Identificação de sentimento em voz por meio da combinação de classificações intermediárias dos sinais em excitação, valência e quadrante

Guilherme B. S. Gering, Patrick M. Ciarelli, Evandro O. T. Salles

[1]Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal do Espírito Santo (UFES) – Vitória, ES – Brasil

guilhermebutzke@outlook.com,{patrick.ciarelli,evandro.salles}@ufes.br

Abstract. *Speech emotion recogntion is commonly performed in categorical classes, such as "sadness" or "joy". According to Russell's map of affection, emotions can also be classified by arousal (excitation), valence, and quadrants. In this work is proposed a method to increase the performance of speech emotion recogntion in categorical classes using classifiers that perform intermediate classification in the classes of valence, excitation and quadrants using a multi-view approach. To combine these results and obtain the final classification, a decision tree is proposed and that increases F1 metrics from 0.73 by Ensemble of three kinds of classifiers to 0.87 in a public database.*

Resumo. *A identificação de sentimento em voz é comumente realizada em classes categóricas como "tristeza" ou "alegria". De acordo com o mapa de afeto de Russell, sentimentos também podem ser classificados por excitação, valência e quadrantes. Neste trabalho é proposto um método para incrementar o desempenho de identificação de sentimentos em classes categóricas utilizando classificadores que realizam classificação intermediária nas classes de excitação valência e quadrantes usando uma abordagem multi-visão. Para combinar esses resultados e obter a classificação final é proposta uma árvore de decisão que aumentou o desempenho F1 de 0,73 do Ensemble de três tipos de classificadores para 0,87 sobre uma base de dados pública.*

1. Introdução

O complexo sinal de voz pode trazer várias informações a respeito da mensagem, do locutor, da linguagem e da emoção transmitida [Livingstone and Russo 2018, Gadhe et al. 2015, Pathak and Kolhe 2016]. Humanos têm uma habilidade natural de reconhecer emoções através da fala. Máquinas, inclusive, podem identificar "quem disse" e o "que foi dito" na fala, além de poderem também identificar sentimentos expressos nas frases [Gadhe et al. 2015].

No campo da saúde, a identificação de sentimento em voz pode monitorar as condições de paciente em reabilitação ao, aconselhamento psicológico, identificação de autismo e identificação de pacientes com stress ou depressão [Reddy and Vijayarajan 2017]. O estudo e o entendimento de emoções se aplica também quando se deseja conhecer o bem-estar de uma pessoa (seja um paciente, usuário, ou cliente) em determinado espaço.

Se entende por Identificação de Sentimentos em Voz (*Speech Emotion Recognition* - SER) o reconhecimento de sentimentos por máquinas. Duas teorias são amplamente utilizadas para classificação de sentimentos. A primeira associa cada sentimento a uma entidade discreta, separável, categorizada em tipos e quantidades (como raiva, medo, tristeza, alegria, etc.). Estas são chamadas de classes categóricas de sentimento [Bestelmeyer et al. 2017]. A segunda teoria avalia cada emoção com um grau de excitação ou de valência (*arousal and valence*), portanto, em um plano bidimensional. Essas classes são ditas contínuas, e as emoções são decompostas em excitação (ou ativação) ou valência em uma escala de valores [Xia and Liu 2017]. Em [Parthasarathy and Busso 2017] também é utilizado dominância (ativa ou passiva) como uma classe contínua de sentimento.

A valência qualifica o sentimento quanto à simpatia, numa escala de sentimentos negativos (desagradáveis) até positivos (agradáveis). Em excitação (ou intensidade), quantificam-se os sentimentos quanto ao nível de ativação provocado pelo mesmo, em uma faixa que vai de baixo (calmo) até alto (excitado) [Russell 1980]. Essa teoria é muito utilizada nos estudos referentes a emoção. A Figura 1 apresenta o modelo bidimensional de emoção descrito que é conhecido na literatura como modelo circumplexo de afeto de Russell (*Russell's Circumplex Model of Affect*) [Russell 1980], ou mapa de afeto. A figura mais a esquerda apresenta o mapa de afeto para um conjunto de sentimentos apresentados no trabalho de Russell. A figura mais a direita apresenta o mapa de sentimentos encontrados na base de dados Berlin [Burkhardt et al. 2005], que foi utilizada como banco de dados de áudio para este trabalho. Além de excitação e valência, neste trabalho utiliza-se a posição dos quadrantes do mapa de afeto como identificação de um sentimento. A nomenclatura dos quadrantes é indicada no mapa de afeto da base Berlin na Figura 1.

