UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
ESCOLA POLITÉCNICA
DEPARTAMENTO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO

PROPOSTA DE PROJETO DE GRADUAÇÃO

Aluno: Bruno Machado Afonso
bruno.ma@poli.ufrj.br

Orientador: Mariane Rembold Petraglia

1. TÍTULO

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.

2. ÊNFASE

Computação

3. 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.

4. 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.

5. 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 do ambiente, 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 capturar essa extensa quantidade de gravações de áudio é uma tarefa 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.

6. OBJETIVO

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 anteriores em sequência para gerar amostras de voz em Ambientes ruidosos.

7. 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 (dirac).

Neste trabalho é proposto 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 é gerado RIR simuladas modificando as propriedades de Tempo de Decaimento (T60) e de razão entre áudio direto e reverberado (DRR). Através dessas duas propriedades, defini-se praticamente todos os RIR possíveis de serem gravados 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 puros com os RIR simulados que foram gerados 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 curta que podem ser introduzidos em qualquer momento da fala, já os ruídos de ambiente são sons constantes 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 do RIR tem por objetivo colocar a amostra de voz em vários ambientes fechados, e já os ruídos ajudam drasticamente no treinamento de redes neurais impedindo que as redes fiquem viciadas em características muito específicas da fala durante o treinamento, pois eles tendem a simular os fatores externos que podem estar envolvidos em uma gravação real.

8. MATERIAIS

- Computador:
  - CPU: Arquitetura amd86, AMD Ryzen 3600X 3.8GHz
  - RAM: 16 GB RAM DDR4
  - HDD: 1 TB
- Software:
  - MATLAB® R2018a (Software não-gratuito, requer licença para uso)
  - ITA-Toolbox, plugin open source para medições acústicas para MATLAB®
- Dados:
  - The Aachen Impulse Response (AIR) Database [7], base de dados com respostas ao impulso gravadas de diferentes ambientes.
  - MUSAN: A Music, Speech, and Noise Corpus [6], base de dados com amostras de ruídos pontuais e ruídos de fundo.

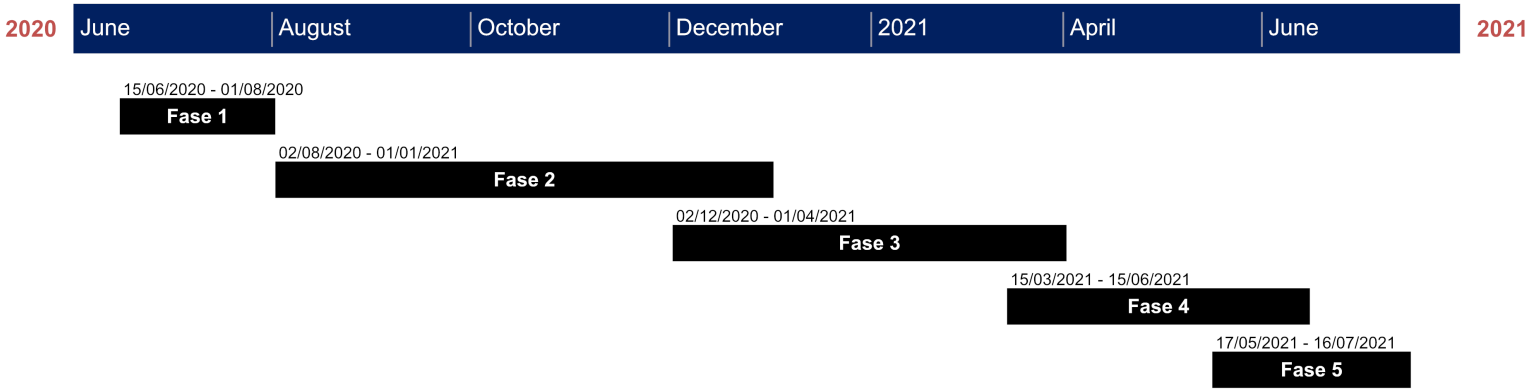

Figura 1: Cronograma do projeto

9. CRONOGRAMA

Apresentado graficamente conforme a Figura 1.

Fase 1: Estudo sobre a Resposta ao Impulso sonoro e sobre os seus principais parâmetros.

Fase 2: Desenvolvimento e implementação do algoritmo proposto [4] de reforço de dados para as RIR.

Fase 3: Desenvolvimento e implementação do algoritmo proposto [5] de incremento de ruídos pontuais e de fundo em amostras de voz.

Fase 4: Desenvolvimento e implementação do algoritmo que cria sinais de voz com RIR e ruídos gerados nas etapas anteriores.

Fase 5: Escrita do texto do projeto de graduação.

Referências Bibliográficas

[1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[2] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "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] F. Xiong, S. Goetze, and B. Meyer, "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] N. J. Bryan, "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] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "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] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 10 2015.

[7] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International*

Conference on Digital Signal Processing (DSP), (Santorini, Greece), pp. 1–4, IEEE, IET, EURASIP, IEEE, 2009.

Rio de Janeiro, 7 de junho de 2021

Bruno Machado Afonso - Aluno

Mariane Rembold Petraglia - Orientador