UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Departamento de Sistemas de Computação

Classificação de emoções no discurso humano

Eduardo Santos Medeiros de Vasconcelos

São Carlos – SP

Classificação de emoções no discurso humano

Eduardo Santos Medeiros de Vasconcelos

Orientador: Jó Ueyama

Monografia referente ao projeto de conclusão de curso dentro do escopo da disciplina SCC0670 – Projeto de Graduação I do Departamento de Sistemas de Computação do Instituto de Ciências Matemáticas e de Computação – ICMC – USP para obtenção do título de Engenheiro de Computação.

Área de Concentração: Inteligência Computacional

USP – São Carlos
31 de maio de 2016

"As invenções são, sobretudo, o resultado de um trabalho teimoso."

(Alberto Santos Dumont)

Dedicatória

Dedico este trabalho à memória de meu avô, Gismundo Pereira de Vasconcelos, Seu Mundinho, que viveu para o serviço à família e aos valores cristãos.

Agradecimentos

Agradeço aos meus pais, Fernando e Lúcia, que me ensinaram, desde muito cedo, através do exemplo, o valor do trabalho honesto e do senso de dever.

Agradeço à minha avó, Cícera, exemplo de simplicidade, entrega e bom humor, que sempre apoiou a escolha do neto de perseguir um sonho a mais de dois mil quilômetros de casa, apesar de não conseguir compreendê-la.

Ao meu irmão, Lucas, por ter conseguido me fazer entender que o que se ama deve vir antes de qualquer coisa na vida.

À Universidade de São Paulo e ao contribuinte do Estado de São Paulo, a quem espero conseguir, um dia, retribuir pela grande oportunidade que me foi oferecida.

Ao meu orientador, Jó Ueyama, que acreditou em mim desde o meu primeiro ano nesta instituição.

A Leandro Mano, em especial, e também a Bruno Faiçal e a Heitor Freitas pela ajuda indispensável no desenvolvimento deste projeto.

Finalmente, a Bia, por todo o apoio e pela serenidade durante as madrugadas em claro que antecederam a entrega deste trabalho.

Resumo

Na área de Interação Humano-Computador (IHC), tornam-se cada vez mais comuns estudos que mostram a importância de se levar em conta o estado emocional do usuário na interação com um sistema computacional. Um dos maiores desafios da área, no entanto, é a aferição de estados emocionais humanos, de forma que um grande número de trabalhos vem sendo feito no âmbito do reconhecimento de emoções. Dentre os diferentes tipos de expressões emocionais estão as expressões vocais. Classificadores do tipo *Support Vector Machine* (SVM) são os mais comuns no estado da arte em reconhecimento de emoções humanas expressas através de elocuções. Ainda assim, é comum a publicação de trabalhos que adotam outras estratégias de classificação de emoções e, apesar da imensa capacidade de generalização dos classificadores SVM isolados, vem crescendo o uso de conjuntos de classificadores para desempenhar tarefas de seriação. Neste trabalho, explora-se o uso de um comitê de classificação no contexto da seriação de emoções humanas a partir de sinais de fala. Utiliza-se uma base de dados pública constituída de gravações de áudio de seis indivíduos diferentes e que contém exemplos das sete emoções humanas básicas. Das gravações de áudio, extrai-se um conjunto de instâncias constituídas de *features* acústicas clássicas e um rótulo correspondente à emoção expressa. O conjunto é, então, utilizado no projeto de um comitê de classificação. O desempenho do comitê obtido é comparado ao de seus componentes e constata-se que ele obtém performance de classificação superior a cada um de seus componentes isolados.

Sumário

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

Segundo Panksepp (1982), a capacidade humana de expressar emoções é compartilhada por todos os mamíferos. Uma região particular do cérebro, denominada lobo límbico, cujo desenvolvimento deu-se muito cedo na evolução dos representantes da classe *Mammalia*, é responsável pela síntese de emoções e pelas respostas comportamentais e psicológicas ao meio. Até mesmo os mamíferos inferiores, ainda que de forma menos complexa, são capazes de manifestar-se emocionalmente.

As emoções permitem que um indivíduo aprecie ou desgoste de experiências, participe de interações sociais e são, por vezes, fatores determinantes em suas decisões cotidianas. De acordo com Zhou et al. (2011), os fatores emocionais das decisões tomadas por um indivíduo são um aspecto muito importante de sua interação com o meio.

O conhecimento acerca do estado emocional de um indivíduo pode ser utilizado para auxiliar seu processo decisório em situações cotidianas. Na verdade, as decisões tomadas por um ser humano não apenas podem ser auxiliadas por uma entidade (humano ou máquina), mas em algumas situações, quando há um estado emocional desfavorável, pode ser até preferível que outra entidade tome a decisão pelo indivíduo. Uma entidade capaz de discernir entre as emoções de um indivíduo poderia, por exemplo, sugerir uma música que aliviasse o estresse associado ao seu estado emocional ou impedir que um indivíduo em estado emocional adverso tome uma má decisão ao conduzir um automóvel ou uma aeronave e acabe causando um acidente. A ideia de endossar a decisão de uma máquina em detrimento da decisão de um humano numa situação de risco iminente pode parecer abismal, mas é preciso lembrar que já é comum confiarmos a máquinas a tomada de decisão em algumas situações muito delicadas. No domínio aeronáutico, por exemplo, os sistemas anticolisão *Traffic Collision Avoidance System* de segunda geração (TCAS II), largamente utilizados nas aeronaves mais modernas, tomam sozinhos decisões inerentes às razões de subida ou descida de aeronaves em rota de colisão de maneira a evitar desastres aéreos.

Já há muito tempo, na psicologia, estuda-se a influência de fatores emocionais nas decisões humanas (LICHTENSTEIN, A. et al., 2008). Na área de Interação Humano-Computador (IHC), no entanto, há apenas alguns anos que o número de estudos envolvendo emoções humanas vem apresentando crescimento. Uma das vertentes desse campo investiga como fazer com que sistemas computacionais reconheçam e reajam a emoções humanas, de forma a suprir as necessidades interativas do usuário (MANO, L. Y. et al., 2015). Um dos grandes desafios da área é que emoções humanas são extremamente difíceis de mensurar e de modelar. Entre os tipos de expressão emocional humana estão as expressões motoras (MAHLKE, S.; MINGE, M., 2008; PETER, C.; URBAN, B., 2012), compartilhadas por todos os grupos humanos e que englobam as expressões faciais, corporais e orais.

Módulos capazes de distinguir entre estados emocionais são a base de qualquer sistema computacional que se proponha a fornecer saídas baseadas em emoções e as expressões motoras, em especial, são alvo de grande interesse da comunidade científica (MANO, L. Y. et al, 2015; ZHOU, F. et al., 2011; SCHULLER, B. et al, 2005; RAMAKRISHNAN, S.; EMARY, I. M. E., 2013; MORRISON, D. et al., 2007; EKMAN, P., 1973). Este trabalho se concentra em um tipo específico de expressão motora, a fala (ou discurso humano). Modelos baseados em SVM são bastante comuns no estado da arte em reconhecimento de emoções humanas expressas através de elocuções (CHAVHAN, Y. et al., 2010; HASSAM, A.; DAMPER, R. I., 2010; BHARGAVA'A, M.; POLZEHL, T., 2012; PEIPEI, S. et al., 2011). A grande vantagem dos modelos baseados em SVM em relação a outras estratégias de classificação é a sua maior capacidade de generalização do aprendizado, o que significa que tendem a obter performance superior a outros classificadores ao determinar a classe de instâncias de dados jamais vistas. Em outras palavras, quando uma quantidade pequena de instâncias de treinamento está disponível, classificadores SVM obterão uma performance superior a outros classificadores na seriação de instâncias de teste. No entanto, Morrison et al. (2012) e Mano et al. (2015), relatam resultados de classificação de expressões motoras ligeiramente superiores aos obtidos por modelos de SVM isolados, fazendo uso de decisão ponderada por comitês de classificação. Inspirado nisso, este trabalho explora o uso de um comitê na seriação de emoções humanas a partir de sinais de fala.

O autor deste trabalho teve a oportunidade de trabalhar como estagiário de verão durante 3 meses em um laboratório de pesquisa na República da Irlanda, em 2014, pelo Programa Ciência Sem Fronteiras, sob a supervisão de uma especialista em reconhecimento de emoções no discurso humano, Drª. Naomi Harte (*Trinity College Dublin*). Durante este período, o autor desenvolveu um projeto cujo objetivo era o reconhecimento de segmentos de risada em sinais de discurso humano utilizando SVM. Após regressar ao Brasil, viu a oportunidade de aproveitar o conhecimento que adquirira no domínio para desenvolver este trabalho junto ao grupo de pesquisa do Prof. Dr. Jó Ueyama, que fora seu orientador em um projeto de Iniciação Científica. Um dos integrantes do grupo, Me. Leandro Yukio Mano Alves, vinha trabalhando com o reconhecimento de emoções expressas pela face empregando comitês de classificação, ao que o autor enxergou a oportunidade de expandir a pesquisa do grupo, através do desenvolvimento deste trabalho, utilizando a expressão vocal de emoções.

Este projeto tem como potenciais contribuições: (i) fomentar o debate acerca do estudo de métodos de classificação de emoções humanas utilizando *features* obtidas a partir de métodos de sensoriamento não invasivos (como é o caso dos sinais de fala); (ii) complementar o trabalho de Mano et al. (2015) de classificação de emoções baseada em expressões faciais no qual se baseia através da constatação de que a mesma abordagem funciona para expressão oral de emoções ou ainda; (iii) constatar que semelhante abordagem não funciona para a expressão oral.

1.2. Objetivos

Este trabalho tem como objetivo conceber um classificador de emoções humanas expressas na fala. O estado da arte em classificação de emoções a partir do discurso humano emprega constantemente modelos de seriação por SVM. O classificador proposto neste trabalho será baseado em um modelo de comitê de classificação, de forma a experimentar, no contexto da classificação de emoções expressas através da fala, a mesma estratégia de classificação recentemente utilizada com sucesso para emoções expressas através da face no principal trabalho em que este projeto se baseia (MANO, L. Y. et al., 2015). Assim, levanta-se a hipótese de que é possível atingir, com um comitê de

classificação, resultados superiores (ou ao menos equiparáveis) aos obtidos pelos modelos que o constituem na classificação de emoções expressas através do discurso humano, assim como o trabalho que inspira este projeto faz para emoções expressas através da face. Adicionalmente, este trabalho tem como objetivo expandir a funcionalidade do modelo de classificador de emoções proposto no trabalho direcionador, permitindo que, futuramente, a configuração de comitê aqui desenvolvida para classificação de expressões orais possa ser cruzada com aquela existente para a classificação de expressões faciais, de maneira a conceber um módulo com capacidade de reconhecimento de emoções superior aos dois comitês individualmente.

1.3. Organização do Trabalho

Este trabalho está organizado em 4 capítulos, sendo este o primeiro. No CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA, serão apresentadas a terminologia básica da área e a literatura relacionada a este projeto. No CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO, o projeto será discutido em detalhes. Sua execução e resultados serão apresentados. Finalmente, no CAPÍTULO 4: CONCLUSÃO, serão apresentadas as conclusões e contribuições deste trabalho, sugeridos os trabalhos futuros e feitas, também, algumas considerações sobre o curso de graduação no qual o autor está atualmente matriculado, bem como será mostrada a relação deste trabalho com o curso.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Neste capítulo são apresentados os conceitos e a terminologia empregados na literatura relacionada à classificação de emoções humanas. É apresentada a fundamentação teórica pertinente ao domínio das emoções humanas através de discussões a respeito da Teoria Componencial das Emoções e do Espaço Emocional Semântico. São discutidos os mecanismos de classificação de emoções, bases de dados de emoções, os tipos de *features* utilizados no trabalho e é discutida, ainda, a captura de expressões orais.

2.2. Teoria Componencial das Emoções

Emoções podem ser entendidas como reações ao meio que abrangem todo o organismo do indivíduo. Devido à enorme complexidade das emoções, estudiosos do tema as dividem em componentes mais simples, ao que se dá o nome de Teoria Componencial das Emoções (FONTAINE, J. R. et al., 2002; SCHERER, K. R., 2005). A Teoria Componencial das Emoções entende que uma emoção é caracterizada pela interação de 5 atributos fortemente inter-relacionados: a avaliação cognitiva, a experiência subjetiva, a tendência comportamental, as reações fisiológicas e as expressões motoras. Os componentes serão discutidos a seguir.

- Avaliação Cognitiva – é o componente emocional relacionado à forma como o indivíduo avalia o meio (FONTAINE, J. R. et al., 2002) através de seus recursos sensoriais, cognitivos e perceptivos (inclui-se aí o raciocínio do indivíduo). Por isso, uma mesma situação pode desencadear emoções diferentes em indivíduos distintos, a depender da interpretação que cada um dá aos eventos aos quais reage (SCHERER, K. R., 2005; MAHLKE, S.; MINGE, M., 2008).
- Tendência Comportamental – é o componente responsável por orientar as ações do indivíduo antes de (ou até durante) um acontecimento (FONTAINE, J. R. et al., 2002), indicando a intensidade das reações emocionais necessárias (EKMAN P.,

2006). Esse componente regula, por exemplo, o quanto o indivíduo se sente confortável para pedir ajuda ou para fazer propostas criativas numa interação social, como a execução de uma tarefa em grupo (FONTAINE J. R. et al., 2002; MAHLKE, S. ; MINGE, M., 2008).

