SISEC 2018: STATE OF THE ART IN MUSICAL AUDIO SOURCE SEPARATION - SUBJECTIVE SELECTION OF THE BEST ALGORITHM

Dominic Ward\textsuperscript{1}, Russell D. Mason\textsuperscript{2}, Chungeun Kim\textsuperscript{1}, Fabian-Robert Stöter\textsuperscript{3}, Antoine Liutkus\textsuperscript{3} and Mark D. Plumbley\textsuperscript{1}

CVSSP\textsuperscript{1} / IoSR\textsuperscript{2}, University of Surrey, Guildford, UK
Inria and LIRMM\textsuperscript{3}, University of Montpellier, France

r.mason@surrey.ac.uk

ABSTRACT

The Signal Separation Evaluation Campaign (SiSEC) is a large-scale regular event aimed at evaluating current progress in source separation through a systematic and reproducible comparison of the participants’ algorithms, providing the source separation community with an invaluable glimpse of recent achievements and open challenges. This paper focuses on the music separation task from SiSEC 2018, which compares algorithms aimed at recovering instrument stems from a stereo mix. In this context, we conducted a subjective evaluation whereby 34 listeners picked which of six competing algorithms, with high objective performance scores, best separated the singing-voice stem from 13 professionally mixed songs. The subjective results reveal strong differences between the algorithms, and highlight the presence of song-dependent performance for state-of-the-art systems. Correlations between the subjective results and the scores of two popular performance metrics are also presented.

1. INTRODUCTION

High-quality separation of the singing voice from accompanying instruments is an important yet difficult task serving many applications, including soloing, karaoke, remixing and upmixing music\textsuperscript{[1]}, and even for studying music production techniques\textsuperscript{[2]}. Despite the plethora of musical applications that audio source separation affords, separation quality is often far from perfect, introducing distortions and artefacts that degrade the perceived quality of the extracted source. Another issue is interference, whereby the unwanted sources remain present to some extent. Nevertheless, notable improvements in signal separation have been made in recent years, particularly due to the surge of deep-learning techniques which have proved to outperform traditional signal processing methods, as documented by the Signal Separation Evaluation Campaign (SiSEC)\textsuperscript{[3]}.

SiSEC was initiated to compare the performance of separation systems on a voluntary and community basis by preparing, organising, deploying and evaluating signal separation challenges that target different domains\textsuperscript{[3]}. The MUS(ric) task of SiSEC is designed to assess systems that extract musical instruments from popular music mixes. It has gained rapid interest from algorithm proponents over recent years, due to the development of high-quality and freely-available audio datasets, as well as its unified platform for training, testing, and submitting source separation algorithms for fair comparison. Separation quality is typically evaluated using objective metrics that quantify distortions between the target and estimated sources; the most common being the blind source separation evaluation (BSS Eval) toolkit\textsuperscript{[4]}, followed by the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolkit\textsuperscript{[5]}. Although these metrics show some predictive ability for singing-voice separation quality\textsuperscript{[6]}, generalisation to other sources and algorithms has been questioned\textsuperscript{[7]}, indicating that further refinement is warranted in order for them to be used for general purpose performance measurement. However, for most applications, subjective quality of the separation result is key; therefore it is imperative that auditory assessments are conducted to measure quality as judged by human listeners.

Accordingly, we have conducted a perceptual experiment using the most recent set of songs and separation algorithms in the SiSEC 2018 MUS task dataset. This paper presents the results of the experiment, with the goal of identifying the separation algorithms that can most successfully extract the singing-voice from a range of song mixtures.

2. SUBJECTIVE ASSESSMENT

2.1. MUSDB18

The MUSDB18 dataset\textsuperscript{[8]} is a freely-available set of 150 professionally mixed songs from a range of genres, each comprising four stereo sources (bass, drums, vocals and a group of other instruments) that sum to produce typical song mixes. For the MUS task, MUSDB18 was divided into 100 songs for training and 50 songs for testing, and 30 systems were evaluated using BSS Eval metrics applied to each of the estimated sources\textsuperscript{[3]}.

