

Referenceless Performance Evaluation of Audio Source Separation using Deep Neural Networks

Emad M. Grais^{1,3}, Hagen Wierstorf¹, Dominic Ward¹, Russell Mason², Mark D. Plumbley¹

¹Centre for Vision, Speech and Signal Processing, University of Surrey, UK

²Institute of Sound Recording, University of Surrey, UK

³Electronics, Communications and Computer Engineering Department, Helwan University, Egypt

Abstract—Current performance evaluation for audio source separation depends on comparing the processed or separated signals with reference signals. Therefore, common performance evaluation toolkits are not applicable to real-world situations where the ground truth audio is unavailable. In this paper, we propose a performance evaluation technique that does not require reference signals in order to assess separation quality. The proposed technique uses a deep neural network (DNN) to map the processed audio into its quality score. Our experiment results show that the DNN is capable of predicting the sources-to-artifacts ratio from the blind source separation evaluation toolkit [1] for singing-voice separation without the need for reference signals.

Index Terms—Performance evaluation, deep learning, audio source separation, BSS-Eval sources-to-artifacts ratio.

I. INTRODUCTION

Audio source separation aims to separate one or more target audio sources from mixture signals [2], [3]. The separated sources often contain distortions, artifacts, and unwanted signals from the other sources in the mixtures. An evaluation of the quality of the separated sources is essential to guide the development of separation algorithms or to select the most suitable algorithm for a given mixture signal or application type. This requires either perceptual evaluation where experienced listeners judge the quality of the estimated sources according to different perceptual attributes [4]–[10], or objective metrics that estimate the proportion of distortions, artifacts, or interference present in the separated sources, by comparing these with the reference clean sources [1], [11].

In experimental situations, the reference sources are usually available for use in evaluating the performance of a certain source separation approach. However, for practical applications of source separation, the mixtures are available but the separate original sources (the reference signals) are not. Without these reference sources being available, the most common objective metrics cannot be employed, and the only way to evaluate the quality of the separated sources is to ask listeners to give scores for the quality of the separated sources. Using listeners to evaluate the quality of the separated sources is time consuming, and often unfeasible, and hence an automated system of evaluating the quality of the separated signal using

neither listeners nor reference signals would be preferable. Such an automated referenceless evaluation method could be useful, for example, for selecting the most appropriate source separation algorithms for soloing or karaoke applications for each song, or automatically evaluating whether the separated signals are of sufficient quality or whether extra work is needed to further improve the quality of the separated signals using post-processing or additional separation techniques, e.g. [12], [13].

The concept of referenceless quality evaluation for processed signals has been introduced in many signal processing domains [14]–[17]. In this paper, we propose a referenceless evaluation method to evaluate the quality of the separated audio sources without using the reference sources. The main idea of the proposed method is to train a deep neural network (DNN) to map the estimated separated sources to the output of a reference-based evaluation metric. The metric used in this paper is the Sources-to-Artifacts Ratio (SAR) from the Blind Source Separation Evaluation (BSS-Eval) toolkit [1]. Our previous work [5] on the perceptual correlation of BSS-Eval metrics showed that SAR and Source-to-Interference Ratio (SIR) are good indicators for sound quality and the perceived interference by other sources, respectively. The SAR is selected as a case study, but it is intended that the proposed method will be applied to other objective metrics, or the results of subjective judgments.

The DNN is first trained to map the separated signals from one or more source separation algorithms to their SAR scores. SAR in the training stage of the DNN is calculated by using the reference signals of each source, following [1]. The trained DNN is then used to estimate the SARs of the separated sources without using any reference signals.

We consider three different scenarios of using DNNs to estimate the SAR values. The first scenario is to evaluate how well a DNN can predict the SAR results for the same single source separation algorithm for which it is trained: we refer to this scenario as a *within-algorithm test*. The second scenario is to evaluate how well a DNN can predict the SAR results for a range of separation algorithms when trained using data from that same set of separation algorithms: we refer to this scenario as an *across-known-algorithms test*. The third scenario is to evaluate how well a DNN can predict the SAR results for a range of separation algorithms when trained using data from a different set of separation algorithms: we refer to this scenario

The first two authors started this work when they were researchers at Centre for Vision, Speech and Signal Processing, University of Surrey, UK. They contributed equally to this work. Hagen Wierstorf is now with audeERING GmbH, Gilching, Germany.

