Caso de estudo de identidade de voz utilizando coeficientes cepstrum nas frequências Mel nas eleições de 2018

Mario Gazziro[1]

[1]Universidade Federal do ABC

mario.gazziro@ufabc.edu.br

Resumo. *Às vésperas da eleição presidencial de 2018 circulou um áudio pela rede social WhastApp atribuído inicialmente a um dos candidatos à presidência, e posteriomente atribuído a um comediante o qual imitava esse candidato em programas de TV. Esse estudo fez uso de técnicas avançadas - uso de coeficientes cepstrum em frequências Mel - para determinar que o áudio vazado apresentou mais chances de pertencer ao candidato (56%) do que ao humorista (44%).*

1. Introdução

No dia 19 de setembro de 2018 deu início a circulação de um áudio nas redes sociais do WhatsApp no qual supostamente o então candidato à presidência da república, Jair Bolsonaro, desfere xingamentos a seu candidato à vice-presidente, a sua enfermeira, e o mais importante, alega estar fazendo parte de um 'teatro', se referindo à questão do atentado cometido contra sua vida ser na verdade um engodo, possivelmente para escapar aos debates e ampliar sua simpatia com o público em geral, visto que devido a esse fato ele deixou de ficar estagnado nas pesquisas e pode finalmente começar dar sinais de enfrentamento a seu opositor, que até então subia vertiginosamente nas inteções de voto.

Embora a importância da elucidação da natureza dessa gravação à época tenha sido de extrema importância para o processo eleitoral, sua análise de veracidade foí tão rápida quanto ineficiente. Apenas algumas horas após o vazamento do áudio, os jornais de notícias reportaram que o mesmo se tratava de conteúdo falso (FAKE) apenas com base nos depoimentos diretos das partes envolvidas (filhos e hospital que atendia o candidato) [Globo 2018], ou seja, grupos que, de forma alguma, eram imparciais na análise do ocorrido, e tinham total interesse em classificá-lo como falso (filhos por protecionismo ao pai e o hospital por auto-preservação).

Simultâneamente outros jornais reportaram análises completamente subjetivas, como a declaração que a voz no áudio tinha um sotaque paulista, alegando que o candidato possuia sotaque carioca [OTempo 2018], deixando completamente de lado o fato de que o candidato é na verdade de naturalidade paulista, o que explicaria o sotaque.

Após apenas 2 dias do ocorrido, a agência da comprovação de notícias falsas 'Comprova' divulgou em matéria no jornal 'O Estado de São Paulo' um laudo técnico encomendado ao instituo IBP Tech de São Paulo, no qual uma análise técnica do tom (pitch) da voz, baseado em frequência fundamental (F0) - classificava mais uma vez o áudio como sendo falso [IBPTech 2018].

Embora seja essa uma técnica muito pobre para tentar identificar um interlocutor, o relatório técnico apresentado e disponível no portal do jornal de notícias apresentou

resultados adulterados e portanto impossíveis de serem reproduzidos, visto que, como mostra a seta vermelha na Figura 1, o tempo em segundos para comparação das amostras está ilegível, sem ao menos nos esclarecer qual trecho exato do intervalo de tempo estão se referindo nas amostras, a fim de que possamos averiguar se nao está sendo comparada a voz do candidato com a voz se seu entrevistador - isso devido ao fato dos peritos que realizaram esse laudo terem curiosamente escolhido um áudio para comparação no qual o candidato era entrevistado por outra pessoa, existindo 2 vozes na amostra de comparação.

Curiosamente esse laudo foi apresentado ao país inteiro sem qualquer questionamento sobre essa adulteração grosseira, a qual impossibilita a qualquer interessado reproduzir tais resultados, alegado veemente como falso.

Dando continuidade a polêmica, um humorista brasileiro foi acusado então (via boatos em redes sociais) de ter sido o autor do áudio [AZEVEDO 2018].

