

Audio denoising in the wild

Relatore: Prof. Simone Bianco

Correlatore: Dott. Paolo Napoletano

Tesi di Laurea Magistrale di:

Fabrizio D'Intinosante

838866



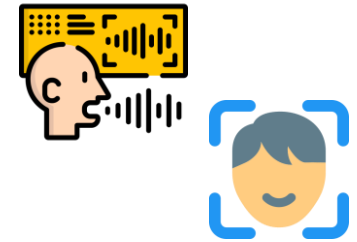
INTRODUCTION

- » premise
- » main purposes
- » basic concepts



audio-related tasks are **very popular** thanks to the amount of existing voice assistants

among all, the most common are **speaker recognition** and **speech recognition**



there are numerous problems for these tasks, **background noise** is one of the main ones [1]



the **primary objective** is the creation of a system for the background noise removal



subsequently verify the possible positive impact of this removal on the performance of a speaker recognition system through an **integrated training**





the **main idea** is to create a noise removal system that includes the following features



neural architecture



limited data
pre-processing



robustness to noise
intensity and duration of
the audio signal



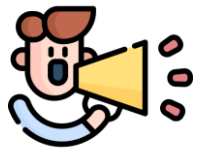
» resources

» procedure

DATASETS



different resources



clean audio
sources



noise
source



audio quality
evaluation

SIWIS [2]

VoxCeleb 2 [3]
($\frac{1}{3}$ test set)

MUSAN [4]

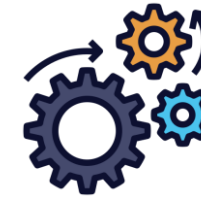
DEMAND [5]



Additive mixing using
different SNRs



the objective is the creation of 3 **distinct** dataset for the model training



SIWIS



VoxCeleb 2

SIWIS+VoxCeleb 2

- 16 kHz re-sampled
- 2.5, 7.5, 12.5 and 17.5 SNR noise addition
- 1 second sliding window
- rebalance through oversampling
- ark, scp storing format

METHODOLOGICAL APPROACH

- » models
- » experimental setup
- » pipeline details



two models
with two tasks



WaveNet for the
denoising task [6]



Implemented **from scratch** using
PyTorch and trained on the 3
different datasets



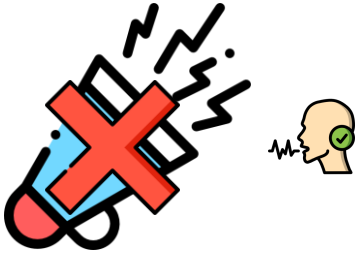
Thin ResNet-34 for the
speaker verification task [7]



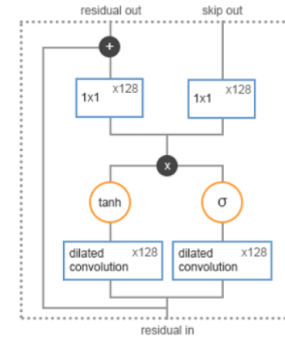
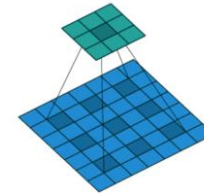
pre-trained on VoxCeleb 1
and VoxCeleb 2 train sets



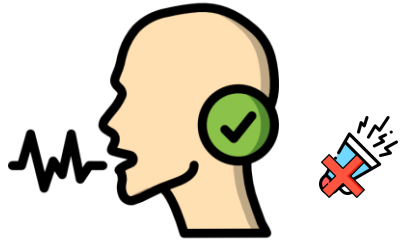
WaveNet characteristics



- **raw** audio as input data
- **Gated Activation Units** (residual blocks)
- non-causal, **dilated** convolutions
- symmetrical 3 x 1 convolutions
- **real value** output with 1 : 1 ratio respect to input
- energy conserving loss

