Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim

Rodrigo Kobashikawa Rosa, Danilo Silva

***Resumo*— Este artigo apresenta o treinamento de um modelo do estado da arte com redes neurais, o Tacotron 2, utilizando um conjunto de dados de voz de código aberto do projeto Common Voice. Foram avaliados os resultados do treinamento do modelo do zero e da aplicação de *transfer learning* a partir de um modelo pré-treinado em inglês. Os resultados mostraram que é possível treinar o modelo com recursos de dados limitados.**

***Palavras-Chave*— Redes neurais, Síntese de voz, Tacotron 2, Português Brasileiro.**

***Abstract*— This paper presents the training of a state-of-the-art neural network model, Tacotron-2, using a open-source voice dataset from the Common Voice project. Results from training the model from scratch and by applying transfer learning of a pre-trained english model were evaluated. The results show that it is possible to train the model with limited data resources.**

***Keywords*— Neural networks, Speech synthesis, Tacotron-2, Brazillian Portuguese.**

I. Introdução

Sistemas tradicionais, como métodos de síntese concatenativa, dominaram as aplicações de síntese de voz por muito tempo. Esses métodos são complexos e trabalhosos de serem implementados, baseando-se no conhecimento de especialistas da área de processamento de voz, e exigem várias etapas de análise e extração de *features* linguísticas e acústicas [1].

Com a popularização do aprendizado profundo para as mais diversas aplicações, também surgiram aplicações na área de síntese de voz. Os modelos de *Text-to-Speech* (TTS) *end-to-end* com redes neurais podem ser treinados diretamente com pares de áudio e texto sem precisar de etapas intermediárias de extração de *features*, pois o próprio modelo é capaz de fazer esse trabalho e também tornando mais fácil condicionar o modelo para outros falantes, linguagens diferentes ou até sentimentos diferentes na fala [2].

Uma dificuldade da pesquisa em TTS com redes profundas para português brasileiro, em comparação com línguas como inglês e mandarim, é a menor disponibilidade de dados abertos. Este trabalho busca treinar o modelo em um conjunto de dados aberto adaptado, utilizando o modelo Tacotron 2 [3], com o *vocoder* Griffin-Lim [4] em vez do WaveNet [5] devido à limitação de recurso computacional para o treinamento, e avaliar os resultados obtidos para o português brasileiro.

II. Tacotron 2

A arquitetura do modelo Tacotron 2 consiste de dois componentes principais: uma rede de predição de espectrogramas mel e um *vocoder* neural WaveNet modificado para ser condicionado a partir de espectrogramas mel. Neste artigo será utilizado apenas o primeiro componente e para a reconstrução da fase do sinal será utilizado o *vocoder* de fase Griffin-Lim.

A rede de previsão de espectrogramas mel é baseada nos modelos *sequence-to-sequence* e inclui um codificador, um decodificador com modelo de atenção, uma rede de pós processamento e uma de predição do *token* de parada. O codificador converte uma sequência de caracteres em uma representação de *feature* oculta resumindo as principais informações para o decodificador consumir e gerar o espectrograma de saída.

A saída do codificador passa pelo modelo de atenção *location-sensitive* [6] que resume a sequência codificada contendo as partes mais importantes da entrada que são relevantes para cada passo do decodificador. O decodificador é uma rede recorrente auto regressiva que prevê cada passo com a informação da predição do passo anterior.

O espectrograma mel predito é somado com a saída da rede de pós processamento que prevê o resíduo com a intenção de melhorar a reconstrução geral. A saída final é utilizada pelo *vocoder* para gerar as formas de onda no domínio do tempo.

III. Metodologia

Após a seleção e a preparação do conjunto de dados foi feita a escolha pelo modelo Tacotron 2, por ser um dos principais modelos atuais e com implementações de código aberto robustas para o experimento. Foram avaliadas duas abordagens diferentes de treinamento: treinando o modelo do zero e aplicando *transfer learning* com base em um modelo pré-treinado em inglês.

A. Conjunto de dados

Diferentemente de linguagens como o inglês que possuem inúmeros conjuntos de dados abertos para o desenvolvimento de sistemas TTS, o mesmo não é visto para o português. Porém, a iniciativa Common Voice [7] tem disponível um imenso corpus de fala multi-idioma aberto para o uso e construído de forma colaborativa. Existem conjuntos de dados para mais de 60 idiomas, sendo que no Common Voice Corpus 4 são 27 horas validadas de áudio para o português brasileiro.

Explorando os dados disponíveis e analisando a distribuição da contribuição de cada falante dentro do conjunto de dados, foi encontrado que um dos falantes representava 34,48% dos dados com 7631 frases validadas e cerca de 6 horas de áudio. Foi necessário adequar os arquivos de áudio para o padrão wav com uma taxa de amostragem de 22.05 kHz.

Rodrigo Kobashikawa Rosa, Danilo Silva, Departamento de Engenharia Elétrica e Eletrônica, UFSC, Florianópolis-SC, e-mails: rodrigo-krosa@gmail.com, danilo.silva@ufsc.br;

B. Treinamento

Foram utilizadas implementações do Tacotron 2 de código aberto para os experimentos. Inicialmente foi treinada a implementação desenvolvida por Rayhane Mama [8] começando o treinamento do zero, diretamente com o conjunto de dados em português, utilizando a GPU disponibilizada no ambiente Google Colab. O treinamento levou em torno de 70 horas para ajustar os parâmetros do modelo durante 70 mil iterações.

