Síntese de Voz Emocional Multi-Idioma para Português Brasileiro: Uma Análise Comparativa de Abordagens de Ajuste Fino

Daniel Oliveira de Brito
UNESP
São José do Rio Preto, Brasil
daniel.o.brito@unesp.br

Sidney Evaldo Leal
Venturus
Campinas, Brasil
sidney.leal@venturus.org.br

Arnaldo Candido Junior
UNESP
São José do Rio Preto, Brasil
arnaldo.candido@unesp.br

Resumo

A síntese de voz emocional multi-idioma para português brasileiro é pouco explorada. Este trabalho investiga diferentes abordagens para incorporar controle emocional em síntese multi-idioma português-inglês, comparando cinco variantes: modelo base YourTTS, ajuste fino com dados emocionais, condicionamento via *tokens* textuais, e arquitetura VECL-TTS com *embeddings* emocionais sob diferentes configurações. Utilizamos *datasets* emocionais em inglês (RAVDESS, Emotional Speech Dataset) e português brasileiro (VERBO), totalizando 14,4 horas, para ajuste fino a partir do modelo YourTTS pré-treinado. A avaliação combinou métricas objetivas (similaridade de *embeddings* emocionais e de falante) com avaliação subjetiva por dez participantes. Os resultados revelam que abordagens arquiteturalmente simples podem alcançar desempenho perceptual comparável ou superior a métodos mais complexos: o YourTTS com ajuste fino obteve a melhor qualidade geral, o condicionamento por *tokens* alcançou a maior similaridade emocional percebida, enquanto o VECL-TTS maximizou o controle emocional objetivo com degradação na qualidade e na similaridade de falante. Observou-se ainda uma competição entre controle emocional e preservação de identidade vocal, bem como discrepâncias entre métricas objetivas e percepção humana. Este trabalho demonstra a viabilidade de transferência emocional multi-idioma para português brasileiro via ajuste fino com recursos limitados.

1 Introdução

Sistemas de síntese de voz (Text-to-Speech, TTS) multi-idioma têm se estabelecido como componentes fundamentais para aplicações de comunicação multilíngue, permitindo a transferência de características vocais e expressivas entre idiomas distintos (Casanova et al., 2023). A incorporação de controle emocional representa desafio adicional significativo, exigindo que o modelo aprenda representações de identidade vocal, conteúdo linguístico e estado emocional (Gudmalwar et al., 2024; Zhu et al., 2023).

No contexto brasileiro, a escassez de sistemas TTS com controle emocional para português brasileiro constitui lacuna importante. Este trabalho investiga diferentes abordagens para incorporar controle emocional em síntese multi-idioma para português brasileiro, comparando: (1) ajuste fino com dados emocionais sem modificações arquiteturais; (2) condicionamento através de *tokens* textuais de emoção; e (3) arquitetura VECL-TTS com *embeddings* emocionais dedicados.

As principais contribuições deste trabalho são: (1) uma análise comparativa entre estratégias de incorporação de emoção em cenários de baixos recursos; (2) uma metodologia para síntese de voz emocional voltada especificamente para o português brasileiro, adaptando arquiteturas estado-da-arte para as particularidades da língua; e (3) a disponibilização pública dos modelos treinados e do código de implementação[1].

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a metodologia, incluindo os modelos avaliados, conjuntos de dados e protocolo experimental; a Seção 4 apresenta os resultados das avaliações subjetivas e objetivas; a Seção 5 discute os achados; por fim, a Seção 6 apresenta as conclusões e direções para trabalhos futuros.

2 Trabalhos Relacionados

2.1 Síntese de Fala Emocional

A síntese de fala emocional tem explorado diferentes paradigmas de controle, desde rótulos discretos até representações contínuas e condicionamento

[1] https://github.com/danielbrito91/vecl-tts

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 638–646

textual. Para uma revisão sobre o tema, considere Barakat et al. (2024).

Controle via *embeddings* discretos e contínuos: Abordagens iniciais utilizaram rótulos discretos (EmoSpeech (Diatlova and Shutov, 2023), EmoQ-TTS (Im et al., 2022)), enquanto trabalhos recentes empregam *embeddings* contínuos. VECL-TTS (Gudmalwar et al., 2024) concatena *embeddings* emocionais e de falante, utilizando perdas adicionais que penalizam divergências entre os *embeddings* do áudio sintetizado e do áudio de referência; DiCLET-TTS (Li et al., 2023) utiliza difusão para transferência emocional inglês-mandarim.

