Avaliação de arquiteturas de síntese de fala generativa com abordagens de espectrograma e fim-a-fim em cenários *low-resource* para clonagem de voz

Bruno C. dos S. Ribeiro[1], Gustavo H. dos S. Figueiredo[1], Leonardo H. da S. Correia[1], Mário Uliani Neto[1], Fernando O. Runstein[1], Ricardo P. V. Violato[1], Marcus Lima[2],

[1]CPQD - Centro de Pesquisa e Desenvolvimento, Campinas, SP, Brasil

[2]Pontifícia Universidade Católica de Campinas, SP, Brasil

Resumo. *O artigo compara modelos de síntese de fala com arquiteturas baseadas em espectrograma e fim-a-fim, com o objetivo de determinar a capacidade de clonagem de voz em cenário low-resource. Foram avaliados conjuntos de treinamento de adaptação com diferentes quantidades de fala para clonagem de uma voz alvo, e o tempo necessário para realizar o treinamento. O modelo VITS mostrou-se mais eficiente, alcançando os melhores resultados no teste de qualidade perceptual no cenário low-resource com dados no idioma português, e completou o treinamento em menos tempo, quando comparado com o Tacotron2.*

1. Introdução

A síntese de fala tem sido um campo de intenso estudo e inovação ao longo dos últimos anos, com avanços significativos impulsionados pelos rápidos progressos na área de inteligência artificial generativa. Dentro deste contexto, diversas abordagens têm sido exploradas, incluindo as arquiteturas baseadas em espectrogramas e as abordagens fim-a-fim.

As arquiteturas Tacotron [Wang et al. 2017] e Tacotron2 [Shen et al. 2018] têm sido amplamente estudadas e aplicadas, demonstrando a capacidade de converter texto em fala natural por meio da geração de espectrogramas intermediários, que são posteriormente transformados em sinais de fala através de vocoders, como as arquiteturas WaveNet [van den Oord et al. 2016] e HiFi-GAN [Kong et al. 2020]. Apesar dos resultados promissores, esses modelos frequentemente requerem grandes quantidades de dados e longos períodos de treinamento para atingir um nível satisfatório de qualidade e naturalidade na fala.

As abordagens mais recentes de síntese de fala, como o modelo VITS (do inglês, *Variational Inference Text-to-Speech*) [Kim et al. 2021], propõem uma estratégia fim-a-fim que elimina a necessidade de um estágio intermediário explícito de geração do espectrograma, combinando de forma eficaz a geração e a codificação do sinal de fala em um único fluxo de trabalho. Este método tem mostrado potencial em reduzir significativamente a quantidade de dados necessários para o treinamento, bem como o tempo total para alcançar resultados de alta qualidade.

A eficiência da síntese de fala em cenários com recursos limitados (*low-resource*) é uma área de interesse crescente, especialmente para idiomas com menor disponibilidade de dados anotados, como o caso do português. Trabalhos recentes têm investigado a

eficácia de diferentes modelos em condições *low-resource*, abordando desafios específicos como a qualidade da fala sintetizada, a adaptabilidade de modelos pré-treinados para novos falantes e a eficiência computacional do processo de treinamento [Lux et al. 2022].

O objetivo deste artigo é comparar as arquiteturas baseadas em espectrogramas e fim-a-fim no contexto de clonagem de voz em português, com ênfase no desempenho do VITS versus o Tacotron2. O objetivo é comparar os modelos em cenários *low-resource* e quantificar o número mínimo de dados e tempo de treinamento necessários para atingir resultados de alta qualidade. Os resultados baseiam-se em métricas de qualidade objetiva e subjetiva, e na análise do tempo de treinamento. Esperamos fornecer *insights* práticos para a escolha e implementação de modelos de síntese de fala com voz personalizada em condições de dados restritos, contribuindo para a eficiência e a acessibilidade da tecnologia de síntese de fala em uma ampla gama de aplicações para o idioma português do Brasil.

2. Metodologia

O treinamento dos modelos foi realizado utilizando duas bases de fala no idioma Português Brasileiro: (i) o *TTS-Portuguese Corpus* [Casanova et al. 2022], composto por textos de domínio público provenientes tanto da Wikipédia quanto do Chatterbot-corpus (um corpus criado originalmente para a construção de *chatbots*), contendo aproximadamente 10 horas e 28 minutos de fala de um único locutor masculino, gravada com taxa de amostragem de 48 kHz e 16 bits, tendo 3.632 áudios no formato WAV linear, com um range de duração de 0,67 a 50,08 segundos (todos os clipes de áudio com duração superior a 20 segundos foram removidos do treinamento); (ii) uma base de fala proprietária do CPQD composta por um locutor masculino contendo 20 minutos de fala, gravada com taxa de amostragem de 48kHz, 16 bits e formato PCM linear, contendo os arquivos de áudio e as transcrições ortográficas correspondentes.

