Análise da Melhoria do Serviço de Voz sobre IP Através da Clonagem de Voz Redução de Dados em Sistemas VoIP Utilizando Clonagem de Voz

OLIVEIRA, G. S. * MIRANDA, I. D. S. †

Submitted: 30 jul.2022. **Approved**: 10 out.2022.

Revista Eletrônica de Ciências Exatas e Tecnológicas
ABR-2023, 4ª Edição, Volume 1.

RESUMO

Sistemas de voz sobre IP (VoIP - voice over internet protocol) possibilitam que a voz trafegue como pacotes de dados em redes de internet, agregando vantagens em relação ao sistema de telefonia convencional, como baixo custo e flexibilidade. No entanto, problemas como a instabilidade ou baixa velocidade da internet ainda são frequentes em algumas regiões, causando cortes e interrupções nas chamadas de voz. Este artigo apresenta uma técnica para redução da taxa de dados transmitidos em sistemas VoIP através da conversão de trechos da fala para texto antes da transmissão e da aplicação de clonagem de voz, no receptor, para reconstituir os trechos enviados como texto. Aplicamos o sistema proposto a 3 áudios de referência e o resultado foi avaliado por 9 indivíduos. A compreensão foi acima de 75% em 88,9% das avaliações realizadas, considerando uma redução de aproximadamente 50% dos dados transmitidos. Isso demonstra o potencial desta tecnologia, apesar de haver diversas possibilidades para aperfeiçoamento.

Palavras-chave: Aprendizagem de máquina. Compressão de Dados. Reconhecimento de Fala. Síntese de Voz.

ABSTRACT

Voice over internet protocol (VoIP) systems allow voice to travel as data packets on internet networks, adding advantages over the conventional telephone systems, such as low cost and flexibility. However, problems such as instability or low internet speed are still frequent in some regions, causing cuts and interruptions in voice calls. This article presents a technique for reducing the rate of data transmitted in VoIP systems by converting speech to text before transmission and applying voice cloning, at the receiver, to reconstitute the excerpts sent as text. We applied the proposed system to 3 reference audios and the result was evaluated by 9 individuals. Comprehension was above 75% in 88.9% of the assessments carried out, considering a reduction of approximately 50% in the transmitted data. This demonstrates the potential of this technology, although there is room for improvement.

Keywords: Machine Learning. Data compression. Speech Recognition. Voice Synthesis.

Sumário

INTRODUÇÃO

Os sistemas VoIP (*Voice over Internet Protocol*) são responsáveis por possibilitar a conversação por meio da internet, onde a voz trafega em pacotes de dados de redes IP via internet, sendo uma alternativa ao sistema de telefonia convencional. Outras vantagens desse serviço correspondem ao seu baixo custo, flexibilidade e facilidade de encontrar e fazer ligações com outras pessoas (SILVA, 2017).

O período de pandemia do Coronavírus (COVID-19) no Brasil e no mundo evidenciou a necessidade do uso das plataformas de comunicação digital como um recurso para o distanciamento social e para realização de trabalho de modo remoto. Por conta disso, ampliou-se

*Gabriel Sampaio de Oliveira. Graduando em Bacharelado em Ciências Exatas e Tecnológicas pela Universidade Federal do Recôncavo da Bahia (UFRB), Brasil. sogabris@aluno.ufrb.edu.br

†Igor Dantas dos Santos Miranda. Possui graduação e mestrado em engenharia elétrica pela Universidade Federal da Bahia (UFBA), onde também se graduou doutor em engenharia industrial. Desde 2013, é professor na área de sistemas elétricos e computacionais da UFRB. igordantas@ufrb.edu.br

ISSN: 2763-8855
RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

mundialmente a utilização dos sistemas VoIP e seus serviços.

