XXX Congresso de Iniciação Científica UNICAMP

ANÁLISE FONÉTICA DE NATURALIDADE DE DIFERENTES TTS EM PORTUGUÊS DO BRASIL

Palavras-Chave: prosódia, fala sintética, text-to-speech, naturalidade da fala

Autores:
RICARDO BUOSO NETO, IEL/UNICAMP
Prof. Dr. PLÍNIO ALMEIDA BARBOSA (orientador), IEL/UNICAMP

INTRODUÇÃO:

Cada vez mais, vozes sintetizadas por computador estão fazendo parte da vida das pessoas. Se encontram tanto em sistemas automatizados de *call centers* e *telemarketing* quanto nos dispositivos pessoais, como é o caso dos assistentes encontrados em *smartphones* e TVs. Por trás dessas vozes que muitas pessoas estão acostumadas e são capazes de reconhecer, existem sistemas robustos de TTS (*Text-to-Speech*) que procuram cada vez mais aprimorar a naturalidade da fala.

No entanto, como demonstrado por Maués (2019) e Leite (2014), entre outros, esses sistemas ainda chamam a atenção do ouvinte por sua falta de naturalidade e caráter de fala dita robótica, sem variação prosódica, isto é, relacionada ao ritmo e entoação da fala (ao tom, chama-se F0, e sua curva prediz nível e estilo de variação prosódica). Maués (2019) aponta a falta de variação prosódica como uma das principais causas da insatisfação em sistemas de atendimento.

Oord et al (2016) mostra que naturalidade é um dos campos com maior foco de pesquisa na área de vozes sintéticas, desencadeando pesquisas como o modelo WaveNet, que mostrou que seu *machine learning* gera resultados mais robustos que os métodos paramétricos e concatenativos.

Com base nos resultados apontados acima sobre a importância da prosódia na naturalidade da fala, a presente pesquisa (financiada pelo CNPq, processo 135069/2021-1) esperou encontrar diferenças prosódicas entre os trechos de fala considerados mais e menos naturais. A principal hipótese é a de que a curva de frequência fundamental (F0) seja a variável com o maior impacto na sensação de naturalidade, portanto, é esperado que os trechos com as notas mais baixas possuam mais divergência de contorno de F0 quando comparado aos mesmos trechos produzidos por locutores humanos (logo, mais naturais) e vice-versa.

METODOLOGIA:

Foi conduzido um experimento de percepção com 24 participantes voluntários para julgar a naturalidade de diversos trechos de áudio. Os sistemas TTS escolhidos foram:

• "WaveNet", novo sistema da Google treinado por *machine learning*. Escolhida por ser uma das vozes mais reconhecíveis, principalmente entre usuários do sistema operacional Android.

• "Felipe" e "Fernanda" são duas vozes da empresa Nuance, especializada em reconhecimento de fala e inteligência artificial.

Também foram escolhidos locutores humanos (um masculino e um feminino). Ambos leram as mesmas frases que foram usadas para gerar os áudios sintéticos e suas curvas de F0 serviram de base para: a) medir a distância entre F0 humano (natural) e o sintético; b) modificar a curva de F0 dos trechos de áudio sintéticos para aproximar seu contorno de F0 dos trechos naturais, para testar se esses trechos modificados são percebidos como mais naturais.

Foram retirados trechos de diversos jornais e adaptados conforme necessário para se conformarem às categorias listadas abaixo**.** Em alguns casos, as frases foram inventadas.

a) Tamanho: curto, longo;

b) Complexidade: sem subordinação, com subordinação;

c) Contexto: assertivo, interrogativo.

O programa Praat foi utilizado como ferramenta para modificar o contorno de F0 das frases sintéticas, tentando aproximá-las das falas humanas gravadas. Para tal, se fez uso do recurso "Manipulation", e os contornos foram modificados manualmente, uma vez que nenhum *script* fez a transferência de forma satisfatória.