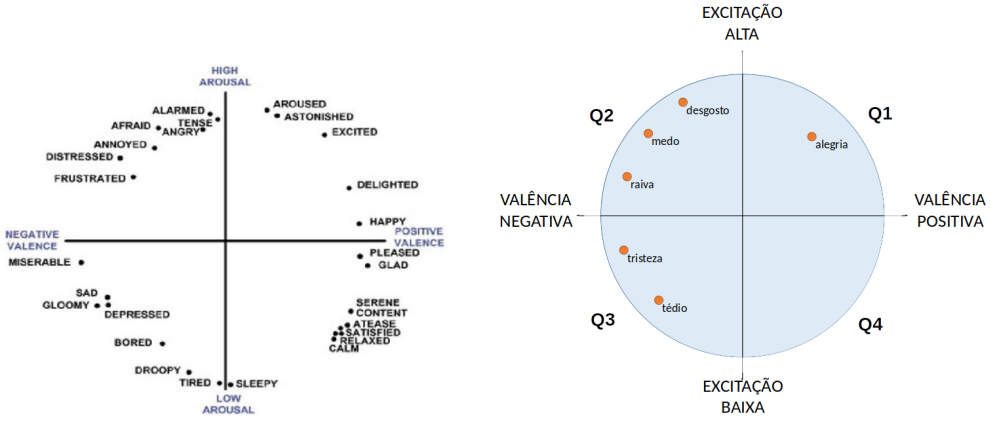

Figura 1. [esq] Modelo Bidimensional de emoções baseado no circumplexo de afeto de Russell e [dir] Modelo Bidimensional de emoções para base da dados Berlin

A proposta deste trabalho é usar informações de excitação, valência e a posição do quadrante do sentimento no mapa circumplexo de emoções para auxiliar na identificação de sentimentos categóricos. Para realizar esta tarefa, é proposta uma metodologia que combina as saídas de vários classificadores, cada qual responsável por realizar a classificação do áudio da voz em diferentes aspectos, como valência, excitação e quadrante, num abordagem que chamamos de multi-aspecto. A classificação dos sentimentos ocorre através de modelos de classificadores distintos que utilizam diferentes carac-

terísticas do sinal de voz para rotulá-los em diferentes aspectos. O uso de diferentes características do sinal para classificação é chamado de abordagem multi-visão de sinal. A principal contribuição teórica deste artigo é objetivada na criação de uma técnica de classificação de sentimentos que combina as abordagens multi-aspecto e multi-visão do sinal de voz.

Este trabalho está organizado como segue. Na Seção 2 é feita uma revisão de literatura. Na Seção 3 é apresentada a metodologia proposta. Os experimentos e análises são realizadas na Seção 4. Por fim, as conclusões são feitas na Seção 5.

2. Revisão de Literatura

2.1. Métodos para Reconhecimento de Sentimentos em Fala

Um sistema reconhecedor de sentimento em fala pode ser compreendido como aquele capaz de extrair informações de voz e destas características pressupor a emoção do falante. Os objetivos principais de um SER são identificar os sentimentos presentes em uma fala e sintetizar a mensagem desejada de acordo com uma mensagem pretendida [Pathak and Kolhe 2016].

Em [Reddy and Vijayarajan 2017] é afirmado que diferentes classificadores podem ser utilizados em aplicações com sinal de voz, como o Modelo de Mistura de Gaussianas, Cadeias de Markov, Redes Neurais Artificiais, Máquinas de Vetores de Suporte (*Support Vector Machine* – SVM) e Redes Neurais Profundas.

Máquinas de Vetores de Suporte têm encontrado resultados muito interessantes na identificação de sentimento em voz. Em [Shen et al. 2011] é descrita uma abordagem em que o sinal da voz é representado pelas características prosódicas de energia, pitch, LPCC (*Linear Prediction Cepstral Coefficient*), e também pelas características espectrais MFCC (*Mel Frequency Cepstral Coefficients*) e LPMCC (*Linear Prediction Mel Frequency Cepstral Coefficients*). Os treinos são realizados com as características individuais e também combinadas, e nos experimentos eles concluem que a combinação das características de Energia, Pitch e LPMCC alcançam melhores resultados. Os testes foram realizados sobre cinco sentimento da base de dados Berlin e a acurácia foi em torno de 82.5%.

Redes neurais convolucionais (*Convolutional Neural Network* – CNN) são arquiteturas de redes neurais que possuem camadas convolucionais que filtram o sinal de entrada extraindo características de alto e baixo nível do sinal que são diretamente passadas para uma camada totalmente conectada [Badshah et al. 2017]. Devido as operações convolucionais, as CNN são muito aplicadas para classificação de imagens ([Zhao et al. 2019]). Em SER, redes convolucionais são aplicadas para extrair características do sinal de voz temporal (sinal unidimensional), ou por vezes também de uma representação do espectrograma do sinal da fala, como pode ser visto em [Zhao et al. 2019] e [Badshah et al. 2017]. As CNN têm sido utilizadas para a identificação de sentimentos com resultados bem promissores, inclusive melhorando os resultados de SVM [Zhang et al. 2018].

