Desenvolvimento de Base de Dados para Treinamento de Redes Neurais de Reconhecimento de Voz Através da Geração de Áudios com Resposta ao Impulso Simuladas por Técnicas de Data Augmentation

Bruno Machado Afonso

bruno.ma@poli.ufrj.br

Departamento de Engenharia Eletrônica e de Computação - Escola Politécnica

Universidade Federal do Rio de Janeiro

14 de julho de 2021

Sumário

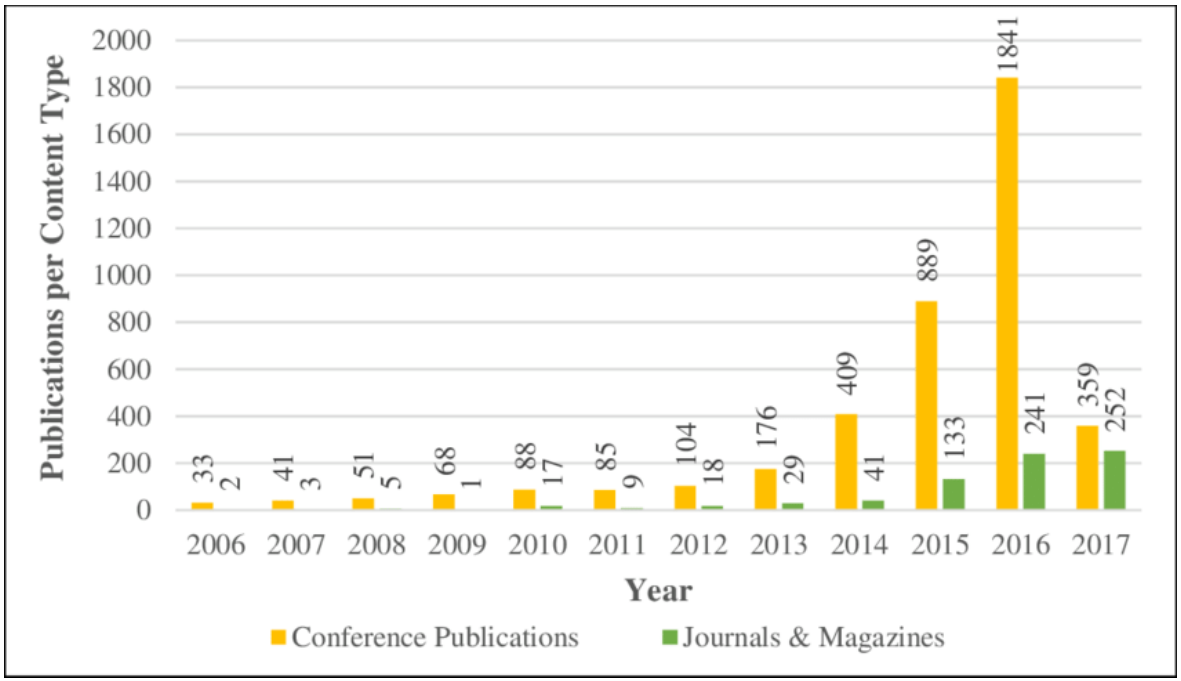
Motivação

Crescimento no número de aplicações de algoritmos de processamento de áudio.

- Detecção e reconhecimento de voz
  - Smartphones
  - Automação residencial
  - Comunicação online
- Cancelamento de eco
- Separação de fontes

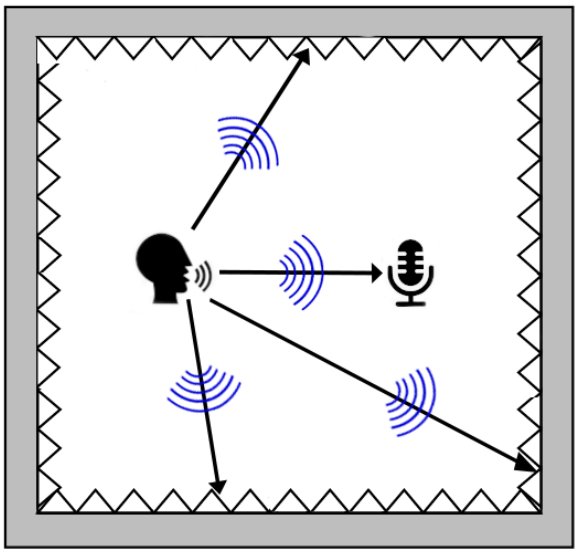
Deep Learning

Aumento no número de artigos que envolvem *deep learning* publicados em grandes conferências.

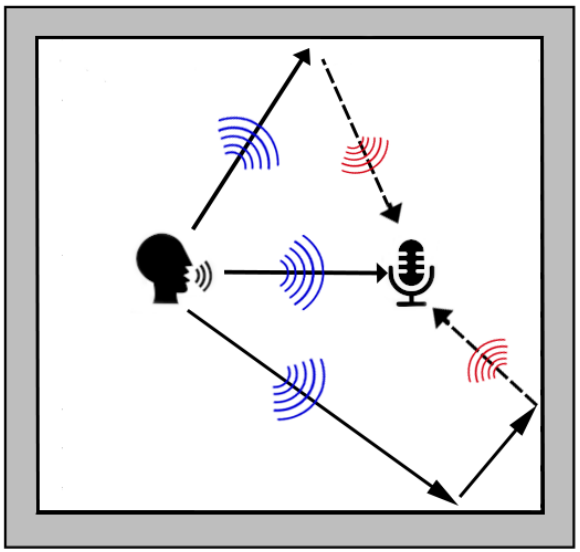


Amostra de Voz em Campo Distante (AVCD)

Sinal de voz anecóico que é corrompido pela reverberação do ambiente fechado e ruído.



(a) Sala anecóica



(b) Sala reverberante

Amostra de Voz em Campo Distante (AVCD)

$$Y(t) = s(t) * h(t) + n(t)$$

$Y(t) \rightarrow$ AVCD
$s(t) \rightarrow$ Amostra de Voz Anecóica
$h(t) \rightarrow$ Resposta ao Impulso de Sala (RIR)
$n(t) \rightarrow$ Sinal de Ruído

Resposta ao Impulso de Sala (RIR)

Representa um modelo acústico de um ambiente para um par fonte/-receptor.

- Razão Direto-Reverberante (DRR)
- Tempo de Reverberação (T60)

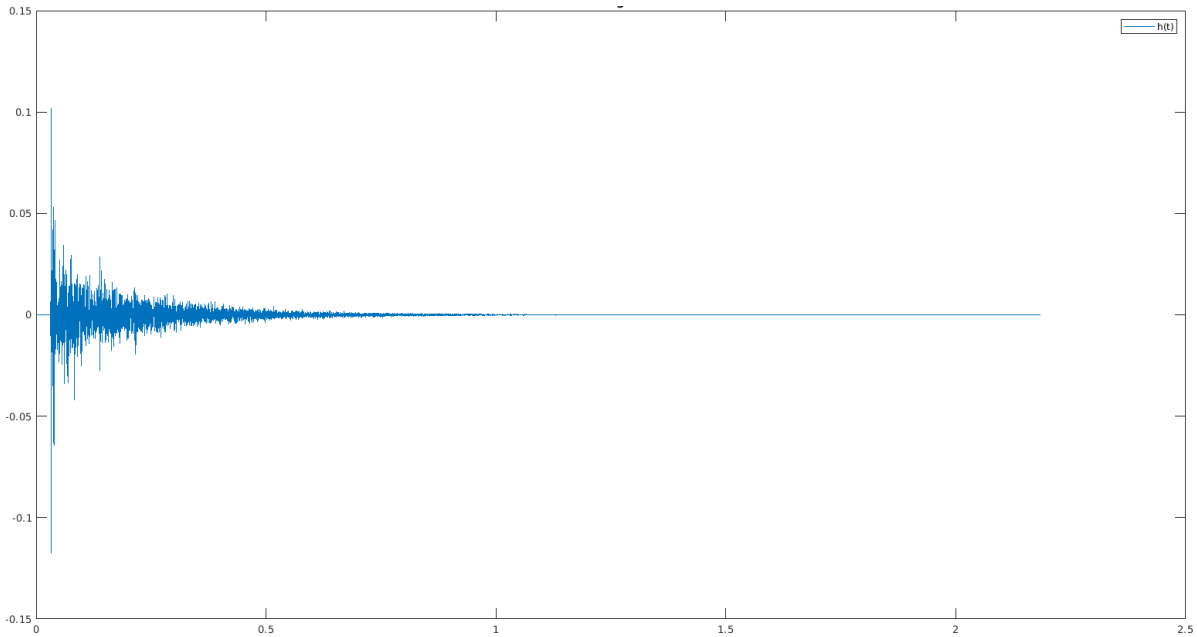
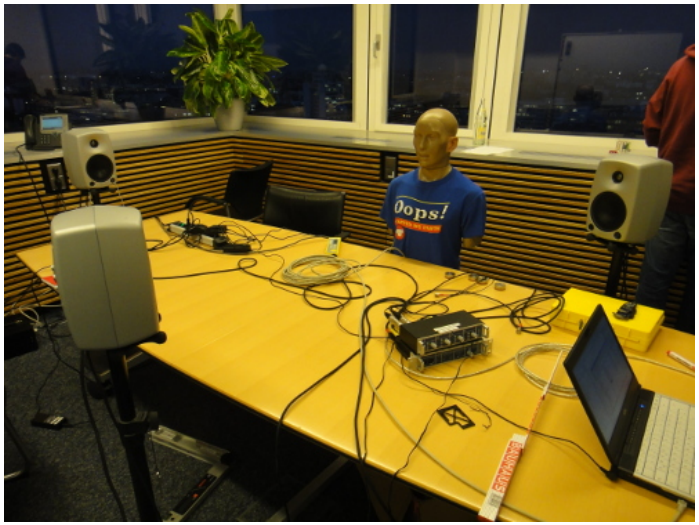


