PREDICTING ALGORITHM EFFICACY FOR ADAPTIVE MULTI-CUE SOURCE SEPARATION

Ethan Manilow, Prem Seetharaman*, Fatemeh Pishdadian*, Bryan Pardo*†

Northwestern University
Electrical Engineering and Computer Science
Evanston, IL, USA
{ethanmanilow, prem, fpishdadian}@u.northwestern.edu, pardo@northwestern.edu

ABSTRACT

Audio source separation is the process of decomposing a signal containing sounds from multiple sources into a set of signals, each from a single source. Source separation algorithms typically leverage assumptions about correlations between audio signal characteristics ("cues") and the audio sources or mixing parameters, and exploit these to do separation. We train a neural network to predict quality of source separation, as measured by Signal to Distortion Ratio, or SDR. We do this for three source separation algorithms, each leveraging a different cue - repetition, spatialization, and harmonicity/pitch proximity. Our model estimates separation quality using only the original audio mixture and separated source output by an algorithm. These estimates are reliable enough to be used to guide switching between algorithms as cues vary. Our approach for separation quality prediction can be generalized to arbitrary source separation algorithms.

Index Terms— spatial, repetition, melody, foreground, background, singing voice separation, prediction

1. INTRODUCTION

Audio source separation is the act of isolating sound sources in an audio scene. Examples include isolating the bass line in a musical mixture, isolating a single voice in a crowd of talkers, and extracting the lead vocal melody from a song.

The problem of extracting a singing voice from a musical mixture of sounds is well studied and many source separation algorithms have been applied to this task. These algorithms depend on a variety of cues. For instance, REPET [1] and REPET-SIM [2] assume that the mixture is made up of a repeating background source (often musical accompaniment) and a non-repeating foreground source (often the singing voice). DUET [3] and PROJET [4] assume each source comes from a unique position in space. Additionally, source separation of the singing voice can be done by leveraging harmonicity and pitch proximity to extract the main melody of a song [5] [6] [7].

Knowing how and when to use any of these algorithms requires an understanding of the mechanisms an algorithm uses and the assumptions it makes about the input mixture. This knowledge requirement is a block on widespread adoption of source separation technology by non-experts. A system that could make recommendations about what auditory situations are best handled by a particular algorithm would be of much use to spreading the adoption of audio source separation technologies.

Additionally, the vast majority of source separation approaches leverage only a *single* cue to perform vocal extraction from a musical mixture. Such an algorithm is only as robust as its cue. If the cue is absent, the algorithm produces poor results. However, humans rely on many cues to parse complex auditory scenes, such as repetition [8], spatial orientation [9], common fate, and spectro-temporal similarity [10]. Humans also adaptively switch between cues [11], depending on the success or failure of that cue at that point in time. A system able to replicate this behavior will be able to capitalize on the strengths of an ensemble of algorithms.

There has been some work in ensemble approaches to source separation. McVicar et al. [12] use a conditional random field to predict a binary mask from a given spectrogram, leveraging a 10-dimensional feature vector built using the output of other source separation approaches such as RPCA [13], HPSS [14], Gabor filtering, REPET [1], and a deep learning source separation approach [15]. Dreidger et al. [16] cascaded multiple audio decomposition techniques to separate out vocals from a mixture.

Our aim is not only to perform source separation using an ensemble technique but also to predict the performance of a source separation algorithm. While prediction of source separation performance has not been extensively studied, there has been significant work in predicting errors in automatic speech recognition (ASR) [17] [18] [19]. The cues used in these works depend on particularities of speech and are not generalizable to source separation of arbitrary sources.

*contributed equally

†This work was supported by NSF Grant 1420971.

The Source to Distortion Ratio (SDR) [20] is a widely used estimate of audio source separation quality that requires access to the ground truth source to rate the quality of a separated source. In this work, we create a system that accepts an audio mixture and a separated source as input and outputs an estimate of the SDR, without need for the ground truth. We then show how to use this system to select the best system from an ensemble of source separation algorithms, each dependent on a different cue.

