

Análise da Robustez das Representações Latentes de Modelos Autossupervisionados para Reconhecimento de Emoções em Sinais de Voz

Mario Guilherme Batista, Cinthya Oestreich Silva, Victor Costa Beraldo, Wesley Beccaro, Miguel Arjona Ramírez

Resumo—Este trabalho investigou a robustez de modelos baseados em aprendizado autossupervisionado (Self-Supervised Learning, SSL) para reconhecimento de emoções na voz, especificamente HuBERT e Wav2Vec2, sob diferentes condições de degradação acústica. Para isso, foi utilizada a base de dados IEMOCAP (*Interactive Emotional Dyadic Motion Capture Database*), combinada com ruídos sintéticos telefônicos (MNRU) e ruídos ambientais da base MUSAN (adicionados em diferentes níveis de intensidade). Os resultados mostraram desempenho global superior do modelo HuBERT, com maior robustez frente às condições degradadas avaliadas. Para ambos os modelos, observou-se perda progressiva de desempenho à medida que a severidade do ruído aumentou. O modelo HuBERT apresentou elevada robustez ao ruído telefônico, enquanto sob ruído de murmúrio e multidão observou-se tendência crescente de confusão entre emoções, particularmente em direção à classe neutra. Já o Wav2Vec2 apresentou padrão de degradação menos consistente, com maior sensibilidade para emoções como raiva e tristeza. Os resultados também sugerem que diferenças arquiteturais entre os modelos influenciam diretamente sua robustez, com o mecanismo de predição mascarada baseado em unidades discretas do HuBERT favorecendo representações mais estáveis sob degradação. Observou-se ainda que determinados tipos de ruído não apenas degradam o sinal, mas introduzem características acústicas capazes de deslocar as representações latentes em direção a outras emoções, como o aumento de energia e variabilidade espectral aproximando certas amostras de padrões associados à neutralidade ou felicidade.

Palavras-Chave—Reconhecimento de emoções, Robustez, Aprendizado Profundo, Aprendizado autossupervisionado.

Abstract—This work investigated the robustness of Self-Supervised Learning (SSL)-based models for speech emotion recognition, specifically HuBERT and Wav2Vec2, under different acoustic degradation conditions. For this purpose, the IEMOCAP (*Interactive Emotional Dyadic Motion Capture Database*) dataset was used in combination with synthetic telephone noise generated through MNRU and environmental noises from the MUSAN corpus, added at different intensity levels.

The results demonstrated an overall superior performance of the HuBERT model, with greater robustness under the evaluated degraded conditions. For both models, a progressive performance degradation was observed as noise severity increased. HuBERT exhibited high robustness to telephone noise, whereas under babble and crowd noise conditions an increasing tendency of confusion between emotions was observed, particularly toward the neutral class. In contrast, Wav2Vec2 presented a less consistent degradation pattern, with higher sensitivity for emotions such as anger and sadness.

The results also suggest that architectural differences between the models directly influence their robustness, with HuBERT's masked prediction mechanism based on discrete units favoring

more stable representations under degradation. Furthermore, it was observed that certain types of noise not only degrade the signal but also introduce acoustic characteristics capable of shifting latent representations toward other emotions, such as increased energy and spectral variability approximating some samples to patterns associated with neutrality or happiness.

Keywords—Speech Emotion Recognition, Robustness, Deep learning, Self-supervised learning.

I. INTRODUÇÃO

O reconhecimento de emoções em sinais de voz (*Speech Emotion Recognition* — SER) é uma aplicação relevante de processamento de sinais de voz e tem recebido crescente atenção com o avanço das técnicas de aprendizado profundo [1]. Esse interesse tem sido impulsionado pelo potencial de aplicação comercial, especialmente em modelos classificadores empregados em sistemas embarcados de assistentes virtuais inteligentes para interação homem-máquina, bem como em ferramentas de *business analytics* voltadas à descrição e previsão de comportamentos a partir de sinais de voz produzidos na interação entre empresas e clientes [2].