Figura: $DRR = -4,5$ dB / $T60 = 1,38$ s

Desafios

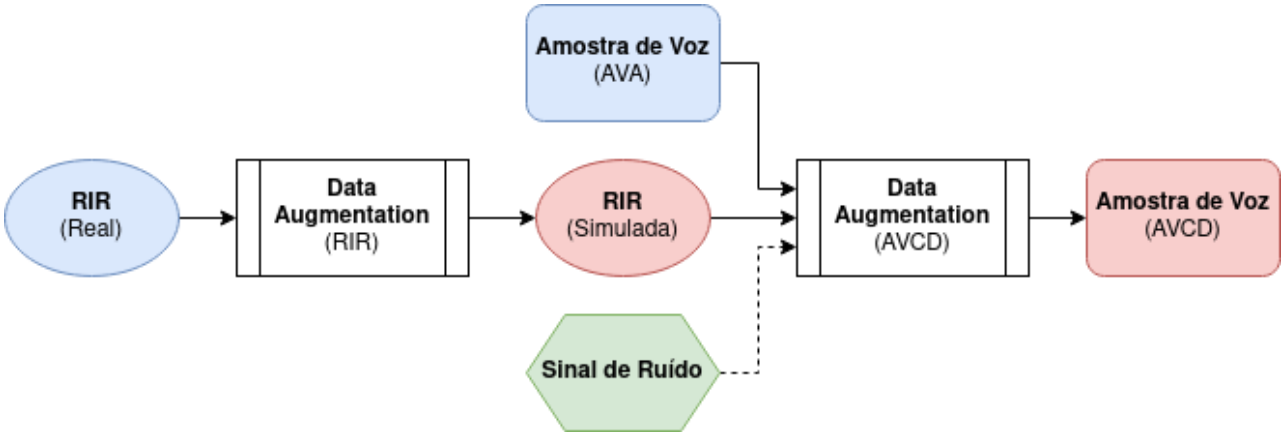
- Baixa quantidade e variedade de bases de dados contendo RIRs anotadas para treinamento de redes de *deep learning*.
- Dificuldade para realizar gravações de RIRs (equipamentos especializados, variedade de ambientes, etc.)



Data Augmentation (DA)

Proposta de duas técnicas de *data augmentation* para gerar AVCDs artificialmente.

- DA para gerar RIRs simuladas (RIRSM)
  - Razão Direto-Reverberante (DRR)
  - Tempo de Reverberação (T60)
- DA para gerar AVCDs, usando RIRSMs e ruídos



Data Augmentation (DA)

As técnicas de DA de RIRSM e AVCDs foram baseadas, respectivamente, nos artigos abaixo.

- [1] - “Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation” , N. J. Bryan, ICASSP 2020
- [2] - “A study on data augmentation of reverberant speech for robust speech recognition” , T. Ko et al, ICASSP 2017

Data Augmentation (DA)

$$h_e(t) = \begin{cases} h(t), & t_d - t_0 \leq t \leq t_d + t_0 \\ 0, & \text{caso contrário.} \end{cases}$$

$$h_l(t) = \begin{cases} h(t), & t < t_d - t_0 \\ h(t), & t > t_d + t_0 \\ 0, & \text{caso contrário.} \end{cases}$$

$h(t) \rightarrow$ RIR
$h_e(t) \rightarrow$ Resposta inicial
$h_l(t) \rightarrow$ Resposta atrasada
$t_d \rightarrow$ Tempo levado pelo impulso sonoro da fonte até o receptor
$t_0 \rightarrow$ Janela de tolerância ($t_0 = 2,5$ ms, definido por [1])

DA - Razão Direto-Reverberante (DRR)

Definição do DRR:

$$DRR_{dB} = 10 \log_{10} \left(\frac{\sum_t h_e^2(t)}{\sum_t h_l^2(t)} \right)$$

DA do DRR:

$$h'_e(t) = \alpha w_d(t) h_e(t) + [1 - w_d(t)] h_e(t)$$

$w_d(t) \rightarrow$ Janela de Hann de duração $2t_0$

DA - Razão Direto-Reverberante (DRR)

Substituindo $h_e(t)$ por $h'_e(t)$ na definição do DRR:

$$\alpha^2 \sum_t w_d^2(t) h_e^2(t) + 2\alpha \sum_t [1 - w_d(t)] w_d(t) h_e^2(t) + \sum_t [1 - w_d(t)]^2 h_e^2(t) - 10^{DRR_{dB}/10} \sum_t h_l^2(t) = 0$$

O parâmetro α desejado é a raiz de maior valor.

DA - Tempo de Reverberação (T60)

Definição do T60:

$$\begin{cases} t_i, \text{ onde } h(t_i) = max(h(t)) \\ t_f, \text{ onde } 10\log_{10}\left(h^2(t_i) - h^2(t_f)\right) = 60\text{dB} \\ \text{T60} = t_f - t_i \end{cases}$$

Modelo de $h_l(t)$:

$$h_m(t) = Ae^{-(t-t_o)/\tau}n(t)u(t-t_o) + \sigma n(t)$$

$A \rightarrow$ Ganho da RIR
$\tau \rightarrow$ Taxa de decaimento
$\sigma \rightarrow$ Desvio padrão do ruído de chão
$n(t) \rightarrow$ Ruído gaussiano padrão
$t_o \rightarrow$ Balor temporal onde $h_l(t)$ tem seu primeiro valor não nulo
$u(t) \rightarrow$ Degrau unitário

DA - Tempo de Reverberação (T60)

Taxa de decaimento:

$$T60 = \ln(1000)\tau T_s$$

$T_s \rightarrow$ Tempo de amostragem

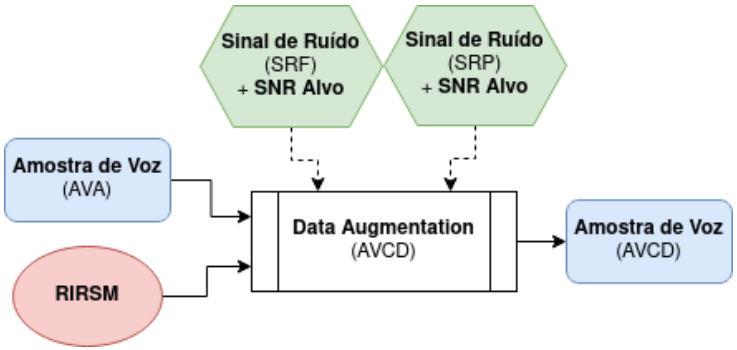
DA do T60:

$$h'_l(t) = h_l(t)e^{-(t-t_o)\frac{\tau-\tau_d}{\tau\tau_d}}$$

RIRSM completa:

$$h'(t) = h'_e(t) + h'_l(t)$$

DA - Amostra de Voz em Campo Distante (AVCD)



