EyetrackingMOS: Proposta de um método de avaliação online para modelos de síntese de fala

Gustavo E. Araújo[1], Julio C. Galdino[1], Rodrigo de F. Lima[1], Leonardo Ishida[1], Gustavo W. Lopes[1], Miguel Oliveira Jr.[2], Arnaldo Candido Jr.[3], Sandra M. Aluísio[1], Moacir A. Ponti[1]

[1] Universidade de São Paulo (USP)

[2]Universidade Federal de Alagoas (UFAL)

[3]Universidade Estadual de São Paulo (UNESP)

{gustavo_evangelista,juliogaldino,guico21,leoishida,gustavowlopes}@usp.br

miguel@fale.ufal.br, arnaldo.candido@unesp, {moacir, sandra}@icmc.usp.br

Abstract. *Evaluating Text-To-Speech (TTS) systems is challenging, as the increasing quality of synthesis makes it difficult to discriminate models' ability to reproduce prosodic attributes, especially for Brazilian Portuguese. Offline evaluation metrics do not capture our genuine reactions to audio stimuli. Therefore, we propose an online evaluation method using eye-tracking. Our experiments with 76 annotators show a reasonable correlation between EyetrackingMOS and MOS, as well as a reduction in the total evaluation time. We believe this metric provides precise and potentialy fast information to complement existing evaluation methods.*

Resumo. *Avaliar sistemas Text-To-Speech (TTS) é um desafio, uma vez que a qualidade crescente da síntese impõe obstáculos em discriminar a capacidade de modelos em reproduzir atributos prosódicos, especialmente para o português brasileiro. Métricas de avaliação offline não medem a reação genuína de avaliadores aos estímulos de áudios. Propõe-se, portanto, um método de avaliação online com rastreamento de globo ocular. Os experimentos com 76 anotadores apontam que há uma correlação razoável entre EyetrackingMOS e MOS, assim como uma redução em sua duração total. Desta forma, acredita-se que esta métrica forneça uma informação precisa e potencialmente rápida para complementar os métodos de avaliação.*

Index Terms: Speech Synthesis Models Evaluation, Portuguese language, spontaneous speech, eyetracking

1. Introdução

Sistemas de texto-para-fala, do inglês *Text-To-Speech* (TTS) buscam vocalizar um texto escrito em níveis próximos a naturalidade de fala humana [Caseli and Nunes 2024]. Os avanços em Aprendizado Profundo impulsionaram o desenvolvimento de tais sistemas. Posteriormente, a utilização de modelos gerativos baseados em fluxo, como os propostos por [Kingma et al. 2016] e [Hoogeboom et al. 2019], tem permitido maior flexibilidade

na manipulação de características prosódicas[1] da fala sintética. Os resultados de modelos do estado da arte já reproduzem a identidade dos locutores com bastante naturalidade em condições mais amplas de dados [Casanova et al. 2022, Tan et al. 2022].

Entretanto, modelos de síntese ainda encontram obstáculos na reprodução de aspectos específicos da expressividade individual de falantes. Estes aspectos podem ser medidos através da entoação, duração e ritmo da fala [Ju et al. 2024], que são de natureza prosódica, o que se agrava em cenários de síntese *zero-shot* [Casanova et al. 2022, Ju et al. 2024]. Neste contexto, sistemas contemporâneos de TTS investigam outras capacidades além da reprodução da identidade de um locutor com naturalidade nos resultados, dentre elas o interesse em manter a naturalidade ao gerar fala nas variantes internacionais de uma língua (**accent-robust**), como o *Synthesizing Multi-Accent Speech By Weight Factorization* (SYNTACC) [Nguyen et al. 2023]. A possibilidade de síntese de fala com sensibilidade de sotaques internacionais, também levanta hipóteses de aplicações para variantes linguísticas regionais de uma dada língua que conta com menos recursos, afim de avaliar se a qualidade se preserva. O português brasileiro é uma língua que contempla uma grande quantidade de variantes, dadas as dimensões continentais do Brasil, e devido a fatores históricos, sociais e culturais [Mota et al. 2023].

