Recuperação Vocal de Pacientes Traqueostomizados Utilizando Processamento de Áudio e Aprendizado de Máquina em Cenário Zero-Shot

Mario Pinto Freitas Filho, João Dallyson Sousa de Almeida, Geraldo Braz Junior
mario.freitas@discente.ufma.comjoao.dallyson@ufma.br,geraldo@nca.ufma.br
Universidade Federal Do Maranhão (UFMA)
Programa de Pôs-Graduação em Ciência da Computação (PPGCC)

ABSTRACT

This work proposes a method for voice reconstruction in individuals who have undergone laryngectomy, integrating advanced audio processing and machine learning techniques. The approach aims to restore features such as timbre, intonation, and prosody, which are often lost when using an electronic larynx, whose sound is constrained by a constant fundamental frequency (F0). To address the lack of public datasets containing voices of tracheostomized patients, a synthetic dataset was created to simulate the acoustic properties of these devices. The developed pipeline comprises three stages: (i) speech analysis, involving the extraction of linguistic content and style; (ii) mapping, combining this information with the mel-spectrogram through techniques such as conditional modulation and diffusion networks, with a particular focus on Flow Matching; and (iii) reconstruction and synthesis, using high-fidelity vocoders. Experiments compared two preprocessing methods—timbre shifter and F0 fixation—evaluated in four training and testing combinations. Results show that the F0 → F0 configuration outperformed the alternative in three out of the four analyzed metrics (MCD of 444.04, LSD of 0.47, and PSNR of 42.27), suggesting that F0 fixation favors voice reconstruction that more closely matches the original signal. These findings highlight the potential of the proposed approach to improve the naturalness and intelligibility of synthesized speech for laryngectomized patients.

KEYWORDS

Saúde Digital e Tecnologias Assistivas, Processamento de Linguagem Natural, IA, Aprendizado de Máquina e Aprendizado Profundo, Acessibilidade

1 INTRODUÇÃO

A reconstrução da voz envolve a replicação das características sonoras de um indivíduo, como timbre, estilo, tonalidade, entonação e inflexões, em áudios distintos. Embora essa técnica possa ser utilizada de maneira ilícita, como em casos de falsificação de identidade, ela também possui um grande potencial para beneficiar pessoas que perderam a capacidade de falar. Um dos principais grupos que se beneficiam dessa tecnologia são os pacientes que passaram por laringectomia, uma cirurgia que resulta na remoção da laringe. Como consequência, esses indivíduos perdem a estrutura necessária para produzir a voz naturalmente e precisam recorrer a dispositivos auxiliares para se comunicar. Entre 2011 e 2020, foram registradas 172.456 traqueostomias em pacientes hospitalizados pelo Sistema Único de Saúde (SUS), representando cerca de 0,15% de todas as internações nesse período [19].

Um exemplo desse dispositivo é a laringe eletrônica, que tenta simular as funções da laringe humana por meio de vibrações geradas pelo aparelho, que se movem com os músculos do pescoço. Contudo, o som produzido por esse dispositivo tende a ser robótico e metálico, uma característica resultante da frequência fundamental (F0) constante gerada pelo aparelho, impedindo a variação natural da voz, deixando-a artificial e monótona. A aplicação de técnicas de processamento de áudio e aprendizado de máquina pode, portanto, representar um avanço significativo na restauração da voz original desses pacientes, proporcionando uma forma mais natural e fiel de comunicação.

A recuperação da voz, para ser eficaz, segue um pipeline de conversão dividido em três estágios principais: Análise da Fala (*Speech Analysis*), Mapeamento (*Mapping*) e Reconstrução e Síntese *(Reconstruction & Synthesis*) [4].

No estágio de Análise da Fala, o objetivo é extrair o conteúdo linguístico, ou seja, identificar o que está sendo dito. Esse processo envolve a extração dos *tokens* da fala diretamente do áudio. Para isso, são utilizadas redes neurais específicas, como o *WAV2VEC2* [3] ou *WavLLM* [7], eficazes para essa tarefa. Além disso, essa fase também abrange a extração do *pitch* e da prosódia, frequentemente referida como extração de estilo. A pergunta aqui é: como a fala soa? Isso inclui o tom, a entonação e o ritmo da fala, e pode ser realizado utilizando métodos como o CREPE [14], ou ainda redes neurais como *StyleGAN* [13] ou *CAMplus* [21].

No estágio de Mapeamento, é feita a combinação da extração do conteúdo linguístico com o estilo de fala, juntamente com a adição de um mel-espectrograma, que servirá como entrada para a rede neural. A junção dessas informações pode ser realizada por meio de técnicas como Concatenação e Camada Linear, Modulação FiLM (*Feature-wise Linear Modulation*) [20] ou *Adaptive Instance Normalization* (AdaIN) [12]. A saída da rede nesse estágio é o mel-espectrograma. Para realizar essa codificação, utiliza-se uma rede neural que recebe como entrada o conteúdo linguístico, o estilo extraído e o mel-espectrograma concatenados. A partir dessa codificação, utiliza-se uma rede de difusão para a reconstrução do áudio. Entre os modelos utilizados para essa tarefa, estão redes generativas como GANs, Variational Autoencoders (VAEs) [4], mas o mais recente, usando o uso do Flow Matching condicional [17] vem se mostrado bem promissor.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