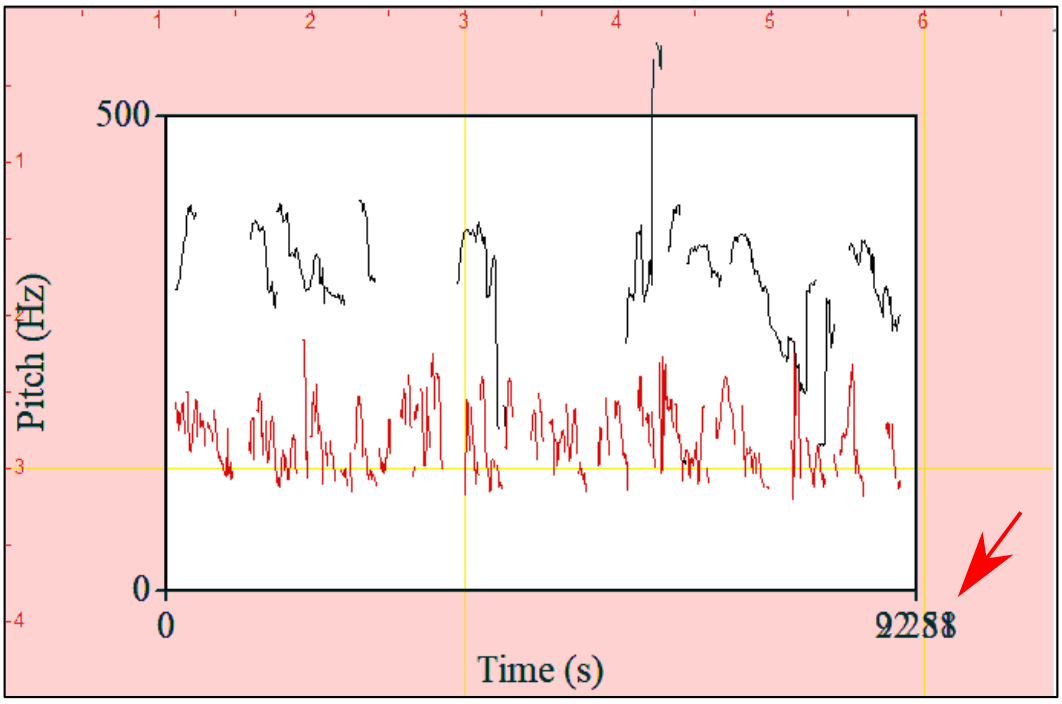


Figure 1. Análise do tom (pitch) para cada amostra no laudo técnico requisitado pelo projeto Comprova [IBPTech 2018]. Seta vermelha indica região adulterada e ilegível do laudo apresentado ao público do país inteiro.

2. Objetivos

Com o objetivo de elucidar o caso em questão de uma maneira técnica e reprodutível por outros - usando processamento de sinais no estado-da-arte - demos inicio ao presente trabalho.

3. Metodologia

Foi adotado o arcabouço de extração de características através de coeficientes Cepstrum na escala de frequencias Mel, tido como estado-da-arte para tal tarefa de identificação de interlocutores [Has 2004] junto a literatura forense internacional (Figura 2).

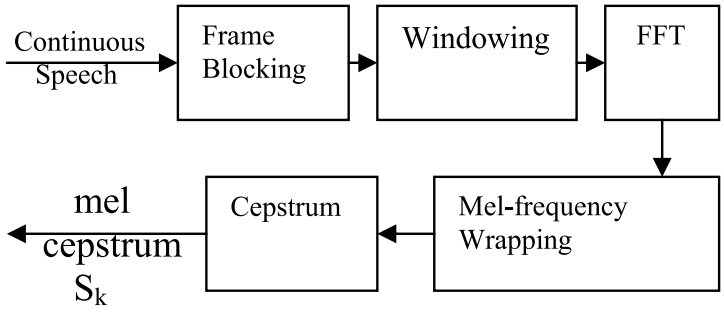


Figure 2. Diagrama em blocos da extração de características usando coeficientes Mel [Has 2004].

Após a extração de tais caractepristicas, foi utilizada a técnica de aprendizado de máquina KNN (K-Nearest Neighbors) para classificação não supervisionada dos resultados.

As figuras e tabelas no restante desse artigo apresentam os resultados obtidos com tal análise e classificação.

Com relação aos áudios de controle, a idéia foi variar ao máximo as amostras do mesmo locutor, por isso para o primeiro áudio foi escolhido uma canção na lingua inglesa, e para o segundo áudio, uma poesia na lingua portuguesa. Uma vez que as técnicas de reconhecimento de assinatura vocal devem identificar características inerentes ao trato vocal do locutor, tais diferenças não devem ser significativas no processo (mudança de lingua ou mesmo mudança de canto para prosa).