2. PROPOSED METHOD

Our basic approach is to train a deep neural network to learn the association between signal to distortion ratio (SDR) and the combination of audio mixture and separated source that would produce the SDR. Once done, the network can be used as an estimator of SDR that does not require the ground-truth source. This allows the construction of an ensemble source separator that combines results from separation algorithms that depend on completely different cues or assumptions about the mixture.

2.1. The Separation Algorithms

We selected the following separation algorithms because they each rely on a very distinct auditory cue to perform separation and therefore, their separation performance should be uncorrelated, making them a good ensemble for a selection algorithm.

PROJET [4] is a spatial source separation approach and relies on differences in direction of arrival between sources to perform source separation. REPET-SIM [21] relies on repetition in the auditory scene to do source separation. It identifies and separates out repetitive sources from the mixture by using the similarity matrix of the mixture spectrogram. Non-repeating sources are left over. MELODIA [6] uses harmonic salience and frequency proximity to extract pitch contours from the auditory scene. In each frame of the spectrogram, the frequencies with the highest energy that correspond to an overtone series (harmonic peaks) are selected. Then consecutive harmonic peaks are streamed into candidate pitch contours using frequency and time proximity. Then, the predominant pitch contours are selected from the candidate pitch contours. From the predominant pitch contour, we derive a time-frequency mask that is applied to the audio.

2.2. Dataset Creation

The goal of our work is to build a model that will let us select which separation algorithm will be most successful for a particular situation. Therefore, to test and train the model, we need a dataset where the results for each approach in our ensemble varies from good to poor. Also this variation in performance should not be correlated between source separation approaches.

We started with DSD100, a data set of music audio created for the evaluation of source separation algorithms [22]. The DSD100 data set consists of 100 multitrack sessions. Each session is split into four sources: vocals, drums, bass, and other. The "vocals" source consists of any vocal material (lead vocals as well as backing vocals) in the mixture. In total, DSD100 amounts to nearly 7 hours of music.

We extracted clips where the vocals are prominent. For an individual multitrack session we selected the 30-second portion of the mixture with the maximum vocal energy content, as measured using RMS amplitude. We selected that same 30-second portion of each unmixed stem/source (vocals, drums, bass, other). We then created mixtures from these 30-second stems using a set of mixing parameters that let us independently vary the salience of the cue on which each of the separation algorithms relies, and thus varying the level of success of each algorithm.

PROJET [4] is a spatial approach and relies on differences in angle between sources to perform source separation. When the vocals are obscured spatially by a background stem, PROJET cannot separate them. We placed each background stem (drums, bass, other) at a unique angle in the mixture. We then varied the angle between the vocal stem and the nearest background stem to enhance or suppress the effectiveness of this cue.

REPET-SIM [21] relies on repetition in the auditory scene to perform source separation. It identifies and separates out repetitive sources from the mixture. Non-repeating sources are left over. In this work, we assumed the vocals to be the non-repeating source and the sum of background stems to be the repeating source. To vary the effectiveness of the repetition cue we synchronously pitch shifted both the background stems and the vocals stem over the course of the 30-second mixture. The mixture glides upwards in key as it continues. Each stem starts at 6 half steps below the original key of the source and ends at 6 half steps above. The pitch shift is done via a phase vocoder. Spectrogram frames that are in different keys will not have energy in the same frequency bins, resulting in small self-similarity values, and hence degrading the performance of REPET-SIM.

We also distorted the background sources by making them clip. The distortion of the background sources reduces self-similarity, harming the efficacy of repetition as a cue to separate the auditory scene. Distortion is common in many forms of music. Clipping may happen intentionally or accidentally in musical performances. Clipping was applied by multiplying the time series audio by a gain of 100 and setting everything beyond -1 and 1 to 1.

The MELODIA-based approach assumes the singing voice corresponds to the predominant melody in the mixture.

