TREINAMENTO E AVALIAÇÃO DE GENERALIZAÇÃO DO VOCODER MELGAN
PROCESSAMENTO DE LINGUAGEM NATURAL - PROF. ADRIANO VELOSO

Thiago Malta Coutinho

thiagomaltac@gmail.com

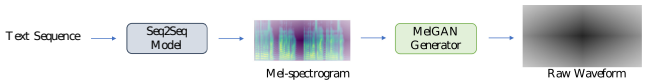
ABSTRACT

Esse trabalho avalia o treinamento da rede Adversarial Generativa MelGan[1] em um Corpus em língua portuguesa e sua capacidade de generalização. A avaliação é feita sobre um conjunto de teste contendo dados do Corpus de treinamento e outros Corpus com dados de outras línguas. Os resultados demonstraram que mesmo com uma quantidade relativamente pequena de iterações de treinamento a rede foi capaz de generalizar para outros Corpus contendo diversos locutores, inclusive locutores de gênero diferente da locutora de treino.

***Index Terms*—** GAN, Generative Adversarial Network, Vocoder, TTS, Text-to-Speech, Natural Language Processing, MelGan

1. INTRODUÇÃO

1.1. Text-to-Speech(TTS)



Result 1

Fig. 1. Problema de geração de áudios a partir de entradas de texto(TTS).

O problema de Text-to-Speech(TTS) consiste em vocalizar áudio a partir de entradas de texto. Até o ano de 2016 as abordagens para o problema eram focadas em modelos concatenativos, que fornecem resultados satisfatórios mas com artefatos de descontinuidade entre fonemas. Em 2016 o modelo WaveNet[2] mostrou que redes neurais são modelos viáveis para o problema de TTS, gerando resultados Estado da Arte(SOTA).

Os modelos TTS que utilizam redes neurais convergem em uma estrutura em comum contendo duas partes: transformação do texto de entrada para espectrogramas-mel(são gerados a partir de uma transformação baseada no logaritmo, escala que melhor representa a percepção do ouvido humano) e transformação dos espectrogramas para áudios de saída. A primeira etapa é feita por um modelo do tipo *Seq2Seq*, predominantemente o Tacotron-2[3]. A segunda fase do problema é abordada por outro tipo de modelo, os *Vocoders*. Os *Vocoders* são redes neurais capazes de transformar espectrogramas-mel em áudios, entre as redes presentes na literatura existem redes auto-regressivas[2][4] e redes não-auto-regressivas[5][6][7].

Em modelos auto-regressivos a inferência da rede possui dependência sequencial e, portanto, produzem áudios com velocidades menores que redes não-auto-regressivas. No entanto inferência em velocidades maiores que tempo real podem ser atingidas. Devido à inferência mais lenta, impossibilitando aplicações de grande porte, modelos não-auto-regressivos tem sido o alvo de novos estudos.