Diversas tarefas de processamento de voz têm investigado a robustez de modelos frente à degradação dos sinais de entrada. Entre essas tarefas, destacam-se o reconhecimento automático de voz, o reconhecimento de palavras-chave [3], a verificação de locutor, a identificação de locutor e o reconhecimento de emoções. Nesses estudos, são analisados os impactos de ruído aditivo, reverberação, compressão e distorções de canal sobre o desempenho de modelos profundos e das representações acústicas extraídas.

Comparando com outras tarefas de classificação de voz, o reconhecimento de emoções pode apresentar maior variabilidade nos rótulos, uma vez que a percepção emocional está frequentemente associada a fatores subjetivos, contextuais e culturais. Dessa forma, a tarefa demanda modelos com maior capacidade de generalização e extração de padrões prosódicos robustos, tornando-se um cenário relevante para a avaliação e comparação de diferentes abordagens de aprendizado de representações [1].

Abordagens recentes baseadas em aprendizado autossupervisionado têm demonstrado avanços significativos na extração de atributos de sinais de voz. O modelo HuBERT [4], por exemplo, é capaz de aprender representações diretamente do sinal de voz por meio de uma estratégia de predição de unidades ocultas mascaradas, permitindo capturar de forma robusta estruturas acústicas e fonéticas relevantes.

Nesse contexto, modelos baseados em aprendizado autossupervisionado de representações (*Self-Supervised Representation Learning*) têm se destacado por aprender representações latentes mais gerais, discriminativas e robustas a degradações acústicas do sinal de voz, avaliadas com ruído aditivo, reverberação e distorções de canal [5], [6], [7]. Modelos como o Wav2Vec 2.0 [8] e o HuBERT [4] demonstraram a capacidade de extrair representações transferíveis a partir de grandes volumes de voz não rotulados, permitindo posterior adaptação a tarefas específicas com menor dependência de dados rotulados.

O presente trabalho investiga os efeitos do ruído em diferentes emoções, camadas de extração de atributos de modelos autossupervisionados e níveis de degradação do sinal, avaliando a estabilidade e a robustez das representações latentes produzidas sob diferentes degradações do sinal de voz. A Seção II detalha a metodologia de avaliação proposta, os bancos de dados avaliados e o métodos de degradação dos sinais empregados no trabalho. A Seção III discute os resultados obtidos e a Seção IV apresenta a conclusão.

II. MATERIAIS E MÉTODOS

Foi utilizado exclusivamente o banco de dados IEMOCAP (*Interactive Emotional Dyadic Motion Capture Database*) [9], amplamente utilizado em reconhecimento de emoções na fala. O conjunto contém cinco sessões de interações entre pares de atores, com sinais de voz organizados por diálogo e seus respectivos rótulos emocionais. Em geral, a sessão 5 é utilizada para teste.

A distribuição das instâncias por emoção é de 1103 exemplos de raiva, 595 de felicidade, 1041 de excitação, 1849 de frustração, 1084 de tristeza e 1708 de neutro, caracterizando um conjunto desbalanceado. Para reduzir esse desbalanceamento, felicidade e excitação foram agrupadas, seguindo o modelo circunflexo de Russell [10], e as emoções menos representativas foram excluídas. Dessa forma, o problema foi definido como uma tarefa de classificação em quatro classes: felicidade, tristeza, neutro e raiva.

Foram utilizados os modelos *Wav2Vec2-Base* e *HuBERT Base* (*hubert-base-ls960*), ambos pré-treinados em sinais de áudio a 16 kHz e disponibilizados pela biblioteca *Hugging Face Transformers*¹. O processamento dos sinais de áudio foi realizado com a biblioteca *librosa*².

Para simular os efeitos do ruído telefônico, foi utilizado o método de referência MNRU (*Modulated Noise Reference Unit*), disponibilizado pela International Telecommunication Union (ITU). O MNRU permite controlar a degradação do sinal de forma padronizada, possibilitando avaliações objetivas e reproduzíveis da qualidade da fala.