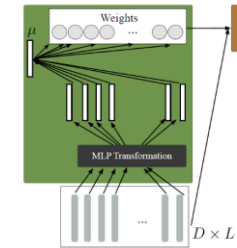
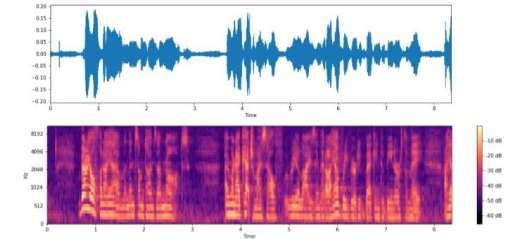
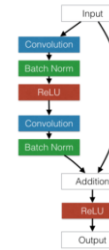


$$\mathcal{L}(\hat{s}_t) = |s_t - \hat{s}_t| + |b_t - \hat{b}_t|$$



Thin ResNet-34 characteristics

- **Mel spectrogram** as input
- classical residual blocks
- thin because has $\frac{1}{4}$ **filters** compared to normal counterpart
- **Self-Attemptive Pooling** layer
- 1 x 512 embedding dimension
- angular prototypical loss




$$L_P = -\frac{1}{N} \sum_{j=1}^N \log \frac{e^{S_{j,j}}}{\sum_{k=1}^N e^{S_{j,k}}}$$



experimental setup



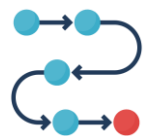
WaveNet

- 3 models (1 for dataset)
- 30 GAUs conv. blocks with 128 filters
- 3 stacks (1 to 512 dilation factor)
- 2048 and 256 filters on final conv layers
- 6.3 mln parameters
- scheduled Adam optimizer
- training and fine-tuning procedure with **variable** n. of epochs 



Thin ResNet-34

- pre-trained with ~500 epochs and scheduled Adam optimizer
- 34 residual conv. blocks
- 1.4 mln parameters
- used as component for the fine-tuning procedure with the WaveNets



pipeline



WaveNet init

first training



fine-tuning **stacking**
with Thin ResNet-34



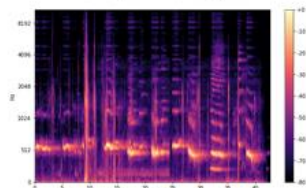
noisy input

WaveNet

enhanced output



Mel spectrogram



energy
conserving
loss



Thin
ResNet-34

$\lambda * \text{MSE loss}$

embedded
output

1
2
3
4
5
6
7
8
⋮
⋮
⋮
509
510
511
512



» measures

» speaker verification

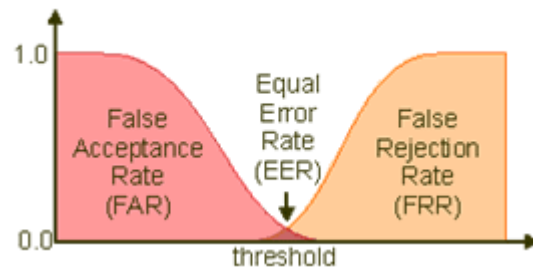
» denosing

RESULTS



two types of measures

speaker verification quality



Equal Error Rate (**EER**)

audio enhancement quality

objective quality measures [8]

- signal distortion (**SIG**)
- background noise distortion (**BAK**)
- overall quality (**OVL**)
- perceptual evaluation of speech quality (**PESQ**)
- segmental signal-to-noise ration (**SSNR**)