Para avaliar a qualidade da fala sintetizada nesses sistemas, são utilizadas diversas métricas. As **métricas subjetivas** como: *Mean Opinion Score* (MOS) [ITU - T 1996], *Crowd MOS* [Ribeiro et al. 2011], *Similarity MOS* (SMOS) [Jia et al. 2019] e *Comparative MOS* (CMOS), por um lado, dependem da opinião e percepção de um grupo de ouvintes humanos. Apesar de importante, este perfil de métricas pode oferecer risco para análise de sotaques a depender da correspondência entre o contexto regional/cultural dos avaliadores e os áudios sintéticos, uma vez que a avaliação será influenciada por seus contextos culturais, linguísticos e experiências individuais. Por outro lado, as **métricas objetivas** como: *Speaker Encoder Cosine Similarity* (SECS) [Casanova et al. 2021], *Prosody Similarity with Prompt*, *Prosody Similarity with Ground Truth* e *Word Error Rate* (WER) [Shen et al. 2023] podem não capturar completamente a percepção humana da qualidade do áudio sobre o desempenho na qualidade de expressividade individual e representatividade de variantes linguísticas e, por isso, complementam a análise subjetiva. A ausência de uma métrica padrão e amplamente aceita dificulta a identificação de tendências e avanços consistentes no campo do TTS, além de dificultar o entendimento de quais modelos são mais adequados para determinados cenários ou requisitos específicos (cf. [Le Maguer et al. 2024]). Ambos os perfis têm sensibilidades a aspectos diferentes e limitações que devem ser avaliadas [Cooper et al. 2024].

Ambas as métricas também podem ser observadas quanto a sua resposta aos estímulos de áudios fornecidos durante a avaliação. Em métodos de **avaliação *offline*** (MOS, CrowdMOS, SMOS e CMOS), o indivíduo pontua apenas após ouvir todo o estímulo, enquanto métodos de **avaliação *online*** permitem que se registre suas impressões à medida que o estímulo é recebido, tendo como objetivo capturar reações genuínas e momentâneas. A avaliação de estímulos de áudio utilizando rastreamento ocular já é amplamente empregada em contextos linguísticos, como na

[1]A prosódia estuda as funções dos suprassegmentos, que são essenciais para a melodia da fala (tom, entoação, tessitura), para a dinâmica da fala (duração, pausa etc.) e para qualidade da voz (volume, registro etc.) [Cagliari 1992].

análise de processamento de linguagem, compreensão auditiva, e percepção fonética [ALMEIDA et al. 2021]. No entanto, sua aplicação na avaliação de sistemas de síntese de fala ainda é pouco explorada. Buscamos preencher essa lacuna, propondo um novo método, EyetrackingMOS, que utiliza o rastreamento ocular para avaliação de qualidade dos áudios forma mais natural, sem que o participante atribua uma nota de forma direta.

As principais contribuições feitas nesse trabalho são sumarizadas como se segue:

1. Proposta de um novo método de avaliação de sistemas de síntese de fala que integra o rastreamento ocular, chamado de EyetrackingMOS;
2. Comparação entre o EyetrackingMOS e uma adaptação do MOS tradicional, destacando suas respectivas vantagens e limitações;
3. Apresentação dos experimentos, detalhes de configuração do modelo e interfaces em um repositório[2], facilitando a replicabilidade em diferentes cenários e promovendo avanços na pesquisa sobre síntese de fala.

2. Revisão sobre métricas subjetivas para análise de sistemas de TTS

Na década de 1990, a *International Telecommunication Union* (ITU) padronizou diversos tipos de testes de audição que eram frequentemente usados na telefonia [ITU - T 1996]. A pontuação baseada em opinião pode ser definida como o valor em uma escala predefinida que um sujeito atribui à sua opinião sobre o desempenho de um sistema [ITU - R 2017, Loizou 2011]. A pontuação média de opinião, do inglês *Mean Opinion Score* (MOS) é um tipo de *Absolute Category Rating* (ACR) [Ribeiro et al. 2011]. A MOS emergiu como o descritor mais popular sobre a percepção da qualidade de mídia. Para o cálculo da MOS, humanos avaliam os áudios sintetizados e naturais e atribuem uma nota de 1 a 5, no qual o valor final corresponde à média das notas de todos os avaliadores. A tabela traduzida com a equação correspondente pode ser vista no repositório[2].

