Evaluation of Fréchet Audio Distance with Bass Sounds

Adelmo Assmar[1], **Flávio Assis**[1*]

[1]LaSiD – Distributed Systems Laboratory
Graduate Program in Mechatronics
Federal University of Bahia
Salvador, Bahia, Brazil

adelmo.assmar@ufba.br, fassis@ufba.br

Abstract. *Objective metrics are important to determine the effectiveness of automatic audio processing, such as in denoising and music source separation. The* Fréchet Audio Distance *(FAD) is an objective metric that can be used to compare two sets of audios based on their statistical properties. An advantage of FAD when compared to traditional metrics is that FAD does not depend on the existence of the original clean version of processed or noisy audios to be evaluated. However, the effectiveness of the use of FAD has been divergent in the literature, requiring thus further investigation. In this paper we study the impact on FAD scores when different reference sets and models are used, for the restricted case of bass sounds. We studied the sensitivity of FAD to audio effects and noises and used it to evaluate denoising processes. We compared the results with a qualified subjective evaluation, carried out by people with experience with professional recording. We show that reference sets must be designed very specifically, as apparently equivalent sets might provide different results. We additionally highlight the usefulness of understanding FAD scores relatively, instead of absolutely.*

Keywords — Fréchet audio distance, automatic audio enhancement, audio evaluation metrics

1. Introduction

The need for assessing the quality of audio appears in different contexts, such as when isolating voices in automatic speech-to-text tools, in Music Source Separation (MSS) and in automatic audio quality improvement. MSS refers to tools and techniques that extract the original constituent parts of a mixed song (stems) and store them in separate audio files [1, 2, 3, 4]. Automatic audio quality improvement consists of tools that automatically generate a better version of a noisy or inaccurate audio, such as an old recording, a speech recorded with low quality equipments, or a stem generated by MSS tools.

Evaluation of audio quality has been generally performed in these contexts in two ways: (a) by human evaluation, when people assess the audios according to their personal feelings based on some audio reference; and (b) by automatic evaluation, using a set of automatically computable metrics. Human evaluation, although subjective, is considered the ground truth measurement in many contexts [5], as humans are the ultimate natural target of many audio applications. However, it is time consuming and might be expensive. Objective metrics can be faster, less costly and are fundamental to the design of automatic audio processing tools.

Many objective metrics have been developed so far. L1, L2 [3], PESQ [6], POLQA [7] and ViSQOL [5] are examples of standard metrics in the context of speech quality. Commonly used metrics in MSS are *Source-to-Distortion Ratio* (SDR), *Source-to-Interference Ratio* (SIR), *Signal-to-Noise Ratio* (SNR), *Source-to-Artifact Ratio* (SAR) [8] and *Scale-Invariant Source-to-Distortion Ratio* (SI-SDR) [9]. These metrics, however, have two drawbacks: (a) they depend on the existence of the original source audio to be used as reference; and (b) it is known that they do not correlate well with human perception and sometimes can be quite divergent [10, 11].

The *Fréchet Audio Distance* (FAD) metric was recently introduced [12], based on a different approach. FAD is a distance measured between statistical distributions obtained from a set of audios to be evaluated and a reference set. Using FAD has one main potential advantage: it can be computed without having the original version of the particular audios to be evaluated. Instead, the reference becomes a set of audios that collectively exhibit the parameters that the audios to be evaluated should have. FAD thus appears as a natural choice of metric to automatic audio quality assessment in many contexts.

FAD has been used in audio evaluation in various contexts with different outcomes. In particular, FAD has been used: to compare processed audios with their original versions (e.g., [13, 10, 14, 11]) or with different audio sets (e.g., [15, 16]); together with a human listening test (e.g., [15, 16, 10, 11]) or without subjective evaluation (e.g., [13]); as a sole metric to evaluate audio quality or associated with other metrics (e.g., [15, 16, 10, 14]); among others. However, the effectiveness of FAD and especially its correlation to human perception has varied substantially, ranging from very weak to strong. Further investigation is thus needed to examine to which extent and how FAD can be used as a metric to automatic audio quality assessment.

FAD is dependent on the reference set used and the model from which *embeddings* are extracted. In this paper, we investigate how the choice of these sets and models might affect the results achieved with FAD. This issue is little discussed in the literature (to our knowledge, only in [17]). We first describe an overview of the usefulness of FAD as an objective metric, as reported in the literature. We then describe results of experiments to study FAD in the specific case of bass audios. Most of the results in previous work are based on datasets composed of different types of instruments, music styles and audio quality. Although this wide range of audio types aim at generalising the results, the great variability of parameters makes the analysis more difficult. Studying FAD with a restricted case allows us to compare the suitability of apparently similar types of datasets, but which yield different FAD scores. We evaluate FAD sensitivity

*This work was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) - Grant Number TIC 0009/2015.

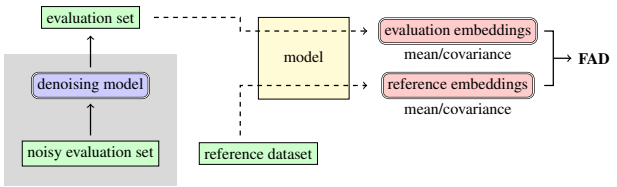

Figure 1: FAD

to effects and noises and as a metric to assess the quality of audio generated from examples of denoising processes. Finally, we compare the obtained FAD scores with the results of a subjective evaluation. Differently from existing work, this subjective evaluation was performed by professional or semi-professional musicians, all of them with experience in professional recordings.