Para reproduzir as limitações de largura de banda da telefonia convencional (*narrowband*), foi aplicado um filtro passa-faixa com frequências de corte em 300 Hz e 3.400 Hz. Em seguida, foi adicionado o ruído controlado pelo MNRU, resultando no sinal degradado utilizado nos experimentos

O diagrama básico do MNRU, descrito na recomendação ITU-T P.810, é composto por dois caminhos: o do sinal (*signal path*) e o do ruído (*noise path*). Neste último, um ruído gaussiano é modulado pelo sinal de entrada e somado ao sinal original. Os ganhos dos dois caminhos são ajustados de forma que a razão entre o sinal e o ruído modulado, expressa em decibéis, seja controlada pelo parâmetro Q , medido na saída do filtro passa-faixa. Considerando ganho unitário no caminho do sinal, a saída pode ser expressa por

$$y[i] = x[i] \left[1 + 10^{-Q/20} N[i] \right], \quad (1)$$

em que $x[i]$ é o sinal de entrada, $N[i]$ é uma sequência gaussiana aleatória, Q representa a razão entre as potências do sinal de voz e do ruído modulado, em dB, e $y[i]$ é o sinal degradado.

A variação do parâmetro Q permite controlar a intensidade do ruído introduzido. Especificamente, foram considerados valores de Q no intervalo de 10 a 25 dB, no qual valores mais altos indicam menor degradação do sinal, possibilitando avaliar a robustez dos modelos em diferentes cenários de qualidade de canal.

Para o acréscimo de murmúrio (*babble noise*), que é a voz concorrente com o enunciador principal, foi utilizado o banco de dados MUSAN [11], composto por um *corpus* trechos de voz, música e ruído. Foram utilizadas 5 amostras aleatórias do *librivox*, disponível na base de dados MUSAN, com mesmo tamanho do áudio selecionado do IEMOCAP, adicionando-o à instância original, com SNR (*Signal-to-Noise Ratio*) 20, 10 e 5 dB.

A fim de simular o efeito de multidão, comum em ambientes público com muitas pessoas, foi utilizado o arquivo de áudio *noise-free-sound-0761*, do conjunto de ruídos do MUSAN, composto de 32s de vozes humanas sem conteúdo linguístico inteligível. Para cada arquivo do IEMOCAP foi adicionado um trecho aleatório de igual duração desse arquivo, com uma SNR de 20, 10 e 5 dB, com o objetivo de avaliar o efeito de diferentes intensidades.

Foi necessário investigar quais das 12 camadas dos modelos HuBERT e Wav2Vec2 produziam vetores com maior capacidade de separabilidade emocional para a tarefa. Isso ocorre porque as diferentes camadas desses modelos representam distintos níveis de abstração das características do sinal de voz, variando desde atributos acústicos de baixo nível até informações linguísticas potencialmente relevantes.

Para comparar as representações latentes geradas antes e após a introdução de ruído, foi utilizada a distância Euclidiana entre vetores. Da mesma maneira foi utilizado o coeficiente de silhueta (*Silhouette Score*) para avaliar a separabilidade das representações latentes extraídas pelos modelos SSL sob diferentes condições de ruído. Essa métrica quantifica o quão próximo cada exemplo está dos elementos de sua própria classe em comparação com os elementos das demais classes, sendo definida por

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (2)$$

em que $a(i)$ representa a distância média entre a amostra i e os demais elementos da mesma classe, enquanto $b(i)$ corresponde

¹<https://huggingface.co/transformers/>

²<https://librosa.org/>

à menor distância média entre a amostra i e os elementos das demais classes. Valores próximos de 1 indicam boa separabilidade entre classes, próximos de 0 indicam sobreposição entre grupos e negativos sugerem que a amostra está mais próxima de outra classe do que da sua própria.

As representações latentes resultantes possuem dimensão fixa igual a 768. A distância média entre os vetores correspondentes aos sinais limpos e ruidosos foi calculada com o objetivo de avaliar quantitativamente o impacto do ruído sobre a representação vetorial produzida pelos modelos.

Após a extração dos vetores, um modelo de regressão logística, treinado sobre as seções 1 a 4 do IEMOCAP, foi utilizado com o objetivo de avaliar a separabilidade entre as classes emocionais representadas pelos vetores gerados pelos modelos autossupervisionados para a seção 5. Foi utilizada como principal métrica de avaliação global a métrica *Unweighted Accuracy* (UA), definida como a média do *Recall* obtido em cada classe, dado por

$$UA = \frac{1}{N} \sum_{i=1}^N Recall_i, \quad (3)$$

em que N representa o número total de classes emocionais consideradas. A métrica *Recall*, que mede a capacidade do modelo em identificar corretamente os exemplos pertencentes a uma determinada classe de interesse, é dada por $Recall = \frac{VP}{VP+FN}$, em que VP representa o número de verdadeiros positivos e FN o número de falsos negativos.

