RESUMO - SOCIOFONÉTICA

ANALISANDO CLONAGENS DE VOZ ATRAVÉS DA SOCIOFONÉTICA

Gabriel Catani (catani@dr.com)

Atualmente, tecnologias de síntese de voz podem ser facilmente acessadas por qualquer pessoa com uma conexão de internet. Através do rápido desenvolvimento de modelos de aprendizado de máquina, essas vozes sintéticas são capazes de enganar até ouvidos treinados, nos fazendo pensar que pertencem a falantes humanos. Esse tipo de tecnologia possui diversas aplicações, das quais se destacam as associadas à acessibilidade: desde a leitura de telas em tempo real, até a possibilidade de dar voz a pessoas com a capacidade fonatória comprometida. Apesar das aplicações benéficas, vozes geradas por meio de Inteligência Artificial (IA) têm sido amplamente utilizadas para fins ilícitos, como golpes, muitos deles pautados na clonagem de voz. Baseando-se em alguns segundos de fala de determinado falante-alvo, certos modelos são capazes de emular a voz de tal falante com uma qualidade surpreendente. Esse tipo de áudio sintético apresenta um desafio para análises forenses, dado que múltiplos parâmetros acústicos da síntese podem ser praticamente idênticos aos do falante-alvo. O desafio é crescente, considerando os altos investimentos na área e frequentes melhorias nos modelos. Ainda assim, os engenheiros e matemáticos tendem a favorecer ajustes finos,

majoritariamente de ordem acústica, a despeito de outros fatores de ordem sociolinguística, que passam desapercebidos. Enquanto a implementação de pequenas variações nos ciclos glotais trazem naturalidade para as vozes clonadas, emergem incoerências dialetais, como o uso de variantes ausentes no idioleto do falante-alvo. Nesse sentido, essa comunicação busca analisar a clonagem de voz através de uma perspectiva linguística, tendo como base o arcabouço teórico-metodológico da Sociofonética. Realizou-se uma análise comparativa detalhada entre áudios gerados sinteticamente (outputs) e seus equivalentes originais (inputs), com o objetivo de explorar as capacidades e limitações dos modelos utilizados, identificando características linguísticas específicas e potenciais indicadores de síntese. O corpus do estudo consiste em gravações da fala de 10 indivíduos brasileiros, com idades entre 18 e 32 anos, divididos por gênero. Cada participante contribuiu aproximadamente com 565 segundos de conversação espontânea, 191 segundos de leitura de frases derivadas dessas conversas e 38 segundos de leitura isolada de palavras. Com base nesse material, foram criados clones vocais utilizando o modelo XTTSv2, testando a influência variável dos hiperparâmetros. A partir disso, na primeira fase, extraiu-se uma variedade de medidas acústicas que se relacionam tanto com aspectos prosódicos quanto segmentais da fala, incluindo descritores associados à frequência fundamental, MFCCs, características formânticas e ritmo. Resultados preliminares sugerem que a distância mel-cepstral e as variações na frequência fundamental podem atuar como indicadores eficazes para a detecção de clones de voz.

Palavras-chave: sociofonética; fonética forense; tts; deepfake.