This paper is structured as follows. In Sect. 2 we describe FAD. In Sect. 3 we discuss related work and present an assessment on the usefulness of FAD. In Sect. 4 we list the tools we have used to build our experiments. In Sect. 5 we describe the datasets we used and the developed denoising models. In Sect. 6 we present how FAD correlates to increasing levels of effects and noises. In Sect. 7 we evaluate FAD using denoising processes as examples and correlate the results with human evaluation. In Sect. 8 we further discuss some specific aspects of the experiments. Finally, Sect. 9 concludes the paper.

2. Fréchet Audio Distance (FAD)

FAD is based on the Fréchet Inception Distance (FID), introduced to evaluate generative models for images [18]. Fig. 1 illustrates how FAD is computed. Statistics are extracted from a reference set of audios with a certain property (*reference dataset*) and from a given set of audios to be evaluated (*evaluation set*). These audios can be, for example, the result of some denoising process applied to a set of original noisy audios (gray area in Fig. 1).

The statistics (mean and covariance) are computed from *embeddings* of datasets extracted from a *model*. In [12], embeddings were extracted from the last layer of the VGGish network [19], used to classify video soundtracks. Other embeddings might be used, as in [17] and in this work.

FAD is basically the Fréchet distance [20] computed over the statistics obtained from the evaluation and reference embeddings. These statistics are expected to reflect suitable properties of the datasets. A proper notion of distance between them should reflect how much of the properties existing in the reference set can be perceived in the set of audios to be evaluated.

Specifically, multivariate Gaussians are computed from the evaluation set embeddings $\mathcal{N}_e(\mu_e, \Sigma_e)$ and the reference embeddings $\mathcal{N}_b(\mu_b, \Sigma_b)$. From [20], it is known that the Fréchet distance between two Gaussians $\mathbf{F}(\mathcal{N}_b, \mathcal{N}_e)$ can be computed as:

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

where tr is the trace of a matrix. As a distance notion, the lower the FAD value, the closer the evaluated and reference datasets are. FAD values are usually referred to as *FAD scores*.

3. Related Work

Since its introduction, FAD has gained considerable momentum. In this section we discuss representative works where FAD was used in the contexts of audio synthesis [13, 15, 16, 21, 22, 23, 24, 25, 26] and audio enhancement [14, 10, 27], two of the main areas of digital audio processing, as well as those works whose goal was to specifically evaluate objective metrics [12, 11, 17]. Our main focus is on discussing in which ways and how effectively FAD has been used as an objective metric to audio evaluation.

Audio synthesis. In [24, 13], the authors used FAD as a metric to evaluate resynthesis of sounds of trumpet, violin and flute. They relied solely on FAD to evaluate their results, i.e. which instrument sound could be best resynthesized and with which parameter values. No correlation to human perception was done.

In [15] the authors used FAD as well as other objective and subjective metrics to evaluate the performance of three generative adversarial network (GAN) models to generate loops. The evaluation based on FAD varied depending on the reference set used and yielded a negative correlation to human perception. Other metrics were used (Inception Score and Diversity Measurement) to reflect different aspects of the datasets other than the statistics used in FAD. FAD was also used to evaluate the similarity between generated audios and the original ones in [25] (various instruments) and [26] (drum sounds).

In [21] the authors evaluated audio generated by four networks: DiffWave [28], DDSP [29] and NSynth [30], using objective and subjective metrics. The authors concluded that subjective listening evaluation in not well reflected in the objective metrics and new metrics of audio quality for automatically generated audios have to be further developed.

In [16] the authors used FAD to evaluate the quality of a textually guided audio generation tool. They considered that FAD correlates well with human perception in terms of audio quality, but they used another metric (KL-Divergence) to evaluate the relationship between the audios and the texts. FAD is also used, together with other metrics, to evaluate audios generated from text descriptions in [22] and [23]. In [23] FAD is considered to correlate with human subjective evaluation in some cases (worst cases) but not in others (best cases).

Audio enhancement. Audio quality enhancement in general is less addressed in the literature, where the main focus has been on the specific case of speech enhancement.

In [14], distortion was added to guitar recordings and a denoising process evaluated using FAD. The denoising process was evaluated with different MSS tools. FAD was computed using the original clean sounds as the reference set and VGGish embeddings. The FAD values varied, showing the best results for different tools depending on the type of effect (overdrive and clipping). The authors conclude that different metrics reflect specific aspects of audio improvement and should be further investigated.

In [10] the authors described an approach to enhance amateur quality recordings. The approach was based on processing mel-spectrograms as pictures. The results were evaluated with a human listening test as well as by objective metrics, including FAD. None of the used objective metrics correlated strongly with human evaluation. The metrics were not sensible to artifacts introduced during the process of converting from spectrograms to waveforms. FAD, however, was sensitive to additive noise.

In [27], the CQT-Diff audio enhancement framework was presented, that improves audios without knowledge of the types of degradation to which the original audios were subjected. The types of degradation evaluated were: bandwidth extension, audio inpainting and declipping. FAD and other objective metrics were used to compare the results with other related tools.