Diversas variações da MOS foram desenvolvidas para atender a diferentes necessidades de avaliação. A *Crowd Mean Opinion Score* (crowdMOS) propõe uma adaptação ao ambiente tradicional de testes MOS, ao utilizar trabalhadores de uma multidão (do inglês, *crowd*) pela internet para realizar avaliações em ambientes não controlados, o que permite maior diversidade de ouvintes a um custo reduzido, embora com desafios em termos de controle de qualidade [Ribeiro et al. 2011]. A *Similarity Mean Opinion Score* (SMOS)[3], por sua vez, foca na avaliação da semelhança entre áudios sintetizados e de referência, sendo útil para medir quão próximo um áudio gerado está de uma voz original em termos de características acústicas e vocais [Ren et al. 2021]. Já a *Comparative Mean Opinion Score* (CMOS) avalia a qualidade relativa entre duas versões de áudio sintetizado, pedindo aos avaliadores que comparem diretamente os áudios e apontem qual deles possui melhor qualidade, utilizando uma escala de -3 a +3 [Ren et al. 2022]. Cada uma dessas variantes da MOS foca em diferentes aspectos da qualidade de áudio, utilizadas de acordo com o que se deseja avaliar. Considerando que estudos têm utilizado a MOS como uma medida de naturalidade da fala em tarefas de síntese (cf. [Sellam et al. 2023], [Choi et al. 2022]), a descrição da característica observada pelo avaliador foi adaptada para a avaliação de naturalidade (veja a coluna 4 da Tabela 1).

[2]Acesso em `https://github.com/GustavoEvangelistaAraujo/EyetrackingMOS-STIL`

[3]Também abreviado por SimMOS na literatura.

3. EyetrackingMOS

O rastreamento ocular é amplamente reconhecido como uma das técnicas mais precisas para a avaliação *online* do processamento linguístico [Mitchell 2004, Kaiser 2013]. Os variados movimentos dos olhos durante o processamento de informações podem ser utilizados para inferir como essas informações são processadas, seja durante a leitura de texto (estímulo de leitura) ou ao observar uma imagem (estímulo visual). O resultado é obtido a partir da porcentagem de tempo em que o avaliador olhou para o lados direito e esquerdo, os quais mostram figuras relacionadas aos conceitos que se deseja medir. Assim, registramos a porcentagem de tempo que o participante permanece com o olhar sobre a figura que representa a fala natural. Esta medida pode ser avaliada em um intervalo de 0% a 100% e mapeada para a escala MOS como apresentado na Tabela 1. Assim como no MOS, ao final é calculada a média das notas de todos os avaliadores.

Tabela 1. Mapeamento entre pontuações do EyetrackingMOS e MOS

Tempo de fixação (%)	Avaliação MOS	Qualidade	Naturalidade
81 a 100	5	Excelente	Extremamente natural
61 a 80	4	Boa	Muito natural
41 a 60	3	Razoável	Razoavelmente natural
21 a 40	2	Pobre	Pouco natural
0 a 20	1	Ruim	Nada natural

4. Materiais e métodos

4.1. Descrição do conjunto de dados

Há uma carência de conjuntos de dados de áudio com variantes linguísticas regionais. Corpus como BRACCENT, utilizado em [Batista 2019], [Ling et al. 2018] e [Ynoguti 1999], não apresentam volume satisfatório de dados, assim como não tratam da fala espontânea. Portanto, foi escolhido para este estudo um recorte de áudios de um grande dataset do Museu da Pessoa[4], um museu virtual e colaborativo de histórias de vida, que são do tipo entrevistas biográficas, compilado pelo projeto Tarsila[5]. Detalhes do recorte preliminar do corpus (MuPe-v1) estão disponíveis no repositório[2].

4.2. Modelo de síntese de fala

O modelo SYNTACC [Nguyen et al. 2023] é uma arquitetura para síntese de fala com múltiplos sotaques baseada no YourTTS [Casanova et al. 2022]. Similarmente ao antecessor, utiliza uma arquitetura de codificação-decodificação baseada em Transformer, onde o codificador recebe a sequência de texto como entrada e gera uma representação intermediária, que é posteriormente processada pelo decodificador para gerar o espectrograma mel, uma representação em espectro da frequência ao longo do tempo que é reconstruída em áudio por um *vocoder*.

Esse modelo implementa as seguintes mudanças: na entrada, a arquitetura concatena 4 *embeddings* de idiomas treináveis em cada caractere de entrada, uma técnica de fatorização de pesos (*weight factorization*), o que permite um treinamento *multi-accent*.

