CLASSIFICAÇÃO DE EMOÇÕES COM BASE EM CARACTERÍSTICAS DA VOZ UTILIZANDO MÁQUINA DE VETOR DE SUPORTE

G. Jablonski* e T. F. Zaruz**

*Graduando em Engenharia da Computação, UFU
**Graduanda em Engenharia Biomédica, UFU
e-mail: gabriel.jablonski.ufu@gmail.com

Resumo: A voz é um dos principais meios de interação humana, e através dela, emoções podem ser transmitidas. O reconhecimento delas por máquinas se torna cada vez mais importante no contexto tecnológico moderno que presenciamos. Para isso, são utilizadas, dentre outras técnicas de *machine learning*, as *support vector machines (SVMs*), a fim de realizar o treinamento e validação de classificadores. Neste trabalho, um classificador de emoções, baseado em uma *SVM*, foi treinado e validado com gravações de voz em inglês da base de dados *RAVDESS*. Os resultados mostram que o *SVM* foi eficiente para classificar e identificar as emoções escolhidas, com uma acurácia média de 80%, apresentando resultados semelhantes quando comparado com outros trabalhos, ou até mesmo com a identificação humana.
Palavras-chave: Voz, Reconhecimento de emoções, SVM.

Abstract: *The voice is one of the main ways of human interaction, and through it, emotions may be conveyed. Emotion recognition by machines is of increasing importance in the modern technological environment we observe. To accomplish this task, among other machine learning techniques, support vector machines (SVM) can be used in order to train and validate classifiers. In the present work, an SVM based emotion classifier was trained and validated based on voice recordings in English from the RAVDESS database. Results indicate that the SVM is efficient to classify and identify the chosen emotions, with an average accuracy of 80%, showing similar results when compared with other researches, or even with labeling done by humans.*
Keywords: *Speech, Emotion recognition, SVM.*

Introdução

Desde os primórdios da humanidade, a voz é usada como meio de comunicação, e, por meio dela e da expressão corporal, as emoções são transmitidas. O ser humano expressa seu estado e a forma como o meio em que está presente o afeta através das emoções [1].

Com a tecnologia presente nos dias de hoje, surge a necessidade crescente da automatização dessa interpretação, facilitando a interação homem-máquina no dia a dia. Um exemplo dessa interação seria a identificação da emoção durante uma ligação de emergência, auxiliando o operador a melhor lidar com a outra pessoa, ou mesmo identificando que a chamada não se trata de uma situação verdadeira.

Apesar de existirem diversas teorias na psicologia acerca da formação das emoções, o que nos interessa conhecer são os tipos de emoções e as características que definem cada uma [2,3]. Essa características são as informações que podem ser inseridas em um computador, permitindo a aprendizagem de máquina.

A aprendizagem de máquina, ou *machine learning*, em inglês, é a área da inteligência artificial que desenvolve técnicas para que as máquinas "aprendam" [2].

Nilofer et al. (2015) afirma que o reconhecimento de emoções compreende os seguintes passos: entrada de áudio, pré-processamento, extração de características, classificação, e saída. No caso deste trabalho, a saída do corresponde à emoção encontrada pelo classificador.

A partir deste conhecimento, e sabendo os aspectos importantes de cada tipo de emoção, um computador pode ser capaz de identificar a emoção transmitida pela voz.

Para este trabalho, foram tomadas as seguintes características para tentativa de identificação das emoções:

- *Pitch*: Frequência principal do sinal de áudio.
- *Loudness*: Energia ou intensidade do sinal.
- *Jitter*: Variação do *pitch*.
- *MFCC*: Coeficientes Mel Cepstrais. Representação paramétrica do espectro de frequências da voz.
- *MFB*: Banco de filtros mel.
- *LSP*: Pares espectrais de linha para representação dos coeficientes de predição linear.

Além das características principais, o método usado também considera as sub-características, conhecidas como funcionais, do vetor correspondente à algumas características, como seus valores máximo e mínimo, desvio, etc.

Para que o computador pudesse aprender a diferenciar as emoções, foi desenvolvido um algoritmo baseado em uma máquina de vetor de suporte, ou *SVM*, do inglês *support vector machine*.

Materiais e métodos

Base de dados – Para o treinamento e validação do classificador, fez-se uso da base de dados *Ryerson Audio-Visual Database of Emotional Speech and Song* (*RAVDESS*) [4], a qual contém no total 7536 gravações de alta qualidade. Elas são disponibilizadas de acordo com os seguintes identificadores:

- Modalidade: áudio e vídeo, somente vídeo, ou somente áudio.
- Canal vocal: fala ou canção.
- Emoção: neutra, calma, felicidade, tristeza, raiva, medo, nojo, surpresa.
- Intensidade emocional: normal ou expressiva.
- Frase: "*Kids are talking by the door*" ou "*Dogs are sitting by the door*".
- Repetição: primeira ou segunda instância.
- Ator: dentre 12 homens e 12 mulheres.