Analysis of FAD. In [12] the authors introduced FAD and evaluate its correlation with a set of audio distortions, including Gaussian noise, frequency filters, quantization, Griffin-Lim distortions and variations in speed. They computed FAD using part of the audios from MagnaTagATune [31] as the reference set and VGGish as the model. The paper reported a correlation between FAD and all the types of distortion. FAD correlated better than SDR with human evaluation, although with outliers for some of the distortions. Neither FAD nor SDR, however, yielded a strong correlation to human evaluation.

In [11], the correlation between human evaluation and the objective metrics FAD, L1, L2, SDR and SI-SDR was investigated, based on the perception of how good MSS tools can generate stems. According to the study, it is not possible to define such a correlation. Different conclusions can be drawn about how effective the tools are depending on which metric is chosen.

To the best of our knowledge, [17] is the only previous work that has a similar goal as ours, i.e. to study the effects on FAD of different reference sets and models. In [17], however, the impact of choosing different sample sizes and the relevance of evaluating FAD scores of individual audios were also investigated. The authors introduced two new reference sets, FMA-Pop and MusCC, as commonly used sets in the literature, such as MusicCaps, contain audios that are of low-quality and that are not music. FMA-Pop contains audios of a large genre diversity. MusCC contains studio-quality professionally mixed music. Embeddings from the following models were studied in the paper: VGGish (audio classifier), CLAP [32] and LAION CLAP [33] (joint audio-text embeddings), MERT [34] (self-supervised audio embedding), CDPAM [35] (audio similarity), EnCodec [36] and DAC [37] (low-rate audio codecs). The paper shows that different models yielded reasonably different results in terms of sensitivity to effects (distortion, filters, reverb, among others) and assessment of audio and music quality. Additionally, correlation between FAD scores and subjective evaluation occurred depending on the specific choice of reference set and model.

In summary, FAD has been so far used in very different scenarios. Correlation to human evaluation has been divergent, ranging from very weak to strong, and is dependent on specific choices of reference set and model. FAD has been considered to be sensitive to some types of noises but not to artifacts generated during audio processing.

We analyse FAD for a specific case (bass stems) and with more homogeneous reference sets. This provides a closer analysis of FAD than those in previous works. We additionally highlight the importance of considering relative instead of absolute FAD scores.

4. Used Tools

We used Open-Unmix to create denoising models. Open-Unmix [4] is an open-source MSS tool, developed by Inria Montpellier and Sony Corporation. It is based on PyTorch and the recurrent bidirectional Long-Short Term Memory (BLSTM) network architecture described in [38]. We used the default configuration, with three BLSTM layers.

We used *Sound eXchange* (SoX) v14.4.2 [39] to add effects and noise to the audio files and to normalize them. SoX is a cross-platform command line utility for audio processing. We additionally used Librosa v0.10.2.post1 [40] to resample and analyse spectrograms of audios.

We used the Google FAD implementation[1] to compute FAD scores with VGGish and Microsoft fadtk [17][2] for other embeddings (specifically, CLAP-2023 [32], CLAP-laion-audio [33] and CDPAM-acoustic) [35].

5. Reference Sets and Denoising Models

5.1. Reference Sets

Our experiments used four reference sets: one containing pure senoids (*RS-senoids*); one containing mixed songs with more than one instrument (*RS-songs*); and two containing only bass audios (*RS-bass-notes* and *RS-bass-lines*). We used an additional set of clean bass tracks to which effects and noises were applied (*RS-clean-audios*). These sets are as follows:

1. *RS-senoids:* it consists of 108 audio files, each containing six seconds of a pure senoid corresponding to the fundamental frequency of a musical note, ranging from the C0 (C note in the 0th octave) to B9 (B note in the 9th octave).
2. *RS-songs:* it consists of 9403 audios from the MagnaTagATune dataset [31]. MagnaTagATune contains 29-second excerpts of songs from a broad range of genres (Classical, Pop, Rock, among others).
3. *RS-bass-notes:* this is part of the IDMT-SMT-Bass dataset[3] [41], created at the Fraunhofer Institute for Digital Media Technology (IDMT). It consists of 2964 audio files containing each two seconds of a single note, played on three different 4-string basses. The dataset contains notes corresponding to the first up to the 13th fret on all strings, with different expression and plucking styles. We used only the audios corresponding to different expression styles.
4. *RS-bass-lines*: this dataset consists of the bass tracks of the 94 songs of the train set of MUSDB [42]. MUSDB is the standard dataset used to evaluate MSS tools and techniques, used in evaluation campaigns, such as the Signal Separation Evaluation Campaign (SiSEC) [1]. Each audio file contains seven seconds of the bass parts of songs from different styles.
5. *RS-clean-audios:* it consists of the bass tracks of 25 songs of the test set of MUSDB. The whole test set contains 50 songs.

All files contain studio quality audios at 16k sampling rate.