2.2. Selecting the stimuli

For the listening test, we limited the selection of test stimuli to the singing-voice (vocals) from the test portion of MUSDB18. As a starting point, we first selected one algorithm from participants with multiple submissions, and then selected the six algorithms (see Table\textsuperscript{[1]} with the highest median source-to-distortion ratios (SDR), a global performance metric provided by the BSS Eval v4 toolkit\textsuperscript{[3]}. The
songs were selected with the goal of obtaining content of varying separability. Among the 50 test songs, 10 were selected to yield a wide range of SDR values; a further three were selected from an informal pilot experiment as containing atypical artefacts. The original vocals were included as reference material for listeners to compare against. Stimuli were segmented into seven-second portions where all instruments were most active \(^9\), and then loudness normalised according to ITU-R BS.1770 \(^10\). was less than 0.05/13, the null-hypothesis was rejected. The standardised residuals (difference between observed count and expected count under the null-hypothesis, divided by square-root of the expected count) were then used to identify algorithms with counts that systematically differed from the expected value, based on a \(z\)-score threshold of 2.64 (corresponding to \( \alpha = 0.05/6 \)). The results are shown in Figure 1.

Table 1: Overview of the six deep-learning based musical audio source separation algorithms evaluated, in descending order of median SDR for the vocals of the MUSDB18 test set.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAK2</td>
<td>Multi-scale multi-band DenseLSTM (^11)</td>
</tr>
<tr>
<td>TAU1</td>
<td>Blending of MMDenseNets and LSTM (UHL3) (^12) [13]</td>
</tr>
<tr>
<td>UHL3</td>
<td>Bi-directional LSTM with 3 BLSTM layers (^12)</td>
</tr>
<tr>
<td>JY3</td>
<td>Denoising auto-encoder with skip connections (^14)</td>
</tr>
<tr>
<td>STL1</td>
<td>Wave-U-Net for end-to-end audio source separation (^15)</td>
</tr>
<tr>
<td>MDL1</td>
<td>Recurrent inference algorithm with masker and denoiser architecture (^16)</td>
</tr>
</tbody>
</table>

2.3. Procedure

Participants of the listening assessment were delegates of the 14th LVA ICA conference \(^1\) and the assessments were undertaken in edit rooms with good acoustic isolation and minimal visual distractions. The stimuli were reproduced over professional-grade studio headphones. Forty-two listeners undertook at least part of the experiment at the conference, and two listeners completed the full experiment online, giving a total of 44 participants. Of these, 34 completed the full experiment.

For each of the 13 songs, participants were asked to select one of six extracted vocals (each corresponding to one of the algorithms listed in Table 1) that they perceived to be most similar to the original vocals. A short training stage involving audio examples informed listeners about the music source separation task and explained what audible deviations should be taken into account as promoting dissimilarity from the original vocals, e.g. the presence of other instruments and artificial noise. The presentation order of both songs (on separate pages) and algorithms (within each page) was randomised for each listener. Participants were able to end the session at any time. \(^2\)

3. RESULTS

The analysis was undertaken using the data from the 34 participants that completed the full experiment. The count data, obtained by summing the number of times each stimulus was selected, were subject to a Chi-Square analysis to test the null-hypothesis that all algorithms are equally likely to be selected for a given song (of 13). If the probability of obtaining a Chi-Square statistic at least as extreme as the one calculated

\[^{1}\text{http://cvssp.org/events/lva-ica-2018/}\]

\[^{2}\text{The test is available at https://cvssp.github.io/sisec-mus-18-pick-the-best/}\]
all separation quality. The loudness-normalised stimuli were input to both metrics, using the vocal accompaniment as the true interfering source. Figure 2 shows the (song-wise) Spearman rank correlation between the observed frequency counts of the test stimuli, and the SDR and OPS measures. The plot suggests that stimuli with higher SDR are more likely to be selected; for seven songs the most frequently selected algorithm (or joint-top) had the highest SDR. OPS is less correlated with the subjective results, including some moderate negative correlations; the algorithm with the highest OPS was most frequently selected in only two songs.

Figure 2: Spearman correlation between the observed frequency counts of the test stimuli, and the corresponding OPS/SDR values, separated by song.

4. CONCLUSIONS

A listening experiment was performed to identify source separation algorithms that can most successfully extract the singing-voice from a range of song mixtures (defined as the separated vocals perceived as most similar to the original vocals), and whether the success is accurately predicted by the objective measures.

The results show that for many of the songs the TAU1 algorithm was most successful, but there was a certain amount of dependence on song, suggesting that specific algorithms may need further enhancement to generalise across genres. For seven songs, the algorithms with notably high counts also had the highest SDR value (BSS Eval toolkit). Spearman correlations between subjective results and SDR were consistently higher than those with OPS (PEASS toolkit), which included some moderate negative correlations. These results indicate that some refinement of objective measures, primarily OPS, is required to accurately reflect the similarity judgements of human listeners.

5. REFERENCES