Neste contexto, um problema frequente é a instabilidade ou baixa velocidade da internet, o que pode causar cortes e interrupções nas chamadas de voz e conferências. Para chamadas VoIP, é necessário pelo menos 100 kbps de banda larga por linha, exigindo pelo menos 1 Mbps de banda dedicada disponível para poder lidar com dez chamadas VoIP de alta qualidade ao mesmo tempo (BAI, 2020). Isso indica que esse problema atinge tanto usuários individuais como empresas. A redução da taxa de dados transmitidos durante as chamadas VoIP é, portanto, uma estratégia imprescindível para melhoria do serviço.

Diversos autores estudaram estratégias para solucionar os problemas relacionados aos serviços VoIP. A mais difundida é o uso de algoritmos de compactação e descompactação de dados digitais, conhecidos como *codecs*. Estes sistemas podem ser aplicados a arquivos de áudio, tornando-os menores sem perder a qualidade da voz (BAI, 2020). Dessa forma, é possível reduzir a quantidade de banda de internet necessária para transmissão.

Atualmente existem diversos algoritmos de *codecs* de voz disponíveis, onde os mais comumente utilizados para VoIP foram padronizados pelo Setor de Padronização de Telecomunicações da União Internacional de Telecomunicações (ITU-T) e pela Força-Tarefa de Engenharia da Internet (IETF) (NOWORATZKY, 2019). O ITU-T desenvolveu a série G de *codecs*, introduzidos no início dos anos 70 quando os sinais de voz digitalizados começaram a ser implementados na telefonia. Os *codecs* da série G são os mais utilizados em dispositivos VoIP.

Ainda que grandes avanços tenham ocorrido na área, verificamos a necessidade e possibilidade de melhoria para o serviço. Este trabalho propõe tratar do problema com uma abordagem nova, utilizando reconhecimento de fala e clonagem de voz. No método proposto, trechos da fala são transformados para texto antes da transmissão e reconstruídos na recepção, utilizando a técnica de clonagem com a própria voz do falante. Esperamos com isso reduzir em aproximadamente 50% a taxa de dados necessária para transmissão VoIP, com perdas toleráveis na qualidade do serviço.

Este artigo está organizado conforme segue. Após essa introdução, na Sessão 2 realizou-se a discussão teórica, detalhando as etapas relevantes para os métodos de reconhecimento de fala e clonagem de voz. Em seguida, apresentamos na Sessão 3 os materiais e métodos que utilizamos em nossa aplicação. Por fim, são apresentados, nas Sessões 4 e 5, os resultados e considerações finais, incluindo sugestões de trabalhos futuros.

FUNDAMENTAÇÃO TEÓRICA

Reconhecimento de Fala

O reconhecimento de fala é um algoritmo baseado na aprendizagem de máquina capaz de traduzir a fala em texto. Uma tecnologia de diversas aplicações, possibilitando funcionalidades como digitação por voz, geração automática de legendas e até a comunicação através da fala com dispositivos eletrônicos. Com esse recurso, também é possível proporcionar inclusão digital para pessoas com deficiência (PCD).

De modo geral, a inteligência artificial do reconhecimento de fala é treinada para receber como entrada a fala humana e, com base nos parâmetros e padrões de fala, encontrar as palavras mais prováveis naquele idioma para transcrever o som em texto.

A Figura 1 exibe o processo realizado para o reconhecimento de fala. Inicialmente o sinal analógico é convertido para digital, em seguida uma unidade de pré-processamento melhora o sinal de fala reduzindo ruídos. Então, o algoritmo de reconhecimento usa modelagem acústica para registrar fonemas, unidades distintas do som da fala que representam e distinguem uma palavra da outra. Na decodificação os fonemas são construídos em palavras e frases compreensíveis usando modelagem de linguagem (TATE, 2021). Simplificando, o reconhecimento de fala combina o sinal da fala de um usuário com outros padrões de voz.

Clonagem de Voz

Um sistema de clonagem de voz recebe como entrada um trecho de voz e um texto. Como saída, ele gera um áudio com uma voz similar à do falante dizendo as palavras contidas no texto fornecido. Com essa técnica, a voz de uma pessoa pode ser imitada com grande semelhança.

