Universidade Federal
do Rio de Janeiro
Escola Politécnica

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

Projeto de Graduação apresentado ao Curso de Engenharia Eletrônica e de Computação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Mariane Rembold Petraglia

Rio de Janeiro
Julho de 2021

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

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO ELETRÔNICO E DE COMPUTAÇÃO

Autor:

Bruno Machado Afonso

Orientador:

Profª. Mariane Rembold Petraglia, Ph. D.

Examinador:

Prof. José Gabriel Rodríguez Carneiro Gomes, D. Sc.

Examinador:

Prof. Julio Cesar Boscher Torres, D. Sc.

Rio de Janeiro

Julho de 2021

Declaração de Autoria e de Direitos

Eu, *Bruno Machado Afonso* CPF *136.151.347-02*, autor da monografia *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*, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na disciplina de Projeto de Graduação da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

2. Excetuam-se do item 1. eventuais transcrições de texto, figuras, tabelas, conceitos e ideias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.

3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.

4. O autor pode, excepcionalmente, encaminhar à Comissão de Projeto de Graduação, a não divulgação do material, por um prazo máximo de 01 (um) ano, improrrogável, a contar da data de defesa, desde que o pedido seja justificado, e solicitado antecipadamente, por escrito, à Congregação da Escola Politécnica.

5. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.

6. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.

7. Por ser verdade, firmo a presente declaração.

_____

Bruno Machado Afonso

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Escola Politécnica - Departamento de Eletrônica e de Computação
Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária
Rio de Janeiro - RJ CEP 21949-900

AGRADECIMENTO

Sempre haverá. Se não estiver inspirado, aqui está uma sugestão: dedico este trabalho ao povo brasileiro que contribuiu de forma significativa à minha formação e estada nesta Universidade. Este projeto é uma pequena forma de retribuir o investimento e confiança em mim depositados.

RESUMO

O tema de reconhecimento de voz se torna cada vez mais relevante graças ao seu amplo uso tecnológico na sociedade, desde assistentes pessoais em *smartphones*, automação residencial, até autenticação por voz para aplicações de segurança.

Uma das características mais importantes neste tema é a detecção da Resposta ao Impulso de ambientes acústicos (RIR), que representa o modelo acústico do ambiente. A RIR é usada no processamento de áudio para identificação e reconhecimento de fontes sonoras em campo distante, que é formada, no caso do tema de reconhecimento de voz, por uma amostra de voz anecoica convoluída com a RIR, acrescida de um ruído.

Um dos desafios no reconhecimento de voz é a estimação da RIR em um sinal de voz em campo distante. Além das técnicas tradicionais de processamento de sinais, diversas soluções de *deep learning* foram propostas para a estimação da RIR, contudo estas acabam sendo limitadas devido à falta de variedade e quantidade de bases de RIRs disponíveis para treinamento de redes neurais.

Neste contexto, o objetivo deste projeto é de desenvolver um algoritmo, usando técnicas de *data augmentation*, que gera amostras de voz em campo distante (AVCD), construindo assim uma base de dados para uso em treinamentos de soluções de *deep learning*. O algoritmo é composto por dois segmentos de *data augmentation*: o primeiro modifica as características de razão direto-reverberante (DRR) e tempo de reverberação (T60), partindo de RIRs reais, gerando RIRs simuladas (RIRSM); o segundo gera AVCDs, convoluindo amostras de voz anecoicas com as RIRSMs e adicionando ruídos à voz reverberada. Ao final do trabalho, são exibidos exemplos de AVCDs geradas pelo algoritmo proposto, analisando se os dados gerados são válidos para uso em treinamento de redes neurais.

Palavras-Chave: Resposta ao Impulso de sala, *data augmentation*, *deep learning*, reconhecimento de voz.

ABSTRACT

Speech recognition is a very relevant topic in the present days due to it's vast technological usage on modern society from personal assistants on smartphones, residential automated systems to voice authentication for security applications.

One of the most important characteristics on this topic is the Room Impulse Response detection (RIR), which represents the acoustic model of the room. The RIR is used on signal processing to identify and recognize far-field audio sources, which for the speech recognition topic, is the anechoic voice sample convolved with the RIR plus noise signal.

One of the challenges when it comes to speech recognition is to estimate the RIR in a far-field voice sample. Beyond the traditional signal processing algorithms, many deep learning solutions are proposed for the RIR estimation, however they end up with limited results due to the lack of variety e quantity of RIR databases available for training.

In this context, the main objective of this project is to develop an algorithm using data augmentation techniques that will generate far-field voice samples, therefore building a database for deep learning training. The algorithm is composed of two segments: the first modifies the real RIRs characteristics of direct-to-reverberant ratio (DRR) and the reverberation time (T60) generating simulated RIRs (RIRSM); the second generates far-field voice samples using the previously created RIRSMs, anechoic voice samples and noise signals. At the end of this work, examples of the generated far-field voice samples by the algorithm are shown and they are analysed to see if they are valid to be used in neural network training.

Key-words: Room Impulse Response, data augmentation, deep learning, voice recognition.

SIGLAS

AIR - Aachen Impulse Response database

AVA - Amostra de voz anecoica

AVCD - Amostra de voz em campo-distante

AVR - Amostra de voz reverberada

DA - *Data Augmentation*

DL - *Deep Learning*

DRR - Razão Direto-Reverberante

RIR - Resposta ao Impulso de Ambiente Acústico

RIRDA - *Data Augmentation* da Resposta ao Impulso de Ambiente Acústico

RIRO - Resposta ao Impulso de Ambiente Original

RIRSM - Resposta ao Impulso de Ambiente Acústico Simulada

SNR - Razão Sinal-Ruído

SRF - Sinal de ruído de fundo

SRP - Sinal de ruído pontual

T20 - Tempo de Reverberação (queda de 20 DB)

T30 - Tempo de Reverberação (queda de 30 DB)

T60 - Tempo de Reverberação (queda de 60 DB)

UFRJ - Universidade Federal do Rio de Janeiro

VA - Voz anecoica

VR - Voz reverberada

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Neste capítulo, será introduzido os principais tópicos do projeto, além de mostrar sua relevância para o escopo da engenharia moderna e as metodologias que são usadas para alcançar seus objetivos. Ao final é descrito a estrutura organizacional do texto.

1.1 Tema

O tema do trabalho é sobre o estudo de uma forma de simular Respostas ao Impulso de Ambientes Acústicos (RIR) com parametrizações diferentes a partir de amostras de RIR gravadas em ambientes reais, e ainda usar a RIR para gerar amostras de áudio em locais simulados a partir de gravações de voz reais.

1.2 Delimitação

O estudo é focado em inferir uma técnica de reforço de dados tanto em amostras reais de RIR quanto nas gravações de voz. Este trabalho está delimitado em apenas modificar amostras reais de áudio, e não gerar amostras simuladas sem uma gravação de base.

1.3 Justificativa

Com o avanço das tecnologias de automação residencial, assistentes pessoais nos *smartphones* e comunicação *online*, o estudo de técnicas de processamento de

áudio (no caso específico deste trabalho, relacionados a voz), tornou-se mais relevante para a sociedade. Uma das características mais importantes a ser detectada no processamento de áudio é a Resposta ao Impulso de salas, que representa o modelo acústico do ambiente, pois através desta é possível extrair informações pertinentes do local em que o áudio foi gravado e também detectar a posição de fontes sonoras e as isolar para reconhecimento. No âmbito da área de reconhecimento de voz, a fala reverberante, ou seja, o sinal de fala combinado com o modelo acústico do ambiente é um dos desafios encontrados para a detecção da voz, tornando a identificação do RIR de vital importância para o reconhecimento de fala [1].

Junto a isso, houve avanços no âmbito do aprendizado de máquina, fornecendo alternativas para os métodos tradicionais de processamento de áudio [2]. Modelos de arquitetura de redes neurais necessitam de um grande volume de dados para que sejam treinados e aprimorados, e um dos mais recentes desafios nessa área é o fato das bases de RIR não serem extensas, conforme esclarecidas no artigo [3], pois realizar uma grande quantidade de gravações de áudio é uma tarefa de alto custo tanto financeiro e temporal, necessitando de equipamento especializado e diversos locais com características de modelo sonoro diferentes e pessoas diversas para amostras de voz.