Controle via linguagem natural: TextrolSpeech (Ji et al., 2024) introduz um corpus (330h, 236k amostras) com descrições naturais de estilo e propõe uma arquitetura baseada em codec com dois estágios: modelo de linguagem autoregressivo para a primeira camada de *tokens* acústicos e não autoregressivo para as camadas subsequentes. ParaEVITS (Jing et al., 2024) combina *computational paralinguistics* (CP) com aprendizagem contrastiva (ParaCLAP), usando modelos de difusão para mapear descrições em *embeddings* de emoção que controlam atributos prosódicos finos. PromotiCon (Lee et al., 2024) e PROEMO (Zhang et al., 2025) também exploram prompts textuais, com PROEMO integrando LLMs para predição de intensidade.

2.2 TTS Multi-Idioma com Controle Emocional

YourTTS (Casanova et al., 2023) estabeleceu o paradigma de zero-shot multi-idioma voice cloning via GlowTTS. METTS (Zhu et al., 2023) estende-se para a síntese emocional multilíngue através de: (i) modelagem multi-escala, separando representações agnósticas à linguagem das dependentes de linguagem; (ii) perturbação de atributos do sinal para desacoplamento de timbre; (iii) correspondência de emoção baseada em quantização vetorial para seleção de referências. Os autores reportam transferência emocional multi-falante e multi-idioma para mandarim-inglês.

Pawar et al. (2025) focam especificamente em sotaques índicos, estendendo Parler-TTS com: alinhamento de fonemas híbrido específico por idioma, *embeddings* de emoção culturalmente sensíveis, e alternância dinâmica entre idiomas via quantização vetorial residual. Com essas modificações, reportam 23,7% de melhoria em precisão de sotaque e 85,3% em reconhecimento emocional.

3 Metodologia

3.1 Modelos Utilizados

Comparamos cinco variantes de modelos de síntese de voz para avaliar diferentes abordagens de controle emocional em contexto multilíngue. Com exceção do modelo base, os demais modelos foram ajustados com o mesmo conjunto de dados descrito em 3.2.1.

1. **YourTTS:** Modelo base sem controle emocional explícito (Casanova et al., 2023), utilizando checkpoint treinado no dataset CML-TTS (Oliveira et al., 2023).

2. **YourTTS com emoção implícita:** Versão ajustada do modelo base em dados emocionais, sem mecanismo explícito de controle.

3. **YourTTS com *tokens* emocionais:** Modelo ajustado utilizando marcadores textuais de emoção precedendo o texto de entrada.

4. **VECL-TTS (Gudmalwar et al., 2024) ($\alpha_{\text{falante}} = 1$):** Implementação com *embeddings* emocionais em que a consistência de falante possui peso inferior à consistência emocional.

5. **VECL-TTS ($\alpha_{\text{falante}} = 9$):** Variante aumentando o peso para consistência de falante.

Para extração de *embeddings*, utilizamos SpeechBrain/IEMOCAP[2] (emoção) e H/ASP (Heo et al., 2020) (falante). Nossa implementação segue a configuração "Ablation 2" do VECL-TTS original, incorporando *embeddings* emocionais via concatenação, sem perda de consistência de conteúdo.

3.2 Conjuntos de Dados

3.2.1 Treinamento

Utilizamos exclusivamente conjuntos de dados com anotações emocionais não presentes no treinamento original.

Inglês:

- **RAVDESS** (Livingstone and Russo, 2018): 1.440 gravações de 24 atores (12 homens, 12 mulheres) produzindo 8 emoções distintas, com um total de 1 hora e 28 minutos.

[2] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

- **Emotional Speech Dataset** (Zhou et al., 2022): 350 sentenças cobrindo 5 emoções, produzidas por 10 locutores, com 13 horas e 24 minutos.

Português Brasileiro:

- **VERBO** (Torres Neto et al., 2018): gravações de 12 locutores (6 homens, 6 mulheres) expressando 7 emoções categóricas, totalizando 1167 amostras e 47 minutos.

3.2.2 Avaliação

Construímos conjunto balanceado com 4 emoções (raiva, alegria, tristeza, neutro) × 2 idiomas × 2 gêneros, totalizando 16 condições. Para cada condição, geramos áudio de referência sintético via Google Veo 3, permitindo maior controle experimental.

3.3 Protocolo Experimental

Geramos 400 amostras distribuídas uniformemente: 80 por modelo, 5 sentenças para cada uma das 16 condições de referência (emoção, idioma e gênero).

3.3.1 Avaliação Subjetiva

Conduzimos avaliação perceptual com dez avaliadores em 40 amostras selecionadas, garantindo representação balanceada de modelos, emoções e idiomas. Cada avaliador classificou as amostras em escala de 1 a 5 segundo três critérios:

- **Naturalidade:** qualidade geral e realismo da fala sintética
- **Similaridade de falante:** preservação das características vocais do áudio de referência
- **Similaridade emocional:** correspondência entre emoção pretendida e expressa

Calculamos médias e intervalos de confiança agregando avaliações de todos os participantes para cada modelo, controlando variabilidade inter-avaliador.

