INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Fréchet Audio Distance: A Reference-free Metric for Evaluating Music Enhancement Algorithms

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, Matthew Sharifi

Google AI, Switzerland

{kkilgour, mzuluaga, droblek, mns}@google.com

Abstract

We propose the Fréchet Audio Distance (FAD), a novel, reference-free evaluation metric for music enhancement algorithms. We demonstrate how typical evaluation metrics for speech enhancement and blind source separation can fail to accurately measure the perceived effect of a wide variety of distortions. As an alternative, we propose adapting the Fréchet Inception Distance (FID) metric used to evaluate generative image models to the audio domain. FAD is validated using a wide variety of artificial distortions and is compared to the signal based metrics signal to distortion ratio (SDR), cosine distance, and magnitude L2 distance. We show that, with a correlation coefficient of 0.52, FAD correlates more closely with human perception than either SDR, cosine distance or magnitude L2 distance, with correlation coefficients of 0.39, -0.15 and -0.01 respectively.

1. Introduction

Music enhancement aims to accomplish two goals: separating a music signal from other, interfering noise signals and improving its quality to sound more like studio recorded music. Imagine a mobile phone recording of Vivaldi's Four Seasons played through low quality speakers in a large, reverberant room where a group of people are having a loud discussion. The resulting recording will not be pleasant to listen to.

Video hosting platforms such as YouTube [1] and Vimeo [2] contain a multitude of amateur musical recordings, often captured with a low quality microphone in a setup very different from a recording studio. Such recordings could potentially benefit from music enhancement.

Existing research has looked into techniques for speech separation [3] and speech enhancement [4] as well as separating music into its instrumental components [5] or removing the vocals to produce a karaoke version of the track [6]. Speech enhancement and separation have been active areas of research for many years. Applications include enhancing mobile device recordings, hearing aids and conference call systems.

For the specific task of music enhancement, we found it challenging to quantitatively compare different approaches or models with respect to the perceived quality of their output.

Standard metrics[1] such as signal to distortion ratio (SDR) and signal to interference ratio (SIR) [7], which are typically used to evaluate signal separation algorithms, are able to determine which music enhancement algorithm produces reconstructed music whose signal is closest to a studio recorded original. However, these metrics do not take into account the perceptual quality of the reconstructed music which sometimes results in reconstructions with a lower SDR being more pleasing to listen to. A further disadvantage is that these metrics are full-reference metrics and require a copy of the studio recorded music that the enhancement algorithm should produce.

Based on the Fréchet Inception Distance (FID), introduced by Heusel *et al.* [8] to evaluate generative models for images, we propose the Fréchet Audio Distance (FAD) for evaluating generated audio. FAD compares statistics computed on a set of reconstructed music clips to reference statistics computed on a large set of studio recorded music. We compare both SDR and FAD against human ratings to evaluate their correlation with perceptual quality.

2. Related Work

In speech enhancement there are three overarching approaches for evaluating the quality of the speech: direct signal comparison methods, human evaluations, and signal-based heuristics which are designed to correlate with human evaluation scores.

The first type of approach compares the enhanced speech signal to a reference signal. This includes basic distance metrics such as *cosine distance* and L_2 *distance* as well as ratio metrics such as signal to noise ratio (SNR), SDR and SIR [7]. These full-reference metrics are agnostic to the type of audio being separated or enhanced and can be used for evaluating music enhancement techniques without any changes.

Throughout this paper, we use the implementation of SDR from Raffel *et al.* [9]. Le Roux *et al.* [10] have recently brought to light some weaknesses of SDR and this implementation in particular. They propose a *scale invariant SDR* as an alternative.

Although useful, signal level metrics do not necessarily predict how a human listener will perceive the reconstructed music. For speech enhancement, perceptual level metrics are regularly used, where human raters are asked to compare speech output with ground truth. Human raters are typically provided with individual audio clips of speech and asked to evaluate the *naturalness* of the speech signal on a five point scale from *5 (very natural, no degradation)* down to *1 (very unnatural, very degraded)* and how intrusive the background noise is from *5 (not noticeable)* down to *1 (very conspicuous, very intrusive)* [11].

The third category of speech enhancement metrics, which are not trivially applicable to evaluating music enhancement approaches, are automatic metrics such as Perceptual evaluation of speech quality (PESQ) [12] and short-time objective intelligibility (STOI) [13] that approximate perceptual level metrics without requiring any human raters. Such metrics are designed to correlate with human evaluation scores for speech quality specifically.

