Uma análise da naturalidade de síntese de fala com redes neurais no português em cenários de restrição

May 24, 2020

1 Introdução

O texto apresentado nessa seção é um adendo ao texto principal com intuito de tratar objetivamente dos objetivos da pesquisa e da metodologia planejada. As seções aqui citadas farão referência ao texto principal visando facilitar a futura junção dos textos. O intuito desse anexo é servir como referência para a questão de pesquisa, as hipóteses levantadas e a metodologia proposta. As seções desse texto serão posteriormente expandidas no texto principal a ser completamente revisado.

1.1 Contextualização

A fala é o meio mais comum de comunicação entre seres humanos[fla94]. A expressão falada é um dos meios que trocamos cotidianamente ideias, emoções e refletimos nossas personalidades[TT12]. Computadores têm se tornado cada vez melhores na complexa tarefa que é interagir conosco usando a fala [CDG+18] tanta na tarefa de compreensão (*Speech to Text* (STT)) quanto na síntese (*Text to Speech* (TTS)).

A síntese de fala já é um problema com vários produtos disponíveis para uso pelo público [1] [2] presente no cotidiano de muitos através das mais diversas aplicações. Entretanto esses modelos costumam permitir apenas síntese de falantes pré-determinados e frequentemente são modelos pagos e proprietários. Recentemente diversas pesquisas demonstraram excelentes resultados na criação de modelos de voz , seja o processo completo de síntese ou parte, a partir de gravações de um ou mais falantes na língua inglesa com modelos neurais. Podemos destacar a grande influência dos trabalhos do Baidu com o Deep Voice[ACC+17, ADG+17, PPG+17, ACP+18] e da Google com o Tacotron [WSRS+17, SPW+17, vdODZ+16] todos com propostas similares de módulos neurais.

1.2 Questão de Pesquisa e Hipóteses

Será que podemos obter a síntese de fala com esses modelos na língua portuguesa para um falante qualquer com resultados similares? Essa mesma síntese pode ser obtida com resultados satisfatórios com restrições de tempo e de dados em seu treinamento? Como podemos medir de maneira quantitativa os resultados de modelos de síntese de modo a facilitar a avaliação qualitativa?

Para validar essas questões devemos responder várias hipóteses como:

1. Podemos treinar os mesmo modelos com dados em português?
2. Podemos medir esses modelos treinados de maneira similar a proposta nos trabalhos?
3. Podemos aproveitar algum conhecimento aprendido de outro conjunto de dados no nosso treinamento?
4. Podemos reduzir o conjunto de dados disponíveis e obter resultados comparavelmente bons?
5. Podemos restringir o tempo de treinamento do modelo e obter resultados comparavelmente bons?

[1] https://cloud.google.com/text-to-speech/
[2] https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

Queremos levantar comparativamente se um modelo pré-treinado em outra língua que já tenha apresentado bons resultados perceptuais pode ser benéfico para a síntese na língua portuguesa observando-se o tempo de treinamento, a convergência da função de erro do modelo e a avaliação perceptual de um grupo de ouvintes. Para responder essas questões estabelecemos uma sequência de decisões e experimentos destrinchados na próxima sessão.

2 Metodologia

Queremos levantar como os atuais algoritmos de síntese neural se comportam perceptualmente na língua portuguesa com restrições de dados e de tempo na geração do modelo. A primeira decisão é a escolha do modelo neural a qual desejamos replicar. Por uma questão de custo de implementação decidi abordar apenas um modelo, o Tacotron 2 [SPW+17] desenvolvido pela Google cuja implementação encontra-se disponível em código aberto[3]. O modelo também foi escolhido pois apresentar os melhores resultados de acordo com o comparativo levantado de modelos de síntese no inglês em Dezembro de 2017[4]. Tendo sido fixado o modelo a ser utilizado estabelecemos também conjuntos de dados a serem utilizados nessa atividade como pode ser visto na Tabela 1.