3.3.2 Avaliação Objetiva

Complementamos a avaliação perceptual com métricas objetivas no conjunto completo de 400 amostras:

- **Similaridade emocional:** similaridade de cosseno entre *embeddings* emocionais do áudio de referência e da síntese (SpeechBrain/IEMOCAP (Ravanelli et al., 2021)).
- **Similaridade de falante:** similaridade de cosseno entre *embeddings* de falante da referência e síntese (H/ASP (Heo et al., 2020)).

4 Resultados

4.1 Análise subjetiva

A Tabela 1 apresenta os resultados subjetivos.

Modelo	MOS Qual.	Sim. Emo.	Sim. Fal.
YourTTS	2,76 [2,50 – 3,05]	2,38 [1,99 – 2,66]	**2,68 [2,29 – 3,04]**
YourTTS: FT	**3,31 [2,96 – 3,68]**	2,70 [2,30 – 3,03]	2,55 [2,12 – 2,91]
YourTTS: FT-T	2,86 [2,66 – 3,10]	**2,74 [2,35 – 3,11]**	2,45 [2,06 – 2,80]
VECL ($\alpha = 1$)	2,54 [2,29 – 2,88]	2,55 [2,09 – 2,90]	2,08 [1,76 – 2,40]
VECL ($\alpha = 9$)	2,12 [1,88 – 2,44]	2,48 [2,01 – 2,85]	2,10 [1,75 – 2,44]

Tabela 1: Avaliação subjetiva com N=10 participantes. Valores: média [IC 90%]. MOS Qual.: qualidade geral; Sim.: similaridade percebida (escala 1-5) para os modelos avaliados (FT: ajustado com dados emocionais; FT-T: ajustado com *tokens* de emoção. Valores de α do VECL para Consistência do Falante)

Qualidade de Áudio

Um teste de Friedman revelou diferenças significativas nas avaliações de qualidade entre os modelos ($\chi^2 = 32,64$, $p = 1,41 \times 10^{-6}$). Testes pareados de Wilcoxon a posteriori com correção de Holm confirmaram que o VECL-TTS ($\alpha = 9$) recebeu avaliações significativamente inferiores em relação a todos os demais modelos ($p < 0,05$), com tamanhos de efeito (d de Cohen) grandes a muito grandes: vs. YourTTS: FT ($d = 1,86$), vs. YourTTS: FT-T ($d = 1,46$), vs. YourTTS base ($d = 1,18$), e vs. VECL-TTS ($\alpha = 1$) ($d = 0,72$). Na direção oposta, o YourTTS: FT destacou-se como o modelo de maior qualidade percebida, com diferenças significativas em relação ao YourTTS base ($p = 0,02$) e ao VECL-TTS ($\alpha = 1$) ($p = 0,02$). As demais comparações não atingiram significância estatística após a correção, incluindo YourTTS base vs. YourTTS: FT-T ($p = 0,41$) e YourTTS: FT-T vs. VECL-TTS ($\alpha = 1$) ($p = 0,09$), sugerindo qualidade perceptual comparável entre esses modelos.

Notavelmente, a comparação entre VECL-TTS ($\alpha = 1$) e VECL-TTS ($\alpha = 9$) indica que o aumento do parâmetro α, embora teoricamente projetado para melhorar a consistência do falante, comprometeu a qualidade de áudio percebida (MOS: 2,54 $\rightarrow$ 2,12, $d = 0,72$, $p = 0,048$) sem produzir benefícios perceptuais correspondentes significativos em outras dimensões (ver análises abaixo).

Similaridade Emocional

Para similaridade emocional, o teste de Friedman não revelou diferenças significativas entre os

modelos ($\chi^2 = 6,01, p = 0,198$), e nenhuma comparação pareada de Wilcoxon atingiu significância após correção de Holm ($p > 0,70$ em todos os pares). Ainda assim, os modelos YourTTS: FT e YourTTS: FT-T apresentaram as maiores avaliações numéricas (MOS = 2,7), com tamanhos de efeito de médio porte em relação ao modelo base ($d = 0,46$ para YourTTS: FT e $d = 0,50$ para YourTTS: FT-T). Os modelos VECL-TTS ($\alpha = 1$ e $\alpha = 9$), por sua vez, situaram-se em posição intermediária, com efeitos pequenos tanto em relação ao modelo base ($d < 0,24$) quanto aos modelos com ajuste fino ($d < 0,32$). Esses resultados sugerem que o ajuste fino com dados emocionais tende a melhorar a similaridade emocional percebida, enquanto o condicionamento por *embeddings* no estilo VECL não proporcionou ganho comparável nessa dimensão.