Layer Name	Input	Conv1	MaxPool1	Conv2	MaxPool2	Conv3	MaxPool3	Dense1	Dense2	Dense3	Dense4	Output
# of Units/Filters	2x8000	32	-	64	-	128	-	256	128	64	32	1
Output shape	(2, 8000, 1)	(2, 7937, 32)	(2, 992, 32)	(2, 961, 64)	(2, 120, 64)	(2, 53, 128)	(2, 6, 128)	(256)	(128)	(64)	(32)	(1)
Filter Size/Stride	-	(1, 64), (1, 1)	(1, 8), (1, 8)	(1, 32), (1, 1)	(1, 8), (1, 8)	(1, 16), (1, 2)	(1, 8), (1, 8)	-	-	-	-	-
Activation function	-	ReLU	-	ReLU	-	ReLU	-	tanh	tanh	tanh	tanh	linear
Notes	2x 1 sec. audio	L2 Reg=0.001	-	L2 Reg=0.001	-	-	-	Dropout, drop=80%	Dropout, drop=50%	-	-	SDR

Table 1: Network architecture. The input to the network is 1 second of PCM audio of the mixture and the associated 1 second of PCM audio from the separated source. The output is an estimate of the SDR. The earlier layers are convolutional feature maps. These features are fed to a series of fully connected layers that predict the source to distortion ratio for the separated source.

Single Algorithm	SDR (dB)	Ensemble	SDR (dB)
PROJET	-1.9 ± 8.7	Random	1.1 ± 6.9
MELODIA	1.3 ± 5.6	Proposed	4.8 ± 4.6
REPET-SIM	3.9 ± 4.0	Oracle	5.1 ± 4.5

Table 2: Mean and standard deviation of the true SDR achieved by using different source separation algorithms on the testing set. Our proposed method for selecting algorithms based on predicted performance approaches the oracle.

Its performance will be low when the vocals are soft and high when the vocals are loud. To vary the effectiveness of the melodic cue, we altered the mixing ratio of the vocals to the background to either -5 dB, 0 dB, or +5 dB.

We extracted a single 30-second excerpt from each track in DSD100. Each of the data augmentations (pitch shifting, clipping, varying vocal source angle, varying mixing ratio) was either applied or not applied to each mixture. Of the total number of mixtures, 1/2 are pitch shifted, 1/2 are clipped, 1/3 are at the same mixing ratio, and 1/5 have the vocal source at the same angle as another source. This results in 60 mixtures for each 30-second excerpt from the four stems ($2 \times 2 \times 3 \times 5$: pitch shifted or not pitch shifted, clipped or not clipped, number of mixing ratios, number of vocal source angles). This resulted in 2820 mixtures built from the development tracks and 2423 mixtures from the testing tracks. The combined dataset contains 5243 mixtures, or about 44 hours of audio. 9 tracks from DSD100 were excluded due to lack of vocal or accompaniment energy in the resulting mixtures.

We ran each of the three source separation algorithms (PROJET [4], REPET-SIM [21], and MELODIA [6]) on every mixture in the dataset, with the goal of separating out the vocals from the accompaniment. The mixtures and separated sources from each algorithm were then downsampled to 8000 Hz and converted to mono signals. This downsampling from 44.1 kHz was done in order to reduce the computational load for our model (see Section 2.3).

Source-to-Distortion Ratio (SDR) [20] is a common benchmark in the audio source separation literature. We divided each 30-second mixture and its corresponding separated vocal track produced by PROJET, REPET-SIM or MELODIA into non-overlapping 1-second segments and computed the SDR over each short segment (30 SDR values per mixture). As a result of our data augmentation approach, we got a wide range of SDR values (e.g. high-quality to low-quality separations). SDR for our dataset ranges between -30 dB and 30 dB. In total, we computed $30 \times 5423 \times 3$ (# of seconds per track, # of mixtures, and # of source separation algorithms) = 488,070 examples. Each example consists of a 1-second mixture, a corresponding 1-second separated vocal sources from one of the three separation approaches, and the SDR for the separated source. Of these 488,070 examples, 270,000 were built from the development tracks in DSD100 and 218070 were built from the test tracks. The first 270,000 examples were used as the training set for our model and the second 218,070 examples as the testing set. We now describe how to predict separation quality without ground truth and how to use the prediction to select the optimal source separation output.

