Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, Matthew Sharifi
Google AI

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 [23] and Vimeo [20] 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 [22] and speech enhancement [12] as well as separating music into its instrumental components [13] or removing the vocals to produce a karaoke version of the track [9]. 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) [21], 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. [7] to evaluate generative models for images, we propose the Fréchet Audio Distance (FAD) for evaluating

---

[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.

generated audio. FAD compares statistics computed on a set of reconstructed music clips to background 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 [21]. 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. [15]. Le Roux et al. [11] 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)* [8].

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) [16] and short-time objective intelligibility (STOI) [19] that approximate perceptual level metrics without requiring any human raters. Such metrics are designed to correlate with human evaluation scores for speech quality specifically.

When the result of a speech enhancement algorithm is primarily intended to be used as the input to a subsequent process, e.g. an automatic speech recognition (ASR) system, then it makes sense to measure its quality using an error metric designed for that subsequent process. Chai et al. [2] developed a method called acoustics-guided evaluation which uses an existing acoustic model and compares the posteriors of an enhanced evaluation set to the posteriors of its aligned, cleaned counterpart. The authors showed that this metric is highly correlated with word error rate (WER).

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 [18] to generate embeddings from an evaluation set of images produced by the GAN and a large set of background images. The Fréchet distance [3] is then computed between multivariate Gaussians estimated on the evaluation embeddings and the background embeddings.

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 metric which is designed to measure how a given audio clip compares to clean, studio recorded music.
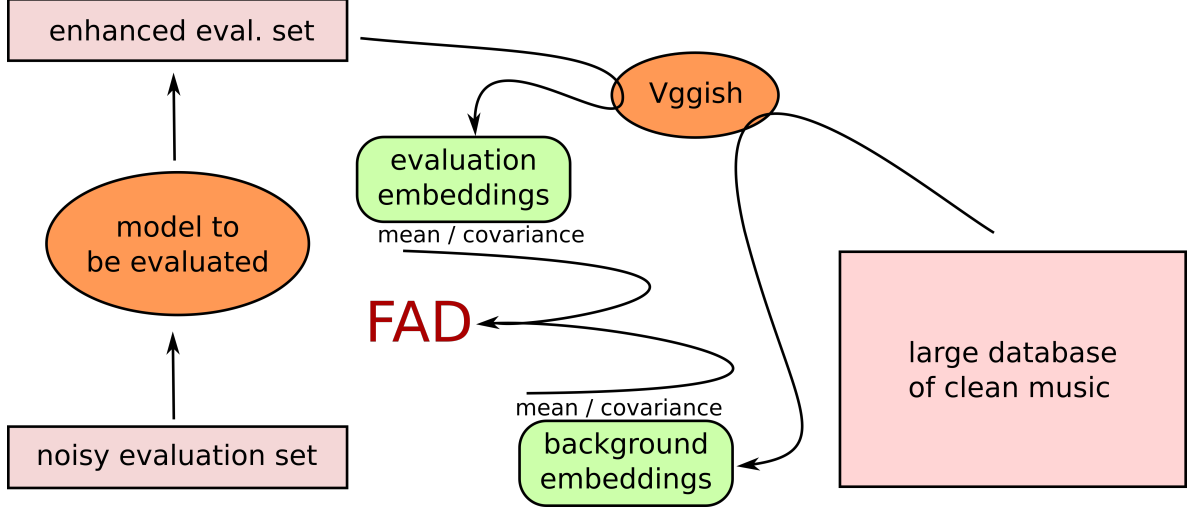
Figure 1: *FAD computation overview: using a pretrained audio classification model, VGGish, embeddings are extracted from both the output of a enhancement model that we wish to evaluate and a large database of background music. The Fréchet distance is then computed between multivariate Gaussians estimated on these embeddings.*

3.1 Definition

Unlike existing audio evaluation metrics, FAD does not look at individual audio clips, but instead compares embedding statistics generated on the whole evaluation set with embedding statistics generated on a large set of clean music (e.g. the training set). This makes FAD a reference-free free metric which can be used to score an 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 [18] to generate embeddings, FAD uses embeddings generated by the VGGish [6] model.

As shown in Figure 1, this gives us a set of background 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 background embeddings $\mathcal{N}_b(\mu_b, \Sigma_b)$. Dowson et al. [3] show that the Fréchet distance between two Gaussians is:

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

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

3.2 FAD Embedding Model

VGGish[2] is derived from the VGG image recognition architecture [17] and is trained on a large dataset of YouTube videos, similar to YouTube-8M [1] 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

---

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

input requirement of $1\,$s is considerably shorter than typical evaluation music clips, we extract $1\,$s windows every t seconds. In Appendix A, we analyze what value should be chosen for t and find that it should be $0.5\,$s, thereby overlapping each window by $50\,\%$.

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.

4 Experimental Setup

To verify the usefulness of our FAD metric, we start by firstly computing the background 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.

Gaussian noise : A distortion signal is sampled from a normal distribution, with $\mu = 0$ and varying σ, and added to the input signal.

Pops : We randomly select p% of the input signal's samples and set half of them to -1 and half to $+1$, or $\pm$abs(`max_input`) if the signal is not normalized.

Frequency filter : The signal is passed through a high or low pass filter with various cutoff frequencies.

Quantization : The signal is reduced from $16\,$bits per sample down to q bits per sample.

Griffin-Lim distortions : The signal is converted to a magnitude spectrogram, and then reconstructed using the Griffin-Lim algorithm [5]. The quality of the reconstructed phase depends on the algorithm's iteration parameter.

Mel encoding : The signal is converted into a mel-scale magnitude spectrogram and back again using the original input phase. We look at two mel variants: narrow, where the mel bins only covers the frequency range from 60 to 6000 Hz, and wide, which covers everything from 0 to 16 000 Hz.

Speed up / slow down : The playback speed of the signal is increased/decreased by a given factor, and as a side effect this also leads to an increase/decrease in its pitch.