1.4 Objetivos

O objetivo deste trabalho é desenvolver um algoritmo capaz de gerar amostras de RIR simuladas para diferentes ambientes a partir de uma RIR real e gerar um banco de dados de amostras de voz convoluídas com as RIR simuladas e com ruídos para uso em treinamento de redes neurais. Dessa forma, têm-se como objetivos específicos:

1. Propor um algoritmo que altere as características da RIR para simular diferentes ambientes com RIR diferentes;

2. Elaborar um algoritmo que faça o acréscimo de ruídos pontuais ou ruídos de fundo em uma amostra de voz;

3. Desenvolver um sistema computacional que aplique ambos os algoritmos an-

teriores em sequência para gerar amostras de voz em ambientes ruidosos.

1.5 Metodologia

Um sinal de voz gravado em um ambiente pode ser interpretado como a junção de três partes: uma amostra de voz pura, sem nenhum fator externo ou reverberação envolvido, convoluída com a Resposta ao Impulso da sala (RIR) onde ocorre a gravação, somada a um sinal de ruído, podendo este ser pontual ou um ruído de ambiente. A RIR representa um modelo acústico do ambiente, que define como um receptor acústico irá receber caso o áudio seja gerado e percebido de dentro deste ambiente. Uma definição de Resposta ao Impulso é a de uma função que registra a pressão sonora temporalmente em um ambiente fechado após uma excitação extremamente curta e cheia de energia (impulso de Dirac).

Neste trabalho é proposta uma forma de gerar RIR simuladas partindo de uma RIR real, ou seja, gravando um áudio que representa um impulso em um ambiente fechado real, e alterando suas propriedades. Reproduz-se o que foi proposto no artigo de data augmentation para respostas ao impulso para estimação do modelo acústico [4], onde são geradas RIRs simuladas, modificando-se as propriedades de Tempo de Decaimento (T60) e de razão entre áudio direto e reverberado (DRR). Através dessas duas propriedades, define-se praticamente todas as RIRs possíveis de serem gravadas artificialmente.

Para gerar as amostras de vozes reverberadas que compõe a base de dados, acompanha-se o que é proposto no artigo de estudo de data augmentation em vozes reverberadas [5], onde são convoluídos sinais de voz anecoicos com as RIRs simuladas que foram geradas anteriormente. Além disso, é acrescentado a essa sinal de voz reverberado ruídos diversos, que são caracterizados de duas formas: ruídos pontuais e de ambiente. Os ruídos pontuais são amostras de áudio curtas que podem ser introduzidos em qualquer momento da fala; já os ruídos de ambiente são sons constantemente presentes ao fundo da gravação para simular um ambiente externo. Os ruídos foram extraídos da biblioteca MUSAN [6].

Através desses dois passos, são gerados vários sinais de vozes reverberados artificialmente. A simulação da RIR tem por objetivo colocar a amostra de voz em

vários ambientes fechados, e a inclusão de ruídos ajudam drasticamente no treinamento de redes neurais, impedindo que as redes fiquem viciadas em características muito específicas da fala durante o treinamento, uma vez que tendem a simular os fatores externos que podem estar envolvidos em uma gravação real.

1.6 Descrição

O capítulo 2 apresenta uma breve análise sobre as principais aplicações do tema e os desafios que este trabalho auxilia na solução.

No capítulo 3 será descrito a metodologia usada para fazer a *data augmentation* de uma RIR já existente.

No capítulo 4 explica-se a metodologia usada para gerar sinais de voz aleatórios a partir de RIRs simuladas anteriormente e da adição de ruídos pontuais ou de fundo.

O capítulo 5 é focado em exibir os resultados obtidos através dos métodos anteriores e demonstrar sua eficácia.

Por fim, o capítulo 6 trata das conclusões que são tiradas sobre este projeto, além de mostrar trabalhos futuros.

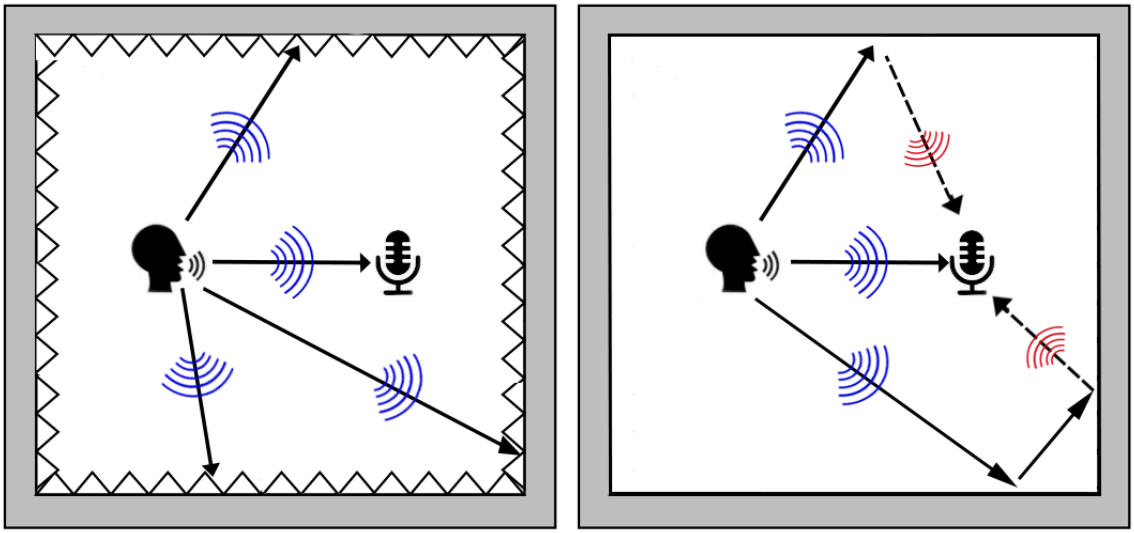
Capítulo 2

Análise de Fontes Sonoras e seus Desafios

Este capítulo é dedicado à introdução do leitor ao principal tópico de estudo do projeto e assim mostrar algumas aplicações onde este é usado, além de apresentar os desafios relacionados à estas aplicações.

2.1 Resposta ao Impulso de Ambiente Acústico e suas Aplicações

Dentre os diversos tópicos na grande área de estudo de sinais de áudio, destaca-se a detecção e reconhecimento de fontes acústicas no espaço físico. Um caso específico deste tópico é sobre sinais de voz gravados em ambientes fechados, onde um ou mais microfones são posicionados na sala afastados da fonte sonora, normalmente uma pessoa que performa a gravação. Estes sinais são corrompidos pela reverberação do ambiente, que surge a partir da sobreposição da onda sonora anecoica que chega ao microfone com a onda sonora atenuada e refletida nas paredes do ambiente fechado.



(a) Sala anecoica (b) Sala reverberante

Figura 2.1: Representação de uma sala anecoica e reverberante

Observa-se na figura 2.1 uma representação de uma sala anecoica, onde o único áudio capturado pelo microfone é a onda sonora direta enviada pela fonte, sem nenhuma reflexão do ambiente; já na sala reverberante, nota-se que o áudio capturado será uma combinação da onda sonora direta com as refletidas nas paredes. Este sinal reverberado pode ser modelado da seguinte forma:

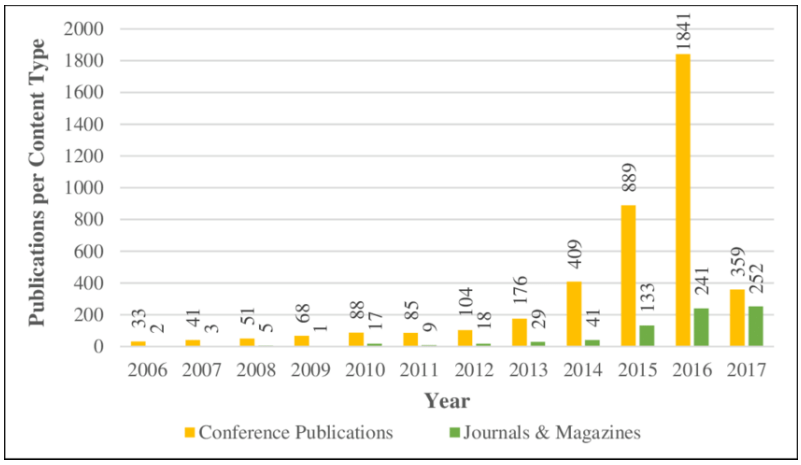
$$Y(t) = s(t) * h(t) + n(t) \tag{2.1}$$

Onde $Y(t)$ representa o sinal de voz em campo distante, $s(t)$ o sinal de voz anecoico, $h(t)$ a RIR e $n(t)$ o sinal de ruído que pode estar presente no ambiente. Dessa forma, é possível inferir que a RIR representa o modelo acústico de uma sala, para uma determinada combinação de fatores do ambiente, incluindo: temperatura e umidade relativa do ar, pressão atmosférica, material das paredes e posicionamento de móveis. Reverberação causa degradação do sinal de voz, levando à perda de clareza na comunicação [7] e à redução da performance de sistemas de reconhecimento de voz [8]. Este problema demonstra a necessidade de identificar dinamicamente o modelo acústico do ambiente para que possam ser mitigadas as perdas nas amostras de voz gravadas e assim facilitar os algoritmos que usam esses sinais.