5.2. Denoising Models

Following the approach presented in [12], we used Open-Unmix to create two models to denoise audios: *DIST-MODEL*, which denoises audios with different levels of distortion; and *WNOISE-MODEL*, which denoises audios containing white noise. These models were only developed to serve as examples of denoising processes to be evaluated using FAD (and, thus, are not meant to be the most effective ones). In particular, we use them to investigate correlation between FAD scores and human evaluation. The models were created as follows:

[1] Available at: `https://github.com/google-research/google-research/tree/master/frechet_audio_distance`

[2] Available at: `https://github.com/microsoft/fadtk`

[3] Available at: `https://www.idmt.fraunhofer.de/en/publications/datasets/bass.html`

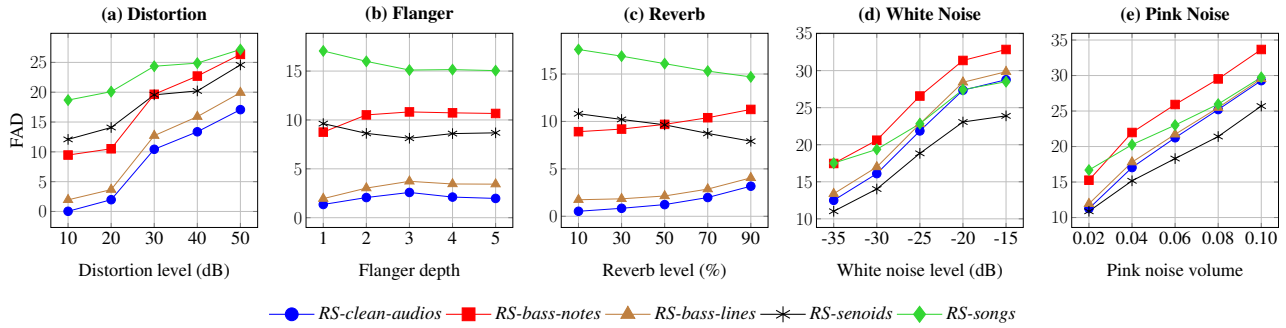

Figure 2: FAD sensitivity to different types of effects and noise. FAD scores computed using VGGish. See Table 1 for a description of the effect levels

1. *DIST-MODEL*: each of the bass tracks of the original 94 train songs of MUSDB was used to generate five new audios, with the following distortion levels: 10 dB, 20 dB, 30 dB, 40 dB and 50 dB. These audios were used as the train set (470 audios). The same distortion levels were applied to 25 of the original MUSDB test audios, to be used as the test set (125 audios). These audios are the ones that are not in *RS-clean-audios*.
2. *WNOISE-MODEL*: a similar procedure as for the DIST-MODEL was used, but now, instead of distortion, we added five leves of white noise: -15 dB, -20 dB, -25 dB, -30 dB and -35 dB. As for *DIST-MODEL*, we generated sets with respectively 470 and 125 audios with white noise to be used as the train and test sets.

We trained the models using a GeForce RTX3080 Galax GPU with the aligned training option of Open-Unmix, which allows us to define an input and the corresponding output file. The input file was the audio with either distortion or white noise. The output file in each case was the corresponding original bass track. The training was configured to execute 1000 epochs with early stop. We experimented with a batch size of 4 and 16. As the FAD scores were close in most cases for both batch sizes, we present in the following the results for the models generated with batch size 4. The volume of all audios was normalized to -0.1 dB.

6. FAD Sensitivity to Effects and Noise

In our first experiment, we analysed the FAD sensitivity to the following effects and noises: distortion, flanger, reverb, white noise and pink noise. The scores in this section were computed using VGGish embeddings.

We generated five audio sets by applying five increasing levels of effects/noises to the 25 audios of *RS-clean-audios*. Thus, each audio set corresponding to a given effect/noise contained $25 \times 5 = 125$ audios. The levels of effect/noise used in the sets are shown in Table 1.

Sensitivity to these types of effects has been previously evaluated [12, 17]. However, we address here the much more restricted case of bass stems to evaluate FAD in scenarios with less variability. This experiment also aimed at verifying whether any of the chosen datasets approximates *RS-clean-audios* and, thus, could be used as a reference for bass tracks when the original versions of audios are not available.

Sensitivity to effects: Figs. 2a, 2b and 2c show the FAD scores for distortion, flanger and reverb, respectively. The scores increase with the increase in the level of distortion, showing a correlation to this effect (Fig. 2a). There was in general a slight increase in the FAD scores with an increase in the intensity of flanger (Fig. 2b) and reverb (Fig. 2c) for *RS-clean-audios*, *RS-bass-notes* and *RS-bass-lines*, but not always. For *RS-senoids* and *RS-songs* there was a negative correlation. These results are consistent with [17, 12].

Effect	Levels
Distortion (dB)	10, 20, 30, 40, 50
Flanger (depth)	1, 2, 3, 4, 5
Reverb (%)	10, 30, 50, 70, 90
White noise (dB)	-15, -20, -25, -30, -35
Pink noise (amplitude ratio)	0.02, 0.04, 0.06, 0.08, 0.10

Table 1: The reverb and flanger levels refer to the parameters of the corresponding sox effects. The pink noise level refers to the *sox synth pinknoise vol* parameter.

The lowest scores occurred for the reference set with the original audios (*RS-clean-audios*), as expected. The FAD scores for *RS-bass-lines* were very close to those of *RS-clean-audios* in all cases. Additionally, in general, except for the case of *RS-songs* and *RS-senoids* and the effects flanger and reverb, the FAD scores varied roughly similarly with the increase in the intensity of the effect/noise for different reference sets. For audios with lower levels of distortion (10 dB and 20 dB), flanger (depth 1) and reverb (levels 10% and 30%), the smallest FAD scores were for the sets containing only bass sounds. In these cases, therefore, FAD was not only sensitive to the level of effect but also to the specific type of sound (bass sound).

