ISSN: 2526-1371

Reconhecimento de Fonemas utilizando Redes Neurais Convolucionais para Transcrição Fonética Automática

Bauke Alfredo Dijkstra[1], **Ionildo José Sanches**[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Monteiro Lobato, s/n - Km 04, CEP 84016-210 - Ponta Grossa - PR – Brasil

{baukead,ijsanches}@gmail.com

Palavras-chave: Reconhecimento de fonemas; Deep learning; Processamento acústico

Resumo. Este trabalho tem como objetivo desenvolver uma técnica de reconhecimento automático de fonemas (RAF) de forma contínua. O reconhecimento de fonemas é a capacidade de extrair características para reconhecer as unidades sonoras das palavras e transcrevê-las. O RAF tem aplicações em diversas áreas como o reconhecimento automático de fala (ASR - *Automatic Speech Recognition*), identificação de locutores, identificação de erros de pronuncia e reconhecimento de emoções. O RAF consiste de três modelos principais de fala: palavras isoladas, palavras concatenadas e da fala contínua. Conforme (Ashwini Kshirsagar, et al. Comparative Study of Phoneme Recognition Techniques. 2012) o reconhecimento de fonemas funciona melhor para palavras isoladas do que em fala contínua. Para realizar o RAF pode-se realizar um pré processamento dos áudios denominado processamento acústico para minimizar ruídos e diferenças entre locutores, e então realizar a classificação dos fonemas utilizando algoritmos de aprendizagem de máquina, com o objetivo de identificar os fonemas. No desenvolvimento desse projeto os testes serão realizados com fonemas extraídos das bases de áudios como a TIMIT *Acoustic-Phonetic Continuous Speech* Corpus que é uma base com fala em inglês e possui transcrições ortográficas, fonéticas e de palavras, e em bases com fala em português como a VoxForge, Sid e LABSM1.4. As bases em português são apenas transcritas em forma ortográficas, portanto torna-se necessário utilizar um programa de grafema para fonema para formar os fonemas em relação ao texto, e por fim validar de forma manual a transcrição de acordo com os áudios. Para realizar o RAF é necessário realizar o processamento acústico dos áudios dessas bases, que podem ser feitas utilizando o janelamento de Hamming e depois aplicar o MFCC (*Mel Frequency Cepstral Coeficients*). Após o processamento acústico é realizado a classificação dos fonemas. Nesse trabalho será utilizado o *framework* Keras para *Deep learning*, uma API (*Application Programming Interface*) de alto nivel para o TensorFlow, para realizar os treinamentos com as bases citadas. Por fim, os resultados serão comparados e analisados para avaliar o modelo obtido com a taxas de reconhecimento.