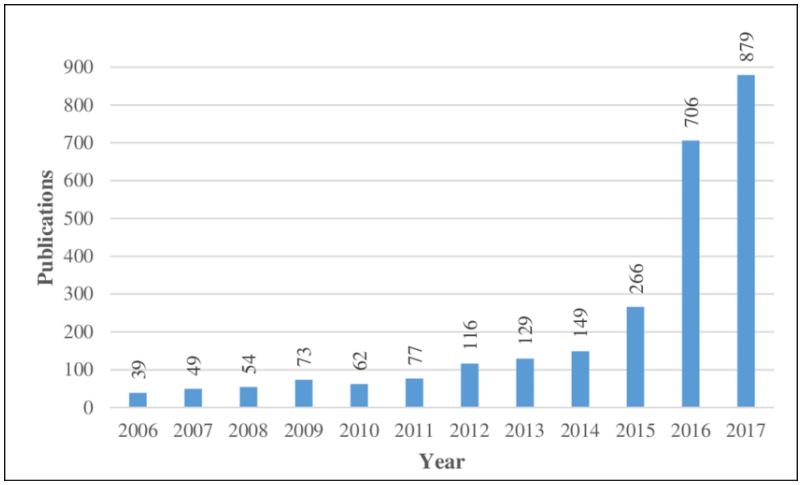
Este projeto é focado no estudo de uma forma de gerar RIRs simuladas a partir de RIRs reais devido à sua importância para diversas aplicações usadas

atualmente na indústria. Uma de suas aplicações é na análise e desenvolvimento de algoritmos de reconhecimento de voz robusta [8], onde é necessário inferir a RIR para que possa ser feita a comparação entre a supressão da reverberação ideal com a cega. Outra aplicação da RIR é para desenvolvimento de algoritmos para localização e separação de fontes sonoras [9], onde as RIRs são usadas no auxílio do mapeamento acústico de ambientes reverberante através de algoritmos de separação de fonte às cegas.

2.2 Desafios correlacionados à RIR



(a) Publicações por ano - IEEE



(b) Publicações por ano - Springer®

Figura 2.2: Gráficos de quantidade de artigos publicados por ano relacionados com *Deep Learning*

É possível notar um recente aumento de pesquisas relacionadas à área de aprendizado de máquina no meio científico (especialmente sobre a subdivisão de *Deep Learning*) [10]. Observando a figura 2.2 [11] nota-se que após 2015, houve um aumento considerável de publicações em conferências do IEEE e o mesmo pode ser constatado para artigos em livros publicados pela editora Springer®. Muitas dessas publicações são dedicadas para áreas de pesquisa relacionadas com áudio [10, 12, 13]; de acordo com o artigo [14], aproximadamente 20% das publicações são voltadas para o tópico de reconhecimento de voz usando técnicas de *Deep Learning* em suas metodologias.

Um dos maiores desafios enfrentados ao utilizar técnicas de *Deep Learning* é de obter uma grande quantidade de dados para treinamento. Pode-ser observar um exemplo disso em [3, 15], onde os autores precisaram agrupar dados de mais de 5 bases contendo RIRs para que fosse possível treinar e avaliar suas redes. No caso de bases de dados que envolvem RIRs, o motivo de não existir uma alta variedade de dados é devido às dificuldades de realizar a gravação dos áudios [16]. Para gerar o impulso sonoro, é necessário de uma fonte sonora (por exemplo, um alto-falante) capaz de realizar uma varredura de senos com o mínimo de distorção possível, ou usar um equipamento para iniciar um som de decaimento rápido e de alta intensidade (por exemplo, um balão estourando) [17]. A gravação do impulso requer microfones que estejam dentro de uma câmara anecoica, capacitando assim o microfone de gravar apenas o som vindo direto da fonte sonora, e não as ondas que são refletidas nas paredes. Não menos importante, para aumentar a quantidade de amostras na base, deve-se não só gravar o áudio não só em diferentes posições no ambiente e distâncias da fonte-microfone, como também preparar estes mesmos procedimentos em ambientes diferentes, levando ao transporte de diversos equipamentos especializados entre localizações físicas.

2.3 *Data Augmentation*

Data Augmentation representa um conjunto de técnicas que são usadas em dados já existentes com o intuito de gerar cópias modificadas que se enquadram para uma determinada aplicação. No contexto de *Deep Learning*, essas técnicas tornam-

se vitais para incrementar artificialmente bases de dados para treinamento que não possuem uma alta variedade de amostras, e isso inclui dados de áudio [18, 19].

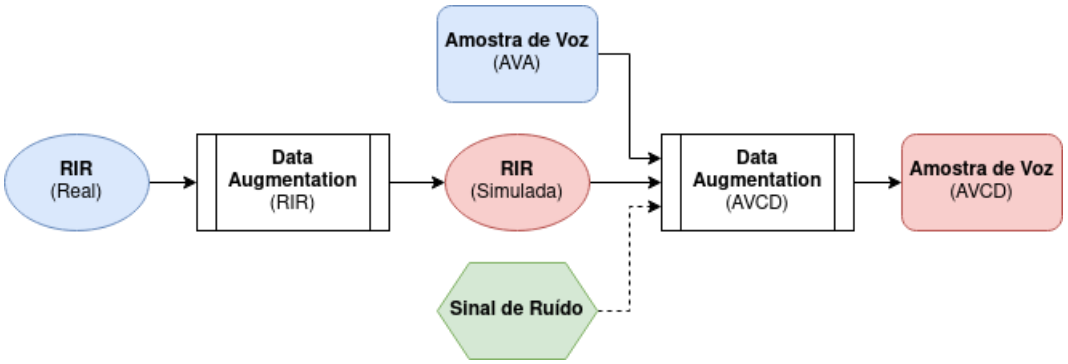


Figura 2.3: Fluxo geral de procedimentos para gerar sinais de voz reverberantes.

No escopo deste trabalho, de acordo com a figura 2.3 usaremos duas técnicas de *Data Augmentation* para gerar amostras de voz reverberantes. Uma das técnicas é voltada para simulação de RIRs, que altera suas propriedades para que possa ser simulado diferentes condições e posições em um determinado ambiente; já a outra técnica é desenvolvida para simular ruídos no ambiente, que adiciona tanto ruídos pontuais em um trecho da amostra de voz, sendo este convoluído com a RIR real ou simulada, quanto um ruído de fundo.

Capítulo 3

Data Augmentation da Resposta ao Impulso do Ambiente

Este capítulo é dedicado ao desenvolvimento do primeiro algoritmo de *Data Augmentation*, onde são geradas RIRs simuladas (RIRSM) a partir de RIRs originais (RIRO). São observados os parâmetros de razão Direto-Reverberante (DRR) e tempo de reverberação (T60), que são inferidos com base em uma RIRO e que serão manipulados pelo algoritmo para gerar RIRs que vão representar modelos acústicos diferentes.

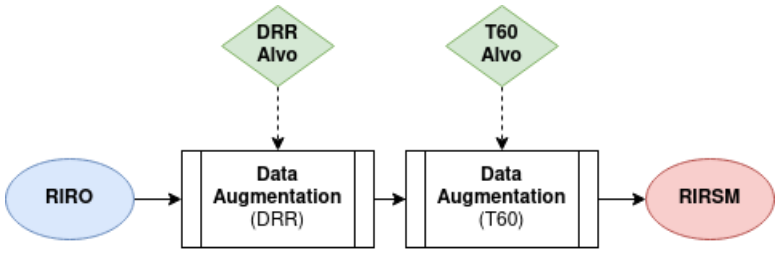


Figura 3.1: Fluxo de procedimentos para gerar a RIRSM.

