

DO NATURAL AO ARTIFICIAL: A ASCENSÃO DA CLONAGEM DE VOZ E O FUTURO DA AUTENTICIDADE VOCAL

FROM NATURAL TO SYNTHETIC: THE RISE OF VOICE CLONING AND THE FATE OF VOCAL AUTHENTICITY

DEL LO NATURAL A LO ARTIFICIAL: EL ASCENSO DE LA CLONACIÓN DE VOZ Y EL FUTURO DE LA AUTENTICIDAD VOCAL

Guilherme Salim de Barros¹

Jonathan Gazzola²

Carlos Caetano de Almeida³

RESUMO: Esse artigo buscou analisa as técnicas de clonagem de voz, através de um estudo exploratório do modelo zero-shot multilíngue XTTS-v2 no cenário do português brasileiro. Foram geradas 30 amostras artificiais com base em três condições de referência, com durações aproximadas de 8, 14 e 28 segundos, em seguida comparadas às gravações humanas utilizando descritores acústico-prosódicos. A condição de 28 segundos demonstrou menor distância média e dispersão no índice exploratório utilizado, sugerindo maior estabilidade na mesma. Embora haja variações entre as frases produzidas e o índice não constitua uma medida direta de naturalidade/identidade vocal, os resultados mostram que pequenas amostras de referência já são capazes de produzir características acústico-prosódicas da fala original. As constatações reforçam a importância de analisar a autenticidade vocal de forma multidimensional, contribuindo para a discussão dos limites e impactos da clonagem e estudos de identificação da fala sintética em português brasileiro.

Palavras-chave: Clonagem de Voz. Síntese de Fala. XTTS-v2.

ABSTRACT: This article aims to analyses voice cloning techniques through an exploratory study of the multilingual zero-shot XTTS-v2 model in the context of Brazilian Portuguese. Thirty synthetic samples were produced based on three reference conditions of approximately 8, 14, and 28 seconds, and subsequently compared with human recordings using acoustic-prosodic descriptors. The reference condition of 28 s showed the lowest mean distance and dispersion in the exploratory index, suggesting greater stability under this condition. While variation was observed across the generated samples and the index does not directly measure vocal naturalness or voice identity, the results still indicate that short reference samples can already produce synthetic speech that approximates acoustic and prosodic properties of the original voice. The findings reinforce the value of multidimensional analysis of vocal identity, contributing to discussing the limitations and impacts of voice cloning and to studies on synthetic speech detection in Brazilian Portuguese.

Keywords: Voice cloning. Speech Synthesis. XTTS-v2.

¹ Discente do curso de Engenharia da Computação, Aluno Orientado — Faculdades de Campinas.

² Docente do curso de Engenharia Agrônômica, Co-Orientador - Universidade Federal de São Carlos.

³ Pós-Doutor em Inteligência Artificial e Agro 4.0, Orientador - Universidade Federal de São Carlos.

RESUMEN: Este artículo buscó analizar las técnicas de clonación de voz, a través de un estudio exploratorio del modelo zero-shot multilingüe XTTS-v2 en el escenario del portugués brasileño. Se generaron 30 muestras artificiales con base en tres condiciones de referencia, con duraciones aproximadas de 8, 14 y 28 segundos, comparadas posteriormente con las grabaciones humanas utilizando descriptores acústico-prosódicos. La condición de 28 segundos demostró menor distancia media y dispersión en el índice exploratorio utilizado, lo que sugiere una mayor estabilidad en la misma. Aunque existen variaciones entre las frases producidas y el índice no constituye una medida directa de naturalidad/identidad vocal, los resultados muestran que pequeñas muestras de referencia ya son capaces de producir características acústico-prosódicas de la habla original. Las constataciones refuerzan la importancia de analizar la autenticidad vocal de forma multidimensional, contribuyendo a la discusión sobre los límites e impactos de la clonación y a los estudios de identificación del habla sintética en portugués brasileño.

Palabras clave: Clonación de voz. Síntesis de habla. XTTS-v2.

INTRODUÇÃO

Com a evolução da Inteligência Artificial (IA), sua aplicação tem se expandido para um número crescente de áreas. Russell SJ e Norvig P (2022) abordam a IA a partir do estudo de sistemas capazes de receber informações do ambiente e executar ações com base nessas entradas. A flexibilidade desse princípio permite que diferentes técnicas de IA sejam aplicadas a problemas diversos, enquanto seu potencial e seus limites empíricos continuam sendo explorados à medida que novos métodos são desenvolvidos. Nesse contexto, técnicas de aprendizado de máquina e, especialmente, de aprendizado profundo têm impulsionado avanços significativos no processamento e na geração de fala.