Pitch preserving speed up / slow down : The playback speed of the signal is increased/decreased by a given factor using a *phase vocoder* [4] which preserves the signal's original pitch.

Reverberations : Multiple dampened copies of the original signal are added using a provided delay.

Pitch up / down : The pitch of the signal is increased/decreased by a provided number of semitones.

All distortions are designed to be unaffected by loudness normalization. The 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. An overview of the parameters used for each distortion can be found in Appendix C.

4.2 Data

For our experiments, we use the Magnatagatune dataset [10], which contains 600 hours of music samples at 16 kHz. We use 540 hours as the background clean music set and 60 hours for evaluation of 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.

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:

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

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

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

where s_d is the distorted audio signal, s_c the corresponding clean audio signal. Please refer to Vincent et al. [21] 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, for normalized signals, a target t, and output o, the L_2 distance $\|t - o\|_2$ is $\sqrt{2 \, \mathtt{cosdist}(t, o)}$, which does not provide us with any further information relevant to the evaluation. 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

An overview of the FAD scores on 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 the same as the FAD score of 0.2 computed on non-distorted clean audio.
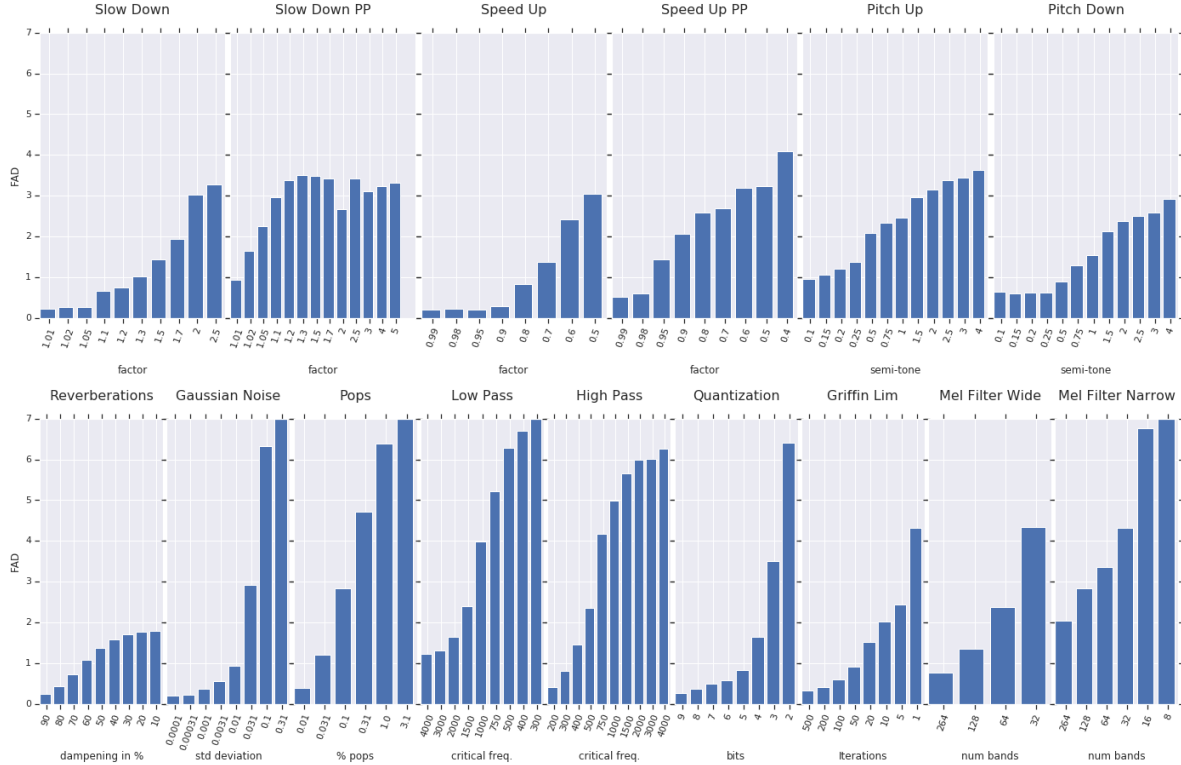
Figure 2: *FAD scores for artificial distortions on the Magnatagatune dataset. The abbreviation PP indicates the pitch preserving variant of a distortion. For comparison, the FAD computed on the non-distorted clean audio is* 0.2.

We observed that distortions with similar FAD scores were of similar subjective quality, e.g. we perceived Gaussian noise with a standard deviation of 0.031 as having roughly the same quality as setting the percentage of pops to 0.1, and slightly worse than quantizing to 4 bits. In Section 5.3 we run a large scale human evaluation in order 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 models.