[4]https://museudapessoa.org/

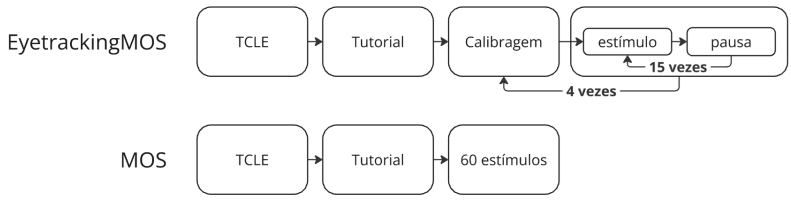
[5]https://sites.google.com/view/tarsila-c4ai/home

Esta abordagem divide os pesos do modelo em componentes compartilhados e específicos para cada variante linguística, otimizando o treinamento em cenários de poucos recursos. Isso possibilita que a síntese de fala seja adaptável para o contexto do português brasileiro, sendo possível obter um controle explícito de sotaques pelo congelamento parcial de pesos atribuídos a ele e portanto permite que a fala seja sintetizada de forma mais específica para cada variante. Detalhes da arquitetura, configurações do modelo e etapa de treinamento também foram disponibilizados no repositório[2].

5. Experimentos

A Figura 1 apresenta o fluxo de interação do usuário neste experimento. Para tanto, foi utilizada a plataforma Gorilla[6], uma plataforma paga, com o objetivo de construção e coleta de tarefas de anotação. A sequência de interfaces e o conjunto de dados dos estímulos também são apresentados no repositório[2].

Figura 1. Fluxograma da interação do usuário em cada experimento



No experimento elaborado neste trabalho, o processo se inicia com a aceitação do Termo de Consentimento Livre e Esclarecido (TCLE). Em seguida, o participante é conduzido a um tutorial, que tem o objetivo de ambientá-lo com o experimento subsequente. Para a captura do vídeo, são utilizadas as câmeras padrões dos dispositivos pessoais (apenas computador e notebook) dos usuários, caso as configurações de iluminação e qualidade de imagem não sejam suficientemente boas para permitir que o participante complete a calibragem sem erros, o participante é impedido de continuar. Em conjunto com uma calibragem recorrente, é possível inferir que a qualidade de rastreamento se mantenha desde o início até o final do experimento. O Gorilla utiliza a biblioteca Webgazer[7] para rastreamento ocular. No caso do EyetrackingMOS, após o tutorial, o usuário passa pela etapa de calibragem, que é dividida em duas partes. Primeiro, é necessário posicionar corretamente o rosto em relação à câmera. Em seguida, o participante deve fixar o olhar em uma sequência de pontos que aparecem aleatoriamente nas extremidades da área útil da tela. São apresentados 10 pontos no total, no qual os 5 primeiros pontos tornam-se uma referência de rastreamento, e os 5 seguintes são repetidos como validação dos anteriores. Caso haja uma discrepância significativa entre a referência e a validação em um dos pontos (considerado como a tolerância do teste), a calibragem é considerada falha, e o usuário precisa repetir o processo. Após uma calibragem bem-sucedida, o participante prossegue e visualiza uma tela com duas imagens vetoriais ilustrativas (um robô e uma figura humana, que trocam de posição de forma aleatória a cada estímulo) enquanto ouve o áudio. A pausa é uma tela subsequente ao final do áudio com apenas um sinal de "+" por 3 segundos, feita para poder reposicionar o olhar do usuário no meio da tela.

[6]https://gorilla.sc/

[7]https://webgazer.cs.brown.edu/

Este ciclo de estímulo e pausa é repetido 15 vezes, e então é feita uma nova calibragem para garantir a qualidade de rastreamento do globo ocular, sendo realizado quatro ciclos completos, que totalizam 60 estímulos.

Por outro lado, no experimento de MOS, após o TCLE e o tutorial, o participante é exposto a 60 estímulos de áudio. Durante o tutorial, são apresentados três exemplos de áudios sintetizados, correspondentes às pontuações 1, 3 e 5, para ajudar o participante a alinhar suas expectativas. O participante pode ouvir cada estímulo mais de uma vez antes de decidir sua pontuação, utilizando a Tabela 1 como referência para todas as 60 amostras de áudio, conforme descrito na literatura de avaliação de modelos de síntese.