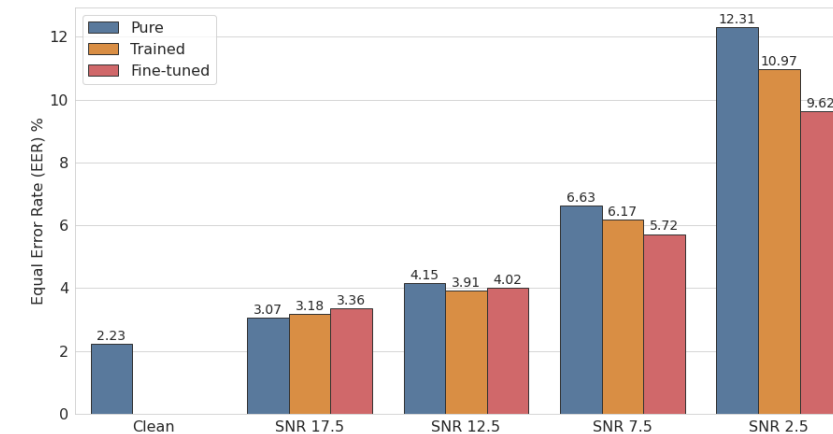
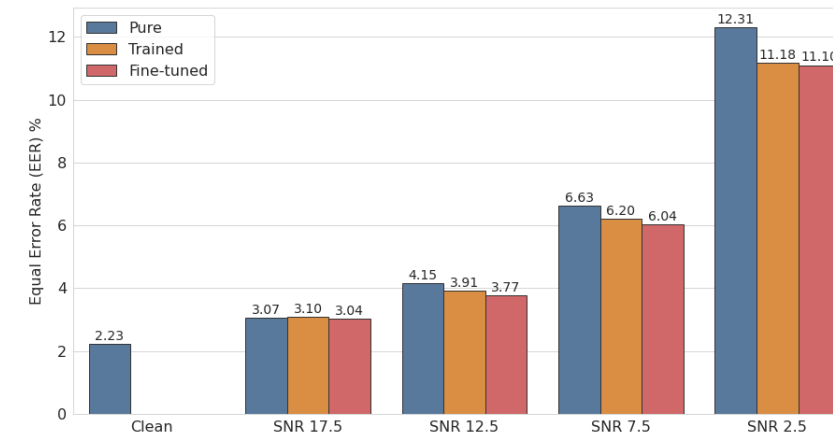
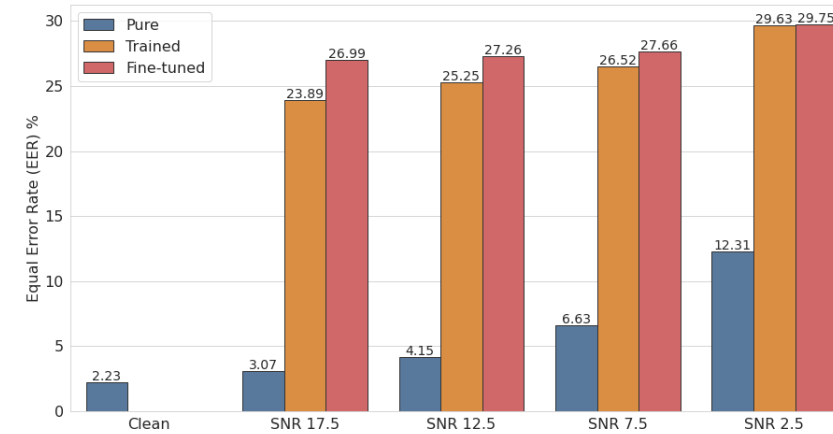
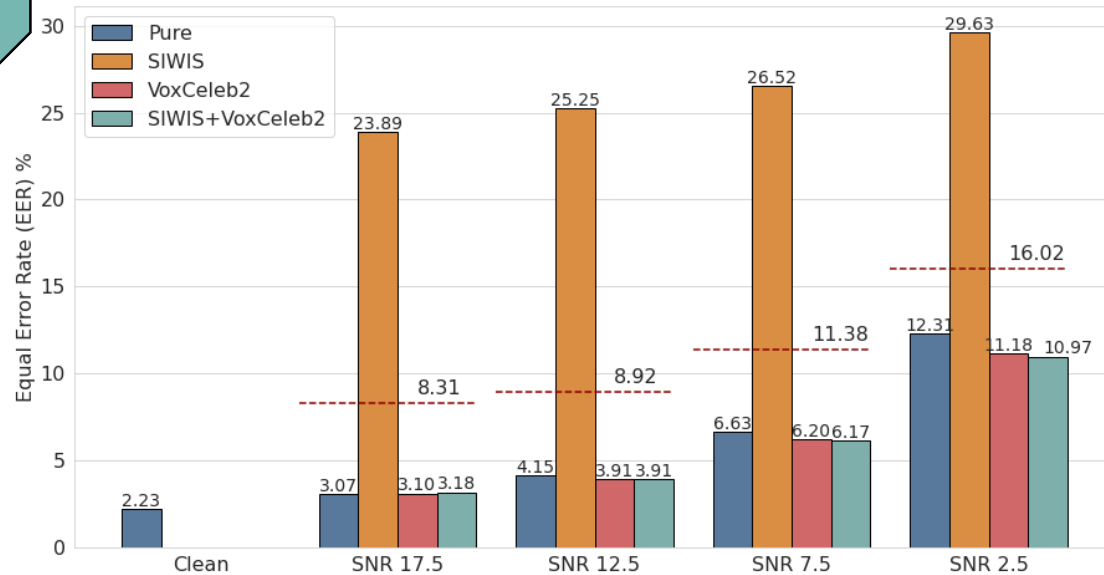


