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 a partir de gravações de um ou mais falantes na língua inglesa[PPG+17, ACP+18, TWPN17, DSSO18, vdODZ+16, ACC+17, ADG+17, PPG+17].

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? (MOS Score)
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?

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.

3 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]. 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. Esses dados possuem formato similar ao de treinamento empenhado no modelo implementado e com pequenos ajustes pode ser facilmente adaptado para adequar-se a necessidade 1. Uma primeira métrica 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. Para responder ao problema 3 podemos nos utilizar de *transfer learning* 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.

[1] `https://cloud.google.com/text-to-speech/`

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

[3] https://github.com/carpedm20/multi-speaker-tacotron-tensorflow

Podemos nos utilizar de *transfer learning* com um modelo pré-treinado no inglês para obter uma convergência mais rápida dos modelos em português com naturalidade similar percebida pelos ouvintes comparativamente a um modelo treinado exclusivamente com o português?

Conseguimos utilizar uma fração dos dados no treinamento com o modelo pré-treinado para obter um menor tempo de treinamento e uma síntese com naturalidade consistente com o mesmo modelo treinado com todos os dados em um período maior de tempo?

4 Datasets

Escolhemos para esse trabalho 3 conjuntos de dados sendo 2 prontamente acessíveis pelo grupo FalaBrasil e o outro extraído diretamente de vídeos do YouTube. Todos os arquivos tem 22.050Hz com 16 bits.

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

5 Métodos

Para responder aos questionamentos acima propomos os seguintes passos:

Tomar a implementação[4] do modelo Tacotron [SPW+17] reconhecido como estado da arte. O primeiro passo para estabelecer um paralelo comparativo treinando o modelo completamente a partir de dados em português usando os mesmos parâmetros que a publicação original (arquitetura de rede, otimizador, taxa de aprendizagem, entre outros) e encerrando o treino no mesmo número de épocas que o modelo pré-treinado.

Tendo esse modelo como comparativo podemos então treinaremos modelos com restrições de tempo e de dados. Para as restrições de tempo determinamos fixar uma quantidade de épocas de treino em 1k, 5k, 33k e 50k épocas. Como nosso intuito é comparar a performance 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. Para os parâmetros de dimensão dos dados propomos dividir os dados de entrada em porcentagens de suas totalidades, mais especificamente 5, 15, 33 e 50% dos dados, a serem escolhidos aleatoriamente dentro de cada conjunto de dados selecionado. Com esses parâmetro queremos gerar um total de 16 modelos para cada conjunto de

[4] https://github.com/carpedm20/multi-speaker-tacotron-tensorflow

dados. Com essas 16 hiperparametrizações fixadas para cada um dos 3 datasets descritos na Tabela 1 temos um total de 48 modelos.

Quantitativamente os modelos são analisados a partir da variação do valor do erro no tempo e o tempo proporcional para atingir aquela quantidade de épocas desejada. Qualitativamente desejamos estabelecer um teste com ouvintes humanos através de um teste de MOS (*Median Opinion Score*) conforme a literatura. Como a quantidade de modelos a ser testada é grande utilizamos uma métrica proveniente da fonoaudiologia para filtrar os modelos mais naturais de modo a oferecer apenas os melhores modelos para os ouvintes finais. O intuito é estabelecer um teste que não cause exaustão ou desistência dos ouvintes de modo a obter resultados consistentes. Além disso podemos estabelecer um indicativo da utilização dessa métrica de naturalidade estabelecida pela fonoaudiologia como futuro parâmetro quantitativo para estudos similares.

Na literatura de fonoaudiologia recomenda-se utilizar as frases estabelecidas no CAPE-V para avaliar a naturalidade da fala. O CAPE-V é composto de 6 frases a serem pronunciadas. Com essas frases e o auxílio do software Pratt é possível obter métricas quanto ao formato da onda, amplitude

Com essas proporções para os três conjuntos de dados listados pretendemos concluir quanto a variabilidade mínima necessária de um mesmo falante e também como o comprimento do conjunto em média afeta a naturalidade final sintetizada.