2. METODOLOGIA

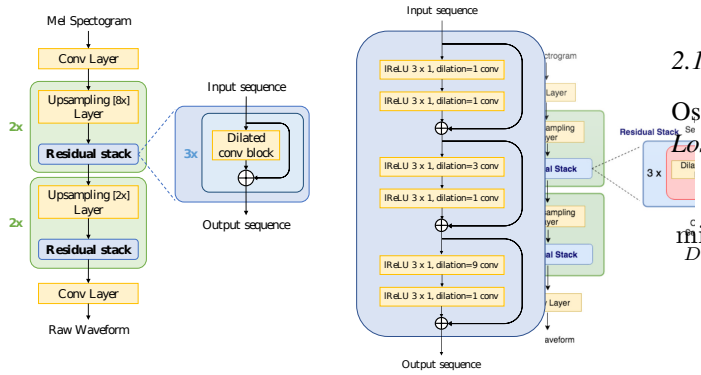
2.1. MelGan

MelGan[1] é uma rede Generativa Adversarial, a ideia é treinar duas redes, uma discriminadora e outra generativa. As duas redes são treinadas em conjunto, uma tentando exercer seu papel melhor que a outra. O objetivo da rede generativa é receber espectrogramas-mel e gerar áudios nos quais a rede discriminativa não consiga distinguir se são reais ou gerados pela rede generativa. O objetivo da rede discriminadora é distinguir entre áudios reais e áudios gerados pela rede adversária, com o menor erro possível. O fim do treinamento acontece quando as duas redes entram em equilíbrio, ou seja, nenhuma das duas consegue melhorar mais. Ao atingir o equilíbrio espera-se que a rede generativa tenha aprendido a gerar saída indistinguíveis das reais, caso a complexidade da rede geradora e discriminadora seja suficiente.

2.1.1. Rede Geradora

A rede geradora é totalmente convolucional. Como os espectrogramas-mel possuem dimensão temporal 256x menor que o áudio de saída, camadas de convolução transposta são utilizadas para fazer *upsampling* das amostras de entrada. Cada camada de convolução transposta é seguida por empilhamentos de blocos residuais[2] com convoluções dilatadas. Camadas compostas por *upsampling* e empilhamentos resid-

uais são empilhadas umas sobre as outras para induzir um viés que relaciona amostras temporalmente afastadas.



(a) Camadas do gerador. (b) Empilhamentos Residuais.

Fig. 2. Estrutura generativa da rede.

Para as camadas de convolução transpostas os autores utilizaram tamanho de kernel divisível pelo stride, para evitar os *Checkerboard Artifacts*[8]. Para cada camada da rede geradora, os autores utilizam *Weight Normalization*[9], uma reparametrização dos pesos da rede:

$$\mathbf{W} = \frac{g}{||\mathbf{v}||}\mathbf{v} \implies ||\mathbf{W}|| = g \tag{1}$$

2.1.2. *Rede Discriminadora*

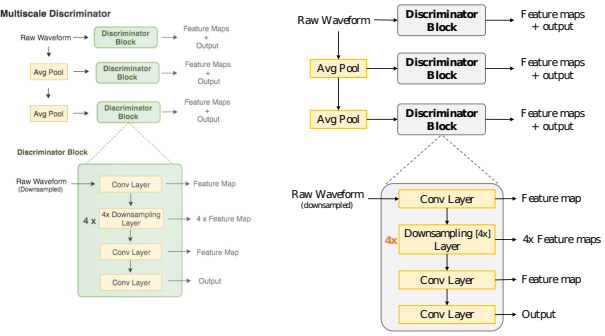


Fig. 3. Estrutura do discriminador com o detalhamento do bloco discriminador.

A rede discriminadora possui três discriminadores com estrutura idêntica mas operando em diferentes frequências temporais. O primeiro discriminador opera sobre o sinal de áudio puro, o segundo sobre o sinal amostrado por um fator de 2 e o terceiro por um fator de 4. Os autores justificam a abordagem dizendo que os áudios possui informação em diferentes níveis. Portanto cada discriminador aprende características para diferentes faixas de frequência do áudio. Operar apenas com um discriminador no áudio puro iria introduzir um viés resultando no aprendizado apenas de *features* baseadas em componentes de baixa frequência.

2.1.3. *Funções objetivo*

Os autores usaram como função objetivo a versão da *Hinge Loss*[] para GAN's:

$$\min_{D_k} \mathbb{E}_x\Big[\min(0, 1-D_k(x))\Big] + \mathbb{E}_{s,z}\Big[\min(0, 1+D_k(G(s,z)))\Big],\ \forall k \tag{2}$$

$$\min_G \mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] \tag{3}$$

Além do sinal do discriminador, os autores utilizam uma função de perda com *Feature Matching*[] para a função objetivo do gerador. O objetivo é minimizar a distância L1 entre o *feature map* do discriminador para áudios reais e sintéticos.