SIWIS

VoxCeleb

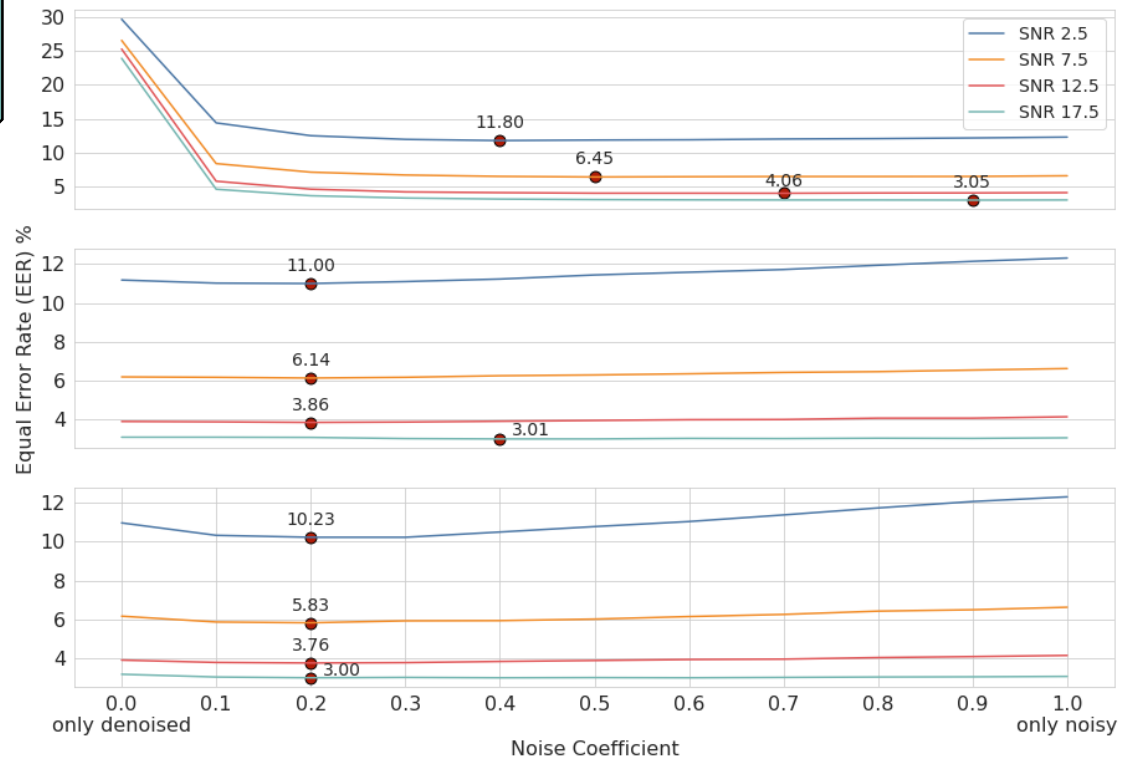
SIWIS
+
VoxCeleb

first training performance

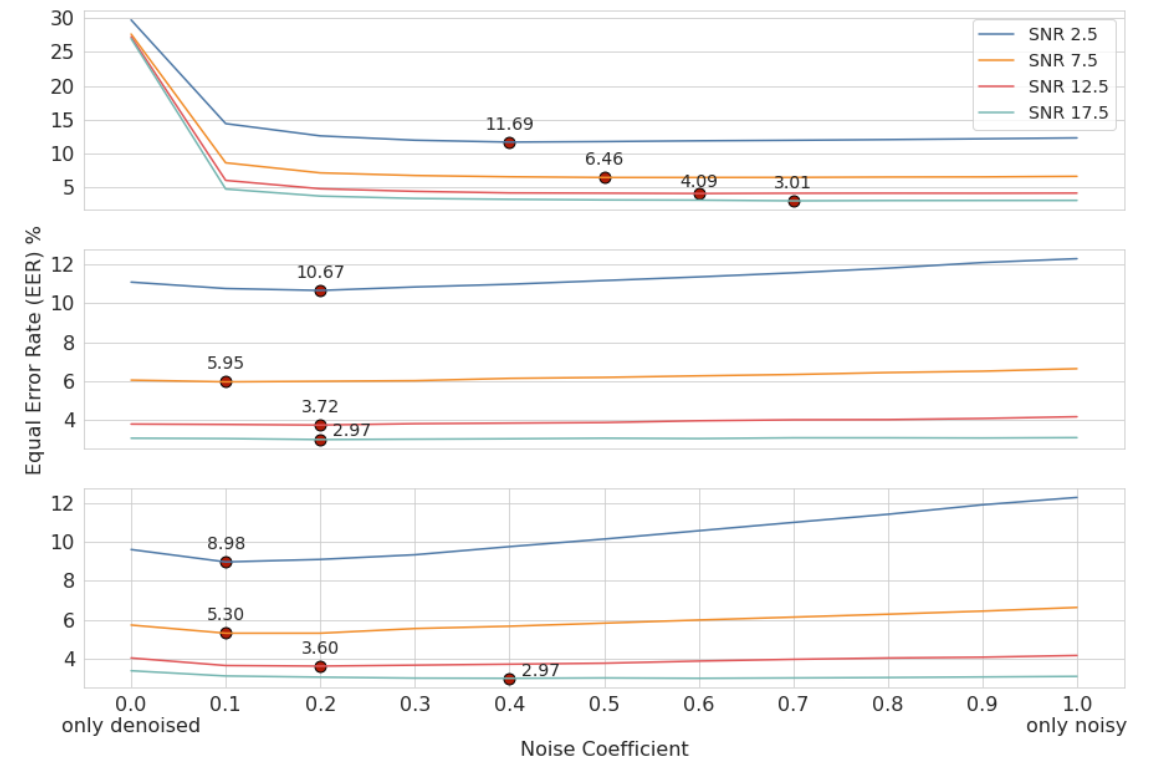




