# Semi-supervised adversarial audio source separation applied to singing voice extraction[1]

Conference submission under review

Daniel Stoller, Sebastian Ewert, Simon Dixon

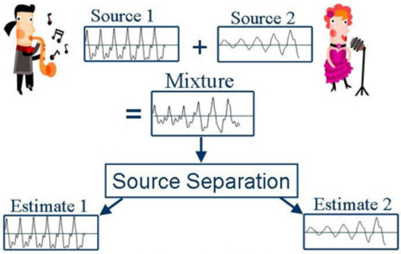
Centre for Digital Music
Queen Mary University London

SIGMA, 20.12.2017

[1] https://arxiv.org/abs/1711.00048

# Music source separation

- Task of MSS: Recover instrument sources from mixtures
- Applications:
  - Karaoke and instrumental versions
  - Remixing
  - Further analysis of sources: Preprocessing for
    - Singer identification
    - Transcription

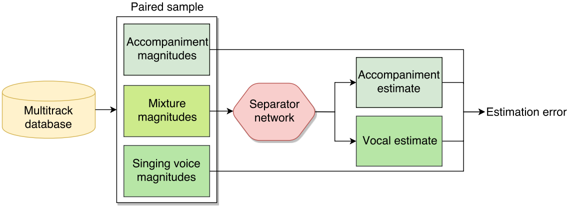


# Generative and discriminative approaches

- Generative approach [2, 6]
  - Model joint $p(s_1, \ldots, s_K, m)$ with source prior $p_\theta(s_1, \ldots, s_K)$ and likelihood $p(m|s_1, \ldots, s_K)$
  - Then given a mixture, infer likely sources (posterior inference)
  - \- Inference slow, models constrained for tractable inference
  - \+ Integration of prior knowledge
- Discriminative approach
  - Train $f_\phi(m)$ to estimate sources directly
  - Supervised training by ERM: $\arg\min_\phi \mathrm{E}_{(m_i, \mathbf{s}_i) \sim p_{\text{data}}}[l(f_\phi(m), \mathbf{s}_i)]$
  - \+ Simple, fast inference
  - \- Unclear how to define $l$

# Current state of the art

- Discriminative approaches [7, 5]
- Training on multitrack datasets
- Use neural network for $f_\phi$
- Use MSE as loss $l$
- Estimation in spectral magnitude domain



# Available data

- Multitracks:
  - DSD100 [4]
  - MedleyDB [1]
  - CCMixter (Vocals only) [2]
  - iKala (Vocals only) [3]
- Solo instrument recordings:
  - Bass: IDMT bass notes [3]
  - Drums: ENST-Drums [4]
  - Vocals: DAMP (30,000 songs) [5]
  - And many more
- Mixtures: Practically infinite