Arquiteturas LSTM (*Long-Short Term Memory*) são utilizadas para classificação de sinais cujo o estado atual tem alta dependência de estados passados por meio de funções com capacidade de armazenar informações relevantes a longo prazo bem como também esquecer informações mais irrelevantes. Em SER, eles tem sido muito utilizadas conecta-

das as camadas de CNN, como pode ser visto em [Zhao et al. 2019] e [Fayek et al. 2017]. A característica das LSTM de armazenar informações a longo prazo se torna poderosa para identificação de sentimentos, uma vez que a característica do sentimento na fala é predominante a longo prazo. O trabalho de [Fayek et al. 2017] apresenta o resultado para várias arquiteturas de CNN construídas, incluindo aquelas que apresentam camadas de LSTM.

Técnicas de *Ensemble* utilizam o treino paralelo de mais de um classificador nos quais as saídas dos modelos são combinadas de forma a garantir uma predição final baseada nos resultados individuais de cada tarefa [Shih et al. 2017]. O trabalho de [Zhang et al. 2018], por exemplo, possui um esquema de *Ensemble* constituído de quatro classificadores e a predição de cada um desses modelos é combinada por um esquema de votação para estabelecer-se uma predição final de classificação.

Na literatura é definido também o termo multi-visão para análise de sentimentos. Em [Tuarob et al. 2014], por exemplo, é apresentada uma metodologia que combina cinco tipos de características que representam diferentes aspectos de semânticas em mensagens para identificação de sentimentos em textos de redes sociais. A abordagem multi-visão em SER pode ser entendida como a análise de um sinal de voz sob o ponto de vista de diferentes características, como por exemplo, a análise espectral em conjunto com a análise de energia da fala. Em [Zhang et al. 2018], por exemplo, é proposta uma arquitetura para reconhecimento de emoção em voz que combina redes convolucionais unidimensional para classificar o sinal de voz no tempo e redes convolucionais bidimensionais para extrair informações do espectro.

Árvores de decisão hierárquicas têm sido utilizadas em SER por voz para classificar sentimentos através de múltiplos classificadores. Em [Lee et al. 2011] é utilizada uma classificação de sentimentos baseadas em árvores de decisão binárias hierárquicas onde cada folha de decisão da árvore identifica individualmente uma classe de sentimentos por meio de classificadores SVM ou por Regressão Logística Bayesiana. A ordem com que as classes são separadas a cada nível da árvore, escolhidas pelos autores, é empírica, baseado no conjunto de alguns modelos testados.

2.2. Base de Dados de Voz

A base de dados de voz para identificação de sentimentos utilizada nesta pesquisa foi a base de dados Berlin [Burkhardt et al. 2005]. A base é do tipo simulada, composta por cinco falantes de sexo masculino e 5 falantes de sexo feminino que totalizam 535 sinais de áudio gravados a uma taxa de 48kHz e reamostrados a 16kHz. Sete categorias de sentimento são representadas nessa base: Raiva (***A**nger*), Tédio (***B**oredom*), Desgosto (***D**isgust*), Medo (***F**ear*), Felicidade (***H**appy*), Tristeza (***S**ad*) e estado Neutro (***N**eutral*). Ao longo do texto, essas classes de sentimentos são por vezes indicadas pela primeira letra do nome do sentimento em inglês. O tempo de cada sinal dura entre 2 a 12 segundos.

3. Metodologia

A abordagem proposta neste trabalho para identificação de sentimentos discretos é realizada em duas etapas. Na primeira etapa, vários classificadores são construídos para identificar o sentimento em voz em diferentes aspectos (excitação, valência, quadrantes, sentimentos categóricos). Um mesmo modelo de classificador é utilizado para classificar

o sentimento sobre diferentes rótulos. A classificação do sinal em cada um destes aspectos é realizada em uma abordagem multi-visão, pois cada tipo de classificador utiliza diferentes características extraídas do sinal de voz para classificá-lo em um tipo de rótulo. O termo multi-visão se deve ao fato do método proposto utilizar diferentes "visões" do sinal de voz, através de diferentes características, para a classificação em emoção. Na segunda etapa do método, as saídas dos classificadores são usadas como características de entrada em uma árvore de decisão, e a saída da mesma é a classificação do sentimento em uma categoria discreta, como raiva, medo ou alegria.

Antes de explicar o método proposto, as partes que constituem o método são apresentadas.

3.1. Tipos de Rótulos

Denomina-se "Tipo de rótulo" o nome dado a um grupo de classes que representam o sinal de voz. Seis tipos de rótulos foram utilizados neste trabalho: 1) rótulos categóricos, que caracterizam os sentimentos individualmente, como em "raiva e tristeza"; 2) rótulos de Excitação; 3) Valência e 4) Quadrante, que caracterizam o sinal de voz quanto a sua posição no modelo circumplexo de afeto (Figura 1); 5) rótulos ADF, que classificam o sinal em *anger*, *disgust* e *fear*, se o sinal for previamente classificado como pertencente ao quadrante 2 (Q2); e 6) rótulos BS, que classificam o sinal em *boredom* e *sadness*, se o sinal for previamente classificado como pertencente ao quadrante 3 (Q3). A justificativa para existência de classificadores com rótulos do tipo ADF e BS está apresentada na Seção 3.3. A lista abaixo apresenta as classes pertencentes a cada tipo de rótulo. Na Tabela 1 é apresentado como cada sentimento discreto está relacionado com as classes dos demais tipos de rótulos. Por exemplo, o sentimento *anger* está relacionado ao nível de excitação alta (H), valência negativa $(-)$ e ao quadrante 2 (Q2). Destaca-se que o quadrante Q3 foi suprimido pois não existe emoção classificada nesse quadrante pela base de dados Berlin utilizada nesse trabalho. Q0 refere-se ao sentimento neutro. A tabela ainda apresenta a quantidade de amostras registradas para cada sentimento. Reparar que esta é uma base desbalanceada.