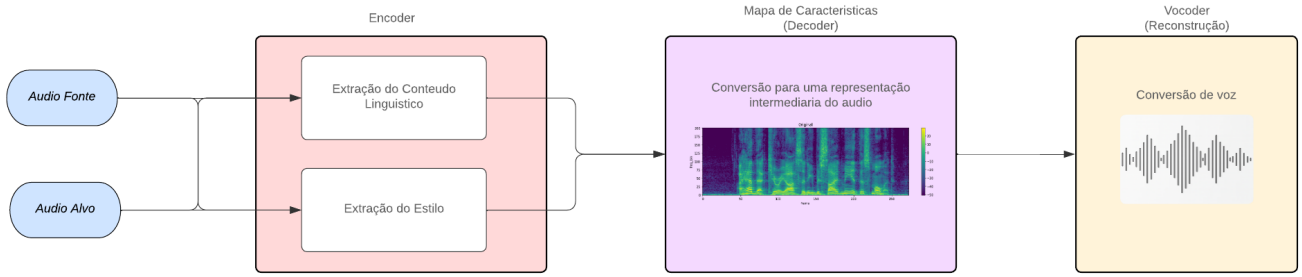

Figure 1: Pipeline de Conversão de Voz

No estágio final de Reconstrução e Síntese, após a obtenção do mel-espectrograma reconstruído, é utilizado um *vocoder* para converter essa representação em um sinal de áudio. *Vocoders* modernos, como o *HiFi-GAN* [15], são treinados para gerar ondas sonoras de alta fidelidade, com qualidade realista. Essa conversão é essencial para garantir que a fala sintetizada tenha uma qualidade natural e inteligível, essencial para a eficácia do processo de recuperação da voz.

Como não foi encontrada na literatura nenhuma base de dados com áudios de pessoas traqueostomizadas, desenvolvemos um dataset sintético inédito para suprir essa lacuna. Esses áudios foram gerados de forma a simular as características sonoras da laringe eletrônica.

A principal contribuição deste estudo é a composição de um método que integra avanços recentes em processamento de áudio e aprendizado de máquina para a reconstrução de voz de indivíduos submetidos à laringectomia. O método combina técnicas de análise de fala, mapeamento de estilo e prosódia, e reconstrução de áudio com redes neurais de difusão — em especial o Flow Matching condicional para gerar uma voz sintetizada com maior proximidade à voz original. Além disso, conduzimos uma avaliação da qualidade da fala reconstruída, utilizando vocoders modernos para garantir a fidelidade e a naturalidade do áudio final.

A estrutura deste trabalho está organizada conforme descrito a seguir: a Seção 1 aborda a motivação e os objetivos do estudo. Na Seção 2, é apresentado o referencial teórico que fundamenta o método adotado. A Seção 4 descreve em detalhes a base de documentos utilizada e o método empregado. Os resultados da sumarização extrativa automática são discutidos na Seção 5. Por fim, a Seção 6 expõe as conclusões do estudo.

2 REFERENCIAL TEÓRICO

2.1 Características da Voz

A voz humana possui características acústicas essenciais como frequência fundamental (F0), intensidade (volume) e timbre, que são influenciadas pela anatomia do trato vocal, pulmões e pregas vocais. Além dessas propriedades, há diferentes qualidades vocais, incluindo voz modal (normal), creaky (rangida), breathy (sussurrada), tense (tensa) e lax (relaxada), que influenciam a percepção da fala. As variações nessas características permitem que cada pessoa tenha uma identidade vocal única e possibilitam a expressão de emoções e intenções. A prosódia, com elementos como ritmo, entonação e acento, complementa a voz ao atribuir significado emocional e linguístico às frases, contribuindo para a comunicação eficaz e compreensível [8].

2.1.1 Frequência Fundamental. A frequência fundamental de um sinal de fala, frequentemente denotada por F0, refere-se à frequência aproximada da estrutura (quase) periódica dos sinais de fala sonora. A oscilação se origina das pregas vocais, que oscilam no fluxo de ar quando adequadamente tensionadas. A frequência fundamental é definida como o número médio de oscilações por segundo e expressa em Hertz. Como a oscilação se origina de uma estrutura orgânica, ela não é exatamente periódica, mas contém flutuações significativas. Em particular, a quantidade de variação no comprimento do período e na amplitude são conhecidas respectivamente como jitter e shimmer . Além disso, o F0 normalmente não é estacionário, mas muda constantemente em uma frase [5].

2.2 Espectrograma

O espectro de Fourier de um sinal revela seu conteúdo de frequência, isso torna um espectro um domínio intuitivo e agradável para se trabalhar, pois podemos examinar os sinais visualmente. Na prática, trabalhamos como sinais discretos de modo que a transformação tempo-frequência corresponde à Transformada Discreta de Fourier. Ela mapeia um comprimento N sinal X_n em uma representação de domínio de frequência de valor complexo X_k de N coeficientes como [5]:

$$X_k = \sum_{n=0}^{N-1} x_n e^{i2\pi \frac{kn}{N}} \tag{1}$$

Para entradas de valor real, os componentes de frequência positiva e negativa são conjugados complexos uns dos outros, de modo que retemos unidades únicas de informação. No entanto, como os espectros são vetores de valores complexos, é difícil visualizá-los como tal, pois sinais de fala são sinais não estacionários. Se transformarmos uma frase falada para o domínio da frequência, obteremos um espectro que é a média de todos os fonemas da frase, enquanto gostaríamos de ver o espectro de cada fonema individual [5].

Para isso, ao dividir o sinal em fragmentos menores, podemos nos concentrar nas propriedades do sinal em um determinado momento. Ao aplicar a janela e a Transformada Discreta de Fourier (DFT) em cada janela, obtemos a Transformada de Fourier de curto

prazo (STFT) do sinal, para um sinal de entrada x_n e janela wn' a transformação é definida como [5]:

$$STFT(x_n)(h,k) = X(h,k) = \sum_{n=0}^{N-1} x_n e^{i2\pi \frac{kn}{N}} \quad (2)$$

A STFT é uma das ferramentas mais utilizadas em análise e processamento de fala. Ela descreve a evolução dos componentes de frequência ao longo do tempo. Outro paralelo com um espectro é que a saída da STFT é de valor complexo, embora, onde o espectro é um vetor, a saída da STFT seja uma matriz. Como consequência, não podemos visualizar diretamente a saída de valor complexo. Em vez disso, as STFTs são geralmente visualizadas usando seus espectros logarítmicos, $20log_{10}(X(h,k))$ Esses log-espectros bidimensionais podem então ser visualizados com um mapa de calor conhecido como espectrograma[5].