In this paper, we propose an automatic metric designed for music enhancement, which is based on the FID metric used to evaluate image-generating GANs. FID uses the coding layer of the Inception network [14] to generate embeddings from an

[1] Throughout this paper, the term metric will be used to mean a measure for quantitative assessment and not necessarily a mathematical measure of distance.

http://dx.doi.org/10.21437/Interspeech.2019-2219

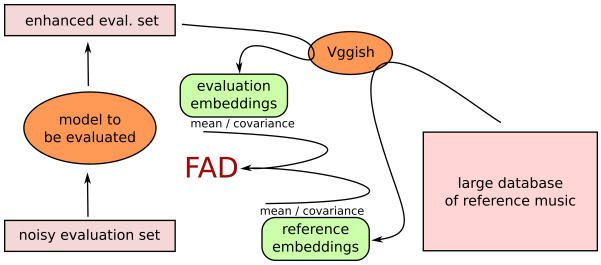

Figure 1: *FAD computation overview: using a pretrained audio classification model, VGGish, embeddings are extracted from both the output of an enhancement model that we wish to evaluate and a large database of background music. The Fréchet distance is then computed between multivariate Gaussians estimated on these embeddings.*

evaluation set of images produced by the GAN and a large set of reference images. The Fréchet distance [15] is then computed between multivariate Gaussians estimated on the evaluation embeddings and the reference embeddings. This approach has also been adapted to videos by Unterthiner *et al.* [16].

3. Fréchet Audio Distance (FAD)

Through our initial work developing techniques for music enhancement, we observed that signal based metrics often disagreed with our own subjective evaluations of the enhanced music. These metrics would penalize enhanced music that differed from the ground truth signal, even when it would sound more like studio quality music to a human listener. To this end, we propose FAD: a reference-free metric which is designed to measure how a given audio clip compares to clean, studio recorded music.

3.1. Definition

Unlike existing audio evaluation metrics, FAD does not look at individual audio clips, but instead compares embedding statistics generated on a full evaluation set with embedding statistics generated on a large set of clean music (e.g. the training set). This makes FAD a reference-free metric which can be used to score an evaluation set where the ground truth reference audio is not available. Where FID uses the activations from a hidden layer in the Inception network [14] to generate embeddings, FAD uses embeddings generated by the VGGish [17] model.

As shown in Figure 1, this gives us a set of reference embeddings from the clean music and a set of evaluation embeddings from the output of the music enhancement model that we wish to evaluate.

We then compute multivariate Gaussians on both the evaluation set embeddings $\mathcal{N}_e(\mu_e, \Sigma_e)$ and the reference embeddings $\mathcal{N}_r(\mu_r, \Sigma_r)$. Dowson *et al.* [15] show that the Fréchet distance between two Gaussians is:

$$\mathbf{F}(\mathcal{N}_b, \mathcal{N}_e) = \|\mu_b - \mu_e\|^2 + tr(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e}) \quad (1)$$

where tr is the trace of a matrix. When comparing algorithms, both the reference embeddings and the evaluation set of noisy signals passed as input to the algorithm are fixed. We often refer to the FAD computed between embeddings of the denoised evaluation set and the reference embeddings as an algorithm's *FAD score*.

3.2. FAD Embedding Model

VGGish[2] is derived from the VGG image recognition architecture [18] and is trained on a large dataset of YouTube videos, similar to YouTube-8M [19] as an audio classifier with over 3000 classes. The activations from the 128 dimensional layer prior to the final classification layer are used as the embedding.

The input to the VGGish model consists of 96 consecutive frames of 64 dimensional log-mel features extracted from the magnitude spectrogram computed over 1 s of audio. Given that the input requirement of 1 s is considerably shorter than typical evaluation music clips, we extract 1 s windows every t seconds. In an initial round of experiments we analyzed what value should be chosen for t and found that 0.5 s worked well, thereby overlapping each window by 50 %. In order to get a stable FAD score, evaluation embeddings are required from at least 25 minutes of audio.

It is worth noting that the input to the existing VGGish model may not be ideal, given that ignoring the phase and using mel-scaled bins could lead to certain distortions going undetected. We investigate this further in Section 5. We performed some initial investigations into using, non-publicly available models as well as models trained explicitly for this purpose and didn't see a significant improvement in metric quality.