first training



fine tuning

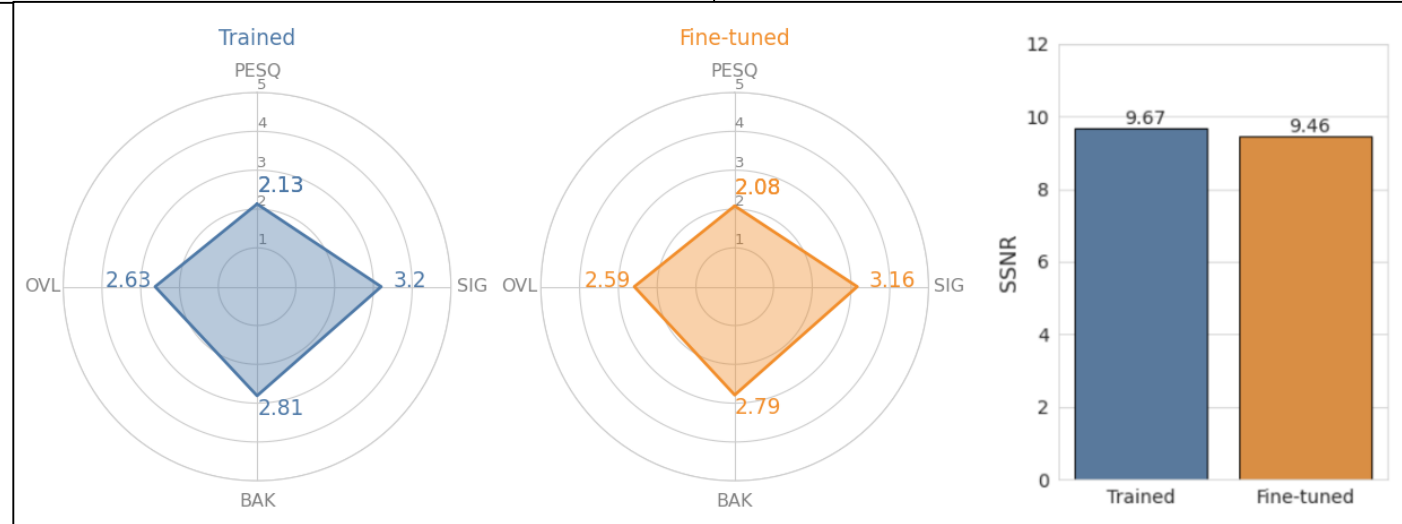
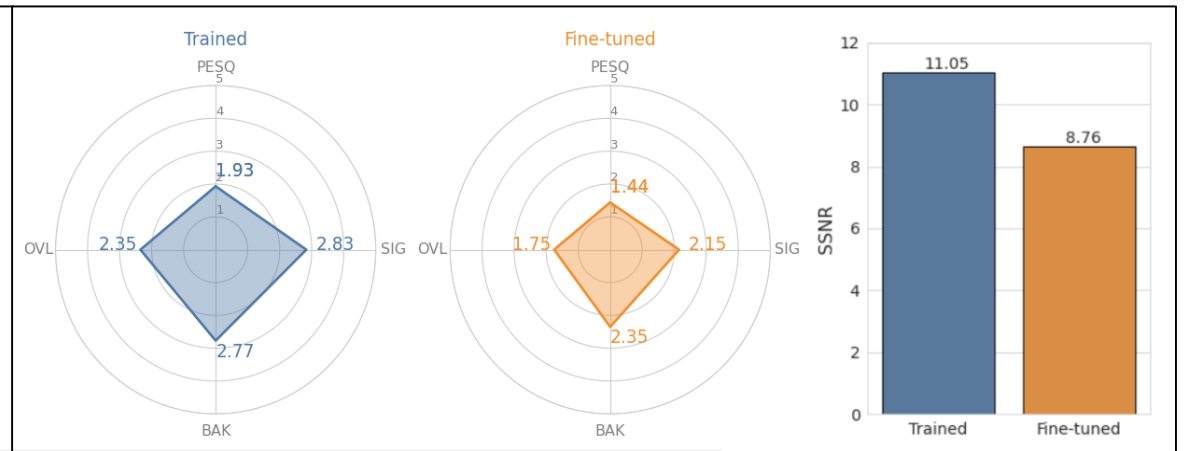