5.1 Comparison to Signal Based Metrics

In this section, we compare how different distortions affect SDR, FAD and cosine distance. Figure 3 shows the SDR of distortions at various parameter configurations and their cosine distance with three distinct groups of distortions clearly visible. The boomerang shaped group consists of *gaussian noise*, *quantization*, *mel filter*, *pops* and *reverberations* which mostly function as additive distortions and do not affect the signal temporally. For these distortions, their SDR score is proportional to the logarithm of their cosine distance.
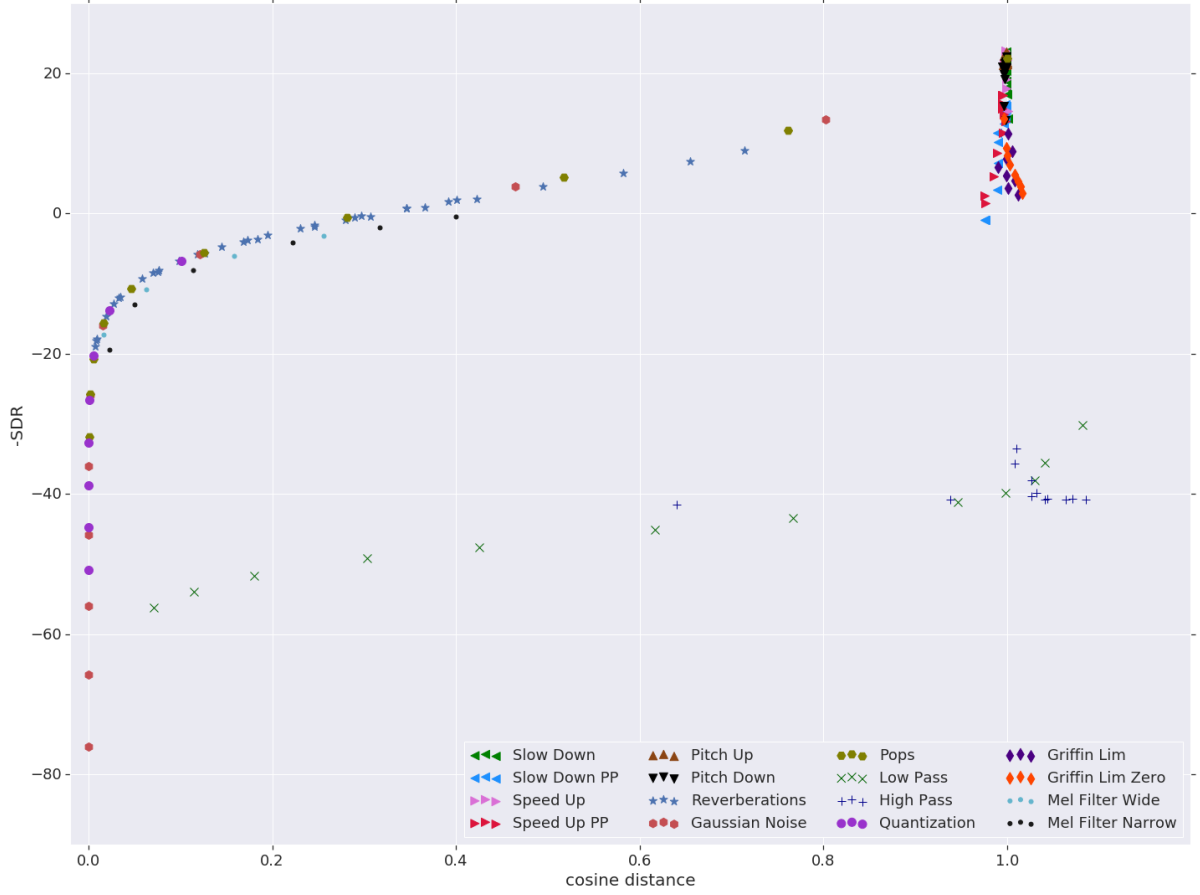
Figure 3: *Comparison of SDR and cosine distance. The abbreviation PP indicates the pitch preserving variant of a distortion. A full overview of the parameter values tested for each distortion can be found in Appendix C.*

The second group, forming a narrow band in top right corner, consists of *speed up/slow down*, *pitch preserving speed up/slow down*, *pitch up/down* and *Griffim-Lim* which displace the signal from its reference by either stretching/compressing the signal or by altering its phase. Each distortion in this group has a cosine distance value of 1, indicating that the signals are completely different as far as cosine distance is concerned. These distortions also result in variable and generally low SDR scores. Because SDR allows for time-invariant filter distortions up to a fixed number of samples T, it can still catch differences between them up to a certain extent.

The final group containing only the high and low pass filters have a cosine distances that are to be expected but surprisingly high SDR scores. This is again due to SDR being insensitive to certain transformations, which is explored in detail by Le Roux et al. [11].

Comparing FAD to cosine distance (Figure 4), we again see two distinct groups of distortions. Along the top of the figure, we find *speed up/slow down*, *pitch preserving speed/slow down*, *pitch up/down*, *Griffin-Lim* and *high pass* all have cosine distance values of 1, regardless of the amount of distortion applied. On the other hand, their FAD values are almost always monotonic and increase when the level of distortion of is increased. The other distortions, *gaussian noise*, *quantization*, *mel filter*, *pops*, *high pass* and *reverberations* appear to be correlated on an individual basis but not between distortions. This implies that, while both metrics can detect these distortions, they rate their severity differently. The cosine distance penalizes *reverberations* and *high pass* more than FAD which is more affected by *gaussian noise*, *quantization*, *mel filter* and *pops*.

Figure 4: *Comparison of FAD and cosine distance. The abbreviation PP indicates the pitch preserving variant of a distortion. A full overview of the parameter values tested for each distortion can be found in Appendix C.*

In Figure 5, we show FAD plotted against magnitude L2 distance. Overall the distortions appear to be individually correlated on log scale, but the two metrics disagree a lot regarding how intense the distortions are. The distortions *gaussian noise*, *quantization*, *mel filter*, *high pass* and *pops* are more highly penalized by FAD, while the others, with the exception of *Griffin-Lim*, are much more highly penalized by magnitude L2 distance. All parameter configurations of the *Griffin-Lim* distortion have magnitude L2 distances that only vary between 0.4 and 0.6 while their FAD scores are more spread out.

The FAD to SDR plot in Figure 6 is more spread out. As before, we see that SDR is almost invariant to the *high pass* and *low pass* distortions. 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.