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{x,s\sim p_{\text{data}}}\left[\sum_{i=1}^{T}\frac{1}{N_i}||D_k^{(i)}(x) - D_k^{(i)}(G(s))||_1\right] \tag{4}$$

Assim, função de perda do gerador é dada por:

$$\min_G\left(\mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] + \lambda\sum_{k=1}^{3}\mathcal{L}_{\text{FM}}(G, D_k)\right) \tag{5}$$

2.2. Corpus

O Corpus utilizado foi obtido nas bibliotecas abertas LibriVox[10] e Projeto Gutenberg[11]. Os áudios e texto foram são do livro A Relíquia, de Eça de Queirós, o áudio está disponível em [12] e o texto em [13]. O Corpus contém 10 horas e 31 minutos de áudio, com locutora feminina falando em língua portuguesa do Brasil. Os arquivos de áudio vem particionados em pedaços que variam entre oito minutos e uma hora.

2.3. Pré-processamento dos dados

Para a rede aprender com mais eficácia, os áudios a serem aprendidos devem ter duração máxima até 30 segundos, áudios maiores são mais difíceis de modelar devido à duração. A partir dessa premissa, os arquivos de áudio do Corpus foram particionados para tamanhos menores.

Para realizar a tarefa de maneira automatizada o alinhamento do texto com o áudio foi feito a partir de uma rotina baseada no algoritmo Virtebi. Os caracteres especiais do tipo UTF-8 foram removidos do texto, junto com parêntesis e

traços, os acentos foram substituídos por texto comum para facilitar o alinhamento. Como o texto vem separado com quebra de linhas em frases e parágrafos, os áudios cortados teriam tamanho próximo ao desejado.

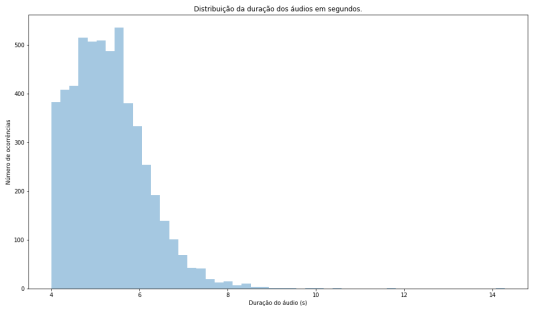


Fig. 4. Distribuição da duração dos áudios do Corpus.

Após o particionamento dos áudios, os arquivos com duração menor que quatro segundos foram filtrados para garantir um tamanho mínimo. A Figura 4 mostra a distribuição final da duração dos áudios do Corpus. O arquivo com maior duração tem 14 segundos, o menor tem 4 segundos e a média está em 5,3 segundos. A duração total do Corpus foi reduzida para 7 horas e 57 minutos, com 5.397 arquivos de entrada.

2.4. Treinamento

A rede foi treinada por 957 mil iterações com os hiper parâmetros especificados na Tabela 1.

Parâmetro	Valor
Número de blocos discriminadores	3
Número de camadas residuais	3
Fator de Downsampling	4
Número de canais mel	80
λ	10
Tamanho do Batch	16

Table 1. Hiper parâmetros utilizados na rotina de treinamento da rede.

Em [1] os autores treinam o algoritmo por 2,5M iterações, até a convergência.

2.5. Avaliação do desempenho do modelo

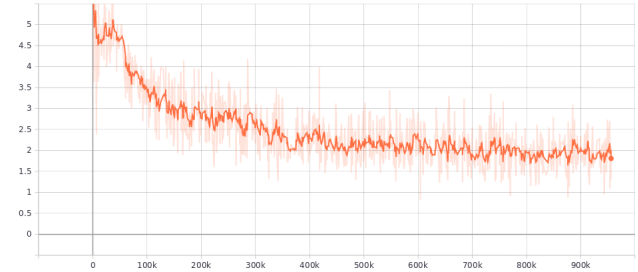
A avaliação do desempenho do modelo é feita qualitativamente nos áudios gerados pela rede, uma métrica de avaliação é o MOS(*Mean Opinion Score*), uma pesquisa de avaliação qualitativa com embasamento estatístico feita com um grande número de ouvintes.