Como o conjunto de dados obtido de 6 horas é relativamente pequeno se comparado aos conjuntos de outras línguas que possuem mais de 20 horas, o segundo experimento foi utilizar *fine tuning* de um modelo pré-treinado em inglês. A implementação desenvolvida por Rayhane não possui modelos pré-treinados disponíveis, portanto foi utilizada a implementação desenvolvida no repositório TensorFlowTTS [9] que conta com um modelo do Tacotron 2 treinado com 65 mil passos e com suporte para o *fine tuning*. Foi treinado o modelo por 30 mil passos, utilizando uma GPU GTX 1060 Ti 6 Gb.

IV. Resultados e discussão

Para a avaliação dos resultados do modelo, foi feita a síntese de frases fora do treinamento, utilizando um conjunto de 200 frases foneticamente balanceadas [10] sintetizadas nos dois modelos. Dois tipos comuns de erros encontrados em sistemas TTS foram avaliados: (1) o número de palavras que foram puladas e (2) erros de pronúncia. Os áudios sintetizados podem ser encontrados no repositório do projeto[1]. Na Fig. 1 pode ser observado um exemplo de um dos espectrogramas mel sintetizados pelo modelo treinado do zero.

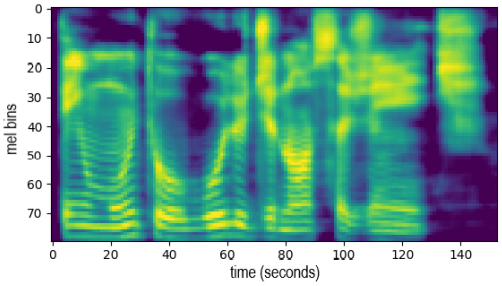

Fig. 1. Espectrograma mel sintetizado a partir de uma frase de avaliação.

O *vocoder* Griffin-Lim foi utilizado para a síntese da forma de onda da voz a partir dos espectrogramas mel, por ser um *vocoder* que necessita de menos recursos computacionais em comparação ao WaveNet. O Griffin-Lim consegue reconstruir iterativamente a fase do sinal com o auxílio da transformada inversa de Fourier de tempo curto, e assim sendo possível recuperar o sinal da forma de onda no domínio do tempo. O áudio gerado pelo Griffin-Lim possui qualidade inferior ao WaveNet, porém para o fim de avaliar a capacidade de geração de áudio do sistema TTS, ele cumpre seu propósito.

As 200 frases sintetizadas correspondem a um total de 1349 palavras. Observou-se que o modelo treinado do zero foi superior em ambas as métricas avaliadas. Um percentual de 0,89% das palavras foram puladas e uma taxa de 3,78% sofreram erros de pronúncia. Em comparação, o modelo utilizando *transfer learning* do inglês teve 18,60% das palavras puladas ou simplesmente sem conseguir sintetizar a frase e 5,91% do total de palavras sintetizadas sofreram erros de pronúncia. Na Tabela I foi feito o resumo das análises obtidas.

TABELA I
Comparação entre os resultados dos modelos do Tacotron 2.

Modelo	**Treinado do zero**	**Transfer learning**
Palavras puladas	12 (0,89%)	251 (18,60%)
Erros de pronúncia	51 (3,78%)	70 (5,19%)

Vale mencionar que o resultado acima difere do obtido em [11], o qual obteve um melhor desempenho a partir de *transfer learning* do inglês. Essa diferença pode ter sido ocasionada por termos utilizado duas implementações diferentes para síntese com e sem *transfer learning*, além do fato de que o trabalho em [11] difere do nosso em diversos aspectos estruturais (implementação do Tacotron 2 utilizada, *vocoder*, conjunto de dados e uso de fonemas como entrada).

V. Conclusões

Nesse estudo sobre a utilização de um conjunto de dados aberto para uso e limitado em relação à quantidade e qualidade das amostras, foram investigadas duas abordagens diferentes para um sistema TTS com o modelo Tacotron 2. Os resultados demonstraram que é possível treinar outras linguagens como o português com poucas mudanças no modelo original. Pela qualidade do áudio sintetizado e os poucos erros de pronúncia e de síntese, o modelo treinado do zero parece satisfatório para aplicações de TTS com uma certa tolerância de erros. Em trabalhos futuros, seria interessante treinar um *vocoder* neural como o WaveNet para poder fazer análises subjetivas da qualidade da voz sintetizada, algo que não feito neste trabalho devido à baixa qualidade de síntese do Griffin-Lim.

[1] https://github.com/kobarion/tacotron2-GL-brazillian-portuguese

Referências

[1] P. Taylor, "Text-to-Speech Synthesis,"Cambridge: Cambridge University Press, 2009.
[2] Y. Wang et al. Tacotron: Towards End-to-End Speech Synthesis. Proc. Interspeech 2017, 4006-4010. 2017.
[3] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,"2018 IEEE International Conference on Acoustics, ICASSP (2018), pp. 4779-4783
[4] D. Griffin e J. Lim, “Signal estimation from modified short-time Fourier transform,” ICASSP (1983).
[5] A. Oord et al., “WaveNet: A generative model for raw audio,” CoRR, vol. abs/1609.03499, 2016.
[6] J. K. Chorowski et al., “Attention-based models for speech recognition,” em Proc. NIPS, 2015, pp. 577–585
[7] R. Ardila et al., "Common Voice: A Massively-Multilingual Speech Corpus". Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 4211—4215
[8] R. Mama, "Tacotron 2,"https://github.com/Rayhane-mamah/Tacotron-2.
[9] "TensorFlowTTS,"https://github.com/TensorSpeech/TensorFlowTTS.
[10] I. Seara, "Estudo Estatístico dos Fonemas do Português Brasileiro Falado na Capital de Santa Catarina para elaboração de Frases Foneticamente Balanceadas,"Dissertação de Mestrado, UFSC, 1994.
[11] E. Casanova, "Síntese de voz aplicada ao português brasileiro usando aprendizado profundo". Trabalho de conclusão de Curso, UTFPR, 2019.