4. Experimental Setup

To verify the usefulness of our FAD metric, we start by computing the reference statistics $\mathcal{N}_b$ over embeddings from a dataset of clean music. We then apply various distortions to our audio clips from the evaluation set and compute statistics on their embeddings. The distortions can be viewed as both artifacts that could possibly be introduced by a music enhancement algorithm, or as interfering noises that were not completely removed. We obtain an FAD score for each parameter configuration of a distortion.

4.1. Artificial Distortions

The intensity of each distortion can be controlled by one or more parameters. We expect that, for a given distortion function, parameter configurations which distort the audio more should have a higher FAD score. We examined the following distortions: *Gaussian noise*, *pops*, *frequency filter*, *quantization*, *Griffin-Lim distortion*, *mel encoding*, *speed change*, *reverberations* and *pitch change*.

All distortions are designed to be unaffected by loudness normalization. Distortions for each test parameter configuration are applied separately and in parallel to each of the audio segments in the evaluation set to generate embeddings. This results in an FAD score for each distortion parameter configuration.

4.2. Data

For our experiments, we use the Magnatagatune dataset [20], which contains 600 hours of music samples at 16 kHz. We randomly split the data into a 540 hours reference clean music set and a 60 hours set for evaluating the metrics. For human evaluations, a 25 minute subset of the 60 hour evaluation set is used, which is split into 300 audio clips of 5 s in length.

[2] VGGish can be downloaded from: https://github.com/tensorflow/models/tree/master/research/audioset

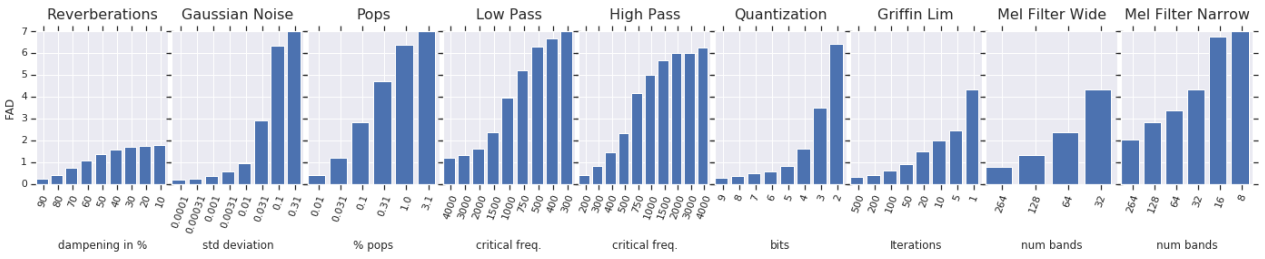

Figure 2: *FAD scores for artificial distortions on the Magnatagatune dataset. For comparison, the FAD computed on the non-distorted clean audio is* 0.2*.*

4.3. Evaluation Metrics

In addition to FAD, we compute the cosine distance, magnitude L2 distance and SDR scores of each parameter configuration of the distortions using:

$$\texttt{SDR}(s_d, s_c) = \frac{\|s_{\texttt{target}}\|^2}{\|e_{\texttt{interf}} + e_{\texttt{noise}} + e_{\texttt{artif}}\|^2} \tag{2}$$

$$\texttt{magnitudeL2}(s_d, s_c) = \||\texttt{stft}(s_d)| - |\texttt{stft}(s_c)|\|_2 \tag{3}$$

$$\texttt{cosdist}(s_d, s_c) = 1 - \frac{s_d \cdot s_c}{\|s_d\|\|s_c\|} \tag{4}$$

where s_d is the distorted audio signal, s_c the corresponding clean audio signal. Please refer to Vincent *et al.* [7] for more details on SDR.

The output range of cosine distance is between 0 and 2, where values closer to 0 indicate that the signals are more positively correlated, values closer to 2 that they are more negatively correlated, and values close to 1 that they are either not at all, or only insignificantly correlated. This follows from the definition of the cosine similarity. We omit the L_2 distance on samples from our analysis because it is equivalent to the cosine distance for normalized signals. Unlike the other metrics where lower values are better, SDR scores signals that are more similar higher. As a result, we plot -SDR to maintain a consistent pattern of lower being considered as better.

4.4. Human Evaluation

For our human-based evaluation, we asked raters to compare the effect of two different distortions on the same 5 s of audio, randomizing both the pair of distortions that they compared and the order in which they appeared. We included the clean original as a pseudo-distortion. The raters were asked *"which audio clip sounds most like a studio produced recording?"* and if they were unable to make a choice after listening to both clips twice, they were able to declare them tied.