A fim de tentar esclarecer ambos os casos, optamos por realizar um confrontamento ao áudio vazado tanto com uma amostra da voz do candidato quanto com uma amostra da voz do humorista.

As amostras foram adquiridas no portal YouTube, sendo a amostra do humorista um trecho de um show no qual ele esta efetivamente imitando o candidato, e a amostra do candidato um trecho de uma entrevista em data muito próxima ao vazamento do áudio, para caracterizar o mesmo estado médico no qual se encontrava na época do vazamento do áudio em questão.

Todos os códigos fonte utilizados e amostras de áudio se encontram na plataforma GitHub [Gazziro 2018].

4. Resultados

A tabela da Figura 6 apresenta os resultados finais. Primeiramente os dados de controle comprovam a eficácia do algoritmo desenvolvido, onde o mesmo locutor foi identificado com percentual de similaridade de mais de 87%.

Uma comparação entre o áudio de controle gerado pelo autor do trabalho e o áudio do canditado apresentou corretamenete uma baixíssima taxa de similaridade de 12%, como era de se esperar, visto serem vozes perceptivelmente diferentes a quaisquer ouvidos.

O resultado mais aguardado, a comparação do áudio do candidato com relação ao áudio vazado apresentou uma taxa de similaridade de 56%. Embora o valor obtido foi

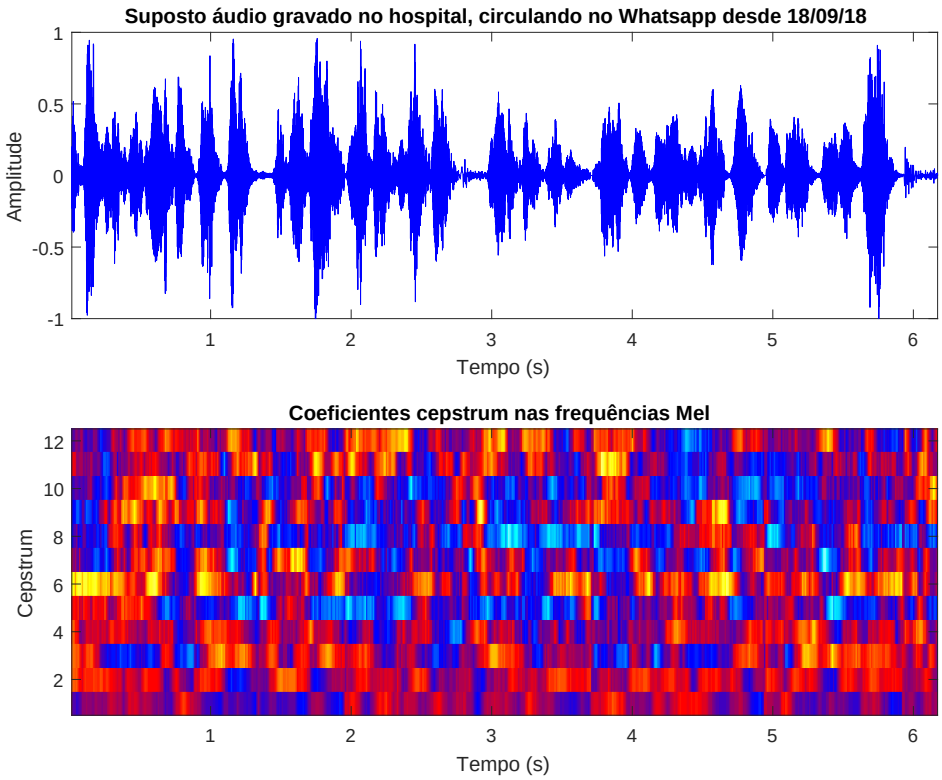


Figure 3. Extração de características do suposto áudio atribuído ao candidato e ao comediante.

atípico para uma classificação positiva dentro do contexto de reconhecimento de padrões (acima de 80%), resta lembrar que a amostra digitalizada possivelmente foi gravada por trás de uma porta, fato que pode ter inferferido na identificação das características do trato vocal original do sujeito avaliado.