Nos últimos anos, as técnicas de clonagem de voz evoluíram rapidamente, mostrando-se capazes de combinar melhorias na naturalidade da fala sintética com a redução da quantidade de dados necessária para adaptar modelos de clonagem a novos falantes. Azzuni H e El Saddik A (2025) mostram que esse avanço acompanha a evolução dos sistemas texto-para-fala (TTS) em aspectos como a qualidade, naturalidade, prosódia e o timbre. Além disso, os autores ressaltam ao revisar a literatura recente que, o avanço de abordagens como *speaker adaptation*, *few-shot* e *zero-shot voice cloning*, constatando que o desenvolvimento da área não se concentra apenas na qualidade da voz sintetizada, mas também na redução da quantia de dados necessária para adequar ferramentas a novos falantes. Essa menor exigência de dados tem tornado a clonagem de voz mais acessível, rápida e eficiente, ampliando suas possibilidades de ser aplicada em setores diversos.

Desta forma, a clonagem de voz pode se utilizar de diferentes tipos de modelos para síntese. Weitzman C (2025) afirma que o processo de clonagem é baseado no procedimento

fornecido pela plataforma Speechify e dividido em quatro fases, sendo: 1 – Coleta de amostra de voz; 2 – Pré-Processamento; 3 – Treinamento de um modelo; 4 – Teste e Sintetização da voz. Os modelos mais tradicionais de TTS são divididos majoritariamente em 3 etapas, sendo:

Normalização de Texto: Normalização de texto para gerar representações fonêmicas internas, baseado na tokenização de frases, ou seja, quebrando seus elementos em blocos, “tokens” e em seguida, realização da expansão sequencial de palavras que não são “padrão”, *non-standard words* (NSWs). Além disso, é nessa fase que palavras homógrafas (palavras de mesma grafia mas se diferenciam pela pronúncia) tem suas pronúncias corretas reconhecidas (JURAFSKY D e MARTIN JH, 2009).

Análise Fonética: Etapa de produção de pronúncia para cada palavra correspondente nas palavras normalizadas na fase anterior e que consiste na elaboração de um dicionário de pronúncias para as palavras de determinada língua e que realiza a conversão de grafema para fonema (g2p) de forma automática, algo que é frequentemente atribuído a Lucassen e Mercer (1984). Atualmente, a conversão grafema-fonema é realizada principalmente por métodos baseados em *deep learning*, com destaque para modelos seq2seq, Transformers e abordagens com modelos pré-treinados (CHENG Z, et al., 2024).

Análise Prosódica: Prosódia é definido como o conjunto de propriedades suprasegmentais da fala, como a entonação, o ritmo, o acento, as pausas e a organização temporal/melódica do enunciado, responsáveis por marcar estrutura, saliências e nuances de sentido (LADD DR, 2014). Em sua análise, existem 3 aspectos imprescindíveis para o TTS tradicional: a estrutura prosódica, a proeminência prosódica e contorno entoacional (entonação).

Atualmente, a análise e a modelagem prosódica em TTS tendem a ser realizadas por abordagens neurais mais integradas, em contraste com os métodos tradicionais baseados em módulos e regras explícitas, como mencionado acima (BARAKAT H, et al., 2024). Entretanto, é essencial entender como os processos ocorrem desde sua origem para avaliarmos a evolução da naturalidade da voz.

A partir do roteiro mostrado, se torna nítido que o processo de clonagem da voz se trata de sintetizar uma voz a partir de algumas amostras da mesma. Diferente de outras abordagens, como o TTS tradicional, que nem sempre preservam características individuais de um falante específico (identidade vocal), ainda que possam ser minimamente expressivos, neste trabalho tem-se como objetivo estudar a clonagem de voz, analisando como características da fala,

sobretudo acústico-prosódicas das amostras geradas por um modelo de clonagem, se aproximam da fala original em diferentes condições.

Do ponto de vista técnico, sistemas modernos de clonagem de voz baseiam-se majoritariamente em arquiteturas neurais capazes de modelar características acústicas da fala a partir de espectrogramas, como o próprio Tacotron2 e vocoders neurais. Esses modelos já são capazes de reproduzir aspectos como timbre e fluidez, aproximando-se bem mais da naturalidade humana que o TTS mais tradicional. No entanto, muitos ainda apresentam limitações claras em relação à consistência prosódica, variação expressiva e demais micro variações da fala, elementos que podem servir como indícios para possível detecção de conteúdo sintético.