For the remaining distortions, we see that each distortion's log FAD scores are correlated with its SDR scores. We observe that the two metrics rate the distortion types differently, with FAD again penalizing *gaussian noise*, *quantization*, *mel filter* and *pops*. SDR on the other hand is more tolerant of them and gives *reverberations*, *Griffin-Lim*, and *pitch preserving speed up/slow down* high scores.
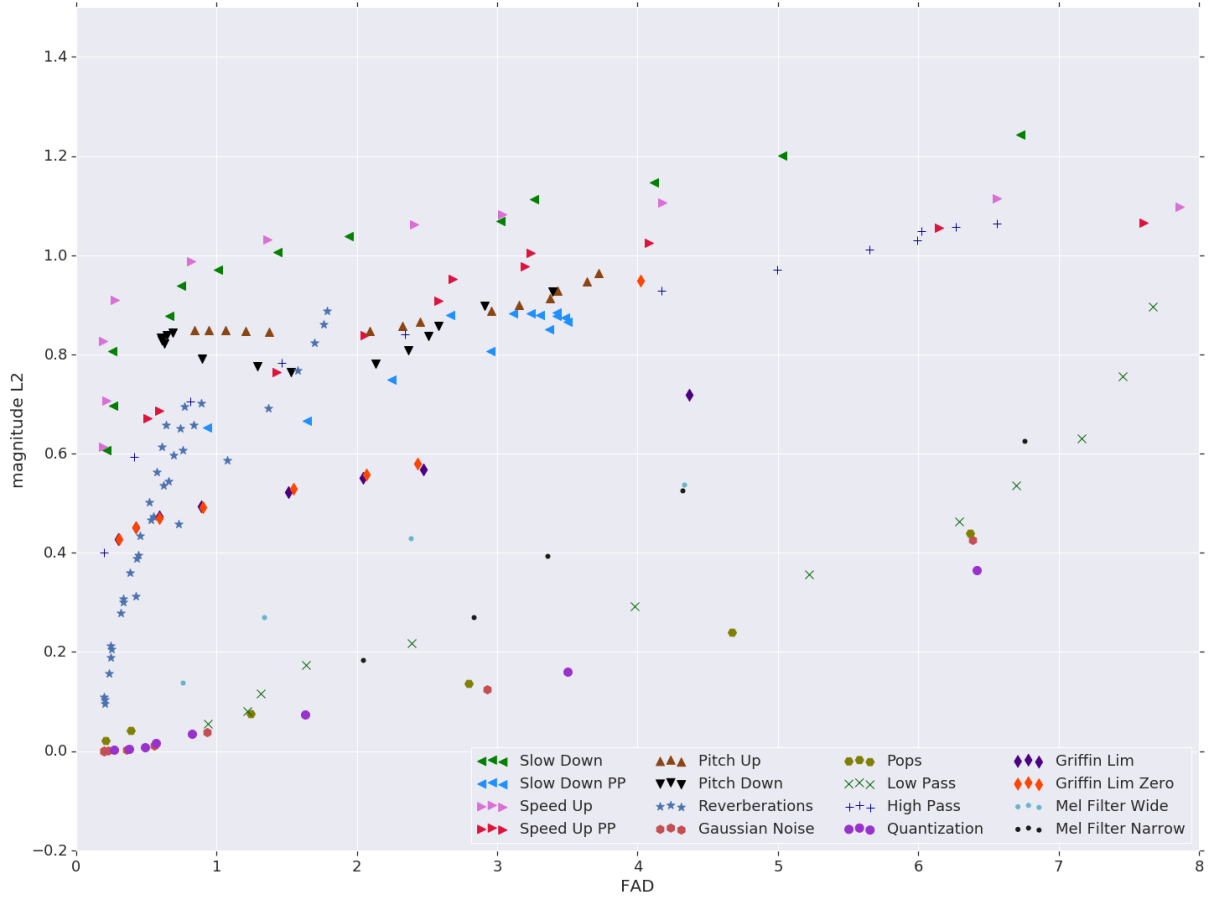
8

Figure 5: *Comparison of FAD and magnitude L2 distance. The abbreviation PP indicates the pitch preserving variant of a distortion. A full overview of the parameter values tested for each distortion can be found in Appendix C.*

5.2 Analysis

Taken as a whole, these comparison plots split the distortions into 4 groups:

SDR-breaking distortions : These are distortions that will lead to very low SDR scores, independent of the distortion parameter configuration. They will generally have high magnitude L2 distances and their cosine distance score will be around 1. Their FAD, on the other hand, appears to be sensitive to these distortions, with a low intensity of distortion having low FAD scores which increases for parameter configurations that cause more intense distortions. For some distortions, FAD appears to plateau by either always having a minimum value no matter how low the distortion parameter, or no longer increasing after a certain maximum distortion parameter. The group includes: *speed up*, *slow down* and *pitch up/down*

Somewhat SDR-breaking distortions : The distortions in this group have low SDR scores that vary with their distortion parameter configuration. While SDR can also detect the intensity of the distortion, their scores will still be very low even for low distortion levels. Their cosine distance is either continuously 1, or behaves similarly to SDR. FAD treats these distortions the same as the SDR-breaking distortions. Their magnitude L2 distances are medium to high. They include: *reverberations*, *Griffin-Lim*, and *pitch preserving speed up/slow down.*

