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. A SER DEFINIDO, D. Sc.

Examinador:

_____

Prof. A SER DEFINIDO, D. E.

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

Este exemplar é de propriedade da Universidade Federal do Rio de Janeiro, que poderá incluí-lo em base de dados, armazenar em computador, microfilmar ou adotar qualquer forma de arquivamento.

É permitida a menção, reprodução parcial ou integral e a transmissão entre bibliotecas deste trabalho, sem modificação de seu texto, em qualquer meio que esteja ou venha a ser fixado, para pesquisa acadêmica, comentários e citações, desde que sem finalidade comercial e que seja feita a referência bibliográfica completa.

Os conceitos expressos neste trabalho são de responsabilidade do(s) autor(es).

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

Inserir o resumo do seu trabalho aqui. O objetivo é apresentar ao pretenso leitor do seu Projeto Final uma descrição genérica do seu trabalho. Você também deve tentar despertar no leitor o interesse pelo conteúdo deste documento.

Palavras-Chave: trabalho, resumo, interesse, projeto final.

ABSTRACT

Insert your abstract here. Insert your abstract here. Insert your abstract here. Insert your abstract here. Insert your abstract here.

Key-words: word, word, word.

SIGLAS

DA - *Data Augmentation*

DL - *Deep Learning*

DRR - Razão Direto-Reverberante

RIR - Resposta ao Impulso de Ambiente Acústico

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.

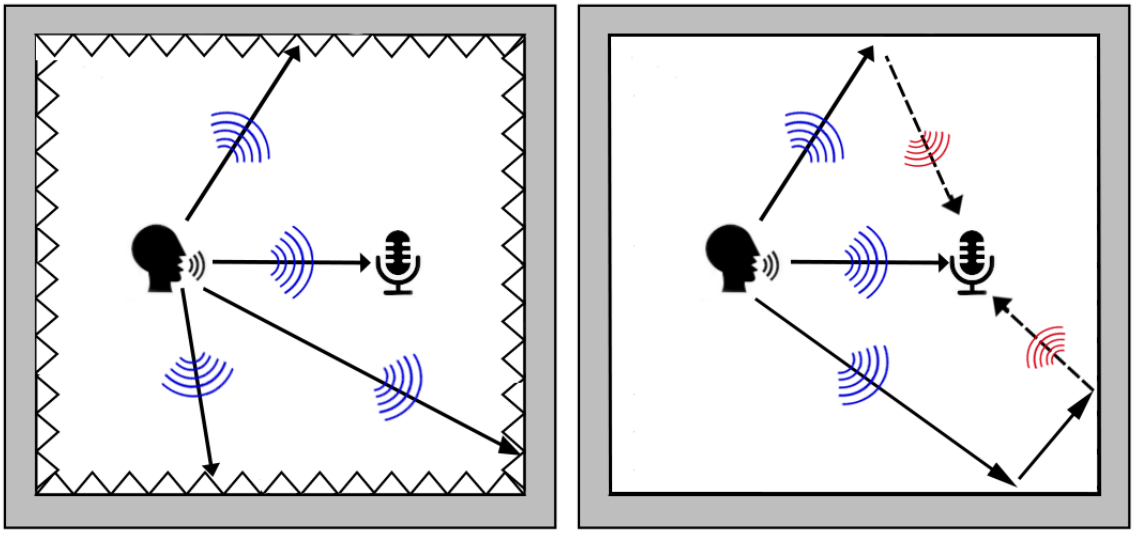
Capítulo 2

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:

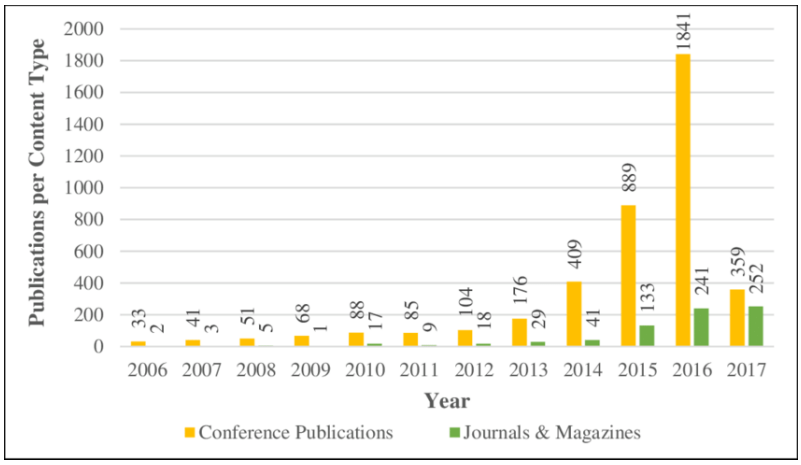
$$Y(t) = s(t) * h(t) + n(t) \tag{2.1}$$

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.

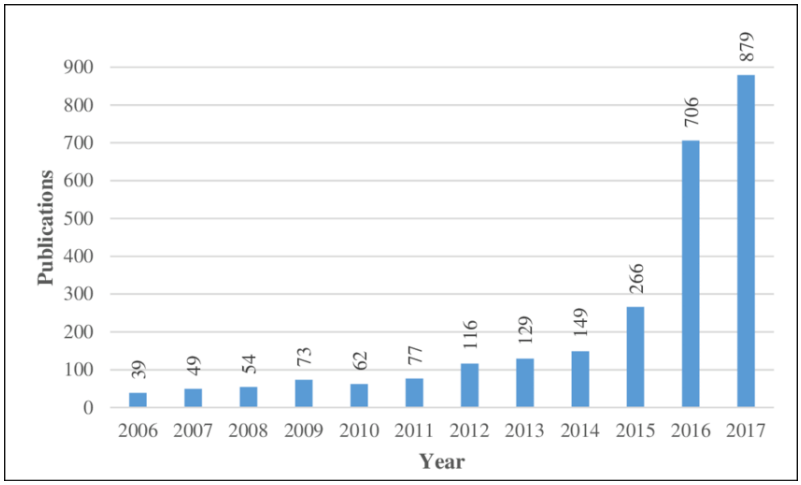
Este projeto é focado no estudo de uma forma de gerar RIRs simuladas a partir de RIRs reais devido à sua importância para diversas aplicações usadas

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].

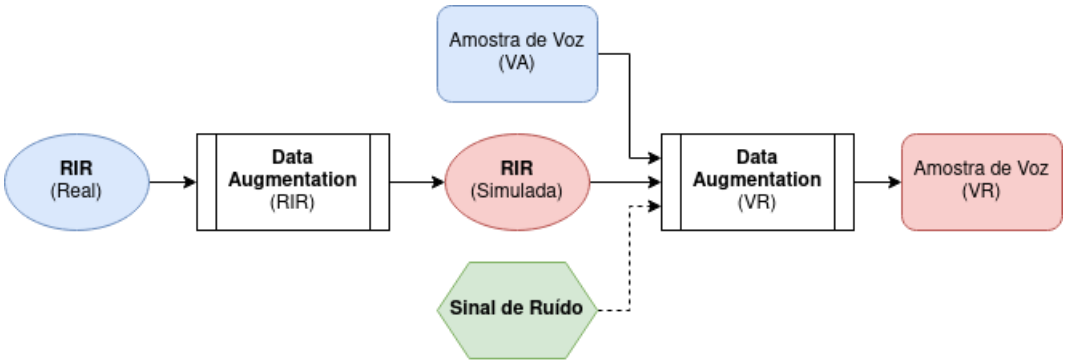


Figura 2.3: Fluxo geral de procedimentos para gerar sinais de voz reverberantes.

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

3.1 Razão Direto-Reverberante (DRR)

3.2 Tempo de Reverberação (T60)

3.3 Comparação entre RIR real e simulada

Capítulo 4

Desenvolvimento de Sinais de Voz Reverberadas Simuladas com Ruídos

4.1 Simulação de fala em campo distante

Capítulo 5

Resultados Experimentais

5.1 Configuração dos parâmetros

5.2 Resultados

Capítulo 6

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.

[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.