XXX ENCONTRO DA SOBRAC 2023
SOCIEDADE BRASILEIRA DE ACÚSTICA
19 a 22 de novembro de 2023
Natal - RN

# AIRCADE: an Anechoic and IR Convolution-based Auralization Data-compilation Ensemble

**Chiodi, T.[1]; dos Santos, A.N.[2]; Martins, P.[2]; Masiero, B.S.[2]**

[1] Engenharia Acústica, Universidade Federal de Santa Maria, Santa Maria, RS, Brasil, tulio.chiodi@eac.ufsm.br
[2] Laboratório de Comunicações Acústicas, Universidade Estadual de Campinas, Campinas, SP, Brasil, masiero@unicamp.br

## Resumo

Neste artigo, um novo conjunto de dados compilados é introduzido, com o intuito principal de servir de recurso para pesquisadores no campo de abordagens algorítmicas de desreverberação orientadas por dados. AIRCADE compreende amostras de fala e canções, juntamente com sons de violão, com anotações originais pertinentes ao Reconhecimento de Emoções e a Recuperação de Informações Musicais. Além disso, também está inclusa uma seleção de amostras de Resposta ao Impulso com valores de Tempo de Reverberação variados, fornecendo uma grande variedade de condições para avaliação. Essa compilação de dados pode ser utilizada juntamente com códigos em Python fornecidos, para geração de conjuntos de dados aurilizados em diferentes tamanhos: *minúsculo*, *pequeno*, *médio* e *grande*. Adicionalmente, os metadados de anotação fornecidos permitem a análise e investigação mais aprofundadas do desempenho de algoritmos de desreverberação sob diferentes condições. Os resultados são validados através do treinamento de um modelo recentemente proposto de Aprendizado Profundo para desreverberação utilizando o *corpus* de dados obtido, seguido da avaliação objetiva do seu desempenho usando métricas de qualidade de fala, inteligibilidade e pureza de sinal. Todos os dados estão sob Licença *Creative Commons* Atribuição 4.0 Internacional).

**Palavras-chave:** aurilização, conjunto de dados, desreverberação, aprendizado de máquina, aprendizado profundo.

## Abstract

In this paper, a new data-compilation ensemble is introduced, primarily intended to serve as a resource for researchers in the field of data-driven dereverberation algorithmic approaches. AIRCADE comprises speech and song samples, together with acoustic guitar sounds, with original annotations pertinent to Emotion Recognition and Music Information Retrieval. Moreover, it includes a selection of Impulse Response samples with varying Reverberation Time values, providing a wide range of conditions for evaluation. This data-compilation can be used together with provided Python scripts, for generating auralized data ensembles in different sizes: *tiny*, *small*, *medium* and *large*. In addition, the provided metadata annotations allow for further analysis and investigation of the performance of dereverberation algorithms under different conditions. Results were validated by training a recently proposed Deep Learning dereverberation model using the obtained data *corpus*, followed by the objective evaluation of the model's performance using speech quality, intelligibility and signal purity metrics. All data is under Creative Commons Attribution 4.0 International License.

**Keywords:** auralization, dataset, dereverberation, machine learning, deep learning.

## 1. Introduction

Reverberation is the persistence of sound in a space after a source ceases to emit it. This is mainly caused by the reflections of the sound waves off the surfaces within the space, which gradually dissipate over time. In this scenario, the Reverberation Time (RT) is considered an important aspect of room acoustics because the characteristics of the space in which reverberation occurs can influence the duration and intensity of the phenomenon [1, 2].

In the context of entertainment, reverberation can positively affect the Human Auditory System (HAS), as it can help to enhance the perceived loudness and richness of sounds. This is because the multiple reflections can create a sensation of spaciousness and immersion that is aesthetically pleasing, particularly in music. However, in the context of communication, excessive reverberation can affect the quality and intelligibility of speech, because the multiple reflections can create a "smearing" effect, which can make it difficult to distinguish individual sounds and syllables. Moreover, a reduction in the overall Signal-to-Noise Ratio (SNR) of a given sound source can mask quiet sounds, making it harder to understand them [3, 4].

Therefore, the effects of reverberation on the HAS can vary widely, depending on the duration and intensity of the phenomenon, the frequency content of the sound source, the individual characteristics of the listener's hearing etc. Generally, it is desirable to optimize the amount of reverberation in a given acoustic scenario to ensure the best possible listening experience for the intended audience. In this case, dereverberation can be used to reduce or eliminate the effects of excessive reverberation from an audio signal. This is typically used in situations where the audio signal has already been recorded in a reverberant environment and the resulting reverberation is unwanted or detrimental to the quality of the recorded audio [5].

In recent years, data-driven dereverberation methods have become increasingly popular owing to their ability to learn complex mappings between anechoic and reverberant signals. While showing promising results, these methods have a wide range of applications including Speech Recognition, Speaker Verification and Music Production [6, 7].

Still, an important challenge for these methods is the need for large amounts of training data with paired clean and reverberant audio signals for training, which ideally should cover a wide range of acoustic environments, microphone types and speaker characteristics, to ensure that a model will generalize well to new scenarios. Examples of dereverberation datasets include the REVERB challenge dataset [8], BUT Speech@FIT Reverb Database [9] and VoiceBank-SLR [10], among others.

Most of these datasets have limitations, such as narrowband anechoic data, i.e., only speech is considered a signal of interest, and the absence of Impulse Response (IR) data, i.e., usually only anechoic and reverberant data are paired for supervised training. Hence, in this study, a new data-compilation ensemble is introduced, primarily intended for training data-driven dereverberation models capable of dealing with full-bandwidth audio signals such as speech, song and instrumental music. Pairs of natural anechoic and IR data are offered, compiled from datasets licensed under Creative Commons Attribution 4.0 International License, together with Python scripts for convolution-based auralization, under the hypothesis that this ensemble could serve as better training and evaluation tool for such algorithms.

Results are validated by training a recently proposed Deep Learning (DL) dereverberation model using the obtained data *corpus*, followed by the objective evaluation of the model's performance using speech quality, intelligibility and signal purity metrics.

The remainder of this paper is organized as follows: Section 2 details the methodology used to select the anechoic and IR samples for synthesizing the auralized data. Section 3 outlines the resulting auralized data with annotations. Section 4 describes the DL dereverberation model used to validate the dataset and provides an objective evaluation of the model's performance. Section 5 presents an overall discussion of the results and Section 6 concludes this study.