2.3 Extração do conteúdo Semântico

A extração do conteúdo linguístico no processamento de áudio acontece no espectrograma. O sinal de fala é representado por características de envelope espectral y(t). como o spectro log-mel. dentro de cada y(t) é passado essa fatia para uma rede neural como WavLLM[7] que ira extrair as características do y(t), $t \in [\ldots, t_0 - 2, t_0 - 1, t_0]$. O resultado desse processo é uma representação linguística de todo o espectrograma, denotada como c(t_0). Esse vetor gerado é então projetado linearmente, servindo como um token, que representa a informação linguística extraída do sinal de fala para posterior processamento.

2.4 Extração de Estilo

Diferentemente da extração do conteúdo linguístico, que ocorre no mel-espectrograma, a extração de estilo acontece diretamente no áudio, pois visa capturar características relacionadas à forma como o fonema é articulado, como entonação, prosódia, inflexões e outras propriedades fonéticas. Essas características estão diretamente ligadas ao modo como a fala é produzida, ou seja, a maneira única de cada pessoa pronunciar palavras e frases. A extração de estilo é realizada por redes treinadas especificamente para reconhecer a voz de indivíduos, extraindo suas características únicas diretamente do áudio.

Uma das redes usadas para essa tarefa é o CAMplus [21], composta por duas partes principais. A primeira parte é o front-end, que utiliza convoluções 2D com conexões residuais, sendo responsável por extrair características detalhadas no domínio tempo-frequência. A segunda parte é baseada no D-TDNN (Deep Time Delay Neural Network) com Context-Aware Masking, uma versão densa da TDNN, chamada D-TDNN. Cada camada do D-TDNN contém um módulo CAM (Context-Aware Masking).

O CAM é um mecanismo de atenção que foca nas partes mais relevantes do áudio para identificar o locutor, ignorando ruídos e variações irrelevantes. A versão CAM++ aprimora o CAM original, funcionando como um mecanismo de atenção, mas com um foco específico em "mascarar" as partes irrelevantes do sinal e destacar as características essenciais para identificar o locutor. O CAM aplica uma máscara de atenção sobre os mapas de características gerados pelas camadas D-TDNN, otimizando a extração de estilo e permitindo uma identificação mais precisa da voz de diferentes indivíduos.

2.5 Condicional Flow Matching

Seja x_1 uma variável aleatória distribuída de acordo com uma distribuição de dados desconhecida $q(x_1)$. Assumimos que temos acesso apenas a amostras dessa distribuição $q(x_1)$, mas não à função de densidade em si. Além disso, consideramos um caminho de probabilidade p_t tal que $p_0 = p$ é uma distribuição simples, por exemplo, a distribuição normal padrão, $p(x) = N(x|0, I)$, e que p_1 é aproximadamente igual, em distribuição, a $q(x)$. Posteriormente, discutiremos como construir tal caminho. O objetivo do *Flow Matching* é alinhar este caminho de probabilidade alvo, permitindo-nos fluir p_0 para p_1 [17].

O *Conditional Flow Matching* (CFM) é uma versão prática do *Flow Matching* (FM). Em vez de trabalhar com probabilidades marginais e campos vetoriais marginais, que são difíceis de calcular, o CFM foca em probabilidades condicionais e campos vetoriais de probabilidades [17].

Para cada amostra dos dados reais $x_1 \sim q(x_1)$, onde x_1 é um valor no conjunto de dados e $q(x_1)$ é a densidade probabilística desse valor, o CFM define uma trajetória de probabilidade que começa no ruído $p_0(x \mid x_1) = p(x)$ e termina em torno de x_1. Essa trajetória visa aproximar a distribuição do dado real a partir de um processo de geração gradual [17].

Ao invés de integrar sobre toda a distribuição marginal $q(x)$, o CFM simplifica o processo ao amostrar um dado x_1 diretamente e calcular a perda associada a essa amostra. Isso torna o cálculo mais viável do ponto de vista computacional, pois evita o custo de calcular a distribuição marginal completa, tornando o processo mais eficiente para tarefas de aprendizado e reconstrução de dados [17].

2.5.1 Transporte Ótimo (OT). O Transporte Ótimo (OT) é um conceito da teoria das probabilidades e matemática que trata de como "mover" uma distribuição de probabilidade de uma forma para outra, minimizando o custo desse movimento.

Agora, no contexto do Flow Matching, estamos usando o conceito de OT para transformar uma distribuição simples (como o ruído) em uma distribuição mais complexa, como uma imagem. Os Campos Vetoriais Condicionais de OT fazem parte desse processo de transformação.

Um campo vetorial condicional é simplesmente um campo que depende de um dado específico. Ou seja, a direção e a magnitude do movimento de cada ponto de dados (neste caso, as amostras) ao longo do processo de transformação são determinadas com base nas informações do dado de entrada (como uma imagem ou um vetor de características).

2.6 Vocoder

O termo vocoder vem das palavras voz e codificador, e refere-se ao processo de aplicação de uma característica semelhante à fala a um som. A ideia básica do vocoder é que qualquer som que tenha uma estrutura formante será percebido como um som de fala. Assim, se modificarmos um som de forma que ele tenha picos no espectro semelhantes a um som de fala, ele será percebido como um som de fala. Muitas das características originais do som serão preservadas,

mas ele terá a interpretação adicional de um som de fala. É como se o som original se tornasse o tom portador do sinal de fala [5].

