: overview

Works within the MARuSS project

- 2018: More complicated combinations
  - *Multi-channel & multi-resolution* Convolutional Autoencoders, in *time domain only*


mix


extracted vocal


accompaniment

  - Multi-stage / multi-neural network combination

The MARuSS project: overview

Works within the MARuSS project

- Most recent work in this AES Convention
  - E. M. Grais and M. D. Plumbley, “Combining Fully Convolutional and Recurrent Neural Networks for Single Channel Audio Source Separation,” in Audio Engineering Society Convention 144, Milan, Italy, 2018
  - **Poster Session P19**: Audio Processing/Audio Education
  - Friday, May 25, 13:15 — 14:45

Key issues in perceptual evaluations and listening test design

Some background

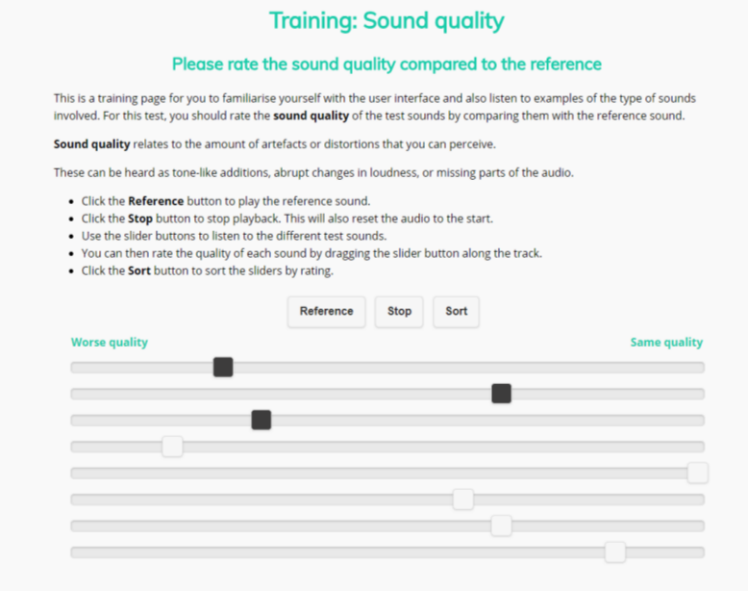
- Why go perceptual?
  - Metric needed to evaluate the source separation performance
  - Conventional measure: BSS-eval (Vincent *et al*. 2006), based on energy ratios of decomposed signals
    - e.g. Source-to-Distortion / Interference / Artifacts
  - Correlation with actual listener responses questioned

- More recent alternative
  - PEASS (Perceptual Evaluation methods for Audio Source Separation) (Emiya *et al*. 2011): application of computational auditory models
  - Correlation with listening test data still questioned (Cano *et al*. 2016)

Key issues in perceptual evaluations and listening test design

Listening test for perceptual evaluation

- Typical listening test design
  - Multi-stimulus
    (e.g. MUSHRA: Multi Stimulus with Hidden Reference and Anchors)
  - Reference: perfectly separated source = original track
  - Anchors: dependent on the quality aspects being asked



Key issues in perceptual evaluations and listening test design

Listening test for perceptual evaluation

- Quality aspects for listening test questions
  - Initiated from energy-based metrics (SDR, SIR, SAR)
  - Typical starting point:
    - Global quality
    - Preservation / distortion of target source
    - Suppression of other sources (interference)
    - Absence of additional artificial noise (artifacts)
  - Anchors are created towards lowest perceived quality
    - e.g., LPF, time-freq frame removal, adding other tracks

Key issues in perceptual evaluations and listening test design

Confusions found from anchor scores

- Physical “loss” resulting in “something new”
  - Target distortion vs artifacts? (Emiya *et al*. 2011, original PEASS work)

interference anchor
(all tracks)

artifact anchor
(target + 99% TF bins removed)

target distortion anchor
(3.5kHz LPF → 20% TF bins removed)

- Source of “interference” – other sources? Or musical noises?
  - Interference vs artifacts (artificial musical noise)? (Ward *et al*. 2018)

reference
(original vocal)

interference anchor
(all tracks)

artifacts anchor
(20% TF bins removed → 3.5kHz LPF + 99% TF bins removed → 3.5kHz LPF)

- Perceptually not independent

Key issues in perceptual evaluations and listening test design

Further steps

- Identify the relevant perceptual dimensions
    - e.g., descriptive extraction / multi-dimensional mapping (Cano *et al*. 2018 ICASSP)
- Find the right descriptors
    - Attempts to use alternative questions (e.g., Simpson *et al*. 2017, Ward *et al*. 2018)
    - Evaluation of SiSEC 2018 dataset under way (Check LVA-ICA 2018 at University of Surrey)

Source separation and repurposing

Can we use these separation techniques at all?