1. Rótulos categóricos: ***A**nger, **B**oredom, **D**isgust, **F**ear, **H**appiness, **S**adness, **N**eutral*
2. Rótulos de Excitação: Alta (**H**), Neutra (0), Baixa (**L**)
3. Rótulos de Valência: Positiva $(+)$, Neutra (0), Negativa $(-)$
4. Rótulos de Quadrante: **Q1**, **Q2**, **Q3**, **Q4** e **Q0** (neutro)
5. Rótulos ADF: **A**, **D**, **F**
6. Rótulos BS: **B**, **S**

3.2. Classificadores

Quatro tipos de classificadores são utilizados neste trabalho para identificação das emoções dos sinais de voz: SVM (1), CNN-LSTM-1D (2), CNN-LSTM-2D (3) e o *Ensemble* (4) destes classificadores. A Figura 2 apresenta a estrutura genérica de um classificador. Os tipos de rótulos associados aos classificadores podem ser qualquer um dos seis tipos listados. Para indicar o tipo de rótulo e o classificador utilizado, é empregada uma termologia. Por exemplo, um classificador valência-CNN-LSTM-1D utiliza a arquitetura CNN-LSTM-1D para categorizar os sentimentos quanto a sua valência (positiva, neutra, negativa).

Tabela 1. Rótulos dos sentimentos categorizados individualmente, por excitação, valência e por quadrantes e quantidade de amostras de cada classe

	Sentimento	Categóricos	Excitação	Valência	Quadrantes	ADF	BS	Qnt.
0	Anger	A	H	(-)	Q2	A	-	127
1	Boredom	B	L	(-)	Q3	-	B	81
2	Disgust	D	H	(-)	Q2	D	-	46
3	Fear	F	H	(-)	Q2	F	-	69
4	Happiness	H	H	(-)	Q1	-	-	71
5	Sadness	S	L	(+)	Q3	-	S	62
6	Neutral	N	0	(0)	Q0	-	-	79

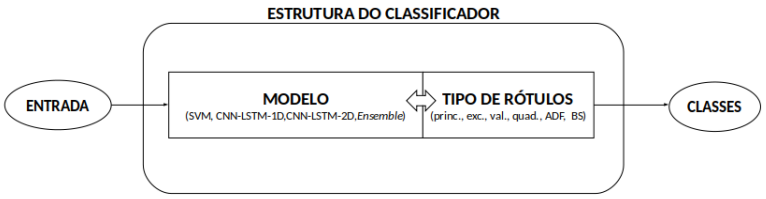

Figura 2. Estrutura genérica de um classificador

A entrada de cada tipo de classificador varia conforme o modelo. Os classificadores SVM utilizados nesta pesquisa para identificar sentimentos em voz foram baseados no artigo de [Shen et al. 2011]. Os autores utilizam como entrada um vetor de características estatísticas (máximo, mínimo, média, taxa de subida e descida, entre outras) extraídas sobre a energia, o pitch, o LPCC além dos coeficientes espectrais de MFCC e LPMCC do sinal de voz. A concatenação dessas características totaliza um vetor de 52 dimensões, que é posteriormente classificado pelo SVM. A descrição das características, o pré-processamento do sinal e o detalhamento da arquitetura do SVM é apresentada em [Shen et al. 2011].

Redes convolucionais associadas a LSTM são outros tipos de classificadores utilizados neste trabalho e foram baseados em [Zhao et al. 2019]. Os classificadores que utilizam redes convolucionais para extrair informações temporais do sinal de voz são aqui referenciados como CNN-LSTM-1D, enquanto que os classificadores que utilizam redes convolucionais para extrair informações espaciais do espectrograma do sinal são chamados de classificadores CNN-LSTM-2D.

No classificador CNN-1D-LSTM, uma janela de 1.2 segundos centralizada na metade do sinal do áudio a ser classificado foi utilizada como vetor de entrada para a rede, nas quais características temporais foram extraídas pela camada convolucional e foram posteriormente treinadas pela camada LSTM. No classificador CNN-LSTM-2D, calculou-se o espectrograma da mesma janela de 1.2 segundos, nas quais características espaciais do espectro foram calculadas pela camada convolucional e posteriormente treinadas na camada LSTM. As arquitetura de CNN-LSTM-1D e CNN-LSTM-2D estão descritas no trabalho de [Zhao et al. 2019], onde podem ser encontrados mais detalhes também sobre o pré-processamento do sinal, e também sobre o método para determinação do espectrograma do sinal.