Mainline distortions : For these distortions, all four metrics are low for low parameter con-
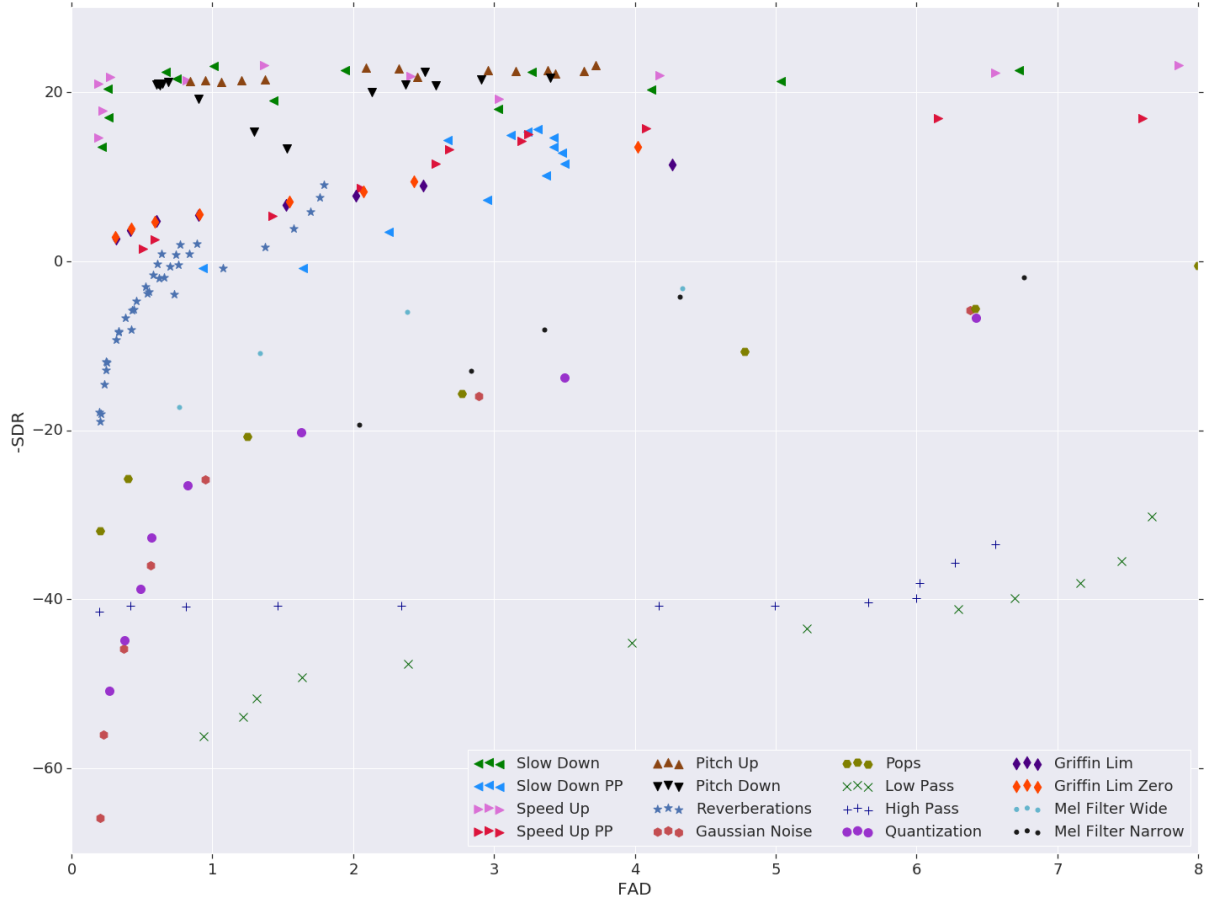
Figure 6: *Comparison of FAD and SDR. The abbreviation PP indicates the pitch preserving variant of a distortion. A full overview of the parameter values tested for each distortion can be found in Appendix C.*

figurations and progressively increase as the amount of distortion increases, although for some of them FAD may still plateau on both the low and high ends. Their rate of increase varies by distortion. This group includes: *gaussian noise*, *quantization*, *mel filter* and *pops*

SDR-tolerant distortions : Unlike FAD, SDR has a hard time detecting *low pass* and *high pass* filters. For the cosine distance *low pass* behaves like the breaking distortions, while *high pass* is detectable and behaves like the mainline distortions. According to the magnitude L2 distance *high pass* behaves more like a somewhat SDR-breaking distortion and *low pass* like a mainline distortion.

5.3 Human Evaluation

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 [14], which estimates a *worth value* for each parameter configuration. The evaluated distortions and their parameter configurations are listed in Appendix C together with their SDR and FAD scores.

Figure 7 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, are plotted against the human evaluation results in Figure 8. With correlation coefficients of -0.15 for cosine distance and -0.01 for magnitude L2 distance, both perform significantly worse than either FAD or SDR. In particular, these two metrics fail at being able to compare between different types of distortions.

6   Conclusion

In this paper, we proposed the reference-free FAD metric for measuring the quality of music enhancement approaches or models 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 a model'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 show 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.

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.

Although our evaluated set of distortions is quite large and diverse, it does not encompass all possible distortions that may occur to signals in either the real world or during enhancement. This is especially true if we wish to adapt FAD to other audio domains. As an area of future work, we would like to evaluate the effectiveness of FAD on further distortions and distortion combinations.

The VGGish embedding model uses log-mel features as input. Future work in this domain should investigate replacing VGGish with models that use other types of input such as raw samples or a complex spectrogram. A key disadvantage of our implementation of FAD is that it only looks at embeddings created on 1 s windows, which means that the metric is unaware of long distance temporal changes within a song. An embedding model which operates on music of variable lengths and computes a single embedding per song may be useful here.