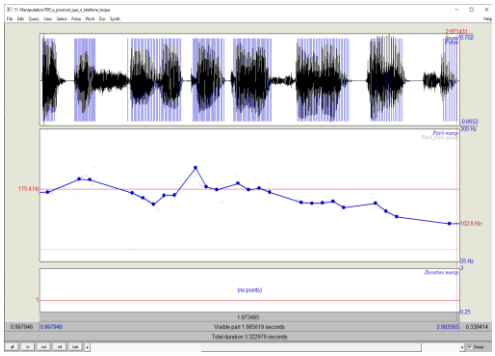

Figura 1 - A linha azul representa o contorno de F0 a ser modificado

O teste de percepção foi feito através do Google Forms. Todos os trechos de áudio foram convertidos em arquivos de vídeo e hospedados no YouTube, para serem exibidos diretamente na página do teste. Para os participantes, foram mostradas apenas as notas qualitativas em texto. Para fins de análise, foi utilizado o valor numérico:

• Muito artificial (-1)

• Pouco artificial (-0,5)

• Razoável (0)

• Relativamente natural (0,5)

• Natural (1)

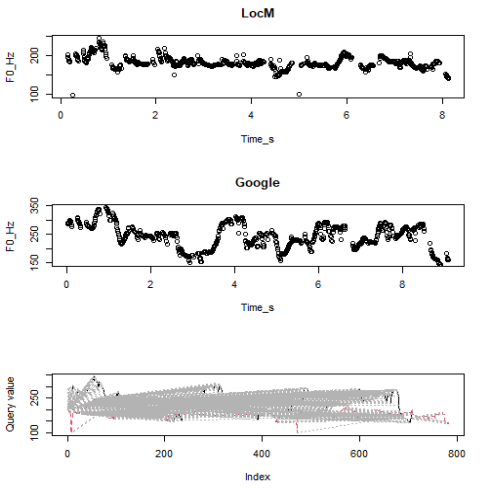

Figura 2 - Curvas de F0 da locutora "LocM" e TTS Google. O terceiro gráfico corresponde à distancia DTW (nesse caso: 18,65)

Para quantificar a diferença entre as curvas de F0 foi empregado o algoritmo *Dynamic Time Warping (*DTW), conhecido por ser eficiente e já ter sido previamente empregado

em estudos de reconhecimento de fala (Senin, 2008). O DTW permite encontrar padrões e formas similares independentemente da distorção temporal. A normalização temporal é necessária, uma vez que a taxa de elocução das gravações varia e os picos e vales de F0 nunca estarão alinhados entre dois falantes diferentes. Com o Praat, cada áudio foi transformado em um arquivo txt em formato de tabela contendo o valor de F0 em função do tempo, para, assim, ser lido pelo DTW no programa de estatística R.

RESULTADOS:

O primeiro resultado observado alinhou-se ao que era esperado: notas baixas para as vozes artificiais e altas para os locutores. Ambos os locutores obtiveram médias parecidas de 0,7. "Fernanda" (Nuance) se saiu melhor que as outras duas vozes sintéticas e Google WaveNet teve as menores notas.

	Nuance Fernanda	Nuance Felipe	Google Wave Net	Loc M	Loc N
Nota média	-0,32	-0,51	-0,6	0,71	0,77
Desvio Padrão	0,255	0,198	0,1	0,15	0,11

Tabela 1 - Notas médias obtidas no experimento online

Incluindo os valores de DTW obtidos anteriormente, os dados foram divididos em grupos de acordo com as variáveis usadas na elaboração inicial das frases, bem como grupos para cada sistema TTS. Regressões lineares simples foram aplicadas, de forma que se observou a relação entre valores de DTW (distância normalizada entre a curva de F0 artificial e a equivalente da mesma frase) do locutor humano, e a nota média obtida no experimento para cada grupo.