Modelo de uma AVCD:

$$S_{cd}[t] = S_a[t] * h[t] + \sum_i n_{pi}[t] * h[t] + n_f[t]$$

$S_a[t] \rightarrow$ Amostra de Voz Anecóica (AVA)
$h[t] \rightarrow$ RIRSM
$n_p[t] \rightarrow$ Sinal de Ruído Pontual (SRP)
$n_f[t] \rightarrow$ Sinal de Ruído de Fundo (SRF)

DA - Amostra de Voz em Campo Distante (AVCD)

Primeira etapa: Adição do SRP

$$S_r[t] = S_a[t] * h[t] + \alpha \text{ offset}(n_{pi}[t] * h[t], o_t)$$

OBS: $SNR_t = SNR(S_r[t], \alpha(n_{pi}[t] * h[t])) \rightarrow$ Razão Sinal-Ruído alvo

$S_a[t] \rightarrow$ Amostra de Voz Anecóica (AVA)
$h[t] \rightarrow$ RIRSM
$n_{pi}[t] \rightarrow$ SRP
$\alpha \rightarrow$ Fator de correção da intensidade de $n_{pi}[t]$ para obter o SNR_t
offset$(X, o_t) \rightarrow$ Deslocamento de X para uma posição dentro do intervalo de $S_a[t]$

DA - Amostra de Voz em Campo Distante (AVCD)

Segunda etapa: Adição do SRF

$$S_{cd}[t] = S_r[t] + \alpha n_f[t]$$

OBS: $SNR_t = SNR(S_{cd}[t], \alpha n_f[t]) \rightarrow$ Razão Sinal-Ruído alvo

$S_r[t] \rightarrow$ Amostra de Voz Reverberada + SRP
$n_f[t] \rightarrow$ SRF

Implementação dos algoritmos

Os algoritmos apresentados foram implementados com a ajuda dos seguintes softwares.

- MATLAB® R2018a
- ITA Toolbox (plugin para MATLAB) [3]

São utilizadas três bases de dados para gerar as RIRSMs e AVCDs.

- Base de amostras de voz anecóicas
- Base de RIRs - Aachen Impulse Response database
- Base de ruídos - MUSAN

Implementação dos algoritmos

Configurações das características desejadas.

Parâmetro	**Faixa**
DRR_{alvo} (dB)	$-6 \leq DRR_{alvo} \leq 18$
$T60_{alvo}$ (s)	$T60_{org} - 1 \leq T60_{alvo} \leq T60_{org} + 1$, onde o limite inferior de $T60_{alvo} = 0.2$
SNR_{alvo}	$3 \leq SNR_{alvo} \leq 20$

Resultados - DRR

Exemplo	Sala RIR	Distância (m)	Amostra de Voz
D1	lecture	7.1	H2-T2
D2	booth	1	H2-T1
D3	office	2	M2-T2

Exemplo	DRR_{org} (dB)	DRR_{alvo} (dB)	DRR_{res} (dB)	ρ_{DRR} (%)
D1	-4,5	10	10	0
D2	4,7	-2	-2	0
D3	0,5	18	18	0

$$\rho_{DRR} = |DRR_{res} - DRR_{alvo}| / DRR_{alvo}$$

Resultados - DRR

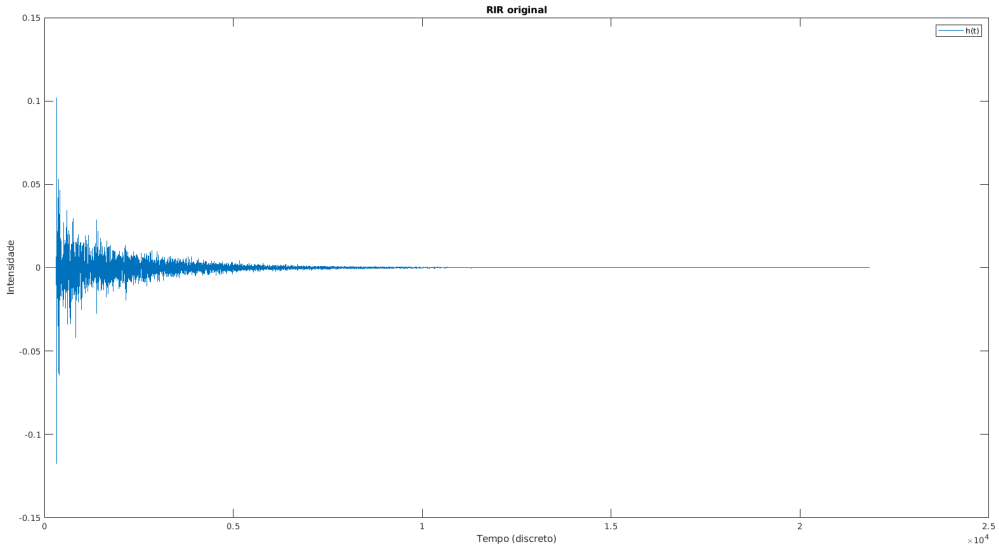
Experimento empírico: sensação subjetiva de “distância”, ordenado de mais para menos distante.

Exemplo	DRR_{org} **(dB)**	DRR_{res} **(dB)**	**Comparação**	**Ordem**
D1	-4,5	10	original	2
D2	4,7	-2	simulado	1
D3	0,5	18	original	3

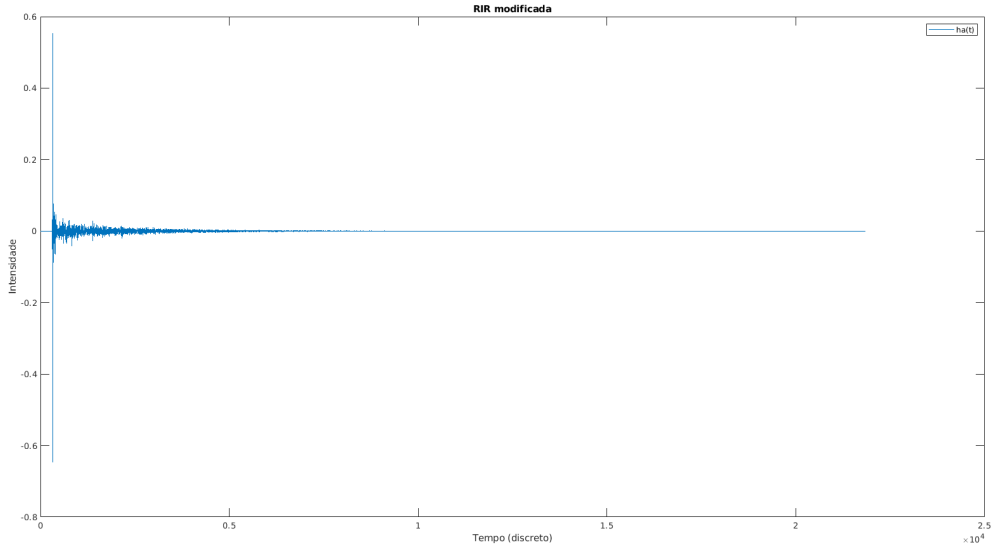
Exemplo D1

Exemplo	DRR_{org} (dB)	DRR_{res} (dB)	Comparação	Ordem
D1	-4,5	10	original	2
D2	4,7	-2	simulado	1
D3	0,5	18	original	3