8   Acknowledgements

Figure 7: *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.*
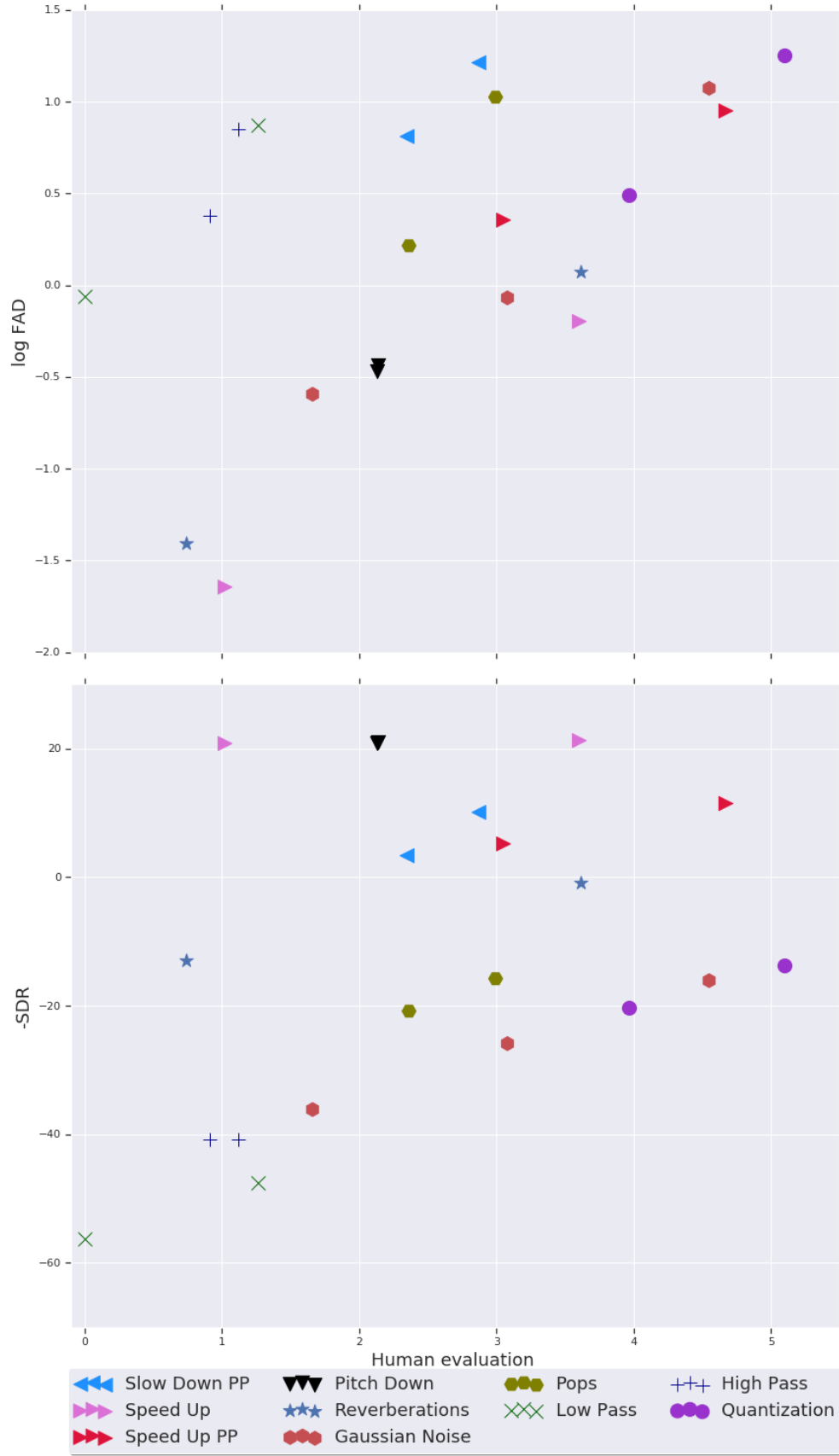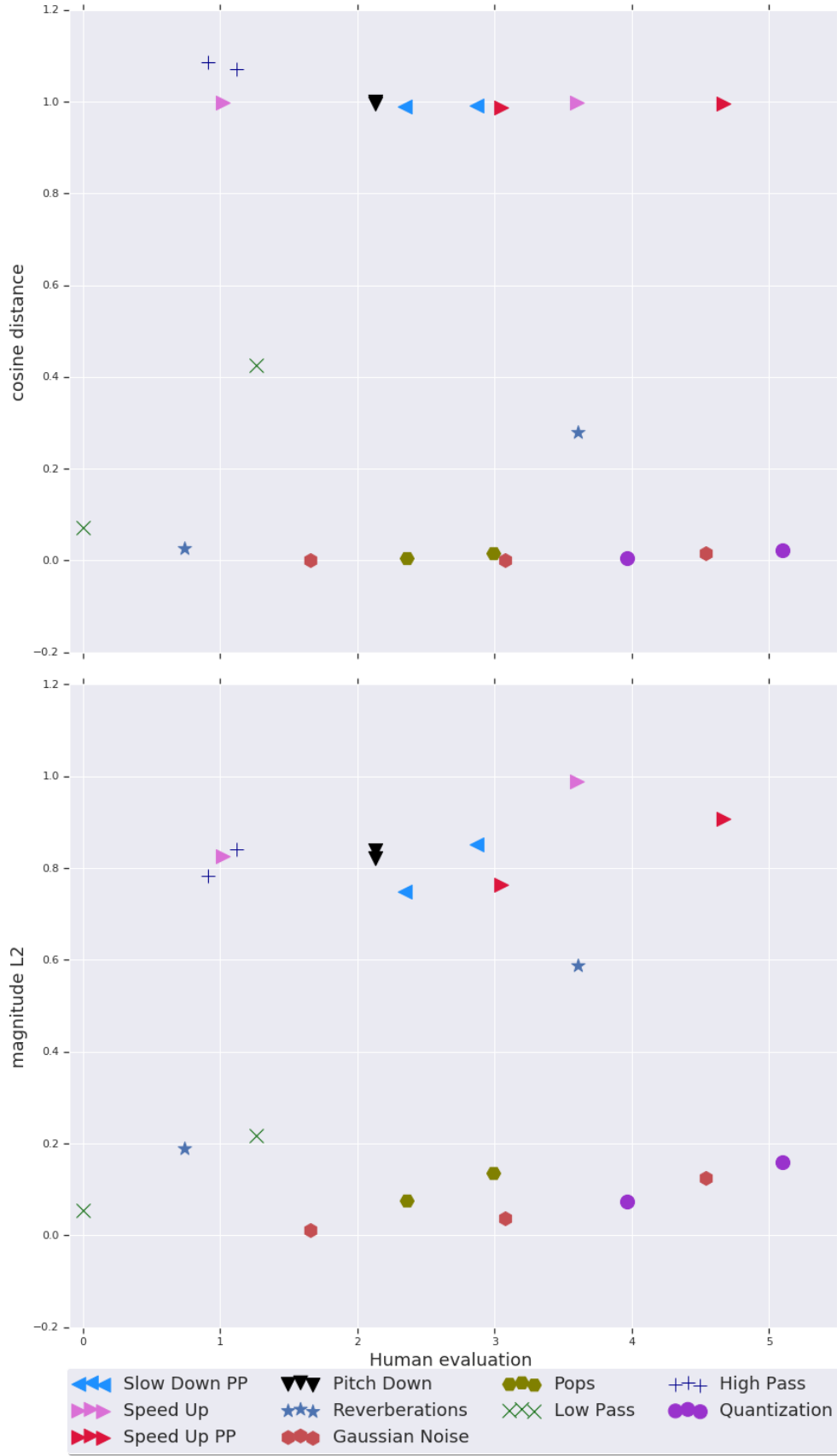
Figure 8: *Results of our human evaluation. The scale on x-axis is the worth value estimated by our Plackett-Luce model. The top plot compares this worth value to the distortions cosine distance and the bottom plot compares it to its magnitude L2 distance.*