De forma complementar, foi utilizada a métrica F1-score, que combina *Recall* e *Precisão* em uma medida única, permitindo avaliar simultaneamente a capacidade de recuperação e a confiabilidade das predições do modelo.

III. RESULTADOS E DISCUSSÕES

Os resultados mostraram que as camadas 2 e 11, para os modelos Wav2Vec2 e HuBERT, respectivamente, obtiveram os maiores valores de acurácia não ponderada UA durante a avaliação inicial de cada modelo, de forma que todos os experimentos foram feitos sobre essas camadas.

A Tabela I apresenta a distância Euclidiana média dos vetores das representações latentes causado pelos ruídos telefônico, de multidão e murmúrio, respectivamente, para diferentes SNR, considerando os vetores extraídos da sessão 5 da base IEMOCAP, em comparação com as representações latentes dos sinais sem ruído. Em todos os cenários avaliados, verifica-se uma tendência consistente de aumento da distância média entre as representações latentes à medida que a SNR diminui.

As representações do modelo Wav2Vec2 apresentam maior deslocamento médio das suas representações no espaço latente, sugerindo que essas representações foram mais sensíveis à presença de ruído e distorções no sinal de voz. Os valores menores apresentados pelo modelo HuBERT indicam maior estabilidade das representações latentes frente às degradações aplicadas.

De forma complementar, a Tabela I mostra valores próximos a 0 para o coeficiente de silhueta, evidenciando sobreposição de vetores de diferentes classes para ambos os modelos no

espaço dimensional. O Wav2Vec2 mantém coeficientes ligeiramente positivos sob ruído telefônico, indicando preservação parcial da geometria relativa entre classes. Os baixos valores observados de forma geral evidenciam limitada separabilidade emocional no espaço latente.

A Figura 1 apresenta o desempenho geral da classificação de emoções para os modelos HuBERT e Wav2Vec2 utilizando o classificador proposto. Observa-se que, para os sinais originais da base de teste, o modelo HuBERT apresentou maior acurácia não ponderada, UA, atingindo aproximadamente 60%, enquanto o modelo Wav2Vec2 obteve cerca de 53%.

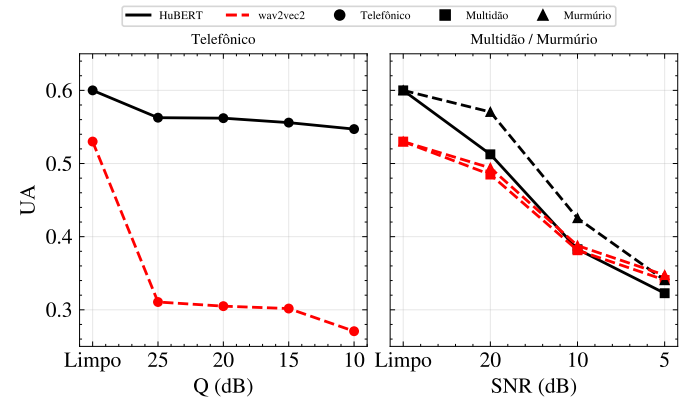


Fig. 1

ACURÁCIA NÃO PONDERADA MÉDIA PARA OS *embeddings* DOS MODELOS HUBERT E WAV2VEC2 SOB DIFERENTES TIPOS E INTENSIDADES DE RUÍDO USANDO O CLASSIFICADOR REGRESSÃO LOGÍSTICA.

Para ambos os modelos, o ruído telefônico apresentado na Figura 1(a) produziu degradação progressiva com o aumento da intensidade. O modelo HuBERT apresentou redução máxima de aproximadamente 5% na UA. Em contrapartida, o modelo Wav2Vec2 sofreu uma redução de aproximadamente 26%, alcançando valores próximos ao desempenho esperado para uma classificação aleatória entre quatro classes (25%).