Dois conjuntos de dados se destacaram por um R^2 elevado: 14% e 56%. Ambos também compartilham as características "Subordinadas" e "Assertivas", e diferem-se quanto ao tamanho de frase. Além disso, o conjunto de "curtas, não subordinadas e assertivas" também possui $R^2 = 8,2\%$, maior que os demais. Nota-se que os três conjuntos com maior valor de R^2 são "assertivos". Dito isso, seu valor está próximo a de outros dois conjuntos, e distante dos dois maiores valores. Por conta disso, não foi levado adiante nas próximas análises. Os dados de 14% e 56% foram unificados. O novo valor de R^2 é de 32% (Figura 3), ainda significativamente maior que os outros, os quais ficam entre 0 e 9%.

Conjunto de dados	R^2
Todos	8,8%
Curtas, Não subord., Assert.	**8,2%**
Longas, Subord., Assert.	6%
Curtas, Subord., Assert.	**14%**
Longas, Subord., Assert.	**56%**
Curtas, Não subord., Interrog.	0,1%
Longas, Não subord., Interrog.	0,7%
Curtas, Subord., Interrog.	0,4%
Longas, Subord., Interrog.	7,5%
Felipe	2%
Fernanda	0,8%
Google	3,7%

Tabela 2 - R^2 da regressão linear de cada grupo. Em negrito estão destacados os maiores valores

Neste trabalho, é assumido o valor de α padrão 0,05. Dessa forma, o teste Shapiro-Wilk revela o valor de $p = 0,4 > 0,05 = \alpha$. Esse valor indica uma alta probabilidade de que a distribuição de resíduos seja normal. Por conseguinte, o teste de homocedasticidade indica que $p = 0,18 > 0,05 = \alpha$.

Os resíduos indicaram dois dados que fogem das tendências observadas, destacando-se em todos os gráficos pelas menores notas e baixo valores de DTW. Felizmente, ambos correspondem à mesma frase, em suas variações "padrão" e "F0 modificado". Assim, pode-se justificar sua remoção do conjunto de dados como um problema específico à frase, preservado em sua versão modificada.

Após remover os dois dados do conjunto, o modelo de regressão linear indica R^2 ajustado de 66% e coeficiente de inclinação significativo ($p = 2{,}72 \times 10^{-4}$), o que explica satisfatoriamente a relação entre os fatores (32% anteriormente).

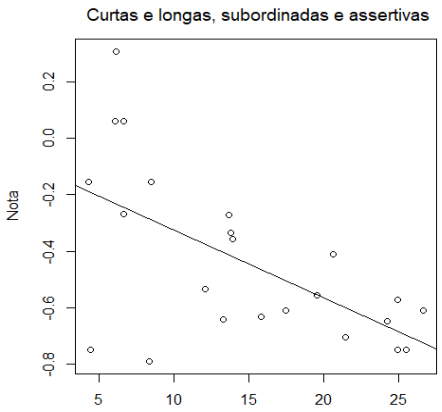

Figura 3 - Regressão linear dos grupos unificados com $R^2 = 32\%$

Os novos resíduos mostram outro dado com bastante influência no conjunto. Por precaução, foram realizados novos testes com o dado removido, resultando em $R^2 = 67\%$. Uma vez que não se destacou no ponto de vista linguístico e parece seguir a previsão do modelo, essa versão foi descartada e a análise final se torna a anterior.

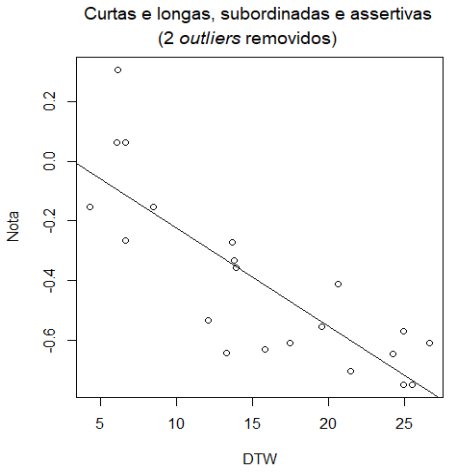

Figura 4 - Regressão com dois dados removidos e $R^2 = 66\%$, representando maior relação entre distância de F0 (DTW) e notas médias

Foi demonstrado que os trechos com menor valor de DTW (portanto, contorno melódico mais próximo do locutor humano de referência) possuem uma relação linear com uma maior nota de naturalidade, mas considerações são necessárias. Em primeiro lugar, a amostra utilizada nesta pesquisa é pequena: após divisão, cada subcategoria possui apenas 12 trechos em cada. Além disso, apenas dois grupos apresentaram relações lineares maiores que 10% sob uma primeira análise, dentro das 12 possíveis subdivisões.