RIR Original



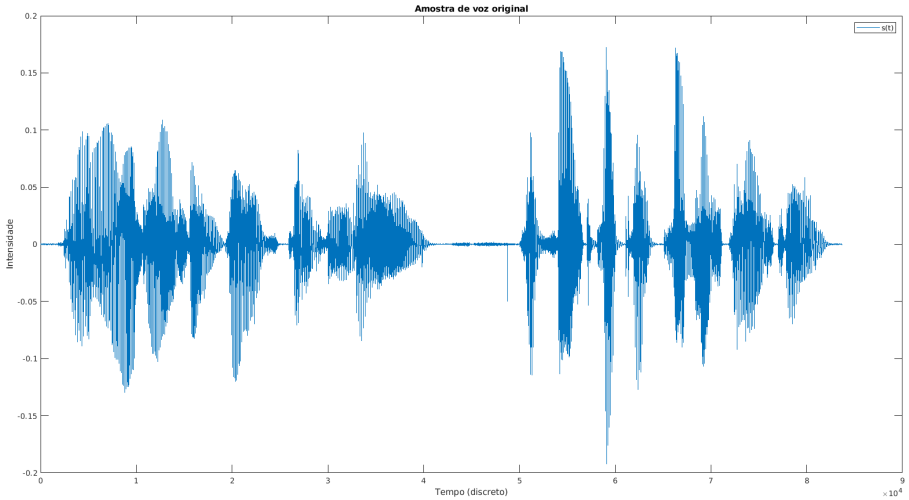
RIR Simulada



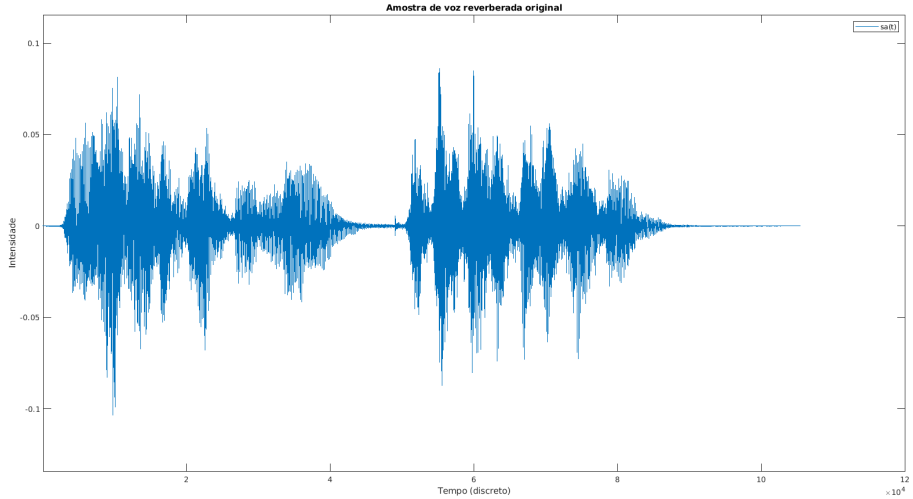
Exemplo D1

Exemplo	DRR_{org} (dB)	DRR_{res} (dB)	Comparação	Ordem
D1	-4,5	10	original	2
D2	4,7	-2	simulado	1
D3	0,5	18	original	3

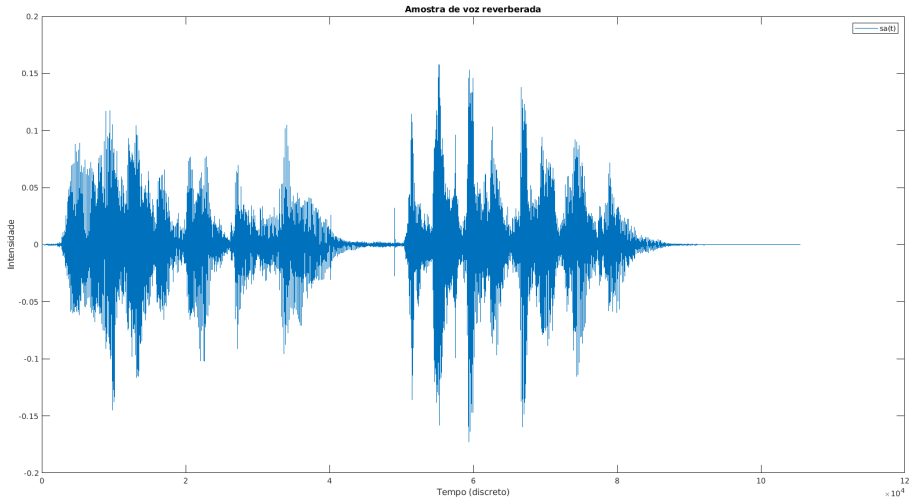
amostra de voz original



amostra de voz reverberada - RIRO



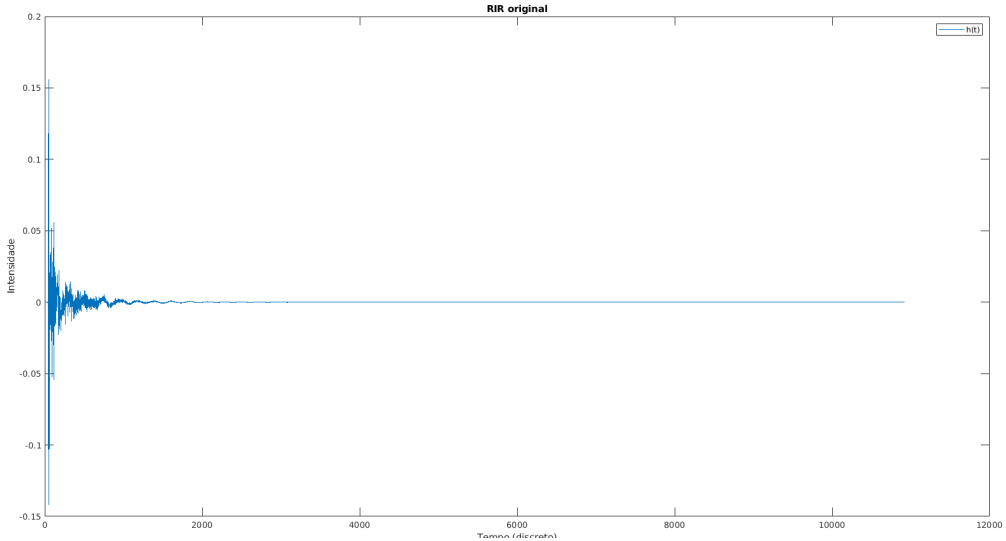
amostra de voz reverberada - RIRSM



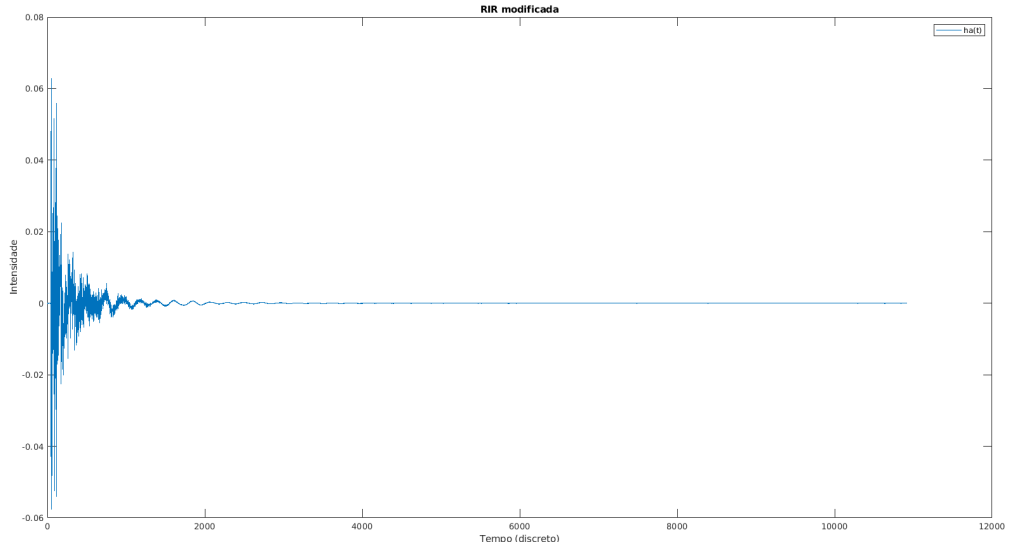
Exemplo D2

Exemplo	DRR_{org} (dB)	DRR_{res} (dB)	Comparação	Ordem
D1	-4,5	10	original	2
D2	4,7	-2	simulado	1
D3	0,5	18	original	3