References

[1] Sami Abu-El-Haija et al. "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675* (2016).

[2] Li Chai, Jun Du, and Chin-Hui Lee. "Acoustics-guided evaluation (AGE): a new measure for estimating performance of speech enhancement algorithms for robust ASR". In: *ArXiv e-prints*, arXiv:1811.11517 (Nov. 2018), arXiv:1811.11517. arXiv: 1811.11517 [eess.AS].

[3] DC Dowson and BV Landau. "The Fréchet distance between multivariate normal distributions". In: *Journal of multivariate analysis* 12.3 (1982), pp. 450–455.

[4] James L Flanagan and RM Golden. "Phase vocoder". In: *Bell System Technical Journal* 45.9 (1966), pp. 1493–1509.

[5] D. Griffin and Jae Lim. "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243. ISSN: 0096-3518. DOI: 10.1109/TASSP.1984.1164317.

[6] Shawn Hershey 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.

[7] Martin Heusel et al. "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.

[8] Yi Hu and Philipos C Loizou. "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2008), pp. 229–238.

[9] Andreas Jansson et al. "Singing voice separation with deep U-Net convolutional networks". In: *ISMIR 2017* (2017).

[10] Edith Law et al. "Evaluation of algorithms using games : The case of music tagging". In: *In Proc. wISMIR 2009*.

[11] Jonathan Le Le Roux et al. "SDR-half-baked or well done?" In: *arXiv preprint arXiv:1811.02508* (2018).

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

[13] Sungheon Park et al. "Music Source Separation Using Stacked Hourglass Networks". In: *arXiv preprint arXiv:1805.08559* (2018).

[14] R. L. Plackett. "The Analysis of Permutations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.2 (1975), pp. 193–202. ISSN: 00359254, 14679876.

[15] Colin Raffel et al. "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.

[16] Antony W Rix et al. "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*. Vol. 2. IEEE. 2001, pp. 749–752.

[17] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[18] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[19] Cees H Taal et al. "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.

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

[21] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation". In: *IEEE transactions on audio, speech, and language processing* 14.4 (2006), pp. 1462–1469.

[22] DeLiang Wang and Jitong Chen. "Supervised speech separation based on deep learning: an overview". In: *arXiv preprint arXiv:1708.07524* (2017).

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

A   Window Step Size

The VGGish embedding model that we are using requires an input of 1 s of audio. When extracting embeddings from a continuous stream of audio we can either partition the stream into 1 s long chunks or extract embeddings from a 1 s moving window every t seconds. Using a small embedding window step length t will provides us with more embeddings, which may result in us being able to estimate the multivariate Gaussians better.

We compare various embedding window step lengths to determine whether smaller values are useful. The results on some of our distortion configurations can be seen in Figure 9. Overall the FAD scores change very little as we reduce the embedding window step length, indicating that computing many embeddings from highly overlapping segments is not necessary. For a couple of distortion types having non-overlapping windows does change the FAD score slightly, and we therefore recommend using an embedding window step length of 0.5 s.
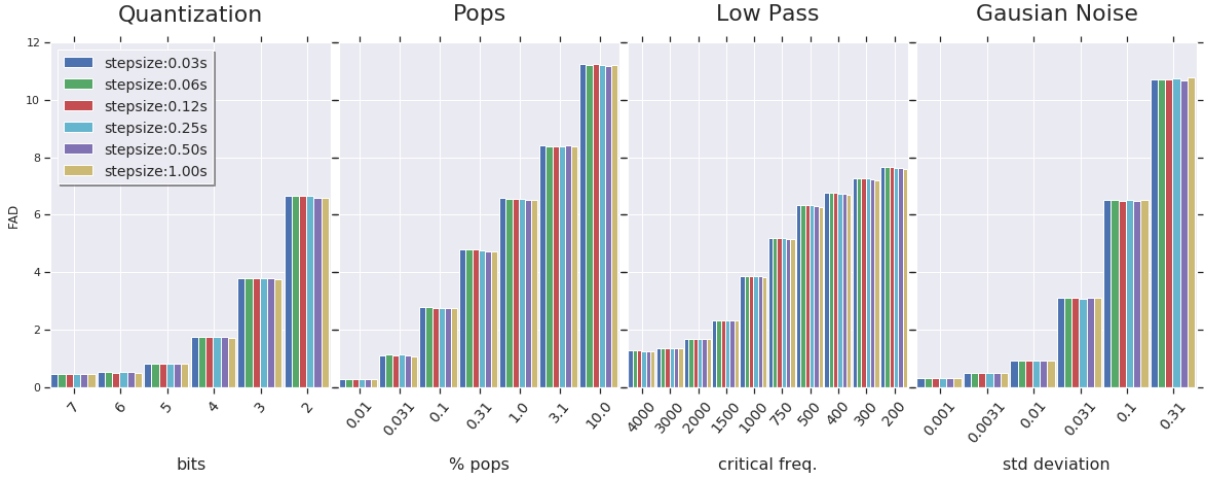

Figure 9: *Overview of the effect of the embedding window step length on a subset of the evaluated distortions.*

B   Evaluation Set Size

As FAD requires measuring the distance between two multivariate Gaussians estimated on sets of embeddings, it can be greatly affected by the size of these sets. Using smaller sets will result in a less accurate estimate of the multivariate Gaussians. In our case, we assume that our set of background embeddings is significantly larger than the set of evaluation embeddings and investigate how large this set needs to be in order to have a stable FAD score.