## 2. Methods

To simulate or recreate an acoustic environment, such as a concert hall or recording studio, using computer algorithms and specialized software, one can resort to a well-known technique called auralization. This involves the measurement or simulation of the physical properties of a space to mimic its acoustic characteristics. Various techniques such as acoustic ray tracing and convolution-based auralization are available to proceed accordingly [11].

From a signal-processing perspective, convolution is a mathematical operation that describes the interactions between two signals. In the context of auralization, convolution can be used to simulate the effects of the acoustic environment on the audio signals. The basic concept involves convolving an audio signal with an IR signal that describes the acoustic characteristics of a room or space. The resulting output produces a new signal that represents the original audio signal after it is modified by the acoustic characteristics of the room [12, 13].

In the context of data-driven dereverberation methods, the choice of natural, synthetic or "joint" datasets — i.e., those obtained by processing combinations of natural recordings with synthetic sounds — directly impacts in-the-wild applicability. This is because, when models are trained mostly on synthetic data, they usually do not generalize well for real-world scenarios. However, most recent studies have used joint combinations — e.g., convolving anechoic data with naturally recorded or synthesized IR data — perhaps because it would be too cumbersome to naturally acquire all the necessary data [14]. Hence, this compilation attempts to achieve a balance between these features, compiling real anechoic and IR data and synthetically producing auralized data ensembles by means of convolution.

### 2.1 Anechoic data

To cover a wide range of full-bandwidth audio signals of interest, three categories were selected for this study: speech, song and instrumental music.

Since RAVDESS [15] is a well-known dataset with subsets of emotional speech and song, it was chosen to cover the first two aforementioned categories. Its speech-only portion comprises 1,440 samples performed by 24 actors (12 male and 12 female), vocalizing two lexically matched statements ("kids are talking by the door" and "dogs are sitting by the door") in a "neutral" North American accent. Each expression is pronounced at two different levels of emotional intensity (normal and strong) with an additional neutral expression, resulting in a total of 60 trials per actor. Speech emotions include calm, happiness, sadness, anger, fear, surprise and disgust.

Its song-only portion is quite similar and comprises 1,012 samples performed by 23 actors singing the same two lexically matched statements. Song emotions include neutral, calm, happiness, sadness, anger and fear. The original sample rate is fixed at 48 kHz, and Figures 1 (a) and (b) illustrate histograms with the original duration of files in each subset used.

Since the acoustic guitar is a popular musical instrument, for a variety of reasons (including its ability to produce polyphonic sound and its musical versatility), a subset from GuitarSet [16] was chosen for this study, here referred to as "audio_mono–mic". It comprises 360 samples performed by six musicians, playing 30 twelve- to sixteen-bar excerpts from lead sheets in a variety of keys, tempos and musical genres. The recording was performed using a Neumann U87 condenser microphone placed approximately 30 cm in front of the $18^{th}$ fret of the guitar. The original sample rate is fixed at 44.1 kHz, and Figure 1 (c) illustrates a histogram with the original duration of the files in this subset.

### 2.2 IR data

IR data was curated from the Open Acoustic Impulse Response (Open AIR) Library, which is a collaborative online database. Because the original metadata in this database provides information about space category, IR duration etc., the samples were chosen to achieve a balance between a selected variety of

open and enclosed spaces, with IRs in the range of a few milliseconds up to 5 s. Figure 1 (d) illustrates a histogram with the original duration of the selected IR data. In total, 65 IRs were selected for this study.

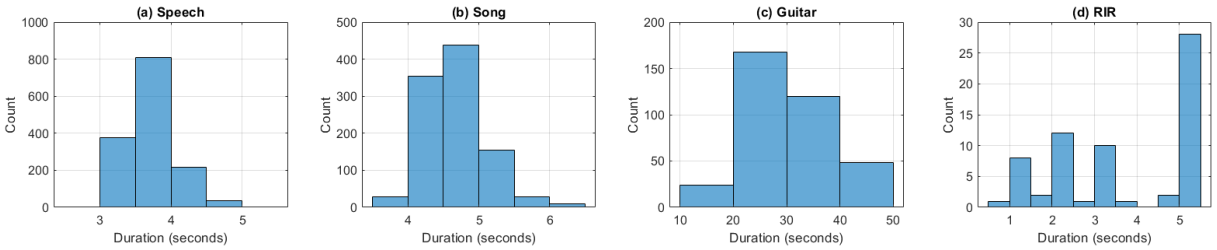


**Figure 1:** Original duration of files in the (a) speech, (b) song, (c) guitar and (d) IR subsets of this compilation.

## 2.3 Data processing

Considering the great difference between the duration of signals in RAVDESS and GuitarSet, the samples in GuitarSet were split into segments of smaller duration — fixed at 5 s — resulting in 2,004 new different samples. Another reason for this decision is that when choosing the length of the anechoic data, it is important to have a balance between the computational cost of the convolution operation and segment length. Moreover, since the sample rates of RAVDESS and GuitarSet are different, the GuitarSet segments were resampled to 48 kHz, thus standardizing this value for all files in this compilation.

The selected IR data was found in a variety of formats, such as B-format, MS Stereo and Mono, at sample rates varying from 44.1 kHz to 96 kHz. Because all anechoic data was compiled in Mono format at 48 kHz, the selected IR data was first converted to Mono and then normalized to prevent files from clipping. Finally, each IR was resampled to 48 kHz, as required.

## 2.4 Data annotation

Because the anecehoic data in this compilation comprises speech and song samples with original annotations pertinent to Emotion Recognition, together with acoustic guitar sounds with original annotations pertinent to Music Information Retrieval (MIR), it was decided to provide users with metadata that can be used to trace back each sample to its original annotations. Hence, this dataset is not limited to dereverberation tasks only, but can also be used in applications such as Emotion Recognition and MIR in more challenging scenarios, i.e., in the presence of convolutive noise.

Since the computational effort in dereverberation tasks is highly intertwined with the RT values of IR data, $RT_{20}$ values were extracted from each IR sample using ITA Toolbox. Figure 2 illustrates a histogram of the extracted $RT_{20}$ values from the selected IR data.