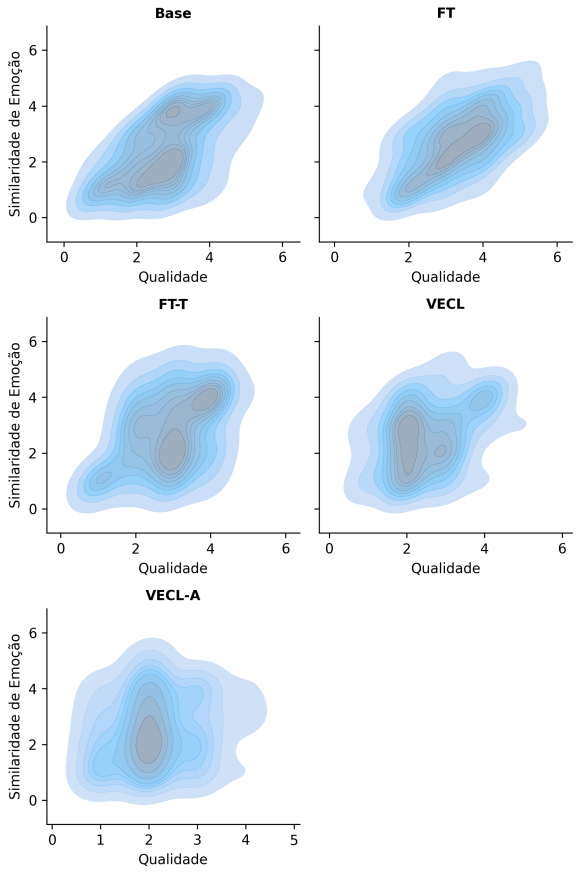

Figura 1: Correlação entre avaliação perceptual de qualidade e emoção

A Figura 1 apresenta a correlação entre similaridade emocional e qualidade do áudio nas avaliações subjetivas. Verifica-se que os modelos com a arquitetura do YourTTS (modelo base, FT, e FT-T) apresentam correlações moderadas e significativas ($r = 0,56$ e $p = 6,25 \times 10^{-8}$, $r = 0,58$ e $p = 1,13 \times 10^{-8}$, e $r = 0,40$ e $p = 0,0002$, respectivamente). O VECL-TTS ($\alpha_{falante} = 1$) também apresentou uma correlação de Pearson estatisticamente significativa ($r = 0,35$ e $p = 0,0013$), diferentemente do VECL-TTS ($\alpha_{falante} = 9$) ($r = 0,12$ e $p = 0,26$).

Similaridade de Falante

Para similaridade do falante, o teste de Friedman revelou diferenças significativas ($\chi^2 = 20,99$, $p = 0,0003$). Comparações pareadas de Wilcoxon com correção de Holm identificaram diferenças significativas entre o YourTTS base e ambas as variantes VECL-TTS ($p = 0,04$ para $\alpha_{falante} = 1$ e $p = 0,02$ para $\alpha_{falante} = 9$, com tamanhos de efeito grandes ($d = 0,87$ e $d = 0,81$, respectivamente). Os modelos sem condicionamento emocional explícito (YourTTS base e YourTTS: FT) obtiveram as maiores avaliações de similaridade (MOS = 2,68 e 2,55), seguidos pelo YourTTS: FT-T (MOS = 2,45) em posição intermediária, com efeitos médios em relação aos modelos VECL ($d = 0,49 - 0,54$), porém sem atingir significância estatística.

As variações do VECL-TTS ($\alpha_{falante} = 1$ e $\alpha_{falante} = 9$) obtiveram novamente avaliações praticamente idênticas (MOS = 2,08 e 2,10, e $d = 0,04$), indicando que o aumento do parâmetro α de consistência do falante não resultou em melhoria perceptual, apesar de produzir leve aumento nas métricas objetivas (0,25 vs. 0,28; Tabela 2). Esta discrepância entre métricas objetivas e percepção humana sugere que diferenças de similaridade de cosseno abaixo de 0,05 podem não ser perceptualmente significativas.

4.2 Análise objetiva

A Tabela 2 apresenta as métricas objetivas calculadas usando *embeddings* pré-treinados para quantificar similaridade emocional e de falante entre áudios sintetizados e referências.

Modelo	Cos. Emoção	Cos. Falante
YourTTS	0,15	**0,39**
YourTTS: FT	0,30	0,35
YourTTS: FT-T	0,34	0,34
VECL-TTS ($\alpha = 1$)	0,53	0,25
VECL-TTS ($\alpha = 9$)	**0,54**	0,28

Tabela 2: Métricas objetivas de similaridade. Cos. Emoção: similaridade de cosseno entre *embeddings* emocionais da síntese e referência. Cos. Falante: similaridade de cosseno entre *embeddings* de falante da síntese e referência. Valores representam médias no conjunto de teste. FT: ajustado com dados emocionais; FT-T: ajustado com *tokens* de emoção; α: parâmetro de consistência do falante no VECL-TTS.