- Reações Fisiológicas – é o componente emocional que permite ao indivíduo revelar suas emoções de maneira consciente ou inconsciente. Esse componente participa da regulação do sistema nervoso central, que controla o sistema motor, do sistema neuroendócrino, que é o responsável pela normalização de alterações no organismo e do sistema nervoso autônomo, responsável pela respiração, circulação sanguínea, digestão, etc. (STEMMLER, G., 2003; SCHERER, K. R., 2005; MAHLKE, S.; MINGE, M., 2008).
- Experiência Subjetiva – este componente se relaciona com a avaliação que o próprio indivíduo faz da emoção que sente e, por isso, está ligado à capacidade de autorregulação emocional (DESMET, P., 2003). Portanto, esse componente pode ser entendido como uma meta experiência. Seu objetivo é monitorar o estado do corpo e a maneira como o indivíduo se relaciona com o ambiente (SCHERER, K. R., 2005) através de suas reações emocionais. Em suma, é um componente de retroalimentação emocional.
- Expressões Motoras – é o componente responsável por fazer notar as tendências comportamentais (SCHERER, K. R., 2005; MAHLKE, S.; MINGE, M., 2008; XAVIER, R. A. C. et al., 2012). As expressões motoras são as expressões faciais, orais e gestuais que correspondem ao estado emocional do indivíduo. Diferentes estados emocionais resultam em expressões motoras distintas: a fala, por exemplo, sofre alterações de velocidade de discurso, intensidade e melodia do som (LEVENTHAL, H., 1984). Na face, por outro lado, observa-se, por exemplo, o franzir de testa, o sorrir e o arregalar de olhos, todos relacionados a estados emocionais.

Este projeto se concentra na identificação de características emocionais inerentes às Expressões Motoras, especificamente à expressão emocional através da fala.

2.3. Espaço Emocional Semântico

Até mesmo os humanos podem encontrar dificuldade em identificar com exatidão o estado emocional de outros indivíduos. Por vezes, até o próprio estado emocional. Assim, para cogitar a classificação de emoções humanas utilizando uma máquina, é preciso definir uma forma de quantificá-las. Há alguns modelos propostos para auxiliar os psicólogos na avaliação de experiências emocionais (RUSSELL, J. A., 1980; SCHERER, K. R., 2005).

De acordo com Russell (1980), existem duas visões sobre a representação de emoções: uma chamada "contínua" e a outra, "categórica". Segundo a visão contínua, um estado emocional caracteriza-se por um conjunto multidimensional de atributos sobre um espaço contínuo. A visão categórica, por outro lado, vê os diferentes estados emocionais como entidades bidimensionais de um espaço contínuo que pode ser representado por um plano cujos eixos denominam-se "valência" (que pode ser positiva ou negativa) e "excitação" (que corresponde ao *nível de energia* associado à emoção). O modelo tem dois domínios principais, que correspondem às emoções "prazerosas" e às "não prazerosas". O modelo de quantificação emocional proposto pela visão categórica chama-se *circumplex* (RUSSELL, J. A., 1980).

Estudos de diferentes grupos culturais (EKMAN, P., 1973) indicam que, das várias emoções que podem ser mapeadas através do modelo *circumplex*, seis constituem um conjunto básico de emoções a partir das quais todos os estados emocionais podem ser derivados: alegria, tristeza, raiva, medo, desgosto e surpresa. Qualquer ser humano, independente do grupo étnico a que pertence, é capaz de perceber estas emoções básicas da mesma forma. Uma maneira de representar o modelo *circumplex* é através de uma circunferência sobre a qual as emoções básicas são dispostas de acordo com seus valores de valência e excitação, sendo que o eixo da valência tem sentido de crescimento do valor "desagradável" para o valor "agradável" e o eixo de excitação tem sentido de crescimento do estado emocional de menor energia para o estado emocional de maior energia. As demais emoções podem ser derivadas destes 6 estados emocionais básicos, variando seus níveis de valência e excitação. A Figura 1 apresenta uma representação do modelo *circumplex* segundo Russell (1980).

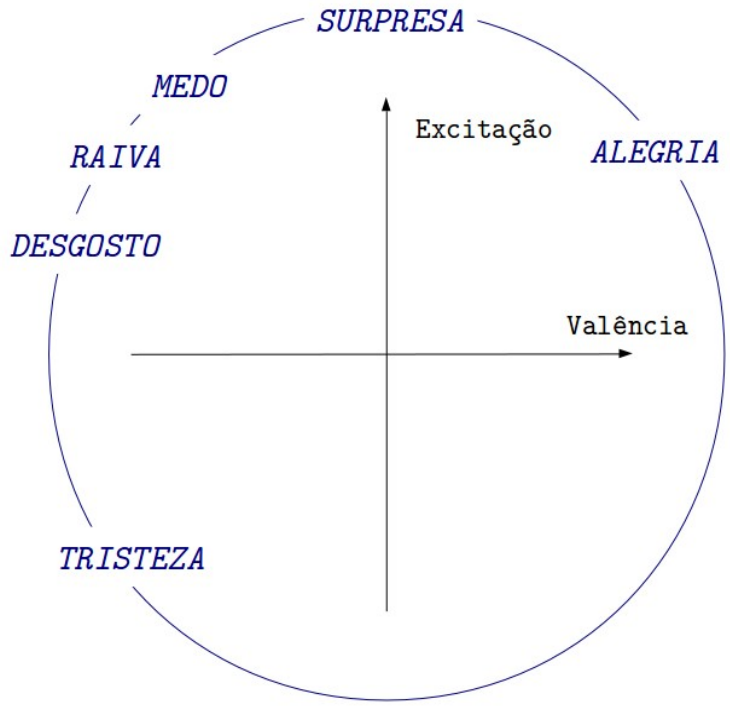

Figura 1 – Representação do circumplex. *Adaptado de Russell (1980).*

No dia a dia é comum que as pessoas utilizem um sistema categórico para identificar emoções, de forma que é mais intuitivo utilizar um modelo semelhante ao *circumplex* na abordagem do problema de classificar emoções humanas computacionalmente. O estado emocional neutro, que também aparece na literatura da área, é adicionado ao modelo utilizado neste trabalho.

Segundo Picard (1998), a difusão da computação afetiva vem motivando cada vez mais pesquisas na análise de estados emocionais. A informação correspondente à resposta emocional do indivíduo já vem sendo usada em inúmeras aplicações.

É sabido que cada emoção humana está relacionada a um tipo de expressão motora e que as expressões motoras compreendem as expressões orais, faciais e corporais (DESMET, P., 2003). Assim, é possível extrair características diferentes de expressões referentes a estados emocionais distintos.

Há décadas que a área de *Machine Learning (*ML) vem provendo à comunidade científica, ferramentas de classificação que podem ser empregadas nos mais diversos

domínios. Os modelos propostos em ML são capazes de generalizar o aprendizado baseado em conjuntos de padrões cujos rótulos (ou classes) são conhecidos para seriar dados cujos rótulos são desconhecidos (DUDA, P. E.; RICHARD, O. H., 1973; CLARK, P.; NIBLETT, T., 1989; CRISTIANINI, N.; SHAWE-TAYLOR, J., 2000; UEYAMA, J. et al., 2014).

2.4. Mecanismos de Classificação de Emoções

O emprego das técnicas propostas em ML vem ganhando espaço na classificação de expressões emocionais humanas (ØHRN, A.; ROWLAND, T., 2000; ZHOU et al., 2011; CHANEL et al., 2009). Algumas das técnicas já empregadas com sucesso no domínio da classificação de emoções (e que são utilizadas neste trabalho) serão apresentadas de maneira sucinta a seguir.

- *k-Nearest Neighbours* (kNN) – modelo de classificador que não assume, de antemão, nenhuma distribuição de probabilidade para os dados a classificar (MANO, L. Y. et al., 2015). Definido o número de vizinhos k com o qual se deseja trabalhar, o kNN encontra as k instâncias de treinamento mais próximas da instância de teste que se deseja classificar e atribui-lhe um rótulo igual ao rótulo predominante entre seus k vizinhos encontrados.
- *Support Vector Machine* (SVM) – é o modelo de classificador que, se bem configurado, resulta na melhor fronteira de separação entre as classes a distinguir (CLARK P.; NIBLETT T., 1989; LITTLEWORT et al., 2011; RAMAKRISHNAN S.; EMARY I. M., 2013). O SVM é um classificador baseado na minimização do erro empírico de classificação, isto é, o erro associado a exemplos desconhecidos, o que lhe garante a maior capacidade de generalização de aprendizado dentre todas as estratégias de classificação.
- Árvore de Decisão – modelo de classificador baseado numa árvore que cresce a partir dos atributos que agregam mais informação às instâncias de treinamento, isto é, quanto mais informação um atributo agregar à seriação de um exemplo de treinamento, mais próximo à raiz da árvore o atributo é colocado (DUDA, P. E.; RICHARD, O. H., 1973).

- Classificador Bayesiano – modelo de classificador baseado em probabilidades. Um classificador Bayesiano afere as dependências condicionais de variáveis aleatórias e as representa sob a forma de um grafo acíclico (FACELI, K., 2011). No contexto deste trabalho, um classificador Bayesiano poderia, por exemplo, verificar que a concentração espectral de energia em bandas de baixa frequência de um sinal de voz resulta numa probabilidade alta de que a instância seja de determinada emoção.
- *Multilayer Perceptron* (MLP) – modelo de classificador baseado em redes neurais. *Multilayer Perceptron* é a generalização do *Perceptron* para classificar dados não linearmente separáveis. Uma rede MLP tem uma camada de entrada, uma ou mais camadas escondidas e uma camada de saída, todas constituídas por unidades fundamentais denominadas neurônios artificiais, conectados por sinapses, e para seu treinamento utiliza-se o algoritmo *backpropagation* (ainda que já existam outras abordagens de treinamento para redes de arquitetura análoga). Apesar de menos comum na literatura, este modelo também é empregado no reconhecimento de emoções (KHANCHANDANI, K. B.; HUSSAIN, M. A., 2009).

A grande quantidade de particularidades de configuração de cada classificador e a complexidade associada à extração de *features* adequadas para treiná-los faz com que dificilmente se consiga resultados de classificação próximos aos 100% de acurácia (CANUTO, A. M., 2001; MONARD, M. C.; BARANAUSKAS, J. A., 2003). O processo de decisão entre modelos de classificadores no domínio da seriação de emoções, geralmente, dá-se pela construção de diversos modelos de classificadores, a escolha daqueles que obtêm os melhores desempenhos e o descarte daqueles cujo desempenho é ruim (CANUTO, A. M., 2001). A seleção de um classificador isolado dentre muitos modelos de classificadores acarreta, por vezes, no descarte de classificadores cujo desempenho é promissor (SCHULLER, B. et al., 2005).

Baseado na ideia de que conjuntos de classificadores podem conduzir a melhoras na capacidade de generalização do aprendizado e também da acurácia de classificação, vem ganhando força o conceito de comitê de classificação (também conhecido na literatura como *ensemble classifier*, combinação de especialistas e classificador modular) (DUDA, R. O. et al. 2012). A garantia de bom desempenho de um comitê está justamente nas diferenças entre seus membros (DUDA, R. O. et al. 2012). Classificadores cujos erros não

são correlacionados, isto é, que não erram para os mesmos tipos de padrões, tendem a apresentar uma correção mútua de seus erros. A "opinião" de outros membros do comitê de classificação tende a "ajudar" os demais classificadores com seus pontos fracos. Assim, um comitê de classificação deve ser formado por classificadores cujas acurácias sejam todas razoáveis e cujos erros de classificação não se deem para os mesmos tipos de instâncias (SCHULLER, B. et al., 2005). A Figura 2 ilustra a arquitetura de um comitê genérico em acordança o trabalho de Schuller et al. (2005).

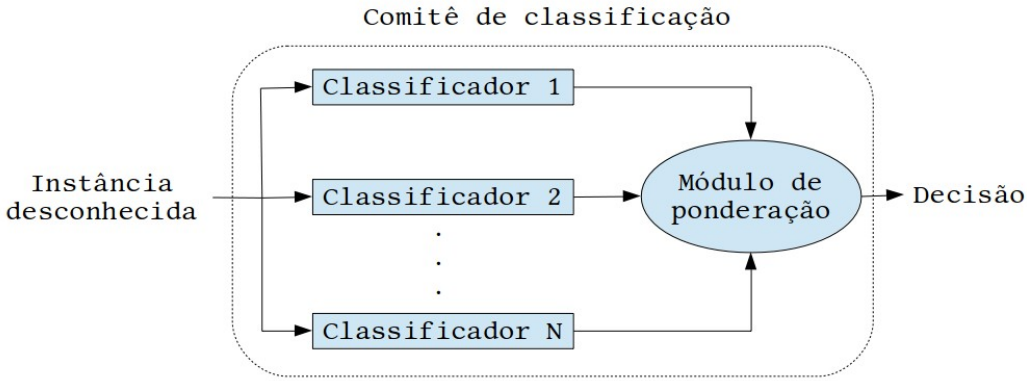

Figura 2 – Esquema de um comitê de classificação. Adaptado de Schuller et al. (2005).