O treinamento foi realizado a partir do repositório do VITS[1], que foi adaptado para a inclusão de fonemas do idioma português do Brasil, realizado através do uso do módulo *Phonemizer*[2] em conjunto com a *pipeline* de preparação de dados.

O treinamento dos modelos base ocorreram ao longo de 80 horas e 2.000 épocas no dataset *TTS-Portuguese Corpus*. A partir do último *checkpoint* gerado pelo modelo base, foram realizados *fine-tunings* trocando os dados de treinamento pela base proprietária com a voz do locutor masculino, usando conjuntos de treinamento com 20, 15, 10 e 5 minutos de fala visando avaliar a quantidade mínima de dados necessários para obter síntese de boa qualidade. O objetivo do *fine-tuning* é adaptar o modelo base para as características da voz alvo, ou seja, realizar a clonagem de voz. Após apenas 1 hora de treinamento de *fine-tuning* usando 20 minutos de fala, foram observados resultados de alta qualidade tanto no VITS como no Tacotron2. A qualidade melhorou ainda mais após 20 horas de treinamento. Ambos utilizaram o vocoder HiFi-GAN, sendo que no caso do Tacotron2 o vocoder foi treinado de forma independente. Para os conjuntos de treinamento menores, a seção 3 apresenta os resultados obtidos.

[1]https://github.com/jaywalnut310/vits/

[2]https://pypi.org/project/phonemizer/3.0.1/

3. Resultados

Para avaliar a qualidade da fala sintetizada resultante foram utilizadas medidas objetivas e subjetivas. As métricas objetivas foram o MCD (do inglês, *Mel-Cepstral Distortion*) e o F0 RMSE (do inglês, *Log-F0 Root Mean Square Error*) [Hayashi et al. 2021]. Para a avaliação subjetiva foi utilizada a métrica MOS (*Mean Opinion Score*), em um experimento que contou com 15 avaliadores não especialistas.

A métrica MCD, calculada por meio do repositório *TTS Objective Metrics*[3], quantifica a distância entre dois sinais de fala. Quanto menor o valor MCD, mais semelhantes são as vozes. A qualidade da voz sintetizada foi avaliada com base no conjunto de teste, com frases separadas para validação. Ao comparar a voz sintetizada resultante do modelo de *fine-tuning* obtido com 20 minutos, com a voz original gravada, obteve-se valores de MCD entre 1.6 e 1.78. A métrica MCD mostra valores próximos de 0, indicando que o modelo é capaz de gerar fala sintetizada próxima da fala gravada. Para o F0 RMSE, aplicada nas mesmas sentenças, foram obtidos valores entre 0.18 e 0.34. Os resultados reforçam a alta qualidade da fala sintetizada.

3.1. Avaliação Subjetiva

Para a avaliação subjetiva foi utilizado o servidor webMUSHRA[4]. Um grupo de 15 avaliadores não especialistas ouviram um conjunto de amostras e atribuiram notas de 0 a 100 com base na naturalidade da voz, sendo 0 nada natural e 100 muito natural. Esse processo permitiu realizar uma análise subjetiva da qualidade do áudio sintetizado, proporcionando uma análise mais fidedigna da percepção humana em relação ao desempenho dos modelos. As avaliações mostram uma melhor qualidade do VITS em relação ao Tacotron2. A Figura 1 mostra o *boxplot* com os dados do teste subjetivo, utilizando áudios sintetizados por modelos obtidos através do *fine-tuning* com diferentes conjuntos de treinamento da voz alvo. Na legenda, 400 representa o conjunto com 20 minutos de fala, 300 indica 15 minutos, 200 indica 10 minutos e 100 indica o conjunto com 5 minutos de fala.

Os resultados indicam que o VITS (C1) consistentemente recebeu avaliações mais altas em comparação ao Tacotron2 (C2). O desvio padrão menor do VITS em comparação ao Tacotron2 em todos os conjuntos de treinamento indica que as opiniões dos usuários sobre a qualidade do áudio gerado pelo VITS são mais consistentes e robustas.

No teste realizado com o conjunto contendo 5 minutos de fala de treinamento, o VITS teve uma média de 71,49 enquanto o Tacotron2 teve 67,49. Essa diferença foi consistente em todos os conjuntos de treinamento (20, 15, 10 e 5 minutos). No entanto, a diferença aumenta com um volume maior de dados, sugerindo que o VITS não apenas produz áudio de melhor qualidade, mas também que melhora mais conforme a quantidade de dados de treinamento aumenta.