O *Ensemble* combina os resultados dos classificadores SVM, CNN-LSTM-1D e CNN-LSTM-2D para predizer uma classe do sinal. O *Ensemble* e uma técnica que combina os resultados de classificação, mas neste texto tratamos ele como um dos quatro tipos de classificador. Os resultados dos classificadors combinados em uma abordagem multivisão é interessante pois a identificação da emoção acontece por meio de diferentes características do sinal: por características estatísticas do sinal classificadas no SVM, por características temporais extraídas da CNN-LSTM-1D e por características espectrais devido à CNN-LSTM-2D. A técnica de votação do tipo moda foi escolhido para predição final da classe do *Ensemble*; ou seja, cada classificador prediz uma classe e a classe mais comum é a predição final do *Ensemble*. Em caso de todas as classes encontradas serem distintas, escolhe-se a classe que foi encontrada pelo modelo que obteve o melhor desempenho na fase de validação dos classificadores. Para exemplificar a técnica, a saída do classificador quadrantes-*Ensemble* é a moda dos resultados obtidos pelos modelos quadrantes-SVM, quadrantes-CNN-LSTM-1D e quadrantes-CNN-LSTM-2D. Espera-se que o *Ensemble* de classificadores apresente desempenho melhor que os classificadores que apresentam apenas um modelo de classificação, justamente por essa técnica analisar o sinal por diferentes "pontos de vista", tendendo a uma predição mais precisa.

A métrica utilizada nesta pesquisa para avaliar o desempenho de cada classificador foi a F1-*score*, que consiste da média harmônica entre a precisão e o *recall* de cada classe. Cada classe classificada apresenta uma *recall* e uma precisão, e portanto, uma métrica F1 própria. A métrica F1 para o classificador como um topo é calculado como uma média resultados F1 de cada classe ponderadas pela quantidade de elementos existente em cada classe (também chamada de suporte). Essa abordagem é conhecida na literatura como F1-*score weighted average*. Quanto maior o valor de F1, melhor, sendo o ideal o valor igual a 1. Embora comumente a métrica acurácia seja utilizada na literatura, ela não é adequada para bases de dados desbalanceadas, sendo melhor usar a métrica F1.

3.3. Árvore de Decisão Hierárquica

A metodologia utilizada para combinar os resultados dos classificadores foi criar uma árvore de decisão hierárquica binária. Cada nível da árvore utiliza as saídas de algum modelo de regressão ou classificação para separar um tipo de classe das demais. O termo binário é dado pois cada nível estabelece se um dado é "sim" ou "não" pertencente a uma classe [Lee et al. 2011, MAO et al. 2010, Liu et al. 2017]

A Figura 3 ilustra uma representação simplificada da estrutura do método proposto. A entrada da estrutura apresentada é um sinal de voz que vai ser classificados em seis tipos de rótulos: individuais, excitação, valência, BS, ADF e quadrantes (ver Tabela 1). Cada bloco recebe o sinal de entrada e rotula o sinal através de quatro tipos de classificadores: SVM, CNN-LSTM-1D, CNN-LSTM-2D e *Ensemble*. Repare que há exceções nos blocos BS e ADF, pois esses recebem apenas os sentimentos do terceiro e do quarto quadrante, respectivamente. Repare que também há uma exceção no bloco de quadrantes, pois eles não possui quatro classificadores, mas apenas um único que combina os resultados de excitação e valência, conforme explicado no final da seção. Reparar que, antes do bloco da etapa da árvore, 76 valores F1 são encontrados: 4 classificadores $\times$ (7 individuais +3 excitacoes +3 valências +2 BS +3 ADF) classes + 1 classificador $\times$ 4 quadrantes.