Cada classificador é treinado de maneira isolada previamente à montagem do comitê. Ao ser apresentada uma instância desconhecida ao comitê de classificação, cada um de seus componentes (Classificadores de 1 a N) emite sua opinião a respeito da instância, ou seja, envia ao módulo de ponderação do comitê o rótulo da classe ao qual acredita que a instância desconhecida pertence. O módulo de ponderação, por sua vez, cruza os rótulos recebidos de cada classificador e emite a decisão final do comitê. O cruzamento de informação é feito seguindo uma regra definida após o treinamento dos classificadores, quando suas acurácias para os dados de treinamento são conhecidas e suas capacidades de generalização podem ser estimadas. Uma maneira de cruzar a informação é atribuir a cada Classificador i do comitê um peso p_i. Uma forma de atribuir pesos aos classificadores é utilizando a Equação 1, usada no trabalho de Mano et al. (2015), onde a_i representa a acurácia do Classificador i.

$$p_i = \frac{a_i}{\sum_{j=1}^{N} a_j} \quad (1)$$

Assim, o módulo de ponderação fornece como decisão a classe cuja somatória dos pesos dos classificadores que a escolheram for maior.

Uma ferramenta muito utilizada em ML e empregada neste projeto é o WEKA (HALL, M. et al., 2009), que implementa uma série de algoritmos de aprendizado de máquina, dentre os quais, os citados aqui, oferece recursos de visualização de dados, análise fatorial de algoritmos, análise estatística de resultados, etc. O WEKA oferece a opção de trabalhar, na fase de configuração dos algoritmos de ML, com uma interface gráfica bastante intuitiva. Posteriormente, na fase de implementação da aplicação que utilizará os modelos obtidos, é possível importar toda a funcionalidade do WEKA para um programa em Java através de uma biblioteca distribuída junto à própria ferramenta.

2.5. Bases de Dados de Emoções

O primeiro passo na elaboração de um modelo de classificador de emoções é a aquisição de uma base de dados de emoções. Como o foco deste trabalho é a classificação de emoções expressas no discurso humano, deseja-se, especificamente, uma base de dados de expressão oral das emoções. No contexto da expressão emocional oral existem dois tipos recorrentes de bases de dados: as de discurso atuado e as de discurso natural (VERVERIDIS, K.; KOTROPOULOS, C., 2003).

As bases de dados de discurso atuado são aquelas nas quais as elocuções são feitas por atores. Cada ator da base pronuncia diversas frases e, a cada frase pronunciada, o ator expressa uma determinada emoção. A grande vantagem de bases de dados desse tipo é sua simplicidade, pois não é necessário preprocessar as gravações disponíveis, uma vez que a emoção expressada numa gravação é conhecida previamente à gravação propriamente dita e, assim, a base de dados pode ser organizada em diretórios que dividem as gravações disponíveis por ator e/ou por emoção expressada.

As bases de dados de discurso natural, por outro lado, são mais complexas. Podem ser constituídas por uma infinidade de tipos diferentes de gravações, como gravações de *broadcast* de rádio, áudio de sinais de televisão, conversas gravadas em escritórios ou residências, palestras, aulas, entrevistas, etc. Não há nenhum controle sobre o teor emocional das gravações. Assim, além das gravações de áudio, a base vem com uma série de arquivos contendo transcrições que informam os intervalos de tempo contidos em cada arquivo nos quais ocorrem expressões orais de emoções, de maneira que previamente à extração de *features*, deve ser feito o *clipping* (isto é, o "recorte") dos arquivos de áudio, para que seja possível trabalhar apenas com os segmentos que contêm expressões orais de emoções e não outros tipos de vocalizações.

Após a análise de algumas bases de dados, optou-se pelo uso da base EMOVO (CONSTANTINI, G. et al., 2014). A base EMOVO contém, ao todo, 588 gravações de áudio em língua italiana, sendo que 6 atores (3 mulheres e 3 homens) pronunciam 98 frases cada um. Cada ator pronuncia 14 frases expressando cada uma das 7 emoções humanas básicas definidas pela literatura: raiva, alegria, tristeza, desgosto, neutralidade, medo e surpresa. A escolha da base EMOVO deu-se por conter gravações de discurso atuado, o que elimina a necessidade de preprocessamento, e por ser uma base de dados aberta, bem documentada, mas, apesar disso, simples. Ademais, as bases de dados com discurso atuado possuem boa aceitação na comunidade científica (VERVERIDIS, K.; KOTROPOULOS, C., 2003).

2.6. *Features* Acústicas

Independente da abordagem do modelo de classificação, se de classificador isolado ou de comitê de classificação, é necessário extrair *features* da base de dados utilizada, a fim de treinar o(s) classificador(es) do modelo.

Dois tipos de *features* foram extraídos da base de dados de fala utilizada neste trabalho. De cada segmento de 25 ms – o intervalo de 25 ms será explicado adiante – dos arquivos de áudio da base foram extraídos 12 coeficientes *cepstrais* de frequência em mel, cuja sigla em inglês é MFCC, de *Mel Frequency Cepstral Coefficient,* e 1 coeficiente de energia do sinal, especificamente do tipo *log energy*. A partir desse conjunto de 13

coeficientes, um conjunto de 13 coeficientes delta foi obtido e do conjunto de 13 coeficientes delta, um conjunto de 13 coeficientes de aceleração foi obtido. Assim, cada instância é composta por 39 *features* mais o rótulo que caracteriza a emoção que a define. Esse *set* com 39 *features* (12 MFCCs, 1 *log energy*, 13 deltas e 13 acelerações) é uma sugestão encontrada na literatura (JURAFSKY, D.; MARTIN, J. H., 2009).

A diferença entre um sinal de voz expressando uma emoção ou outra está, sobretudo, na forma como o espectro de energia do sinal se comporta. Emoções cuja expressão vocal é caracterizada por exaltações e brados, como a raiva e o medo, tendem a carregar mais energia do que emoções cuja expressão é mais branda, como o desgosto e a tristeza, por exemplo. É exatamente isso que mostram os espectrogramas exibidos na Figura 3, de uma gravação de áudio da frase em italiano "Gli operari si alzano presto." ("Os operários levantam-se cedo.") pronunciado por um indivíduo do sexo feminino com diferentes conteúdos emocionais (para as 6 emoções humanas básicas).

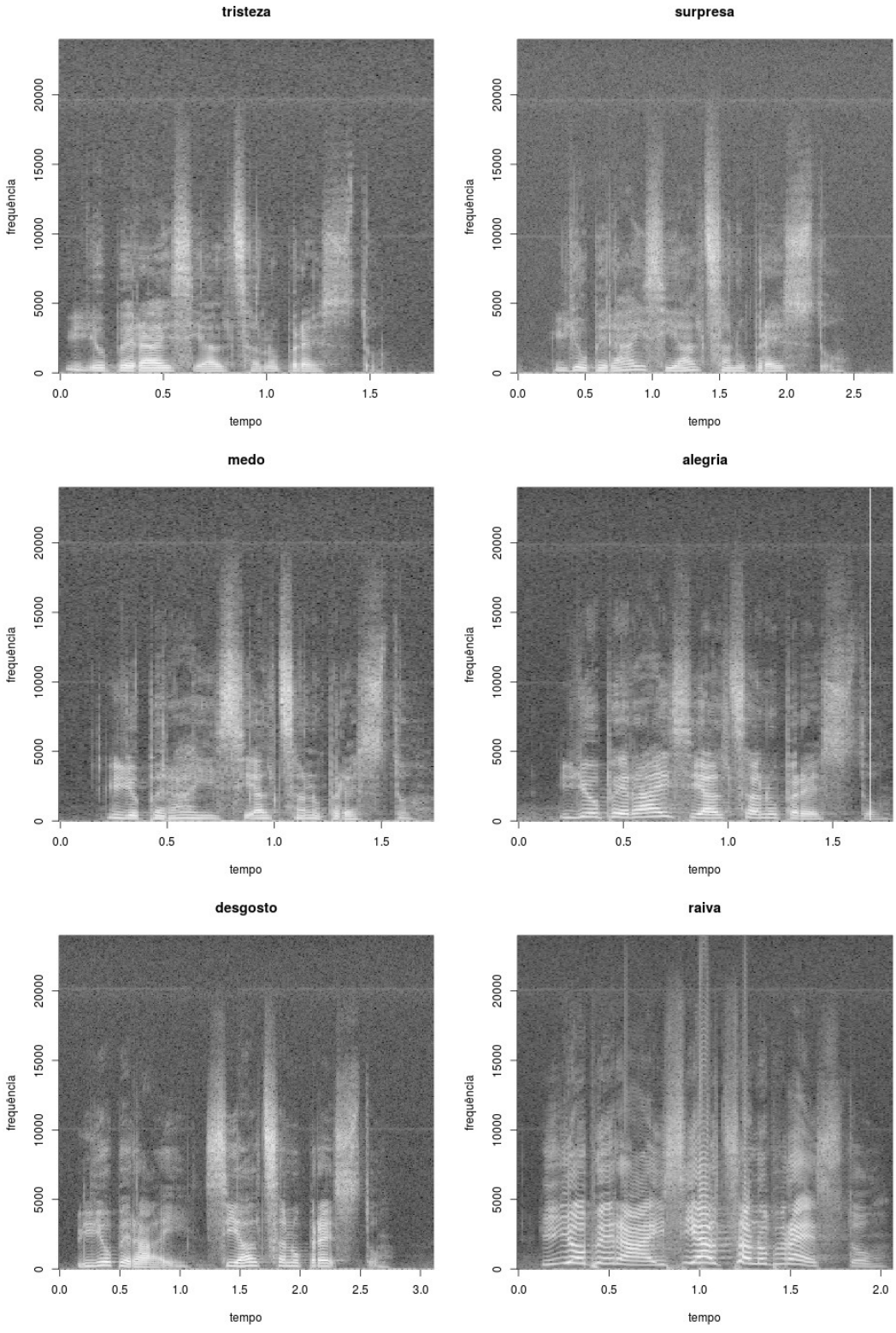

Figura 3 – Espectrogramas da frase "Gli operari si alzano presto." para as emoções básicas.

Detalhes adicionais acerca das *features* e de sua extração serão apresentados nos tópicos seguintes.

2.6.1. Coeficientes *Cepstrais* de Frequência em Mel

Os coeficientes *cepstrais* de frequência em mel (MFCCs, na sigla em inglês) são um dos tipos de *feature* mais comumente utilizados em aplicações de classificação de emoções expressas no discurso humano (PAO, T. et al., 2005), bem como em aplicações de processamento de voz (JURAFSKY, D.; MARTIN, J. H., 2009). Nesta seção descreve-se o processo de extração de MFCCs a partir de um sinal de voz. Discute-se, também, o tipo de informação que carregam.

Suponha-se que o sinal de áudio x(n) contínuo no tempo foi amostrado utilizando a frequência de amostragem $f_{a,}$, resultando no sinal x[n] discreto no tempo, constituído por N amostras igualmente espaçadas de T_a no domínio temporal. É exatamente assim que um sinal de áudio (por exemplo, uma elocução) é representado em um dispositivo digital. A extração de coeficientes MFCC do sinal de áudio x[n] é composta por 6 etapas, sumarizadas na Figura 4.

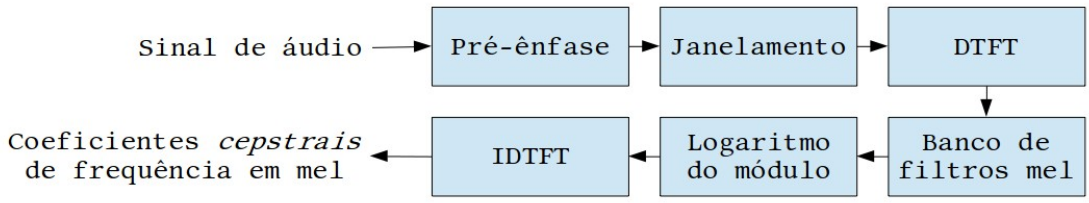

Figura 4 – Diagrama de blocos da extração de MFCCs. Adaptado de Jurafsky e Martin (2009).

O primeiro passo na extração de MFCCs de um sinal é a filtragem *high-pass*, também conhecida por pré-ênfase. O propósito da pré-ênfase é aumentar a amplitude do sinal nas bandas de alta frequência e diminuí-la nas bandas de baixa frequência, pois devido à forma como o trato vocal humano emite sons, as frequências baixas tendem a reter mais energia (vide tópico 2.6.2. Energia do Sinal de Voz) do que as frequências altas.