A divisão das listas de áudios para avaliação foi realizada considerando dois tipos de áudios: sintetizados e naturais. Esses áudios foram organizados em duas listas: Lista A e Lista B, que foram atribuídas aos participantes de forma equilibrada. 30 áudios naturais foram colocados na Lista A, enquanto seus correspondentes sintetizados foram alocados na lista B. Da mesma forma, 30 áudios sintetizados foram incluídos na Lista A, com os seus correspondentes naturais na lista B. Quanto aos participantes, conforme [Loizou 2011], a proporção de avaliações subjetivas deve ser de 10 especialistas para 20 não especialistas. Foram escolhidos 76 anotadores dentre 28 especialistas e 48 não especialistas, distribuídos entre 4 grupos de 19 participantes. Ambos experimentos foram elaborados desta mesma forma, o que assegurou uma diversidade de perspectivas nas avaliações, permitindo uma análise comparativa abrangente entre as opiniões de especialistas e não especialistas sobre os áudios apresentados.

6. Resultados preliminares

A Figura 2 ilustra a relação entre os valores mensurados pelo EyetrackingMOS, convertidos para a escala MOS, e os valores mensurados diretamente pelo MOS. Cada ponto azul representa um par de medidas, com o eixo horizontal correspondendo aos valores do EyetrackingMOS convertidos para a escala MOS e o eixo vertical representando os valores obtidos diretamente pelo MOS. A linha verde traçada no gráfico indica a linha de tendência linear, mostrando a direção geral da correlação entre as duas variáveis. A Figura 3 ilustra a dispersão das pontuações obtidas tanto pelo MOS quanto pelo rastreamento ocular (EyetrackingMOS) para áudios reais e sintetizados. Em ambos os testes, os participantes conseguiram separar razoavelmente bem os áudios reais dos sintetizados.

Figura 2. Gráfico de dispersão entre EyetrackingMOS e MOS

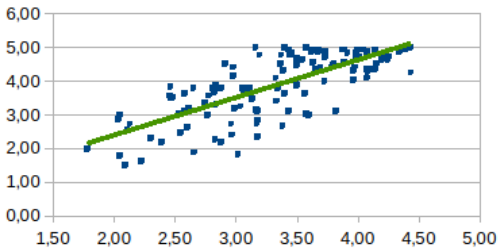
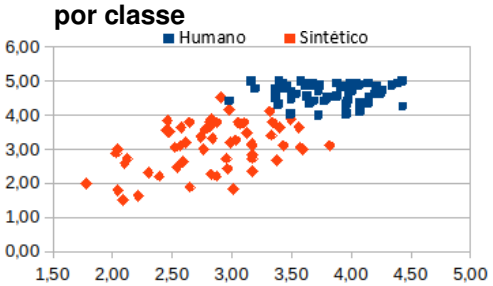


Figura 3. Gráfico de dispersão por classe



No gráfico de dispersão por MOS (Figura 4), observa-se uma distinção clara entre os áudios reais, que tendem a receber pontuações mais altas, e os sintetizados, que

se concentram nas faixas intermediárias e baixas. No entanto, no gráfico de dispersão por rastreamento ocular (Figura 5), a separação entre áudios reais e sintetizados é menos evidente. Essa maior dispersão nos resultados do rastreamento ocular é esperada, já que nesse método os estímulos são percebidos apenas uma vez, enquanto nos métodos *offline* como o MOS, o anotador pode ouvir o estímulo repetidas vezes antes de tomar sua decisão, resultando em uma separação mais clara entre os tipos de áudio. Assim, o rastreamento ocular oferece uma avaliação mais detalhada, capturando variações mais sutis na percepção da qualidade dos áudios.

Figura 4. Gráfico de dispersão por MOS

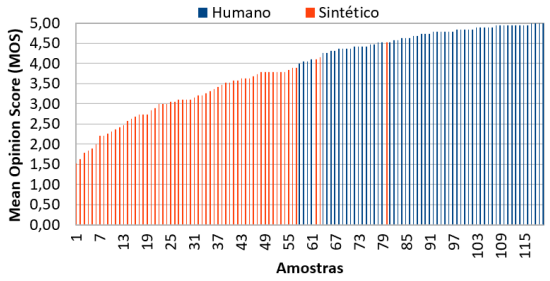
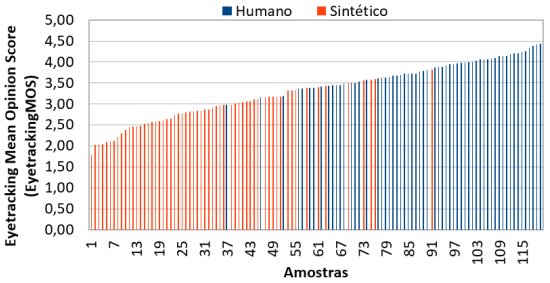