Similaridade Emocional

Os resultados da Tabela 2 demonstram progressão clara na capacidade de controle emocional entre as diferentes abordagens. O YourTTS (modelo base) alcançou similaridade emocional de 0,15, refletindo a ausência de mecanismo explícito de controle emocional. O ajuste fino com dados emocionais (YourTTS: FT) elevou este valor para 0,30, indicando que a exposição a exemplares emocionais durante o treinamento permite captura implícita de características expressivas.

A incorporação de *tokens* textuais como marcadores de emoção (YourTTS: FT-T) resultou em similaridade de 0,34, superando marginalmente a abordagem de emoção implícita. Este resultado sugere que sinalizadores explícitos no texto de entrada facilitam o condicionamento emocional do modelo.

As variantes VECL-TTS apresentaram desempenho superior, com similaridade emocional de 0,53 ($\alpha = 1$) e 0,54 ($\alpha = 9$). A diferença mínima entre as configurações $\alpha = 1$ e $\alpha = 9$ no controle emocional indica que o parâmetro de balanceamento tem impacto limitado na preservação de características emocionais, sugerindo que o *embedding* emocional contribui independentemente do peso relativo atribuído ao *embedding* de falante.

A Figura 2 apresenta a distribuição da similaridade emocional para cada modelo segmentada por emoção (raiva, alegria, neutro e tristeza). Nota-se que as abordagens avaliadas se destacaram na emoção "raiva", apresentando medianas elevadas e distribuições concentradas.

Similaridade de Falante

A Figura 3 apresenta a distribuição da similaridade de falante para cada modelo. A análise revela padrão inverso ao observado para controle emocional. O YourTTS (modelo base) obteve similaridade de 0,39, estabelecendo referência para preservação de características vocais sem controle emocional explícito.

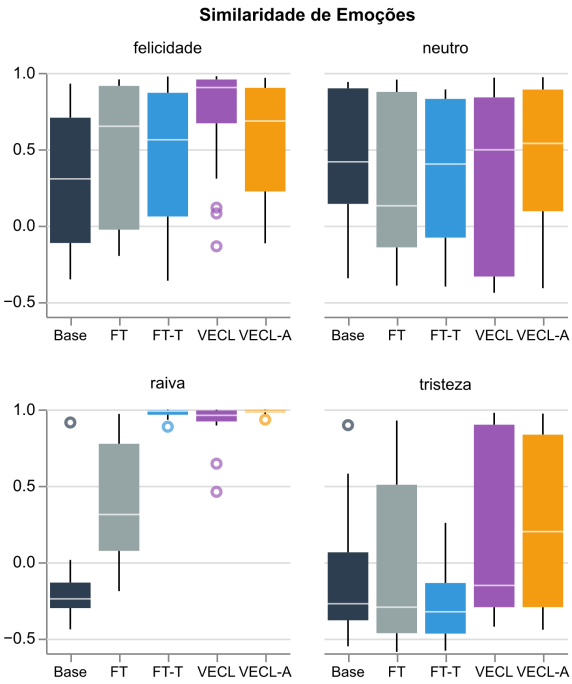

Figura 2: Similaridade emocional entre modelos usando embeddings do IEMOCAP.

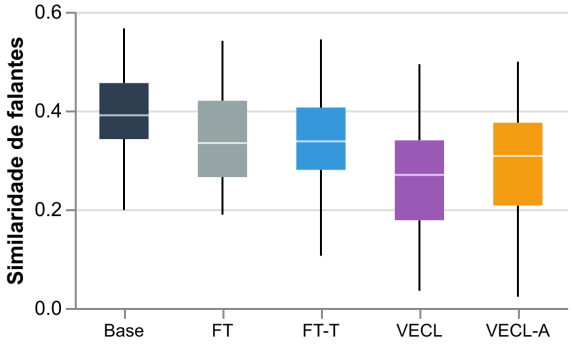

Figura 3: Similaridade dos falantes entre modelos usando *embeddings* do H/ASP.

O ajuste fino com dados emocionais (YourTTS: FT) resultou em leve degradação para 0,35. A abordagem com *tokens* emocionais (YourTTS: FT-T) manteve similaridade de 0,34, comparável ao modelo com emoção implícita.

As variantes VECL-TTS apresentaram redução mais acentuada na similaridade de falante: 0,25 para $\alpha = 1$ e 0,28 para $\alpha = 9$. O aumento de α de 1 para 9 resultou em melhoria modesta (0,25

para 0,28; $\Delta = +0,03$), sugerindo que o rebalanceamento via hiperparâmetro oferece recuperação parcial, mas insuficiente, da similaridade de falante. Esta melhoria objetiva não se traduziu em diferença perceptual significativa (MOS = 2,08 e 2,10, $d = 0,04$), evidenciando possível limiar de percepção abaixo do qual variações em similaridade de cosseno não são detectáveis por avaliadores humanos.