Para avaliar a generalização do modelo um Corpus de teste foi construído a partir de áudios de diferentes Corpus. A estrutura dos dados está descrita na Tabela2:

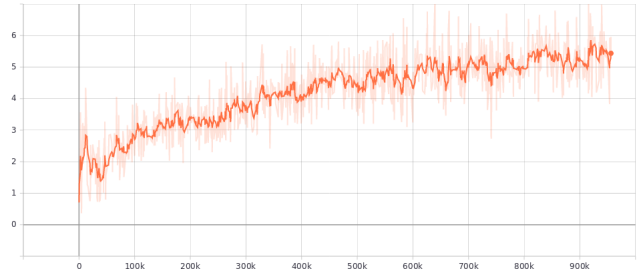
Número de Amostras	Gênero do(a) Locutor(a)	Língua	Ref.
15	Feminino	PT-BR	[12]
8	Feminino	EN-US	[14]
3	Feminino	PT-BR	[15]
8	Masculino	PT-BR	[15]
4	Feminino	PT-PT	[16]
4	Feminino	PT-BR	[17]

Table 2. Composição do Corpus de teste.

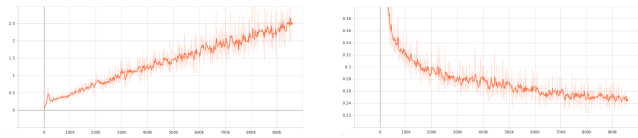
3. RESULTADOS



(a) Loss do discriminador.



(b) Loss do gerador.



(c) Loss do Feature Matching. (d) Loss do espectrograma-mel.

Fig. 5. Funções de perda do treinamento da rede MelGan.

As funções de perda do discriminador e gerador apresentaram as características esperadas, uma é similar ao inverso da outra. Além disso, nenhuma rede apresentou crescimento ou decrescimento abrupto na função de perda, podendo gerar *Mode Collapse*. A função de perda da reconstrução do espectrograma-mel(definida pela distância L1 entre a reconstrução do espectrograma e o original), decresceu com

o aumento das iterações de treinamento. A loss do *Feature Matching* apresentou característica de crescimento constante.

4. CONCLUSÕES

Esse trabalhou explorou o treinamento do vocoder MelGan e analisou qualitativamente os resultados obtidos em um Corpus com diferentes locutores em 3 línguas.

Apesar do número de iterações de treino da rede ser consideravelmente menor que o recomendado pelos autores do modelo[1], a rede apresentou resultados satisfatórios. Os áudios gerados a partir de espectrogramas da mesma locutora do conjunto de treinamento apresentaram poucos artefatos e som limpo. Para outros locutores os áudios apresentaram mais artefatos e qualidade baixa, no entanto os sons são inteligíveis e demonstram a capacidade de generalizar e característica generativa da rede.

5. REFERENCES

[1] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[2] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[3] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2018.

[4] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," 2018.

[5] Wei Ping, Kainan Peng, and Jitong Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," 2018.

[6] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis, "Parallel wavenet: Fast high-fidelity speech synthesis," 2017.

[7] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *CoRR*, vol. abs/1811.00002, 2018.

[8] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[9] Tim Salimans and Diederik P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," 2016.

[10] LibriVox, "https://librivox.org/," .

[11] Projeto Gutenberg, "https://www.gutenberg.org/wiki/pt_principal," .

[12] A Reliquia, "https://librivox.org/a-reliquia-by-jose-maria-de-eca-de-queiros/," .

[13] A Reliquia(texto), "https://www.gutenberg.org/wiki/pt_principal," .

[14] LJSPEECH, "https://keithito.com/lj-speech-dataset/," .

[15] Eça de Queirós Cartas de Inglaterra, "https://librivox.org/cartas-de-inglaterra-by-jose-maria-de-eca-de-queiros/," .

[16] Eça de Queirós Contos, "https://librivox.org/contos-by-jose-maria-de-eca-de-queiros/," .

[17] Eça de Queirós O defunto, "https://librivox.org/o-defunto-by-jose-maria-de-eca-de-queiros/," .