2.3. Predicting SDR without true sources

The official method to calculate the SDR, as defined in the Blind-Source-Separation Evaluation toolkit [20] requires the separated vocal track, the true vocal track, and the true accompaniment track. We present a system that can estimate SDR without requiring the true vocal track or the true accompaniment track. We do this by training a deep neural network that maps from 1 second of mixture and 1 second of separated vocal track to its corresponding SDR. This allows estimation of the SDR without access to the ground truth sources.

2.3.1. Neural Network Architecture

Our network consists of a series of convolutional layers followed by a series of fully connected layers that are trained via standard back-propagation to perform regression. The input to the network is single-channel (mono) time-domain audio for 1 second of mixture and 1 second of a corresponding separated vocal track, sampled at 8000 Hz. The output of the network is the SDR value for the separated vocal track. The full network architecture is detailed in Table 1. We applied batch normalization after the Input, Conv1, Conv2, and Conv3 layers. The model was trained using Stochastic Gradient Descent with a learning rate of $.001$ and momentum set to $.9$. The cost function was mean squared error between the predicted SDR and the actual SDR. We trained the network for 50 epochs with a batch size of 128. Of the training set,

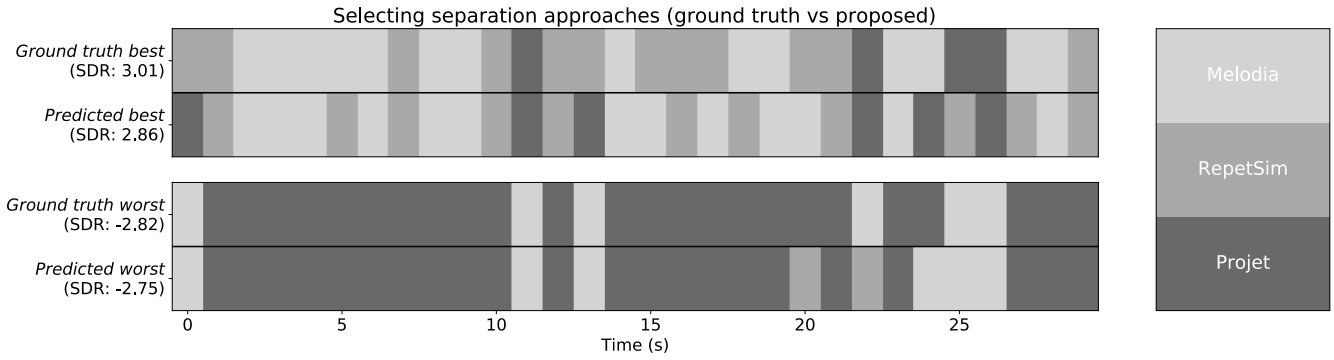

Figure 1: Predicted source separation approaches over time for an example mixture from the DSD 100 dataset. The top graph shows the ground truth best source separation approach to use for every 1-second segment versus the approach predicted as best by our system. The bottom graph shows the ground truth and predicted worst algorithms to use. We see that our system accurately rejects PROJET for this mixture, and correctly chooses either MELODIA or REPET-SIM as the best algorithm.

10% were held out as a validation set. The remaining 90% were used for training. The network with the best performance on the validation set was kept.

3. EVALUATION

The performance of the trained predictor was evaluated using the testing set, which contains 72,690 1-second mixtures. For a single mixture, we had 3 corresponding vocal source estimates, one from each algorithm. Our goal was to select the vocal estimate with the highest SDR. In the case of the trained network, the one with the highest predicted SDR value is selected. The trained network receives each mixture and a corresponding separated vocal track as input and outputs the predicted SDR. We compared this case to:

1. A random selection of algorithms according to a uniform distribution.
2. Always selecting one of the algorithms (PROJET, REPET-SIM, MELODIA).
3. Selecting algorithms based on the ground truth SDR ("Oracle" selector), *i.e.*, the best any system could do.