RIR Original



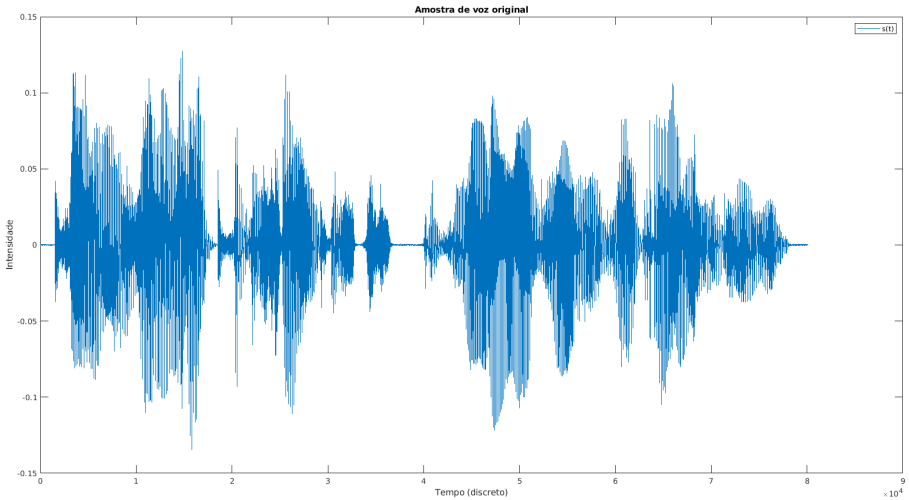
RIR Simulada



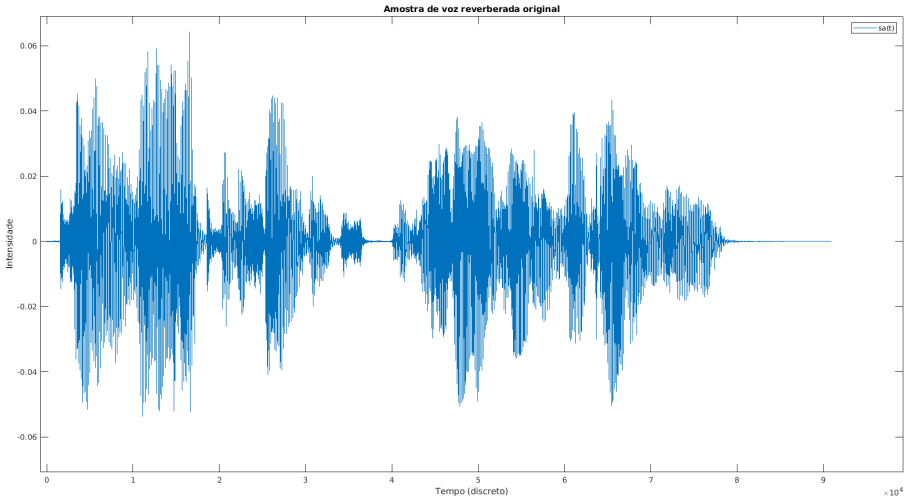
Exemplo D2

Exemplo	DRR_{org} (dB)	DRR_{res} (dB)	Comparação	Ordem
D1	-4,5	10	original	2
D2	4,7	-2	simulado	1
D3	0,5	18	original	3

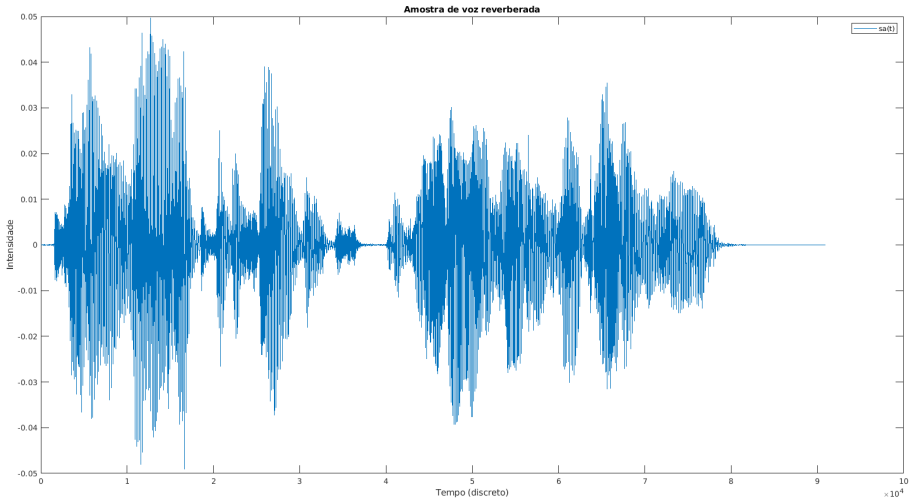
amostra de voz original



amostra de voz reverberada - RIRO



amostra de voz reverberada - RIRSM



Resultados - T60

Exemplo	Sala RIR	Distância (m)	Amostra de Voz
T1	lecture	7.1	M2-T1
T2	booth	1	H1-T2
T3	office	2	H2-T2

Exemplo	$T60_{org}$ (s)	$T60_{alvo}$ (s)	$T60_{res}$ (s)	ρ_{T60} (%)
T1	1,38	1,15	1,01	12.1
T2	1,01	1,88	1,89	0,5
T3	0,75	0,61	0,60	1,6

$$\rho_{T60} = |T60_{res} - T60_{alvo}|/T60_{alvo}$$

Resultados - T60

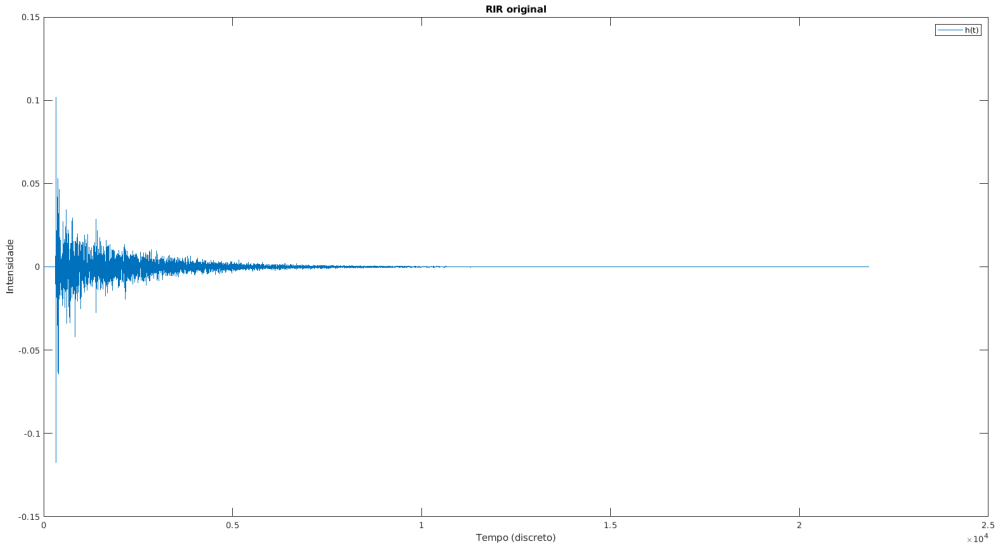
Experimento empírico: sensação subjetiva de “eco”, ordenado de mais para menos ecoante.

Exemplo	$T60_{org}$ **(s)**	$T60_{res}$ **(s)**	**Comparação**	**Ordem**
T1	1,38	1,01	original	2
T2	1,01	1,89	simulado	1
T3	0,75	0,60	original	3

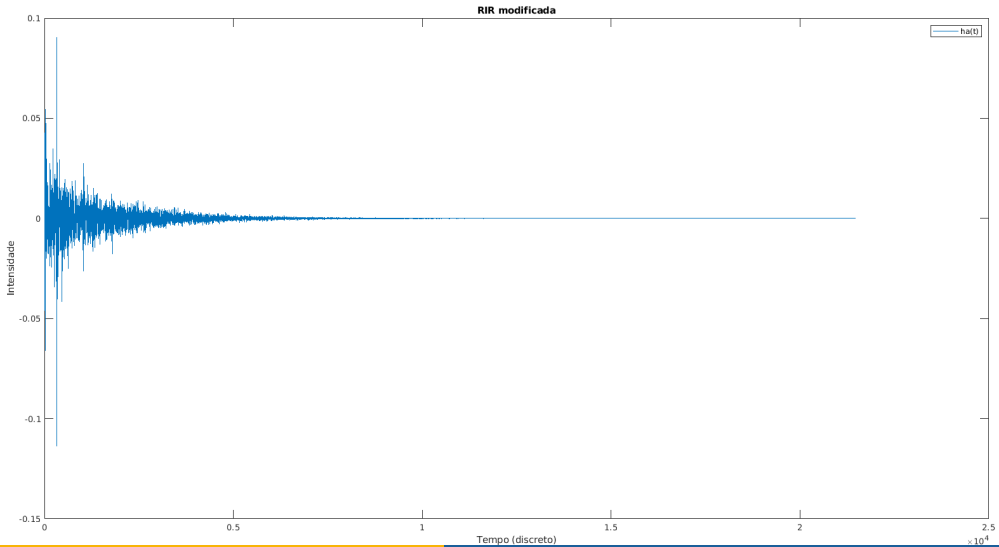
Exemplo T1

Exemplo	$T60_{org}$ (s)	$T60_{res}$ (s)	Comparação	Ordem
T1	1,38	1,01	original	2
T2	1,01	1,89	simulado	1
T3	0,75	0,60	original	3