A Tabela II mostra os valores de *Recall* e F1 para ambos os modelos. Observa-se variabilidade da performance para cada emoção e modelo. Para os áudios sem ruído, a emoção neutra foi a que mostrou-se mais separável para os modelos HuBERT e Wav2Vec2 (com *Recall* 65% e 55%). A Figura 2 apresenta o comportamento da performance por emoção através da métrica *F1-score* para os diferentes tipos de ruído e intensidade. Observa-se uma grande variabilidade da evolução para cada emoção, modelo e ruído.

Sob degradação de canal telefônico, o modelo HuBERT apresentou robustez considerável para as emoções raiva, felicidade e tristeza, com quedas de F1-score de apenas 4%, 5% e 2%, respectivamente, entre a condição limpa e a condição mais severa ($Q = 10$). Para a emoção neutra, a queda foi mais expressiva, atingindo 17%. A análise da matriz de confusão indica que essa degradação decorre principalmente da transferência progressiva dessa emoção para a classe felicidade.

Para o modelo Wav2Vec2, a degradação sob ruído telefônico foi significativamente mais severa, especialmente para as emoções raiva e tristeza, cujos valores de F1-score se aproximaram

TABELA I

DISTÂNCIA EUCLIDIANA MÉDIA E COEFICIENTE DE SILHUETA DAS REPRESENTAÇÕES LATENTES PARA DIFERENTES TIPOS E INTENSIDADES DE RUÍDO.

Ruído	SNR (dB)	MNRU (dB)	Distância Euclidiana		Coeficiente de Silhueta	
			HuBERT	Wav2Vec2	HuBERT	Wav2Vec2
Telefônico	—	25	1,40	2,35	−0,0089	0,0051
Telefônico	—	20	1,45	2,35	−0,0088	0,0051
Telefônico	—	15	1,53	2,39	−0,0093	0,0038
Telefônico	—	10	1,73	2,56	−0,0111	0,0016
Multidão	20	—	0,90	1,41	−0,0125	−0,0110
Multidão	10	—	1,93	2,91	−0,0180	−0,0252
Multidão	5	—	2,69	3,80	−0,0184	−0,0252
Murmúrio	20	—	1,05	0,99	−0,0087	−0,0026
Murmúrio	10	—	2,16	2,18	−0,0090	−0,0036
Murmúrio	5	—	2,70	2,76	−0,0160	−0,0062

TABELA II

DESEMPENHO POR EMOÇÃO EM CONDIÇÃO DE ÁUDIO SEM RUÍDO.

Métrica	Raiva	Felicidade	Neutro	Tristeza
HuBERT Recall	0,58	0,57	0,65	0,60
Wav2Vec2 Recall	0,52	0,54	0,55	0,51
HuBERT F1	0,60	0,59	0,61	0,61
Wav2Vec2 F1	0,50	0,57	0,54	0,49

de zero. Observando-se a matriz de confusão, verifica-se que a emoção raiva passa a ser redistribuída de forma aproximadamente equilibrada entre as classes felicidade e neutro, enquanto 64% das instâncias da classe tristeza são classificadas como felicidade e 35% como neutro. Esse fenômeno também se reflete nas classes felicidade e neutro, que apresentam manutenção ou leve aumento no *Recall*, mas redução no F1-score (quedas de 14% e 13%, respectivamente), evidenciando viés de classificação induzido pelo ruído. Esses resultados indicam que o modelo HuBERT apresenta maior robustez à degradação telefônica, enquanto o Wav2Vec2 demonstra maior fragilidade das fronteiras de decisão, com progressiva perda de desempenho.

Sob ruído de multidão, a degradação foi mais severa para o modelo HuBERT quando comparada ao canal telefônico, com quedas percentuais de F1-score de 9%, 11%, 28% e 32% para as emoções neutro, felicidade, tristeza e raiva, respectivamente. A análise da matriz de confusão mostra que, nessas condições, o modelo tende progressivamente a inferir a classe neutro independentemente da emoção original, indicando colapso parcial da discriminação emocional.

Para o modelo Wav2Vec2, a queda de F1-score entre a condição limpa e o cenário mais degradado (SNR = 5 dB) foi de 6%, 7%, 33% e 39% para as emoções tristeza, neutro, felicidade e raiva, respectivamente. A matriz de confusão revela tendência de concentração das inferências na classe tristeza, o que explica a aparente maior robustez observada especificamente para essa emoção. Esses resultados mostram que o ruído de multidão não apenas reduz o desempenho global, mas induz vieses sistemáticos de classificação, deslocando as inferências para classes emocionalmente menos expressivas, como neutro, ou dominantes no espaço latente.