Os formantes são picos no envelope espectral do sinal. Ou seja, se observarmos a forma bruta (ou macro) do espectro de magnitude de um sinal de fala, ele apresentará um pequeno número de picos, especialmente na região entre 300 e 3500 Hz. A localização (e a amplitude) desses picos corresponde a vogais específicas ou, inversamente, cada vogal possui uma constelação única e identificadora de picos formantes [5].

Vocoders modernos utilizam redes adversárias generativas (GANs) ou autoencoder para a síntese de voz, convertendo mel-espectrogramas em ondas sonoras brutas de alta qualidade. Esses vocoders são principalmente empregados na construção da fala a partir de texto, um processo conhecido como TTS (Text-to-Speech). Nesse contexto, o modelo gera um mel-espectrograma a partir de um texto de entrada, e o vocoder transforma esse espectrograma em um sinal de áudio. O vocoder é treinado para reproduzir as propriedades do falante, garantindo que a voz sintetizada tenha as características desejadas, como entonação, timbre e estilo de fala.

O HiFi-GAN [15] é um exemplo notável desse tipo de vocoder, utilizando redes adversárias para gerar ondas sonoras realistas, oferecendo uma conversão de alta fidelidade que preserva as nuances da voz, essencial para sistemas de síntese de fala natural e com boa qualidade perceptual. A combinação de redes adversárias e vocoders permite uma síntese de voz mais fluida e expressiva, sendo um avanço significativo em comparação com métodos anteriores.

3 TRABALHO RELACIONADOS

Na literatura, são poucos os trabalhos que abordam especificamente a recuperação de voz em pessoas que utilizam a laringe eletrônica. Um desses trabalhos é o de [22], que, na ausência de um dataset específico, optou por criar um dataset sintético. Esse dataset foi desenvolvido a partir da premissa de que o som produzido pela laringe eletrônica possui um F0 constante, ou seja, uma frequência fundamental que permanece inalterada, resultando em oscilações uniformes ao longo do tempo.

Com base nessa característica, [22] utilizou um dataset de áudios e realizou um pré-processamento para tornar todos os F0 constantes, simulando, assim, o som gerado pela laringe eletrônica. Esse procedimento permitiu que os áudios gerados se assemelhassem bastante àqueles produzidos por pessoas que utilizam esse dispositivo. A partir dessa base de dados, foi possível realizar a transferência de áudio, ou seja, converter a voz com F0 constante para a voz original do falante, restaurando, assim, aspectos mais naturais e individuais da fala. Essa abordagem abriu portas para métodos mais avançados de recuperação de voz, contribuindo para pesquisas nesse campo.

3.1 SEED-VC

A arquitetura Seed-VC foi concebida para superar os principais desafios do voice conversion em cenários zero-shot, especialmente problemas como vazamento de timbre (timbre leakage), representação insuficiente do timbre e a discrepância entre os processos de treinamento e inferência. O cerne da Seed-VC é um framework que combina um transformador baseado em difusão (diffusion transformer) com estratégias de manipulação do timbre durante o treinamento.

No início do pipeline, é utilizada uma abordagem de flow matching, em que o treinamento é orientado para alinhar as distribuições de características acústicas entre o áudio de origem e o de destino. Isso é feito por meio de um campo vetorial dependente do tempo, aprendido pela rede, que gradualmente transforma as características da origem nas do destino, promovendo um caminho suave entre essas distribuições.

O principal componente da Seed-VC é o diffusion transformer, um modelo transformador com múltiplas camadas e atenção multicabeças, otimizado para o processo de denoising do esquema de difusão. Esse transformador incorpora avanços como skip connections ao estilo U-Net (sem downsampling da sequência temporal) e embeddings posicionais rotatórios para melhorar a generalização. Além disso, a informação do tempo é inserida tanto como token de prefixo quanto como elemento de normalização adaptativa nas camadas do transformador, permitindo à rede modelar adequadamente a evolução temporal do processo de difusão.

Um diferencial central da Seed-VC é o uso de um timbre shifter externo durante o treinamento. Esse módulo que pode ser implementado por modelos de voice conversion em regime zero-shot, como o OpenVoice ou o AutoVC, modifica o timbre do áudio de origem, de modo que o extrator de conteúdo semântico opere sobre um sinal livre do timbre original do locutor. Essa estratégia minimiza o vazamento de timbre e torna as condições de treinamento mais próximas às de inferência, nas quais o conteúdo e o timbre são, de fato, provenientes de fontes distintas.

4 MATERIAIS E MÉTODO

Neste trabalho, usaremos como extrator de conteúdo linguístico o XLSR [2], para a extração do estilo sera o CAM++ [21], como arquitetura principal que receberá todas essas informações temos o SEED-VC [18]. O seed vc é usado em conjunto com o condicional flow matching para a reconstrução do mel espectrograma.

4.1 Arquitetura do Modelo

Um dos objetivos dessa arquitetura é possibilitar o funcionamento em regime one-shot, ou seja, realizar a conversão de voz a partir de somente uma amostra de áudio na etapa de inferência. No contexto de voice conversion, essa abordagem envolve a distinção entre duas arquiteturas principais: uma projetada para o treinamento e outra voltada para a inferência.

A arquitetura de treinamento tem como objetivo ensinar o modelo a reconstruir o mel-espectrograma do áudio-alvo. Para isso, é necessário apenas um único áudio de entrada, a partir do qual são extraídas as informações de estilo e conteúdo linguístico. O modelo é treinado para gerar o mel-espectrograma correspondente ao conteúdo da fala, condicionado a esses embeddings.

Por outro lado, a arquitetura de inferência requer dois áudios distintos: o áudio-fonte, de onde se extrai o conteúdo semântico da fala, e o áudio-alvo, utilizado para obter o estilo vocal desejado. Após a geração do mel-espectrograma com base nesses condicionamentos, o resultado é passado por um vocoder, que realiza a conversão do espectrograma gerado em forma de onda, produzindo assim o áudio final convertido com o estilo do falante-alvo.