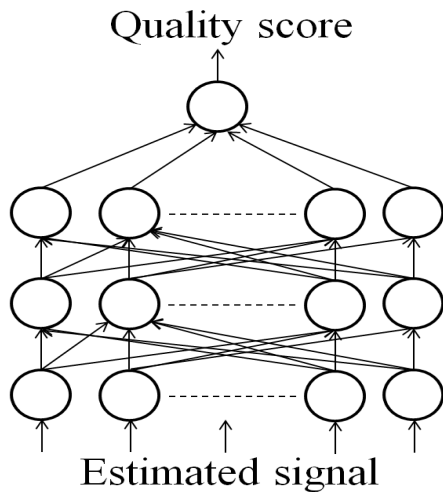


Fig. 1. The deep neural network structure used in this work. The input is the estimated separated signal and the output is its corresponding quality score (SAR, in dB).

as an *across-unknown-algorithm test*.

II. THE BLIND SOURCE SEPARATION EVALUATION TOOLKIT

The Blind Source Separation Evaluation (BSS-Eval) toolkit [1] is the most frequently used tool for evaluating source separation algorithms. BSS-Eval decomposes the error between the reference/target source and the extracted/separated source into a target distortion component reflecting spatial or filtering errors, an artifacts component pertaining to artificial noise, and an interference component associated with the unwanted sources. The salience of these components is quantified using three energy ratios: source Image-to-Spatial distortion Ratio (ISR), Sources-to-Artifacts Ratio (SAR), and Source-to-Interference Ratio (SIR). A fourth metric, the Source-to-Distortion Ratio (SDR), measures the global performance (all impairments combined). Computing these metrics depends mainly on comparing the reference signals and their corresponding estimated signals from the source separation system for each source. Without the reference sources, the BSS-Eval toolkit cannot provide information regarding the quality of the estimated sources.

III. DEEP NEURAL NETWORK FOR REFERENCELESS SAR PREDICTION

In this paper we use a deep neural network to predict the BSS-Eval SAR scores from the output signals of a source separation system. The DNNs we use are fully connected feed forward neural networks as shown in Fig. 1. SAR was selected as a case study: it has been shown to be an indicator of the magnitude of perceptual artifacts in the separated signals [5], [6].

The DNN is trained to map the extracted features of the separated sources to their corresponding SAR values. In this

training stage of the DNN, we assume the reference signals are available. Given the reference or clean signals and their corresponding estimated signals from the source separation technique, the SAR is calculated using BSS-Eval [1]. We extract features from the separated sources and use these features as input to the DNN. The features we use in this work are the mel-frequency spectrogram (MFS), which are calculated by converting the spectrograms of the estimated signals to a mel-frequency scale with 128 frequency channels. The training of the DNN parameters is done by minimizing the mean-square-error between the SAR values estimated by the DNN and their corresponding SAR values calculated using BSS-Eval.

The trained DNN is then used to estimate the SAR values for a new set of separated sources without using the reference signals. The MFS features are extracted from the separated sources and fed to the trained DNN to estimate the SAR values of the input features.

IV. EXPERIMENTS

We undertook a pilot study to predict the sources-to-artifact ratio (SAR) as provided by BSS-Eval. The audio data and the source separation algorithms were taken from the SiSEC-2016-MUS-task challenge [18]. The dataset consists of 100 stereo songs, though four of them are corrupted so were removed. Each song is a mixture of vocals, bass, drums, and other musical instruments. The SiSEC-2016-MUS-task involved separating these four sources from each song in the dataset. In total, 24 different source separation algorithms with differing performance were submitted to this challenge. The following submitted source separation algorithms are blind source separation algorithms: DUR [19], KAM [20], OZE [21], RAF [22], JEO [23], and HUA [24], and the following submitted algorithms are supervised source separation algorithms using deep neural networks: STO [25], UHL [3], NUG [26], CHA [27], GRA [28], and KON [29]. The separated signals using the Ideal Binary Mask (IBM) [18] are also included in this data. More details about each algorithm can be found on the SiSEC-2016 website [29]. These source separation algorithms produced separated signals with a wide range of SAR values (from -10 dB to 20 dB).

In our experiments we aimed to predict the SAR for the vocal separated from each song for all the source separation algorithms that were submitted to this challenge. We tested three different scenarios of varying difficulty:

- Test 1: The DNN model was used to predict the SAR for the source separation algorithm for which it had been trained. We call this test a *within-algorithm test*. This was conducted separately for each separation algorithm to examine any algorithm-dependence in the results.
- Test 2: The DNN model was trained using data from all 24 source separation algorithms simultaneously, then used to predict SAR values of each of the 24 source separation algorithms. We call this test an *across-known-algorithms test*.