- Suppression of unwanted sources
    - James Clarke, 2017 AES Berlin Convention, Beatles at the Hollywood Bowl

(Image from “The Beatles At The Hollywood Bowl, The Lost Live Album”, Feature Story, onabbeyroad.com)

(Image from http://www.meetthebeatlesforreal.com/2014/08/the-hollywood-bowl-in-1964.html)

Source separation and repurposing

Can we use these separation techniques at all?

- Repurposing can make the individual source degradations unnoticeable
  - Level remixing: Wierstorf *et al*. 2017
    - Vocal level after separation could be increased by up to +6dB


separated


Remixed, 0dB


Remixed, +6dB


Remixed, +12dB

  - Spatial remixing (upmixing): some previous studies

Cobos et al. 2008
Barry and Kearney 2009
FitzGerald 2011

    - Scene width can be manipulated
    - Artifacts / interferences → spatial fluctuation or localization ambiguity (demo)

Source separation and repurposing

Spatial remixing - demo

(For downloadable tool check
3D Tune-In EU project: http://3d-tune-in.eu/)

Summary

- Source separation research
  - Deep learning has provided promising results
  - Performance enhancement through novel techniques
- Evaluation of source separation
  - Relevance of BSS-eval / PEASS under questions
  - Need for more representative perceptual metrics
  - Confusions in the quality attributes require further investigations
- Source separation towards repurposing
  - Non-perfect separation is still acceptable
  - Excessive interferences/artifacts now lead to spatial degradation

Thank you!

- This project is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).

http://cvssp.org/events/lva-ica-2018/

References

- Vincent, E., Gribonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech and Language Processing, 14(4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005

- Emiya, V., Vincent, E., Harlander, N., & Hohmann, V. (2011). Subjective and Objective Quality Assessment of Audio Source Separation. Audio, Speech, and Language Processing, IEEE Transactions on, 19(7), 2046–2057. https://doi.org/10.1109/TASL.2011.2109381

- Cano, E., Fitzgerald, D., & Brandenburg, K. (2016). Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. European Signal Processing Conference, 2016–November(1), 1758–1762. https://doi.org/10.1109/EUSIPCO.2016.7760550

- Ward, D., Wierstorf, H., Mason, R. D., Grais, E. M., & Plumbley, M. D. (2018). BSS eval or peass? Predicting the perception of singing-voice separation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Alberta, Canada.

- Cano, E., Libbetrau, J., FitzGerald, D., & Brandenburg, K. (2018). The Dimensions of Perceptual Quality of Sound Source Separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, Alberta, Canada.

- A. J. R. Simpson, G. Roma, E. M. Grais, R. D. Mason, C. Hummersone, and M. D. Plumbley, “Psychophysical Evaluation of Audio Source Separation Methods,” in *Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings*, P. Tichavský, M. Babaie-Zadeh, O. J. J. Michel, and N. Thirion-Moreau, Eds. Cham: Springer International Publishing, 2017, pp. 211–221.

- Wierstorf, H., Ward, D., Mason, R., Grais, E. M., Hummersone, C., & Plumbley, M. D. (2017). Perceptual Evaluation of Source Separation for Remixing Music. In Audio Engineering Society Convention 143. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=19277

- Cobos, M., Lopez, J. J., Gonzalez, A., & Escolano, J. (2008). Stereo to Wave-Sield Synthesis Music Up-mixing: An Objective and Subjective Evaluation. *2008 3rd International Symposium on Communications, Control, and Signal Processing, ISCCSP 2008*, (March), 1279–1284. https://doi.org/10.1109/ISCCSP.2008.4537423

- Barry, D., & Kearney, G. (2009). Localization Quality Assessment in Source Separation-Based Upmixing Algorithms. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=15187

- FitzGerald, D. (2011). Upmixing from mono - A source separation approach. In *2011 17th International Conference on Digital Signal Processing (DSP)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICDSP.2011.6004991