Por fim, o resultado da comparação do áudio do humorista com a amostra vazada gerou uma taxa de similaridade de 44%. Embora muito próxima da taxa obtida com o teste da voz do candidato, e ainda abaixo do padrão para classificação positiva em reconhecimento de padrões, devemos lembrar que o sucesso das imitações em humoristas de grande talento se devem justamente ao controle das contrações de seu próprio trato vocal, a partir de muito treino e prática, o que explica tal proximidade com um sistema de reconhecimento baseado nesses quesitos.

5. Conclusão

O índice obtido de 56% indica que pode ser plausível a atribuição do áudio da amostra ao candidato, o que tornaria o áudio vazado verdadeiro, visto que um renomado e talentoso imitador só conseguiu atingir 44% de similaridade.

References

(2004). *SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS, Hasan, R. et al. in 3rd International Conference on Electrical Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh*.

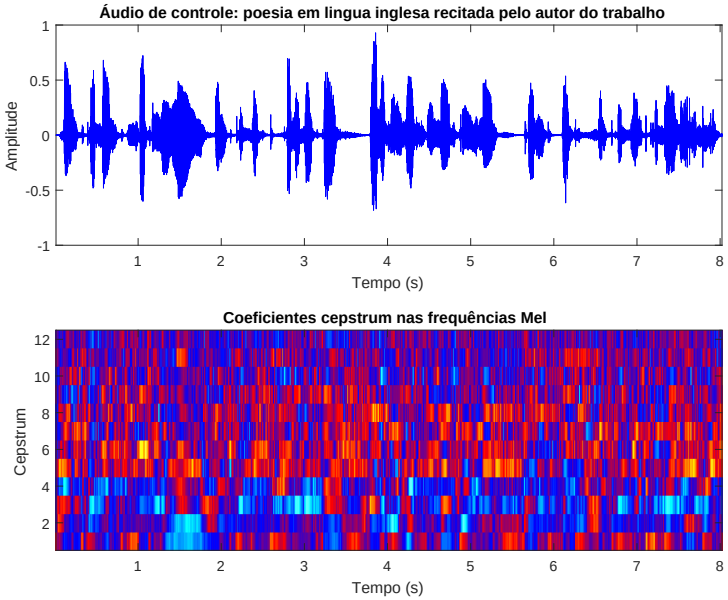
AZEVEDO, P. (2018). Marcelo adnet nega ter imitado a voz de bolsonaro em hospital e denuncia fake news. https://portalovertube.com/2018/09/27/marcelo-adnet-nega-ter-imitado-a-voz-de-bolsonaro-em-hospital-e-denun

Gazziro, M. (2018). Voiceid. https://github.com/mariogazziro/VoiceID.

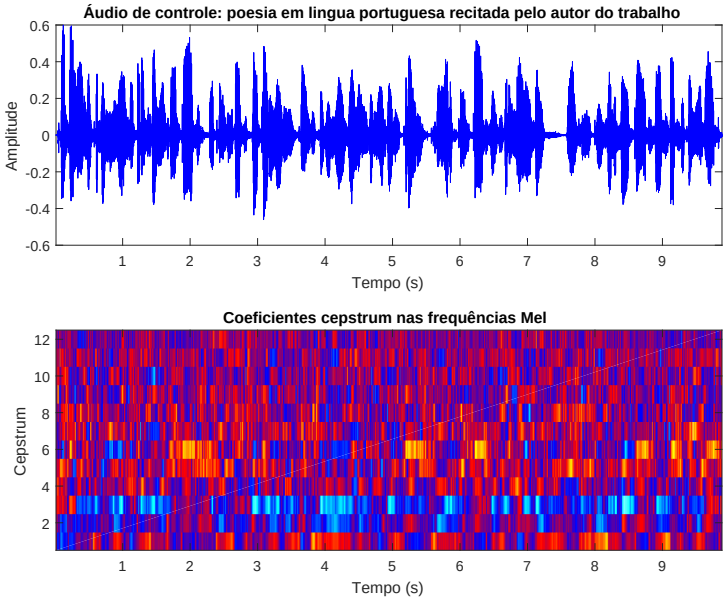
Globo (2018). É fake áudio que mostra bolsonaro gritando em hospital. https://g1.globo.com/fato-ou-fake/noticia/2018/09/20/e-fake-audio-que-mostra-bolsonaro-gritando-em-hospital.ghtml.