5 Discussão

5.1 Comparação entre as arquiteturas

Os resultados demonstram que as técnicas avaliadas foram capazes de representar emoções nas sínteses de fala, corroborando trabalhos anteriores (Gudmalwar et al., 2024; Zhu et al., 2023). No entanto, emergem diferenças importantes entre abordagens de diferentes complexidades.

A incorporação direta de *embeddings* emocionais através de concatenação com representações de falante constitui mecanismo para controle emocional, como demonstrado tanto pelo VECL-TTS original quanto por nossa implementação. Todavia, a abordagem VECL-TTS não superou métodos mais simples em avaliações perceptuais, apesar de alcançar melhor desempenho em métricas objetivas de similaridade emocional (0,53-0,54 vs. 0,34 para *tokens*).

Os resultados da Tabela 1 sugerem um *trade-off* entre as dimensões avaliadas. Enquanto o YourTTS: FT obteve melhor equilíbrio geral (MOS = 3,31; Sim. Fal. = 2,55), o YourTTS: FT-T se destacou pela maior similaridade emocional (2,74), mas com menor qualidade geral percebida (2,86) e menor similaridade de falantes (2,45), indicando que o condicionamento emocional explícito pode comprometer outras dimensões da síntese.

Contexto com trabalhos recentes e diferenças metodológicas: Gudmalwar et al. (Gudmalwar et al., 2024) compararam VECL-TTS com YourTTS (similar ao nosso experimento com emoção implícita), METTS (Zhu et al., 2023) e M3 (Shang et al., 2021), mas não avaliaram abordagens de marcadores de estilo, lacuna que nosso trabalho endereça através de *tokens* textuais de emoção.

Uma diferença metodológica importante contextualiza os resultados apresentados: METTS e VECL-TTS empregaram treinamento do zero em *datasets* com mais de 40h. Em contraste, nosso trabalho consiste em ajuste fino do YourTTS pré-treinado, em um conjunto de 14,4h dados emocionais. Esta escolha explica tanto vantagens quanto limitações discutidas abaixo.

O ajuste fino alcançou MOS de qualidade competitivo (3,31 vs. 3,95 do METTS) utilizando <25% dos dados emocionais e com menor tempo de treinamento, demonstrando viabilidade para cenários de recursos limitados. Entretanto, nosso VECL-TTS apresentou similaridade de falante objetiva substancialmente inferior (0,25-0,28 vs. 0,86 reportado por Gudmalwar et al. (2024)), sugerindo que a representação timbre-emoção pode requerer: (1) retreinamento extensivo permitindo reorganização do espaço latente; (2) maior volume de dados e iterações para convergência completa das perdas de consistência. Nosso trabalho também se valeu majoritariamente de dados em inglês, sendo a avaliação realizada com síntese em português, ao passo que Gudmalwar et al. (2024) e Zhu et al. (2023) possuíam dados nos idiomas de interesse.

Simplicidade arquitetural: O sucesso relativo de *tokens* emocionais (FT-T: Sim. Emo. = 2,74) sugere que o ajuste fino beneficia-se de modificações leves em vez de reestruturações arquiteturais profundas. Abordagens que minimizam perturbação do espaço latente pré-treinado - como *tokens* que atuam no nível de entrada em vez de *embeddings* de alta dimensão concatenados internamente - parecem mais adequadas para transferência de aprendizado.

Aplicabilidade: Os resultados indicam que o ajuste fino multilíngue constitui estratégia viável particularmente para: (1) línguas de médios recursos como português brasileiro; (2) prototipagem rápida de sistemas emocionais; (3) aplicações em que a qualidade do modelo base deve ser preservada. Comparando com os resultados de literatura, o treinamento completo permanece preferível quando desempenho máximo é crítico e recursos abundantes estão disponíveis, mas nosso trabalho demonstra que resultados úteis são alcançáveis com abordagem mais acessível a línguas menos favorecidas.

5.2 Discrepância entre Métricas Objetivas e Subjetivas

A comparação entre Tabelas 1 e 2 revela discrepâncias entre resultados objetivos e subjetivos. Para similaridade emocional, VECL-TTS alcançou melhor desempenho objetivo (0,53-0,54) mas não superou YourTTS: FT-T em avaliação subjetiva (2,48-2,55 vs. 2,74). Para similaridade de falante, o aumento de α melhorou métricas objetivas (+0,03)

sem impacto perceptual (MOS = 2,08 e 2,10).

Essas discrepâncias sugerem que: (1) métricas baseadas em *embeddings* podem não capturar completamente dimensões perceptualmente relevantes da similaridade emocional e vocal; (2) existe limiar abaixo do qual variações em similaridade de cosseno não são perceptíveis a avaliadores humanos; (3) a qualidade geral do áudio e similaridade de falante pode modular a percepção de similaridade de emoções, conforme observado na Figura 1, potencialmente comprometendo a avaliação de outras dimensões perceptuais. A ausência de correlação no VECL-TTS ($\alpha_{falante} = 9$) - modelo de menor qualidade percebida - pode ser explicada pela baixa variabilidade nas avaliações de qualidade ($\sigma = 0,77$, comparado com $\sigma = 0,95 - 1,02$ nos demais modelos), comprimindo as notas em uma faixa restrita, dificultando a detecção de associações.