Neekhara P, et al. (2021) apresentou um trabalho intitulado “*Expressive Neural Voice Cloning*” no qual desenvolveu um modelo de clonagem de voz, fazendo o uso de redes neurais, que, com apenas algumas amostras, permitiu um controle preciso de diversos aspectos da fala, elevando o nível de expressividade e precisão da voz sintetizada. O estudo introduziu uma visão diferente sobre a clonagem de voz. Entretanto, esse tipo de método nem sempre preserva ou controla com precisão características específicas da fala do indivíduo copiado, como o tom da voz, velocidade da fala, entonação e emoção, o que é necessário para uma replicação realista da voz de um indivíduo. O estudo ainda mostrou o que deveria ser feito para preencher o gargalo da falta de naturalidade em modelos de clonagem de voz da época. Mas isso não se trata do suficiente: para propor uma voz mais natural, o artigo propõe que o ponto central da Clonagem de Voz Expressiva e Neural é justamente não apenas clonar a identidade vocal, isto é, o timbre do falante, mas adicionar expressividade ao resultado mesmo sem que o falante tenha sido usado no treino do modelo.

O artigo argumenta claramente que são necessárias mais técnicas e ferramentas para que se copie o estilo e prosódia do falante em questão. E é neste ponto que entra a Síntese de Fala Expressiva, no qual utilizaram uma arquitetura composta por três ferramentas: **Mellotron**, que permite um maior controle sobre o estilo e a prosódia da fala sintetizada; **Speaker encoder**, que tem a função de perca usada para aproximar embeddings de um mesmo locutor e afastar embeddings de locutores diferentes, permitindo representar de forma mais precisa a identidade vocal de cada falante; **Waveglow**, que converte os espectrogramas mel em ondas de áudio escutáveis.

Utilizando dessa arquitetura, os autores foram capazes de implementar duas técnicas: *Zero-Shot* e *Model Adaptation*. O *Zero-Shot*, sendo um tipo de aprendizado *n-shot*, se refere à natureza do problema no processo de aprendizado e que requer apenas dos áudios de referência para extrair o *embedding* e, logo em seguida, replicar a voz. O *Model Adaptation*, por outra via, é responsável por ajustar o modelo previamente treinado com um pequeno conjunto de dados de texto e áudio do novo falante. O trabalho de pesquisa termina concluindo que essa metodologia não foi suficiente para capturar bem a expressividade quando o modelo era treinado com diversas vozes, acabando por gerar uma fala majoritariamente neutra e que a clonagem de voz mais expressiva, mais próxima da fala original, exige mais do que meramente a identidade vocal do falante original.

Logo, com o que foi apresentado acima e dada essa movimentação mundial em prol de regulamentação, o tema se torna de alta relevância, assim como procurar entender o seu funcionamento, sua ascensão, seus limites e impactos da mesma. Para isso, o intuito deste artigo se contém em estudar métodos de clonagem de voz e sua evolução, explorando um modelo de livre acesso para testagem e pesquisa, o XTTS-v2, modelo multilíngue e zero-shot (CASANOVA E, et al., 2024). Neste trabalho, portanto, foi analisado o quanto o modelo disponível para pesquisa pode se aproximar de gravações humanas no contexto do português brasileiro, sobretudo quanto a suas características acústico-prosódicas, em diferentes condições de referência.

MÉTODOS

Para experimentação das ferramentas de clonagem neural de voz, o estudo exploratório teve como base modelos pré-treinados e um corpus próprio em português brasileiro. Os testes se iniciaram pela preparação e análise dessas gravações, providas por apenas um único falante (o próprio autor da pesquisa). Em seguida, foi utilizado um sistema neural de TTS já pré-treinado com outra voz, apenas para estudo e referência técnica. Apenas após esses dois passos, iniciou-se a clonagem *zero-shot*, a produzindo a partir de um conjunto com diferentes condições de referência da voz original. Com o intuito de encontrar similaridades com a voz humana original, as amostras produzidas foram depois comparadas às gravações originais através de descritores acústico-prosódicos, buscando analisar como podem indicar naturalidade mas sem estabelecer medidas definitivas para isso.