O fluxo de treinamento tem início com o processamento do áudio-fonte, a partir do qual são extraídas duas informações fundamentais:

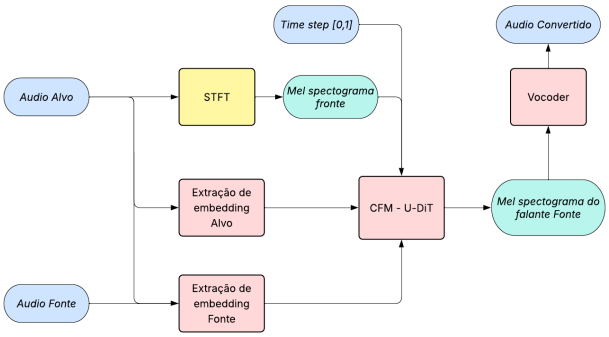

Figure 2: Estrutura do método. Essa mesma estrutura também é usada na inferência. O método tem 2 blocos principais para a extração tanto do estilo quanto do conteúdo linguístico. O bloco "extração de embeddings alvo" acontece a extração de estilo e extração do conteúdo linguístico, visto com detalhe na figura 4. Isso acontece porque queremos que o áudio final tenhas as caracterizas estelicas da voz do áudio alvo com o conteúdo linguístico do áudio fonte, logo, a "extração de embeddings fonte" se concentra em extrair o conteúdo linguístico do audio fonte como mostrado na figura 5

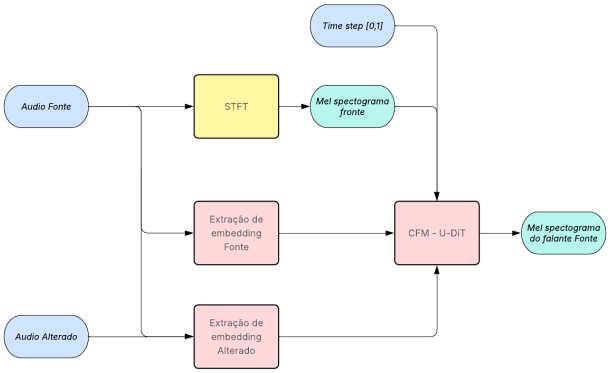

Figure 3: Estrutura usada durante o treinamento do método

o estilo do falante, por meio de um extrator de estilo baseado em [21], que gera um embedding representando características vocais; e o conteúdo linguístico, obtido como um embedding semântico. Além disso, o áudio é convertido para o domínio do mel-espectrograma por meio de uma transformação via STFT. Em seguida, o sinal é submetido ao timbre shifter, responsável por remover as características do timbre original da voz, mantendo apenas o conteúdo.

Os embeddings de estilo e de conteúdo linguístico são então concatenados com a codificação do passo temporal e utilizados como entrada para o modelo, como mostrado na Figura 4.

A partir do áudio modificado (sem timbre), é novamente extraído o embedding de conteúdo linguístico, o qual será utilizado na função de condicionamento no bloco cond. Essa função é composta predominantemente pelos embeddings do áudio alterado,

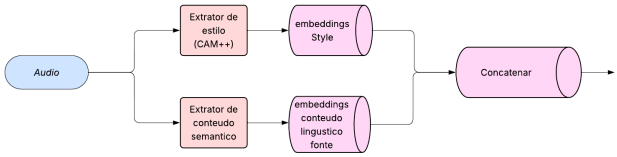

Figure 4: Estrutura Interna do bloco Extração de embeddings alvo obs: Esse mesmo bloco é usado na inferência como o bloco de nome "Extração de Embedding Alvo"

mas com uma parte inicial — definida por prompt len — substituída pelos embeddings correspondentes do áudio original. Essa substituição tem como objetivo guiar o modelo na reconstrução do mel-espectrograma do áudio original, assegurando que o conteúdo semântico da fala seja preservado.

O valor de prompt len é amostrado individualmente para cada instância do batch, variando entre 1 e a largura temporal do mel-espectrograma. Ele define quantos tokens do início do cond serão provenientes da fala original. Além disso, em 10% dos casos, prompt len é igual a zero, simulando um cenário de geração completamente não-condicionada.

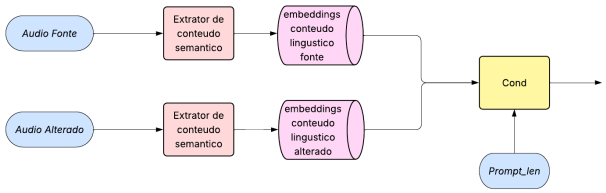

Figure 5: Estrutura Interna do bloco extração de embeddings fonte obs: Esse mesmo bloco é usado na inferência com o nome de "Extração de Embedding Fonte", sendo o áudio fonte o áudio alvo e o áudio alterado o áudio fonte

Com as informações do áudio fonte concatenadas, os embeddings do áudio alterado configurados e o mel espectrograma extraído, podemos alimentar a arquitetura U-DiT, que será responsável pela reconstrução do mel espectrograma a partir desses dados. A ground truth utilizada nesse processo é o mel espectrograma do áudio fonte. Como o U-DiT é baseado em Conditional Flow Matching, ele requer a definição de time steps para realizar a reconstrução progressiva, partindo de uma distribuição Gaussiana até alcançar a estrutura final do mel espectrograma.

A camada do U-DiT é composta por 13 blocos DiTs, usando conexões que lembram a U-net

No bloco DiT, o passo atual da reconstrução é processado por uma MLP [10], cuja saída é utilizada nas camadas de scale e shift para condicionar cada etapa ao time step correspondente do processo de reconstrução do mel espectrograma.