Assim, o intuito da filtragem é fazer com que a informação disponível nos componentes de baixas frequências do sinal manifeste-se mais intensamente no modelo acústico. A filtragem é realizada por um filtro *Finite Impulse Response* (FIR) de primeira ordem com coeficiente de memória $0,9 \leq a \leq 1,0$ definido pela Equação 2.

$$y[n]=x[n]+ax[n-1] \tag{2}$$

O segundo passo na extração de MFCCs é o janelamento do sinal. A natureza do espectro de uma elocução muda muito rápido, de forma que um sinal de voz humana é dito "não estacionário", o que significa que suas propriedades estatísticas não são constantes no decorrer do tempo. No entanto, pode-se assumir que o sinal é estacionário para intervalos de tempo muito pequenos – de alguns milissegundos – e assim, a partir desses intervalos pequenos, extrair informação concisa a respeito do sinal. Uma forma de obter esses intervalos aproximadamente estacionários é submeter o sinal a um janelamento.

Para a operação de janelamento, a literatura define como padrão a janela de Hamming com duração de 25 ms. O janelamento do sinal x[n] é feito tomando cada segmento $s_i[n]$ (i = 0, 1, 2, …) de 25 ms do sinal e multiplicando-o por uma janela de Hamming w[n] de 25 ms de duração. Cada segmento $s_i[n]$ de 25 ms é obtido concatenando os últimos 10 ms do segmento que o precede, $s_{i-1}[n]$, com os próximos 15 ms do sinal, exceto para o segmento $s_0[n]$, para o qual toma-se os primeiros 25 ms do sinal, de forma que se define uma sobreposição de 10 ms entre dois segmentos consecutivos. O produto de cada segmento $s_i[n]$ por uma janela w[n] fornece um *frame* do sinal. A sequência formada por todos os *frames* do sinal resulta no "sinal janelado" y[n] de x[n]. Uma janela de Hamming é definida segundo a Equação 3, na qual L representa o número total de amostras da janela e n o índice da enésima amostra da janela.

$$w[n]=\begin{cases} 0,54-0,46\cos\left(\frac{2\pi n}{L}\right), & 0 \leq n \leq L-1 \\ 0, & c.c. \end{cases} \tag{3}$$

Na prática, a janela de Hamming atenua as bordas dos segmentos sobre os quais é aplicada, evitando descontinuidades entre *frames* consecutivos. Para ilustrar, ao aplicar sobre o sinal constituído por 500 amostras aleatórias com valores entre 0,0 e 1,0 mostrado

na Figura 5, a janela de Hamming constituída por L = 500 amostras, apresentada na Figura 6, obtém-se o sinal apresentado na Figura 7.

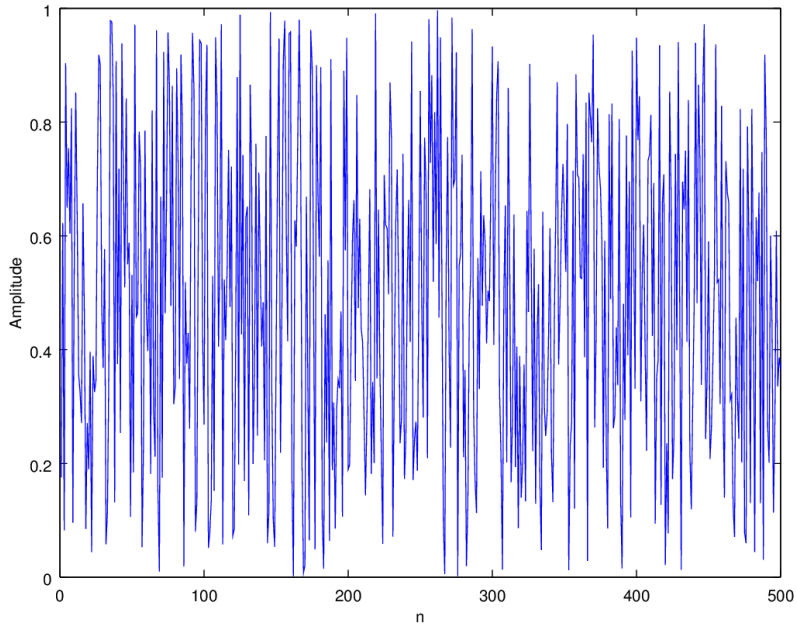

Figura 5 – Sinal digital constituído por 500 amostras aleatórias.

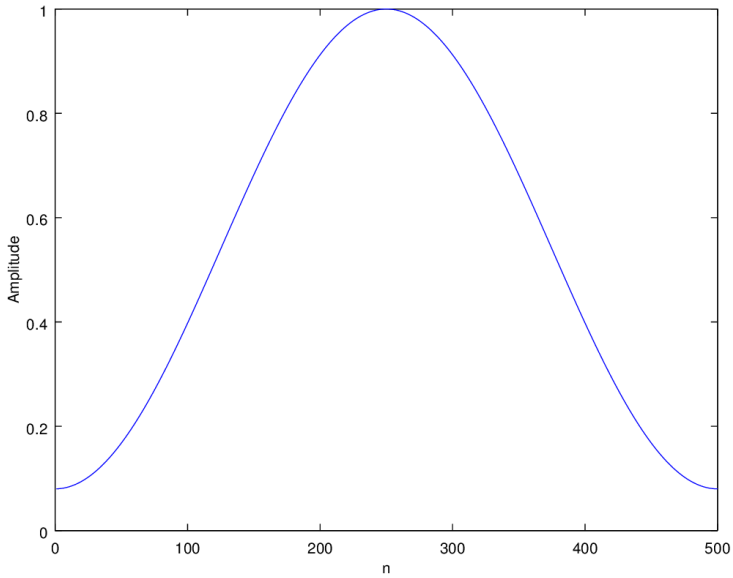

Figura 6 – Janela de Hamming típica, com comprimento 500.

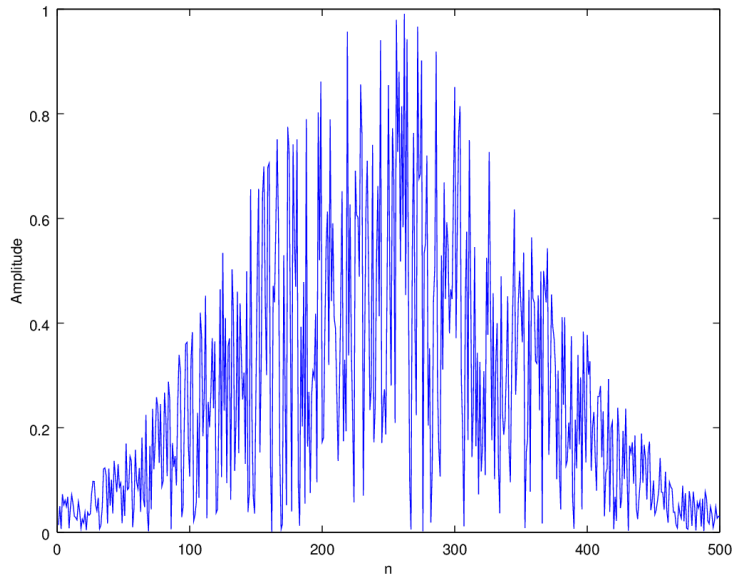

Figura 7 – Sinal resultante da aplicação de uma janela de Hamming.

O terceiro passo na extração de MFCCs é o cômputo da Transformada de Fourier no Tempo Discreto (DTFT, na sigla em inglês) de y[n], que fornece informação a respeito da distribuição de energia do sinal sobre seu espectro de frequências. Formalmente, a DTFT X[ω] de um sinal discreto x[n] de N amostras é definida segundo a Equação 4, onde ω representa um valor de frequência no espectro do sinal, n é o índice da enésima amostra de x[n] e j é a unidade imaginária.

$$X[\omega] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}\omega n} \tag{4}$$

Em seguida, no quarto passo da obtenção de MFCCs, o espectro de frequências Y[ω] obtido do cômputo da DTFT de y[n] é submetido a um banco de filtros mel, ou um banco de filtros *em escala* mel. A escala mel é uma escala de frequências baseada na percepção auditiva humana (SCHULLER, B.; BATLINER, A., 2014). Os seres humanos têm percepção auditiva melhor em baixas frequências, o que significa que variações de amplitude numa onda sonora são percebidas melhor entre os componentes de baixas frequências do que de altas frequências da onda, e a escala mel é definida levando esse fato em consideração. A conversão entre um valor de frequência na escala mel (f_{mel}) para o valor correspondente em Hz (f_{Hz}) é regida pela Equação 5.

$$f_{mel} = 1127 \cdot \ln\left(1 + \frac{f_{Hz}}{700}\right) \quad (5)$$

É importante ressaltar que uma elocução humana é constituída de um conjunto de diferentes amplitudes distribuídas sobre um espectro de frequências, resultante da filtragem realizada pelo trato vocal sobre a onda sonora gerada pelo fluxo intermitente de ar originado da interação da rajada de ar vinda dos pulmões de um falante com suas cordas vocais (FURUI, S., 2000).

Finalmente, um banco de filtros mel é constituído por uma série de filtros triangulares distribuídos segundo a escala mel sobre um eixo de frequências em Hz, de forma que cada filtro triangular atua como um acumulador sobre uma banda de frequência. Ao submeter o espectro Y[ω] a um banco de filtros mel, seus componentes de frequência são ponderados pelos coeficientes dos filtros triangulares do banco e, para cada banda de frequências, ou seja, para cada filtro do banco, obtém-se um coeficiente espectral denominado *melspec*, que é a soma ponderada dos componentes de uma banda de frequência específica. Assim, os coeficientes *melspec* agregam informação a respeito de como a densidade espectral de um sinal está distribuída sobre o espectro de frequências e são tão numerosos quanto os filtros do banco. A Figura 8 apresenta um exemplo de banco de filtros mel.

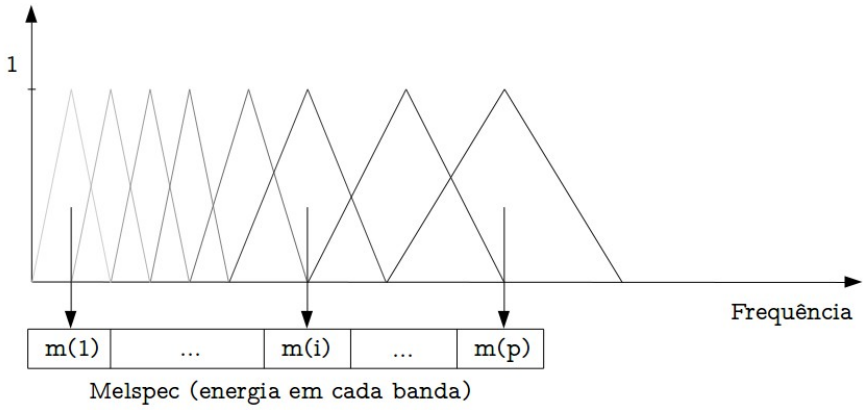

Figura 8 – Típico banco de filtros mel. Adaptado de Jurafsky e Martin (2009).

Na quinta etapa da extração dos MFCCs, calcula-se o logaritmo do valor absoluto dos coeficientes *melspec*. Em seguida, na sexta e última etapa do processo de extração, calcula-se a Transformada Inversa de Fourier no Tempo Discreto (IDTFT) dos valores obtidos na quinta etapa. Coeficientes cepstrais de frequência em mel levam esse nome porque estão relacionados com o conceito de *cepstro* (tradução livre de *cepstrum*) c[n] de um sinal x[n] de N amostras, que é definido pela Equação 6.

$$c[n] = \sum_{n=0}^{N-1} \log\left(\left|\sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}\omega n}\right|\right) e^{j\frac{2\pi}{N}\omega n} \tag{6}$$

A Equação 7 define o *cepstro* de maneira mais compacta:

$$c[n] = IDTFT\{\ \log|DTFT\{x[n]\}|\ \} \tag{7}$$

Definindo de maneira simples, pode-se dizer que MFCCs carregam a informação referente ao formato do trato vocal quando da elocução (JURAFSKY, D.; MARTIN, J. H., 2009), isto é, da função de transferência do filtro que deu origem aos fonemas da elocução em questão.

2.6.2. Energia do Sinal de Voz

Outro tipo de *feature* que aparece de maneira muito corriqueira em trabalhos de classificação de emoções humanas é a energia do sinal de voz no domínio do tempo. Mais especificamente, o logaritmo da energia do sinal de voz.

De acordo com Jurafsky e Martin (2009), a energia E(x) de um sinal discreto genérico x[n] é dada pela Equação 8.

$$E(x) = \sum_{n=0}^{N-1} x[n]^2 \tag{8}$$

Logo, a *feature* logaritmo da energia do sinal de voz LOGE(x) (amplamente conhecida simplesmente por *log energy*) é definida pela Equação 9.

$$LOGE(x) = \log\left(\sum_{n=0}^{N-1} x[n]^2\right) \quad (9)$$

Essencialmente, LOGE(x) é o logaritmo da somatória das potências instantâneas do sinal de áudio a cada amostra e, assim, é de se esperar que tenha valores mais altos para elocuções mais "enérgicas", como aquelas que expressam a raiva e outras emoções cuja expressão oral é caracterizada por exaltações e brados (DESMET, P., 2003).

É importante ressaltar que o cômputo dos coeficientes *log energy* deve acontecer numa correspondência 1:1 com os segmentos de 25 ms obtidos pelo janelamento do sinal de voz quando da extração dos coeficientes MFCC, pois o conjunto de 12 MFCCs obtidos para cada segmento é concatenado ao coeficiente LOGE correspondente no processo de formação das instâncias.