O fluxo de processamento da clonagem de voz utilizado em nossa pesquisa se dá como mostrado na Figura 2. Onde o codificador de fala é o responsável por extrair e codificar o timbre da voz do áudio que foi recebido, gerando uma incorporação para ser condicionado ao sintetizador. O sintetizador também recebe uma frase em formato de texto e transforma em um espectrograma gerado pela concatenação da incorporação do áudio de referência e a sequência de fonemas. Finaliza-se com o vocoder que recebe o espectrograma gerado pelo sintetizador e o codifica para gerar o áudio final. Para uma

ISSN: 2763-8855
RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

Figura 1 – Etapas do processo de reconhecimento da fala (MIRANDA; CHITALE; COHEN, 2019, modificado).

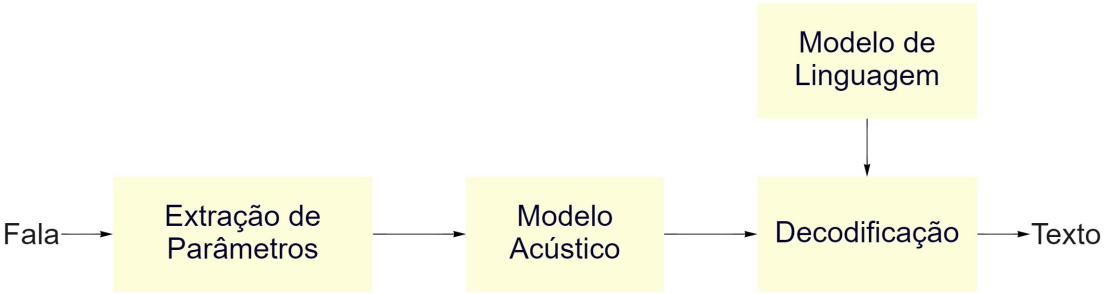
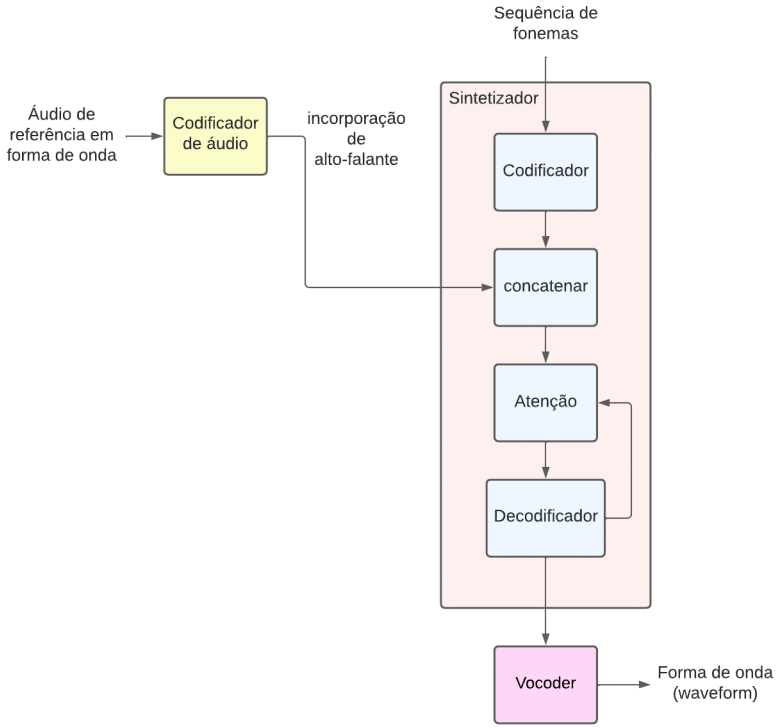


Figura 2 – Diagrama de etapas de execução da clonagem de voz (JEMINE, 2019, modificado).



melhor compreensão sobre a sequência de passos a serem realizados na execução da clonagem de voz aqui utilizada, recomendamos o estudo de (JEMINE, 2019).

