Reconhecimento de Emoções em Sinais de Fala Usando Transferência de Aprendizado

Sergio Pinto Gomes Junior, Flávio Rainho Ávila e Michel Pompeu Tcheou

***Resumo*— Neste artigo, propõe-se um sistema de reconhecimento de emoções em sinais de fala usando redes profundas convolucionais com as técnicas de transferência de aprendizado e aumento de base de dados. A transferência de aprendizado foi realizada a partir de uma rede residual (ResNet) de 34 camadas treinada para a base ImageNet. O aumento da base foi realizada alterando-se o *pitch* e alargando-se no tempo as amostras de sinais de fala das classes de felicidade, tristeza, raiva e neutra da base IEMOCAP (*Interactive Emotional Dyadic Motion Capture*). O sistema proposto foi capaz de classificar corretamente 81,26% das amostras.**

***Palavras-Chave*— Reconhecimento de emoções, Processamento da Fala, Redes Neurais Convolucionais, Transferência de Aprendizado.**

***Abstract*— In this article, we propose a system of speech emotion recognition using deep convolutional networks with transfer learning and data augmentation. The transfer learning was performed from a 34-layer residual network (ResNet) trained for the ImageNet database. The data augmentation was performed by altering the pitch and time-stretching the speech signal samples of the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database for the happiness, sadness, anger and neutral classes. The proposed system was able to correctly classify 81.26% of samples.**

***Keywords*— Emotion Recognition, Speech Processing, Convolutional Neural Networks, Transfer Learning.**

I. Introdução

A fala tem se tornado um meio de interação entre o ser humano e os computadores cada vez mais importante, dado que ela é o meio mais rápido e natural de comunicação. Este advento é consequente ao grande enfoque dado desde o final dos anos cinquenta à pesquisa relativa ao reconhecimento automático de fala por máquinas, o qual busca converter um discurso humano em uma sequência de palavras [1]. Podemos perceber a evolução dessa área por meio do aumento da complexidade dos assistentes virtuais dos sistemas operacionais que utilizamos. Entretanto, nenhum desses assistentes consegue se comunicar com tanta naturalidade quanto um ser humano, pois ainda não possui a habilidade de compreender as emoções do falante ou usuário. A busca por essa naturalidade proporcionou a criação de um campo de pesquisa relativamente recente, o reconhecimento de emoções na fala, entendido como a classificação automática do estado emocional do falante através do sinal proveniente de sua fala [1].

Sergio Gomes Junior, Flávio Ávila, Michel Tcheou, Programa de Pós-Graduação em Engenharia Eletrônica (PEL), Universidade do Estado do Rio de Janeiro (UERJ), e-mail: sergiopgjunior@gmail.com, flavio.avila@uerj.br, mtcheou@uerj.br.

A maneira convencional de criar um mecanismo capaz de reconhecer a emoção da fala pode ser dividida em quatro principais etapas. Primeiramente, precisamos definir um modelo adequado de representação de emoções [2]. Dois modelos são geralmente encontrados na prática. O primeiro modelo é de classes discretas, como as seis categorias de emoções de Ekman, incluindo raiva, desgosto, medo, felicidade, tristeza e neutra. Já o segundo modelo possui uma abordagem de dimensão de valor contínuo e é formado por dois eixos: o eixo de ativação, caracterizado por uma escala bipolar como calmo/excitado; e o eixo de valência, que é caracterizado por uma escala bipolar como positivo/negativo [4].

Uma vez definido o modelo de representação das emoções, é necessário a aquisição de dados e rotulá-los de forma adequada. Existem diversas bases de dados que foram criadas para a pesquisa nessa área. O grande desafio nessa etapa se dá pela subjetividade e a incerteza dos rótulos. Inclusive os seres humanos geralmente discordam em algum grau sobre qual emoção está presente no discurso de outras pessoas [5]. Em algumas bases de dados, o desempenho do reconhecimento humano atinge na média apenas 65% de acerto [6].

Outra etapa importante do processo de reconhecimento de emoções de fala é a extração de características dos sinais que sejam capazes de refletir de maneira fidedigna o conteúdo emocional. Os parâmetros de fala podem ser agrupados em quatro categorias: contínuos, qualitativos, espectrais e parâmetros baseados no operador de energia Teager [1].