5. Results

A representative overview of the FAD scores of some of the distortions described in Section 4.1 with various parameters is shown in Figure 2. Overall, the FAD scores of the distortions generally behave as expected, with FAD scores increasing as the magnitude of the distortion is increased. For the Gaussian noise distortion, the low FAD scores for very small standard deviations are reasonable because such distortions are also barely detectable to a human. Their FAD scores of $0.2-0.3$ are almost

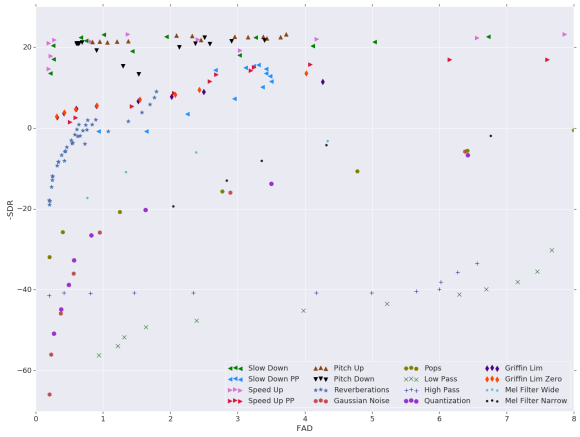

Figure 3: *Comparison of FAD and SDR. The abbreviation PP indicates the pitch preserving variant of a distortion.*

the same as the FAD score of 0.2 computed on non-distorted clean audio.

We observed that distortions with similar FAD scores were of similar subjective quality, e.g. for a music signal with an average root mean square value of 0.18, we perceived Gaussian noise with a standard deviation of 0.031 as having roughly the same effect on quality as adding 16 pops per second, and slightly worse than quantizing it to 4 bits. In Section 5.2 we run a large scale human evaluation to validate our subjective observations.

We verified that using an embedding model which only looks at a mel-scale magnitude spectrogram could still be useful in identifying phase distortions. Removing the phase and reconstructing the signal using Griffin-Lim is noticeable to humans, but often results in audio with an acceptable quality given a sufficient number of iterations. With an iteration parameter of 5, the Griffin-Lim distortion had an FAD score of 2.4. This steadily decreased when the iteration parameter was increased, plateauing out at around 0.31 after 500 iterations.

Applying a mel filter is also detectable using FAD. A wide mel filter with 64 bins results in an FAD score of 2.4, while using 32 bins increases the FAD score to 4.3. Even using 256 bins results in detectable FAD scores for both the narrow and wide variants. These last two results highlight the usefulness of FAD in detecting distortions and irregularities in music signals, and indicate that it should prove useful in evaluating music enhancement algorithms.

5.1. Comparison to Signal Based Metrics

In this section, we compare how different distortions affect SDR and FAD. In Figure 3, we see that SDR is almost invariant to the *high pass* and *low pass* distortions. This is due to SDR being insensitive to certain transformations, which is explored in detail by Le Roux *et al.* [10].

Because FAD does a very good job of detecting these distortions, they form a band along the bottom of the plot. Another band containing *speed up*, *slow down* and *pitch up/down* along the top of the plot are the distortions that consistently get a low SDR score regardless of their intensity, while FAD increases with an increase in intensity.

The remaining distortions form a group of offset logarithm curves, which indicate that each distortion's log FAD scores are correlated with its SDR scores. The offsets between the curves show that the two metrics rate the distortion types differently, with FAD penalizing *Gaussian noise*, *quantization*, *mel filter* and *pops*. SDR is more tolerant of these distortions and, on the other hand, gives *reverberations*, *Griffin-Lim*, and *pitch preserving speed up/slow down* high scores.