Sigla	Nome	Tópico	Descrição Técnica
CONST	Constituição	Leitura da constituição nacional cujo texto e arquivos de áudio original foram processados pela equipe FalaBrasil de modo a adequar-se às necessidades do estudo de fala	Segmentados em aproximadamente 30 segundos, falante masculino, em ambiente de gravação de rádio. 9000 arquivos; aproximadamente 9h de gravações
LAPS	LAPS BenchMark	Corpus de voz utilizado para avaliação de desempenho de sistemas LVCSR.	700 frases, o corpus possui 35 locutores com 20 frases cada, sendo 25 homens e 10 mulheres, o que corresponde a aproximadamente 54 minutos de áudio
YT	Youtube (Cellbit)	Áudios extraídos de vídeos do YouTube de um *youtuber* no estilo *vlog*. Legendas manualmente inseridas nos vídeos pelo próprio autor dos vídeos.	23 vídeos com legenda publicados entre 04/07/2017 e 27/03/2018 totalizando aproximadamente 23h de áudio. O locutor é um youtuber do sexo masculino com 21 anos. Os arquivos foram seccionados baseado nos tempos de cada linha inteira do arquivo de legenda. 4203 arquivos (2.9±4.2s).

Table 1: Descrição dos Datasets

Esses dados foram selecionados fruto de uma pesquisa de dados anotados em português com amostras de falantes e frases. Os dados CONST e LAPS foram obtidos através do projeto FalaBrasil [NSKA08]. Originalmente os dados de CONST foram obtidos da Câmara dos Deputados[5] mas como foram posteriormente anotados, alinhados e seccionados por frase mantemos a referência do grupo FalaBrasil. Os dados YT foram obtidos de um canal do YouTube especificado na tabela com o intuito de fornecer uma fonte de áudio menos estruturada, menos formal e de fácil acesso público permitindo a qualquer um replicar o experimento com outro falante e uma estratégia comparativa similar.

Fazendo um mapeamento direto com os questionamentos levantados em 1.2 propomos os experimentos e avaliações abaixo. Cada tópico está propondo alguma solução para o problema levantado na questão respectiva.

[3] https://github.com/carpedm20/multi-speaker-tacotron-tensorflow
[4] https://paperswithcode.com/sota/speech-synthesis-on-north-american-english
[5] http://bd.camara.gov.br/bd/handle/bdcamara/1708

1. Os dados apresentados possuem formato similar ao utilizado no modelo pré-treinado de referência. Sendo assim com pequenos ajustes pode ser facilmente adaptado para responder o questionamento 1

2. Para estabelecer-se uma linha de comparação treinaremos um modelo para cada conjunto de dados apresentado com 100% dos dados de cada. Uma primeira métrica proposta a ser avaliada para responder ao questionamento 2 é o MOS Score que consiste na média de opiniões coletadas de ouvintes que avaliaram a síntese do modelo perceptualmente. Essa métrica é utilizada no trabalho escolhido e facilitará a comparação de resultados

3. Para responder ao questionamento 3 podemos nos utilizar de *transfer learning*[6] com um modelo pré-treinado no inglês. O intuito de se aproveitar do *transfer learning* é observar se as características aprendidas pelo modelo em uma língua podem ser aproveitadas em outra

4. Para responder o questionamento 4 propomos utilizar frações (5, 15, 33 e 50%) dos dados apresentados para comparar seus resultados com o modelo base treinado com 100% dos dados

5. Para responder o questionamento 5 propomos treinar o modelo em uma quantidade fixa de épocas de 1k, 5k, 33k e 50k

Como o intuito do problema 3 é utilizar do *transfer learning* no auxílio do treino inicial do modelo não nos é interessante propor épocas muito maiores que nosso modelo de referência. Como nosso modelo de referência é treinado em 100k épocas estabelecemos 50k como faixa de corte máxima. Para as frações de dados propostas (1k, 5k, 33k e 50k) escolheremos aleatoriamente dentro da totalidade de cada conjunto selecionado. Para facilitar a reprodução a ordem dos arquivos usados e a semente aleatória usada na seleção aleatória serão disponibilizados. Tendo em vista esses parâmetros definidos queremos gerar um total de 21 modelos para cada conjunto de dados usando as restrições de tempo e . Com essas 21 hiperparametrizações fixadas para cada um dos 3 datasets descritos na Tabela 1 temos um total de 63 modelos.

Novamente para cada solução proposta na última lista enumerada fazemos abaixo um mapeamento das atividades respectivas.

1. Conforme mencionado são necessários alguns pequenos ajustes nos dados para estarem conforme o modelo espera. Nessa atividade implementamos o código necessário para utilizar os dados levantados como entrada do modelo escolhido.

2. Estando os dados alinhados com o que o modelo espera podemos treinar a versão base do modelo que recebe 100% dos dados de cada um dos três conjuntos propostos. Esse primeiro modelo é treinado até 100k épocas, sendo as frações de tempo intermediárias salvas nesse mesmo treino.