A última etapa consiste no método para classificação das emoções baseadas nos parâmetros extraídos da fala. Várias técnicas de aprendizagem de máquina vêm sendo utilizadas para esta tarefa, tais como o *Gaussian Mixture Model* (GMM) [7]–[9], as redes neurais artificiais (RNA) [10]–[12], e, recentemente, as redes neurais profundas, como a *Convolutional Neural Network* (CNN) [15] e a *Recurrent Neural Network* (RNN) [16].

Em literatura recente, é possível verificar o uso de redes neurais profundas aplicadas ao reconhecimento de emoções em fala e também uma tentativa de criar sistemas fim-a-fim, onde os parâmetros da fala são extraídos e selecionados automaticamente pelo próprio sistema. Em [22], foi proposto um sistema baseado em redes neurais recorrentes que obteve uma acurácia geral de 63,9%. Já em [17], os pesquisadores desenvolveram um sistema de reconhecimento de emoções de fala fim-a-fim, formado por uma combinação de uma rede neural convolucional e uma rede LSTM (*Long Short-Term Memory*), alimentado por espectrogramas das amostras de fala, que atingiu uma acurácia de 68%. Ambos os artigos usam a base de dados *Interactive Emotional Dyadic Motion Capture*

(IEMOCAP), mais especificamente as classes de felicidade, tristeza, raiva e neutra.

O presente trabalho propõe o emprego de uma rede profunda convolucional em conjunto com as técnicas de transferência de aprendizado e aumento da base de dados para a tarefa de reconhecimento de emoções em sinais de fala. A transferência de aprendizado foi realizada a partir de uma ResNet de 34 camadas treinada para a base ImageNet. O aumento da base foi realizada alterando-se o *pitch* e alargando-se no tempo os sinais de fala das classes de felicidade, tristeza, raiva e neutra da base IEMOCAP.

Na Seção II, são apresentados detalhes a respeito da base de dados IEMOCAP, utilizada nos experimentos realizados neste trabalho. A Seção III apresenta a configuração dos experimentos utilizando a CNN. Os resultados dos experimentos são apresentados e discutidos na Seção IV. Por fim, a Seção V apresenta as conclusões deste trabalho.

II. Base de Dados IEMOCAP

A base IEMOCAP (*Interactive Emotional Dyadic Motion Capture Database*) contém dados audiovisuais, e foi desenvolvida por pesquisadores da *University of Southern California* (UCS) cujo objetivo principal era criar uma base de dados de sinais de fala por meio de muitos indivíduos capazes de expressar emoções genuínas [18]. Os autores perceberem que uma das maiores limitações da área de estudo das expressões das emoções é a falta de bases de dados com interações genuínas, sem roteiro e contexto específicos [18].

Para alcançar esse objetivo, o conteúdo da base foi cuidadosamente selecionado. Portanto, os atores foram solicitados a trabalhar com duas abordagens [18]. Na primeira, foi proposto aos participantes a memorização e o ensaio de roteiros. O uso de roteiros fornece uma maneira de restringir o conteúdo semântico e emocional da base. Uma vez que essas emoções são expressas dentro de um contexto adequado, elas são mais propensas a serem transmitidas de uma maneira genuína, em comparação com gravações de frases isoladas. Na segunda abordagem, os atores improvisaram com base em cenários hipotéticos projetados para provocar emoções específicas. Os tópicos para os cenários espontâneos foram selecionados seguindo as orientações fornecidas por [19]. Indivíduos que foram solicitados a lembrar de situações no passado que lhes provocaram determinadas emoções. Os cenários hipotéticos foram baseados em algumas situações comuns como perda de um amigo, separação, etc. Nesse cenário, os sujeitos estavam livres para usar suas próprias palavras para se expressarem. Ao conceder aos atores liberdade de expressão de suas emoções, foi possível induzir uma genuína percepção das emoções. A base de dados contou com dez atores. Ao todo, cinco homens e cinco mulheres foram escolhidos após uma audição. Os indivíduos gravaram, em duplas formadas por um homem e uma mulher cinco sessões. Após isto, os diálogos foram segmentados por turnos, onde cada turno foi definido como segmento contínuo por ator.