O corpus das vozes foi estruturado com 16 áudios no formato WAV e, para melhor organização, foram divididos em 3 subconjuntos: de referência neutra longa, usado posteriormente para condicionar os modelos de clonagem; 10 diferentes frases de avaliação, utilizadas como referência humana nas comparações com as amostras sintéticas geradas; por fim, 5 amostras expressivas de uma mesma frase, com diferentes entonações e interpretações (neutra, alegre, triste, com raiva e dramática), utilizadas apenas para caracterização inicial do corpus e não para a experimentação do modelo principal.

As gravações foram feitas em um ambiente silencioso, com baixo ruído e sem a presença de outros falantes, para que houvesse a mínima interferência. Os arquivos foram salvos com taxa de amostragem de 48kHz e codificação PCM de 16 bits, sendo tratados como mono nas etapas de processamento para os modelos utilizados. Inicialmente, todo o corpus foi auditado em relação à duração, número de canais e integridade dos arquivos.

A análise acústica também foi feita, sendo extraído descritores como a duração, energia, taxa de cruzamento por zero (ZCR), MFCCs (Coeficientes Cepstrais nas Frequências de Mel), frequência fundamental (Fo) e proporção de frames contendo voz. O Fo foi estimado com o algoritmo pYin, deixando os trechos não vozeados e de silêncio fora do cálculo da frequência fundamental. Além disso, é necessário ressaltar que, embora diversos descritores tenham sido extraídos, os essenciais para as principais comparações entre voz humana e sintética foram a duração, Fo, energia e vozeamento.

Além disso, uma amostra longa foi preparada para referência do modelo de clonagem. Após remover o silêncio e normalizar o sinal do áudio, produziu-se 3 condições de referência, de 8, 14 e 28 segundos, sendo os três trechos dessa mesma gravação. Isso foi feito justamente para comparar o desempenho do modelo tendo mais conteúdo fonético de condicionamento do modelo.

Antes da clonagem efetiva, foi utilizado um pipeline neural de Tacotron2 e Waveglow já pré-treinados e adaptados para a voz de um falante fixo, apenas com o intuito de analisar o processo da síntese do TTS neural. Em seguida se realizou o verdadeiro processo, utilizando o modelo XTTS-v2: multilingue e pré treinado, usado em clonagem *zero-shot* mas sem fine-tuning específico para o novo falante. O sistema recebeu textos em português brasileiro e uma das 3 condições de referência anteriormente preparadas.

As 10 frases de avaliação foram produzidas em cada uma das condições de referência, isto é: com áudio de 8, 14 e 28 segundos, o que resultou num total de 30 amostras do modelo

XTTS-v2. Prosseguindo, os áudios gerados passaram pela mesma extração acústica empregada nas gravações humanas e comparados com suas referências correspondentes, principalmente quanto à duração, média, variabilidade de Fo, proporção de frames vozeados do áudio e energia.

Em busca de deixar claro essas diferenças, calculou-se um índice exploratório de distância acústico-prosódica, um valor que combinou todas as diferenças já normalizadas dos descritores avaliados. Para cada par, foi calculada a diferença absoluta em duração, média e desvio-padrão de Fo, RMS e proporção de frames vozeados. Cada um dos componentes foi dividido pelo maior valor observado nas 30 amostras, e o índice se correspondeu à média (não ponderada) dos valores normalizados. Nesta normalização feita, os menores valores representaram então uma proximidade maior entre a voz humana e sua respectiva versão produzida artificialmente pelo modelo. O índice foi utilizado apenas como uma medida exploratória para a comparação e não como uma métrica definitiva para a autenticidade vocal. Importante salientar que um protocolo de avaliação perceptiva foi estruturado, porém não foram reunidas avaliações suficientes para que fizessem parte da pesquisa. Por último, devido a todas as frases terem sido geradas em cada uma das condições de referência estabelecidas anteriormente, as análises foram feitas de forma pareada, usando de estatística descritiva incluindo média, mediana, desvio-padrão e valores mínimo e máximo.

RESULTADOS

A experimentação se iniciou pela auditoria do corpus gravado, sendo processadas no total 16 gravações: 10 frases de avaliação, 5 amostras expressivas e uma referência longa. Tendo durações entre 2,88 e 30,81 segundos, os áudios possuíam inicialmente uma taxa de amostragem de 48 kHz e PCM de 16 bits. Ao terem seu Fo analisado, ressalta-se que uma das gravações teve um pico inesperado, sugerindo possível erro na estimação, porém a inspeção manual não encontrou erro evidente de estimação, logo o arquivo foi mantido juntamente aos outros áudios e utilizado para análise.