2.6.3. Coeficientes Delta e Aceleração

Uma característica do *cepstro* é que ele varia de *frame* para *frame*, mesmo dentro do espaço de tempo que define um mesmo fonema (JURAFSKY, D.; MARTIN, J. H., 2009). Assim, a maneira como o *cepstro* varia também é uma característica do discurso, que deve ser acrescentada a um modelo que utilize MFCCs e *log energy* como *features*. Os coeficientes delta e aceleração são obtidos tomando respectivamente as derivadas de primeira e de segunda ordem do conjunto de 13 coeficientes obtido da concatenação dos 12 MFCCs com o coeficiente *log energy* de cada *frame*. Obtêm-se 13 coeficientes delta e 13 coeficientes de aceleração, totalizando, assim, 39 *features* por instância.

2.7. Captura de Expressões Orais

Conforme discutido anteriormente, este trabalho se concentra na classificação de emoções expressas através da fala e utiliza as *features* expostas na seção 2.6. Features Acústicas. Considerando que as 7 emoções (6 emoções básicas mais o estado emocional neutro) que constituem o conjunto de categorias de emoções a seriar neste trabalho podem ser mapeadas pelo modelo *circumplex* das emoções exposto na seção , é possível, então,

extrair características acústicas (*features*) de um sinal de áudio através das quais a categoria de expressão emocional que ele representa possa ser inferida.

Uma ferramenta de extração de *features* bastante utilizada na literatura é o openSMILE (EYBEN, F. et al., 2013), desenvolvido no Instituto de Comunicação Homem-Máquina da Universidade Técnica de Munique. O openSMILE funciona baseado em *scripts* de configuração da extração que se deseja fazer. A ferramenta recebe como argumentos o caminho do *script* de extração a utilizar e o caminho do arquivo de áudio do qual extrai as *features* descritas. O extrator deve ser chamado de um programa auxiliar, escrito na linguagem de preferência do usuário, para iterar sobre o conjunto de arquivos de áudio dos quais as *features* devem ser extraídas. Para este trabalho foi utilizada a linguagem *Shell script* (sh) do Linux. O sh, devido ao seu propósito fundamental de servir às operações corriqueiras num sistema operacional, é capaz de lidar de maneira muito simples com a iteração, a renomeação de arquivos e toda sorte de operações de apoio que o openSMILE precisa.

No *script* de extração do openSMILE são descritos os *pipes* necessários para extrair cada *feature*. Por exemplo, para a extração das *features* MFCC, LOGE, Delta e Aceleração, conjunto utilizado neste trabalho, o *pipe* de extração é comum até a operação de janelamento. Em seguida, o *pipe* sofre uma bifurcação em dois fluxos: um fica responsável por continuar a extração dos coeficientes MFCC (isto é, extrair o *melspec* do sinal) e o outro pela extração dos coeficientes LOGE. Finalmente, os fluxos independentes devem ser unidos a fim de extrair os Deltas e Acelerações do conjunto. A Figura 9 ilustra o *pipe* de extração descrito.

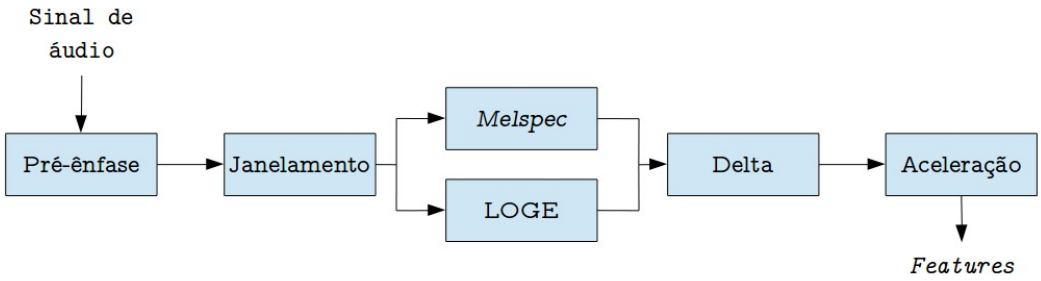

Figura 9 – Representação do pipe *de extração de* features *utilizado no openSMILE.*

O openSMILE retorna um arquivo contendo o conjunto de *features* extraídas que, dentre outros formatos, pode estar em ARFF, formato utilizado pelo WEKA, a ferramenta de ML utilizada neste projeto e citada na seção 2.4. Mecanismos de Classificação de Emoções.

2.8. Considerações Finais

Neste capítulo discorreu-se sobre a terminologia empregada na área de reconhecimento de emoções humanas. Os aspectos teóricos pertinentes às emoções humanas, sua composição e sua categorização foram discutidos. Discutiu-se, ainda, os aspectos relevantes a respeito das bases de dados de emoções, das *features* acústicas mais comuns no domínio, dos modelos de classificadores e da captura de expressões orais. No capítulo seguinte descrever-se-á em detalhes o desenvolvimento deste trabalho.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

Neste capítulo o projeto será descrito em detalhes. Será apresentada a metodologia adotada e serão apresentados e discutidos os resultados obtidos de cada etapa do processo de desenvolvimento do trabalho. Adicionalmente, será discutido como o modelo de comitê proposto precisou ser revisto. Ao final, discutir-se-ão as principais dificuldades de execução e limitações do trabalho desenvolvido e serão sintetizadas as lições aprendidas durante o decorrer deste projeto.

3.2. Projeto

O objetivo deste trabalho é obter um sistema computacional capaz de classificar emoções humanas expressas através do discurso. Particularmente, adotou-se um tipo de sistema de classificação denominado comitê de classificação, pelos motivos já explicitados nas seções anteriores. Conforme discutido, um comitê de classificação utiliza diversos modelos de classificadores individuais diferentes, de forma que as características positivas de cada um são absorvidas pelo comitê, ao que se espera que o comitê de classificação tenha performance superior à de seus componentes individuais na seriação de instâncias de discurso humano quanto ao tipo de emoção que expressam.

O primeiro passo no processo de desenvolvimento de semelhante sistema é a obtenção de uma base de dados de emoções. Assim, a primeira etapa do processo de desenvolvimento deste projeto é a busca por uma base de dados de emoções adequada. Particularmente, busca-se uma base de dados de discurso humano especializada em expressão emocional na fala.

Obtida uma base de dados adequada, o passo seguinte é a escolha e extração de *features* adequadas a partir das instâncias de fala disponíveis na base. Segue a escolha dos

classificadores que comporão o comitê de classificação. O problema de escolha de classificadores é um problema de análise fatorial dos modelos de classificação desejados. Utiliza-se um conjunto de treinamento reduzido para avaliar a performance de várias configurações de vários algoritmos de classificação. As configurações que obtêm melhor performance são selecionadas para compor o comitê.

Após a escolha das configurações de classificadores adequadas, resta o treinamento definitivo dos classificadores, com conjuntos de treinamento muito maiores do que aqueles usados na fase de análise fatorial. Obtidos os modelos, a etapa final é fazer com que trabalhem juntos dentro de um comitê de classificação, composto pelos modelos de classificadores devidamente treinados. Nesta etapa, deve ser escolhida uma equação de ponderação adequada para os classificadores que compõem o comitê de classificação. Espera-se que o comitê de classificação concebido obtenha desempenho superior a todos os seus componentes. No entanto, para o primeiro comitê de classificação elaborado neste projeto, não foi isso que aconteceu. Um de seus componentes, o classificador kNN, acabou obtendo performance superior ao comitê, algo que tornaria completamente inútil o uso de um comitê de classificação.

A razão do classificador kNN obter desempenho superior ao comitê foi investigada, o modelo de comitê foi reelaborado e com isso foi obtido um segundo modelo de comitê de classificação, este sim superior a todos os seus componentes, inclusive ao seu componente kNN (o mesmo que se mostrou superior ao primeiro comitê obtido).

3.3. Descrição das Atividades Realizadas

Nesta seção serão descritas em detalhe as etapas da metodologia descrita na seção anterior.

3.3.1. Pesquisa de Base de Dados de Emoções no Discurso

A pesquisa de bases de dados de emoções expressas no discurso humano foi feita essencialmente utilizando ferramentas de busca na Internet. Constatou-se que existe uma infinidade de bases de dados de emoções. No entanto, bases de dados de expressão emocional exclusivamente oral são mais escassas do que bases de dados de expressão facial, que dos três tipos de expressões motoras (da Teoria Componencial das Emoções), sem dúvida é aquele que é alvo do maior número de estudos e que possui maior acervo de bases de dados disponíveis para uso em pesquisa.

Ainda assim, algumas bases de dados de discurso com foco nas emoções expressas pelas instâncias que contêm foram encontradas nas mais diversas línguas. Algumas contêm instâncias rotuladas dos dois tipos de expressões motoras (ou seja, imagens de expressões faciais associadas a emoções, bem como gravações de áudio com conteúdo emocional). A maior parte das bases de dados encontradas considera ou as 6 emoções básicas do Espaço Emocional Semântico ou as 6 emoções acrescidas do estado emocional neutro. As principais bases de dados encontradas e suas características serão listadas abaixo e, ao final desta seção, a base de dados escolhida para ser usada neste projeto será discutida mais detalhadamente.

- *Toronto Emotional Speech Set* (TESS) – esta base de dados, compilada pelo Departamento de Psicologia da Universidade de Toronto (DUPUIS, K.; PICHORA-FULLER, M. K., 2010), contém 2800 exemplos de elocuções humanas, sendo que duas atrizes de idades 26 e 64 anos cuja língua materna é o inglês pronunciam, cada uma, 1400. São 200 elocuções para cada uma das 7 emoções utilizadas neste trabalho (6 emoções básicas mais o estado emocional neutro). As 200 sentenças têm o formato “Say the word...”, e segue-se uma palavra que termina a sentença. A grande desvantagem desta base de dados é o acesso a ela, que deve ser feito através de um ambiente chamado “Tspace”, um repositório de conteúdo com direitos autorais da Universidade de Toronto que, apesar de ser aberto, não facilita o acesso aos itens ali compartilhados. Além disso, a TESS não é muito clara em sua documentação quanto ao tipo de licença que utiliza.

- *Berlin Database of Emotional Speech* (Emo-DB) – base de dados de emoções utilizada em muitos trabalho na área de classificação de emoções no discurso (BURKHARDT, F. et al., 2005). Trata-se de uma base de dados compilada pelo Departamento de Ciência da Comunicação, Instituto de Discurso e Comunicação da Universidade Técnica de Berlim. Apesar da extensa documentação e indiscutível qualidade, rapidamente a possibilidade de usá-la foi descartada por se tratar de uma base de dados, por ora, restrita.
- *An Italian Emotional Speech Corpus* (EMOVO) – conforme já citado, a EMOVO é uma base de dados de discurso emotivo em língua italiana compilada devido ao esforço de representantes de vários departamentos da Universidade de Roma. De fato, a base EMOVO é a primeira base de dados em língua italiana de emoções expressas no discurso. A EMOVO, assim como a TESS, emprega atores para pronunciar frases, atribuindo-lhes conteúdo emocional. Ao todo, gravações de 6 atores constituem a base, sendo 3 mulheres e 3 homens. Cada ator pronuncia 98 frases de conteúdo semântico incoerente, que não predispõe o reconhecimento de nenhuma emoção em particular, totalizando 588 gravações. Os mesmos 6 estados emocionais discutidos quando se tratou, aqui, do Espaço Emocional Semântico, acrescidos do estado emocional neutro, são as emoções das quais trata a base. Assim, há 14 elocuções de cada emoção para cada um dos 6 atores. A base possui documentação simples, ainda que muito satisfatória, e o acesso a ela é aberto e igualmente simples.

Das três bases de dados cogitadas, optou-se pela base EMOVO. O motivo da escolha foi principalmente a simplicidade do acesso à base, que é livre e se dá através do *download* de um arquivo *zipado* que já contém todo o material necessário ao uso da base, incluindo sua documentação, diferente das demais bases de dados analisadas cujo acesso, quando livre, sempre se mostrou atribulado por toda sorte de fatores. A base separa as instâncias por ator, de forma que em cada um de seus diretórios são colocadas as gravações correspondentes a um ator. Os nomes dos arquivos WAV da EMOVO identificam a emoção e a elocução a que correspondem. Apesar da simplicidade, talvez por se tratar da primeira base de dados de discurso emotivo compilada em língua italiana, a EMOVO é uma base de

dados muito organizada e bem documentada, ainda que todo o conteúdo, com exceção dos *papers* publicados a seu respeito, esteja em italiano.

O conteúdo da base EMOVO foi validado por ouvintes humanos em um teste proposto por seus criadores a 24 sujeitos, separados 12 a 12 em ambientes distintos e isolados, pelos quais a classificação do conteúdo emocional das elocuções gravadas deveria ser feita após ouvi-las. A Tabela 1, retirada do *paper* seminal da base EMOVO (CONSTANTINI, G. et al., 2014), mostra a matriz de confusão (em porcentagens arredondadas correspondentes às emoções reconhecidas) obtida do teste com um dos grupos de 12 indivíduos, que é estatisticamente equivalente à tabela do outro grupo.