Na maioria das bases de dados, os atores são solicitados a enunciar uma frase expressando uma determinada emoção a qual é utilizada posteriormente como rótulo para essa enunciação [1]. Uma desvantagem desse método é não garantir que a expressão oral gravada reflita a emoção alvo. Para evitar esse problema, a IEMOCAP foi rotulada com uma combinação de avaliações subjetivas. Para tal, alguns alunos da USC avaliaram o conteúdo emocional dos turnos gravados utilizando um esquema de anotações discretas baseadas em categorias e anotações baseadas em atributos contínuos. No esquema discreto, os avaliadores utilizaram categorias, tais como felicidade, tristeza, etc. Já para o esquema contínuo, as gravações foram avaliadas com os atributos de ativação, valência e dominância. A etapa de avaliação do conteúdo emocional em anotações discretas baseadas em categorias contou com seis avaliadores. Cada turno foi classificado por três pessoas diferentes. Por questões de simplicidade, foi adotado o método de voto majoritário para a atribuição do rótulo ao turno, se a categoria com o maior número de votos fosse única.

Em um primeiro momento, os autores da base de dados definiram que ela seria dividida em quatro classes, onde cada classe representaria as emoções presentes nas encenações dos turnos, que são: raiva, tristeza, felicidade, frustração e um estado emocional neutro. Entretanto, devido a naturalidade com a qual os atores desenvolveram as representações, os autores adicionaram às classes inicialmente definidas as emoções de desgosto, de medo, de surpresa e de animação.

III. Reconhecimento de emoções com Redes Profundas Convolucionais

Com base nos avanços científicos mais recentes e relevantes da área [2], propõe-se um sistema de reconhecimento de emoções em sinais de fala, ilustrado na Figura 1. Inicialmente, geram-se imagens de espectrogramas a partir dos sinais da base IEMOCAP. A forma de um espectrograma pode ser associada com a forma do trato vocal, a qual varia conforme o estado emocional [25]. Baseado nessa associação, os espectrogramas das amostras de fala foram utilizados neste trabalho, visto que eles possibilitam a utilização de uma ferramenta de processamento de imagens para um problema de processamento de fala. Essas imagens são divididas em conjuntos de treinamento (70%) e validação (30%) aos quais aplica-se uma rede neural convolucional (do Inglês, *Convolutional Neural Network* - CNN), responsável pelo reconhecimento da emoção contida na fala. Utiliza-se neste trabalho a rede convolucional ResNet (*Residual Network*) com 34 camadas [24].

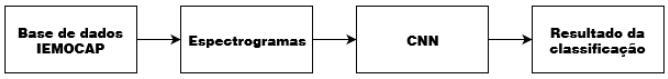

Fig. 1. Diagrama do sistema de reconhecimento proposto

As redes convolucionais são amplamente utilizadas na área de visão computacional, sendo atualmente o estado da arte em tarefas diversas como a classificação de imagens, a segmentação semântica, a transferência de estilo, a detecção de objetos, o reconhecimento facial, entre outras. A ideia é explorar o sucesso dessa arquitetura para a aplicação de reconhecimento de emoções em fala, fazendo as devidas

adaptações. A primeira delas é produzir uma imagem que represente o sinal de fala. Nessa caso, adota-se a imagem de um espectrograma na escala perceptiva *mel*, designada como *mel-espectrograma* [25]. A segunda é realizar a transferência de aprendizado a partir de uma ResNet ajustada para classificação de imagens da base ImageNet [24]. A terceira é o aumento da base de dados, alterando-se o *pitch* e alargando-se o tempo dos sinais de fala da IEMOCAP.

O sistema foi desenvolvido na linguagem de programação *Python* com o auxílio da biblioteca *Fast.ai* [23], que por sua vez, faz uso de técnicas utilizadas no treinamento da rede como a Descida do Gradiente Estocástico com Reinicializações (*Stochastic Gradient Descent with restarts* - SGDR) [20] e o *Dropout* [21].

A. Aumento da base de dados

Dado que a CNN em questão, ResNet, necessita de uma grande quantidade de dados para a etapa de treinamento, decidimos realizar testes também com o aumento da base. Para isso, foram alteradas algumas características das amostras de áudio originais, e os resultados dessas modificações foram acrescentados à base. A primeira modificação foi realizada com o auxílio de um algoritmo de deslocamento de *pitch* [26], onde cada amostra de áudio teve seu *pitch* deslocado de 4 semitons. Esta alteração tornou os sons mais agudos. Na segunda modificação, com o auxílio de um algoritmo de alargamento do tempo [26], a escala do tempo das amostras de áudio originais foram modificadas com um fator de alargamento igual a 1,5. Por último, criamos novas amostras de áudio, extraindo o silêncio dos arquivos originais [27], utilizando novamente o detector de atividade vocal com um limiar de 23 dB. A Tabela I exibe a quantidade de amostras das bases original e aumentada, separadas pelas classes consideradas neste trabalho. Observe que a base aumentou em quatro vezes.

TABELA I
QUANTIDADE DE AMOSTRAS POR CLASSE DA BASE DE DADOS IEMOCAP AUMENTADA

	Quantidade de amostras	
Classe	Original	Aumentada
Felicidade	595	2380
Tristeza	1084	4336
Raiva	1103	4412
Neutra	1708	6832

B. Taxa de aprendizado

Neste trabalho, foram utilizadas taxas de aprendizado diferenciais, um procedimento onde as camadas mais altas da rede variam mais do que camadas mais profundas durante o treinamento [28]. A criação de modelos de aprendizagem profunda a partir de arquiteturas pré-existentes permite a realização de tarefas de visão computacional de forma mais eficiente. A maioria dessas arquiteturas é treinada no *ImageNet* e, dependendo da similaridade de seus dados com as imagens no *ImageNet*, esses pesos precisarão ser alterados com maior ou menor intensidade. Quando se trata de modificar esses pesos, as últimas camadas do modelo geralmente precisam de mais alterações, enquanto os níveis mais profundos que já estão bem treinados para detectar recursos básicos, como bordas e contornos,e nesse caso, precisarão de menos alterações.

O procedimento adotado para definir a taxa de aprendizado consiste em iniciar o treinamento da rede com uma taxa muito baixa, por exemplo 1^{-8}, e, em seguida, aplicar o método do gradiente descendente uma única iteração e anotar o valor da função custo resultante. Após isso, repete-se o procedimento um determinado número de vezes sempre usando uma taxa de aprendizado duas vezes maior que a anterior. Então plota-se o valor da função custo em função da taxa de aprendizado. A curva resultante é usada para guiar a escolha da taxa de aprendizado a ser usada para o treinamento completo. Diversos critérios podem ser adotados, mas a recomendação é escolher aquela taxa a partir do qual a função custo começa a decair com maior intensidade. Isso não é o mesmo que escolher o ponto mínimo da curva, e o motivo é que esse ponto mínimo corresponde a um valor a partir do qual a função custo começa a aumentar, não sendo portanto a taxa ideal para o treinamento. Atribui-se a última camada o valor encontrado e, seguindo as boas práticas, define-se cada taxa de aprendizado como 10 vezes menor do que a posterior.

C. Transferência de aprendizado

Uma rede neural profunda com muitas camadas e grande quantidade de pesos em geral requer uma quantidade elevada de dados de treinamento. Para lidar com esse desafio, lança-se mão da transferência de aprendizado, que consiste no uso de uma rede pré-treinada em uma base grande, idealmente associada a um problema similar, e na adaptação algumas camadas à base de interesse. Normalmente, as primeiras camadas, relacionadas aos atributos mais fundamentais e gerais, se mantêm inalteradas quando treinadas em bases similares.

Sendo assim, foram realizados novos testes utilizando a transferência de aprendizado com os valores dos pesos definidos para a ResNet de 34 camadas, treinada para a base de dados ImageNet, considerando um problema de classificação de imagens [24]. Ainda que o problema em questão difira do de reconhecimento de imagens, observamos empiricamente que os atributos aprendidos são úteis para o problema de reconhecimento de emoções através da fala.

Neste trabalho, são avaliados dois modelos de transferência de aprendizado. Em um primeiro momento, testou-se a rede após o treinamento apenas da camada totalmente conectada, responsável pela classificação propriamente dita, e mantendo fixos os valores dos pesos das camadas anteriores. Em um segundo momento, a rede foi treinada por completa, com os pesos inicializados com os valores do pré-treino, porém usando taxas de aprendizado diferentes para cada camada, de forma que as camadas finais fossem modificadas mais intensamente do que as iniciais.

IV. RESULTADOS

No primeiro experimento computacional, a ResNet foi treinada com os pesos inicializados de forma aleatória, e foi

gerada a curva de custo por taxa de aprendizado, que pode ser vista na Figura 2. A partir dessa curva, definiu-se a taxa de aprendizado da rede como 10^{-2}, seguindo procedimento detalhado na Seção III.B. Em seguida, a rede foi treinada e validada com essa taxa em quinze épocas. O resultado desta etapa está apresentado na Tabela II, onde pode-se verificar que a acurácia geral da rede não variou muito entre as épocas, e seu valor final foi de 0,598608. Além disso, o resultado do custo de validação apresenta *overtraining* desde o início do treinamento desta rede. Essa situação ocorre devido a complexidade da rede e o pequeno número de amostras da base.

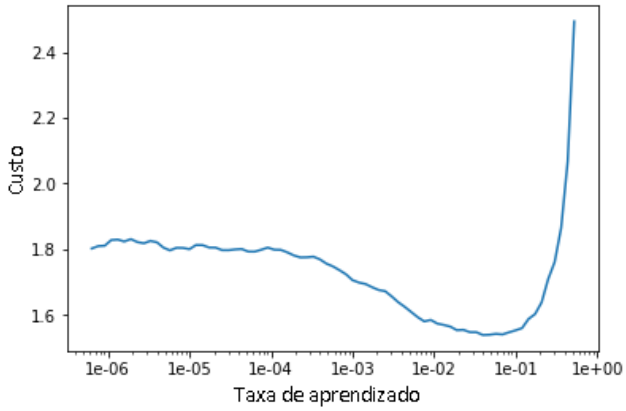

Fig. 2. Custo em função da taxa de aprendizado da rede ResNet com pesos inicializados aleatoriamente

TABELA II
RESULTADO DA CLASSIFICAÇÃO, INICIALIZANDO-SE OS PESOS ALEATORIAMENTE, COM A TAXA DE APRENDIZADO 10^{-2}

Época	Custo de Treino	Custo de validação	Acurácia
1	1,286629	1,027866	0,590487
2	1,088441	1,033717	0,589327
3	0,967714	1,044087	0,605568
4	0,911319	1,188079	0,596288
5	0,853787	1,182761	0,609049
6	0,797502	1,080468	0,604408
7	0,700675	1,098917	0,588167
8	0,616391	1,178904	0,569606
9	0,511176	1,250570	0,598608
10	0,423265	1,456313	0,588167
11	0,329858	1,369980	0,597448
12	0,257469	1,426652	0,585847
13	0,203169	1,504504	0,596288
14	0,159058	1,507039	0,598608
15	0,137549	1,505375	0,598608

No segundo experimento computacional, foram utilizadas a base de dados IEMOCAP aumentada aplicada a uma ResNet pré-treinada com os dados da base ImageNet. Em seguida, definiu-se a taxa de aprendizado da rede como 10^{-2} por meio da análise da curva que relaciona essa taxa com o custo da rede pré-treinada exibida na Figura 3. De posse desse valor, a camada *Fully connected* foi treinada em cinco épocas. Em seguida, realizou-se processo de validação. Os resultados das etapas de treinamento e de validação podem ser verificados na Tabela III. Observa-se que na quinta época alcança-se uma acurácia de 0,671492.

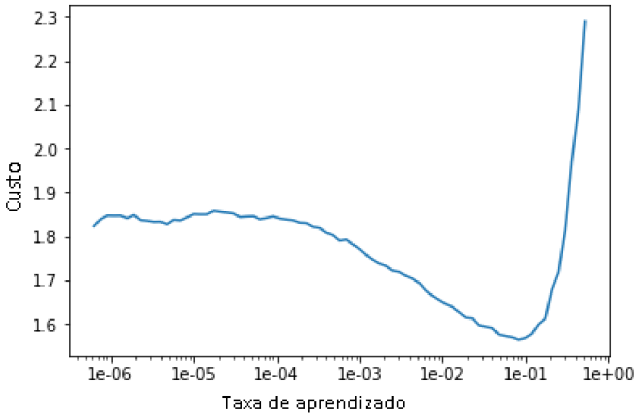

Fig. 3. Custo em função da taxa de aprendizado da rede ResNet pré-treinada com os dados da base ImageNet aplicada a base de dados IEMOCAP aumentada

TABELA III
RESULTADO DA CLASSIFICAÇÃO DA REDE RESNET PRÉ-TREINADA COM OS DADOS DA BASE IMAGENET APLICADA A BASE DE DADOS IEMOCAP, COM A TAXA DE APRENDIZADO IGUAL A 10^{-2} AUMENTADA

Época	Custo de treino	Custo de validação	Acurácia
1	1,201774	1,545130	0,483296
2	1,352582	1,126734	0,553174
3	1,144411	1,012508	0,619432
4	0,976383	0,899688	0,643653
5	0,873673	0,836139	0,671492