No período de testes das ferramentas de síntese neural, Tacotron2 e Waveglow, foi gerada uma amostra curta em inglês de aproximadamente 1,02 segundo. A testagem foi usada para estudo técnico do processo de TTS, todavia como o modelo usava de uma voz pré-estabelecida, a saída não foi incluída em comparações posteriores de clonagem.

Para o processo de clonagem em si, foram geradas dez frases em fala sintética pelo modelo XTTS-v2, utilizando de cada uma das condições de referência: 8, 14 e 28 segundos. Com

isso, foram produzidas 30 amostras (10 para cada uma das condições), sem que houvesse falhas de geração registradas.

Para observar e comparar conjuntamente as divergências entre as gravações originais e os áudios gerados pelo modelo, foram analisadas todas as variações de duração, média, F_0 , energia RMS e de proporção de frames vozeados. Todos os valores foram reunidos no índice (novamente, meramente exploratório) de distância acústico-prosódica anteriormente descrita.

Tabela 1: Resultados descritivos do índice exploratório de distância acústico-prosódica

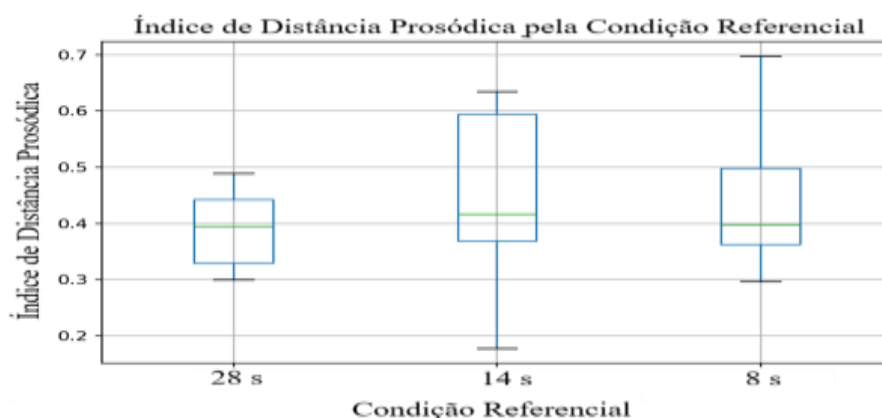
Condição referência	Média	Mediana	Desvio-padrão	Mínimo	Máximo
8 s	0,448	0,398	0,136	0,296	0,698
14 s	0,444	0,416	0,157	0,177	0,635
28 s	0,391	0,395	0,071	0,300	0,489

Fonte: elaboração própria (2026)

Na comparação dos índices gerados em cada uma das três condições, a de 28 segundos de referência apresentou a menor distância média (de 0,391), como mostrado na tabela acima. Enquanto isso, as de 14 e 8 segundos mantiveram maiores médias, respectivamente com médias de 0,444 e 0,448. Ademais, a condição mais longa obteve uma menor dispersão entre as frases, atingindo desvio padrão de 0,071, enquanto a média resultou em 0,157 e a menor em 0,136. No quesito a mediana, entretanto, os valores foram próximos: 0,398 para a condição de 8 segundos e 0,395 para a de 28 segundos.

8

Figura 1 - Gráfico *boxplot* de distribuição do índice exploratório de distância acústico-prosódica.



Fonte: elaboração própria (2026)

Ao analisar o gráfico, é notável que, nas condições de 28 segundos, houve resultados mais concentrados, enquanto as condições menores tiveram maior variação. Entretanto, mesmo que isso seja verdade em um aspecto mais amplo, a sobreposição entre as distribuições mostra que nenhuma duração obteve as menores distâncias em todas as frases.

Além de não haver unanimidade em todas as métricas gerais, houveram outros fatores limitantes nesta comparação. Uma das amostras de 14 segundos não apresentou uma estimativa válida de F_0 , outro ponto a ser denotado são diferenças consideráveis evidenciadas entre o nível de energia dos áudios originais para os artificiais, particularmente no ganho e na normalização de sinais. Dados esses fatores, os descritores devem ser interpretados com cuidado, e não de forma isolada, sendo necessária uma inspeção auditiva por parte do(s) pesquisador(es).