5.3 Competição entre Controle Emocional e Identidade Vocal

Os resultados evidenciam a dificuldade de maximizar simultaneamente controle emocional e preservação de identidade vocal. As abordagens de ajuste fino (YourTTS: FT e FT-T) mantêm melhor equilíbrio, conforme elencado na Tabela 2, com degradação limitada na similaridade de falante (0,34–0,35) em troca de ganhos moderados em expressividade emocional (0,30–0,34).

A abordagem com *tokens* emocionais (YourTTS: FT-T) alcança 0,34 em similaridade emocional e 0,34 em similaridade de falante, valores equilibrados que podem ser preferíveis em aplicações em que ambas as dimensões são críticas. A degradação observada pode sugerir que a representação emocional pode dominar o espaço latente na concatenação entre *embeddings* com representação de emoção e do falante. Gudmalwar et al. (2024) encontraram resultados semelhantes nesse sentido, observando uma maior similaridade de falantes perceptual no modelo YourTTS quando em comparação com o VECL-TTS.

5.4 Análise por Emoção segundo o Modelo Circumplexo de Russell

Os resultados agregados mascaram variações importantes no desempenho entre emoções específicas. Utilizamos o modelo circumplexo de Russell (Russell, 1980) — que organiza emoções pelos eixos de valência (positiva/negativa) e ativação (alta/baixa) — para investigar sistematicamente essas diferenças. A Figura 2 revela padrões distintos: raiva e alegria (alta ativação) versus tristeza (baixa ativação) e neutro (valências intermediárias).

Emoções de Alta Ativação (Raiva e Alegria): Raiva demonstrou o melhor desempenho entre todas as emoções, com VECL-TTS alcançando similaridade próxima a 1,0, devido a características acústicas salientes. Alegria também apresentou desempenho elevado, com medianas consistentemente altas para todos os modelos, beneficiando-se igualmente de correlatos acústicos pronunciados da alta ativação.

Emoções de Baixa Ativação (Neutro e Tristeza): Neutro apresentou desempenho intermediário, com variabilidade entre modelos. Tristeza revelou o pior desempenho entre todas as emoções avaliadas, com medianas próximas ou abaixo de zero, especialmente para modelo base e modelos com ajuste fino. Este padrão sugere dificuldade sistemática com estados de baixa ativação: características acústicas sutis são mais desafiadoras para síntese controlada que manifestações prosódicas pronunciadas de alta ativação.

A análise evidencia que a dimensão de ativação é fator determinante para qualidade de síntese emocional: emoções de alta ativação (raiva, alegria) são substancialmente mais facilmente controláveis que emoções de baixa ativação (neutro, tristeza), independentemente de valência. Abordagens futuras poderiam explorar augmentação específica de dados de baixa ativação ou estratégias de treinamento focadas nesta dimensão.

6 Conclusões e Trabalhos Futuros

Este estudo investigou diferentes abordagens para incorporar controle emocional em síntese de voz multi-idioma para português brasileiro, comparando métodos de ajuste fino simples, condicionamento por *tokens* textuais e arquitetura VECL-TTS com *embeddings* emocionais dedicados. Os resultados revelam três achados principais que contribuem para o entendimento de síntese emocional: (1) abordagens arquiteturalmente mais simples podem alcançar desempenho perceptual comparável ou superior a métodos mais complexos em contexto de ajuste fino; (2) observou-se uma competição entre controle emocional e preservação de identidade vocal; (3) métricas objetivas baseadas em *embeddings* nem sempre se correlacionam com percepção humana, havendo evidências de que a qualidade geral do áudio pode modular a avaliação perceptual

das demais dimensões.

Para aplicações práticas em português brasileiro, os resultados sugerem que a escolha da abordagem deve considerar os requisitos específicos da aplicação. O modelo YourTTS: FT oferece o melhor equilíbrio qualidade-identidade (MOS 3,31; similaridade de falante 2,55), sendo adequado para aplicações que priorizam naturalidade e preservação vocal. A abordagem YourTTS: FT-T, com melhor similaridade emocional percebida (2,74), pode ser preferível quando expressividade emocional é crítica. A estratégia de ajuste fino demonstrou-se particularmente viável para cenários de recursos limitados.