- Test 3: The DNN model was trained using data from 17 source separation algorithms simultaneously, then used to predict SAR values for the remaining 7 source separation algorithms not used in the training. We call this test an *across-unknown-algorithm test*.

The 96 available songs (uncorrupted) from SiSEC-2016 dataset were split into 67 training songs and 29 test songs, all processed by the algorithms used in the tests. As the perceptual quality varies over time for musical signals, the SAR was calculated every 117 milliseconds (ms) over a time window of 464 ms on a 116 second (s) excerpt of every song. Calculating the SAR over short time windows also gives more data for training the DNN than calculating SAR over the whole songs. The goal of the trained DNNs was to predict the time-varying SAR for every song and source separation algorithm in the test dataset. The DNNs were deep fully connected feed forward neural networks as shown in Fig. 1, consisting of three hidden layers using a rectified linear unit (ReLU) activation function for all but the last layer, which used a linear activation function. The number of nodes in each hidden layer was 500. The input features were calculated as follows: the stereo inputs were converted to mono by taking the average between the two channels; the spectrogram was calculated and converted to mel-frequency spectrograms (MFS) with 128 frequency channels. We stacked 40 neighbour MFS frames to form the inputs of the DNN with dimension $40 \times 128 = 5120$ MFS values, where 40 is the number of stacked frames, and each frame contains 128 frequency bands.

To evaluate how well the DNNs could predict the SAR values without using the reference signals, we compared the estimated SAR as output from the DNNs with the SAR values calculated from the BSS-Eval toolkit using the reference signals; the average absolute error and the correlation between these were used to evaluate the performance of the DNN accuracy.

V. RESULTS

Table I shows the mean absolute error in dB and the mean Pearson correlation between the referenceless estimated SAR values using DNNs and the calculated SAR using BSS-Eval with reference signals (reference SAR) for the three scenarios (Test 1 to Test 3).

A. Test 1: the within-algorithm test

Test 1 was intended to be a case where a DNN could be trained individually for a given separation algorithm, and hence should give the most favourable results as the DNN is customised for a single case. For this, we independently trained 24 DNNs: one for each source separation algorithm. Each DNN in this case was used to estimate the SAR for the separation algorithm for which it was trained. The same set of training songs and the same set of test songs was used for each algorithm, with no overlap between the two sets of songs. The error in the predictions was calculated as the difference between the predicted SAR from each DNN, and the reference SAR for the same separated signal. The mean

Algorithm	Test1		Test2		Test3	
	Error	Corr.	Error	Corr.	Error	Corr.
CHA	1.2	0.82	1.5	0.83	0.7	0.89
GRA2	1.4	0.87	1.5	0.86	1.3	0.92
GRA3	1.3	0.80	1.6	0.81	1.7	0.89
IBM	1.3	0.90	2.9	0.86	3.1	0.93
JEO1	0.8	0.89	1.3	0.76	0.9	0.89
KAM1	1.2	0.83	1.2	0.79	0.9	0.87
KAM2	0.9	0.81	1.0	0.75	0.6	0.85
KON	1.3	0.90	1.3	0.88	1.3	0.92
NUG1	1.4	0.89	1.1	0.88	0.5	0.95
NUG2	1.3	0.89	1.1	0.88	0.5	0.96
NUG3	1.4	0.89	1.2	0.89	0.8	0.95
OZE	1.0	0.72	1.1	0.73	0.9	0.80
RAF1	0.9	0.75	1.3	0.72	1.2	0.78
STO1	1.1	0.90	1.0	0.87	0.5	0.94
UHL3	1.5	0.86	1.8	0.85	1.5	0.93
NUG4	1.5	0.89	1.2	0.89	1.6	0.92
UHL2	1.5	0.84	1.7	0.85	1.5	0.90
DUR	1.2	0.75	1.7	0.72	3.7	0.74
HUA	0.8	0.66	1.1	0.61	4.4	0.30
JEO2	0.8	0.95	1.1	0.93	1.6	0.93
RAF2	1.0	0.77	1.1	0.73	1.4	0.70
RAF3	1.0	0.82	1.4	0.78	2.0	0.79
STO2	1.1	0.90	1.0	0.88	1.1	0.88
UHL1	1.4	0.85	1.3	0.86	1.5	0.86