SIWIS



VoxCeleb



SIWIS+VoxCeleb

DISCUSSION





some considerations on the obtained results

- 1 as expected, the **greater** the SNR, the **greater** the classification error
- 2 the **larger** and more varied the dataset, the **better** the performance
- 3 fine-tuning brought a real **benefit** to performance
- 4 the mixing procedure is **effective** for each model, for each SNR, both before and after the fine-tuning
- 5 audio quality **degrades** for each model
- 6 an improvement in classification **does not necessarily coincide** with an improvement in audio quality



🔊 final considerations

🔊 future work

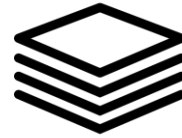
CONCLUSIONS



in conclusion...



there is a **real improvement** in performance



the stacking strategy has shown **encouraging results**

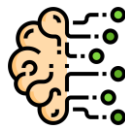


despite the limited resources,
the modesty of the results
bodes well

possible improvements...



more computational resources



use of
generative models
(CycleGANs)



improve denoising by
classifying the type of noise



try to apply denoising
techniques on **non-audio**
signals (i.e. electrocardiogram)

THANK YOU



REFERENCES

- [1] N. Singh, A. Agrawal, and R. A. Khan. "Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades". In: Global Journal of Enterprise Information System 9.3 (2017), pp. 45–52.
- [2] Pierre-Edouard Honnet et al. "The SIWIS French Speech Synthesis Database Design and recording of a high quality French database for speech synthesis". In: (Jan. 2017).
- [3] A. Nagrani, J. S. Chung, and A. Zisserman. "VoxCeleb: a large-scale speaker identification dataset". In: (2017).
- [4] D. Snyder, G. Chen, and D. Povey. "MUSAN: A Music, Speech, and Noise Corpus". In: (2015).
- [5] C. Valentini-Botinhao. "Noisy speech database for training speech enhancement algorithms and TTS models". In: (2017).
- [6] D. Rethage, J. Pons, and X. Serra. "A Wavenet for Speech Denoising". In: (2017).
- [7] J. S. Chung et al. "In defence of metric learning for speaker recognition". In: (2020).
- [8] P. Krishnamoorthy. "An Overview of Subjective and Objective Quality Measures for Noisy Speech Enhancement Algorithms". In: IETE Technical Review 28.4 (2011), pp. 292–301.



Dataset	Train set split	Test set split
SIWIS	90 %	10 %
VoxCeleb2	70 %	30 %

Dataset	Portion	n. of samples	avg. samples per speaker
SIWIS	train	7,344	~ 333
	test	816	~ 37
VoxCeleb2	train	8,448	~ 71
	test	2,113	~ 17
SIWIS + VoxCeleb2	train	15,792	~ 112
	test	2,929	~ 21

Dataset	Portion	n. of samples	avg. samples per speaker
SIWIS	train	7,344	~ 333
	test	816	~ 37
VoxCeleb2	train	9,027	~ 76
	test	2,287	~ 19
SIWIS + VoxCeleb2	train	25,334	~ 180
	test	5,482	~ 39

Model	n. of epochs training	n. of epochs fine-tuning
SIWIS	75	83
VoxCeleb	92	66
SIWIS+VoxCeleb	87	82



Università degli Studi di Milano – Bicocca

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

Audio denoising in the wild

Relatore: Prof. Simone Bianco

Correlatore: Dott. Paolo Napoletano

Tesi di Laurea Magistrale di:

Fabrizio D'Intinosante

838866