RIR Original



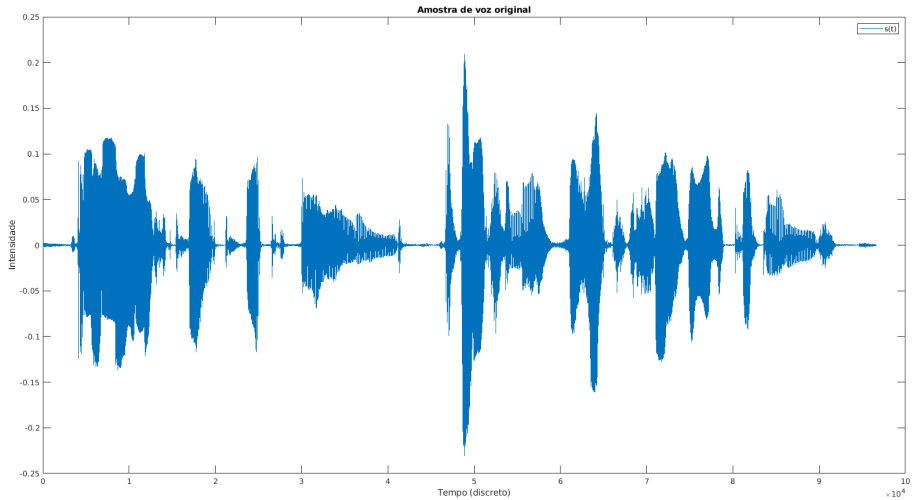
RIR Simulada



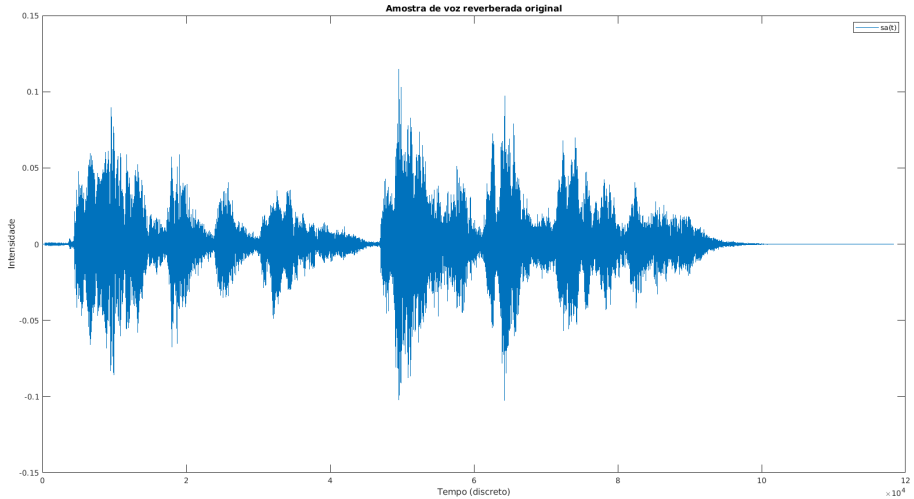
Exemplo T1

Exemplo	$T60_{org}$ (s)	$T60_{res}$ (s)	Comparação	Ordem
T1	1,38	1,01	original	2
T2	1,01	1,89	simulado	1
T3	0,75	0,60	original	3

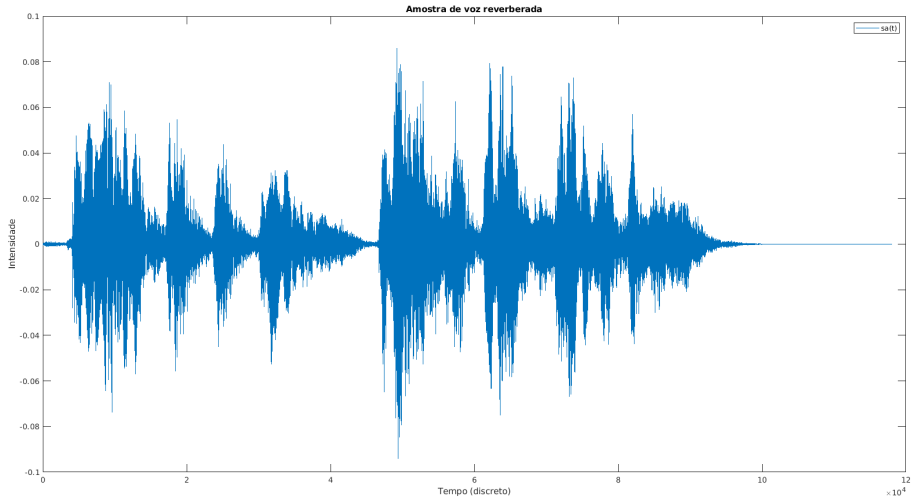
amostra de voz original



amostra de voz reverberada - RIRO



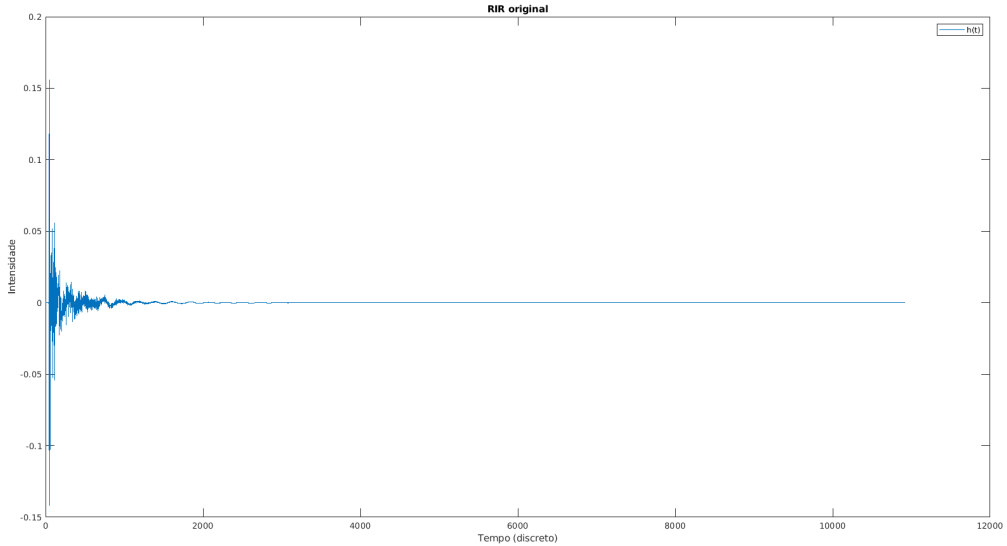
amostra de voz reverberada - RIRSM



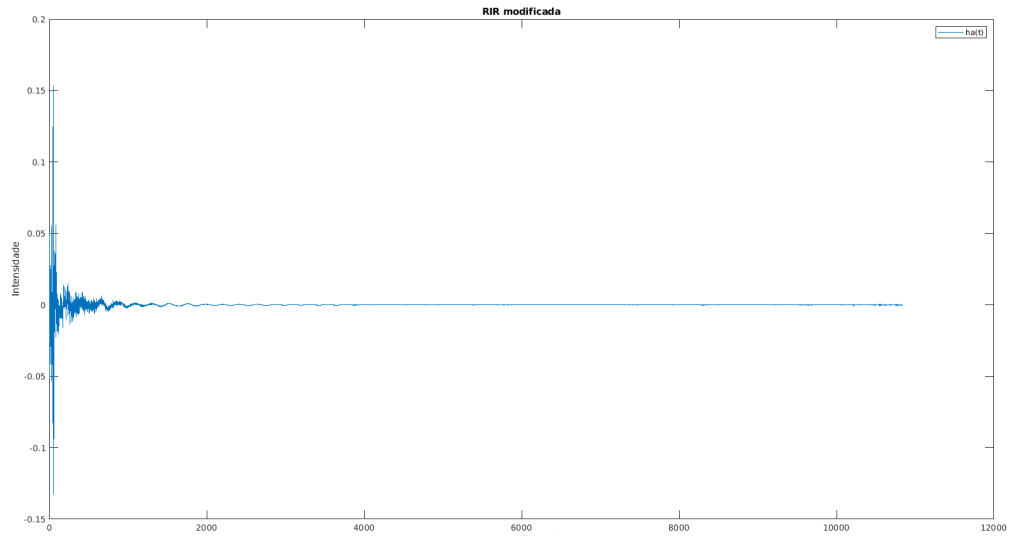
Exemplo T2

Exemplo	$T60_{org}$ (s)	$T60_{res}$ (s)	Comparação	Ordem
T1	1,38	1,01	original	2
T2	1,01	1,89	simulado	1
T3	0,75	0,60	original	3