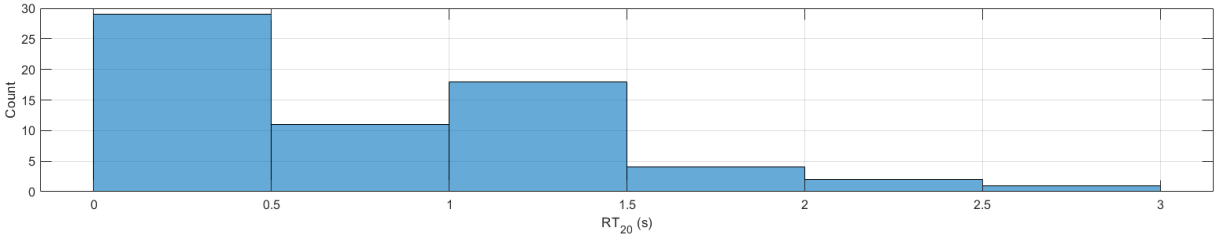


**Figure 2:** Extracted $RT_{20}$ values from the selected the IR data.

## 3. Results

The processed anechoic and IR data is hosted at Zenodo, with an approximate total file size of 1.3 GB. For simplicity, all samples in the compilation were renamed, e.g., guitar_XXXX, rir_XXXX, song_XXXX and speech_XXXX, in which XXXX can range from 0000 to 9999. To synthesize different versions of the auralized data ensemble, the reader is referred to GitHub, where Python scripts are available for

downloading the base data-compilation and then synthesizing a chosen version of the auralized data ensemble. Table 1 illustrates the differences between all versions, detailing the number of songs, speech, guitar, IR and auralized samples in each version, together with their respective total file size and duration.

**Table 1:** # of anechoic, IR and resultant auralized data samples, together with their respective total duration and file size for each ensemble version.

| Version | Tiny | Small | Medium | Large |
|---|---|---|---|---|
| **Song** | 100 samples | 500 samples | 1,012 samples | 1,012 samples |
| **Speech** | 100 samples | 500 samples | 1,012 samples | 1,440 samples |
| **Guitar** | 100 samples | 500 samples | 1,012 samples | 2,004 samples |
| **IR** | 5 samples | 9 samples | 33 samples | 65 samples |
| **Auralized** | 1,500 samples | 13,500 samples | 100,188 samples | 289,640 samples |
| **Total duration** | 3.2 h | 30.41 h | 221.77 h | 658.08 h |
| **Total file size** | 1.1 GB | 10.5 GB | 76.6 GB | 227.5 GB |

## 4. Validation

To validate the results presented Section 3, a recently proposed DL dereverberation model was chosen to be trained on different versions of AIRCADE, as following described.

### 4.1 Baseline model

Originally proposed by Ren *et al.* [17], the chosen DL dereverberation model combines a multi-channel U-Net with a neural beamformer. It takes multi-channel audio signals as input and uses the Short-Time Fourier Transform (STFT) as feature extraction method. This Time-Frequency (T-F) representation is subsequently encoded and decoded by the U-Net, enabling the learned parameters to dynamically adjust the weights of the neural beamformer. This model was the winner of the 1st IEEE MLSP Data Challenge, which then became the baseline model for the challenge's second edition, in 2022. Figure 3 illustrates its architecture, which contains three main modules: (a) encoder blocks for extracting high-level features, (b) decoder blocks for reconstructing the size of input features and (c) skip-connections for concatenating each layer in the encoder blocks with the corresponding layers in the decoder blocks. Its PyTorch original implementation available on GitHub.

### 4.2 Modifications to the baseline model

Since AIRCADE does not comprise multi-channel signals, modifications have been made to the original implementation of the Beamforming U-net, to accommodate it for single-channel audio input. Since some authors suggest that single-channel speech enhancement solutions can be extrapolated to multi-channel scenarios simply by processing each channel independently, e.g., [18–21], this modification is crucial, by promoting compatibility across a diversity of audio formats.

Besides that, PyTorch's Dataset & DataLoader classes were integrated to optimize loading, allowing samples to be processed in batches. These classes also allow audio samples to be padded to a consistent duration and resampled to fixed sample rates. This promotes a continuous and efficient flow of data preparation. It also facilitates the customization of batch sizes and allows for data shuffling, which improves the model generalization by presenting data in a randomized order during each epoch. Furthermore, a dynamic set of sample rates was implemented, allowing the model to accept any sample rate without crashing. These adjustments enable seamless automatic handling of audio signals with varying lengths and sample rates, eliminating the need for manual adjustments. Altogether, these modifications promote flexibility, performance and memory-usage efficiency.

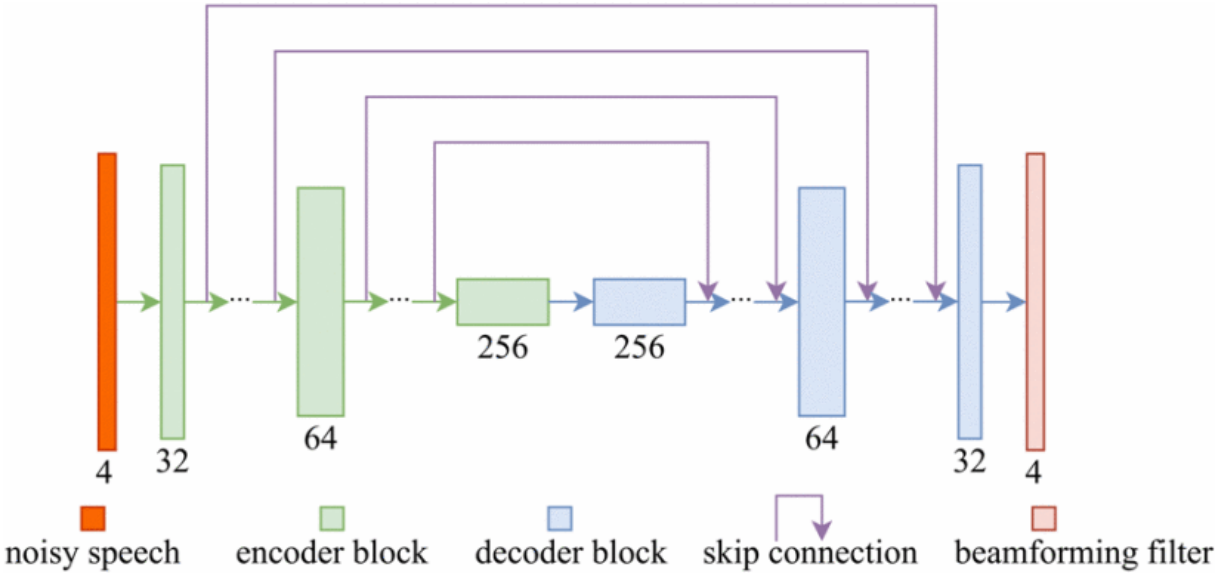


**Figure 3:** Multi-channel U-Net with neural beamformer, as proposed by Ren *et al.* [17].

### 4.3 Feature extraction and hyper-parameter settings

To prepare the data for training, a downsample to 16 kHz was adopted. This was done because higher sample rates result in greater data volumes, thus demanding more processing power and storage. Hence, adopting a lower sample rate can increase the efficiency, which is especially important when computational resources are limited.