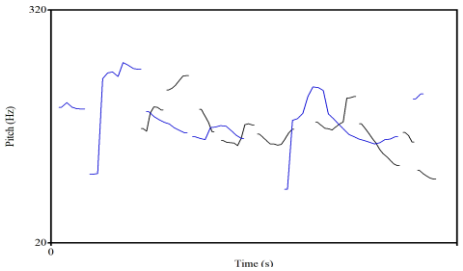

Figura 5 - o F0 da locutora sobe no final da interrogativa, enquanto "Fernanda" (Nuance) abaixa, apontando uma fraqueza do sistema

Explicar a má performance das frases interrogativas de modo geral é relativamente simples. O julgamento subjetivo mostrou que todos os sistemas tiveram grande dificuldade em capturar a subida de F0 em frases interrogativas características do português (Madureira, 1994). Nessa situação, a manipulação com o grau necessário para aproximar os dois valores de F0 causaria distorções acústicas perceptíveis no nível segmental. Poderiam ser obtidas melhoras satisfatórias somente ressintetizando os trechos com controle total de F0, o que não foi possível nos sistemas utilizados.

Além disso, não está claro por qual motivo as frases curtas + longas, subordinadas e assertivas obtiveram uma relação linear tão maior (14% e 56%, respectivamente) que as frases não subordinadas que compartilham as características curtas + longas e assertivas (8,2% e 6%).

CONCLUSÕES:

Foi encontrada uma relação considerada significativa entre nota de naturalidade e valores de DTW nos dados explorados a fundo na seção anterior. Esse resultado era esperado e está de acordo (admitindo-se também a menor complexidade metodológica) com os achados iniciais de Oord (2006), que revelou que dois modelos de seu TTS WaveNet, tiveram notas de naturalidade maior quando o modelo foi treinado também em contornos de F0 quando comparado ao modelo sem esse treino, apesar de qualidade segmental idêntica.

Quanto ao motivo de somente esse conjunto de dados apresentar uma relação desse nível, nada se pode afirmar, já que diversos fatores podem alterar esses resultados, tal como quais frases foram utilizadas em cada conjunto (Hoffman et al, 1999). Em todo caso, explicar 66% dos resultados é uma forte indicação de que os sistemas TTS utilizados no experimento podem, em condições específicas, apresentar resultados mais naturais quando comparados às frases julgadas menos satisfatórias, apenas com uma melhor curva de F0. Um aprofundamento se faz necessário para entender qual ou quais fatores estão em jogo, causando o efeito observado que difere entre grupos de frases. Por exemplo: nível segmental, conteúdo semântico ou algum outro fator de construção sintática, já que sistemas TTS são geralmente usados (e otimizados) para passagens conectadas, como em notícias (Hoffman, 1999).

BIBLIOGRAFIA

Barbosa, P. A; Madureira, S. **Manual de Fonética Acústica Experimental: Aplicações a Dados do Português**. Cortez Editora, 2015.

Hoffmann, R.; Hirschfeld, D.; Jokisch, O.; Kordon, U.; Mixdorff, H.; Mehnert, D.; **Evaluation of a Multilingual TTS System With Respect to the Prosodic Quality.** Technische Universität Dresden, 1999.

Leite, H. M. d. A. **Proposta de Metodologia de Avaliação de Voz Sintética Com Ênfase no Ambiente Educacional.** Unicamp, Campinas, 2014.

Maués, M. P. **Um Olhar Sobre os Assistentes Virtuais Personificados e a Voz Como Interface**. PUC Rio, Rio de Janeiro, 2019.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K.; **WaveNet: A Generative Model for Raw Audio**. Google, London, 2016.