The results for this experiment are shown in Table 2. Our prediction of SDR is reliable enough to discern between competing separation approaches on the same mixture and selects the best approach with close to oracle performance. The regression between predicted SDR values and ground truth SDR values for each example in the testing set has a slope of .85 with an r-squared value of .86, indicating a statistically significant linear correlation between predicted and ground truth SDR. The mean squared error of the regression is 12.2 dB. However, given a pair of different SDR values (a, b) for two one-second segments, where $a > b$, the model predicts that $a > b$ with 89.8% accuracy. Therefore, our model does well when estimating the relative quality of source separation output.

Next, we compare the predicted best approaches to the the approaches produced by the oracle. We measure the rate of success of the algorithm selector using precision and recall. The precision of picking the best algorithm when the best algorithm is PROJET, REPET, or MELODIA are 65.3%, 71.31%, and 70.2%, respectively. The recall rates are 87.5%, 80.3%, and 78.08%. The precision of picking the worst algorithm when the worst algorithm is PROJET, REPET, and MELODIA are 88.83%, 63.5%, and 68.7%, respectively. The recall rates are 79.6%, 92.4%, and 88.9%. In short, the predictor performs reasonably well when predicting the best algorithm and does a very good job of avoiding algorithms that had poor performance. Figure 1 shows the output of our system on a sample mixture. [1]

4. CONCLUSION

We have presented an approach to using a deep neural network to estimate the source separation quality for three source separation algorithms, each leveraging a different cue: repetition, spatialization, and harmonicity/pitch proximity. This estimate does not require the ground truth separated sources. This lets us use our source separation quality estimator to combine source separation approaches to make an ensemble source separator that performs better than any of the component single-cue source separation algorithms. Our approach for quality estimation can be generalized to arbitrary source separation algorithms. Future work will involve investigating how to use deep neural networks to evaluate and predict audio quality as well as finding more ways of fusing source separation algorithms. We can also use this work to train source separation models without ground truth.

[1] Audio examples at https://interactiveaudiolab.github.io/demos/multicue.

5. REFERENCES

[1] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 73–84, 2013.

[2] Z. Rafii and B. Pardo, "Online repet-sim for real-time speech enhancement," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 848–852, IEEE, 2013.

[3] S. Rickard, "The duet blind source separation algorithm," *Blind Speech Separation*, pp. 217–237, 2007.

[4] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," 2016.

[5] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[6] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[7] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[8] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 1188–1193, 2011.

[9] C. Darwin and R. Hukin, "Auditory objects of attention: the role of interaural time differences.," *Journal of Experimental Psychology: Human perception and performance*, vol. 25, no. 3, p. 617, 1999.

[10] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[11] A. Schwartz, J. H. McDermott, and B. Shinn-Cunningham, "Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 357–368, 2012.

[12] M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Learning to separate vocals from polyphonic mixtures via ensemble methods and structured output prediction," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 450–454, IEEE, 2016.

[13] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 57–60, IEEE, 2012.

[14] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[15] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1562–1566, 2014.

[16] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 126–130, IEEE, 2015.

[17] P.-S. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng, "Predicting speech recognition confidence using deep learning with word identity and score features," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7413–7417, IEEE, 2013.

[18] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: predicting asr error from temporal properties of speech signal," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7423–7426, IEEE, 2013.

[19] L. Gillick, Y. Ito, and J. Young, "A probabilistic approach to confidence estimation and evaluation," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, pp. 879–882, IEEE, 1997.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix.," in *ISMIR*, pp. 583–588, 2012.

[22] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, vol. 9237 of *Latent Variable Analysis and Signal Separation*, (Liberec, France), pp. 387–395, Aug. 2015.