3. Para a comparação com o *transfer learning* já temos um modelo disponível no mesmo repositório mencionado com um modelo pré-treinado no inglês com o dataset LJ Speech[7]. Ese modelo foi treinado até 100k épocas, por isso esse mesmo valor foi escolhido no modelo anterior visando facilitar a comparação.

4. Com esse modelo selecionado temos de treinar os 16 modelos restantes, que correspondem as combinações das 4 frações de dados com os 4 valores fixados de épocas.

5. Tendo então os 63 modelos treinados propomos um filtro dos melhores modelos através de uma métrica objetiva que quantifique a qualidade da voz. A literatura de fonoaudiologia disponibiliza diferentes versões da *Acoustic Voice Quality Index*(AVQI) onde cada nova versão tenta refinar a atual equação. Propomos usar essa métrica para filtrar apenas os melhores dos 63 modelos tanto para reduzir o número de modelos a serem ouvidos por falantes humanos. Buscamos também estabelecer um padrão comparativo com alguma referência da literatura para a escolha das frases já que não existe padrão apontado na síntese no português.

6. Por fim tendo filtrado apenas os melhores modelos desejamos utilizar o questionário com o mesmo padrão de questões dos trabalho relacionados (Disponível no Apêndice F em Jia et al.[JZW+18]).

[6] https://machinelearningmastery.com/transfer-learning-for-deep-learning/
[7] https://keithito.com/LJ-Speech-Dataset/

Para isso propomos duas atividades distintas de coletas de respostas. A primeira imita os trabalhos relacionados através do uso da ferramenta Amazon Turk, uma plataforma onde voluntários recebem incentivo monetário para responder questionários. Essa proposta é interessante por dois motivos: fomenta o questionamento da naturalidade da fala sendo ouvida por falantes de outras línguas e obtêm resultados mais rápidos devido ao incentivo monetário. A segunda proposta consiste em replicar esse questionário em um formulário online (e.g. MonkeySurvey) e coletar resultados do meio acadêmico, amigos e participantes espontâneos da comunidade.