		Emoção reconhecida						
		Neutro	Desgosto	Alegria	Medo	Raiva	Surpresa	Tristeza
Emoção atuada	Neutro	93%	1%	0%	0%	4%	0%	2%
	Desgosto	3%	67%	2%	6%	10%	6%	6%
	Alegria	2%	4%	66%	7%	7%	10%	4%
	Medo	2%	7%	2%	74%	3%	3%	9%
	Raiva	1%	1%	1%	3%	92%	1%	1%
	Surpresa	1%	3%	4%	1%	1%	81%	9%
	Tristeza	2%	2%	1%	3%	0%	0%	92%

Tabela 1 – Matriz de confusão do experimento de validação da base EMOVO.

O resultado da validação permite concluir que as gravações atuadas, de fato, podem ter o conteúdo emocional reconhecido por humanos, o que torna a base própria para uso em pesquisa científica.

3.3.2. Escolha e Extração de *Features* Acústicas

Selecionada a base de dados, o próximo passo do trabalho era a seleção de *features* adequadas. A escolha de *features* se deu através da pesquisa em literatura da área. Conforme já mencionado, as leituras de trabalhos no domínio do reconhecimento de emoções permitiram concluir que dois tipos de *features* são verdadeiramente corriqueiras na área: MFCCs e coeficientes *log energy*, mais os respectivos coeficientes delta e aceleração. Um conjunto de 39 *features*, seguindo uma sugestão da literatura, conforme já discutido, composto por 12 MFCCs, 1 coeficiente *log energy* e os respectivos coeficientes delta e aceleração foi compilado utilizando o openSMILE com o *pipe* de extração já apresentado. Detalhes adicionais do *pipe* de extração utilizado para extrair as *features* no openSMILE serão fornecidos a seguir.

Na fase de pré-ênfase do processo de extração foi escolhido um parâmetro de memória "a" (para o filtro *high-pass*) com valor 0,97. Vale lembrar que o parâmetro "a" é o coeficiente que multiplica o fator com atraso unitário (z^{-1} no domínio Z) no numerador da função de transferência do filtro. Os valores de duração e sobreposição das janelas aplicadas aos sinais de entrada foram respectivamente 25 ms e 10 ms, valores de uso comum na literatura, e a função de janelamento escolhida foi a janela de Hamming com ganho unitário. A aplicação de uma janela de Hamming com ganho unitário implica que tudo que a operação de janelamento faz é obter *frames* com bordas atenuadas em relação ao centro, mantendo os trechos do sinal próximos a ele inalterados.

Na fase seguinte, de extração de *melspec*, foram especificadas 26 bandas de filtragem para o banco de filtros em escala mel e a filtragem foi feita em relação ao espectro de potências do sinal, conforme especifica a literatura da área. A frequência de corte inferior do banco de filtros foi definida em 0 Hz e a frequência de corte superior em 8 kHz, mais do que suficiente para extrair a informação pertinente à distribuição da densidade espectral de energia do discurso humano (sob condições normais) para uma aplicação de seriação de emoções expressas no sinal de fala. Basta lembrar que a filtragem utilizada em linhas telefônicas convencionais corta as frequências superiores a 3,4 kHz. Ainda assim, é completamente possível identificar as emoções de um interlocutor numa conversa telefônica. Na verdade, sabe-se que a banda de frequências de um sinal de discurso humano, sob condições normais, está entre (aproximadamente) 500 Hz e 2 kHz.

Finalmente, a série *melspec* foi truncada no coeficiente de número 12 para obter apenas os 12 coeficientes MFCC desejados. O coeficiente *log energy* foi obtido especificando a opção "log" para a função de extração de energia do openSMILE (*cEnergy*) tomando como entrada a saída da chamada de janelamento. Os coeficientes MFCC e *log energy* foram concatenados (função *cVectorConcat*) previamente à chamada da função de derivada da ferramenta (*cDeltaRegression*). Da função *cDeltaRegression* são obtidos os coeficientes delta da concatenação MFCC + *log energy*. A saída da chamada a *cDeltaRegression* é então passada a outra instância da mesma função. Disto obtêm-se os coeficientes de aceleração das *features* espectrais do sinal.

O resultado da extração do openSMILE para a base de dados foi gravado em um arquivo ARFF, formato utilizado pelo WEKA e ao qual o openSMILE suporta como formato de saída da extração. O arquivo obtido pelo openSMILE continha um total de 189.567 instâncias devidamente rotuladas segundo a emoção representada. Optou-se por balancear a base de dados, isto é, fazer com que o número de instâncias com cada um dos 7 rótulos (emoções básicas + estado neutro) fosse igual. A Tabela 2 mostra a distribuição original do número de instâncias por emoção da extração realizada pelo openSMILE.

Emoção	Número de instâncias
Desgosto	36.249
Alegria	24.291
Neutro	23.898
Medo	25.913
Raiva	24.473
Surpresa	26.142
Tristeza	28.601
TOTAL	**189.567**

Tabela 2 – Quantidade de instâncias por emoção obtidas da base EMOVO.

Como o número mínimo de instâncias para uma emoção era 23.898, para o estado emocional neutro, optou-se por escolher 20.000 instâncias de cada emoção para compor a base a ser usada no treinamento dos modelos. O número 20.000 foi escolhido simplesmente para facilitar o posterior particionamento da base de dados na fase de treinamento, de forma que a base de dados obtida tivesse 140.000 instâncias rotuladas no total. Observe-se que os números grandes de instâncias obtidos aqui estão em concordância com o intervalo de janelamento selecionado, pequeno, de 25 ms e ainda com sobreposição de 10 ms (muitos *frames* são obtidos), ainda que apenas 588 gravações constituam o conjunto de elocuções da base.

3.3.3. Elaboração de Modelo e Experimentos de Classificação

No passo seguinte à extração de *features* das instâncias de gravações da base de dados, deve ser realizada a seleção dos modelos de classificadores a utilizar no comitê de classificação. Uma maneira de fazer isso é compilar um subconjunto (muito menor) do

conjunto de *features* obtidas pelo openSMILE e realizar uma análise de desempenho fatorial dos algoritmos de aprendizado a utilizar no comitê. Em seguida, as configurações dos algoritmos com melhor desempenho são selecionadas e treinadas com a base de dados completa para compor os modelos finais componentes do comitê de classificação. Para este trabalho três subconjuntos, um com 700, outro com 3.500 e um com 7.000 instâncias, foram compilados, contendo, respectivamente, 100, 500 e um milhar de cada emoção.

As instâncias de cada emoção para a análise fatorial foram escolhidas de maneira sequencial a partir do topo do conjunto completo de instâncias para cada emoção e, em seguida, embaralhadas e colocadas em um novo arquivo ARFF do WEKA. Os classificadores experimentados foram submetidos a diversas combinações de suas variáveis de configuração, treinados e avaliados para cada uma das configurações diferentes. A estratégia de treinamento adotada foi o *split* simples com porcentagem de treinamento de 90% da base, em seleção aleatória. Por se tratar de uma fase de análise prévia, apenas uma repetição de cada configuração foi executada.

Inicialmente, 41 configurações diferentes de algoritmos foram testadas, sendo 20 configurações de SVM (implementação SMO no WEKA), 5 configurações de Árvore de Decisão (implementação J48 no WEKA), 12 configurações de kNN (implementação IBk no WEKA), 3 configurações de classificador Bayesiano (implementação NaiveBayes no WEKA). Finalmente, um classificador de referência (implementação ZeroR no WEKA) – que na prática escolhe o rótulo da primeira classe que aparece no conjunto de treinamento e "chuta" todos os exemplos de teste com aquele rótulo – foi testado também, como uma maneira de saber como estava o comportamento dos demais modelos em relação ao pior modelo de classificação possível. Cada configuração de algoritmo recebeu um índice, através do qual ela é referenciada. A Tabela 3 apresenta os resultados da análise de desempenho fatorial (em termos de acurácia) das configurações testadas. Em seguida, os detalhes de configuração dos modelos com os melhores resultados, apenas, mostrados em negrito na Tabela 3, estão listados.

	Acurácia (%)		
Configuração	100 instâncias/emoção	500 instâncias/emoção	1.000 instâncias/emoção
ZeroR1	13,04	14,29	14,29
NaiveBayes1	34,78	31,14	28,14
NaiveBayes2	**49,28**	**32,57**	**30,14**
NaiveBayes3	36,23	31,14	30,43
IBk1	**88,41**	**78,86**	**74,43**
IBk2	88,41	78,86	74,43
IBk3	88,41	78,86	74,43
IBk4	66,67	71,43	66,43
IBk5	71,01	76,29	72,71
IBk6	71,01	76,29	72,71
IBk7	55,07	62,29	58,86
IBk8	66,67	68,57	64,86
IBk9	65,22	68,86	64,29
IBk10	43,48	52,29	52,86
IBk11	49,28	60,57	61,57
IBk12	46,38	60	60,43
J481	36,23	35,43	35,29
J482	47,83	39,43	38,43
J483	**49,28**	**39,43**	**38,43**
J484	49,28	39,43	38,29
J485	49,28	39,43	38,29
SMO1	13,04	32,29	33,29
SMO2	13,04	32,29	33,29
SMO3	13,04	32,29	33,29
SMO4	13,04	32,29	33,29
SMO5	13,04	20,29	31,29
SMO6	43,48	47,14	38,71
SMO7	57,97	53,14	48,57
SMO8	71,01	60,29	54,86
SMO9	**68,12**	**61,14**	**61**
SMO10	68,12	60,29	59,71
SMO11	72,46	57,71	52,43
SMO12	65,22	58,29	60,14
SMO13	60,87	54,57	58,57
SMO14	60,87	52,86	57,14
SMO15	60,87	52,57	55,29
SMO16	60,87	52,57	54,71
SMO17	60,87	52,57	54,71
SMO18	60,87	52,57	54,71
SMO19	60,87	52,57	54,71
SMO20	60,87	52,57	54,71

Tabela 3 – Resultados da análise de desempenho fatorial de classificadores.

A análise da Tabela 3 mostra que as melhores configurações de cada algoritmo de treinamento ensaiado são a “NaiveBayes2”, a “IBk1”, a “J483” e a “SMO9”, por obterem acurácias consistentemente iguais ou superiores às demais configurações do mesmo algoritmo em ao menos dois dos três testes realizados. Para os casos onde duas ou mais configurações obtiveram resultados iguais, tomou-se a configuração com menor índice como sendo a melhor. Pode, a princípio, parecer contraditório que o algoritmo IBk tenha

obtido desempenho, no geral, superior ao algoritmo SMO, uma vez que, sendo uma implementação de SVM, o SMO deveria ter resultados consistentemente superiores aos demais algoritmos. Uma análise simples da capacidade de generalização de aprendizado dos diferentes algoritmos pode ser feita através da variação da porcentagem de *split* de treinamento (e consequentemente de teste) e avaliação da acurácia da melhor configuração de cada algoritmo obtida da análise fatorial com cada *split*. Os valores obtidos por cada um dos algoritmos podem ser comparados graficamente para averiguar qual deles possui a melhor capacidade de generalização de aprendizado (em outras palavras, qual obtém a melhor performance na seriação de instâncias de teste quando *poucos* dados de treinamento lhe são fornecidos). A Figura 10 apresenta o gráfico de uma análise conduzida dessa forma. Tomou-se o conjunto de 700 instâncias e o *train split* foi variado entre 10% e 90% com incrementos de 10% e os resultados são condizentes com o que se espera – que o SVM obtenha resultados substancialmente superiores aos demais algoritmos para *train splits* pequenos.

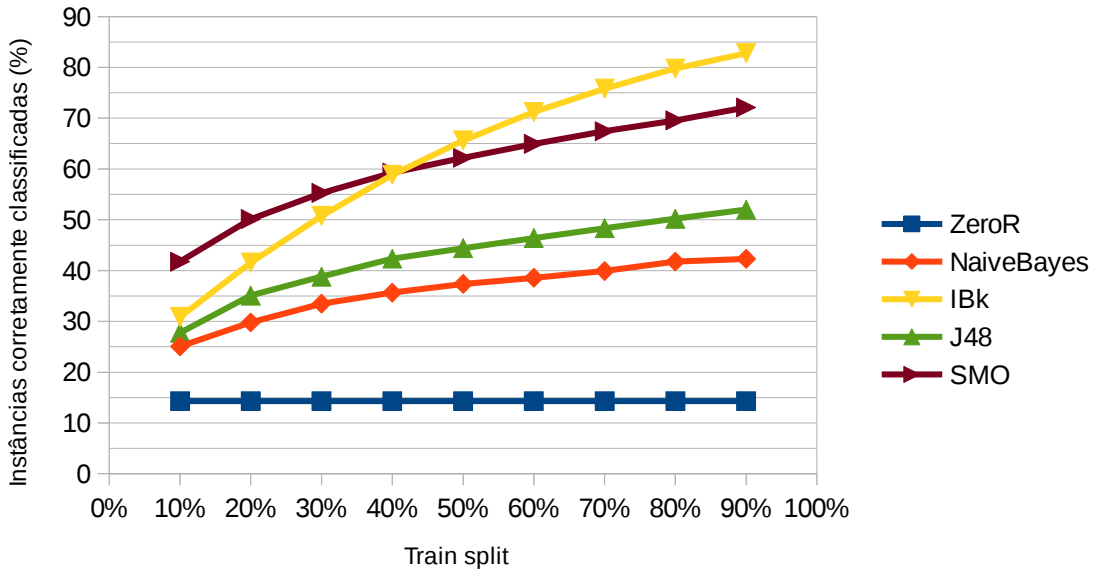

Figura 10 – Gráfico de comparação da capacidade de generalização dos modelos componentes do comitê.