As métricas quantitativas do treino serão:

- O tempo de treino de cada modelo sob as mesmas condições
- O valor da função de perda na última iteração e no menor valor alcançado durante o treino avaliado simultaneamente no conjunto de treino no conjunto de validação

5.1 Proposta de Sentenças 1

Finalizados os treinos dos modelos e podendo-se finalmente estabelecer a síntese selecionamos 3 sentenças completas de cada conjunto de dados presentes nos arquivos originais para permitir um conjunto de controle. Selecionamos ainda 3 trava línguas pela dificuldade usual que eles representariam para um falante humano e pela presença de termos inexistentes nos conjuntos de treino[5]. Conforme o teste nas publicações originais o ouvinte será orientado a ouvir o áudio pelo menos 2x antes de determinar em qual dos 5 níveis de naturalidade ele se encaixa.

Totalizando assim 12 sentenças a serem sintetizadas por 5 modelos, totalizando 60 sentenças sintetizadas e 9 sentenças originais. A partir desses dados a proposta é estabelecer um comparativo da evolução do MOS médio à medida que adicionamos frações dos datasets originais. A partir de um teste de significância estatística desejo determinar o nível de significância com a qual pode-se afirmar que houve alguma melhoria nas médias entre os modelos.

Com o teste podemos afirmar então se as médias do MOS score de fato crescem e se crescem qual é esse nível de significância. A partir dessa informação também somos capazes de apontar os limites de variação mínimos do conjunto de dados para a evolução com significância da média perceptual dos ouvintes. Podemos por fim apontar também na dimensão do tempo o impacto do corte no conjunto de dados e seu respectivo impacto no resultado do MOS de naturalidade do modelo final.

5.2 Proposta de Sentenças 2

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[6]. Entretanto essas sentenças foram pensadas de maneira anglocêntrica e podem não ser adequadas ao português. Seguindo a literatura de fonoaudiologia recomenda-se usar as frases do CAPE-V para detecção de problemas de fala [7] [8] [9] totalizando assim 6 frases. Podemos então avaliar a síntese através da análise AVQI 03.01 proposta na literatura de fonoaudiologia para detecção de desvios e vícios de fala[10]. Esse método se utiliza

[5] http://www.fonoaudiologia.med.br/voz/7-teste-sua-diccao
[6] http://www.cs.columbia.edu/ hgs/audio/harvard.html
[7] https://www.pucsp.br/laborvox/dicas_pesquisa/downloads/CAPEV.pdf
[8] https://www.asha.org/uploadedFiles/members/divs/D3CAPEVprocedures.pdf
[9] http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-17822019000100303&lng=en&nrm=iso&tlng=en
[10] https://journals.sagepub.com/doi/abs/10.1177/0003489416636131?journalCode=aora

do Software Praat que é vastamente utilizado no campo de fala para detecção de parâmetros técnicos, como os necessários para equação conforme podemos ver abaixo

$$AVQI_{03.01} = (4.152 - 0.177 * CPPs - (0.006 * HNR) - (0.037 * Shim) + (0.941 * ShdB) + 0.01 * Slope + (0.093 * Tilt)) * 2.8902 \quad (1)$$

Para consolidar então os resultados obtidos do AVQI podemos executar uma pesquisa buscando obter o MOS dessas 30 sínteses (6 frases * 5 modelos). Com os resultados desse questionário podemos validar a correlação do MOS com o AVQI e correlacionar o volume de dados disponível a nota do AVQI e ao valor atribuído de média de MOS de Naturalidade com seus respectivos níveis de significância.

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.

[DSSO18] Tijana Delić, Sinisa Suzic, Milan Sečujski e Vladimir Ostojic. Deep neural network speech synthesis based on adaptation to amateur speech data. *IcETRAN 2018*, 06 2018.

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

[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.

[TWPN17] Yaniv Taigman, Lior Wolf, Adam Polyak e Eliya Nachmani. Voice synthesis for in-the-wild speakers via a phonological loop. *CoRR*, abs/1707.06588, 2017.

[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.