---

[2] https://members.loria.fr/ALiutkus/kam/

[3] https://://www.idmt.fraunhofer.de/en/business_units/m2d/smt/bass.html

[4] https://perso.telecom-paristech.fr/grichard/ENST-drums/

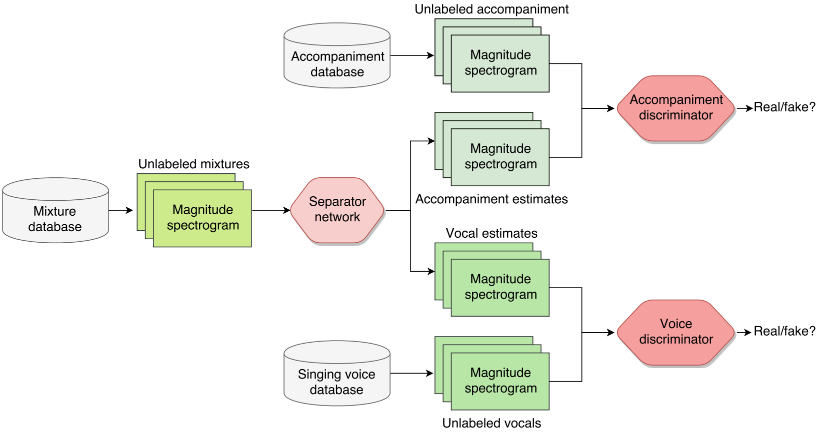
[5] https://ccrma.stanford.edu/damp/

# Discussion of state of the art

\+ Stable, reasonable complexity and results
- Overfitting since multitrack data is quite limited
- Cannot make use of solo source recordings and mixtures
- Loss function
- **Goal**: Learn from all data, combining discriminative and generative strengths

# Outline

# Intuition


Unlabeled accompaniment
Accompaniment database
Magnitude spectrogram
Accompaniment discriminator
Real/fake?
Magnitude spectrogram
Accompaniment estimates
Unlabeled mixtures
Mixture database
Magnitude spectrogram
Separator network
Vocal estimates
Magnitude spectrogram
Voice discriminator
Real/fake?
Singing voice database
Magnitude spectrogram
Unlabeled vocals

# Derivation of unsupervised loss

- Optimal separator $q_\phi(\mathbf{s}|m) = \delta(f_\phi(m) - \mathbf{s})$ would estimate real posterior perfectly: $q_\phi(\mathbf{s}|m) = p(\mathbf{s}|m)$
- Thus marginal separator output ${}^{\text{out}}q_\phi(\mathbf{s}) = E_{m \sim p_\text{m}}\ q_\phi(\mathbf{s}|m)$ is equal to true source marginal $p_\text{s}(\mathbf{s}) = E_{m \sim p_\text{m}}\ p(\mathbf{s}|m)$
- With source marginals ${}^{\text{out}}q_\phi^k(s^k) = \int_{\{s^1,\ldots,s^K\} \setminus \{s^k\}}\ {}^{\text{out}}q_\phi(\mathbf{s})$: ${}^{\text{out}}q_\phi^k \stackrel{!}{=} p_\text{s}^k,\ \forall\ k = 1,\ldots,K$
- Necessary condition for optimal separator
- Loss: Minimise divergence between source outputs: $L_\text{u} = \sum_{k=1}^{K} D[{}^{\text{out}}q_\phi^k || p_\text{s}^k]$

# Overall approach

- Supervised loss: MSE:
  $L_{\mathrm{s}} = \frac{1}{M}\sum_{i=1}^{M} ||f_\phi(m_i) - \mathbf{s}_i||_2$
- Unsupervised loss:
  $L_{\mathrm{u}} = \sum_{k=1}^{K} D[^{\mathrm{out}}q_\phi^k || p_{\mathrm{s}}^k]$
- Additive loss $L_{\mathrm{add}}$: MSE between sum of sources from $f_\phi(m)$ and input $m$
- Total loss:
  $L = L_{\mathrm{s}} + \alpha L_{\mathrm{u}} + \beta L_{\mathrm{add}}$

# Outline

# Divergence minimization with GANs

- Generative adversarial nets: Powerful unsupervised method
- Discriminator estimates divergence D between generator and real distribution
- Generator minimises divergence D

$\Rightarrow$ We use one discriminator per source to estimate the Wasserstein distance $W[^{\text{out}} q_{\phi}^{k} || p_{s}^{k}]$

# Experimental setup

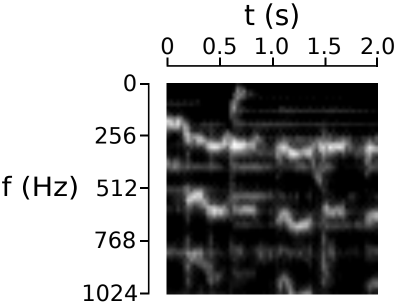
- DSD100 as training, validation and test set
- MedleyDB, iKala, CCMixter as unlabelled, validation and test set
- Avoids dataset bias
- Train supervised and semi-supervised model with early stopping
- U-Net as separator, DCGAN as discriminator
- With and without accompaniment discriminator