A análise da Figura 10 mostra que o SMO, mesmo com um *split* de treinamento de 10%, já é capaz de generalizar o aprendizado a ponto de atingir uma acurácia próxima a

43%, enquanto os demais algoritmos (com exceção do ZeroR, obviamente, que serve apenas para referência) têm acurácias entre 25% e 33%. À medida que o *train split* é incrementado, o quadro se mantém até o *train split* de 30%, quando a acurácia atingida pelo IBk é de 50%, a do NaiveBayes e do J48 ainda estão abaixo dos 40% e o SMO figura com uma acurácia de 55%. Para a porcentagem de *split* de 50%, o IBk supera a acurácia atingida pelo SMO. A ordem de capacidade de generalização dos algoritmos, então, se mantém a mesma até o *train split* de 90%, quando o ensaio é encerrado.

Este ensaio simples mostra o quanto a comparação de performance de classificação de algoritmos pode ser enganosa se olharmos apenas para o valor de acurácia obtido. O autor deste documento teve a oportunidade de conversar com um especialista em ML durante o curso deste trabalho (Dr. Rodrigo Fernandes de Mello – ICMC/USP) e ele destacou que até mesmo entre os pesquisadores da área de aprendizado de máquina este engano é bastante comum e acrescentou o quanto é difícil convencer alguém de que uma acurácia de teste mais alta não é, por si só, uma métrica segura para a performance de uma estratégia de classificação.

A performance de um algoritmo, então, não deve ser medida em termos tão somente de sua acurácia. Isso é especialmente verdade para o kNN, cuja estratégia de aprendizado é de "espalhar" seus exemplos de treinamento sobre um espaço n-dimensional (sendo "n" a dimensão das instâncias de treinamento) e, posteriormente, na fase de teste, as instâncias desconhecidas têm suas distâncias (por exemplo: distâncias cartesianas) a cada um dos elementos espalhados no espaço n-dimensional calculadas. O rótulo da instância (ou o rótulo predominante no conjunto de instâncias) de treinamento mais próxima à instância desconhecida é atribuído a ela. Dessa forma, é intuitivo ver que o kNN pode até ter uma performance superior ao SMO, mas isso depende dele já conhecer a maior parte do conjunto de dados que lhe serão apresentados. No mundo real isso não acontece. Potencialmente, infinitas instâncias de teste podem ser apresentadas a um classificador, o que significa que uma métrica muito mais segura para a avaliação de sua performance é a sua capacidade de generalização. A capacidade de generalização de um algoritmo pode ser medida em termos do erro que ele comete ao classificar instâncias desconhecidas, ou seu "erro empírico". Outra métrica importante para modelos de aprendizado de máquina é a sua curva *receiver operating characteristic* (curva ROC), que mede o quão boa é a

performance de classificação binária (ou ainda *one-vs-all*) de um modelo em termos de suas taxas de verdadeiro positivo e de falso positivo para diferentes valores de fronteira de separação entre as classes. A área debaixo da curva ROC de um modelo pode ser usada para saber se um modelo provê uma classificação de qualidade e não apenas tem uma acurácia alta. A área debaixo da curva ROC é conhecida como AUC (de *area under curve*).

Voltando aos modelos selecionados para compor o comitê, os parâmetros dos modelos de classificação que resultaram nas melhores performances na fase de análise de desempenho fatorial serão listados a seguir, em termos dos parâmetros passados para o WEKA.

- NaiveBayes
  - *debug: False*
  - *displayModelInOldFormat: False*
  - *useKernelEstimator: True*
  - *useSupervisedDiscretization: False*
- IBk
  - *KNN: 1*
  - *crossValidate: False*
  - *debug: False*
  - *distanceWeighting: No distance weighting*
  - *meanSquared: False*
  - *nearestNeighbourSearchAlgorithm: LinearNNSearch*
  - *windowSize: 0*
- J48
  - *binarySplits: False*
  - *confidenceFactor: 0.2*

- *debug: False*
- *minNumObj: 2*
- *numFolds: 3*
- *reducedErrorPruning: False*
- *saveInstanceData: False*
- *seed: 1*
- *subtreeRaising: True*
- *unpruned: False*
- *useLaplace: False*

- SMO
  - *buildLogisticModels: False*
  - *c: 100.0*
  - *checksTurnedOff: False*
  - *debug: False*
  - *epsilon: 1.0E-12*
  - *filterType: No normalization/standardization*
  - *kernel: NormalizedPolyKernel*
    - *cacheSize: 250007*
    - *checksTurnedOff: False*
    - *debug: False*
    - *exponent: 2.0*
    - *useLowerOrder: False*
  - *numFolds: -1*

- *randomSeed: 1*
- *toleranceParameter: 0.001*

As configurações de classificador listadas acima, então, foram treinadas (novamente com o WEKA) e os modelos obtidos foram salvos. Utilizou-se a estratégia *split* simples de instâncias novamente. Dessa vez, no entanto, 10 repetições do experimento (com 10 conjuntos diferentes de instâncias) foram feitas, a fim de obter valores de variância dos resultados obtidos e verificar sua robustez. Assim, 10 modelos de cada configuração de classificador foram obtidos. Cada um dos 10 conjuntos de instâncias continha 1.000 exemplos de cada emoção, selecionados da base de dados de maneira sequencial e embaralhados a seguir, totalizando 7.000 instâncias por conjunto ou 70.000 exemplos no total (correspondendo a 50% da base de dados disponível). O motivo da não utilização de 100% da base de dados, que continha 140.000 exemplos, será discutido na seção 3.5. Dificuldades e Limitações. O valor de *split* utilizado para cada um dos 10 conjuntos de treinamento foi de 50% *train* (50% *test*) em seleção aleatória.

Treinados os classificadores e salvos os modelos, estes foram importados para uma aplicação Java através da API do WEKA. Os classificadores, então, tiveram suas acurácias, erros (*root mean squared*, RMSE) e AUCs calculados para o *split* de teste. A Tabela 4 mostra, das 10 iterações de *split* simples, os valores médios e as respectivas variâncias (σ^2) obtidos para acurácia, RMSE e AUC após os testes dos modelos.

	$Acurácia_m$	σ^2(Acurácia)	$RMSE_m$	σ^2(RMSE)	AUC_m	σ^2(AUC)
NaiveBayes	39,81%	2,80E-03	2,04E-02	5,48E-07	0,77	1,13E-03
SMO	62,60%	2,21E-03	1,65E-02	1,08E-06	0,88	5,19E-04
J48	44,92%	2,29E-03	1,95E-02	4,17E-07	0,71	1,23E-03
IBk	71,82%	1,83E-03	1,42E-02	1,31E-06	0,86	1,31E-03

Tabela 4 – Estatísticas de desempenho dos modelos componentes do comitê.

A análise da Tabela 4 mostra que, assim como nos testes preliminares, o melhor resultado de acurácia foi obtido pelo IBk. O SMO, conforme o esperado, é o classificador com melhor AUC na média e portanto tem a melhor qualidade de classificação, ainda que

apresente acurácia inferior ao IBk. Em seguida, era necessário fazer com que os classificadores trabalhassem juntos através de sua união em um comitê de classificação.

Conforme já discutido, um comitê funciona fazendo com que cada classificador emita sua opinião a respeito de uma instância de teste desconhecida, ou seja, classifique a instância e, em seguida, cruzando as opiniões de todos os classificadores de acordo com uma equação que atribui pesos à opinião de cada um, de forma que a resposta do comitê seja escolhida de acordo com a confiabilidade que a equação atribui a cada componente. Em termos mais simples, cada classificador "vota" numa classe à qual acredita que o exemplo desconhecido pertence e, em seguida, o módulo de ponderação do comitê emite uma resposta baseado no peso que o voto de cada classificador tem sobre a resposta do grupo.

O comitê foi programado na própria aplicação Java usada para a execução dos classificadores com os *splits* de teste. O peso p_i atribuído a cada classificador i foi obtido pela Equação 10, onde a_i e AUC_i são os valores calculados de acurácia e AUC do classificador i e N é o número total de classificadores do comitê, no caso 4.

$$p_i = \frac{a_i \cdot AUC_i}{\sum_{j=1}^{N} a_j \cdot AUC_j} \tag{10}$$

É fácil notar que a somatória dos pesos distribuídos entre os classificadores é igual à unidade. Assim, se pode pensar na atribuição de pesos como uma operação de distribuição de responsabilidade entre os classificadores do comitê, como se o módulo de ponderação do comitê de classificação confiasse $p_i \cdot 100\%$ da resposta que emite ao classificador i. É importante notar que ao utilizar na equação de atribuição de pesos os valores de acurácia e de AUC dos classificadores, estamos perdendo acurácia no resultado do comitê quando avaliado sobre os dados de teste disponíveis, mas estamos emitindo uma resposta melhor, simplesmente porque a AUC é uma métrica mais confiável no que diz respeito à qualidade da resposta de um modelo de classificação, ainda que, obviamente, a acurácia de um classificador sobre as instâncias de teste não deixe de ser uma métrica importante.

Utilizando a Equação 10, 10 conjuntos de pesos (isto é, 10 vetores de 4 elementos) foram calculados a partir dos dados de acurácia e AUC da iteração correspondente e cada um dos 10 conjuntos de pesos foi utilizado para avaliar a performance de classificação do comitê naquela iteração. Após a execução, observou-se que, em média, o comitê obteve uma acurácia de 70,88% com variância de 1,61E-03, com um erro de teste (*root mean squared*) associado de 1,71E-02 e variância de 1,65E-06. Assim, a performance do comitê pode ser dita insatisfatória uma vez que, por mais que o erro associado à classificação das instâncias desconhecidas seja menor do que o de todos os classificadores que o compõem, a acurácia do comitê é inferior à de um deles, o IBk, que obteve acurácia média de 71,82% para as mesmas instâncias de teste. Espera-se de um comitê de classificação uma performance definitivamente melhor do que todos os seus componentes, pois apenas isso justifica a complexidade adicionada ao projeto de um sistema de classificação que utiliza um comitê em vez de um classificador isolado, simplesmente.

O resultado obtido na avaliação da performance do comitê projetado, a princípio, pareceu muito estranha ao autor deste documento. *Como seria possível que um comitê de classificação tivesse obtido performance inferior a um de seus componentes isolados? A equação de atribuição de pesos não deveria, por si só, cuidar da atenuação dos pesos de classificadores que porventura viessem a "derrubar" a performance do comitê?* No entanto, após ter a oportunidade de conversar com um pesquisador da área (Me. Leandro Yukio Mano Alves – ICMC/USP), a razão do resultado obtido ficou mais clara. Segundo ele, para que um comitê de classificação funcione de maneira satisfatória (ou seja, supere todos os seus componentes), é necessário que estes tenham performances parecidas: uma regra prática para o projeto de comitês de classificação é utilizar classificadores com acurácias de teste não muito mais distantes do que 10% uma da outra. Caso contrário, os classificadores de menor acurácia tendem a prejudicar a performance coletiva. Observando novamente os resultados apresentados na Tabela 4, claramente o comitê projetado não atende a esta regra. Os modelos IBk (o de melhor acurácia) e SMO (o de segunda melhor acurácia) têm entre eles uma diferença um pouco inferior a 10%, mas do SMO para o terceiro melhor modelo (J48), a diferença de acurácia é de quase 20%. Assim, procurou-se buscar um novo modelo de classificador, com acurácia mais próxima àquelas do IBk e do SMO para tomar o lugar dos classificadores NaiveBayes e J48 no comitê de classificação.

A busca por outros classificadores levou à escolha de uma rede *Multilayer Perceptron* (MLP), com ocorrência menos comum na literatura. A mesma análise fatorial preliminar (da Tabela 3) foi feita para diferentes configurações de rede MLP. Os resultados obtidos são mostrados na Tabela 5. A melhor configuração encontrada é mostrada em negrito.

	Acurácia (%)		
Configuração	100 instâncias/emoção	500 instâncias/emoção	1000 instâncias/emoção
MLP1	**73,91**	**58,57**	**57,43**
MLP2	66,67	61,71	52,14
MLP3	68,12	61,43	51,57
MLP4	62,32	59,71	54
MLP5	71,01	64,29	51,71
MLP6	65,22	56,85	51,29
MLP7	72,46	64	52,14
MLP8	63,79	62,29	55,43
MLP9	72,46	60,57	55,14

Tabela 5 – Resultados da análise de desempenho fatorial da rede MLP.

Os parâmetros da melhor configuração de MLP serão listados a seguir, em termos dos parâmetros passados para o WEKA.