No terceiro experimento computacional, foi realizado o treinamento da rede por completo com a base IEMOCAP aumentada, inicializando-a com a rede ResNet pré-treinada com a ImageNet. A seguir, a partir da curva de custo versus taxa de aprendizado na Figura 4, define-se taxa como 10^{-7}. Após essa etapa, a rede foi treinada e validada. Os valores para o custo de treinamento, validação e acurácia da rede podem ser vistos na Tabela IV. Na décima época, alcança-se uma acurácia de 0,812639 na classificação, demonstrando a eficácia do método de aumento da base de dados e da transferência de aprendizado.

Com o objetivo de avaliar as classes de maneira separada, a matriz de confusão dos testes é apresentada na Tabela V.

V. CONCLUSÕES

Neste trabalho foi investigado o uso das técnicas de transferência de aprendizado e aumento da base de dados aplicadas a uma rede neural convolucional para o problema de reconhecimento de emoções a partir da fala. Restringiram os procedimentos de treinamento e validação às classes de felicidade, neutra, raiva e tristeza contidas na base de dados IEMOCAP. O sistema proposto neste trabalho superou o resultado apresentado em [17] em 13 pontos percentuais.

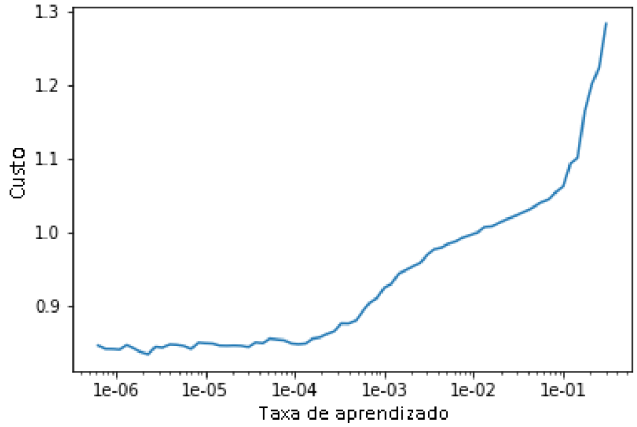

Fig. 4. Custo em função da taxa de aprendizado com a rede treinada por completo com a base IEMOCAP aumentada, inicializando-a com a rede ResNet pré-treinada com a ImageNet

TABELA IV

RESULTADO DA CLASSIFICAÇÃO DA REDE TREINADA POR COMPLETO COM A BASE IEMOCAP AUMENTADA, INICIALIZANDO-A COM A REDE RESNET PRÉ-TREINADA COM A IMAGENET, COM A TAXA DE APRENDIZADO IGUAL A 10^{-7}

Época	Custo de treino	Custo de validação	Acurácia
1	0,878089	0,836610	0,669543
2	0,891139	1,045950	0,588530
3	0,905717	1,353459	0,586860
4	0,809399	0,880154	0,692929
5	0,703324	0,818643	0,704343
6	0,554468	0,702610	0,739699
7	0,382100	0,650087	0,771715
8	0,210504	0,705921	0,791481
9	0,107573	0,711677	0,811804
10	0,075634	0,740088	0,812639

TABELA V

MATRIZ DE CONFUSÃO

		Classe Real			
		Felicidade	Neutra	Raiva	Tristeza
Classe Prevista	Felicidade	63,00%	4,57%	1,65%	2,53%
	Neutra	23,89%	82,07%	10,35%	9,32%
	Raiva	5,71%	5,79%	84,82%	1,73%
	Tristeza	7,40%	7,57%	3,18%	86,42%
Geral		**81,26%**			

REFERÊNCIAS

[1] EL AYADI, Moataz; KAMEL, Mohamed S.; KARRAY, Fakhri. "Survey on speech emotion recognition: Features, classification schemes, and databases". *Pattern Recognition*, v. 44, n. 3, p. 572-587, 2011.

[2] SCHULLER, Björn W. "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". *Communications of the ACM*, v. 61, n. 5, p. 90-99, 2018.

[3] EKMAN, Paul; SORENSON, E. Richard; FRIESEN, Wallace V. "Pan-cultural elements in facial displays of emotion". *Science*, v. 164, n. 3875, p. 86-88, 1969.