A entrada do bloco DiT passa inicialmente por uma camada de normalização, que é então condicionada ao passo atual. Em seguida, essa saída é processada por um módulo de Multi-Head

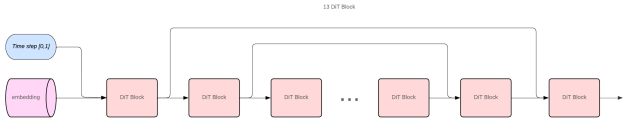

Figure 6: Estrutura do U-DiT com 13 blocos

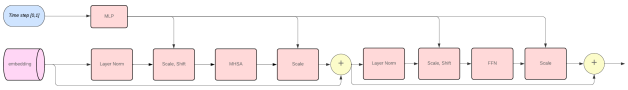

Figure 7: Estrutura Interna do DiT Block

Self-Attention (MHSA) [6], cuja saída também é condicionada ao mesmo passo de reconstrução. Essa saída condicionada é então concatenada com a entrada original (os embeddings) e novamente normalizada com base no time step atual.

Após isso, os dados passam por uma camada Feed-Forward Network (FFN), seguida por outra etapa de condicionamento via scale e shift, e por fim são novamente concatenados com a última concatenação feita, preservando as informações acumuladas ao longo do bloco.

Na fase de inferência, são necessários dois áudios distintos: o áudio-fonte, do qual se extrai o conteúdo linguístico (ou seja, o que está sendo dito), e o áudio-alvo, que fornece o estilo vocal (ou seja, como está sendo dito). A arquitetura utilizada nesta etapa mantém a mesma estrutura empregada no treinamento, com a diferença de que, ao final do processo, o mel-espectrograma gerado que combina o conteúdo do áudio-fonte com o estilo do áudio-alvo é passado por um vocoder. Esse vocoder realiza a reconstrução da forma de onda, resultando em um novo áudio que preserva o conteúdo semântico original, mas com as características vocais do falante-alvo.

4.2 Base de Dados

O dataset utilizado é o Mozilla Common Voice, um dos maiores e mais diversos corpora de fala disponíveis em domínio público. Ele é composto por milhares de horas de gravações de áudio transcritas em dezenas de idiomas, coletadas e validadas por meio de crowdsourcing. Os participantes gravam sentenças apresentadas pelo sistema e validam as gravações de outros usuários, garantindo qualidade e diversidade de locutores, sotaques, idades e gêneros. As gravações são divididas em conjuntos de treino, validação e teste, evitando que o mesmo locutor apareça em mais de um conjunto. O Common Voice é amplamente utilizado em tarefas de reconhecimento e conversão de fala, por ser aberto, sustentável e continuamente expandido pela comunidade global [1].

Devido à ausência de datasets contendo áudios de pessoas traqueostomizadas que utilizam laringe eletrônica, modificamos o dataset Mozilla Common Voice para simular esse tipo de áudio e criar um dataset sintético. Para reproduzir as características da laringe eletrônica, ajustamos os áudios de forma a manter o F0 constante, eliminando variações na periodicidade. Esse procedimento reflete o comportamento real do dispositivo, que, por ser eletrônico, emite sempre a mesma frequência fundamental. E isso pode ser visto também em [22] que utiliza desse mesmo artificio para criação de um dataset sintético.

Foi utilizado um total de 19.002 áudios em português de falantes nativos, sendo 15.202 destinados ao treinamento, 1.900 à validação e 1.900 ao teste. Para cada áudio, existe também uma versão com F0 constante; entretanto, essa modificação é aplicada de forma online durante o treinamento.

4.3 Experimentos

Os experimentos foram conduzidos em duas linhas principais. Na primeira, utilizou-se a arquitetura base do Seed-VC, conforme descrito na seção anterior. Para a extração do conteúdo semântico foi utilizado o XLSR [2].

Na segunda linha experimental, o componente Timbre Shifter foi substituído por um módulo de modificação do F0, responsável por tornar a frequência fundamental do áudio constante. Essa abordagem foi inspirada em [22], que propõe essa técnica para simular o padrão de fala de usuários de laringe eletrônica, cujo som apresenta F0 invariável devido à natureza do dispositivo.

A frequência fundamental (F0) representa a estrutura periódica do sinal de fala, sendo as variações em sua periodicidade responsáveis pelas características individuais da voz humana. Como o sistema fonador humano é orgânico e sujeito a variações fisiológicas, cada pessoa apresenta flutuações únicas na F0, o que resulta em timbres e entonações distintos. Em contrapartida, dispositivos mecânicos, como a laringe eletrônica, produzem um sinal com periodicidade constante, resultando em uma frequência fixa e, consequentemente, em um som monótono e artificial.

Para simular essa característica em experimentos de reconstrução vocal, os áudios são ajustados para uma frequência fundamental fixa de 80 Hz, de modo a reproduzir o padrão de vibração contínua e uniforme gerado pela laringe eletrônica. Essa abordagem permite aproximar os sinais sintéticos das propriedades acústicas observadas na fala produzida por esse tipo de dispositivo.

Todos os experimentos foram conduzidos com batch size 8 ao longo de 10 épocas de treinamento.

4.4 Métricas de avaliação

A Distância Cepstral de Mel (MCD) é uma métrica objetiva usada para quantificar a similaridade espectral entre a fala gerada e a fala de referência [16]. Ela mede a diferença média nos coeficientes cepstrais na escala de Mel (MFCCs) entre os dois sinais, fornecendo uma indicação da distorção espectral.

O MCD é frequentemente utilizado na avaliação de síntese de fala, cujo objetivo é produzir uma fala sintetizada que soe o mais próxima possível da fala humana natural. Um valor menor de MCD indica uma correspondência mais próxima entre as características espectrais da fala sintetizada e da fala de referência, e, portanto, uma melhor qualidade de síntese [16].

O log spectral distance LSD compara os espectros de magnitude (em dB) dos dois sinais, frame a frame. Ele calcula o erro logarítmico ponto a ponto entre os espectros e a média ao longo do tempo. Quanto menor o LSD, mais próximos estão os espectros e são de melhor qualidade [9].