Moreover, to represent each sample in the T-F domain, the Short-Time Fourier Transform (STFT) was used, with 512 frequency bins and a hop size of 128, to control the overlap between adjacent audio frames.

For training, the learning rate was fixed at 0.001, and the Adam algorithm was used for optimization. Batch size was fixed at 21 and a patience of 30 epochs before triggering early stopping was set. The loss function used is the Mean Absolute Error (MAE), which measures the $\ell_1$ distance between the predicted and target values.

### 4.4 Computational resources and training

Training of the modified U-net with the *tiny* version of the dataset was carried out on a personal computer (PC) equipped with an AMD Ryzen 9 5900X processor and an NVIDIA GeForce RTX 3090 Graphics Processing Unit (GPU) with 24 GB of Video Random Access Memory (VRAM). Moreover, training of the modified U-net with the *small* and *medium* versions of the dataset was carried out on the National Center for High-Performance Computing in São Paulo (CENAPAD-SP)'s powerful computational environment "lovelace". It comprises 65 processing nodes, of which two are equipped with an NVIDIA Tesla A100 GPU, each with 40 GB of VRAM.

Because the computational resources of the "lovelace" environment (memory and time) were exceeded, this study does not present the results for the *large* version. Table 2 lists the number of epochs for each trained model variant together with their respective training times and losses.

Figure 4 illustrates the training and validation loss curves for the (a) *tiny*, (b) *small* and (c) *medium* model variants, in which the curves are smoothed by taking the median over a ten-element sliding window.

### 4.5 Objective evaluation

To evaluate the enhancement promoted by each model variant, the following three objective metrics were adopted:

**Table 2:** # of epochs trained, training time and losses for each model variant.

| Version | Tiny | Small | Medium |
|---|---|---|---|
| **GPU** | GeForce RTX 3090 | Tesla A100 | Tesla A100 |
| **# of epochs trained** | 104 | 302 | 496 |
| **Training time** | 1 h 32 min | 55 h | 168 h |
| **Training loss** | 0.017621 | 0.011184 | 0.011633 |
| **Validation loss** | 0.019248 | 0.014325 | 0.013949 |
| **Testing loss** | 0.019938 | 0.01405 | 0.013724 |

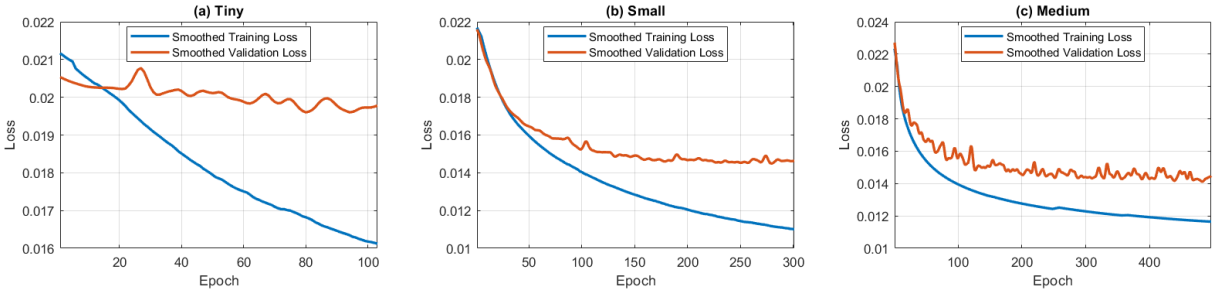


**Figure 4:** Smoothed training and validation loss curves for each model variant.

- **Perceptual Evaluation of Speech Quality (PESQ)**: follows Recommendation P.862 of the International Telecommunication Union (ITU), and is analogous to the Mean Opinion Score (MOS) in the sense that it outputs a numerical score ranging from –0.5 to 4.5. The higher the PESQ score, the better.

- **Short-Time Objective Intelligibility (STOI)**: follows its original implementation, detailed in [22], by comparing the short-term spectral envelopes of a clean speech signal and its degraded version affected by noise. The comparison is based on a model of the HAS, which outputs a numerical score between 0 and 1. The higher the STOI score, the better.

- **Signal-to-Distortion Rario (SDR)**: measures the ratio between the power of the original audio signal and the distortion introduced by a processing algorithm [23, 24]. It outputs a value in dB, in which the higher the gain, the better.

Figures 5, 6 and 7 show the results of the objective evaluation of the *tiny*, *small* and *medium* model variants using each of the aforementioned metrics and their corresponding test sets.

### 4.6 Audio examples

Figures 8, 9 and 10 illustrate examples of clean, auralized and enhanced versions of speech, song and guitar signals obtained using the *tiny*, *small* and *medium* model variants, respectively.

## 5. Discussion

Although model complexity did not change between the training of different model variants, Table 2 shows that as the dataset increases in the number of samples, losses are further minimized for model training, validation and testing. The downside is that it requires considerably more training time and epochs; however, the loss function minimization becomes gentler.

Furthermore, Figure 4 shows that, with more training data, each model variant converges and overfits later than its previous counterpart, perhaps because there is simply more to learn. However, it is arguable that a model that requires more epochs to overfit might have better generalization capabilities.