MATERIAIS E MÉTODOS

Sistema Proposto

O sistema proposto consiste em codificar parte da informação transmitida pela fala através do formato de texto, antes da transmissão. Na recepção, o texto transmitido é convertido em áudio, utilizando as características de voz do falante, e o trecho da fala pode então ser reconstituído.

Para realização dessa tarefa, são realizadas as etapas de segmentação, reconhecimento de fala, reconstrução da voz e concatenação. A Figura 3 exibe o fluxo de execução das etapas do sistema proposto, onde, a segmentação do áudio é realizada escolhendo como pontos de divisão os momentos de silêncio. Os segmentos são separados, intercaladamente, em dois grupos: o grupo dos segmentos que serão enviados como voz e o grupo dos segmentos que serão enviados como texto.

Os segmentos do primeiro grupo devem ser enviados como voz diretamente para o receptor. Em um sistema real, esses segmentos podem ser comprimidos, utilizando um algoritmo de CODEC de áudio. Não realizamos essa etapa no ambiente simulado por conveniência. No receptor, esses segmentos são utilizados para treinar a IA responsável pela clonagem da voz.

Para o segundo grupo de segmentos, um sistema de reconhecimento de fala é utilizado para transformar os trechos de áudio para texto. No receptor, esses segmentos são convertidos para áudio novamente através do processo de clonagem de voz, criando um áudio

ISSN: 2763-8855
RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

sintetizado a partir do que foi dito pelo falante.

Por fim, os segmentos transmitidos como áudio originais e os segmentos sintetizados são concatenados para reconstruir a fala original. A mesma sequência dos segmentos deve ser utilizada para manter a consistência da mensagem. Como a clonagem de voz não produz resultados perfeitos, uma redução na qualidade é esperada. Porém, com a intercalação entre trechos clonados e trechos originais, espera-se que a redução na qualidade do áudio não seja significativa.

Para o reconhecimento de fala, utilizamos a biblioteca SpeechRecognition (ZHANG, 2017), que recebe trechos de fala e os envia para o serviço de reconhecimento de fala da empresa Google e devolve o texto detectado (ROCKIKZ, 2022). Este sistema possui uma acurácia de 84% (VAILSHERY, 2022).

A clonagem de voz foi implementada utilizando o framework SV2TTS (Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis) (JEMINE, 2019). Esse sistema possui uma estrutura de aprendizagem profunda de três estágios que permite criar uma representação numérica de uma voz a partir de alguns segundos de áudio. Essa representação é usada para condicionar um modelo de texto para fala e é treinado para generalizar novas vozes. Como o sistema de clonagem foi pré-treinado para falantes da língua inglesa, restringimos nossos experimentos a esse idioma.

Avaliação de Desempenho

Para avaliarmos o desempenho do sistema proposto, optamos por uma avaliação participante, contando com a análise de pessoas. Isso porque existem aspectos subjetivos na percepção da voz que não seriam possíveis de analisar de forma objetiva. Desta forma, convidamos 9 pessoas de ambos os gêneros, todas falantes de inglês, sendo 2 nativos e 1 linguista. Eles avaliaram nossa pesquisa por meio de um formulário com questões objetivas que foram orientadas por um padrão internacional na avaliação de áudio descrito em ITU-R BS.1116-1 (ITU, 1994).

Utilizamos três áudios de fala com aproximadamente 30 segundos cada, composto por uma mulher adulta (áudio 1), um homem adulto (áudio 2) e um homem idoso (áudio 3). Os arquivos foram convertidos para o formato WAV, com taxa de amostragem de 16kHz. Cada arquivo foi manualmente segmentado utilizando o software Audacity e produzindo uma média de 10 segmentos de fala com duração de 2 a 4 segundos.

RESULTADOS

Com base nos resultados alcançados pela pesquisa, podemos perceber segundo a tabela 1 a avaliação da distorção dos áudios modificados, ou seja, dos áudios finais gerados pelo sistema proposto. Ao analisarmos os dados da tabela, podemos observar que nenhum dos áudios foram avaliados como "Irritante" ou "Muito Irritante". Também podemos perceber que, em alguns dos áudios, a distorção chega a ser imperceptível. Isso nos leva a concluir que esta diferenciação deve-se mais à percepção auditiva das pessoas que participaram dos testes do que a tecnologia em si.