Para construir a árvore, foi proposto um algoritmo que busca o melhor resultado

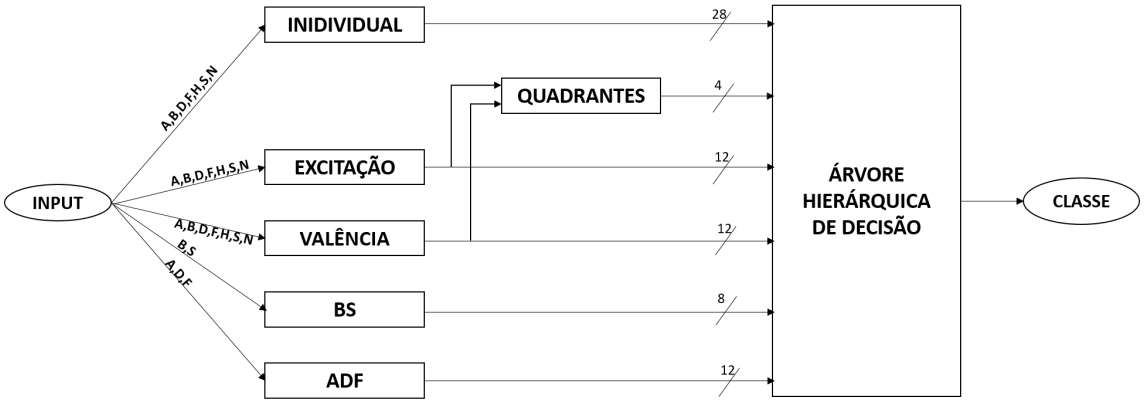

Figura 3. Estrutura do método proposto. Em uma primeira etapa o sinal é classificado em diferentes tipos de rótulos por quatro tipos de classificadores. Na segunda etapa, as saídas dos classificadores são avaliados por uma árvore de decisão.

F1 obtido com dados de validação dos classificadores. Identifica-se qual classificador e qual tipo de rótulo tiveram maior desempenho. Essa informação é usada para construir um nó de decisão. O classificador utilizado para fazer a decisão é descartado e o procedimento é repetido para encontrar outro nó de decisão abaixo desse.

Por exemplo, a Figura 4 apresenta um esquema de árvore construída com os dados de validação dessa pesquisa. Entres os 76 valores de F1 encontrados, o maior valor foi encontrada para a classe "D" obtida pelo classificador excitação-SVM. Portanto, o primeiro nível de decisão da árvore vai utilizar esse classificador para dizer se o sentimento de entrada é "D" (hipótese afirmativa) ou se não é (hipótese contrária). Do segundo nível da árvore em diante, o algoritmo vai procurar o maior resultado F1 entre as classes que ainda não foram classificadas. No exemplo da Figura 4, após separar "D", o resultado de maior F1 encontrado foi "ALTA" prevista pelo classificador excitação-*Ensemble*. Portanto, este classificador foi utilizado para predizer se um sentimento é "ALTA" em hipótese afirmativa ou se não é "ALTA" em hipótese contrária. Analisando a hipótese afirmativa, apenas os sentimentos "A","F" e "H" são "ALTOS". Dentre os classificadores que rotulam essas classes, o classificador quadrantes-*Ensemble* é o que melhor separa as mesmas, e esse esquema de busca e separação acontece até o ponto em que cada nó final da árvore classifique sentimentos categóricos. Isso faz com que, automaticamente, um sentimento classificado como excitação, valência e quadrantes sempre tenham que passar por outros classificadores para "separar" as emoções pertencentes a seus quadrantes. Caso haja empate nos desempenhos de F1, priorizou-se utilizar aqueles classificadores que consegue separar a maior quantidade de sentimentos individuais; mas outras estratégias poderiam ser utilizadas, como escolher para nível de decisão aquela classe que pertence ao classificador de maior desempenho médio.

Os classificadores com os tipos de rótulos ADF e BS foram construídos para separar as classes pertencentes somente aos quadrantes Q2 e Q3, e eles nunca são usados antes da árvore rotular o sentimento a um desses quadrantes. Estes classificadores foram utilizados especificamente para aprimorar os resultados da árvore, isso pois, uma vez que se classifica um sentimento pertencente a Q2 ou Q3, é necessário um modelo sequente para classificar os sentimentos nestes quadrantes. Classificadores que rotulam sentimen-

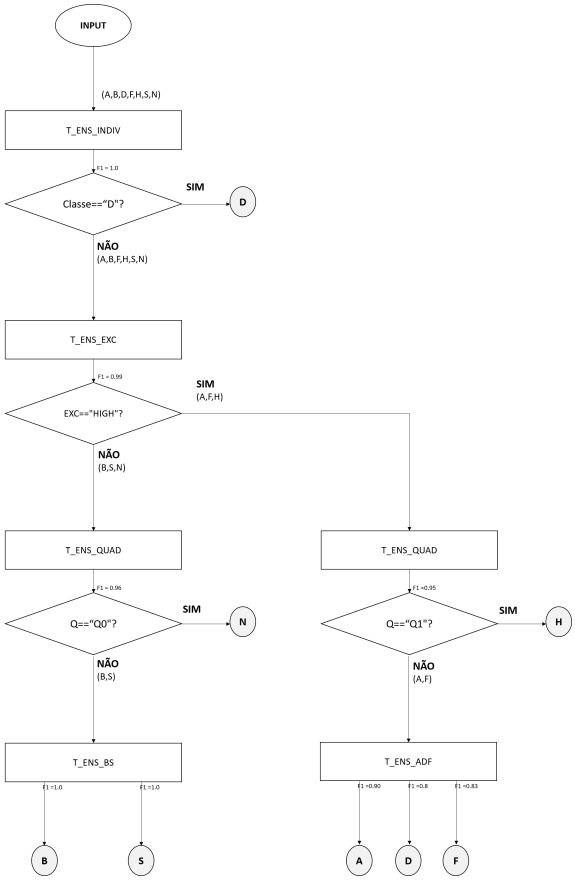

Figura 4. Árvore Hierárquica de decisão para classificação de sentimentos

tos individuais poderiam ser usados para essa abordagem. Contudo, experimentalmente verifica-se que se torna mais assertivo utilizar modelos treinados apenas com os tipos de sentimentos pertencentes aos quadrantes ao usar modelos que contemplam o treino de todos os tipos de sentimentos.