Este trabalho é uma implementação dos passos demonstrados no artigo [4]. A figura 3.1 especifica o fluxo de procedimentos implantados por este algoritmo, onde a DRR e T60 alvos são os valores escolhidos pelo usuário que determinam as características da RIRSM.

Antes de explicar os métodos usados, é necessário definir duas funções;

$$h_e(t) = \begin{cases} h(t), & t_d - t_0 \leq t \leq t_d + t_0 \\ 0, & \text{caso contrário,} \end{cases} \tag{3.1}$$

$$h_l(t) = \begin{cases} h(t), & t < t_d - t_0 \\ h(t), & t > t_d + t_0 \\ 0, & \text{caso contrário,} \end{cases} \tag{3.2}$$

onde t representa o tempo discreto, t_d o tempo que as ondas sonoras diretas, ou seja, sem reflexão, levam da fonte até o destino de gravação, t_0 a janela de tolerância, neste caso definida com o valor 2,5 ms [4], $h(t)$ uma RIR, $h_e(t)$ a resposta adiantada e $h_l(t)$ a resposta atrasada. Neste algoritmo, t_d é determinado da forma:

$$\begin{cases} t_d = t_{max}, \\ t_{max}, \text{ onde } h(t_{max}) = max(h(t)) \end{cases} . \tag{3.3}$$

3.1 Razão Direto-Reverberante (DRR)

A DRR representa a razão entre a energia sonora da resposta ao impulso que atinge o alvo diretamente e a energia reverberante, ou seja, que é refletida pelas paredes do ambiente fechado. Este parâmetro é definido da forma:

$$DRR_{dB} = 10 \log_{10} \left(\frac{\sum_t h_e^2(t)}{\sum_t h_l^2(t)} \right). \tag{3.4}$$

Para obter a DRR alvo desejada, aplica-se um fator de ganho escalar α na resposta adiantada $h_e(t)$. De acordo com [4], para evitar descontinuidades durante o cálculo do fator α, refatora-se a resposta adiantada em duas parcelas, uma que representa a janela direta no pico de intensidade de $h(t)$ e outra que representa uma janela de resíduo de $h_e(t)$ formando, assim,

$$h'_e(t) = \alpha w_d(t) h_e(t) + [1 - w_d(t)] h_e(t), \tag{3.5}$$

onde $w_d(t)$ representa uma janela de Hann de duração de 5 ms, pois a janela de tolerância $t_0 = 2,5$ ms. Substituindo na equação 3.4 $h'_e(t)$ por $h'_e(t)$ e ajustando com a expressão 3.5, obtem-se a seguinte equação quadrática;

$$\alpha^2 \sum_t w_d^2(t) h_e^2(t) + 2\alpha \sum_t [1 - w_d(t)] w_d(t) h_e^2(t) + \sum_t [1 - w_d(t)]^2 h_e^2(t) - 10^{DRR_{dB}/10} \sum_t h_l^2(t) = 0 \quad (3.6)$$

O α desejado será a raiz de maior valor. Uma ressalva deste procedimento é que se deve atentar para não escolher um DRR_{dB} que não seja muito menor que o original, pois após a transformação do $h_e(t)$ para $h'_e(t)$, dependendo do valor de α, é possível incidir em um caso onde $max(h'_e(t)) < max(h_l(t))$, tornando a RIRSM impraticável.

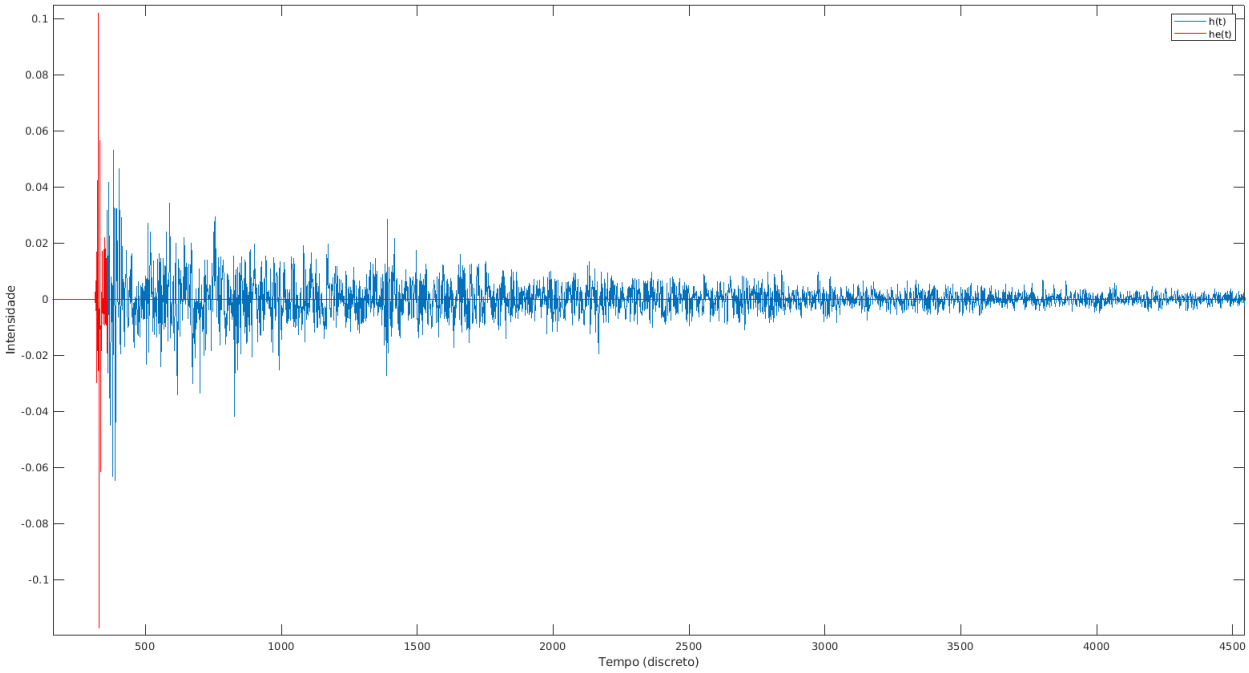


Figura 3.2: Um exemplo de $h(t)$ com $h_e(t)$, onde é feita a DA do DRR, marcado em vermelho.

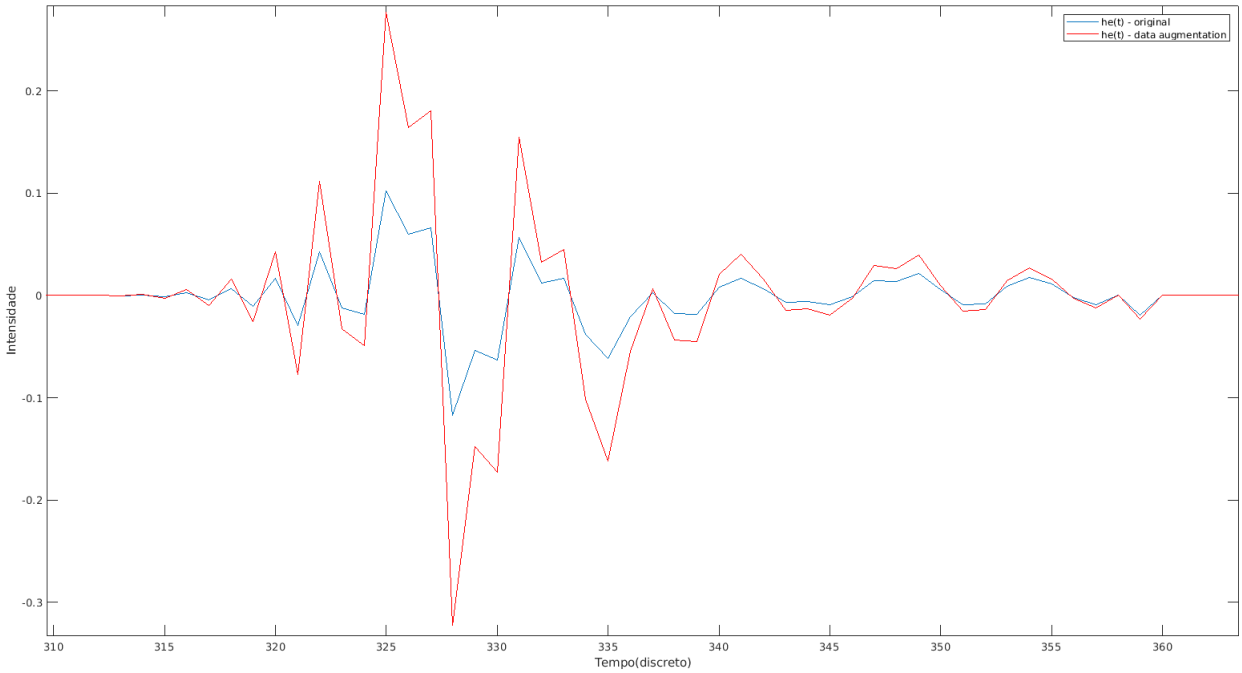


Figura 3.3: Um exemplo de $h_e(t)$ com e sem DA de DRR, original em azul ($DRR = -4,5$) e com DA em vermelho ($DRR = 4$).