Para este estudo, a fim de reduzir a complexidade do problema, tendo este trabalho um cunho exploratório, levou-se em consideração apenas a modalidade exclusiva de áudio e o canal vocal de fala, selecionando também apenas as cinco emoções: neutra, calma, tristeza, felicidade, e raiva. As gravações foram agrupadas levando em consideração apenas a emoção correspondente, desconsiderando os identificadores restantes.

Das 7536 gravações, foram então usadas apenas 856, distribuídas conforme a Tabela 1 e o gráfico da Figura 1.

Tabela 1: Distribuição das gravações por emoção.

Emoção	Número de gravações
Neutra	94
Calma	192
Tristeza	190
Felicidade	189
Raiva	191

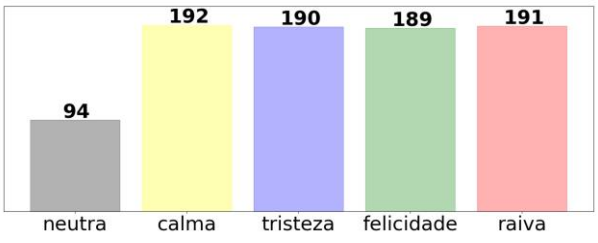

Figura 1: Gráfico de barras da distribuição das gravações por emoção.

Extração de características – Para extração das características listadas anteriormente, utilizou-se a ferramenta *openSMILE* [5], desenvolvida na linguagem de programação *C++*, pela empresa *audEERING* (Munique, Alemanha), que a disponibiliza na modalidade de código aberto (*open source*).

O funcionamento da ferramenta é descrito em uma documentação fornecida pela empresa [6], a qual lista informações sobre as características que podem ser extraídas, e sobre os arquivos de configuração, os quais correspondem a um elemento essencial do *openSMILE*, identificando as características que serão extraídas, e as etapas que serão tomadas (tipo de processamento, janelamento, etc.).

Dentre os arquivos de configuração listados, consta o denominado `emobase2010.conf` [7], desenvolvido para o desafio paralinguístico da conferência *INTERSPEECH* 2010. Essa configuração conta com as características citadas, sendo um total de 1582 atributos finais, levando em consideração as características principais, e os funcionais.

Antes da etapa de treinamento do classificador, o vetor de características foi padronizado através da equação (1).

$$x_i' = \frac{x_i - \bar{x}}{\sigma} \quad (1)$$

onde x_i corresponde a um elemento do vetor de características, $\bar{x}$ é a média do vetor, σ é o desvio padrão, e x'_i é a amostra padronizada.

Vale ressaltar que, devido à alta qualidade das gravações da base *RAVDESS*, optou-se por não realizar nenhum tipo de pré-processamento sobre elas, como remoção de ruídos, equalização, etc.

Balanceamento das classes – Pela Figura 1, é evidente a susceptibilidade desse conjunto de amostras a problemas de desbalanceamento de classes [8], comum em diversas aplicações de aprendizagem supervisionada. Visando melhorar a proporção de amostras da classe da emoção neutra, duplicar suas amostras é uma possibilidade. Porém, isso potencialmente levaria ao efeito conhecido como *overfitting*, recorrente em técnicas de classificação, no qual o classificador seria condicionado inadequadamente por essas amostras repetidas, tal que novas amostras introduzidas teriam uma chance elevada de serem classificadas incorretamente, mesmo possuindo características parecidas com as das amostras originais.

Levando isso em consideração, foi aplicado o algoritmo *Synthetic Minority Over-sampling Technique* (*SMOTE*) [9], o qual consiste em "sintetizar" amostras de uma classe minoritária. De forma simplificada, essa sintetização se baseia na multiplicação de uma constante aleatória (entre 0 e 1) pelo vetor diferença correspondente a duas amostras do conjunto original, escolhidas também aleatoriamente dentro da classe

desejada. Isso se repete até que seja atingido um número adequado de amostras na classe.

As Figuras 2 e 3 correspondem à visualização do *SMOTE* em duas dimensões, ou seja, quando o vetor característica possui duas entradas.

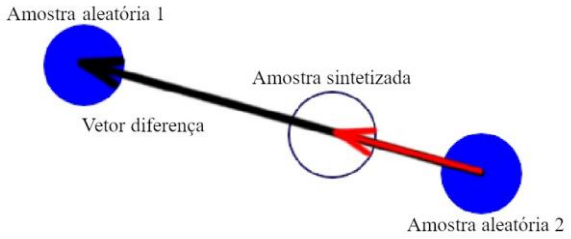

Figura 2: Exemplo de sintetização de amostra em duas dimensões. A sintetização da amostra depende de um valor entre 0 e 1 escolhido aleatoriamente.