RIR Original



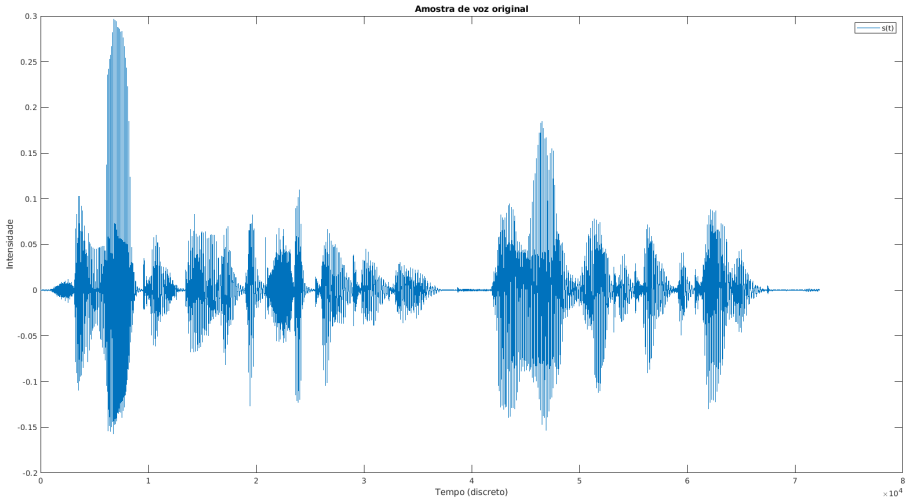
RIR Simulada



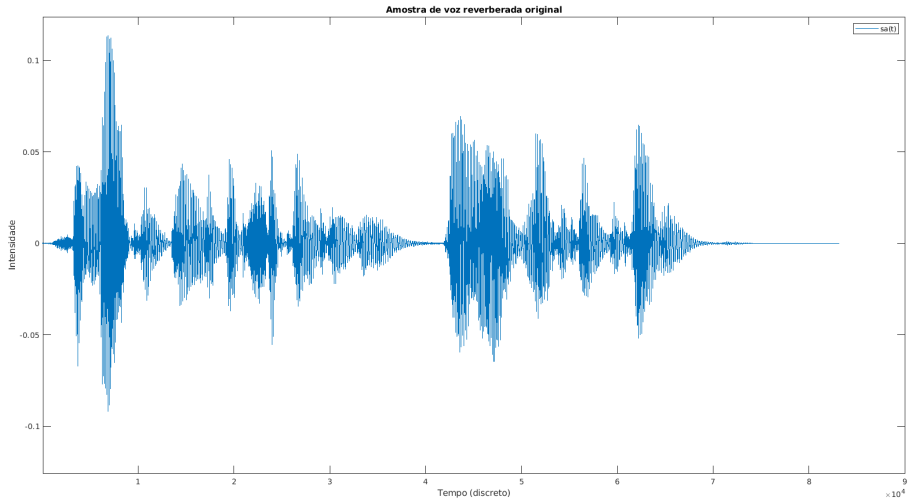
Exemplo T2

Exemplo	$T60_{org}$ (s)	$T60_{res}$ (s)	Comparação	Ordem
T1	1,38	1,01	original	2
T2	1,01	1,89	simulado	1
T3	0,75	0,60	original	3

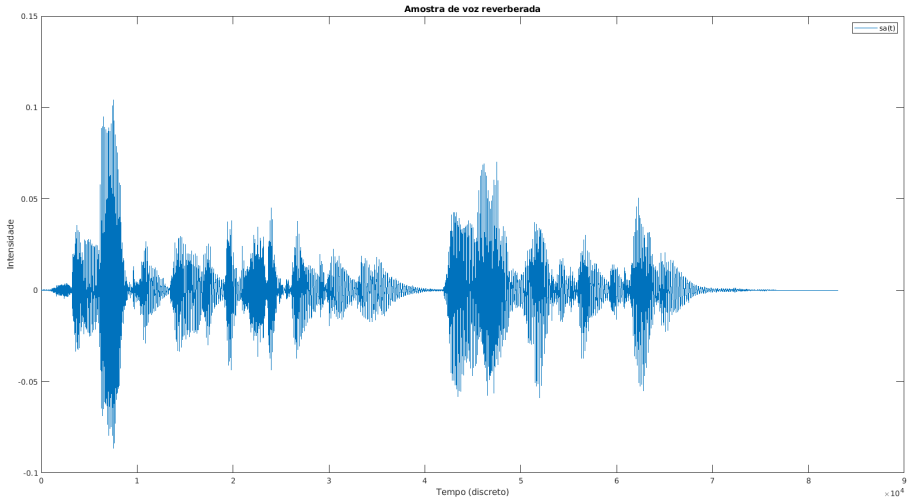
amostra de voz original



amostra de voz reverberada - RIRO



amostra de voz reverberada - RIRSM



Resultados - AVCD

Exemplo	Sala RIR	Distância (m)	AVA	SRP	SRF
N1	lecture	7.1	M2-T1	RP-6	RF-1
N2	booth	1	H2-T1	RP-12	RF-4
N3	office	2	H1-T1	RP-4	RF-4
N4	meeting	1.7	M1-T2	RP-11	RF-2
N5	stairway	1	H2-T1	RP-7	RF-4

Ex.	DRR_{org} (dB)	DRR_{res} (dB)	$T60_{org}$ (s)	$T60_{res}$ (s)	SNR_{alvo}
N1	-4,5	17	1,38	0,56	5
N2	4,7	17	1,01	1,39	10
N3	0,5	14	0,75	0,60	14
N4	6,0	16	0,81	1,16	19
N5	5,0	18	2,70	3,68	3

Resultados - AVCD

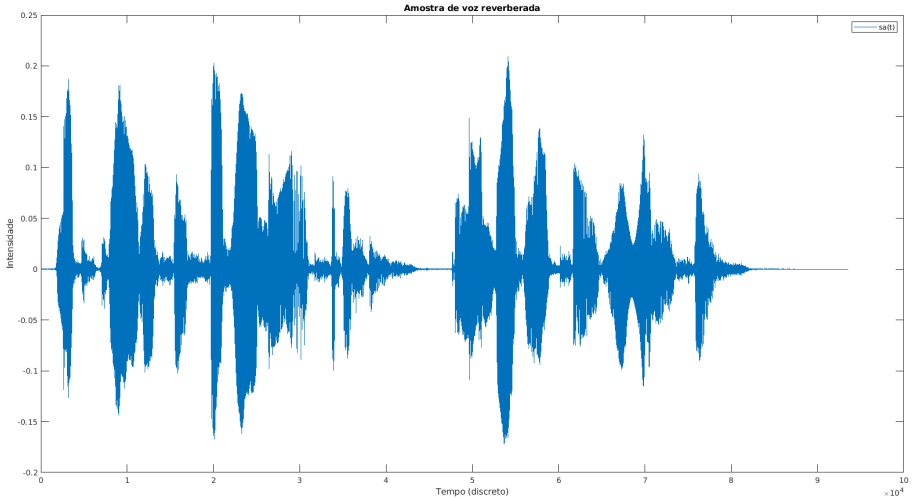
Experimento empírico: análise subjetiva de nível de ruído, ordenado de mais para menos ruidoso.