IBPTech (2018). Parecer técnico ibp18072. https://politica.estadao.com.br/blogs/estadao-verifica/wp-content/uploads/sites/690/2018/09/Parecer-T%C3%A9cnico-IBP18072.pdf.

OTempo (2018). Eleicoes 2018. https://www.otempo.com.br/hotsites/elei%C3%A7%C3%B5es-2018/suposto-%C3%A1udio-de-jair-bolsonaro-reclamando-do-hospital-%C3%A9-falso-1.2034264.

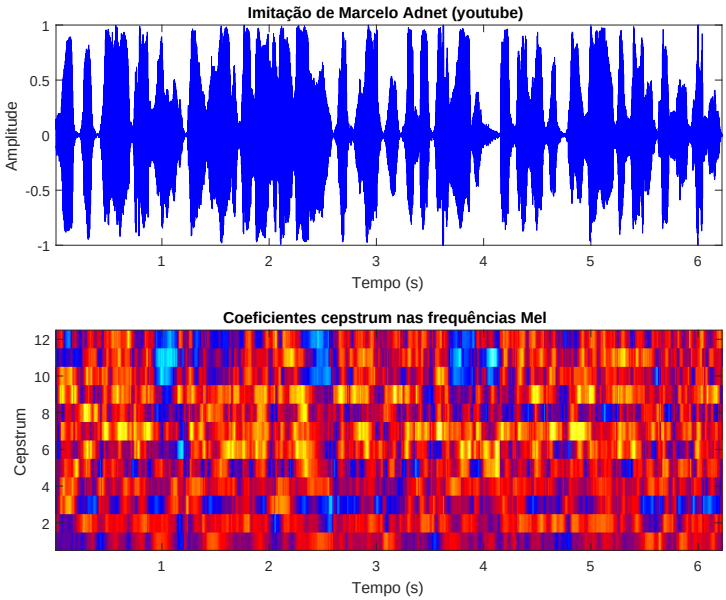


(a) Controle 1: Áudio em lingua inglesa gerado pela voz do autor do presente trabalho.

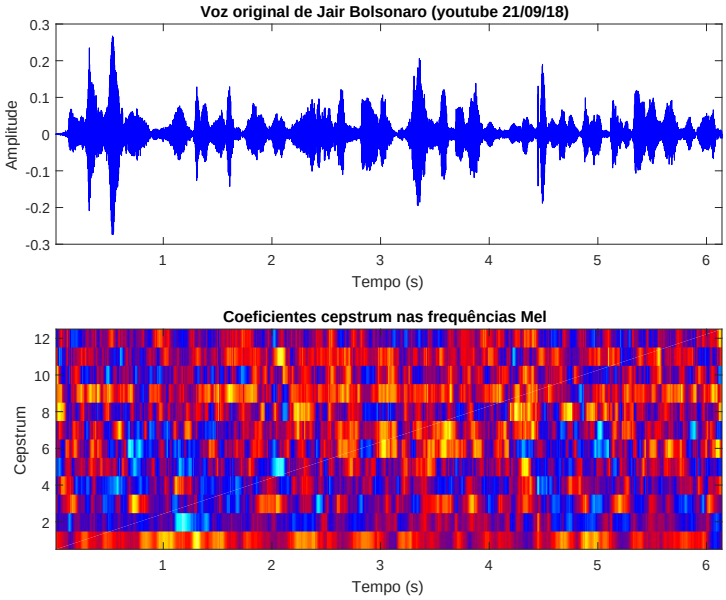


(b) Controle 2: Áudio em português gerado pela voz do autor do presente trabalho.

Figure 4. Amostras de controle do mesmo autor para testes do software.



(a) Imitador.



(b) Candidato.

Figure 5. Extração de características dos áudios do candidato e do imitador.

Resultados			
CONTROLE/TESTE	Áudio 1	Áudio 2	% de similaridade via KNN
CONTROLE	controle 1	controle 2	87,2%
CONTROLE	candidato	controle 2	12,8%
TESTE	candidato	amostra	55,7%
TESTE	imitador	amostra	44,3%

Figure 6. Tabela final com resultados, demonstrando que a amostra testada, o suposto áudio vazado, tem mais chances de pertencer ao candidato (56%) do que ao humorista (44%).