TABLE I
THE MEAN ABSOLUTE ERROR IN DECIBELS AND THE MEAN PEARSON CORRELATION BETWEEN THE REFERENCELESS ESTIMATED SAR VALUES USING DNNs AND THE CALCULATED SAR USING BSS-EVAL WITH REFERENCE SIGNALS (REFERENCE SAR) FOR EACH SOURCE SEPARATION ALGORITHM. THE HORIZONTAL LINE SEPARATES THE ALGORITHMS USED FOR TRAINING (ABOVE THE LINE) AND THOSE USED FOR TESTING (BELOW THE LINE) IN TEST 3.

absolute error between the predicted and reference SAR was 1.2 dB, and ranged from 0.8 dB to 1.5 dB for each separation algorithm. The correlation between the predicted and measured SAR ranged from 0.66 to 0.95 for each algorithm, with an average over the 24 algorithms of 0.84.

Compared to the range of SAR values of -10 dB to 20 dB, the mean absolute error of 1.2 dB represents 4% of the range. This suggests that the SAR values estimated without using a reference could be used to discriminate between the performance of some combinations of algorithm and song. However, it may not be able to discriminate between the average results of some of the algorithms in the SiSEC-2016-MUS-task [18], and hence further refinement is required.

B. Test 2: the across-known-algorithms test

Test 2 was intended to be a case where a single DNN was trained using a set of separation algorithms, and this used to attempt to predict the results of any separation algorithm included in its training set. This requires a more generalised set of predictions compared to Test 1, and hence was intended to be a more challenging test. The single DNN was trained using the same training set of songs employed in Test 1, though this time using the results from all 24 source separation algorithms. The trained DNN was then used to evaluate the separated vocal signals from the test set songs individually for each of the same 24 source separation algorithms. The results are shown in Table I: the mean absolute error between the predicted and

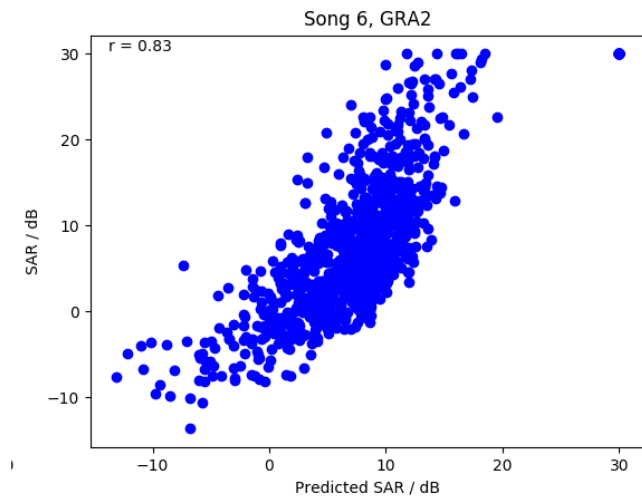


Fig. 2. The correlation between the estimated and reference SAR values for a song separated by source separation algorithm GRA2.

reference SAR was 1.4 dB, and ranged from 1.0 dB to 2.9 dB for each separation algorithm. The correlation between the predicted and measured SAR ranged from 0.61 to 0.93 for each algorithm, with an average over the 24 algorithms of 0.82.

As an example of the correlation between the estimated and actual SAR results, Fig. 2 shows the correlation between the estimated and reference SAR values for a song separated by source separation algorithm GRA2 [18]. As can be seen from the figure, the estimated SAR values are highly correlated with the reference SAR.

Compared to the range of SAR values of -10 dB to 20 dB, the mean absolute error of 1.4 dB represents nearly 5% of the range. Though the performance is less accurate for this more challenging test, even the worst-case mean absolute error of 2.9 dB indicates that the referenceless SAR prediction could be used to discriminate between the performance of some combinations of algorithm and song, but again further refinement is required.

C. Test 3: the across-unknown-algorithm test

Test 3 was intended to be a case where a single DNN was trained using a set of separation algorithms, and this used to attempt to predict the results of any separation algorithm, including those not included in its training set. This requires further generalisation of the results, to both songs and algorithms outside of the training set, and is the most challenging of the tests used. For this, the first 17 source separation algorithms in Table I were used for training and validation, and the last 7 algorithms (separated by a horizontal line in Table I) were used for testing; the training and testing were again undertaken using a separate sets of songs. In addition, the DNN was tested separately for each source separation algorithm using solely the songs from the test set, with the results shown in Table I. The mean absolute error between the predicted and reference SAR was 2.3 dB, and ranged from

1.1 dB to 4.4 dB for each separation algorithm in the test set, and from 0.5 dB to 3.1 dB for each separation algorithm in the training set. The average correlation between the predicted and measured SAR time series was 0.74, with a range of 0.3 to 0.93 for the test set and 0.78 to 0.96 for the training set.