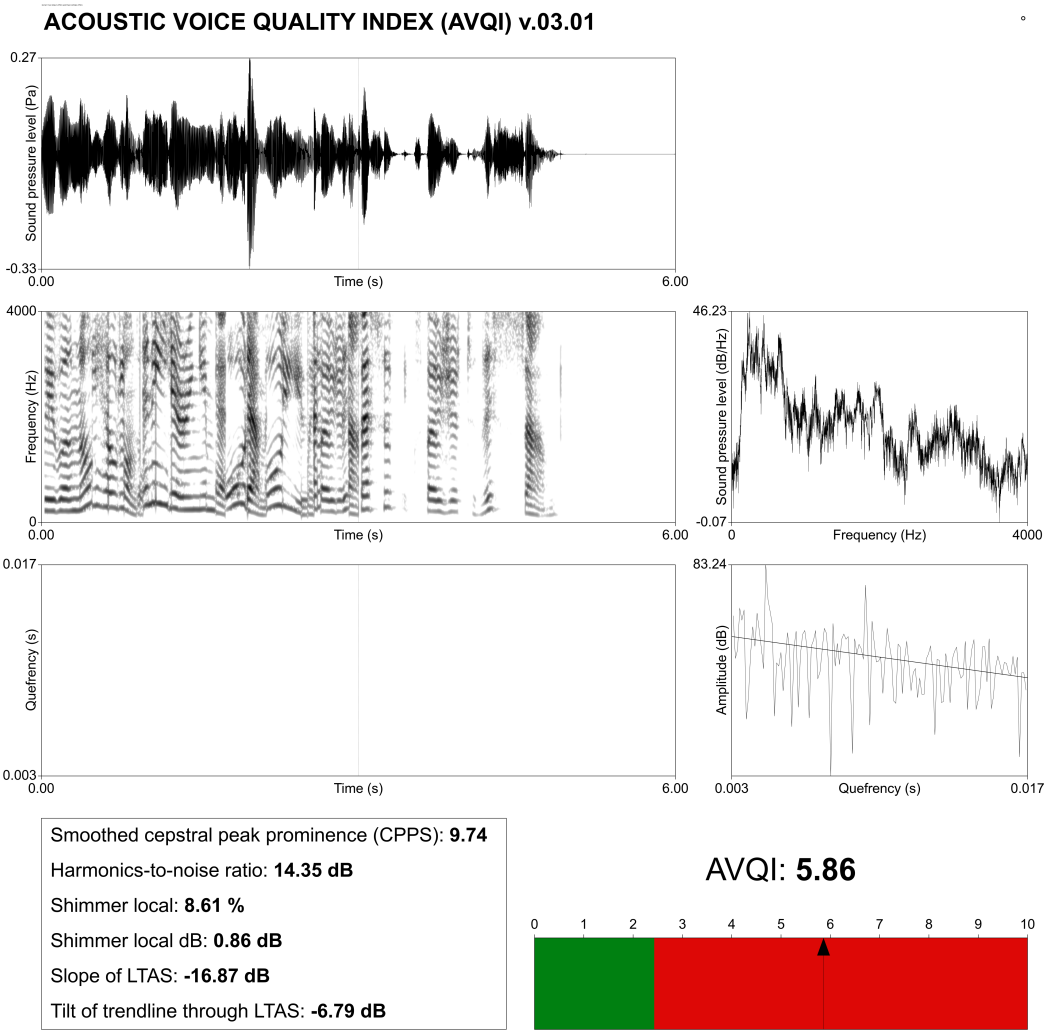

Figure 1: Exemplo de saída do Praat com os gráficos intermediários e o AVQI Score calculado

Quantitativamente os modelos são analisados a partir da variação do valor da função de erro no tempo e o tempo para cada quantidade de épocas estudada. Qualitativamente propomos um questionário com ouvintes humanos através do teste de MOS (*Median Opinion Score*) conforme a literatura. As referências originais se usam das frases de Harvard extraídas do apêndice do: IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements de 1969[8]. Entretanto essas sentenças foram pensadas de maneira anglocêntrica e podem não ser adequadas ao português. Seguindo a literatura

[8]http://www.cs.columbia.edu/ hgs/audio/harvard.html

de fonoaudiologia recomenda-se usar as frases do CAPE-V para detecção de problemas de fala [9] [10] [11][12] totalizando assim 6 frases. O intuito da filtragem dos modelos com o AVQI reduzir a exaustão ou desistência dos ouvintes para consistência dos resultados. Através do comparativo entre os modelos filtrados podemos também validar ambas amétricas como métrica de naturalidade. O CAPE-V é composto de 6 frases que com o auxílio do software Praat[13] é possível obter o AVQI da fala, conforme podemos ver na Imagem 1. Com essas 6 sentenças serão sintetizados 378 exemplos de fala para serem avaliados pelo primeiro filtro, o AVQI. O AVQI gera um valor para cada uma das sentenças sintetizadas. Para determinar o AVQI de um certo modelo faremos uma análise estatística descritiva simples para determinar entre optar pela comparação das médias ou das medianas de AVQIs entre modelos.

Esse filtro permite reduzir o número de modelos a serem avaliados por ouvintes humanos. Fixando-se a fração de dados a ser estudada teremos 5 modelos e fixando-se o número de épocas teremos 4 modelos totalizando assim 9 modelos para cada conjunto de dados. Utilizando as mesmas frases sintetizadas no AVQI teremos um total de 54 sentenças para cada conjunto de dados ou 162 sentenças ao todo. O intuito de filtrar as senteças que serão avaliadas por ouvintes humanos é reduzir a fatiga e possíveis desistências do questionário devido a um tempo muito extenso para completá-lo.