Ao ser realizada a inspeção auditiva dos áudios gerados, foram também notadas aproximações feitas pelo modelo quanto ao timbre tido como referência, gerando um timbre semelhante, mas não que copiasse perfeitamente a identidade vocal do falante em questão. Além disso, algumas das gerações, principalmente nas condições de 8 e 14 segundos, foram geradas com entonação certamente artificial. Essa avaliação, porém, foi realizada meramente em conjunto com a inspeção do experimento e não por um grupo de ouvintes, por isso foi considerada apenas como uma observação qualitativa.

9

Num aspecto amplo, os resultados indicam que, sob a condição de referência mais longa, foram produzidos valores mais concentrados no índice exploratório utilizado, mesmo que essa tendência não tenha sido precisamente consistente nas diferentes frases usadas como base. Mas mais do que determinar condições ideais, a experimentação evidenciou que aproximar a fala sintética da natural depende de várias características intrínsecas da voz original e do material inclusive fornecido ao modelo de clonagem. Tais diferenças são retomadas a seguir, em relação a prosódia, naturalidade e aos atuais limites da clonagem de voz.

DISCUSSÃO

Obtidos os dados da experimentação com o modelo XTTS-v2, foi possível realizar clonagem *zero-shot* nas três condições de testagem propostas e, mesmo que não tivesse sido perfeita, inclusive com apenas 8 segundos de amostragem da voz original. Entretanto, mesmo com as atuais ferramentas de clonagem sendo capazes de realizar uma clonagem *zero-shot* em um cenário como este, o modelo em questão obteve uma menor distância acústico-prosódica média e resultados, no geral, mais concentrados sob a referência de 28 segundos (a mais longa).

Esse comportamento não ocorreu de forma consistente nas diferentes fases experimentadas, o que leva a experiência a sugerir que uma maior quantidade de material de referência ainda pode contribuir significativamente para maior estabilidade de características da fala, sendo associada nesse corpus a uma menor dispersão do índice exploratório. Com isso, todavia, não é possível declarar que automaticamente é possível gerar uma clonagem mais próxima da voz humana ao se ter maiores e volumosos dados.

A variação detectada entre as condições estabelecidas denota que a acurácia e qualidade de uma clonagem de voz não podem ser atribuídas através de um único descritor. Em Neekhara P, et al. (2021), já havia a distinção entre a representação da identidade vocal do falante e os componentes ligados a estilo e prosódia, ao passo que em Bakkouche L, et al. (2025) e Bakkouche L, et al. (2026) é evidenciado o fato de que a naturalidade percebida, a equivalência com a identidade do falante original e aspectos prosódicos podem se comportar de maneiras distintas. Os descritores analisados nesta pesquisa permitiram avaliar características acústicas e prosódicas, mas não foram suficientes para afirmar com completude que uma voz soa mais natural ou não. Essa distinção é importante para a consolidação da noção de autenticidade vocal discutida neste trabalho.

Apesar disso, mesmo com as limitações encontradas durante o processo de pesquisa e experimentação, a geração de amostras através de referências curtas demonstra o quanto os sistemas atuais já evoluíram, sendo capazes de gerar uma fala condicionada a um novo falante através de poucas amostras. Tais resultados correspondem à atual tendência observada na literatura recente, marcada pelos avanços significativos de clonagem *few-shot* e *zero-shot*, além da redução cada vez maior da quantidade de dados necessária para se adaptar um modelo a novas vozes (AZZUNI H e EL SADDIK A, 2025).

Assim, realizar esse processo tendo pequenas amostras de referência auxilia na compreensão do porquê o avanço da clonagem de voz não é meramente discutido como um ganho técnico. Em demais campos e, sobretudo, o artístico, especialmente em atividades que dependem mais diretamente da voz, Tsiavos V e Kitsios F (2025) identificam que seu uso também levanta questões sobre autoria, autenticidade, privacidade e, possivelmente, a substituição de profissionais. Isso reforça a necessidade de critérios éticos serem mais discutidos, bem como formas de proteção para o trabalho humano

No artigo e também podcast “The Voice Actor and Their Double”, o ator e dublador americano Adam Paul relata sobre sua experiência de clonar sua própria voz com as melhores

ferramentas que tinha à disposição, e teve resultados os quais ele mesmo diz ter “se convencido pelo clone”, e isso ainda em 2023. O mesmo relata que a voz clonada soa até “objetivamente melhor, mais agradável” que a própria dele, de forma mais limpa e clara, e também admite que a diferença entre a voz orgânica e a sintética já é muito pequena nos dias atuais (PAUL A, 2023). Paul argumenta que a tecnologia não possui bagagem subjetiva ou emocional de um ser humano e jamais poderia substituir um ator real, mas ainda assim não vê a tecnologia como antagonista à arte, destacando-a como bem útil para a vida dos profissionais. Todavia, enfatiza que as IAs necessitam de um direcionamento mais ético, deixando para o jurídico a responsabilidade do controle autoral da voz do artista (PAUL A, 2023). Tal depoimento demonstra novamente o avanço da qualidade da clonagem de voz nos dias atuais.