Este estudo apresenta limitações importantes. Embora a avaliação subjetiva com $N = 10$ participantes tenha permitido identificar diferenças significativas em qualidade e similaridade de falante, o poder estatístico permanece limitado para detectar diferenças sutis, como as observadas em similaridade emocional. O conjunto de dados emocionais é substancialmente menor que *datasets* utilizados em trabalhos com treinamento completo, potencialmente subestimando o desempenho máximo alcançável pela arquitetura VECL-TTS. Adicionalmente, o desempenho inferior sistemático em emoções de baixa ativação (neutro, tristeza) sugere limitações dos métodos atuais para capturar características prosódicas sutis.

Investigações futuras devem explorar: (1) estratégias de treinamento específicas para emoções de baixa ativação, incluindo aumento de dados e perdas auxiliares focadas na dimensão de ativação do modelo circumplexo; (2) mecanismos arquiteturais mais sofisticados para preservar a identidade vocal e o controle emocional; (3) treinamento do zero da arquitetura VECL-TTS com maior volume de dados emocionais, permitindo a reorganização completa do espaço latente e a convergência extensiva das perdas de consistência.

Este trabalho constitui uma contribuição para a síntese de voz emocional em português brasileiro, demonstrando a viabilidade da transferência multi-idioma de controle emocional e fornecendo uma análise comparativa de diferentes abordagens arquitetônicas.

7 Agradecimentos

Este trabalho foi realizado com o apoio do Centro de Inteligência Artificial (C4AI-USP), da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo #2019/07665-4) e da IBM Corporation. Este projeto também contou com o apoio do Ministério da Ciência, Tecnologia e Inovação, 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.

References

Huda Barakat, Oytun Turk, and Cenk Demiroglu. 2024. Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):11.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2023. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone. *Preprint*, arXiv:2112.02418.

Daria Diatlova and Vitaly Shutov. 2023. EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech. *Preprint*, arXiv:2307.00024.

Ashishkumar Gudmalwar, Nirmesh Shah, Sai Akarsh, Pankaj Wasnik, and Rajiv Ratn Shah. 2024. VECL-TTS: Voice identity and Emotional style controllable Cross-Lingual Text-to-Speech. *Preprint*, arXiv:2406.08076.

Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung. 2020. Clova Baseline System for the VoxCeleb Speaker Recognition Challenge 2020. *Preprint*, arXiv:2009.14153.

Chae-Bin Im, Sang-Hoon Lee, Seung-Bin Kim, and Seong-Whan Lee. 2022. EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6317–6321.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. TextrolSpeech: A Text Style Control Speech Corpus With Codec Language Text-to-Speech Models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305.

Xin Jing, Kun Zhou, Andreas Triantafyllopoulos, and Björn W. Schuller. 2024. Enhancing Emotional Text-to-Speech Controllability with Natural Language Guidance through Contrastive Learning and Diffusion Models. *Preprint*, arXiv:2409.06451.

Ji-Eun Lee, Seung-Bin Kim, Deok-Hyeon Cho, and Seong-Whan Lee. 2024. PromotiCon: Prompt-based Emotion Controllable Text-to-Speech via Prompt Generation and Matching. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1151–1156.

Tao Li, Chenxu Hu, Jian Cong, Xinfa Zhu, Jingbei Li, Qiao Tian, Yuping Wang, and Lei Xie. 2023. DiCLET-TTS: Diffusion Model Based Cross-Lingual Emotion Transfer for Text-to-Speech — A Study Between English and Mandarin. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3418–3430.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. CML-TTS A Multilingual Dataset for Speech Synthesis in Low-Resource Languages. *Preprint*, arXiv:2306.10097.

Pranav Pawar, Akshansh Dwivedi, Jenish Boricha, Himanshu Gohil, and Aditya Dubey. 2025. Optimizing Multilingual Text-To-Speech with Accents & Emotions. *Preprint*, arXiv:2506.16310.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Zengqiang Shang, Zhihua Huang, Haozhe Zhang, Pengyuan Zhang, and Yonghong Yan. 2021. Incorporating Cross-Speaker Style Transfer for Multi-Language Text-to-Speech. In *Proc. Interspeech 2021*, pages 1619–1623.

José R. Torres Neto, Geraldo P. R. Filho, Leandro Y. Mano, and Jó Ueyama. 2018. VERBO: voice emotion recognition database in portuguese language. *Journal of Computer Science*, 14(11):1420–1430.

Shaozuo Zhang, Ambuj Mehrish, Yingting Li, and Soujanya Poria. 2025. PROEMO: Prompt-Driven Text-to-Speech Synthesis Based on Emotion and Intensity Control. *Preprint*, arXiv:2501.06276.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional Voice Conversion: Theory, Databases and ESD. *Preprint*, arXiv:2105.14762.

Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, Hongbin Zhou, Heng Lu, and Lei Xie. 2023. METTS: Multilingual Emotional Text-to-Speech by Cross-speaker and Cross-lingual Emotion Transfer. *Preprint*, arXiv:2307.15951.