Sensitivity to noise: Figs. 2d and 2e show FAD scores for audios with increasing levels of white and pink noise, respectively. An increase in the level of noise led to an increase in the FAD scores, much more regularly than for the case of effects (Figs. 2a-c).

For noises, the lowest FAD scores were obtained for *RS-senoids*. Surprisingly, these scores were lower even than for *RS-clean-audios*, i.e., the original audios. For effects (Figs. 2a-c) the FAD scores were lower when computed with datasets containing only bass sounds (*RS-bass-lines*). Our interpretation is that the effects generate variations in the signal that are around the original bass frequencies, whereas the white and pink noise generate signals in frequencies in a much broader range.

In particular, the results show us that the FAD scores computed for *RS-bass-lines* were always close or very close to *RS-clean-audios*, for both effects and noises.

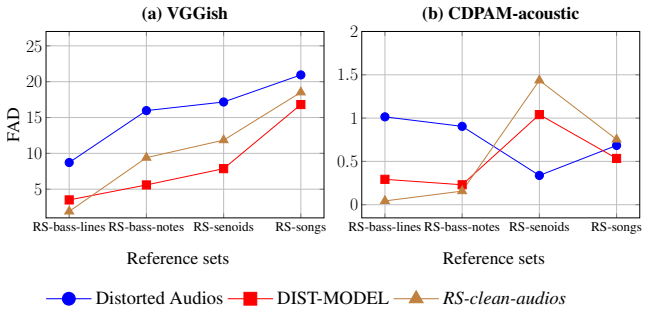

Figure 3: FAD scores (distortion), using VGGish (a) and CDPAM-acoustic (b)

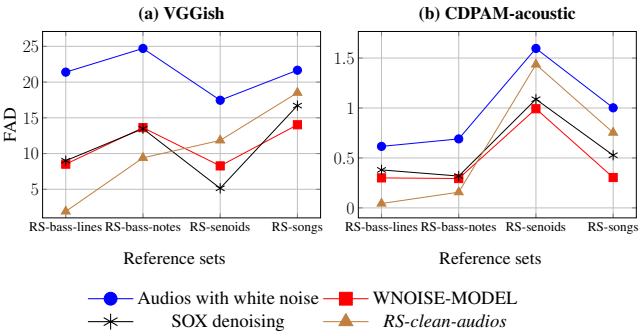

Figure 4: FAD scores (white noise), using VGGish (a) and CDPAM-acoustic (b)

7. Using FAD to Evaluate Denoising

In the second experiment, we studied FAD as a metric to evaluate an automatic audio improvement process. We used denoising as an example, using *DIST-MODEL* and *WNOISE-MODEL*.

We computed FAD for denoised audios using different reference sets and two embeddings: VGGish and CDPAM-acoustic [35]. VGGish was originally used in [12], the paper that introduced FAD. In Sect. 7.2, we report on results we obtained with some of the embeddings supported by fadtk, specifically: CLAP-2023 [32], CLAP-laion-audio [33] and CDPAM-acoustic. We chose these embeddings as they were not developed for the specific case of speech. Additionally, embeddings based on CLAP yielded the highest correlation to human evaluation in [17]. In this section, we focus on the cases of VGGish and CDPAM-acoustic, as the behaviour of FAD for these embeddings is important for the discussion in Sect. 8.

7.1. FAD for Denoised Audios, Different Reference Sets and Embeddings

The audios with distortion were processed using *DIST-MODEL*. Fig. 3 shows the FAD scores for the distorted audios (blue circles), the generated denoised audios (red squares) and the original clean audios (brown triangles), for *RS-bass-lines*, *RS-bass-notes*, *RS-senoids* and *RS-songs* using VGGish (Fig. 3a) and CDPAM-acoustic (Fig. 3b).

For distortion and VGGish (Fig. 3a) the FAD scores were lower for the denoised audios using DIST-MODEL (red) than for the distorted audios (blue) for all reference datasets. That in principle reflects an audio improvement. However, for the case of CDPAM-acoustic (Fig. 3b) the FAD scores for distorted audios (blue) were lower than even the original clean audios for *RS-senoids* and *RS-songs*. For both VGGish and CDPAM-acoustic, the FAD scores for *RS-clean-audios* are the lowest ones for *RS-bass-lines*. Additionally, for this reference set, the relation between the scores for *RS-clean-audio*, denoised and distorted audios were as expected, i.e., the lowest score for *RS-clean-audios*, the highest score for the distorted audios, and an intermediate score for denoised audios. However, *RS-clean-audios* was not always the set with the lowest scores. In Fig. 3a, for example, the sets resulted from DIST-MODEL were closer (lower scores) to *RS-bass-notes*, *RS-senoids* and *RS-songs* than *RS-clean-audios*.

The audios with white noise were processed using *WNOISE-MODEL* and the SOX `noisered` function [39]. This function eliminates noise from an audio according to a noise profile, obtained by using the SOX function `noiseprof`. We used `noisered` with parameter 0.1 and equalized the audios to the 10-800 Hz range. Fig. 4 shows the FAD scores for the noisy audios (blue circles), the generated denoised audios using WNOISE-MODEL (red squares), denoised audios using SOX (black stars) and the original clean audios (brown triangles), for *RS-bass-lines*, *RS-bass-notes*, *RS-senoids* and *RS-songs* using VGGish (Fig. 4a) and CDPAM-acoustic (Fig. 4b).