4. Conclusão

O objetivo principal deste trabalho foi comparar as arquiteturas de síntese de fala generativa com abordagens de espectrograma (Tacatron2) e fim-a-fim (VITS) em cenários

[3]https://github.com/AI-Unicamp/TTS-Objective-Metrics
[4]https://github.com/audiolabs/webMUSHRA/

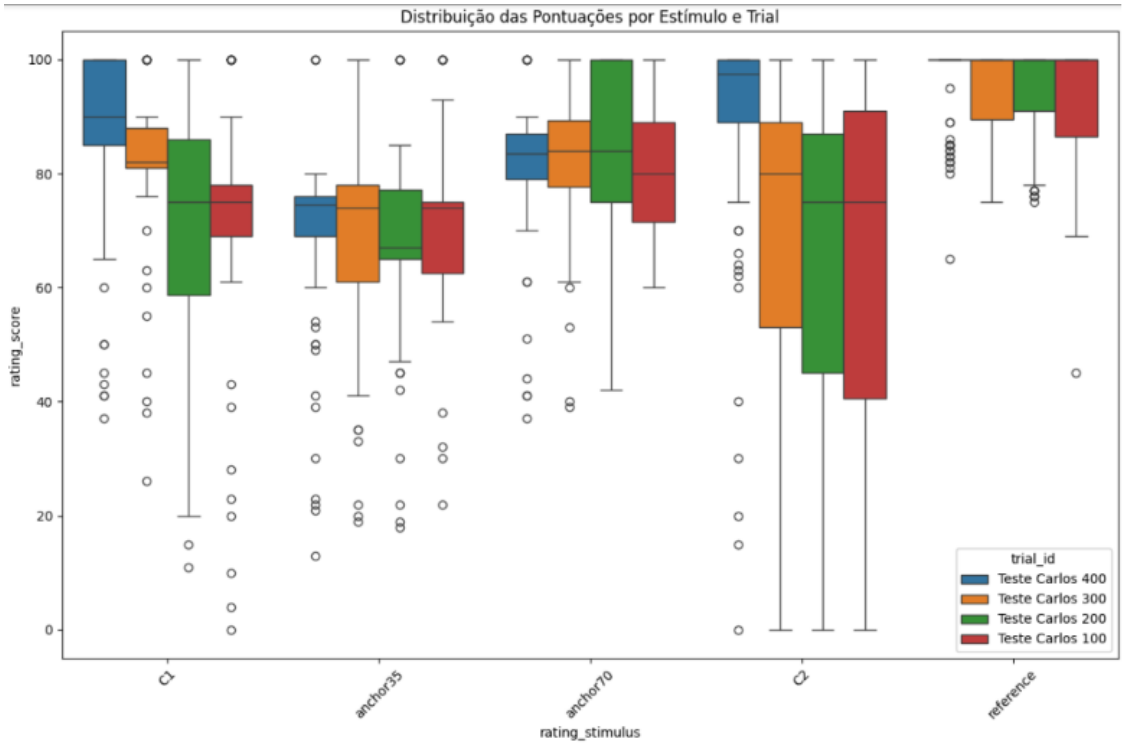

Figura 1. Boxplot da distribuição das pontuações por estímulo. C1 representa o VITS, e C2 representa o Tacotron2.

low-resource, com uso de até 5 minutos de fala no treinamento de *fine-tuning*, para clonagem de voz; ou seja, avaliar a capacidade de adaptação dos modelos base pré-treinados fazendo uso de dados limitados de uma nova voz personalizada.

O modelo VITS, quando treinado com 20 minutos, mostrou resultados com alta qualidade após apenas 1 hora de treinamento. Por outro lado, o Tacotron2, sob as mesmas condições, apresentou maior variabilidade e menor consistência na qualidade do áudio sintetizado. Mesmo quando treinado com 5 minutos o VITS apresentou boa qualidade e baixa variância. Ao comparar o tempo de treinamento, o modelo VITS mostrou-se mais eficiente, alcançando bons resultados em menos tempo e com menos dados em relação ao Tacotron2.

Os resultados indicam que o VITS não só oferece uma síntese de fala de melhor qualidade, com maior similaridade à voz original e menor variância entre as amostras sintetizadas, mas também é mais eficiente em termos de tempo de treinamento em cenários *low-resource*.

5. Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Referências

Casanova, E., Junior, A. C., Shulby, C., Oliveira, F. S. d., Teixeira, J. P., Ponti, M. A., and Aluísio, S. (2022). Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese. *Language Resources and Evaluation*, 56(3):1043–1055.

Hayashi, T., Yamamoto, R., Yoshimura, T., Wu, P., Shi, J., Saeki, T., Ju, Y., Yasuda, Y., Takamichi, S., and Watanabe, S. (2021). Espnet2-tts: Extending the edge of tts research.

Kim, J., Kong, J., and Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech.

Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.

Lux, F., Koch, J., and Vu, N. T. (2022). Low-resource multilingual and zero-shot multispeaker tts.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis.