Datasets Reais para Reconhecimento de Emoção em Voz

Pedro Suguino[1], Bianca Scuracchio[2], Neelakshi Joshi[3], Josue J G Ramos[3]

[1]pedrohsuguino123@gmail.com,[2]bianca-cs@hotmail.com, [3]{njoshi, jgramos}@cti.gov.br

[1]Divisão de Sistemas Ciberfísicos - DISCF
CTI/MCTI Renato Archer – Campinas/SP

Abstract. *The objective of this project is related to the use of information in voice to detect the emotion present in the speaker's voice, Speech Emotion Recognition as it is called , which will enable more empathetic interactions in the area of HRI (Human-Robot Interaction). In order to train AI systems to such intent, datasets with examples of these emotions are required, the examples can be real or performed by actors. In the present case, the objective was to raise the existence of real datasets for the Brazilian Portuguese language and this work presents preliminary results of the search for real datasets for Brazilian portuguese.*

Resumo. *O objetivo do projeto está relacionado ao uso de informações em voz para detectar a emoção presente na voz da pessoa, a chamada Reconhecimento de Emoção em Voz, que possibilitará na área de IHR (Interação Humano Robô) interações mais empáticas. Para treinar sistemas de IA para tal intento, são requeridos datasets com exemplares das emoções, sendo que os exemplares podem ser ser reais ou realizados por atores. No presente caso o objetivo foi levantar a existência de datasets reais para a língua portuguesa do Brasil e este trabalho apresenta resultados preliminares da busca de datasets reais para o Português do Brasil.*

1. Introdução

A conexão entre interação humano robô e o reconhecimento de emoções é direta, visto que ela está relacionada ao objetivo da computação afetiva que é [1] levar em consideração as emoções e os "estados de espírito" para a confecção de hardwares e de softwares para interação com as pessoas. A computação afetiva engloba vários campos do conhecimento, como Informática, Educação, Psicologia, Sociologia, Inteligência Artificial, dentre outros para criar instrumentos com condições de interagir com as pessoas e promover interações que levem em consideração os r aspectos emocionais das pessoas para promover diálogos que levem em consideração o canal emocional no diálogo entre as pessoas e a máquina e vice versa, promovendo interações mais aceitáveis para as pessoas.

Nas últimas décadas a área de computação afetiva vem crescendo bastante, com a automatização de vários processos, como assistentes virtuais, surgiu interesse por detecção automática de emoções, que hoje tem várias aplicações reais. Um dos canais por onde pessoas expressam emoções é a voz através do tom, velocidade, intensidade, etc. Essas características são estudadas e processadas no campo de Speech Emotion Recognition (SER), categorizando e identificando emoções na fala.

Na detecção de emoções estão disponíveis modelos contínuos e discretos. O modelo discreto mais conhecido é o Ekman [2] que considerou as seguintes emoções: alegria,

raiva, medo, nojo, surpresa, tristeza, mais o neutro. Esse modelo é considerado transcultural e foi desenvolvido em 1972 a partir das observações de uma tribo em Papua Nova Guiné, que mostrou que estes identificaram o mesmo conjunto de emoções que outros indivíduos de outras culturas. Uma outra forma de representação são os modelos multidimensionais.

No modelo contínuo de Russel [3] que é composto por duas dimensões arousal e valência (afetiva), que podem ter valores positivos e negativos, a projeção dos valores num plano cartesiano dá origem ao modelo circumplexo de emoção de Russel, onde por exemplo (Scielo), com arousal alto e valência positiva tem-se as emoções associadas à Alegria e à Excitação enquanto que em valência negativa estão as emoções associadas ao Medo e à Raiva. Na forma contínua, normalmente são usadas de duas a três dimensões: Ativação, valência e dominância. Ativação diz respeito à energia da emoção, em outras palavras, quão calma ou agitada uma emoção em questão torna uma pessoa. Valência descreve se uma emoção é positiva ou negativa, agradável ou desagradável. E dominância, que não aparece em todas as vezes, se refere a quão dominante ou submissa a pessoa se sente. Em datasets que usam essa forma de rotulação cada enunciado é um ponto nesse espaço bi- ou tridimensional, dessa maneira a classificação é mais precisa, e pode-se dizer mais próxima à realidade, uma vez que estados emocionais não podem ser feitos de somente uma emoção. Dito isso, essa descrição é consideravelmente contraintuitiva, o que, além de dificultar a interpretação da informação encontrada, pode tornar mais difícil a rotulação dos enunciados, uma vez que se pode ter dúvidas e discordâncias quanto a algumas dimensões em um enunciado. Além disso, algumas das emoções reconhecidas por nós podem ser representadas de forma indistinguível, como medo e nojo, e outras podem ser difíceis de categorizar, como surpresa, que pode ter valência positiva ou negativa. [4]

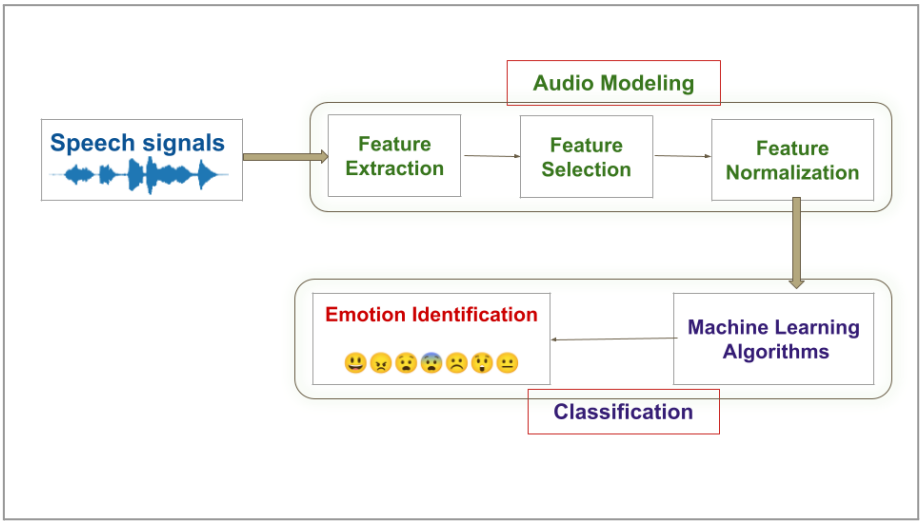

Fig. 1 Diagrama de um Sistema de Reconhecimento de Emoções por voz

Este trabalho apresenta no que segue a caracterização dos diferentes tipos de dados para SER, depois apresenta resultados da busca datasets para SER, apresenta o dataset da CMU e resultados preliminares obtidos com a análise deste e por fim a conclusão dos trabalhos.

2. Emoção na Voz: Tipos de Datasets - do atuado ao real

Segundo Joshi-Lars(2022), como para todo sistema de aprendizado supervisionado é requerido uma grande quantidade de dados para o treinamento de emoções a partir da fala. Os dados de emoção na fala podem ser obtidos de diferentes formas, podendo ser atuados, ou provocados por alguma forma de estímulo, ou naturais. No caso de sons atuados, um ator simula a emoção falando textos pré definidos que são gravados, na maioria das vezes, em condições ideais como pouco ruído, uma só voz falando, entre outros aspectos. No caso de emoções provocadas são obtidas pela indução por meio de estímulos para que a pessoa que fale expresse a emoção em causa. Da mesma forma que nas atuadas, as emoções provocadas são gravadas em condições ideais. A autora observa que o aprendizado dos datasets atuados e provocados são mais fáceis.

De outro lado, segundo Joshi-Lars(2022), os sons reais são mais difíceis de serem obtidos e as fontes usuais podem ser conversas, vídeos na web, shows, enfim todo conjunto de fontes disponíveis no dia a dia que sejam possíveis de serem registrados. As principais dificuldades estão relacionadas a correta rotulagem da emoção embutida na voz, na presença de uma única voz, a presença de ruídos e restrições legais.

Autores, como Plutchik defendem a tese de que há um número de emoções básicas, ele descreve oito: medo, raiva, tristeza, felicidade, nojo, antecipação, surpresa e aprovação. Ele ainda argumenta que emoções são raramente encontradas em sua forma pura na natureza, e que todas as outras emoções são compostas pela combinação das oito emoções básicas. Em SER, geralmente não são usadas todas as oito emoções, as mais recorrentes são raiva, felicidade, tristeza e neutro. Nesse tipo de classificação, cada trecho de áudio é classificado entre alguma dessas emoções. Essa é uma maneira mais simplificada e intuitiva de separação, o que é positivo para aplicações no mundo real, porém como datasets naturais são em geral muito inconsistentes em sua composição, em muitos casos uma maioria dos enunciados acaba caindo em uma categoria, neutro sendo a mais comum, além disso as diferenças entre emoções presentes de estudo para estudo tornam bastante difícil a comparação entre os dados. Outro ponto negativo é que em alguns casos uma análise discreta não possibilita determinar a intensidade de uma emoção sentida, como diz [4], o que leva múltiplas emoções consideravelmente diferentes em sua natureza a serem classificada com o mesmo rótulo, o que ocorre comumente com o rótulo neutro, quando alguma emoção não se mostra intensa o suficiente.

No que segue são apresentados alguns datasets de emoção em voz, em particular o ano em que este foi disponibilizado a língua e a forma de representação da emoção utilizada

Dataset	Ano	Lingua	Emoções/ dimensões
[6]SEMAINE	2011	Inglês	Valência, ativação, poder, antecipação, intensidade
[7]RECOLA	2013	Francês	Valência, ativação
[8]CHEAVD	2016	Chinês	26 emoções não básicas
[9]FAU-Aibo	2008	Alemão	Felicidade, surpresa, empatia, desamparo, irritação, raiva, tédio, reprimido, neutro
[10]SUSAS	1997	Inglês	Nível de estresse e neutro
[11]MSP-Podcast	2019	Inglês	Valência, ativação, dominância. Raiva, tristeza, felicidade, surpresa, medo, nojo, desprezo, neutro
[12]CMU-Mosei	2018	Inglês	Felicidade, tristeza, raiva, medo, nojo, surpresa
[13]CMU-Moseas	2020	Espanhol, português, alemão, francês	Felicidade, tristeza, raiva, medo, nojo, surpresa. Sentimento, subjetividade, atributos
[14]Hebrew Emotional Corpus	2000	Hebraico	Raiva, medo, felicidade, tristeza, nojo
[15]Lee and Narayanan	2003	Inglês	Negativo, não-negativo
[16]Schuller et al.	2003	Inglês, alemão	Raiva, nojo, medo, surpresa, felicidade, tristeza, neutro
[17]Jiang et al.	2005	Mandarim	Felicidade, tristeza, neutro
[18]Morrison et al.	2006	Mandarim	Raiva, neutro
[19]VAM-Audio	2008	Alemão	Valência, ativação, dominância
[20]Hindi Dialect Speech Corpus	2011	Hindi	Raiva, nojo, medo, felicidade, tristeza, neutro
[21]LEGO database	2012	Inglês	Raiva
[22]UUDB	2011	Japonês	Valência, ativação, dominância, credibilidade, interesse, positividade
[23]NNIME	2017	Mandarim	Valência, ativação. Raiva, felicidade, tristeza, frustração, surpresa, neutro

Tabela. 1 Datasets com seu ano, língua e emoções ou dimensões de anotação.

Recentemente datasets que usam um modelo contínuo de rotulação estão tornando-se mais comuns, mas ainda, grande parte desses estão em estudos que contemplam ambas as formas de classificação, uma vez que, como comenta [5], são obtidas informações complementares de cada enunciado, já que ambas as maneiras de classificação têm imperfeições. Há ainda estudos que apresentam outras perspectivas, algumas com um objetivo mais pontual, como níveis de estresse [10], e algumas outras que consideram emoções não prototípicas, ou não básicas [8], chegando a dispor de outras dimensões menos comuns, como sentimento [13] ou credibilidade [22], disponibilizando mais perspectivas.

3. SER em Português Brasileiro

Como pode ser visto na tabela acima, a maior parte dos datasets disponíveis na internet estão em inglês e apesar de haver variedade entre línguas e essa variedade ainda crescer à medida que novos datasets são compilados e lançados, em PB (português brasileiro), ainda há somente dois databases para SER. Sobre eles, é elaborado mais a seguir.

A. VERBO

O Voice Emotion Recognition dataBase in Portuguese Language (VERBO) [24] é um database atuado para SER, consistindo de 1167 frases gravadas por 12 atores. Ele contempla 6 das emoções básicas do modelo de Ekman, seguindo um modelo discreto de classificação.

O VERBO foi o primeiro database de SER para PB, sendo criado em 2018 ele proporcionou um cenário que possibilitou o estudo de SER, que em português brasileiro era uma área relativamente inexplorada. Em consequência, devido à importância da linguagem utilizada em modelos de reconhecimento de emoção no geral, a área teve uma chance melhor de crescer no Brasil.

B. CORAA

O Corpus of Annotated Áudios SER Sentiment Analysis Dataset (CORAA SER) [25] é um dataset de SER construído no projeto TaRSila, que visa aumentar a gama de datasets de áudio para português brasileiro para as áreas como Automatic Speech Recognition (ASR) e Multi-Speaker Text-to-Speech Synthesis (TTS). O dataset usou como base o C-ORAL-BRASIL 1 [26] corpus, um dataset do projeto C-ORAL-BRASIL, e rotulou dados desse para integrar seu dataset.

O CORAA SER version 1.0 é composto de 50 minutos de áudio divididos em três classes para separação de emoções, sendo elas: neutro, feminino não-neutro e masculino não-neutro, o primeiro contém segmentos onde não foram detectadas emoções bem definidas e os demais onde foram detectadas emoções bem definidas.

O dataset foi utilizado no SE&R 2022 Workshop como parte da International Conference on the Computational Processing of Portuguese (PROPOR 2022) como um desafio de reconhecimento de emoção em voz.

Como pode ser constatado, o cenário de datasets de SER em português brasileiro está em crescimento nos últimos anos, porém ainda é limitado em número de horas,

sobretudo em datasets espontâneos, e nesses a pequena diversidade de rótulos e anotações em classificação de emoções, como o COORA SER, que utiliza um modelo pouco distinto de identificação entre as emoções. Porém se expandirmos o escopo para incluir português europeu, temos mais opções.

4. O Dataset CMU-MOSEI

Um dos maiores datasets de emotion recognition disponíveis é o CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [12] que é um dataset espontâneo que disponibiliza texto áudio e vídeo extraídos de segmentos de vídeos de usuários do YouTube. Ele contém anotações tanto de emoções, que utilizam uma escala Likert e um modelo discreto de Ekman para classificação de emoções, quanto de sentimentos, que utilizam uma escala Likert de negatividade-positividade para avaliação dos sentimentos. O dataset compreende mais de 23000 sentenças de 1000 falantes distintos, sendo comparativamente vasto em relação a outros datasets na data em que foi lançado.

Algum tempo depois, nas mesmas linhas do do CMU-MOSEI, a CMU criou também o CMU-Multimodal Opinion Sentiment, Emotions and Attributes (CMU-MOSEAS) [13], compreendendo dessa vez sentenças em espanhol, portugûes, alemão e francês. As frases utilizadas foram novamente adquiridas de vídeos do YouTube. Emoção e sentimento são anotadas de forma similar ao CMU-MOSEI, porém são introduzidos atributos, como dominante, persuasivo e relaxado, entre vários outros, para avaliação. Esse dataset disponibiliza 40000 sentenças, sendo maior que o anterior CMU-MOSEI.

Ambos os datasets são multimodais, deixando disponíveis áudio, texto e vídeo dos trechos coletados do YouTube.

5. Estudos Preliminares com o Dataset CMU-MOSEAS

A CMU disponibiliza para a mais fácil utilização dos seus datasets a CMU-Multimodal SDK. Esse software consiste de uma série de ferramentas para facilitar o pré-processamento de datasets suportados e a construção de modelos neurais, padronizando o alinhamento de datasets multimodais. O Multimodal Data SDK disponibiliza funções para modelar os dados disponibilizados pelos datasets suportados assim como para processamento de outros com dados importados. Junto a ele são disponibilizados os dados não processados dos datasets disponíveis.

Foram realizados estudos preliminares desses datasets, em específico o CMU-MOSEAS. Na linguagem de interesse, que era o português, o MOSEAS proporciona 1000 vídeos com mais de 34 mil frases sendo 10 mil delas anotadas os quais foram gravados com 399 falantes de português europeu.

6. Conclusões

A área de detecção de emoções por voz já é ampla, e os materiais de estudo, diversos. Mais databases naturais aparecem frequentemente, e recentemente muitas delas são

multimodais, além de apresentarem mais formas de rotulação, sendo que rotulação contínua vem se tornando mais comum [2], muitas vezes acompanhada também de uma análise discreta, apresentando informações complementares, que se prova uma escolha mais cara e trabalhosa, por vezes, porém apresenta mais conteúdo e insights, e permite melhor comparação com outros estudos.

7. References

[1] WIKIPÉDIA. **Wikipédia**. 2010. Disponível em: <https://pt.wikipedia.org/wiki/Computa%C3%A7%C3%A3o_afetiva>. Acesso em: 10 dez. 2022.

[2] WIKIPÉDIA. **Wikipédia**. 2008. Disponível em: <https://pt.wikipedia.org/wiki/Paul_Ekman>. Acesso em: 10 dez. 2022.

[3] RAMOS, Danilo; BUENO, José Lino Oliveira. A percepção de emoções em trechos de música ocidental erudita. Per Musi, n. 26, p. 21–30, 2012. Disponível em: <http://dx.doi.org/10.1590/s1517-75992012000200003>.

[2] SHAH FAHAD, Md; RANJAN, Ashish; YADAV, Jainath; *et al*. A survey of speech emotion recognition in natural environment. Digital signal processing, v. 110, n. 102951, p. 102951, 2021. Disponível em: <http://dx.doi.org/10.1016/j.dsp.2020.102951>.

[3] LISETTI, C. L. Affective computing: By Rosalind Picard. Cambridge, Mass,: MIT Press, 1997. Pp. xxii=252. $27.50 cloth. Pattern analysis and applications: PAA, v. 1, n. 1, p. 71–73, 1998. Disponível em: <http://dx.doi.org/10.1007/bf01238028>.

[4] AKÇAY, Mehmet Berkehan; OĞUZ, Kaya. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech communication, v. 116, p. 56–76, 2020. Disponível em: <http://dx.doi.org/10.1016/j.specom.2019.12.001>.

[5] BUSSO, Carlos; BULUT, Murtaza; LEE, Chi-Chun; *et al*. IEMOCAP: interactive emotional dyadic motion capture database. Language resources and evaluation, v. 42, n. 4, p. 335–359, 2008. Disponível em: <http://dx.doi.org/10.1007/s10579-008-9076-6>.

[6] MCKEOWN, G.; VALSTAR, M.; COWIE, R.; *et al*. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing, v. 3, n. 1, p. 5–17, 2012. Disponível em: <http://dx.doi.org/10.1109/t-affc.2011.20>.

[7] RINGEVAL, Fabien; SONDEREGGER, Andreas; SAUER, Juergen; *et al*. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *In*: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). [s.l.]: IEEE, 2013.