Seguindo os dados atuais sobre o crescimento da clonagem de voz e o depoimento do ator hollywoodiano, é nítida a necessidade de regulamentação das IAs nesse âmbito, bem como a proteção de vozes. Essa luta por maior regulamentação vem ocorrido inclusive aqui no Brasil, como pode ser visto pela Dublagem Viva, um movimento lançado em janeiro de 2024 por profissionais da área de dublagem: dubladores, diretores, tradutores, mixadores e outras funções correlatas, justamente para defender a regulamentação do uso de IA no setor audiovisual. No manifesto, o grupo pede que modelos de IA não sejam usados para substituir dubladores no setor, também defende o consentimento e remuneração pelo uso de vozes, proteção dos direitos autorais e conexos, sem contar a preservação da língua portuguesa brasileira e participação de todos os agentes do setor na construção das regras (DUBLAGEM VIVA, 2026). Tamanha mobilização, que repercutiu inclusive internacionalmente ao ser reconhecida por países europeus, mostra que, com o avanço da acessibilidade dos sistemas de clonagem de voz, a discussão deixa de se resumir à semelhança acústica da voz sintética com a original, passando a incluir origem e autorização para utilização da mesma.

Em Oliveira JF (2024), é tratado diretamente sobre os efeitos da IA generativa na indústria fonográfica, com foco na mimetização de vozes de artistas, trazendo detalhes intrínsecos sobre os mesmos. A advogada afirma que a voz é um bem juridicamente sensível e economicamente central, citando o art. 90 da LDA (Lei de Direitos Autorais), que assegura ao profissional direito de proibir o uso de sua voz e interpretação (BRASIL, 1998). Além disso, pode-se relacionar essa discussão à Lei Geral de Proteção de Dados (LGPD), principalmente quando características da voz são usadas como dados biométricos para identificação de uma pessoa (BRASIL, 2018). Dessa forma, a clonagem e uso de vozes sem consentimento fortalecem

a importância de, além de regulamentar, serem criados mecanismos que sejam capazes de averiguar a autenticidade da voz.

Portanto, é evidente que as mesmas características que tornam a síntese mais próxima da fala humana (acústicas, prosódicas, cepstrais ou mesmo neurais) também criam novos desafios para a identificação de conteúdo artificial.

Pesquisas mais recentes na área de detecção de *spoof* (ou *deepfake*) mostram que a detecção tem ascendido para mais do que simples aspectos da síntese artificial, considerando diversos descritores acústicos, prosódicos e demais que possam estar ligados a naturalidade da fala humana. Kim T, et al. (2025), como exemplo mais recente, colocam a naturalidade como uma informação útil para melhorar os detectores de *deepfake*. Nesse sentido, as diferenças acústico-prosódicas analisadas neste estudo não representam um método de detecção, mas evidenciam características que podem ser mais investigadas para auxiliar as atuais técnicas de identificação de fala sintética.

CONSIDERAÇÕES FINAIS

Ao decorrer de todo o processo da pesquisa foi estudada a progressão dos métodos de clonagem de voz, desde os primeiros indícios de TTS com elementos prosódicos até a modelos atuais de alta precisão e baixa necessidade de grandes amostras de dados como o TTS Neural, Zero, One e Few Shot. Essa análise histórica foi essencial para denotar a grande evolução das técnicas de clonagem de voz, seus principais elementos e quais características da fala se distanciam ou se aproximam da voz original.

Diante do atual estágio de desenvolvimento tecnológico dessas ferramentas, o problema central deste trabalho consistiu em investigar de que forma as características (acústicas e prosódicas) da fala podem ser replicadas, podendo possivelmente servir como indicadores para a otimização de modelos de clonagem e para a identificação de vozes em próximos estudos. Na experimentação, foi realizado um estudo exploratório em português brasileiro utilizando o XTTS-v2, com o qual foram geradas 30 amostras a partir de três condições estabelecidas como referência, com durações de aproximadamente 8, 14 e 28 segundos. Durante o processo de comparação acústico-prosódica, a condição mais longa apresentou a menor distância média e a menor dispersão no índice exploratório utilizado.