Uma última observação é feita sobre a construção dos classificadores rotulados em quadrantes. Conforme pode ser visto na Tabela 1, existe uma correspondência entre os rótulos de excitação, valência e quadrante. Por exemplo, sentimentos rotulados como excitação "alta" e valência "positiva" são sentimentos pertencentes ao primeiro quadrante. Neste trabalho é proposto usar a combinação de excitação e valência para a descoberta do quadrante: Os melhores resultados F1 de excitação são combinados com os melhores resultados F1 de valência para obter a classificação do quadrante. Em caso de uma combinação em que apresente um sentimento existente a Q4, não existe na base de dados Berlin, pode se utilizar algum classificador-individual de bom desempenho para predizer o quadrante do sentimento.

4. Resultados

Inicialmente, os dados da base de dados foram divididos em conjuntos de treino (80%), validação (10%) e teste (10%). Os dados de treino foram utilizados no treinamento dos classificadores. Os dados de validação foram utilizados para calcular os desempenhos F1 para cada classe dos classificadores. Os resultados obtidos nesta etapa estão ilustrados na

Tabela 2. Por motivos de simplificação, os resultados do classificador em quadrantes foi colocado na coluna "*Ensemble*", mas se sabe que ele é uma combinação dos classificadores excitação e valência.

Observando os resultados da tabela, verifica-se que as tarefas de *Ensemble* apresentam, para maioria das classes, desempenhos melhores se comparados ao desempenho individual de cada classificador.

Tabela 2. Resultados de F1 dos dados de validação para cada classe

CLASSIFIC.	ROTULOS	SVM	CNN-1D	CNN-2D	ENS.	CLASSIF.	ROTULOS	SVM	CNN-1D	CNN-2D	ENS.
INDIV.	A	0,64	0,86	0,53	0,62	EXCIT.	H	0,99	0,93	0,97	1,00
	B	0,75	0,52	0,70	0,75		L	0,71	0,68	0,83	0,83
	D	0,44	0,10	0,61	1,00		O	0,78	0,51	0,82	0,92
	F	0,18	0,42	0,49	0,67		MEDIA:	0,90	0,84	0,92	0,96
	H	0,50	0,20	0,46	0,67	VALEN.	N	0,74	0,88	0,96	0,92
	S	0,50	0,61	0,83	0,67		O	0,67	0,44	0,91	0,92
	N	0,77	0,32	0,81	0,87		P	0,52	0,13	0,94	0,84
	MEDIA:	0,58	0,58	0,62	0,73		MEDIA:	0,67	0,79	0,94	0,90
QUAD.	Q0	-	-	-	0,96	ADF-Q2	A	0,90	0,89	0,92	0,90
	Q1	-	-	-	0,84		D	0,80	0,57	0,86	0,80
	Q2	-	-	-	0,94		F	0,83	0,67	0,93	0,83
	Q3	-	-	-	0,91		MEDIA	0,88	0,78	0,92	0,88
	MEDIA:	-	-	-	0,92	BS-Q3	B	1,00	0,86	0,56	1,00
							S	1,00	0,67	0,85	1,00
							MEDIA	1,00	0,80	0,78	1,00

Nota: Os resultados para os classificadores individuais e *Ensemble* foram suprimidos, uma vez que o único classificador com rótulo de quadrante utilizado é combinação de classificadores de excitação e valência.

Após a etapa inicial, a árvore foi construída baseada nos resultados destes dados de validação e usando a metodologia apresentada na Seção 3.3. Uma vez que a estrutura da árvore foi montada, dados de testes foram avaliados na estrutura proposta como também nos classificadores individuais usando a métrica F1.

A Tabela 3 compara o desempenho dos classificadores para os sentimentos rotulados em classes individuais com o classificador *Ensemble* e também com o resultado obtido pela árvore, para os dados de validação e teste. Observa-se novamente que o classificador *Ensemble* consegue, para maioria das classes, desempenho maior que o obtido para cada tipo de classificador. Isso indica que a técnica de multi-visão, através do *Ensemble*, foi favorável para classificação dos rótulos. Comparando os resultados da árvore com os resultados dos demais classificadores, verifica-se que para os dados de validação a árvore consegue os melhores resultados. Por outro lado, embora ele não consiga sempre os melhores resultados para os dados de teste, na média ele consegue melhores resultados que qualquer outro classificador individual. Isso indica que a combinação dos classificadores sob os seis tipos de rótulos usando uma árvore de decisão auxiliaram na classificação de sentimentos individuais.