3.2 Tempo de Reverberação (T60)

O T60, definido na equação 3.7 representa a duração de tempo que leva para a energia sonora da RIR no alvo decair 60 dB comparado à sua intensidade máxima. Geralmente, devido à dificuldade de medir uma queda de 60 db, o fator medido é o T20 ou o T30 e depois multiplica-se os seus valores por 3 e 2, respectivamente, para obter o T60.

$$\begin{cases} t_i, \text{ onde } h(t_i) = max(h(t)) \\ t_f, \text{ onde } 10\log_{10}\left(h^2(t_i) - h^2(t_f)\right) = 60\text{dB} \\ \text{T60} = t_f - t_i. \end{cases} . \tag{3.7}$$

Para realizar modificações na RIR, é necessário modelar a resposta atrasada; de acordo com [4] um modelo normalmente usado é de um ruído gaussiano exponencialmente decadente, acrescido de um ruído de chão,

$$h_m(t) = Ae^{-(t-t_o)/\tau}n(t)u(t-t_o) + \sigma n(t), \tag{3.8}$$

onde A representa o ganho da resposta ao impulso, τ a taxa de decaimento, σ_m o ganho do ruído de chão, $n(t)$ um ruído gaussiano padrão (média nula e desvio padrão unitário), t_o o valor temporal onde a resposta atrasada tem o seu primeiro valor não nulo e $u(t)$ um degrau unitário. Neste trabalho, diferente da implementação do algoritmo em [4], é considerado apenas a taxa de decaimento do espectro de frequência por completo da RIR, ao invés de dividi-la em subbandas e analisar diferentes taxas para cada.

Os parâmetros A, τ e σ são estimados de acordo com o padrão definido na ISO 3382-1 [20]. Seja $T60_d$ o valor de T60 alvo para DA, é possível inferir a taxa de decaimento através da equação

$$T60_d = \ln(1000)\tau_d T_s, \tag{3.9}$$

onde τ_d representa a taxa de decaimento alvo e T_s o intervalo de amostragem. A DA do tempo de reverberação é feita multiplicando-se $h_l(t)$ pela exponencial

$$h'_l(t) = h_l(t)e^{-(t-t_o)\frac{\tau-\tau_d}{\tau\tau_d}}. \tag{3.10}$$

Por fim, a RIRSM completa, $h'(t)$, pode ser representada pela equação

$$h'(t) = h'_e(t) + h'_l(t). \tag{3.11}$$

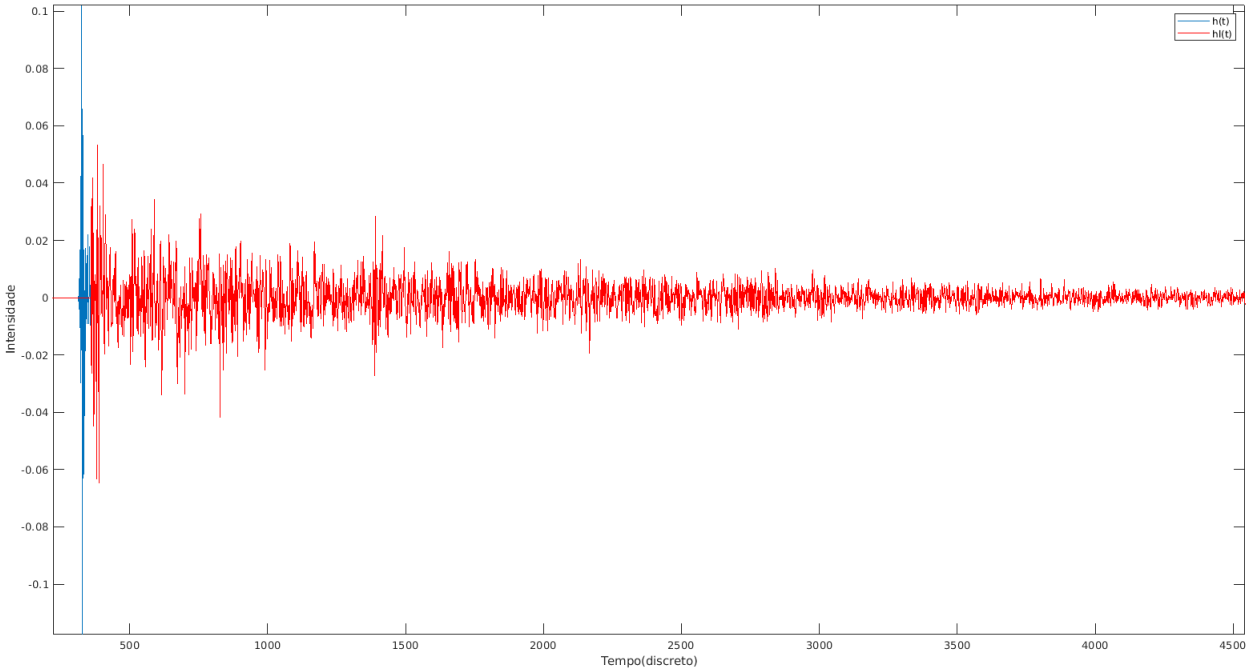


Figura 3.4: Um exemplo de $h(t)$ com $h_l(t)$, onde é feita a DA do T60, marcado em vermelho.

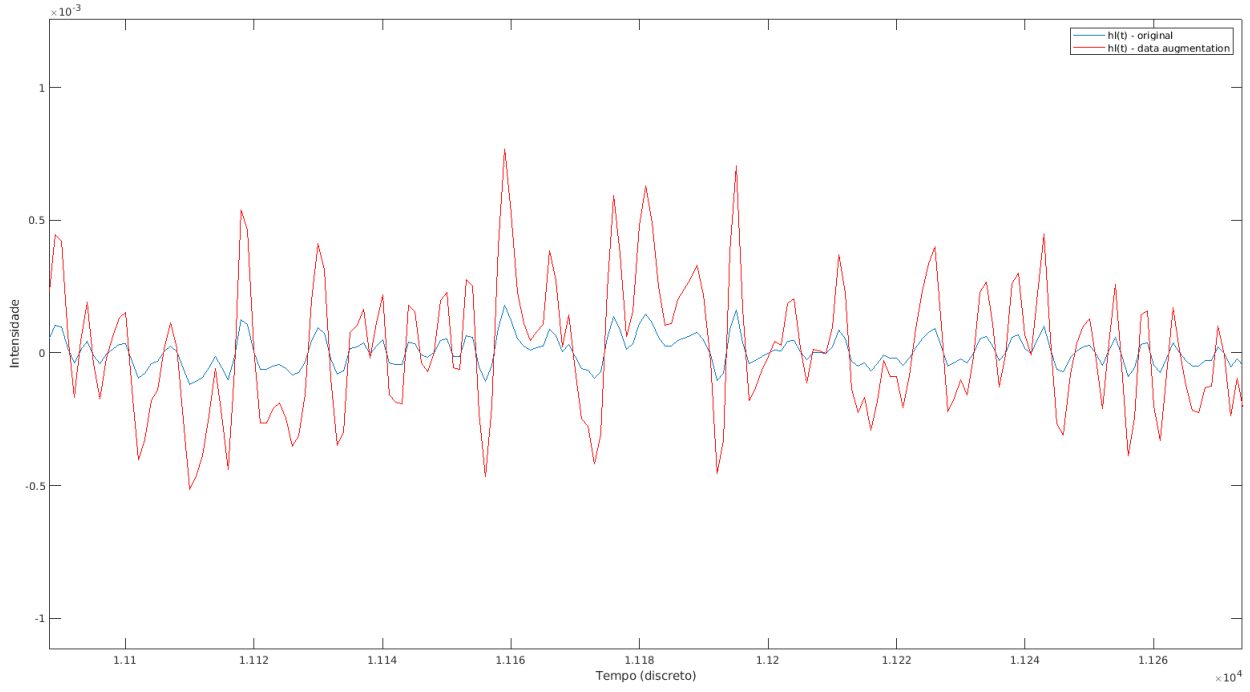


Figura 3.5: Um exemplo de um trecho ao final de $h_l(t)$ com e sem DA de T60, original em azul ($T60 = 1,38$) e com DA em vermelho ($T60 = 2,6$).