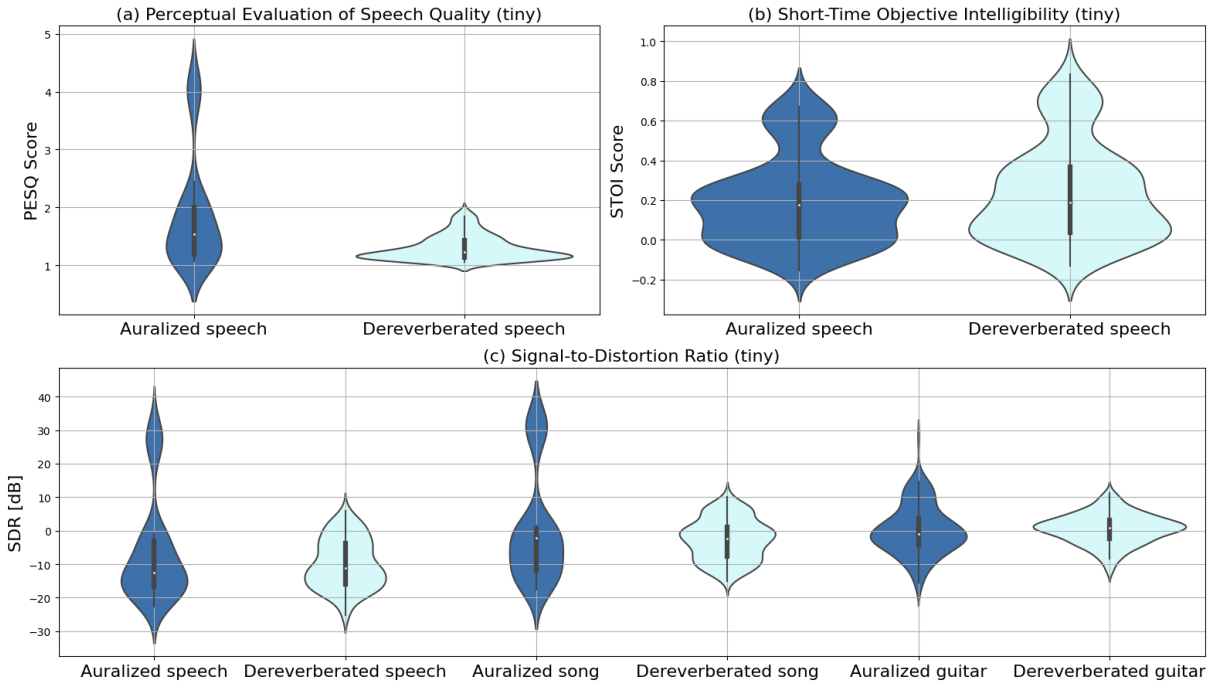


**Figure 5:** Violin plots for the objective evaluation metrics regarding the *tiny* model variant.

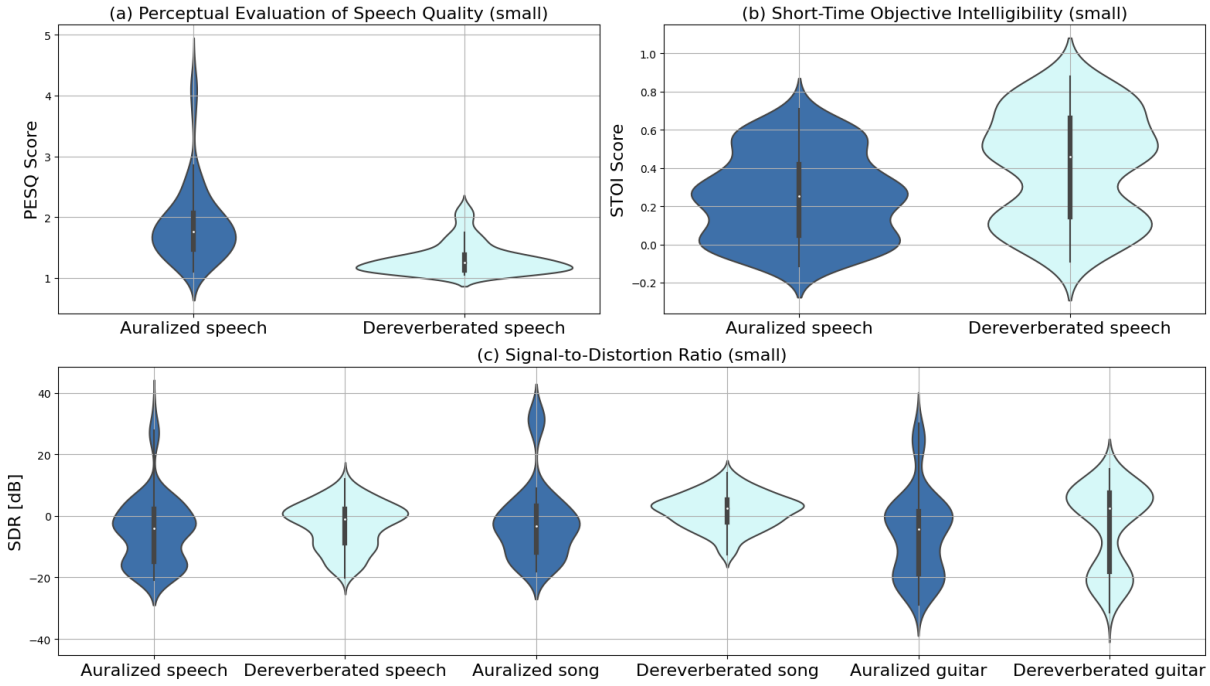


**Figure 6:** Violin plots for the objective evaluation metrics regarding the *small* model variant.

As for the objective evaluation, Figures 5, 6 and 7 evince an escalating decrease in the PESQ score of the enhanced speech signals, while showing progressive improvements in both the STOI score and in the SDR gain of all types of signals (speech, song and guitar).

By looking at Figures 8, 9 and 10, the smearing effect of reverberation is observable between all clean and auralized signals. Moreover, it is evident that the convolutive noise is filtered out by the dereverberation models; however, parts of the signals of interest are also suppressed. This results in audible artifacts (see Section 4.6), which may explain why the PESQ scores decrease between the auralized and enhanced samples. Furthermore, a reduction in the background noise power and signal duration is also observable between the auralized and enhanced signals, which becomes more noticeable for each model variant.