As expected, the performance was less accurate for this test, though the worst-case error would still allow discrimination between some combinations of algorithm and song.

VI. CONCLUSIONS

In this paper we introduced a novel referenceless evaluation method to assess a range of audio source separation systems without the need for the original sources. We used a deep neural network to predict the sources-to-artifacts ratio (SAR) [1] of singing-voice recordings extracted from music mixtures of varying genres. Our experimental results show that the DNNs were capable of predicting the SAR without the reference signals, in most cases resulting in an error that was low enough (mostly <1.5 dB) to allow discrimination between the performance of some combinations of algorithm and song, and with a high correlation (mostly >0.80) between the computed SAR from BSS-Eval that uses the reference signals. This work indicates that the idea of using DNNs to predict the output of objective source separation evaluation toolkits without the use of reference signals produces useful results, and can be extended to train the DNNs to predict the other metrics of the BSS-Eval or predict perceptual related quality scores.

ACKNOWLEDGMENT

This work is supported by grants EP/L027119/2 and EP/N014111/1 from the UK Engineering and Physical Sciences Research Council (EPSRC).

REFERENCES

- [1] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–69, Jul. 2006.
- [2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1066–1074, Mar. 2007.
- [3] 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 *Proc. ICASSP*, 2017, pp. 261–265.
- [4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.
- [5] D. Ward, H. Wierstorf, R. Mason, E. M. Grais, and M. D. Plumbley, "BSS Eval or Peass? predicting the perception of singing-voice separation," in *Proc. ICASSP*, 2018, pp. 596 – 600.
- [6] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *Proc. AES 143*, 2017.
- [7] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *Proc. ICASSP*, 2016, pp. 619–623.
- [8] P. Coleman, Q. Liu, J. Francombe, and P. Jackson, "Perceptual evaluation of blind source separation in object-based audio production," in *Proc. LVA/ICA*, 2018.
- [9] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proc. EUSIPCO*, 2016.

- [10] I.-R. BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Tech. Rep., Tech. Rep., 2015.
- [11] R. Huber and B. Kollmeier, "PEMO-Q: a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.
- [12] E. M. Grais and H. Erdogan, "Spectro-temporal post-enhancement using MMSE estimation in NMF based single-channel source separation," in *Proc. InterSpeech*, 2013.
- [13] D. Williamson, Y. Wang, and D. Wang, "A two-stage approach for improving the perceptual quality of separated speech," in *Proc. ICASSP*, 2014, pp. 7034–7038.
- [14] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.
- [15] S. Fu., Y. Tsao, H. Hwang., and H. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. InterSpeech*, 2018.
- [16] C. Spille, S. D. Ewert, B. Kollmeier, and B. T.Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech and Language*, vol. 48, pp. 51–66, 2018.
- [17] E. Manilow, P. Seetharaman, F. Pishdadian, and B. Pardo, "Predicting algorithm efficacy for adaptive multi-cue source separation," in *Proc. WASPAA*, 2017, pp. 274–278.
- [18] A. Liutkus, F. Stoter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 Signal Separation Evaluation Campaign," in *Proc. LVA/ICA*, 2017, pp. 323–332.
- [19] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Trans. on Selected Topics on Signal Processing*, vol. 5, no. 6, pp. 1118–1133, Oct. 2011.
- [20] A. Liutkus, D. FitzGerald, Z. Rafii, and L. Daudet, "Scalable audio separation with light kernel additive modelling," in *Proc. ICASSP*, 2015, pp. 76–80.
- [21] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, Oct. 2012.
- [22] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 71–82, Jan. 2013.
- [23] I.-Y. Jeong and K. Lee, "Singing voice separation using RPCA with weighted l1-norm," in *Proc. LVA/ICA*, 2017, pp. 553–562.
- [24] P. Huang, S. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, 2012, pp. 57–60.
- [25] F.-R. Stoter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. ICASSP*, 2016, pp. 126–130.
- [26] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [27] P. Chandna, M. Miron, J. Janer, and E. Gomez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. LVA/ICA*, 2017, pp. 258–266.
- [28] E. M. Grais, G. Roma, A. J. R. Simpson, and M. D. Plumbley, "Single channel audio source separation using deep neural network ensembles," in *Proc. AES-140*, 2016.
- [29] URL, "<https://www.sisec17.audiolabs-erlangen.de>," 2017.