As described in Section 4.2, we split our evaluation data into 5 s long audio clips. We are able to extract around 40 000 of these 5 s long audio clips from our full evaluation set. We apply distortions with various parameter combinations and extract embeddings using an embedding window step length of 0.5 s.

As possible evaluation set sizes, we consider k audio clips, with k being either 100, 200, 300, 400, 500, 1000, 5000 or 10 000. For each of these possible sizes, we compute the FAD scores of our distortions at various parameter combinations multiple times using different subsets of evaluation audio clips, allowing us to examine how the evaluation set size affects the variance in FAD.

Different distortion types and configurations will have different expected variances, e.g. *Gaussian noise* with a stddev of 0.1 and *speed up* of 1 %. To compensate for this, we therefore we compute the index of dispersion $D = \frac{\sigma^2}{\mu}$ for each distortion configuration which normalizes the variances by the mean.

The average index of dispersion across all distortion configuration can be seen in Figure 10. The very high value for 100 indicates that 100 audio clips or 8 minutes and 20 seconds of audio is not enough to compute a stable FAD score. An ideal amount would be about 5000 audio clips or around 7 hours. This is a lot of data for evaluation purposes and will often not be available. While not as stable as larger evaluation set sizes, we begin to get usable results from about 300 audio clips or 25 minutes of audio.
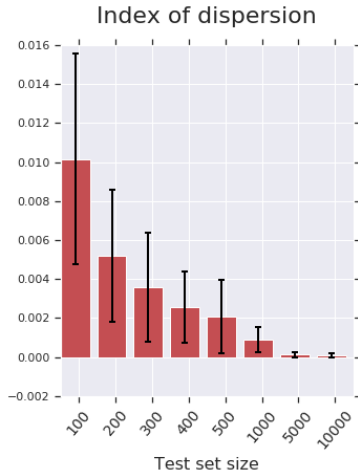

Figure 10: *Average index of dispersion for the test evaluation set sizes across all examined distortion parameter configurations.*

C Evaluated Distortion Parameter Configurations

Distortion	Parameter	Parameter values
Slow Down	factor	1.01, 1.02, 1.05, 1.1, 1.2, 1.3, 1.5, 1.7, 2, 2.5, 3, 4, 5
Slow Down PP	factor	1.01, 1.02, 1.05, 1.1, 1.2, 1.3, 1.5, 1.7, 2, 2.5, 3, 4, 5
Speed Up	factor	0.99, 0.98, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.2, 0.1
Speed Up PP	factor	0.99, 0.98, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.2, 0.1
Pitch Up	semi-tone	0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 5
Pitch Down	semi-tone	0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 5
Reverberations	dampening delay: 1 s echos: 3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Reverberations	dampening delay: 0.5 s echos: 3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Reverberations	dampening delay: 0.25 s echos: 3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Reverberations	dampening delay: 0.25 s echos: 5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Gaussian Noise	std deviation	0.0001, 0.00031, 0.001, 0.0031, 0.01, 0.031, 0.1, 0.31
Pops	% pops	0.0001, 0.00031, 0.001, 0.0031, 0.01, 0.031, 0.1, 0.31
Low Pass	critical freq.	4000, 3000, 2000, 1500, 1000, 750, 500, 400, 300
High Pass	critical freq.	200, 300, 400, 500, 750, 1000, 1500, 2000, 3000, 4000
Quantization	bits	9, 8, 7, 6, 5, 4, 3, 2
Griffin Lim	iterations	500, 200, 100, 50, 20, 10, 5, 1
Griffin Lim Zero	iterations	500, 200, 100, 50, 20, 10, 5, 1
Mel Filter Wide	num. bands	264, 128, 64, 32
Mel Filter Narrow	num. bands	264, 128, 64, 32, 16, 8

Table 1: *List of all the examined distortions and their parameter configurations.*

Distortion	Parameters	Worth	FAD	SDR
low pass	critical frequency: 5000	-0.00	0.94	56
reverberations	dampening: 0.2 delay: 1 s echos: 3	-0.74	0.24	13
high pass	critical frequency: 400	-0.92	1.46	41
speed up	factor: 0.95	-1.02	0.19	-21
high pass	critical frequency: 500	-1.12	2.34	41
low pass	critical frequency: 1500	-1.26	2.39	48
added gaussian noise	stddev: 0.0031	-1.66	0.55	36
pitch down	semi-tone: 0.25	-2.13	0.63	-21
pitch down	semi-tone: 0.1	-2.13	0.65	-21
slow down pp	factor: 1.05	-2.35	2.25	-3
pops	percentage: 0.00031	-2.36	1.24	21
slow down pp	factor: 1.2	-2.87	3.37	-10
pops	percentage %: 0.001	-2.99	2.80	16
added gaussian noise	stddev 0.01	-3.00	0.94	26
speed up pp	factor: 0.95	-3.05	1.43	-5
speed up	factor: 0.8	-3.60	0.82	-21
reverberations	dampening: 0.4 delay: 0.25 s echos: 5	-3.61	1.08	1
quantization	bits: 4	-4.00	1.63	20
added gaussian noise	stddev 0.031	-4.54	2.93	16
speed up pp	factor: 0.8	-4.67	2.58	-12
quantization	bits: 3	-5.10	3.50	14

Table 2: *List of human evaluated distortions, their parameter configurations, worth value estimated by our Plackett-Luce model and both their SDR and FAD scores.*