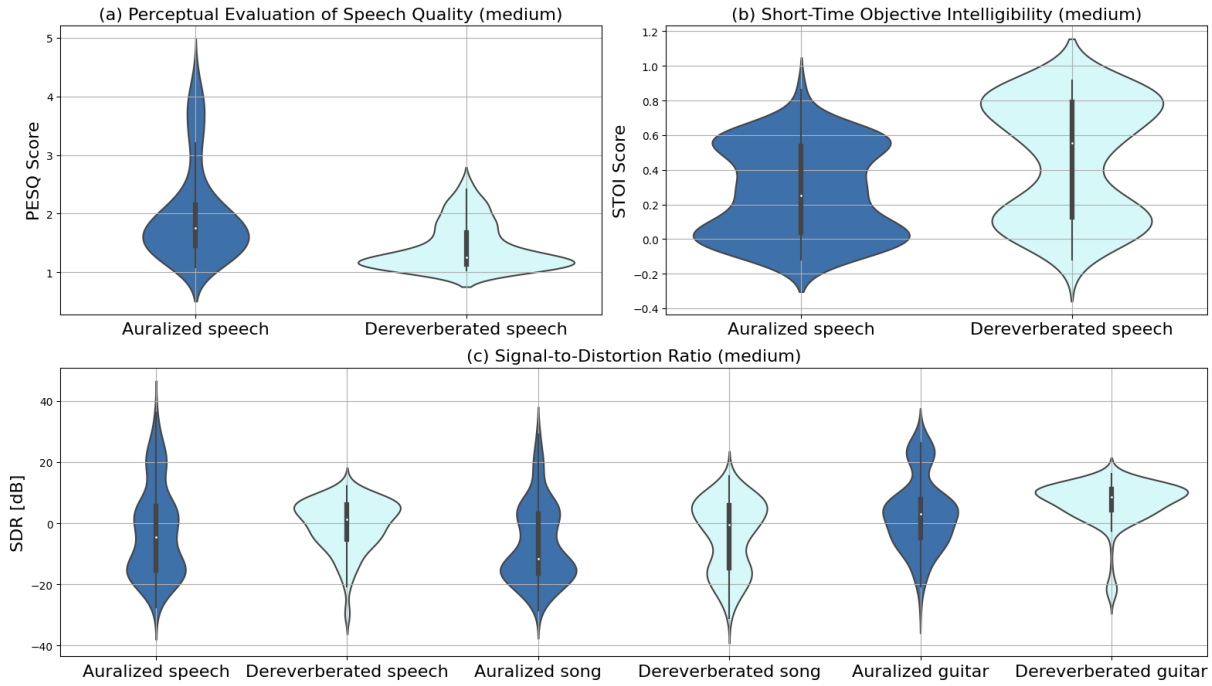


**Figure 7:** Violin plots for the objective evaluation metrics regarding the *medium* model variant.

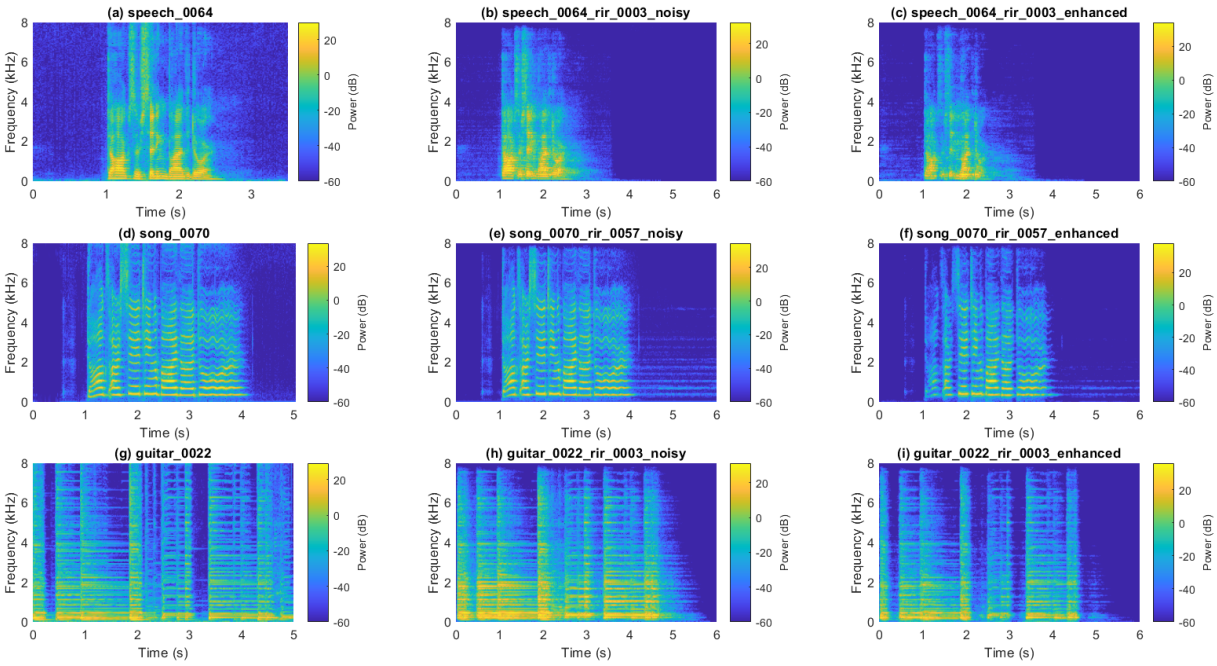


**Figure 8:** Examples of clean (left column), auralized (center column) and enhanced (right column) samples for the test set corresponding to the *tiny* model variant. Top row represents speech, center row represents song and the bottom row represents guitar signals. To listen these samples, the reader is referred to this link.

Although it was not possible to train the model variants with samples at 48 kHz owing to computational limitations, future endeavors in this sense are not discarded.

Finally, regarding the training of a *large* model variant, the observable tendency is that it will minimize losses even further than the others, yet at a more moderate rate.

This evidence prompts further investigation, and future research endeavors could provide insights into whether model efficiency and generalization capabilities reside in dataset size and quality or model complexity.