3 Cronograma

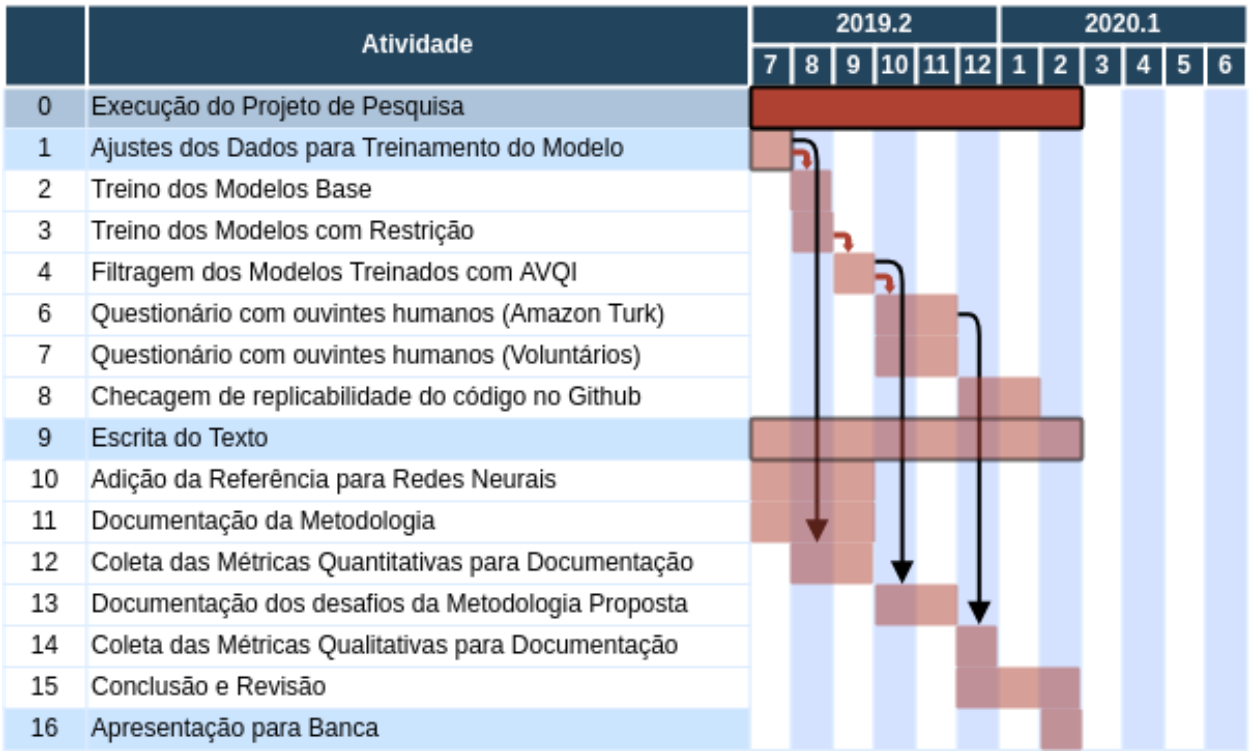

Figure 2: Proposta de Cronograma para Execução do Projeto de Pesquisa

Das atividades mapeadas várias podem ser feitas simultaneamente, como os treinos com diferentes bases de dados e diferentes restrições e os questionários com os dois tipos de público. A escrita do texto é facilmente a atividade mais desafiadora para mim atualmente. Preciso documentar todo meu processo detalhadamente

[9] https://www.pucsp.br/laborvox/dicas_pesquisa/downloads/CAPEV.pdf
[10] https://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf
[11] http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-17822019000100303&lng=en&nrm=iso&tlng=en
[12] http://www.scielo.br/pdf/codas/v31n1/2317-1782-codas-31-1-e20180082.pdf
[13] https://github.com/praat/praat

garantindo que as referências corretas sejam apontadas e que seja um material compreensível, mesmo para alguém sem domínio total do assunto.

References

[ACC+17] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Greg Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta e Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech. *CoRR*, abs/1702.07825, 2017.

[ACP+18] Sercan Ömer Arik, Jitong Chen, Kainan Peng, Wei Ping e Yanqi Zhou. Neural voice cloning with a few samples. *CoRR*, abs/1802.06006, 2018.

[ADG+17] Sercan Ömer Arik, Gregory F. Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman e Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *CoRR*, abs/1705.08947, 2017.

[CDG+18] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew P. Aylett, João P. Cabral, Cosmin Munteanu e Benjamin R. Cowan. The state of speech in HCI: trends, themes and challenges. *CoRR*, abs/1810.06828, 2018.

[fla94] *Voice Communication Between Humans and Machines*. National Academies Press, jan 1994.

[JZW+18] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno e Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *CoRR*, abs/1806.04558, 2018.

[NSKA08] Nelson Neto, Patrick Silva, Aldebaro Klautau e Andre Adami. Spoltech and ogi-22 baseline systems for speech recognition in brazilian portuguese. *Lecture Notes in Computer Science Computational Processing of the Portuguese Language*, página 256–259, 2008.

[PPG+17] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman e John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *CoRR*, abs/1710.07654, 2017.

[SPW+17] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis e Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017.

[TT12] Manjul Tiwari e Maneesha Tiwari. Voice - how humans communicate? *Journal of Natural Science, Biology and Medicine*, 3(1):3, 2012.

[vdODZ+16] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior e Koray Kavukcuoglu. Wavenet: A generative model for raw audio. Em *Arxiv*, 2016.

[WSRS+17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark e Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. *CoRR*, 2017.