Desenvolvimento de base de dados para treinamento de redes neurais de reconhecimento de voz através da geração de aúdios com resposta ao impulso simuladas por técnicas de data augmentation [RESUMO]

Bruno Machado Afonso

18 de Abril de 2021

1 Introdução

Com o avanço das tecnologias de automação residencial, assistentes pessoais nos smartphones e comunicação online, o estudo de técnicas de processamento de áudio (nesse caso específico, voz), torna-se cada vez mais relevante para a sociedade. Junto a isso, houve grandes avanços no âmbito do aprendizado de máquina, fornecendo alternativas para os métodos tradicionais de processamento de áudio.

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 é de capturar essa extensa quantidade de gravações de áudio, pois 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.

A proposta dessa tese é de gerar uma base de dados, contendo amostras de voz reverberadas em diversos ambientes, para treinamento de redes neurais, partindo de amostras de voz puras e resposta ao impulso de salas já existentes. A tese propõe em utilizar técnicas de data augmentation para gerar um grande volume de vozes reverberadas à partir de um pequeno conjunto de amostras.

2 Resposta ao Impulso de uma sala (RIR) e técnicas de Data Augmentation

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 à um sinal de rúido, 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 ao longo do tempo em um ambiente fechado após uma excitação extremamente curta e cheia de energia (dirac).

Nessa tese 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. Seguimos o que foi proposto no artigo de data augmentation para respostas ao impulso para estimação do modelo acústico [1], onde geramos RIR simuladas modificando as propriedades de Tempo de Decaimento (T60) e de razão entre áudio direto e reverberado (DRR). Através dessas duas propriedades, conseguimos definir praticamente todos os RIR possíveis de serem gravados de forma artificial.

Para gerar as amostras de vozes reverberadas que compõe a base de dados, seguimos o que é proposto no artigo de estudo de data augmentation em vozes reverberadas [2], onde convoluímos sinais de voz puros com os RIR simulados que foram gerados anteriormente. Além disso, é acrescentado a essa sinal de voz reverberado rúidos diversos, que são caracterizados de duas formas: ruídos pontuais e ruídos de ambiente. Os ruídos pontuais são amostras de aúdio 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 [3].

Através desses dois passos, conseguimos gerar 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.

Referências

[1] 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.

[2] 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.

[3] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 10 2015.