[8] LI, Ya; TAO, Jianhua; CHAO, Linlin; *et al*. CHEAVD: a Chinese natural emotional audio–visual database. Journal of ambient intelligence and humanized computing, v. 8, n. 6, p. 913–924, 2017. Disponível em: <http://dx.doi.org/10.1007/s12652-016-0406-z>.

[9] STEIDL, Stefan. Automatic classification of emotion-related user states in spontaneous children's speech. [s.l.]: Logos Verlag Berlin, 2009.

[10] HANSEN, John H. L.; SAHAR, E. Getting started with SUSAS: a speech under simulated and actual stress database. [s.l.: s.n.], 1997.

[11] LOTFIAN, Reza; BUSSO, Carlos. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE transactions on affective computing, v. 10, n. 4, p. 471–483, 2019. Disponível em: <http://dx.doi.org/10.1109/taffc.2017.2736999>.

[12] ZADEH, Amir. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. ACL, 2018.

[13] ZADEH, Bagher. MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, v. 2020, p. 1801–1812, 2020.

[14] AMIR, Noam. Analysis of an emotional speech corpus in Hebrew based on objective criteria This. [s.l.: s.n.], 2000.

[15] LEE, C. M.; NARAYANAN, S. Emotion recognition using a datadriven fuzzy inference system. *In*: European conference on speech and language processing (EUROSPEECH). Geneva, Switzerland: [s.n.], 2003, p. 157–160.

[16] SCHULLER, B.; RIGOLL, G.; LANG, M. Hidden Markov model-based speech emotion recognition. *In*: 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698). [s.l.]: IEEE, 2003.

[17] JIANG, Dan-Ning; ZHANG, Wei; SHEN, Li-Qin; *et al*. Prosody analysis and modeling for emotional speech synthesis. *In*: Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. [s.l.]: IEEE, 2006.

[18] MORRISON, Donn; WANG, Ruili; DE SILVA, Liyanage C. Ensemble methods for spoken emotion recognition in call-centres. Speech communication, v. 49, n. 2, p. 98–112, 2007. Disponível em: <http://dx.doi.org/10.1016/j.specom.2006.11.004>.

[19] GRIMM, Michael; KROSCHEL, Kristian; NARAYANAN, Shrikanth. The Vera am Mittag German audio-visual emotional speech database. *In*: 2008 IEEE International Conference on Multimedia and Expo. [s.l.]: IEEE, 2008.

[20] RAO, K.; SHASHIDHAR, G. Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech. [s.l.: s.n.], 2013.

[21] SCHMITT, Alexander. A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. [s.l.: s.n.], 2012.

[22] MORI, Hiroki; SATAKE, Tomoyuki; NAKAMURA, Makoto; *et al*. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. Speech communication, v. 53, n. 1, p. 36–50, 2011. Disponível em: <http://dx.doi.org/10.1016/j.specom.2010.08.002>.

[23] CHOU, Huang-Cheng; LIN, Wei-Cheng; CHANG, Lien-Chiang; *et al*. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. *In*: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). [s.l.]: IEEE, 2017.

[24] TORRES NETO, José R.; FILHO, Geraldo P. R.; MANO, Leandro Y.; *et al*. VERBO: Voice emotion recognition dataBase in Portuguese language. Journal of computer science, v. 14, n. 11, p. 1420–1430, 2018. Disponível em: <http://dx.doi.org/10.3844/jcssp.2018.1420.1430>.

[25] TaRSila. **GitHub**. 2022. Emotion Recognition from Brazilian Portuguese Informal Spontaneous Speech. Disponível em: https://github.com/rmarcacini/ser-coraa-pt-br. Acesso em: 30 ago. 2022.

[26] RASO, Tommaso; MELLO, Heliana. The C-ORAL-BRASIL I: Reference corpus for informal spoken Brazilian Portuguese. *In*: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, p. 362–367.