No caso do ruído de murmúrio, o modelo HuBERT apresentou quedas de F1-score de 14%, 21%, 25% e 42% para as emoções neutro, raiva, felicidade e tristeza, respectivamente. Assim como observado no ruído de multidão, a matriz de confusão indica tendência crescente de transferência das amostras para a classe neutro.

Já para o modelo Wav2Vec2, as quedas foram de 1%, 7%, 18% e 37% para as emoções neutro, felicidade, tristeza e raiva, respectivamente. Nessa condição, observa-se que mais da metade das instâncias de praticamente todas as classes passam a ser classificadas como felicidade, enquanto aproximadamente 20% são classificadas como neutro. Esse comportamento explica a menor degradação aparente nessas duas classes, caracterizando novamente viés de classificação e não necessariamente maior robustez do modelo. Esse comportamento evidencia que ruídos com características espectrais semelhantes à fala, como o murmúrio, não apenas degradam a informação acústica original, mas podem deslocar sistematicamente as representações latentes em direção a emoções específicas.

IV. CONCLUSÕES

Este trabalho apresentou uma análise da robustez de modelos autossupervisionados para reconhecimento de emoções na voz sob diferentes condições de degradação acústica. No nível das representações latentes, o modelo HuBERT mostrou-se mais estável que o Wav2Vec2 para os diferentes cenários avaliados, apresentando menor deslocamento médio em relação à condição limpa. Para ambos os modelos, observou-se aumento progressivo da distância entre as representações degradadas e o sinal original à medida que a severidade do ruído aumentou.

A análise do coeficiente de silhueta indicou baixa separabilidade intrínseca entre as classes no espaço latente, com tendência de maior deterioração sob ruídos mais severos, em especial para multidão e murmúrio. Apesar disso, os resultados mostraram que degradação de desempenho nem sempre está associada exclusivamente à perda de estrutura geométrica das representações, mas também à fragilidade das fronteiras de decisão do classificador de regressão logística.

No nível de desempenho, o modelo HuBERT apresentou maior robustez global em termos da métrica da acurácia não-