5.2. Human Evaluation

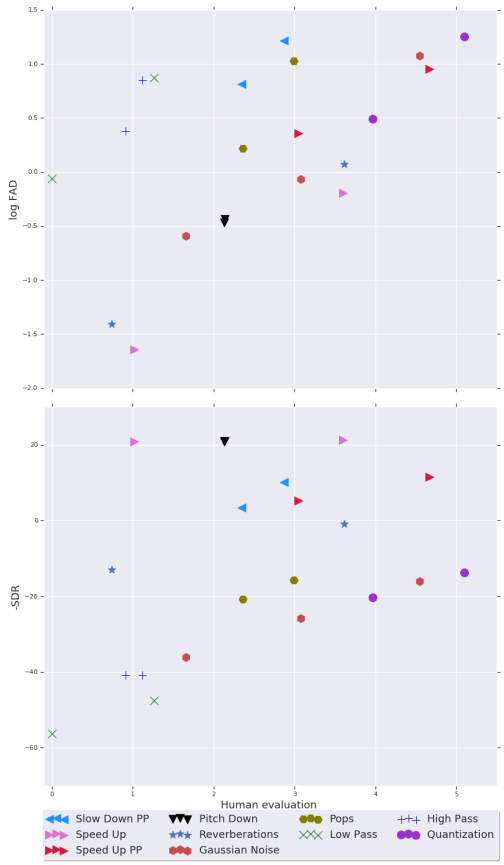

Figure 4: *Results of our human evaluation. The scale on x-axis is the worth value estimated by our Plackett-Luce model. The top plot compares this worth value to the distortions FAD score and the bottom plot compares it to SDR.*

Due to the time-consuming nature of the human evaluation, we only evaluated 10 distortions with total of 21 parameter configurations on 300 audio segments (25 minutes) requiring 69 300 pair-wise comparisons. After some training, the 20 raters were able to compare and rate two 5 s segments in under 40 s. The collected set of pair-wise evaluations was then ranked using a Plackett-Luce model [21], which estimates a *worth value* for each parameter configuration.

Figure 4 plots the worth values estimated by our Plackett-Luce model against both SDR and FAD scores. Neither of the plots shows a perfect correlation. SDR, with a correlation coefficient of 0.39, performs very poorly on *speed up*, *pitch preserving speed up/slow down*, *reverberations* and *pitch down* while correlating quite well with the other distortions.

The plot against FAD also shows some outliers, most noticeably *high pass* and *low pass*. They are, however, still somewhat correlated and overall FAD, with a correlation coefficient of 0.52, correlates better than SDR with how humans rate distortions. The other two examined metrics, cosine distance and magnitude L2 distance both perform significantly worse than either FAD or SDR with correlation coefficients of -0.15 and -0.01.

6. Conclusion

In this paper, we proposed the reference-free FAD. metric for measuring the quality of music enhancement algorithms by comparing statistics of embeddings generated by their output to statistics of embeddings generated on a large set of clean music. Unlike other metrics, FAD can be computed using only an algorithm's enhanced music output, without requiring access to either the original clean music or noise signal.

By testing a large, diverse set of artificial distortions, we have shown that FAD can be useful in measuring the intensity of a given distortion. We compared it to traditional, signal based evaluation metrics such as SDR, and found that FAD can be particularly useful for distortions which always lead to low SDR scores independent of the distortion intensity. Our evaluation using human raters showed that FAD correlated better with human ratings than SDR. These results highlight the usefulness of FAD as metric in measuring the quality of enhanced music and we hope to see others adopt it to report their results.

The source code for FAD is available on GitHub[3]. We encourage those working on audio generation tasks to try it out.

7. Future Work

Our goal was to develop a useful metric for evaluating music enhancement models and we have evaluated FAD as such. However, we suspect that FAD may also prove useful for evaluating a myriad of other audio enhancement and audio generation algorithms.

8. Acknowledgements

The authors would like to thank Javier Cabero Guerra, Pierre Petronin, Trisha Sharma and the rater team for conducting our large-scale human evaluation. We thank Kevin Wilson for the insightful comments that have greatly improved this publication. We further thank Marvin Ritter, Félix de Chaumont Quitry, Dan Ellis, Dick Lyon, Marco Tagliasacchi, Sammy El Ghazzal, David Ramsay, and the Google Brain Zürich team for their support and helpful conversations.

[3]Code available here: https://github.com/google-research/google-research/tree/master/frechet_audio_distance

9. References

[1] *YouTube*, `http://www.youtube.com`.

[2] *Vimeo*, `http://www.vimeo.com`.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *arXiv preprint arXiv:1708.07524*, 2017.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd. Boca Raton, FL, USA: CRC Press, Inc., 2013, ISBN: 1466504218, 9781466504219.

[5] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," *arXiv preprint arXiv:1805.08559*, 2018.

[6] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *ISMIR 2017*, 2017.

[7] 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.

[8] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[9] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, Citeseer, 2014.

[10] J. L. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" *arXiv preprint arXiv:1811.02508*, 2018.

[11] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, IEEE, vol. 2, 2001, pp. 749–752.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, IEEE, 2010, pp. 4214–4217.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[16] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, 2017, pp. 131–135.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[20] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games : The case of music tagging," in *In Proc. wISMIR 2009*.

[21] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 2, pp. 193–202, 1975, ISSN: 00359254, 14679876.