Analisamos também, conforme demonstra a tabela 2, o nível de compreensão dos participantes referente aos áudios modificados. Podemos observar que em apenas um dos áudios houve uma compreensão abaixo de 50% por parte de 11,1% dos participantes. Este resultado demonstra que o sistema proposto atingiu o objetivo, mesmo com uma perda parcial da compreensão, algo esperado devido ao processo de sintetização da voz.

Dentre os áudios analisados, observamos que o áudio 3, cujo autor foi um idoso, gerou mais incômodo visto a avaliação entre os participantes, com uma taxa de 55,6%. Acreditamos que o fato do sistema utilizado funcionar melhor para pessoas mais jovens significa talvez que, como o sistema é pré-treinado e não temos conhecimento dos dados utilizados para este treinamento, o mesmo tenha tido um foco maior no treinamento com pessoas mais jovens. Vale ressaltar também que ruídos externos, assim como a dicção do falante também podem afetar o sistema de reconhecimento de fala.

Como a quantidade de dados nos trechos enviados como texto é relativamente pequena em relação aos trechos equivalentes de áudio, é possível reduzir significativamente a taxa de dados transmitidos.

Utilizando o áudio 3 do nosso experimento como referência, avaliamos a redução de dados com o sistema proposto. Para o experimento descrito na seção dos Materiais e Métodos, o áudio selecionado foi segmentado e 50% dos segmentos foram transformados em texto. Considerando o *codec* VoIP Série G, que possui taxa de transmissão de 100 kbps, e a taxa média de transmissão da parte textual observada para esse áudio, de 121 bps, a taxa de transmissão seria reduzida para 49,94%. A Figura 4 exibe estimativas de redução da taxa de transmissão para outros percentuais de trechos convertidos para texto, considerando as mesmas taxas de transmissão média descritas acima.

ISSN: 2763-8855
RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

Figura 3 – Diagrama de fluxo das etapas de processamento do sistema proposto, desde a voz original até a voz reconstituída no receptor. A etapa de Transmissão emula o canal de comunicação que consideramos ser ideal.

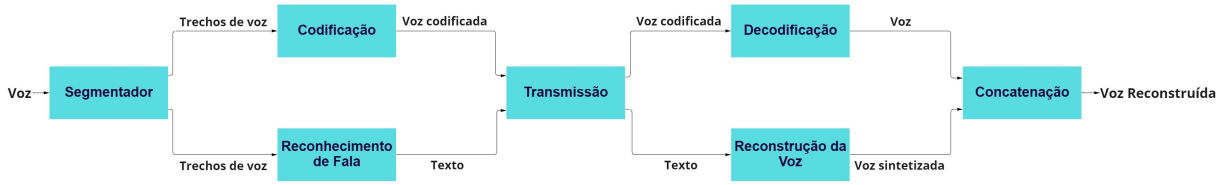
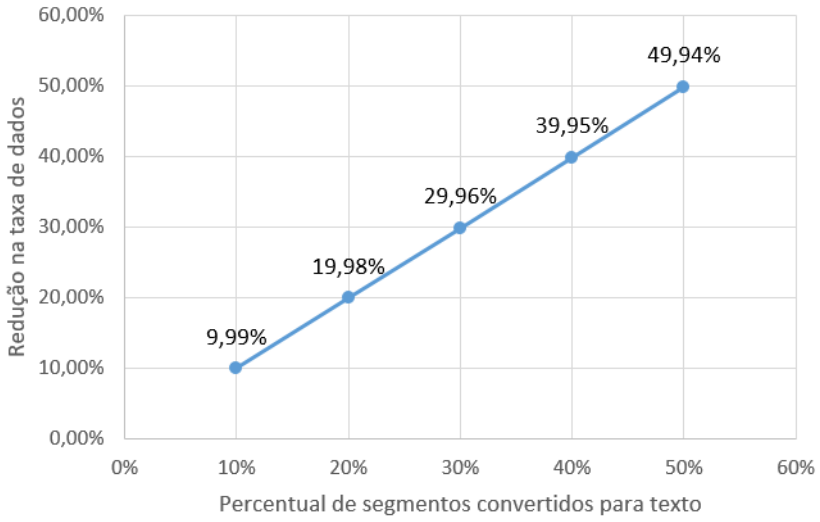