Olhando para a Tabela 3 é possível verificar que algumas classes apresentam desempenho inferior nos dados de teste comparado aos dados de validação. Os principais deles são as classes D ("disgust") e F ("fear"), ambas possuem poucas amostras de dados na base Berlin, conforme pode ser visto na tabela 1. Devido a esse desbalanceamento da base, classes minoritárias costumam apresentar resultados mais discrepantes. Pretende-se amenizar esse problema em trabalhos futuros aplicando técnicas como a de *data augmentation* para aumentar a quantidade de dados de treino e validação.

Tabela 3. Resultados de F1 dos dados de validação e teste para as classes individuais

	SVM	CNN-1D	CNN-2D	ENS.	ARVORE	SVM	CNN-1D	CNN-2D	ENS.	ARVORE
A	0,64	0,86	0,53	0,62	**0,89**	0,68	0,86	0,73	**0,89**	0,88
B	0,75	0,52	0,70	0,75	**1,00**	0,57	0,52	0,55	0,88	**0,92**
D	0,44	0,10	0,61	1,00	**1,00**	0,00	0,10	0,87	**0,97**	0,50
F	0,18	0,42	0,49	0,67	**0,83**	0,31	0,42	0,39	**0,95**	0,77
H	0,50	0,20	0,46	0,67	**0,95**	0,42	0,20	0,47	**0,94**	0,88
S	0,50	0,61	0,83	0,67	**0,80**	1,00	0,61	0,92	0,92	**1,00**
N	0,77	0,32	0,81	0,87	**0,96**	0,73	0,32	0,65	**0,94**	0,89
MEDIA:	0,58	0,58	0,62	0,73	**0,92**	0,58	0,58	0,64	0,73	**0,87**

5. Conclusões e Trabalhos Futuros

Neste trabalho é proposto um método de identificação de sentimentos discretos em voz, como medo e alegria. O método combina, por meio de uma árvore de classificação, os resultados de diferentes classificadores treinados para realizar a classificação em vários tipos de rótulos usando diferentes características extraídas dos sinais, usando uma abordagem multi-visão.

Os desempenhos encontrados pela árvore para identificação dos sentimentos foram, na média, superiores a todos os outros classificadores individuais, incluindo um *Ensemble* de classificadores, mostrando a efetividade do método proposto. Também foi observado que o *Ensemble* se mostrou vantajoso em comparação ao desempenho de cada classificador individual. Esses resultados mostram que a análise multi-visão do problema auxiliou no processo geral de classificação.

Considerando que a base de dados Berlin é desbalanceada, em trabalhos futuros pretende-se aplicar algum método para amenizar esse problema. Pretende-se ainda também aplicar a metodologia utilizada para outras bases de dados de tamanhos maiores, já que a base Berlin possui poucas amostras.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 1769586. Os autores gostariam de agradecer ao financiamento do projeto de pesquisa da Fundação de Amparo à Pesquisa do Espírito Santo (FAPES) 93/2017.

Referências

Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5.

Bestelmeyer, P. E., Kotz, S. A., and Belin, P. (2017). Effects of emotional valence and arousal on the voice perception network. *Social Cognitive and Affective Neuroscience*, 12(8):1351–1358.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A Database of German Emotional Speech. *Interspeech*, (January):1517–1520.

Fayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92:60–68.

Gadhe, R. P., Nilofer, S., Waghmare, V. B., Shrishrimal, P. P., and Deshmukh, R. R. (2015). Emotion Recognition from Speech: A Survey. *International Journal of Scientific & Engineering Research*, 6(4):632–635.

Lee, C. C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234(December 2016):11–26.

Livingstone, S. R. and Russo, F. A. (2018). *The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english*, volume 13.

MAO, Q., WANG, X., and ZHAN, Y. (2010). Speech Emotion Recognition Method Based on Improved Decision Tree and Layered Feature Selection. *International Journal of Humanoid Robotics*, 07(02):245–261.

Parthasarathy, S. and Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-Task learning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017-Augus:1103–1107.

Pathak, S. and Kolhe, V. (2016). A Survey on Emotion Recognition from Speech Signal. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(7):447–450.

Reddy, A. P. and Vijayarajan, V. (2017). Extraction of Emotions from Speech - A Survey. *International Journal of Applied Engineering Research ISSN*, 12(16):973–4562.

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

Shen, P., Changjun, Z., and Chen, X. (2011). Automatic Speech Emotion Recognition using Support Vector Machine. *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference*, 2:621–625.

Shih, P.-Y., Chen, C.-P., and Wu, C.-H. (2017). SPEECH EMOTION RECOGNITION WITH ENSEMBLE LEARNING METHODS Po-Yuan. pages 2756–2760.

Tuarob, S., Tucker, C. S., Salathe, M., and Ram, N. (2014). An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of Biomedical Informatics*, 49:255–268.

Xia, R. and Liu, Y. (2017). A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space. *IEEE Transactions on Affective Computing*, 8(1):3–14.

Zhang, S., Zhang, S., Huang, T., and Gao, W. (2018). Convolutional Neural Network and Discriminant. 20(6):1576–1590.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47:312–323.