Figura 5. Gráfico de dispersão por rastreamento ocular



Os resultados apresentados na Tabela 2 indicam uma correlação razoável entre o EyetrackingMOS e o MOS, com uma métrica R^2 de 56%, sugerindo que o EyetrackingMOS explica 56% da variância observada no MOS. O desvio padrão do erro entre as duas métricas é de 0,72 unidades, mostrando que, em geral, elas tendem a ser próximas, com uma diferença média de menos de uma unidade. Além disso, o MOS tende a classificar um número maior de áudios com a nota máxima ou valores próximos, enquanto o EyetrackingMOS oferece uma análise mais detalhada, por sua escala ser de 0 a 100, o que é observado em áudios de alta qualidade. Essa dispersão indica que, embora exista uma correlação razoável entre as duas métricas, conforme evidenciado pela inclinação positiva da linha de tendência, as medidas não são perfeitamente alinhadas, refletindo diferenças na maneira como cada método capta e avalia a qualidade dos áudios.

Tabela 2. Medidas de performance estatística

Medida	Valor	Interpretação geral
Pearson	0.744	Correlação moderada
Mean Squared Error (MSE)	0.710	Erro médio baixo
Rooted Mean Squared Error (RMSE)	0.844	Erro médio baixo
R^2	0.553	Explica 55% da variância
Spearman	0.714	Correlação moderada

Também foi realizada uma análise da concordância entre os avaliadores dentro de seus respectivos grupos, utilizando o coeficiente de Kendall's W para avaliar a consistência das respostas (Tabela 3). Em resumo, o grupo EyetrackingMOS apresentou maior consistência nas avaliações, com alta concordância na maioria dos estímulos, enquanto o grupo MOS demonstrou uma maior variabilidade, com concordância que variou de alta até nenhuma, indicando possíveis desafios na avaliação uniforme dos

estímulos por este grupo. Com relação ao tempo, EyetrackingMOS e MOS tomaram em média 12:07min e 12:30min dos participantes, respectivamente. As medianas foram de 11:38min e 10:41min, respectivamente. Nota-se que o teste MOS tende a ser em torno de 1 minuto mais rápido que o EyetrackingMOS que pode ser justificada pelo tempo das 4 calibrações do rastreamento ocular.

Tabela 3. Medidas de concordância para cada grupo de experimentos

Grupo	Intervalo de Kendall's W	Interpretação geral
EyetrackingMOS	0.6719 a 0.9579	Alta concordância geral, algumas variações
MOS	0.0000 a 0.9474	Grande variação, alta concordância a nenhuma concordância

7. Conclusão e trabalhos futuros

Conforme os resultados preliminares, o EyetrackingMOS e MOS têm uma correlação razoável. Paralelamente, a utilização de uma medida de avaliação subjetiva com rastreamento ocular oferece vantagens significativas, uma vez que permite capturar reações genuínas e síncronas aos estímulos apresentados. Além disso, o controle mais rigoroso sobre a quantidade de estímulos recebidos por cada participante pode reduzir a variação na concordância e aumentar a quantidade de estímulos por sessão. A reação mecânica ocular também pode reduzir variações na concordância, causadas pelas diferentes interpretações das descrições de pontuação de métricas subjetivas. A escala de 0 a 100 para cada indivíduo oferece uma avaliação mais detalhada e precisa, permitindo uma maior granularidade na análise das respostas, ao contrário das escalas limitadas a poucos pontos. Embora a produção dessa medida seja mais complexa e demorada, o benefício de obter uma análise mais transparente das reações dos participantes justifica seu uso como complemento do MOS tradicional.