[4] BRADLEY, Margaret M.; LANG, Peter J. "Measuring emotion: the self-assessment manikin and the semantic differential". *Journal of behavior therapy and experimental psychiatry*, v. 25, n. 1, p. 49-59, 1994.

[5] DEVILLERS, Laurence; VIDRASCU, Laurence; LAMEL, Lori. "Challenges in real-life emotion annotation and machine learning based detection". *Neural Networks*, v. 18, n. 4, p. 407-422, 2005.

[6] NWE, Tin Lay; FOO, Say Wei; DE SILVA, Liyanage C. "Speech emotion recognition using hidden Markov models". *Speech communication*, v. 41, n. 4, p. 603-623, 2003.

[7] BREAZEAL, Cynthia; ARYANANDA, Lijin. "Recognition of affective communicative intent in robot-directed speech". *Autonomous robots*, v. 12, n. 1, p. 83-104, 2002.

[8] SLANEY, Malcolm; MCROBERTS, Gerald. "Baby ears: a recognition system for affective vocalizations."*Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. IEEE, 1998. p. 985-988.

[9] SCHULLER, Björn. "Towards intuitive speech interaction by the integration of emotional aspects". *IEEE International Conference on Systems, Man and Cybernetics. IEEE*, 2002. p. 6 pp. vol. 6.

[10] NICHOLSON, Joy; TAKAHASHI, Kazuhiko; NAKATSU, Ryohei. "Emotion recognition in speech using neural networks". *Neural computing and applications*, v. 9, n. 4, p. 290-296, 2000.

[11] PETRUSHIN, Valery A. Emotion recognition in speech signal: experimental study, development, and application. In: Sixth International Conference on Spoken Language Processing. 2000.

[12] HENDY, Nermine Ahmed; FARAG, Hania. "Emotion recognition using neural network: A comparative study."*Proceedings of World Academy of Science, Engineering and Technology. World Academy of Science, Engineering and Technology (WASET)*, 2013. p. 791.

[13] BURKHARDT, Felix et al. "A database of German emotional speech". *Ninth European Conference on Speech Communication and Technology*. 2005.

[14] SHAMI, Mohammad T.; KAMEL, Mohamed S. Segment-based approach to the recognition of emotions in speech. In: 2005 IEEE International Conference on Multimedia and Expo. IEEE, 2005. p. 4 pp.

[15] ZHANG, Shiqing et al. "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching". *IEEE Transactions on Multimedia*, v. 20, n. 6, p. 1576-1590, 2018.

[16] TZIRAKIS, Panagiotis et al. "End-to-end multimodal emotion recognition using deep neural networks". *IEEE Journal of Selected Topics in Signal Processing*, v. 11, n. 8, p. 1301-1309, 2017.

[17] SATT, Aharon; ROZENBERG, Shai; HOORY, Ron. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms". *INTERSPEECH*. 2017. p. 1089-1093.

[18] BUSSO, Carlos et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, v. 42, n. 4, p. 335, 2008.

[19] WALLBOTT, Harald F. et al. "Experiencing emotion: A cross-cultural study". *Cambridge University Press*, 1986.

[20] LOSHCHILOV, Ilya; HUTTER, Frank. SGDR: STOCHASTIC GRADIENT DESCENT WITH WARM RESTARTS. Learning, v. 10, p. 3.

[21] SRIVASTAVA, Nitish et al. "Dropout: a simple way to prevent neural networks from overfitting". *The Journal of Machine Learning Research*, v. 15, n. 1, p. 1929-1958, 2014.

[22] LEE, Jinkyu; TASHEV, Ivan. High-level feature representation using recurrent neural network for speech emotion recognition. In: Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[23] HOWARD, Jeremy et al. "Fast.ai", *https://github.com/fastai/fastai*, 2018

[24] HE, Kaiming et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770-778, 2016.

[25] RABINER, Lawrence R. et al. Introduction to digital speech processing. Foundations and Trends in Signal Processing, v. 1, n. 1–2, p. 1-194, 2007.

[26] DRIEDGER, Jonathan; MÜLLER, Meinard. A review of time-scale modification of music signals. Applied Sciences, v. 6, n. 2, p. 57, 2016.

[27] SOHN, Jongseo; KIM, Nam Soo; SUNG, Wonyong. A statistical model-based voice activity detection. IEEE signal processing letters, v. 6, n. 1, p. 1-3, 1999.

[28] HOWARD, Jeremy; RUDER, Sebastian. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, 2018.