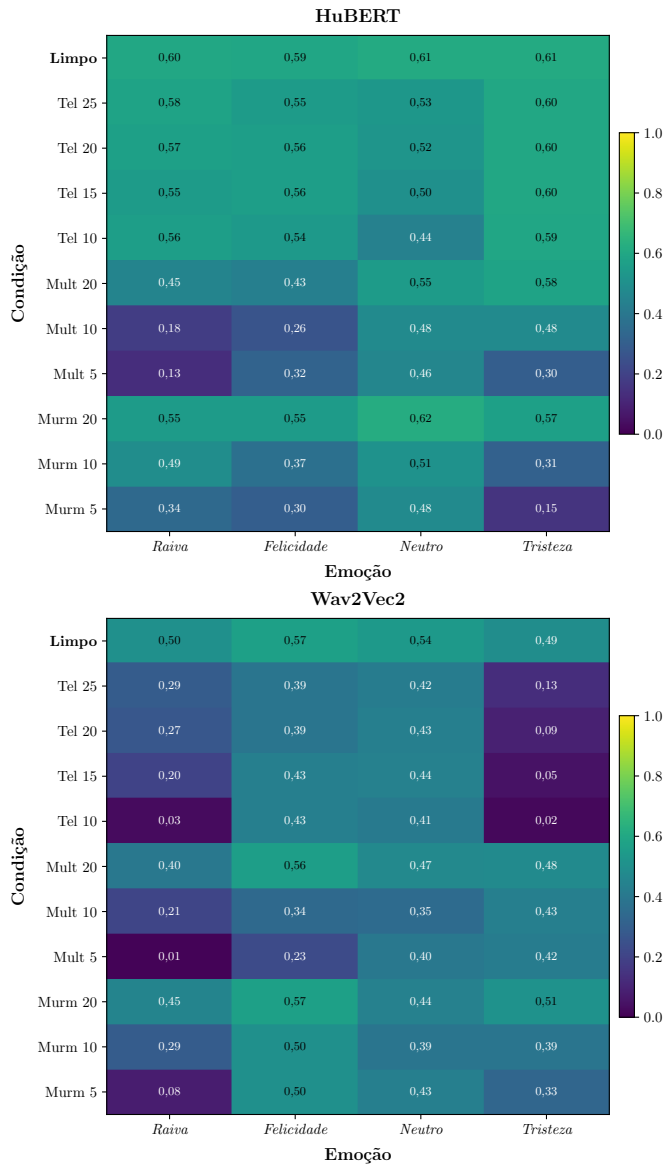


Fig. 2

F1-SCORE PARA OS DIFERENTES TIPOS DE RUÍDO E INTENSIDADES PARA CADA EMOÇÃO.

ponderada, mantendo desempenho superior em praticamente todos os cenários avaliados. Observou-se ainda que o impacto do ruído depende fortemente da emoção analisada. Emoções como raiva e tristeza mostraram elevada sensibilidade a determinados tipos de degradação, enquanto a classe neutra apresentou maior estabilidade. A análise conjunta das métricas de desempenho *Recall* e F1-score evidenciou que aumentos aparentes de desempenho em algumas classes podem refletir viés de classificação induzido pelo ruído, e não necessariamente maior robustez do modelo.

De forma geral, os resultados indicam que modelos autossupervisionados apresentam potencial significativo para a tarefa de reconhecimento de emoções em condições degradadas, mas que sua robustez depende simultaneamente da arquitetura empregada, do tipo de ruído e da emoção considerada. Entretanto,

a análise focou sobre vetores extraídos de determinadas camadas que, à primeira vista, mostraram desempenho superior para o conjunto de treinamento e teste sem ruído. Trabalhos futuros podem explorar como a performance varia para diferentes profundidades e níveis de informação acústica e prosódica associadas a essas representações.

AGRADECIMENTOS

Os autores agradecem ao Programa de Bolsas Itaú (PBI) do Centro de Ciência de Dados (C2D), com apoio do Itaú Unibanco S.A. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 2025/24833-9. O trabalho também foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Processo nº 404081/2023-1.

Disclaimer. As opiniões, conclusões e recomendações expressas neste material são de responsabilidade exclusiva dos autores e não refletem necessariamente o posicionamento do Itaú Unibanco ou do Instituto de Ciência e Tecnologia Itaú (ICTi). Este documento não constitui nem tem a intenção de constituir recomendação de investimento ou prestação de qualquer serviço de investimento. Tampouco deve ser interpretado como oferta de compra ou venda, solicitação de oferta de compra ou venda, ou recomendação para aquisição ou alienação de valores mobiliários ou quaisquer outros instrumentos financeiros. Adicionalmente, todos os dados utilizados neste estudo estão em conformidade com a Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018 – LGPD).

REFERÊNCIAS

- [1] M. B. Akçay and K. Oğuz, "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.
- [2] B. W. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [3] P. H. Pereira, W. Beccaro, and M. A. Ramírez, "Evaluating Robustness to Noise and Compression of Deep Neural Networks for Keyword Spotting," *IEEE Access*, vol. 11, pp. 53 224–53 236, 2023.
- [4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [5] Y. Wang, J. Li, H. Wang, Y. Qian, C. Wang, and Y. Wu, "Wav2vec-Switch: Contrastive Learning from Original-noisy Speech Pairs for Robust Speech Recognition," *arXiv preprint arXiv:2110.04934*, 2021.
- [6] Y. Song, D. Kim, N. Madhu, and H.-G. Kang, "On the Disentanglement and Robustness of Self-Supervised Speech Representations," *International Conference on Electronics, Information, and Communication (ICEIC)*, 2024.
- [7] V. Mitra, V. Kowtha, H.-Y. S. Chien, E. Azemi, and C. Avendano, "Pre-trained Model Representations and their Robustness against Noise for Speech Emotion Analysis," *arXiv preprint arXiv:2303.03177*, 2023.
- [8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2008.
- [10] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [11] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv preprint arXiv:1510.08484*, 2015. [Online]. Available: <https://arxiv.org/abs/1510.08484>