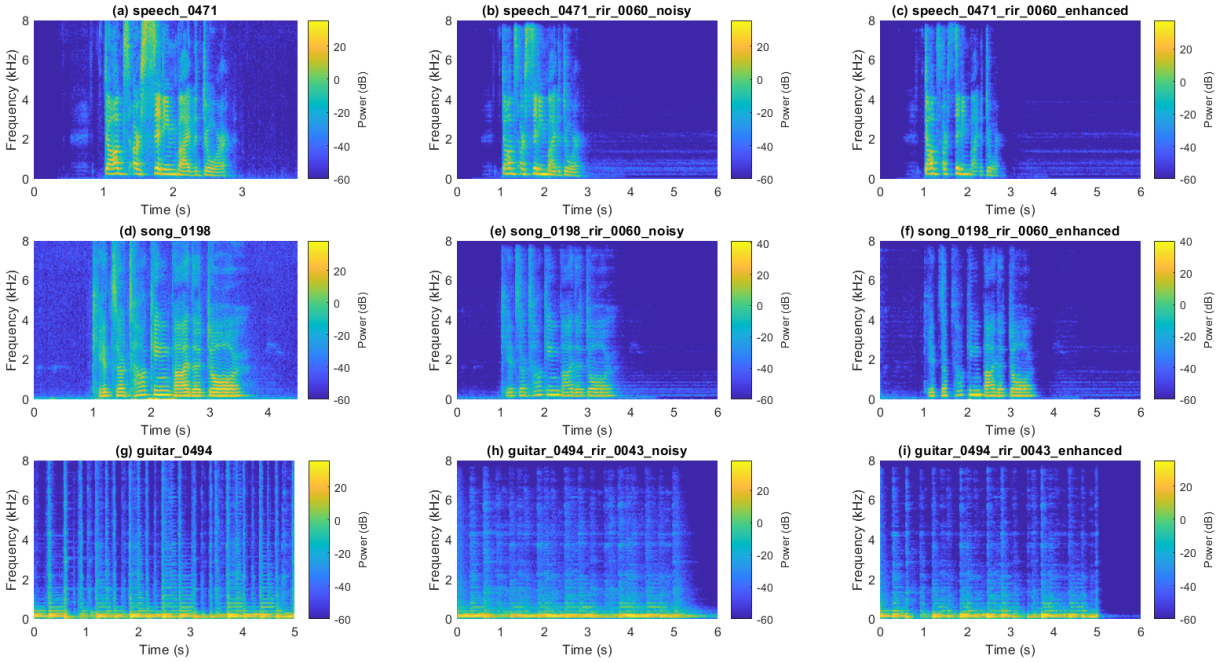


**Figure 9:** Examples of clean (left column), auralized (center column) and enhanced (right column) samples for the test set corresponding to the *small* model variant. Top row represents speech, center row represents song and the bottom row represents guitar signals. To listen these samples, the reader is referred to this link.

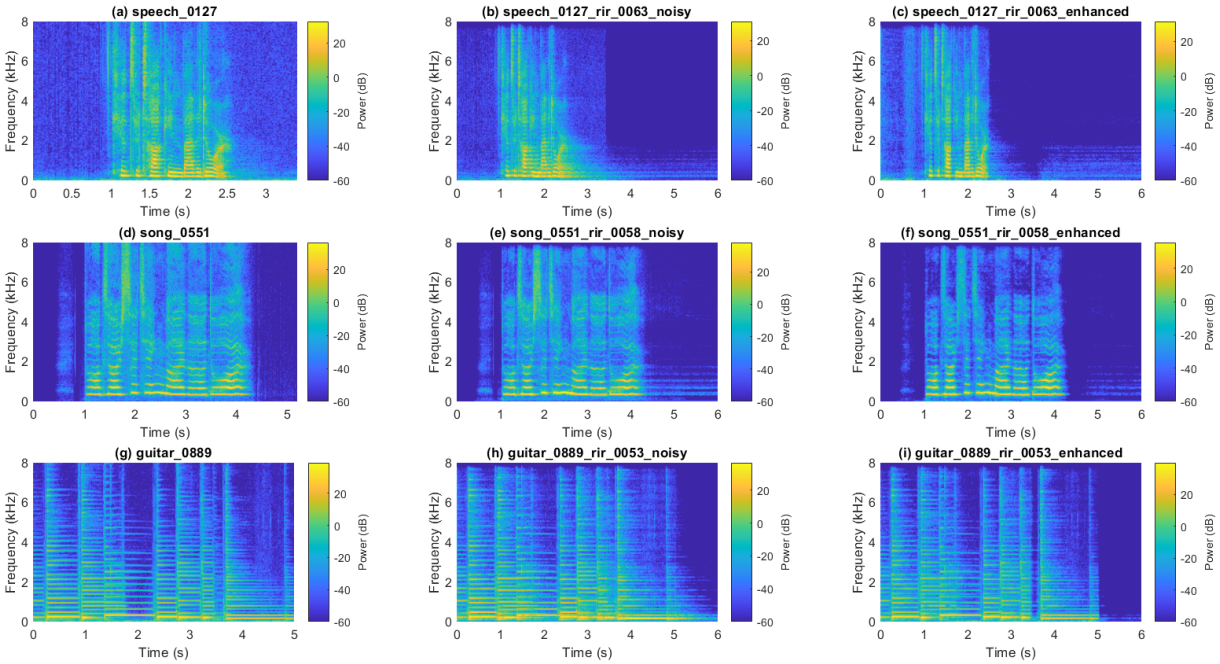


**Figure 10:** Examples of clean (left column), auralized (center column) and enhanced (right column) samples for the test set corresponding to the *medium* model variant. Top row represents speech, center row represents song and the bottom row represents guitar signals. To listen these samples, the reader is referred to this link.

## 6. Conclusion

Overall, the data-compilation ensemble presented in this work provides a diverse and comprehensive set of acoustic scenes for use in dereverberation tasks as well as in other audio signal processing applications, such as Emotion Recognition and MIR. By combining different types of signals of interest, including speech, song and acoustic guitar sounds, together with a variety of IR data, a challenging dataset is provided for researchers working on dereverberation and related tasks.

The dataset is available in different sizes, from a *tiny* version with limited data to a *large* version with

almost 300,000 samples, allowing users to choose the most suitable version for their specific research needs. The dataset also includes metadata that can be used to trace each sample back to its original annotations, facilitating the use of the dataset for tasks such as Emotion Recognition and MIR.

One of the key aspects of this study is the investigation of the trade-off between the dataset size and model generalization. It is essential to note that as the dataset expands, the model's ability to reduce the effects of reverberation gradually improves. However, there may be diminishing returns beyond a certain dataset size. Therefore, researchers should carefully consider the trade-offs between dataset size and model complexity, particularly in resource-constrained environments.

Furthermore, an escalating decrease in the perceptual quality of enhanced audio is evinced by the introduction of artifacts. However, the STOI and SDR metrics showed positive outcomes, indicating that the dereverberation process enhances the speech intelligibility and reduces the impact of background noise. Thus, depending on the specific use case, the benefits of dereverberation may outweigh the reduction in the perceived audio quality. Therefore, this tradeoff is a critical consideration when working with DL dereverberation models.

From the results of this study, possible future research endeavors include exploring the impact of training with 48 kHz samples for applications that demand higher audio fidelity, investigating the scalability of the dataset concerning available computing power, and experimenting with DL architectures tailored for dereverberation, which may enable the development of more specialized models for niche applications.

We hope that this dataset will be useful for researchers working in dereverberation and related fields, and encourage its use in future research. We also believe that the diversity and variability of the dataset can facilitate the development of more robust and generalizable algorithms for dereverberation and other audio signal-processing tasks.

Researchers are encouraged to explore, use, and build on this dataset to foster collaboration and advance the boundaries of audio signal processing.

## 7. Acknowledgements

This work was partially supported by the São Paulo Research Foundation (FAPESP), grants #2017/08120-6, #2019/22795-1 and #2022/16168-7.

This work was partially developed with the support of the National Center for High-Performance Computing in São Paulo (CENAPAD-SP), a project of the University of Campinas (UNICAMP) together with the Research and Projects Financing (FINEP) and Ministry of Science and Technology (MCTI).