Capítulo 4

Desenvolvimento de Sinais de Voz Reverberadas Simuladas com Ruídos

Este capítulo é dedicado ao desenvolvimento do segundo algoritmo de *Data Augmentation*, onde são geradas as amostras de voz reverberadas em campo-distante (AVCDs) usando: amostras de voz anecoicas (AVAs), RIRSM, sinais de ruído pontuais (SRPs) e de fundo (SRFs).

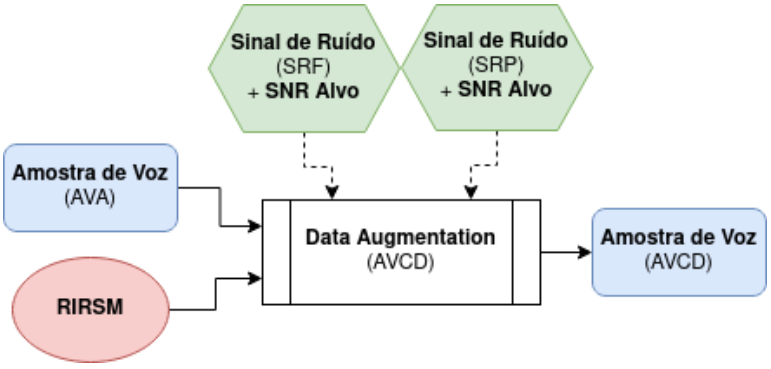


Figura 4.1: Fluxo de procedimentos para gerar a AVCD.

Este trabalho é uma implementação dos passos demonstrados no artigo [5]. A figura 4.1 especifica o fluxo de procedimentos implantados por este algoritmo, onde o SRP, SRF e SNR alvos são aleatoriamente escolhidos, dentro de uma base de dados de ruídos e uma faixa de valores definidas pelo usuário, que determinam as características da AVCD.

4.1 Simulação de fala em campo distante

Sinais de voz em campo-distante são tipicamente compostos por uma combinação de VR, SRP (assumindo que a fonte do ruído pontual encontra-se no mesmo ambiente da VR) e SRF (assumindo que não é afetado pelo modelo acústico do ambiente). É possível modelar uma AVCD conforme a equação

$$S_{cd}[t] = S_a[t] * h[t] + \sum_i n_{pi}[t] * h[t] + n_f[t], \tag{4.1}$$

onde $S_{cd}[t]$ representa a AVCD, $S_a[t]$ a AVA, $h[t]$ a RIR, $n_{pi}[t]$ o i-ésimo SRP e $n_f[t]$ o SRF. Neste trabalho, diferente da implementação do algoritmo em [5], é considerado apenas uma única RIR para gerar a AVCD, ou seja, os ruídos pontuais são convoluídos com a mesma RIR que é usada para a fonte de voz.

Abaixo segue o algoritmo que é usado para gerar sinais de voz em campo-distante simuladas.

Input: fl_p : Flag de inclusão de ruído pontual

Input: fl_g : Flag de inclusão de ruído de fundo

Input: m : Quantidade de ruídos pontuais

Input: SNR_{up} : Limite superior de SNR

Input: SNR_{dw} : Limite inferior de SNR

$S_r[t] \leftarrow S_a[t] * h[t]$: Convolução da RIR com AVA

if $fl_p = true$ **then**

  for $i = 1$ *até* m **do**

    Escolha aleatória de um ruído pontual $n_{pi}[t]$ da biblioteca de ruído.

    Escolha aleatória de uma SNR Alvo SNR_t compreendida dentro do intervalo $[SNR_{dw}, SNR_{up}]$.

    Dedução do fator α a partir da SNR_t para corrigir a intensidade de $n_{pi}[t]$.

    Escolha aleatória de offset o_t compreendida dentro do intervalo $(0, \text{duração(t)})$.

    $S_r[t] \leftarrow S_r[t] + \alpha \text{offset}(n_{pi}[t] * h[t], o_t)$: Adição de SRP na AVR.

  end

end

if $fl_g = true$ **then**

  Escolha aleatória de um ruído de fundo $n_f[t]$ da biblioteca de ruído.

  Escolha aleatória de uma SNR Alvo SNR_t compreendida dentro do intervalo $[SNR_{dw}, SNR_{up}]$.

  Dedução do fator α a partir da SNR_t para corrigir a intensidade de $n_f[t]$.

  Estender ou encurtar $n_f[t]$ até que duração$(n_f[t])$ = duração$(S_r[t])$

  $S_r[t] \leftarrow S_r[t] + \alpha n_f[t]$: Adição de SRF na AVR.

end

Algoritmo 1: Procedimentos para gerar AVCD

Neste trabalho, o algoritmo de geração de AVCD usa as RIRSM geradas através do primeiro algoritmo, diferente do que foi implantado em [5], onde foram geradas RIRs de forma completamente digital [21], ou seja, sem usar RIRs reais como base para *Data Augmentation*. Nota-se também que o algoritmo permite habilitar ou não o uso de cada tipo de ruído para que possa aumentar a variedade de dados gerados, além de acomodar mais propósitos de treinamentos de *Deep Learning*.

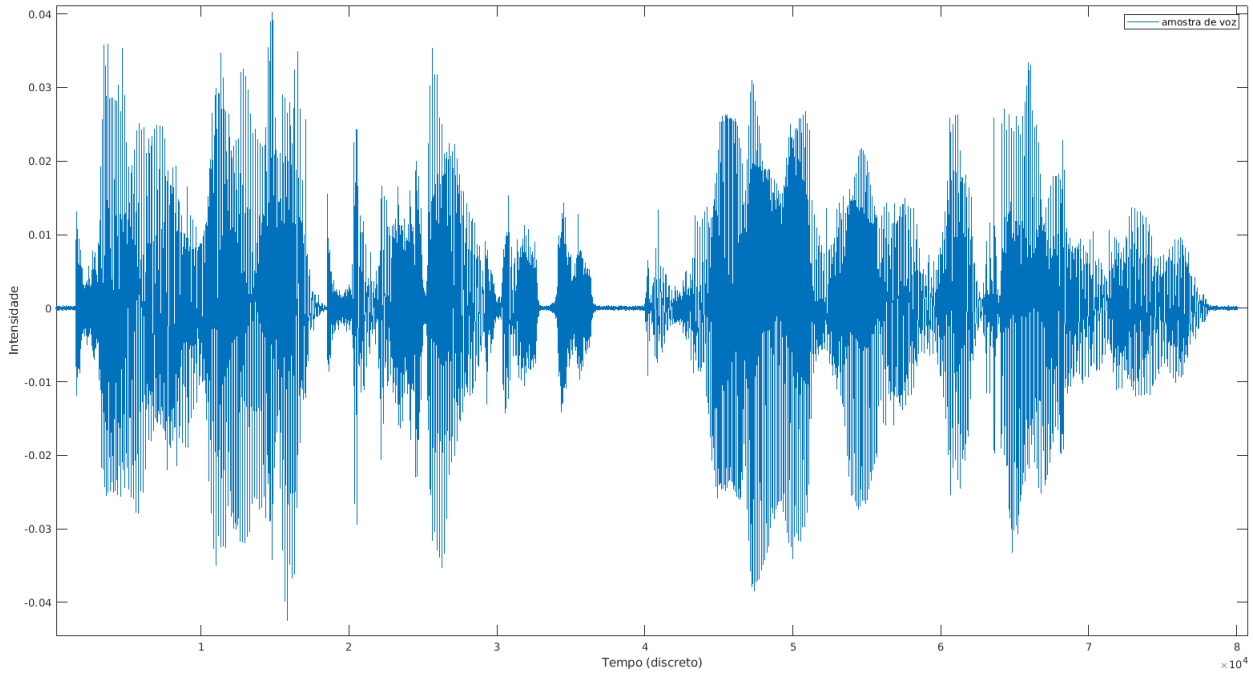


Figura 4.2: Exemplo de amostra de voz anecoica.