Exemplo	SNR_{alvo} **(s)**	**Ordem**
N1	5	3
N2	10	4
N3	14	1
N4	19	5
N5	3	2

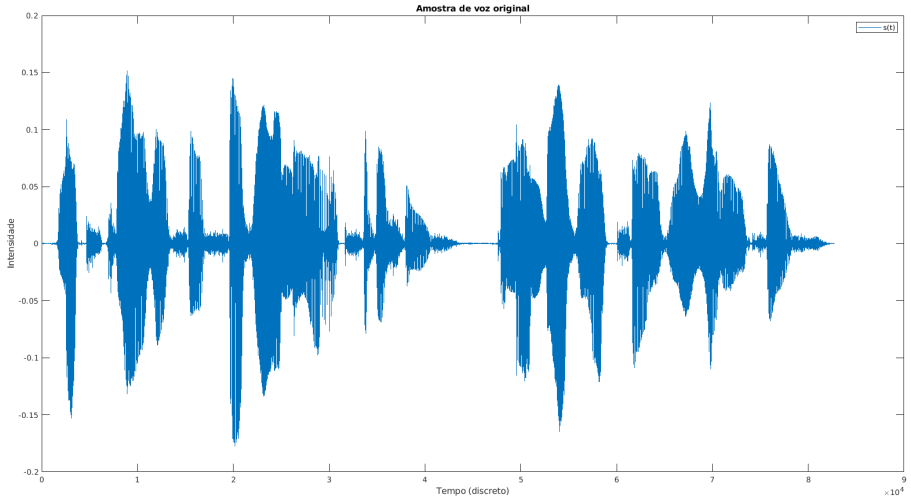
Exemplo N4

Exemplo	SNR_{alvo} (s)	Ordem
N1	5	3
N2	10	4
N3	14	1
N4	19	5
N5	3	2

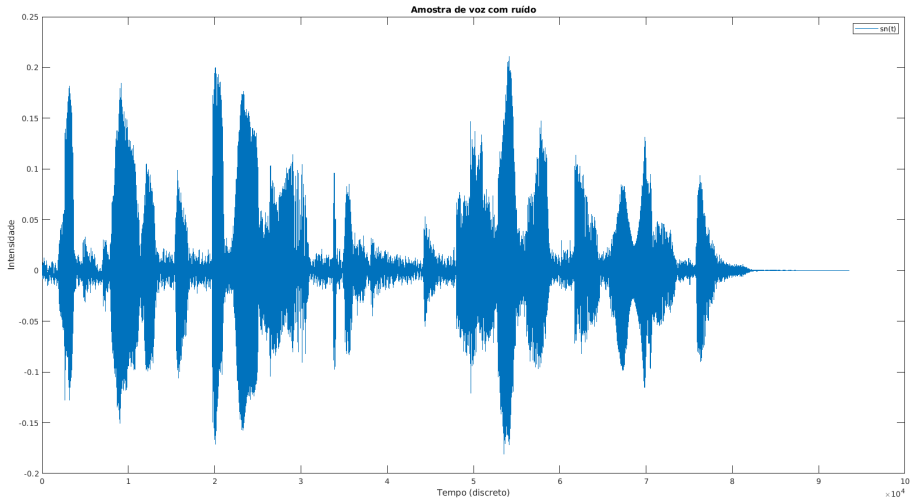
amostra de voz reverberada



amostra de voz original



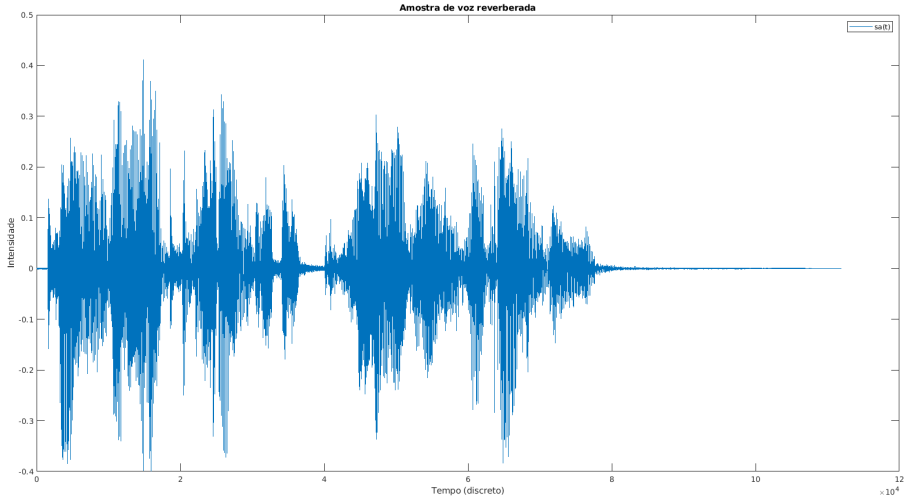
amostra de voz em campo distante



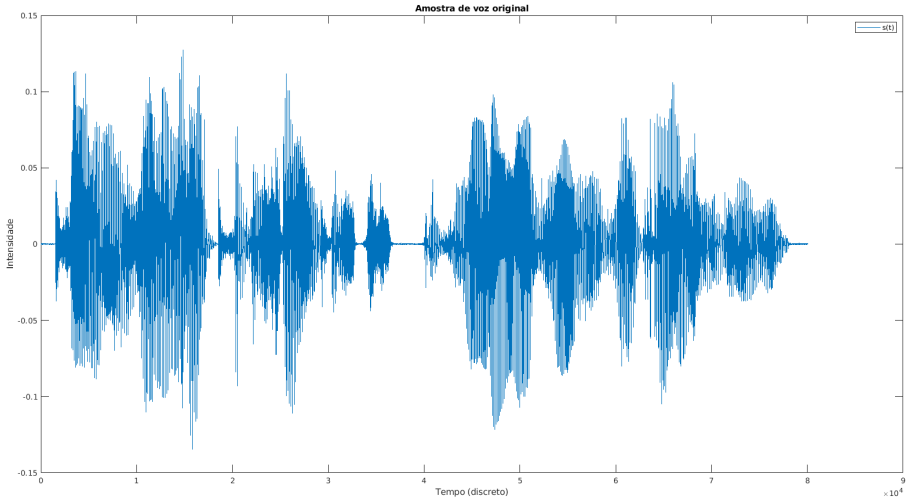
Exemplo N5

Exemplo	SNR_{alvo} (s)	Ordem
N1	5	3
N2	10	4
N3	14	1
N4	19	5
N5	3	2

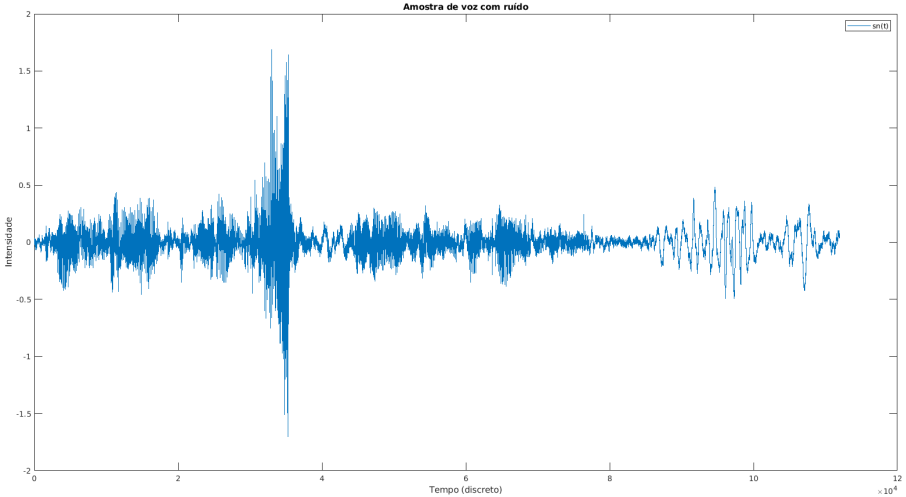
amostra de voz reverberada



amostra de voz original



amostra de voz em campo distante



Conclusões

- Em grande parte, os resultados alcançados estão condizentes com os valores esperados.
- Discrepância nos valores de T60 podem ser explicados pelas diferenças de implementação entre este projeto e [1].
- Avaliação empírica das sensações subjetivas de “distância” e “eco” condizentes com as modificações esperadas.

Trabalhos Futuros

- Implementação de uma metodologia de *data augmentation* de T60 mais próxima à usada no artigo [1].
- Comparação entre as RIRs geradas com a metodologia implementada e RIRs geradas através de programas de simulação acústicas (RAIOS [4]).
- Proposta de um modelo de rede de *deep learning* para estimação de T60 e DRR em AVCDs para observação da eficácia das RIRs como aprimoradoras do treinamento de redes neurais.

Obrigado!

Referências I

[1] N. J. Bryan. “Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation”. Em: **ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. 2020, pp. 1–5. DOI: 10.1109/ICASSP40776.2020.9052970.

[2] T. Ko et al. “A study on data augmentation of reverberant speech for robust speech recognition”. Em: **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. 2017, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.

[3] Marco Berzborn et al. “The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing”. Em: 43th Annual German Congress on Acoustics, Kiel (Germany), 6 Mar 2017 - 9 Mar 2017. 2017. URL: http://publications.rwth-aachen.de/record/687308.

Referências II

[4] Roberto Tenenbaum et al. “Hybrid method for numerical simulation of room acoustics: Part 2-validation of the computational code RAIOS 3”. Em: **Journal of the Brazilian Society of Mechanical Sciences and Engineering** 29 (abr. de 2007). DOI: 10.1590/S1678-58782007000200013.