Peak Signal-to-Noise Ratio (PSNR) é uma métrica amplamente utilizada para medir a qualidade de reconstrução de imagem no

Métricas	(i) Timbre → Timbre	(ii) Timbre → F0	(iii) F0 → Timbre	(iv) F0 → F0
MCD ↓	464.25 ± 9.35	460.21 ± 7.25	447.84 ± 6.19	444.04 ± 8.17
LSD ↓	0.48 ± 0.009	0.48 ± 0.0021	0.50 ± 0.006	0.47 ± 0.0079
PSNR ↑	42.01 ± 0.16	41.96 ± 0.099	42.05 ± 0.096	42.27 ± 0.14
LPIPS ↓	0.00042 ± 0.000052	0.00046 ± 0.000047	0.00067 ± 0.000061	0.00066 ± 0.000072

Table 1: Resultados para cada configuração de treino e teste: (i) Timbre shifter → Timbre shifter, (ii) Timbre shifter → F0 constante, (iii) F0 constante → Timbre shifter, (iv) F0 constante → F0 constante.

nosso caso o mel espectrograma. O PSNR mede a razão entre o valor máximo possível de um sinal e o ruído que afeta a fidelidade da sua representação [11].

Learned Perceptual Image Patch Similarity (LPIPS) mede a distância perceptual entre duas imagens com base em ativações internas de redes neurais profundas (como VGG ou AlexNet) no nosso caso usamos a AlexNet. Ao contrário de métricas tradicionais como PSNR e SSIM, ela não compara pixels diretamente, mas sim representações extraídas por redes treinadas. LPIPS baixo imagens perceptualmente similares, LPIPS alto imagens mais diferentes para o olho humano, e no contexto da conversão de voz ele serve para se a imagem gerada continua no domínio do mel espectrograma [23].

5 RESULTADOS E DISCUSSÃO

Foram realizados dois tipos de experimentos. No primeiro, utilizou-se a estratégia de remover o timbre do áudio, gerando um áudio neutro, a partir do qual se busca reconstruir o áudio original, neste caso, representado pelo mel espectrograma de referência. O segundo tipo de treinamento segue a mesma lógica, porém substitui o módulo responsável pela remoção do timbre (timbre shifter) por uma função que transforma a frequência fundamental (F0) do áudio em uma constante, resultando em um áudio com timbre fixo.

Durante os testes, seguiu-se o mesmo princípio adotado no treinamento. Foram avaliadas quatro combinações: (i) treinamento com timbre shifter e teste com timbre shifter; (ii) treinamento com timbre shifter e teste com F0 constante; (iii) treinamento com F0 constante e teste com timbre shifter; e (iv) treinamento com F0 constante e teste com F0 constante.

Foi utilizada a estratégia de hold-out para cada tipo de experimento, treinando o modelo cinco vezes tanto na configuração com timbre shifter quanto na configuração com F0 constante, mantendo os mesmos parâmetros de treinamento em todas as execuções.

A comparação dos resultados na tabela 1 evidencia que a abordagem baseada na fixação da frequência fundamental (F0 constante) apresenta desempenho superior à técnica com timbre shifter. No cenário F0 → F0, foram obtidos os melhores valores em três das quatro métricas avaliadas: MCD de 444.04, LSD de 0.47 e PSNR de 42.27, superando consistentemente os resultados dos demais cenários. Embora o LPINPS não tenha sido o menor, seu valor (0.00066) permaneceu próximo aos melhores obtidos, indicando que não houve perda significativa nesse aspecto. Esses resultados sugerem que a fixação do F0, ao eliminar variações de periodicidade, contribui para uma reconstrução de voz mais fiel e próxima do sinal original.

Outro ponto relevante diz respeito aos resultados apresentados na métrica LPIPS. Não foi observada uma diferença significativa entre os diferentes testes realizados. Isso ocorre porque a métrica utiliza convoluções para avaliar a semelhança entre imagens, capturando informações mais gerais da estrutura visual. Como todas as imagens comparadas pertencem ao mesmo domínio (espectrogramas), o LPIPS tende a apresentar pouca variação.

5.1 Discussão Ética

O avanço das técnicas de reconstrução e conversão de voz, embora promissor para a reabilitação da fala em pessoas laringectomizadas, levanta questões éticas relevantes, sobretudo no que diz respeito ao uso indevido dessas tecnologias. A capacidade de reproduzir a voz de um indivíduo com alto grau de fidelidade pode ser explorada de forma maliciosa, resultando em casos de falsificação de identidade, manipulação de áudios e desinformação. Tais riscos tornam essencial a implementação de estratégias de mitigação que garantam o uso responsável e seguro dessas ferramentas.

Entre as principais medidas de mitigação, destaca-se a necessidade de diretrizes éticas claras para o desenvolvimento e compartilhamento de modelos de reconstrução de voz. Isso inclui o uso de datasets anonimizados, com consentimento explícito dos participantes e restrições quanto à finalidade de uso. Além disso, é fundamental promover a transparência dos modelos por meio de documentação detalhada e auditorias independentes, assegurando que os sistemas sejam utilizados exclusivamente em contextos clínicos e de pesquisa.

Outra estratégia envolve o desenvolvimento de mecanismos de watermarking ou assinaturas digitais embutidas nos áudios sintetizados, permitindo a rastreabilidade e a identificação de conteúdos gerados artificialmente. Por fim, recomenda-se criar comitês interdisciplinares envolvendo profissionais de tecnologia, ética, medicina e direito, capazes de avaliar continuamente os impactos sociais e morais decorrentes do uso dessas tecnologias.

Dessa forma, ao mesmo tempo em que se busca o avanço científico e a melhoria na qualidade de vida de pessoas com perda de voz, é imprescindível garantir que tais inovações sejam conduzidas dentro de um marco ético sólido, que assegure a integridade, a privacidade e a dignidade dos indivíduos beneficiados.