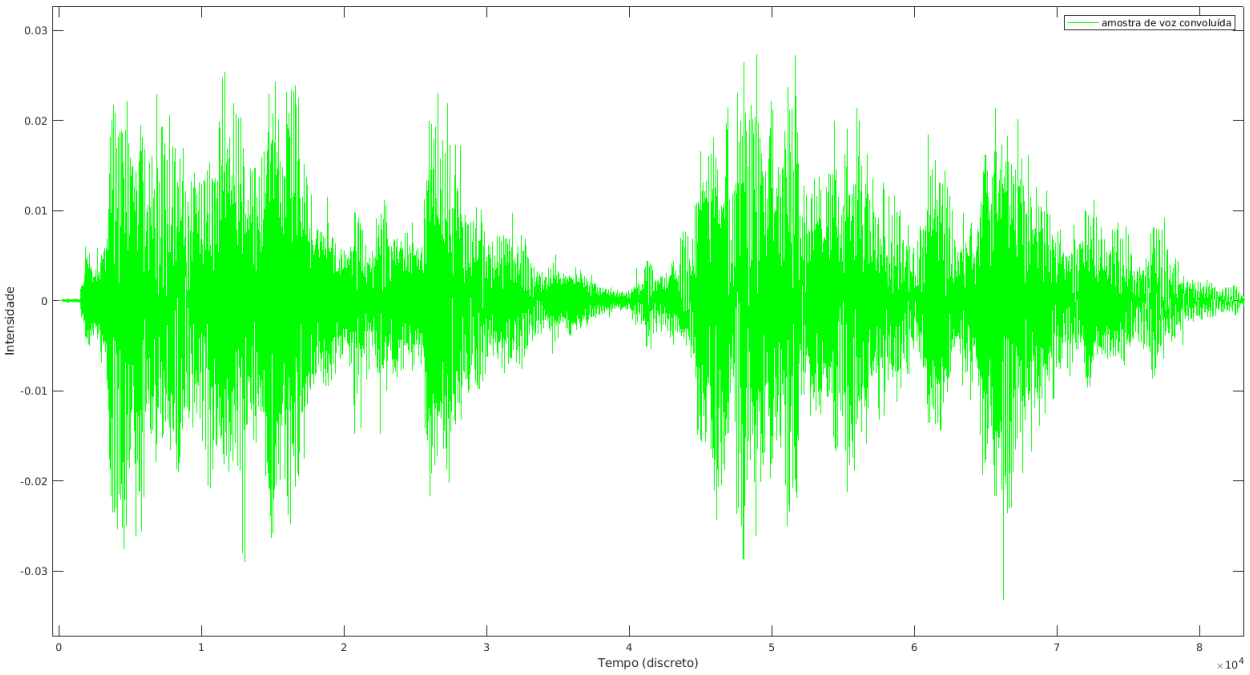


Figura 4.3: Exemplo de amostra de voz reverberante, convoluída com uma RIRSM.

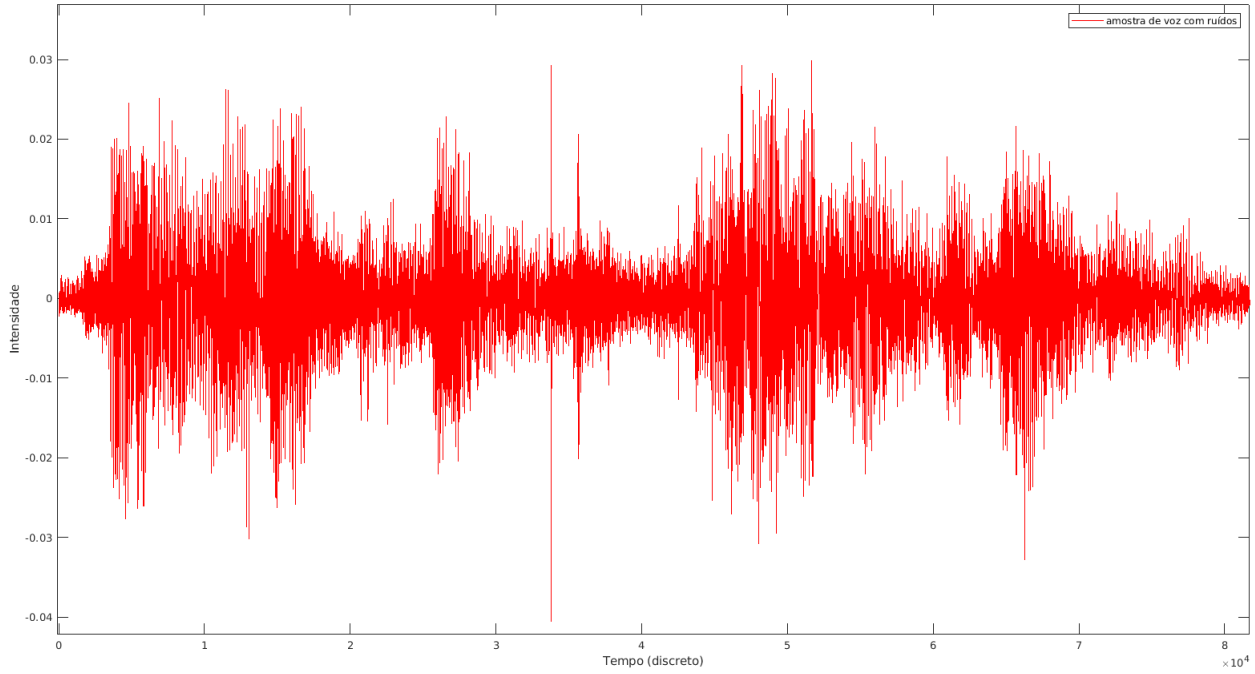


Figura 4.4: Exemplo de amostra de voz em campo distante, representado pela voz reverberada mais os ruídos adicionados pelo segundo método de DA.

Capítulo 5

Bases de Dados

Este capítulo apresenta as bases de dados que serão usadas para gerar os resultados experimentais. Para a aplicação da metodologia deste trabalho, é necessário três fontes de dados:

- base de dados com amostras de voz anecoicas para convolução com as RIRSMs,
- base de dados com RIRs reais para o primeiro nível de *Data Augmentation*, gerando as RIRSMs,
- base de dados com SRPs e SRFs para o segundo nível de *Data Augmentation*, gerando as AVCDs.

5.1 Base de amostras de voz anecoicas

A base de AVAs usada consiste na leitura de textos em inglês por 4 pessoas diferentes (duas vozes masculinas e duas femininas). Os arquivos de áudio são disponibilizados em formato WAV, com frequência de amostragem de 16 KHz, e cada gravação tem duração em torno de 5 a 6 segundos; no caso deste trabalho, foram concatenadas duas frases por pessoa na mesma gravação devido ao tempo de duração dos ruídos pontuais que serão adicionados para a geração de AVCDs.

A tabela 5.1 descreve as gravações usadas neste projeto.

Tabela 5.1: Descrição dos textos pronunciados por locutor.

Nome	Código	Texto
Homen 1 - Texto 1	H1-T1	*This food is too spicy he complained. Young man can be very arrogant and rude.*
Homen 1 - Texto 2	H1-T2	*So Marcus owned a big shipping company. Their eyes met across the table.*
Homen 2 - Texto 1	H2-T1	*Time is running out for the scientists. If you knew Julie like I know Julie.*
Homen 2 - Texto 2	H2-T2	*Your new dress is breathtaking darling. Her first book was published last year.*
Mulher 1 - Texto 1	M1-T1	*Among them are canvases from a young artist. Building from the ground up is very costly.*
Mulher 1 - Texto 2	M1-T2	*Next year we will see several more exibitions. The number of works on view will increase.*
Mulher 2 - Texto 1	M2-T1	*An enourmous quake rocked the island. Eventually he hopes to solve all the problems.*
Mulher 2 - Texto 2	M2-T2	*Eventually he hopes to solve all the problems. Faulty installation can be blamed for this.*

5.2 Base de RIRs - Aachen Impulse Response database (AIR)

A base de AIR [22] é um conjunto de respostas ao impulso sonoras que foram medidas em diversas salas. O objetivo dessa base é de fornecer dados para estudos de algoritmos de processamento de sinais para ambientes reverberados.