Figura 4 – Estimativa da redução na taxa de dados transmitidos em função do percentual de segmentos convertidos para texto.



Considerações finais

Este artigo teve como foco de estudo a possibilidade de redução de dados transmitidos em sistemas de Voz sobre IP com a utilização da transformação da fala para texto e a clonagem de voz.

Através das avaliações dos participantes da pesquisa sobre os áudios gerados pelo sistema proposto, concluímos que esta tecnologia poderá ser viável. Para isso, a capacidade de processamento dos celulares precisará aumentar ou os algoritmos devem ser simplificados. Também constatamos que, ainda que com leves incômodos para a audição em partes dos trechos modificados, o entendimento e a compreensão do texto falado em áudio não foi comprometida.

Os resultados obtidos nesse estudo são encorajadores, abrindo caminhos para utilização eficiente de VoIP em áreas remotas como zonas rurais, navios, aviões e ônibus intermunicipais. Como possibilidades de trabalhos futuros, sugerimos o teste do sistema proposto com um número maior de áudios e participantes na avaliação, experimentos com transmissão real em uma rede VoIP e pré-treinamento dos modelos para português com um amplo espectro de vozes.

Referências

BAI, J. *How much data does VoIP use? Tips to forecast your VoIP bandwidth*. [S.l.], 2020. Disponível em: <https://www.nextiva.com/blog/voip-data-usage.html>. Citado na página 2.

ITU, R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. *ITU-R Recommendation BS*, v. 1116, 1994. Citado na página 4.

JEMINE, C. Real-time-voice-cloning. *University of Liége, Liége, Belgium*, 2019. Citado 2 vezes nas páginas 3 e 4.

MIRANDA, A.; CHITALE, P.; COHEN, J. *How to Build Domain Specific Automatic Speech Recognition Models on GPUs*. [S.l.], 2019. Disponível em: <https://developer.nvidia.com/blog/how-to-build-domain-specific-automatic-speech-recognition-models-on-gpus/>. Citado na página 3.

NOWORATZKY, D. *SpeechRecognition 3.8.1*. [S.l.], 2019. Disponível em: <https://info.teledynamics.com/blog/the-wonderful-world-of-voice-codecs>. Citado na página 2.

ROCKIKZ, A. *How to Convert Speech to Text in Python*. [S.l.], 2022. Disponível em: <https://www.thepythoncode.com/article/using-speech-recognition-to-convert-speech-to-text-python>. Citado na página 4.

SILVA, J. P. Voip (voz sobre ip): vantagens em relação a telefonia convencional no brasil. Universidade Tecnológica Federal do Paraná, 2017. Citado na página 1.

ISSN: 2763-8855
RECET Revista Eletrônica de Ciências Exatas e Tecnológicas

TATE, L. *The Difference Between Speech And Voice Recognition*. [S.l.], 2021. Disponível em: <https://www.kardome.com/blog-posts/difference-speech-and-voice-recognition>. Citado na página 2.

VAILSHERY, L. *Speech-to-Text transcript accuracy rate among leading companies 2021*. [S.l.], 2022. Disponível em: <https://www.statista.com/statistics/1133833/speech-to-text-transcript-accuracy-rate-among-leading-companies/>. Citado na página 4.

ZHANG, A. *SpeechRecognition 3.8.1*. [S.l.], 2017. Disponível em: <https://pypi.org/project/SpeechRecognition>. Citado na página 4.

ISSN: 2763-8855

RECET Revista Eletrônica de Ciências Exatas e Tecnológicas