For white noise, the FAD scores for denoised (red squares and black stars) and clean audios (brown triangles) were lower than the scores for the noisy audios (blue circles) in all cases, as expected (Fig. 4a-b). There was in general a correlation between the scores for the original noisy, denoised (for both WNOISE-MODEL and SOX) and clean audios, except for the case of *RS-clean-audios* and *RS-senoids* in Fig. 4a.

Similarly to distortion, the FAD scores for denoised audios are lower than those for *RS-clean-audios* for the cases of *RS-senoids* and *RS-songs*. The scores for *RS-clean-audios* are low for *RS-bass-lines* and the relation between the scores for clean, denoised and noisy audios for *RS-bass-lines* are as expected. The FAD scores for DIST-MODEL and SOX were very similar in all cases.

As before (Sect. 6), *RS-bass-lines* exhibited the closest distances to the original clean audios, according to FAD.

7.2. Closer Look at the FAD Scores for *RS-bass-lines*

Fig. 3 and 4 presented the FAD scores for the set of denoised audios considering the different levels of distortion and noise (in each case) together. Figs. 5 and 6 show the FAD scores for the denoised audios for each level of distortion and white noise separately, thus presenting a closer look at the scores for the denoising process. Once more, we present the scores for VGGish and CDPAM-acoustic. In these figures, however, we only use *RS-bass-lines* as the reference set, as the FAD scores computed using this reference set were close to those computed using *RS-clean-audios*, as previously discussed (Sects. 6 and 7.1). Recall that one of the advantages of using FAD is that there is no need to have the specific original clean versions of audios to be processed. In this case, we in particular evaluate how *RS-bass-lines* could be used as a reference set for high quality bass audios.

The FAD scores for the denoised audios (red squares) are in general very low in comparison to the distorted audios (blue circles), except for the case of 10 dB and 20 dB and VGGish (Fig. 5a). These graphics lead to the conclusion that, according to FAD, the denoised process yielded good results, as the scores for the denoised audios represent a small distance to *RS-bass-lines*. Regarding white noise, Fig. 6 shows that WNOISE-MODEL (red squares) and SOX (black stars) have generated audios which yielded close FAD scores. Again, FAD leads to the conclusion

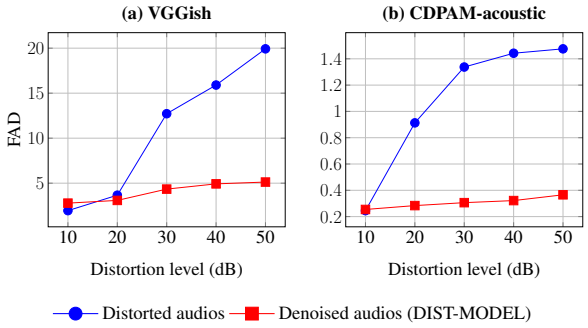

Figure 5: FAD scores when removing distortion with DIST-MODEL, using *RS-bass-lines* and VGGish (a) and CDPAM-acoustic (b)

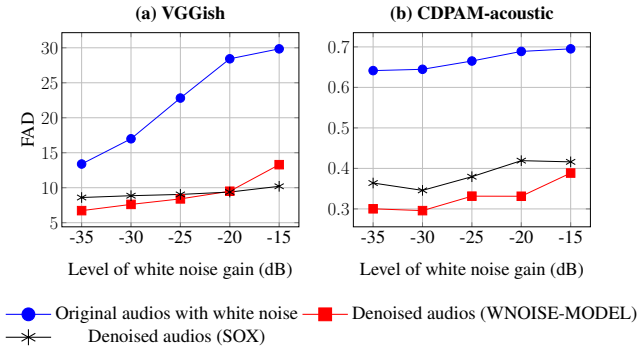

Figure 6: FAD scores for audios denoised using WNOISE_MODEL and SOX, with *RS-bass-lines*, VGGish (a) and CDPAM-acoustic (b)

that both denoising processes might have provided good results, as the distances to *RS-bass-lines* are small.

Fig. 7, however, shows normalized FAD scores for each level of distortion (Fig. 7a) and white noise (Fig. 7b), for embeddings from different models: in addition to VGGish and CDPAM-acoustic, we also show the scores for CLAP-2023 and CLAP-laion-audio. The values were normalized as follows: each score in the figure is the FAD score for the corresponding level of distortion and model divided by the FAD score of *RS-clean-audio* for that model. The normalized scores thus reflect how far they are from the FAD scores of the original clean audios using the respective model. The reference dataset is always *RS-bass-lines*.

For distortion (Fig. 7a) there is a slight increase in the normalized FAD scores with the increase in the distortion level, but higher scores were exhibited with CDPAM-acoustic. For white noise (Fig. 7b), the relative increase was higher for CDPAM-acoustic and VGGish. The scores obtained with CLAP-laion-audio and CLAP-2023 were very close and similar to those for VGGish and distortion (Fig. 7a). The FAD scores were higher, when CDPAM-acoustic was used (Fig. 7a-b).