6 CONCLUSÃO

Constata-se que ainda há uma escassez de pesquisas voltadas ao tratamento do problema enfrentado por pessoas que perderam a laringe. Durante esta investigação, não foi identificado nenhum dataset contendo amostras de indivíduos que utilizam laringe eletrônica, o que motivou o uso de bases sintéticas, como demonstrado em [18], abordagem também adotada no dataset [1].

Outro aspecto relevante foi a comparação entre dois tipos de treinamento: um empregando a arquitetura original proposta por

[18], com timbre shifter, e outro substituindo esse módulo por uma função que fixa a frequência fundamental (F0) do áudio. Essa modificação resultou em melhorias nas métricas, com ganhos de até 20 pontos e redução no desvio padrão, especialmente na métrica MCD, indicando maior estabilidade e qualidade na reconstrução vocal.

Como trabalhos futuros, sugere-se construir uma base de dados real composta por gravações de pessoas submetidas à laringectomia e que utilizem a laringe eletrônica como meio de comunicação. Além disso, propõe-se a construção colaborativa de um dataset em parceria com hospitais e universidades, fomentando a criação de uma base pública e diversificada que possa apoiar novas pesquisas na área. Também sugere-se analisar comparativamente diferentes modelos de conversão de voz (Voice Conversion), a fim de avaliar o desempenho dessas abordagens tanto em datasets sintéticos quanto em dataset real.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA).

REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670 [cs.CL] https://arxiv.org/abs/1912.06670

[2] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. arXiv:2111.09296 [cs.CL] https://arxiv.org/abs/2111.09296

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL] https://arxiv.org/abs/2006.11477

[4] Anders R. Bargum, Stefania Serafin, and Cumhur Erkut. 2023. Reimagining Speech: A Scoping Review of Deep Learning-Powered Voice Conversion. arXiv:2311.08104 [cs.SD] https://arxiv.org/abs/2311.08104

[5] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Marieum Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, Paavo Alku, and Mohammad Hassan Vali. 2022. *Introduction to Speech Processing* (2 ed.). https://doi.org/10.5281/zenodo.6821775

[6] Hao Chen, Xiaoqi Cao, Xiyan Zhang, Zhenyu Wang, Bingjing Qiu, and Kehong Zheng. 2023. Automatic segmentation framework of X-Ray tomography data for multi-phase rock using Swin Transformer approach. *Scientific Data* 10 (11 2023). https://doi.org/10.1038/s41597-023-02734-7

[7] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (Oct. 2022), 1505–1518. https://doi.org/10.1109/jstsp.2022.3188113

[8] Mauricio de Cunto. 2023. A VOZ HUMANA E AS SUAS CARACTERÍSTICAS -UM RESUMO ESCLARECIMENTOS SOBRE A METODOLOGIA DO EXAME DE IDENTIFICAÇÃO FORENSE DE VOZ/FALA PROF. ENGº MAURÍCIO R. DE CUNTO (OUTUBRO DE 2023 -VERSÃO 4). (10 2023). https://doi.org/10.13140/RG.2.2.10780.26246

[9] Per Enqvist and Johan Karlsson. 2008. Minimal Itakura-Saito distance and covariance interpolation. In *2008 47th IEEE Conference on Decision and Control*. 137–142. https://doi.org/10.1109/CDC.2008.4739312

[10] M.W Gardner and S.R Dorling. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* 32, 14 (1998), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

[11] Alain Horé and Djemel Ziou. 2010. Image Quality Metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*. 2366–2369. https://doi.org/10.1109/ICPR.2010.579

[12] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. arXiv:1703.06868 [cs.CV] https://arxiv.org/abs/1703.06868

[13] In-Sun Hwang, Sang-Hoon Lee, and Seong-Whan Lee. 2022. StyleVC: Non-Parallel Voice Conversion with Adversarial Style Generalization. In *2022 26th International Conference on Pattern Recognition (ICPR)*. 23–30. https://doi.org/10.1109/ICPR56361.2022.9956613

[14] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. CREPE: A Convolutional Representation for Pitch Estimation. arXiv:1802.06182 [eess.AS] https://arxiv.org/abs/1802.06182

[15] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv:2010.05646 [cs.SD] https://arxiv.org/abs/2010.05646

[16] R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Vol. 1. 125–128 vol.1. https://doi.org/10.1109/PACRIM.1993.407206

[17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747 [cs.LG] https://arxiv.org/abs/2210.02747

[18] Songting Liu. 2024. Zero-shot Voice Conversion with Diffusion Transformers. arXiv:2411.09943 [cs.SD] https://arxiv.org/abs/2411.09943

[19] NAZARIO, LUIZA CASCAES MAGAJEWSKI, FLÁVIO RICARDO LIBERALI PIZZOL, NATALIA DAL SALOTI, MATHEUS HENRIQUE DA SILVA MEDEIROS, and LEONARDO KFOURI. 2022. Tendência temporal da utilização da traqueostomia em pacientes hospitalizados pelo Sistema Único de Saúde no Brasil no período de 2011 a 2020. *Revista do Colégio Brasileiro de Cirurgiões* (2022).

[20] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2017. FiLM: Visual Reasoning with a General Conditioning Layer. arXiv:1709.07871 [cs.CV] https://arxiv.org/abs/1709.07871

[21] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. arXiv:2303.00332 [cs.SD] https://arxiv.org/abs/2303.00332

[22] Yaogen Yang, Haozhe Zhang, Zexin Cai, Yao Shi, Ming Li, Dong Zhang, Xiaojun Ding, Jianhua Deng, and Jie Wang. 2023. Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion. *Biomedical Signal Processing and Control* 80 (2023), 104279. https://doi.org/10.1016/j.bspc.2022.104279

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924 [cs.CV] https://arxiv.org/abs/1801.03924