# Results
## Performance

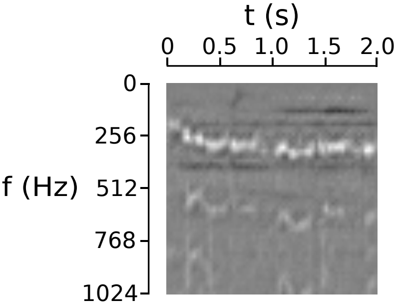
| | Test set | | | DSD100 | | | MedleyDB | | | CCMixter | | | iKala | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Baseline | V | VA | Baseline | V | VA | Baseline | V | VA | Baseline | V | VA | Baseline | V | VA |
| SDR Inst. | 8.09 | **8.89** | 8.55 | **11.11** | 10.75 | 10.76 | 9.40 | 9.60 | **9.65** | 10.65 | **11.09** | 10.89 | 6.34 | **7.71** | 7.13 |
| SDR V. | 6.80 | 7.28 | **7.47** | **3.74** | 3.17 | 3.54 | 2.48 | 2.43 | **3.00** | 3.25 | 3.52 | **3.70** | 9.50 | 10.47 | **10.52** |
| SIR Inst. | 12.03 | 12.58 | **12.67** | **14.46** | 13.56 | 13.86 | 12.18 | 12.07 | **12.74** | 15.99 | 15.49 | **16.08** | 10.42 | **11.79** | 11.57 |
| SIR V. | 13.72 | 14.00 | **14.45** | **10.03** | 9.92 | 10.49 | 9.40 | 9.21 | **9.48** | 8.39 | 8.94 | **9.35** | 16.98 | 17.44 | **17.90** |
| SAR Inst. | 11.27 | **12.05** | 11.40 | 14.20 | **14.60** | 14.10 | 13.94 | **14.23** | 13.45 | 12.84 | **13.69** | 13.24 | 9.43 | **10.42** | 9.70 |
| SAR V. | 8.54 | 9.00 | **9.04** | **5.50** | 4.84 | 5.12 | 4.71 | 4.69 | 5.20 | **6.43** | 6.17 | 6.17 | 10.81 | **11.83** | 11.73 |

Figure: Mean test set performance comparison on the test set and subsets using the supervised baseline, a vocal discriminator (V) and both vocal and accompaniment discriminators (VA)

# Results
## Qualitative




(a) Separator estimate $x$




(b) $\nabla_x D(x)$

Figure: (a) A separator voice estimate $x$. (b) Gradients of the voice discriminator output with respect to the input $x$.

# Summary

- Current SotA methods only use multi-track data
- Our approach also uses solo source recordings for improved source prior
- Combines discriminative and generative approach/loss
- Performance improvement in singing voice separation experiment

# Future work

- More realistic dataset setup
- Multi-instrument separation
- Better discriminator architecture

R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello.
MedleyDB: A multitrack dataset for annotation-intensive MIR research.
In *in Proc. the 15th International Society for Music Information Retrieval Conference (ISMIR*, 2014.

A. T. Cemgil, C. Févotte, and S. J. Godsill.
Variational and stochastic inference for bayesian source separation.
*Digital Signal Processing*, 17(5):891–913, 2007.

T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang.
Vocal activity informed singing voice separation with the ikala dataset.
In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 718–722. IEEE, 2015.

A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave.
The 2016 signal separation evaluation campaign.
In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 323–332, 2017.

A. A. Nugraha, A. Liutkus, and E. Vincent.
*Multichannel audio source separation with deep neural networks*.
PhD thesis, Inria, 2015.

A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval.
Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs.
*IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, 2007.

S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji.
Improving music source separation based on deep neural networks through data augmentation and network blending.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265, March 2017.