7.3. Subjective Evaluation

We conducted a subjective evaluation of the denoising process to verify whether there was a correlation between human perception and FAD scores. Eight professional or semi-professional musicians with experience with studio recordings were interviewed. Each of them were asked five questions. Each question had two audios: one with a denoised audio, and another with the corresponding clean audio. The denoised audio was the result of using either DIST-MODEL or WNOISE-MODEL for one of the five corresponding levels of distortion or noise. After hearing both audios, each person was asked to evaluate the denoised audio according to audio quality, using a Likert scale from 1 to 5, where 1 and 5 indicated that the audio had quality that was, respectively, distant and close to the quality of the corresponding clean audio. Intermediate values expressed intermediate levels of audio quality. Each question contained an audio randomly chosen from the set of denoised audios.

The Mean Opinion Score (MOS) was 1.6 ± 0.3 and 2.6 ± 0.8, for distortion and white noise, respectively, for a 95% confidence interval. Although the interview contained a small number of responses, 29 for distortion and 11 for white noise, it expressed that in general the quality of the denoised audios was distant from the original clean audios for distorted audios. For white noise, the average value was almost the exact middle point between the extremes, but with a high variation.

8. Discussion

Specificity of reference sets. Using FAD demands a very specific choice of parameters. In [17], the reference sets used (MusCC, FMA-Pop and MusicCaps) have different properties. While MusCC has studio-quality audios, MusicCaps and FMA-Pop contain, respectively, audios of less quality and from a large genre diversity. The reference sets used in our experiments contain the main intended property for the target audios (high quality) and thus they could all be considered equivalent in this regard. Additionally, *RS-bass-notes* and *RS-bass-lines* should be equivalent for FAD calculation, as both contain only bass sounds. However, *RS-bass-lines* was the only dataset that exhibited a short distance to the original clean audios (Sects. 6 and 7). These examples highlight that a reference set must be chosen very specifically. According to our experiments, *RS-bass-lines* is a candidate as reference set to evaluate audio quality of bass stems.

Interpretation of absolute and relative FAD scores. In general, the FAD scores presented in Figs. 5 and 6 for the denoised audios are low or very low in comparison to those for the distorted and noisy ones in all cases. The FAD sensitivity to distortion and noise, as indicated in Sect. 6 and in [12, 14, 17], leads us to the interpretation that the denoising processes provided good results, i.e., removed much of the distortion and noise. However, the outcome of the human listening test yielded a low and an average MOS for distortion and white noise, respectively, indicating distance from the original clean audios. Considering the subjective evaluation as a ground truth and the fact that the interviewed people have experience with professional recordings, we conclude that the generated denoised audios were close to the original audios statistically (low FAD scores), but they lack quality, probably by containing artifacts introduced during the denoising process, to which FAD was not sensitive.

In relative terms, the normalized scores presented in Fig. 7 show that, for the case of distortion, the FAD scores increased with an increase in the distortion level. The scores were close for the different models, except for the case of CDPAM-acoustic (Fig. 7a). In the case of white noise, the increase in the normalized FAD scores was moderate to large for VGGish and CDPAM-acoustic (Fig. 7b).

We computed the Root-Mean-Square (RMS) values for each of the songs in *RS-clean-audios* for the case of denoising from distortion, as calculated by *Librosa* (not shown here due to space limitation). In general, the RMS values are roughly similar. This could be interpreted as an indication (although rough) that

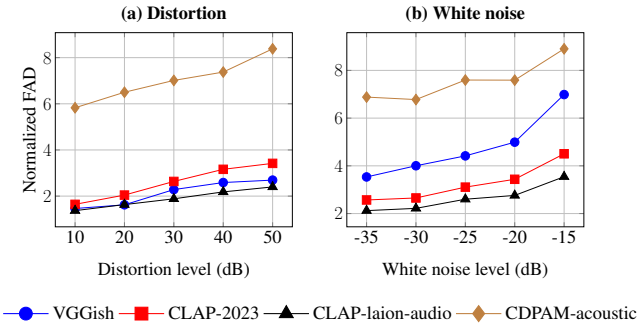

Figure 7: Normalized FAD scores when removing distortion (a) and white noise (b) with DIST_MODEL and WNOISE_MODEL, using *RS-bass-lines* and different models

the resulting denoised audios are somewhat similar, what is consistent with the approximate FAD scores shown in Figs. 5 and 6 (red squares and black stars) and indicates that the models were not sensitive to artifacts introduced in the denoising process.

These results suggest that FAD scores might be more easily interpretable as relative than as absolute values. Additionally, in our experiments CDPAM-acoustic was the model that has best reflected the subjective evaluation.

9. Conclusion

This paper had two goals. First, to present the many ways FAD has been used, how effective its use has been and how it has correlated to human evaluation, as reported in the literature. Then, to investigate the effect on FAD scores when different reference sets and embeddings are used for a specific case, i.e., with more homogeneous audio sets than considered in previous works. With that, we intended to avoid natural sources of different audio features that could lead to a wider range of FAD scores, such as songs of diverse music styles.

Correlation between FAD scores and human evaluation has been reported variably, ranging from very weak to strong. Additionally, FAD is reportedly to be sensitive to some audio features, but to fail in some cases as a general objective metric to audio quality. In our experiments, FAD scores yielded correlation to different levels of distortion and pink and white noise, but were not sensitive to reverb and flanger. These results are consistent with [12] and [17].