Ela é composta primariamente por RIRs binaurais medidas com ou sem uma cabeça falsa de manequim em locais com diferentes propriedades acústicas; é importante frisar também que a base possui gravações com diferentes distâncias entre a fonte sonora e os microfones para a mesma sala, gerando outros tempos de reverberação. A base também possui gravações em diferentes ângulos azimutais com o objetivo de auxiliar algoritmos de detecção de direção da fonte sonora; para o escopo

deste projeto, tais RIRs não serão usadas.

Tabela 5.2: Configurações de RIRs disponíveis na AIR.

Sala	**Descrição**	**Canais**	**Cabeça**	**Distâncias**
Booth	cabine de estúdio	E/D	S/N	0,5/1/1,5
Office	escritório comercial	E/D	S/N	1/2/3
Meeting	sala de reuniões	E/D	S/N	1,45/1,7/1,9/2,25/2,8
Lecture	sala de aula	E/D	S/N	2,25/4/5,56/7,1/8,68/10,2
Stairway	escadaria aberta	E/D	S/N	1/2/3
Aula Carolina	igreja de área 570m^2	E/D	S/N	1/2/3/5/15/20

Os ambientes em que foram feitas as gravações de RIRs e suas respectivas configurações são definidos na tabela 5.2. Todos os ambientes usados possuem gravações com ambos os canais esquerdo e direito (E/D) e com configuração com ou sem a cabeça falsa. As RIRs foram salvas como vetores binários de precisão dupla de ponto flutuante (formato MAT, que pode ser importado via MATLAB®).

5.3 Base de ruídos - MUSAN

A base de MUSAN (*A Music, Speech, and Noise Corpus*) [6] consiste em um conjunto de músicas de diversos gêneros, amostras de voz de doze línguas e uma variedade de ruídos técnicos e não-técnicos. Ela foi criada primariamente para auxiliar no treinamento de modelos voltados para detecção de atividade de voz, contudo ela é usada também para teste de algoritmos processamento de sinais na área de áudio, por exemplo, reconhecimento de voz e orador. Uma das vantagens dessa base é o fato dela ser uma compilação de áudios com fontes em domínios públicos, facilitando a distribuição dos áudios para uso da comunidade científica.

No escopo deste projeto, será usado somente a seção de ruídos da base, composta por seis horas de áudio no total. A seção de ruídos é composta por sons técnicos de curta duração, que são usados como SRPs no segundo processo de *Data Augmentation*, e por sons de ambiente, usados como SRFs no mesmo processo.

Tabela 5.3: Descrição dos tipos de ruídos pontuais usados da base MUSAN.

Código	**Descrição**
RP-1	miado de gato
RP-2	madeira sendo lixada
RP-3	buzina de automóvel
RP-4	porta abrindo
RP-5	grampeador
RP-6	teclado de forno de microondas
RP-7	*zipper* sendo fechado
RP-8	latido de cão
RP-9	batendo em uma porta
RP-10	espirro
RP-11	campainha
RP-12	vibrador de celular

Tabela 5.4: Descrição dos tipos de ruídos de fundo usados da base MUSAN.

Código	**Descrição**
RF-1	avião decolando em aeroporto
RF-2	sala de máquinas
RF-3	estática
RF-4	sons de floresta

As tabelas 5.3 e 5.4 indicam os ruídos separados da base para gerar AVCDs. Os arquivos de áudio são disponibilizados em formato WAV, com frequência de amostragem de 16 KHz.

Capítulo 6

Resultados Experimentais

6.1 Configuração dos parâmetros

6.2 Resultados

Capítulo 7

Conclusões

Referências Bibliográficas

[1] HAEB-UMBACH, R., HEYMANN, J., DRUDE, L., *et al.*, "Far-Field Automatic Speech Recognition", *Proceedings of the IEEE*, v. 109, n. 2, pp. 124–148, 2021.

[2] MOKGONYANE, T. B., SEFARA, T. J., MODIPA, T. I., *et al.*, "Automatic Speaker Recognition System based on Machine Learning Algorithms". In: *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, pp. 141–146, 2019.

[3] XIONG, F., GOETZE, S., MEYER, B., "Joint Estimation of Reverberation Time and Direct-To-Reverberation Ratio from Speech Using Auditory-Inspired Features". In: *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, 2015.

[4] Bryan, N. J., "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2020.

[5] Ko, T., Peddinti, V., Povey, D., *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017.

[6] SNYDER, D., CHEN, G., POVEY, D., "MUSAN: A Music, Speech, and Noise Corpus", 2015, http://www.openslr.org/17/, visitado última vez em 07/06/2021, arXiv:1510.08484v1.

[7] BEUTELMANN, R., BRAND, T., "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners", *The Journal of the Acoustical Society of America*, v. 120, n. 1, pp. 331–342, 2006.

[8] MAAS, R., HABETS, E. A., SEHR, A., *et al.*, "On the application of reverberation suppression to robust speech recognition". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 297–300, 2012.

[9] HELI, H., ABUTALEBI, H. R., "Localization of multiple simultaneous sound sources in reverberant conditions using blind source separation methods". In: *2011 International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 1–5, 2011.

[10] NASSIF, A., SHAHIN, I., ATTILI, I., *et al.*, "Speech Recognition Using Deep Neural Networks: A Systematic Review", *IEEE Access*, v. PP, pp. 1–1, 02 2019.

[11] VARGAS, R., RUIZ, L., "DEEP LEARNING: PREVIOUS AND PRESENT APPLICATIONS", *Journal of Awareness*, v. 2, pp. 11–20, 11 2017.

[12] HAEB-UMBACH, R., WATANABE, S., NAKATANI, T., *et al.*, "Speech Processing for Digital Home Assistants: Combining Signal Processing With Deep-Learning Techniques", *IEEE Signal Processing Magazine*, v. 36, n. 6, pp. 111–124, 2019.

[13] TZINIS, E., VENKATARAMANI, S., WANG, Z., *et al.*, "Two-Step Sound Source Separation: Training On Learned Latent Targets". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, 2020.

[14] ALAM, M., SAMAD, M., VIDYARATNE, L., *et al.*, "Survey on Deep Neural Networks in Speech and Vision Systems", *Neurocomputing*, v. 417, pp. 302–321, 2020.

[15] PARADA, P. P., SHARMA, D., WATERSCHOOT, T., *et al.*, "Evaluating the Non-Intrusive Room Acoustics Algorithm with the ACE Challenge", *ArXiv*, v. abs/1510.04616, 2015.

[16] VIROSTEK, P., "The Quick & Easy Way to Create Impulse Responses", 2014, https://www.creativefieldrecording.com/2014/03/19/the-quick-easy-way-to-create-impulse-responses/, visualizado pela última vez em 16/06/2021.

[17] NAIR, V., "Recording Impulse Responses", 2012, https://designingsound.org/2012/12/29/recording-impulse-responses/, visualizado pela última vez em 16/06/2021.

[18] SALAMON, J., BELLO, J. P., "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification", *IEEE Signal Processing Letters*, v. 24, n. 3, pp. 279–283, 2017.

[19] LU, R., DUAN, Z., ZHANG, C., "Metric learning based data augmentation for environmental sound classification". In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2017.

[20] ISO:, "ISO 3382-1:2009 - Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces", https://www.iso.org/standard/40979.html.

[21] ALLEN, J. B., BERKLEY, D. A., "Image method for efficiently simulating small-room acoustics", *The Journal of the Acoustical Society of America*, v. 65, n. 4, pp. 943–950, 1979.

[22] JEUB, M., SCHäFER, M., VARY, P., "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms". In: *Proceedings of International Conference on Digital Signal Processing (DSP)*, pp. 1–4, IEEE, IET, EURASIP, Santorini, Greece, 2009.