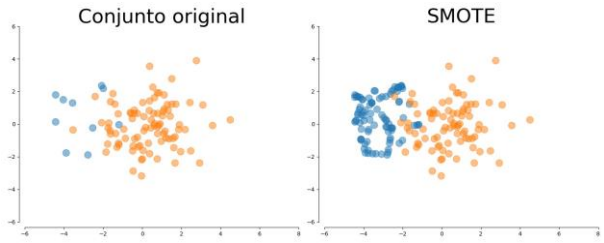

Figura 3: Exemplo da aplicação do SMOTE em um conjunto de amostras em duas dimensões. Fonte: [10].

Dessa forma, para o treinamento e validação do classificador, todas as classes apresentaram 192 amostras cada, totalizando 960 amostras. Ou seja, foram sintetizadas 98 amostras para a emoção neutra, 2 para tristeza, 3 para felicidade, e 1 para raiva.

Classificação – Tendo em vista o alto desempenho observado no uso das *SVMs* para classificações em inúmeros estudos, ela foi escolhida como o classificador para este trabalho.

Em essência, a *SVM* busca encontrar um hiperplano capaz de separar um conjunto de amostras em classes pré-determinadas. Caso não exista no espaço n-dimensional do conjunto um hiperplano capaz de realizar essa separação, é feita uma transformação do conjunto, tal que ele será levado a uma dimensão na qual essa separação existe [11]. Um exemplo dessa transformação, de duas para três dimensões, pode ser visualizado na Figura 4.

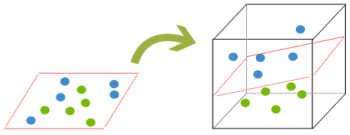

Figura 4: Exemplo de hiperplano obtido após transformação do conjunto de amostras. Fonte: [12].

A implementação da *SVM* adotada é a do pacote *scikit-learn* [13], presente na linguagem de programação *Python*, também usado para as análises estatísticas. Os parâmetros relevantes para este trabalho são: núcleo (*kernel*) linear, e $C = 0{,}001$. Este valor de C foi tomado como ótimo nesta aplicação, após testes realizados com diversos valores.

Resultados

Para a etapa de treinamento e validação, as 960 amostras foram separadas em dois grupos: 720 amostras (75%) para o treinamento; e 240 (25%) para a validação. Essa etapa foi realizada 8 vezes, com os grupos sempre selecionados de forma aleatória dentre as amostras disponíveis.

A distribuição das acurácias dos classificadores resultantes de cada treinamento pode ser visualizada no gráfico da Figura 5, com algumas estatísticas relevantes na Tabela 2.

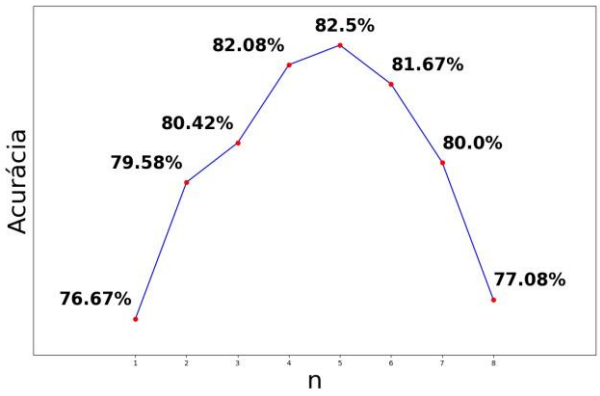

Figura 5: Distribuição das acurácias obtidas.

Tabela 2: Valores estatísticos correspondentes à distribuição das acurácias obtidas.

Estatística	Valor (%)
Média	80,00
Desvio padrão	2,04
Valor máximo	82,50
Valor mínimo	76,67

A Tabela 3 mostra a matriz de confusão correspondente à média das validações de cada repetição.

Tabela 3: Matriz de confusão correspondente às validações realizadas. A linha indica a classe correta, e a coluna é a classe encontrada pelo classificador.

-	neutra	calma	tristeza	felicidade	raiva
neutra	**93,41**	4,15	0,14	1,58	0,72
calma	18,16	**79,09**	0,00	2,75	0,00
tristeza	8,53	0,71	**76,53**	7,11	7,11
felicidade	12,96	16,46	6,88	**62,21**	1,48
raiva	2,45	0,68	7,07	4,35	**85,46**

Discussão

De acordo com a Tabela 3, algumas observações podem ser feitas. A emoção neutra apresentou um percentual de 93,41% de acerto, o que corresponde à identificação correta da emoção neutra por parte da *SVM*. Em relação às outras emoções, esse índice é visivelmente mais elevado. Acredita-se que isso seja devido à aplicação do *SMOTE*. Como a sintetização foi bem mais discreta para as outras classes (apenas 1 amostra sintetizada para raiva, por exemplo), o seu efeito foi maior na neutra.

A excelente identificação da emoção raiva também deve ser ressaltada, em que a *SVM* identificou corretamente 85,46% dos casos. Este resultado pode ter sido alcançado devido às diferenças bem mais evidentes para as características da voz (altura, velocidade, etc.), identificadas entre a raiva e as outras emoções, sendo que essas diferenças apresentaram-se nas características escolhidas.

A felicidade foi a que obteve o pior resultado, tendo identificados corretamente apenas 62,21% dos casos. Essa emoção, quando comparada às outras, apresentou uma grau maior de confusão com a calma (16,46%), e com a neutra (12,96%). Esse resultado pode indicar que as características escolhidas não são as mais adequadas para diferenciar a felicidade da calma e da neutra, já que existe uma semelhança inerente entre elas (velocidade e altura menos elevadas, por exemplo).

Destaca-se também que o algoritmo não confundiu a emoção calma com as emoções de tristeza e raiva. Porém, em 18,16% dos casos a calma foi identificada como neutra, semelhante a como mencionado para a felicidade. Apesar disso, ela obteve o segundo melhor índice de classificação correta, com um resultado de 79,09%.

Conclusão

A *SVM*, a princípio, mostrou-se eficiente na identificação de emoções, obtendo bons resultados quando comparado com outros trabalhos, e com o uso de outros métodos, como o dos K-vizinhos mais próximos (*KNN*), modelos de misturas de gaussianas (*GMM*), e modelos ocultos de Markov (*HMM*) [3].

Além disso, quando comparado com identificações realizadas por pessoas, há algoritmos que apresentaram resultados equiparáveis, com média de reconhecimento de 79,34%, sendo que a média de reconhecimento realizado por pessoas é de 81,3% [2].

Referências

[1] MATTE, A. C. F. ; VIEIRA, J. M.; MEIRELES, A. R.; ARANTES, P. **Emoção na fala: uma análise crítica.** Caderno de Discussão do Centro de Pesquisas Sociossemióticas , São Paulo - SP, v. 10, n.1, p. 1, 2004.

[2] ROSA, J. da. **Reconhecimento automático de emoções através da voz.** 2017. 98 f. TCC (Graduação) - Curso de Sistemas de Informação, Universidade Federal de Santa Catarina, Florianópolis, 2017.

[3] IRIYA, R. **Análise de sinais de voz para reconhecimento de emoções.** 2014. 100 f. Dissertação (Mestrado) - Curso de Ciências, Departamento de Engenharia de Sistemas Eletrônicos, Escola Politécnica da Universidade de São Paulo, São Paulo, 2014.

[4] LIVINGSTONE S. R.; RUSSO F. A. **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): e0196391. https://doi.org/10.1371/journal. pone.0196391

[5] EYBEN F.; WENINGER F.; GROSS, F.; SCHULLER, B.; **Recent developments in openSMILE, the munich open-source multimedia feature extractor,** in Proceedings of the 21st ACM international conference on Multimedia - MM '13, 2013, pp. 835–838.

[6] WENINGER, F. **openSMILE: open-Source Media Interpretation by Large feature-space Extraction.** Disponível em: <https://www.audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf>. Acesso em: 07 de dez. 2018.

[7] XINGYU, N. **emobase2010.conf.** Disponível em: <https://github.com/naxingyu/opensmile/blob/master/config/emobase2010.conf>. Acesso em: 05 dez. 2018.

[8] BATUWITA, R.; PALADE, V. **Imbalanced Learning**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[9] N. V. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. **SMOTE: Synthetic Minority Over-sampling Technique**. Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002.

[10] G. Lemaitre, F. Nogueira, D. Oliveira, C. Aridas. **SMOTE + ENN**. Disponível em: <https://561-36019880-gh.circle-artifacts.com/0/home/ubuntu/imbalanced-learn/doc/_build/html/auto_examples/combine/plot_smote_enn.html>. Acesso em: 05 dez. 2018.

[11] KIM, E. **The Kernel Trick.** Disponível em: <http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html>. Acesso em: 17 de dez. 2018.

[12] BAMBRICK, Noel. **Support vector machines for dummies:** A simple explanation. Disponível em: <http://blog.aylien.com/support-vector-machines-for-dummies-a-simple/>. Acesso em: 07 dez. 2018.

[13] PEDREGOSA F. et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2012.