Como trabalhos futuros, pretende-se experimentar diferentes tecnologias/plataformas de captação ocular para comparar a precisão da captação. Também é importante obter dados estatísticos com uma distinção das pontuações fornecidas entre os grupos de especialistas e não especialistas. Além disso, a seleção de variáveis deve ser refinada, como, por exemplo, calcular a fixação no espaço intermediário entre as imagens, o que pode oferecer uma compreensão mais detalhada das reações dos participantes. Por fim, explorar maneiras de realizar esses testes gratuitamente, seja por meio de parcerias, uso de plataformas de *crowdsourcing* ou outras abordagens que reduzam os custos e ampliem o acesso aos participantes.

8. Agradecimentos

Este trabalho foi realizado no Centro de Inteligência Artificial (C4AI-USP), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, bolsa #2019/07665-4) e da IBM Corporation. O projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Agradecimentos também são dirigidos ao Programa de Excelência Acadêmica (PROEX) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), nº 88887.841258/2023-00.

Referências

ALMEIDA, R. A. S. d., OLIVEIRA JR., M., and COZIJN, R. (2021). *Paradigma do Mundo Visual: Método de Rastreamento Ocular*, chapter 5. Blucher Open Access.

Batista, N. A. R. (2019). Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. Master's thesis, Unicamp.

Cagliari, L. C. (1992). Prosódia: algumas funções dos supra-segmentos. *Cadernos de estudos linguísticos*, 23:137–151.

Casanova, E., Shulby, C., Gölge, E., Müller, N. M., de Oliveira, F. S., Junior, A. C., da Silva Soares, A., Aluisio, S. M., and Ponti, M. A. (2021). Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. (2022). Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.

Choi, Y., Jung, Y., Suh, Y., and Kim, H. (2022). Learning to maximize speech quality directly using mos prediction for neural text-to-speech. *IEEE Access*, 10:52621–52629.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4):161–183.

Hoogeboom, E., Van Den Berg, R., and Welling, M. (2019). Emerging convolutions for generative normalizing flows. In *International conference on machine learning*, pages 2771–2780. PMLR.

ITU - R (2017). ITU-T Rec. P.10/G.100 (11/2017): Vocabulary for performance, quality of service and quality of experience. Recommendation P.10/G.100, International Telecommunication Union. https://www.itu.int/rec/T-REC-P.10-201711-I/en.

ITU - T (1996). Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., and Wu, Y. (2019). Transfer learning from speaker verification to multispeaker text-to-speech synthesis.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., and Zhao, S. (2024). Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models.

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In Podesva, R. J. and Sharma, D., editors, *Research Methods in Linguistics*, pages 135–168. Cambridge University Press, Cambridge.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Le Maguer, S., King, S., and Harte, N. (2024). The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech Language*, 84:101577.

Ling, L., Fernandes Tavares, T., Barbosa, P., and Batista, N. (2018). Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets.

Loizou, P. C. (2011). *Speech Quality Assessment*, pages 623–654. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mitchell, D. C. (2004). On-line methods in language processing: introduction and historical review. In Carreiras, M. and Clifton Jr., C., editors, *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond*, pages 15–32. Psychology Press.

Mota, J. A., Ribeiro, S. S. C., and de Oliveira, J. M. (2023). *Atlas Linguístico Do Brasil: Comentários às Cartas Linguísticas 1-V. 3*. Ed. Universidade Estadual de Londrina.

Nguyen, T.-N., Pham, N.-Q., and Waibel, A. (2023). Syntacc: Synthesizing multi-accent speech by weight factorization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ren, Y., Hu, C., Tao, X., Zhao, Z., Zhang, X., Li, Q., Lei, L., Zhou, S., Liu, J., and Liu, S. (2021). Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Ren, Y., Zhao, Z., Tan, X., Yi, J., Cheng, Y.-L., Yang, J., Qin, T., and Liu, T.-Y. (2022). Naturalspeech: End-to-end text to speech synthesis with human-level quality. In *Advances in Neural Information Processing Systems*.

Ribeiro, F., Florêncio, D., Zhang, C., and Seltzer, M. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2416–2419. IEEE.

Sellam, T., Bapna, A., Camp, J., Mackinnon, D., Parikh, A. P., and Riesa, J. (2023). Squid: Measuring speech naturalness in many languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. (2023). Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Soong, F., Qin, T., Zhao, S., and Liu, T.-Y. (2022). Naturalspeech: End-to-end text to speech synthesis with human-level quality.

Ynoguti, C. A. (1999). *Reconhecimento de Fala Contínua Utilizando Modelos Ocultos de Markov*. PhD thesis, Unicamp.