Os resultados dos experimentos também colocam em vista que a autenticidade vocal não pode simplesmente ser determinada a partir da proximidade de características acústicas ou

prosódicas isoladas, como duração, Fo e energia da voz. O índice que foi utilizado durante o desenvolvimento deste trabalho possui caráter exploratório e não mede diretamente identidade vocal ou naturalidade percebida, porém demonstra a viabilidade técnica da clonagem *zero-shot* com ferramentas acessíveis e pequenas amostras de voz e também expõe limites importantes na comparação entre fala natural e sintética.

Nos dias atuais, a crescente acessibilidade dessas tecnologias amplia as discussões, que ultrapassam seu desempenho técnico, envolvendo questões legais como o consentimento, a autoria, o uso da identidade vocal e a proteção do trabalho de profissionais da voz. Além disso, movimentos como a Dublagem Viva e a literatura recente mostram que a autenticidade também envolve compreender a origem e a autorização de uma voz, e não somente o quanto ela se aproxima acusticamente de um falante. Assim sendo, neste trabalho torna-se evidente que a investigação futura das características acústicas, prosódicas e de identidade pode contribuir para estudos de detecção de fala sintética e deepfakes de áudio, especialmente no português brasileiro.

AGRADECIMENTOS E FINANCIAMENTO

Este trabalho foi financiado pelo Centro de Ciências Agrárias da UFSCar (CCA) - projeto FAI RTI-CCA.

13

REFERÊNCIAS

- AZZUNI H, EL SADDIK A. Voice Cloning: Comprehensive Survey. arXiv, 2025, 579p.
- BAKKOUCHE L, et al. Finding the Human Voice in AI: Insights on the Perception of AI-Voice Clones from Naturalness and Similarity Ratings. In: Proceedings of Interspeech 2025, 2025: 2190–2194.
- BAKKOUCHE L, et al. What determines the success of AI voice-cloned speech? Prosodic and acoustic evidence on three TTS systems. *Phonetica*, 2026: 1-21.
- BARAKAT H, et al. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024; 2024(11): 1-34.
- BRASIL, 1998. LEI nº 9.610, de 19 de fevereiro de 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências. Brasília. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9610.htm. Acesso em: 9 set. 2026.
- BRASIL, 2018. LEI nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 9 set. 2026.

CASANOVA E, et al. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In: Proceedings of Interspeech 2024. 2024: 4978–4982.

CHENG S, et al. A Survey of Grapheme-to-Phoneme Conversion Methods. Applied Sciences, 2024; 14(24): 11790.

DUBLAGEM VIVA. MANIFESTO pela regulamentação da inteligência artificial na dublagem. [s. d.]. Disponível em: <https://dublagemviva.com.br/index.php/manifesto/>. Acesso em: 9 set. 2026.

JURAFSKY D, MARTIN JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed. Upper Saddle River: Pearson Prentice Hall, 2009; 988p.

KIM T, et al. Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection. In: Proceedings of Interspeech 2025. 2025; 5318–5322.

LADD DR. Defining prosody. Simultaneous Structure in Phonology. Oxford: Oxford University Press, 2014; 200p.

NEEKHARA P, et al. Expressive Neural Voice Cloning. in: Proceedings of the 13th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, 2021; 157: 252–267.

OLIVEIRA JF. A mimetização de vozes por inteligência artificial veiculadas no TikTok e as repercussões em direitos autorais na indústria fonográfica. Trabalho de Conclusão de Curso (Bacharelado em Direito) - Faculdade Nacional de Direito, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2024; 81 p.

PAUL A. The Voice Actor and Their Double: Working as a Voice Actor and Teaching Voice Acting in the Age of AI Voice Cloning. Tradition Innovations in Arts, Design, and Media Higher Education, 2023; 1(1), 1-19.

RUSSELL, Stuart J.; NORVIG, Peter. Artificial Intelligence: A Modern Approach. 4th ed. Pearson, 2022; 1016p.

TSIAVOS V, KITSIOS F. The digital transformation of the film industry: How Artificial Intelligence is changing the seventh art. Telecommunications Policy, 2025; 49(8): 103021.

WEITZMAN C. HOW to create an AI voice from a sample. Speechify. Disponível em: <https://speechify.com/blog/how-create-ai-voice-from-sample/>. Acesso em: 9 set. 2026.