We showed the importance of choosing a reference set very specifically. FAD scores were low when *RS-bass-lines* was used as a reference set. Using *RS-bass-notes* did not yield comparably low FAD scores, even though it also contains only bass sounds. We additionally showed that relative FAD values might be more easily interpretable than absolute ones. Among the studied models, CDPAM-acoustic provided the results that best reflected the subjective tests, carried out by people with experience with professional recordings.

As a distance measurement between statistical distributions, FAD surely can be used as a practical metric to evaluate audio quality. However, reference sets must be carefully designed and users must be aware of which combinations of reference sets and models are sensitive to the specific desired target audio properties to be evaluated. FAD has been more sensitive to some effects and less to others, like reverb and flanger. Thus, further investigation is needed to more thoroughly understand which set of metrics and parameters are needed to guide the development of automatic audio improvement tools.

References

[1] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, and K.-W. Cheuk. Music demixing challenge 2021. *Frontiers in Signal Processing*, 1, Jan 2022.

[2] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research.

[3] A. Défossez, N. Usunier, L. Bottou, and F. Bach. Music source separation in the waveform domain, 2021. arXiv:1911.13254.

[4] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-Unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 2019.

[5] A. Hines, J. Skoglund, A.C. Kokaram, and N. Harte. ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

[6] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001.

[7] J.G. Beerends, C. Schmidmer, J.Berger, M.Obermann, R. Ullmann, J.Pomy, and M. Keyhl. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384, june 2013.

[8] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[9] J.L. Roux, S. Wisdom, H. Erdogan, and J.R. Hershey. SDR - half-baked or well done?, 2018. arXiv:1811.02508.

[10] N. Kandpal, O. Nieto, and Z. Jin. Music enhancement via image translation and vocoding. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3124–3128, 2022.

[11] E. Rumbold. A critical analysis of objective evaluation metrics for music source separation audio quality. Master's thesis, Northwestern University - Computer Science Department, Aug. 2022. Master Thesis - Northwestern University - Computer Science Department - Technical Report NU-CS-2022-11.

[12] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. In *Proceedings of INTERSPEECH*, Graz, Austria, Sept. 2019.

[13] F. Caspe, A. McPherson, and M. Sandler. DDX7: Differentiable FM synthesis of musical instrument sounds. In *Proc. of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[14] J. Imort, G. Fabbro, M.A. M. Ramírez, S. Uhlich, Y. Koyama, and Y. Mitsufuji. Distortion audio effects: Learning how to recover the clean signal, 2022. arXiv:2202.01664.

[15] T.-M. Hung, B.-Y. Chen, Y.-T. Yeh, and Y.-H. Yang. A benchmarking initiative for audio-domain music generation using the freesound loop dataset, 2022. arXiv:2108.01576.

[16] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi. AudioGen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335, 2024.

[18] M. Heusel, R. Hubert, U. Thomas, and N. Bernhard. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[19] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[20] D.C. Dowson and B.V Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[21] A. Vinay and A. Lerch. Evaluating generative audio systems and their metrics. In *Proc. of the 23rd Int. Society for Music Information Retrieval Conf. (ISMIR)*, Bengaluru, India, 2022.

[22] A. Agostinelli, T.I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. MusicLM: Generating music from text, 2023. arXiv:2301.11325.

[23] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez. Simple and controllable music generation. In *Proc. of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[24] B. Hayes, C. Saitis, and G. Fazekas. Neural waveshaping synthesis. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021.

[25] J. Nistal, S. Lattner, and G. Richard. DarkGAN: Exploiting knowledge distillation for comprehensible audio synthesis with gans. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021.

[26] S. Rouard and G. Hadjeres. CRASH: Raw audio score-based generative modeling for controllable high-resolution drum sound synthesis. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2021.

[27] E. Moliner, J. Lehtinen, and V. Välimäki. Solving audio inverse problems with a diffusion model. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[28] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *Proc. of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[29] J. Engel, L. Hantrakul, C. Gu, and A. Roberts. DDSP: Differentiable digital signal processing. In *Proc. of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.

[30] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proc. of the 34th International Conference on Machine Learning (PLMR)*, 2017.

[31] E. Law, K. West, M. Mandel, M. Bay, and J.S. Downie. Evaluation of algorithms using games: the case of music annotation. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009.

[32] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang. Clap learning audio concepts from natural language supervision. In *Proc. of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[33] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[34] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu. MERT: acoustic music understanding model with large-scale self-supervised training, 2024. arXiv:2306.00107.

[35] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP 2021*, jun 2021.

[36] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression, 2022. arXiv:2210.13438.

[37] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar. High-fidelity audio compression with improved rvqgan, 2023. arXiv:2306.06546.

[38] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–265, 2017.

[39] C. Bagwell et al. Sound eXchange (SoX), 2015. Available at: `http://sox.sourceforge.net`.

[40] B. McFee, M. McVicar, D. Faronbi, I. Roman, et al. librosa/librosa: 0.10.2.post1, 2024. Available at: https://doi.org/10.5281/zenodo.11192913.

[41] J. Abeßer, H. Lukashevich, and G. Schuller. Feature-based extraction of plucking and expression styles of the electric bass guitar. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2290–2293, 2010.

[42] Z. Rafii, A. Liutkus, F.-R. Stöter, S.I. Mimilakis, and R. Bittner. The MUSDB18 corpus for music separation, December 2017. Available at https://doi.org/10.5281/zenodo.1117372.