## Referências

1. BERANEK, Leo. *Concert Halls and Opera Houses*. [S.l.]: Springer New York, 2004. doi: 10.1007/978-0-387-21636-2. ISBN 978-1-4419-3038-5.

2. LAPOINTE, Yannick. Understanding and crafting the mix : The art of recording: Canadian journal of music. *Intersections*, v. 31, n. 1, p. 209–213,221, 2010. Copyright - Copyright Canadian University Music Society 2010; Última atualização em - 2023-09-01. Disponível em: https://www.proquest.com/scholarly-journals/understanding-crafting-mix-art-recording/docview/1024799804/se-2.

3. GELFAND, Stanley A. *Hearing*. [S.l.]: CRC Press, 2017. doi: 10.1201/9781315154718. ISBN 9781315154718.

4. LYON, Richard F. *Human and Machine Hearing: Extracting Meaning from Sound*. [S.l.]: Cambridge University Press, 2017. doi: 10.1017/9781139051699.

5. NAYLOR, Patrick A.; GAUBITCH, Nikolay D. (Ed.). *Speech Dereverberation*. [S.l.]: Springer London, 2010. doi: 10.1007/978-1-84996-056-4. ISBN 978-1-84996-055-7.

6. XU, Yong; DU, Jun; DAI, Li-Rong; LEE, Chin-Hui. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 23, n. 1, p. 7–19, 2015. doi: 10.1109/TASLP.2014.2364452.

7. HERSHEY, John R.; CHEN, Zhuo; ROUX, Jonathan Le; WATANABE, Shinji. Deep clustering: Discriminative embeddings for segmentation and separation. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2016. p. 31–35. doi: 10.1109/ICASSP.2016.7471631.

8. KINOSHITA, Keisuke; DELCROIX, Marc; YOSHIOKA, Takuya; NAKATANI, Tomohiro; HABETS, Emanuel; HAEB-UMBACH, Reinhold; LEUTNANT, Volker; SEHR, Armin; KELLERMANN, Walter; MAAS, Roland; GANNOT, Sharon; RAJ, Bhiksha. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. [S.l.: s.n.], 2013. p. 1–4. doi: 10.1109/WASPAA.2013.6701894.

9. SZöKE, Igor; SKáCEL, Miroslav; MOšNER, Ladislav; PALIESEK, Jakub; ČERNOCKý, Jan. Building and evaluation of a real room impulse response dataset. *IEEE Journal of Selected Topics in Signal Processing*, v. 13, n. 4, p. 863–876, 2019. doi: 10.1109/JSTSP.2019.2917582.

10. FU, Szu-Wei; YU, Cheng; HUNG, Kuo-Hsuan; RAVANELLI, Mirco; TSAO, Yu. Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 7412–7416. doi: 10.1109/ICASSP43922.2022.9747180.

11. KLEINER, Mendel; DALENBäCK, Bengt-Inge; SVENSSON, Peter. Auralization-an overview. *J. Audio Eng. Soc*, v. 41, n. 11, p. 861–875, 1993. Disponível em: http://www.aes.org/e-lib/browse.cfm?elib=6976.

12. ALLEN, Jont B.; BERKLEY, David A. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, v. 65, p. 943–950, 4 1979. ISSN 0001-4966. doi: 10.1121/1.382599.

13. OPPENHEIM, Alan V.; WILLSKY, Alan S.; NAWAB, S. Hamid. *Signals & Systems (2nd Ed.)*. USA: Prentice-Hall, Inc., 1996. ISBN 0138147574.

14. SANTOS, Arthur dos; OLIVEIRA, Pedro de; MASIERO, Bruno. A retrospective on multichannel speech and audio enhancement using machine and deep learning techniques. In: *Proceedings of the 24th International Congress on Acoustics*. Gyeongju, South Korea: [s.n.], 2022. p. 173–184. Disponível em: https://ica2022korea.org/data/Proceedings_A05.pdf.

15. LIVINGSTONE, Steven R.; RUSSO, Frank A. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Zenodo, 2018. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. Disponível em: https://doi.org/10.5281/zenodo.1188976.

16. XI, Qingyang; BITTNER, Rachel M.; PAUWELS, Johan; YE, Xuzhou; BELLO, Juan P. *GuitarSet*. Zenodo, 2019. Changes from version 1.0.1: - Added Key Annotation in each jams file - Removed impossible notes (negative frets) - Updated Pitch Contour annotations to have the correct ‘’index’‘ field in ‘obs.value‘. - Cleaned up .zip files. Disponível em: https://doi.org/10.5281/zenodo.3371780.

17. REN, Xinlei; CHEN, Lianwu; ZHENG, Xiguang; XU, Chenglin; ZHANG, Xu; ZHANG, Chen; GUO, Liang; YU, Bing. A neural beamforming network for b-format 3d speech enhancement and recognition. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. [S.l.: s.n.], 2021. p. 1–6. doi: 10.1109/MLSP52302.2021.9596418.

18. BABY, Deepak; BOURLARD, Hervé. Speech dereverberation using variational autoencoders. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2021. p. 5784–5788. doi: 10.1109/ICASSP39728.2021.9414736.

19. WESTHAUSEN, Nils L.; HUBER, Rainer; BAUMGARTNER, Hannah; SINHA, Ragini; RENNIES, Jan; MEYER, Bernd T. Reduction of subjective listening effort for tv broadcast signals with recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 29, p. 3541–3550, 2021. doi: 10.1109/TASLP.2021.3126931.

20. MA, Lu; YANG, Song; GONG, Yaguang; WU, Zhongqin. *Noise Classification Aided Attention-Based Neural Network for Monaural Speech Enhancement*. 2021.

21. KOIZUMI, Yuma; KARITA, Shigeki; WISDOM, Scott; ERDOGAN, Hakan; HERSHEY, John R.; JONES, Llion; BACCHIANI, Michiel. *DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement*. 2021.

22. TAAL, Cees H.; HENDRIKS, Richard C.; HEUSDENS, Richard; JENSEN, Jesper. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2010. p. 4214–4217. doi: 10.1109/ICASSP.2010.5495701.

23. VINCENT, E.; GRIBONVAL, R.; FEVOTTE, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 14, n. 4, p. 1462–1469, 2006. doi: 10.1109/TSA.2005.858005.

24. SCHEIBLER, Robin. Sdr — medium rare with fast computations. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2022. p. 701–705. doi: 10.1109/ICASSP43922.2022.9747473.