- MultilayerPerceptron
  - *GUI: False*
  - *autoBuild: True*
  - *debug: False*
  - *decay: False*
  - *hiddenLayers: a*
  - *learningRate: 0.3*
  - *momentum: 0.2*
  - *nominalToBinaryFilter: True*
  - *normalizeAttributes: True*

- *normalizeNumericClass: True*
- *reset: True*
- *seed: 0*
- *trainingTime: 500*
- *validationSetSize: 0*
- *validationThreshold: 20*

A capacidade de generalização da rede MLP foi avaliada da mesma maneira que os algoritmos da primeira versão de comitê obtida, variando a porcentagem de *train split* do mesmo conjunto de 700 instâncias anteriormente usado para esse propósito entre 10% e 90% com saltos de 10% e comparando o gráfico de porcentagem de acertos da rede MLP para cada *train split* com o dos outros algoritmos do comitê (agora apenas o IBk e o SMO). A Figura 11 apresenta a comparação das capacidades de generalização dos algoritmos nesses termos.

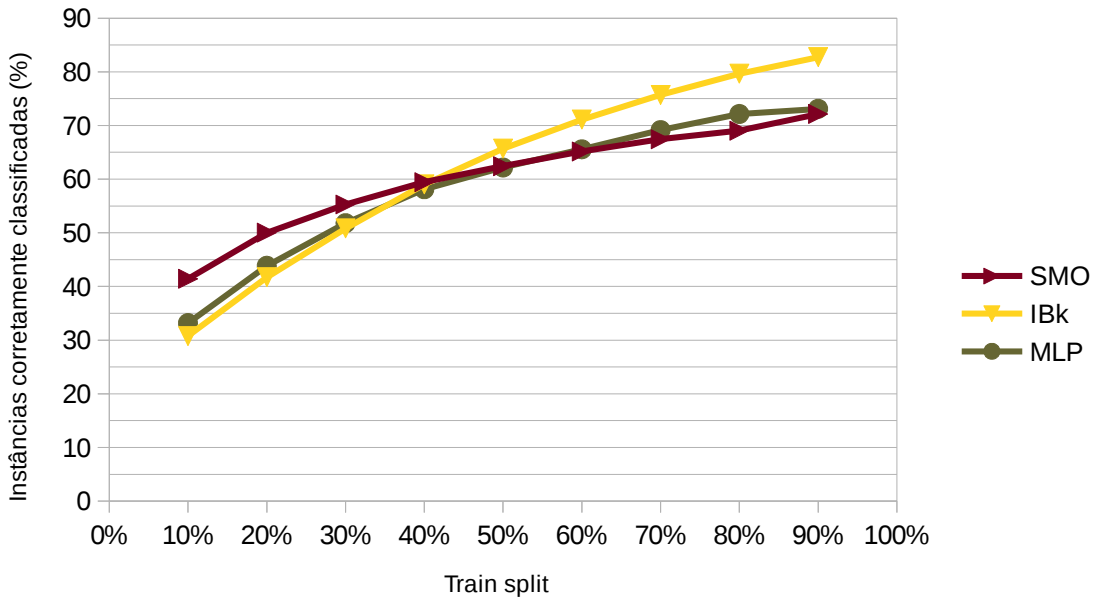

Figura 11 – Comparação das capacidades de generalização dos modelos componentes do novo comitê.

Observa-se que a rede MLP tem capacidade de generalização muito próxima ao IBk entre as porcentagens 10% e 40% de *train split*, sendo superado pelo SMO. A partir do *train split* de 40%, ponto em que o IBk se iguala ao SMO em termos da porcentagem de instâncias corretamente classificadas, a rede MLP passa a acompanhar de perto a performance do SMO. De qualquer maneira, a configuração obtida de rede MLP, sem dúvida alguma é um modelo de classificação que se assemelha muito mais aos modelos IBk e SMO em todos os aspectos analisados do que os modelos NaiveBayes e J48 que lhe cederam lugar no comitê.

Finalmente, 10 modelos de MLP com as configurações acima foram treinados utilizando abordagem idêntica à usada para obter os modelos de classificador utilizados no primeiro comitê projetado: 10 repetições de *split* simples (50% *train*, 50% *test*), cada uma realizada sobre um dos mesmos 10 conjuntos de instâncias utilizadas anteriormente, com 7.000 instâncias cada, sendo 1.000 correspondentes a cada emoção, totalizando 70.000 instâncias.

	$Acurácia_m$	σ^2(Acurácia)	$RMSE_m$	σ^2(RMSE)	AUC_m	σ^2(AUC)
MLP	58,33%	1,72E-03	1,75E-02	1,24E-06	0,87	4,00E-04

Tabela 6 – Estatísticas de desempenho da rede MLP a compor o novo comitê de classificação.

A Tabela 6 apresenta os valores obtidos para acurácia média, erro de teste médio (*root mean squared*) e AUC média e as respectivas variâncias para a rede MLP após realizado o seu treinamento com cada um dos 10 conjuntos de instâncias.

Observa-se que a MLP, de fato, atingiu uma performance muito mais semelhante aos dois melhores modelos de classificador do primeiro comitê projetado, tanto em termos de acurácia quanto de RMSE de teste e de AUC. Assim, era esperado que a substituição dos classificadores NaiveBayes e J48 pelo classificador MLP, obtendo o comitê de classificação SMO + IBk + MLP fizesse com que a performance do comitê de classificação superasse a de seus componentes individuais.

A substituição foi implementada no programa Java, o novo cálculo de pesos foi realizado e o comitê foi testado para cada um dos 10 conjuntos de *split* novamente. Obteve-se com isso uma acurácia média para o comitê de 72,86% com variância de 1,29E-

03 e RMSE de teste associado de 1,42E-02 e variância 1,20E-06. Assim, o segundo comitê projetado obteve performance satisfatória, superando a performance de todos os seus componentes individuais.

3.4. Resultados Obtidos

Com a segunda configuração de comitê elaborada, foi alcançado o objetivo inicial de obter um comitê de classificação cuja performance fosse satisfatória. Os resultados de acurácia e RMSE para a configuração final de comitê de classificação (isto é, a segunda) serão apresentados a seguir, de maneira gráfica, para cada uma das 10 repetições de teste com *split* simples realizadas. A Figura 12 apresenta a comparação do desempenho do comitê de classificação com seus componentes em termos de suas acurácias.

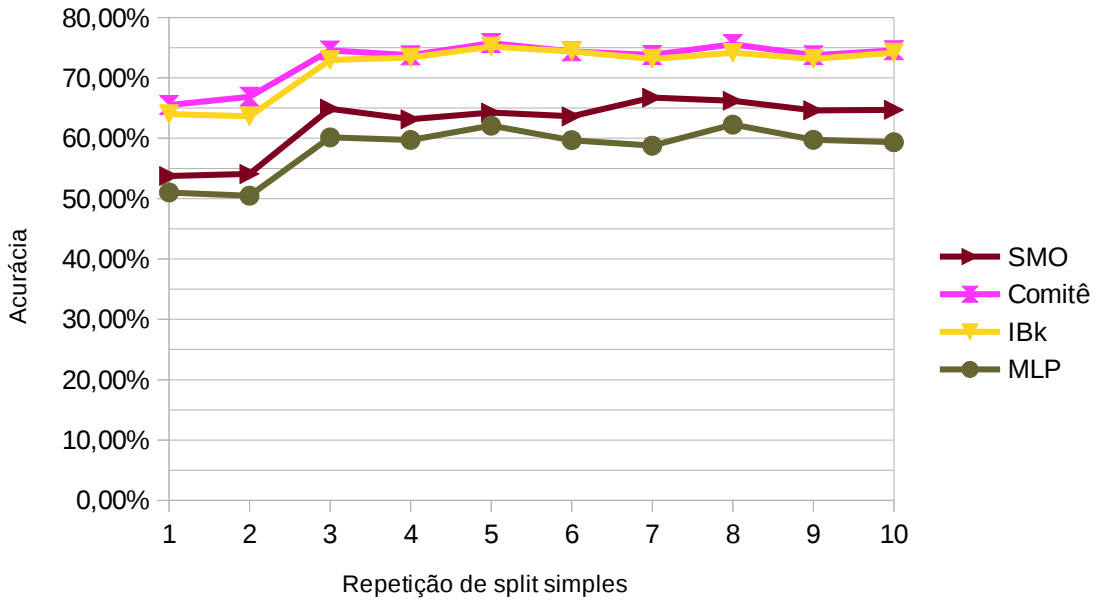

Figura 12 – Comparação entre o comitê de classificação e seus componentes em termos de acurácia.

A Figura 13 apresenta a comparação do comitê de classificação com seus componentes em termos de seus RMSEs de teste.

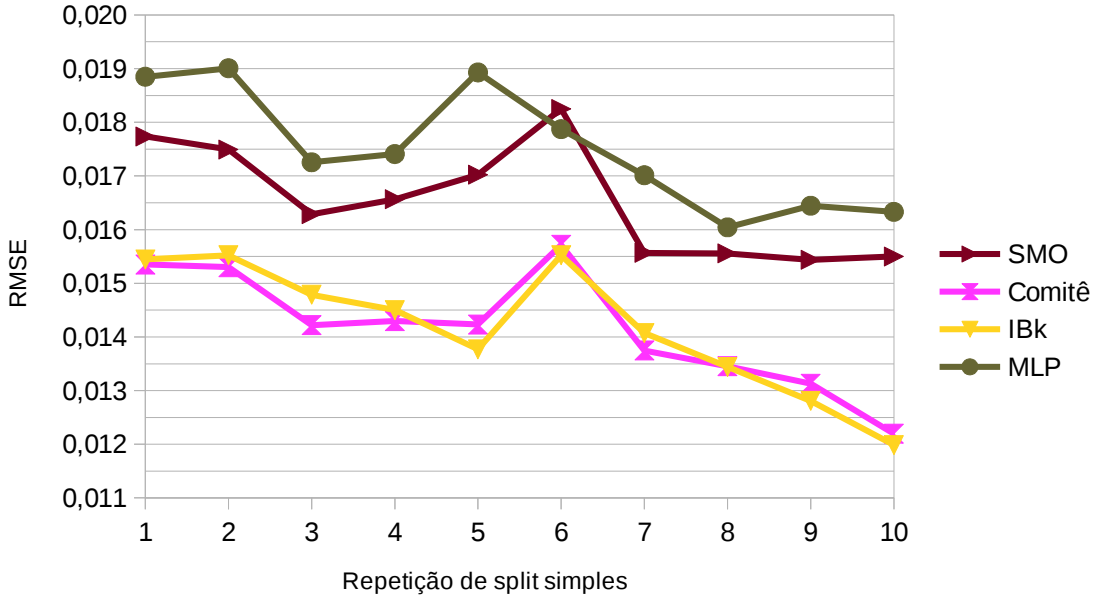

Figura 13 – Comparação entre o comitê de classificação e seus componentes em termos de RMSE.

A análise dos resultados apresentados na Figura 12 e na Figura 13 somente reforça o que já era sabido olhando para os valores médios de acurácia e RMSE do comitê e de seus componentes: a performance do comitê de classificação é superior à de todos os seus componentes individualmente, mas apenas ligeiramente, pois sua performance avaliada para os casos de teste está sempre muito próxima à do IBk.

O comitê de classificação proposto neste trabalho conseguiu melhorar a acurácia média em relação ao melhor dos classificadores isolados (IBk) em 1,04% (de 71,82% para 72,86%), mantendo o valor médio de RMSE em 1,42E-02. O ganho de pouco mais de 1%, dada a maior complexidade de se projetar um sistema de classificação empregando o conceito de comitê, que demanda muito mais do projetista do que o uso de um classificador isolado pode parecer pouco atrativo. No entanto, é preciso levar em consideração que os componentes individuais do comitê proposto neste projeto passaram por muito pouco refinamento antes de serem postos para trabalhar em grupo, bem como as *features* selecionadas, que são de igual importância para o bom desempenho de qualquer sistema de classificação.

Como pôde ser observado, um comitê de classificação minimamente bem projetado, construído respeitando um limite razoável para as diferenças na performance de

seus componentes, consegue obter resultados melhores do que aqueles de seus componentes individuais. Por outro lado, um comitê de classificação construído unindo classificadores de performance muito boa a classificadores de performance muito ruim, tende a obter performance de classificação pior do que a de ao menos um de seus componentes. É interessante notar como o comportamento coletivo dos classificadores, que piora diante da existência de membros "ruins", é semelhante ao comportamento de grupos humanos, cuja performance na execução de tarefas também tende a piorar se há no grupo indivíduos menos aptos.

3.5. Dificuldades e Limitações

A maior dificuldade encontrada durante a condução deste trabalho foi o enorme gasto de tempo com o treinamento dos modelos de classificadores, em especial os modelos de SVM (SMO). Ainda que tenha sido feita a tentativa de utilizar *clusters* para executar o treinamento, o tempo de espera pela finalização do treinamento de alguns modelos se mostrou inviável, ao que foi feita a opção por utilizar menos instâncias da base de dados, de forma que apenas 50% da base disponível foi utilizada no trabalho.

O autor não chegou a explorar a utilização de *features* prosódicas, outro importante parâmetro linguístico que pode ser extraído de sinais de discurso e que poderia alterar os resultados obtidos pelos classificadores, uma vez que a prosódia do discurso está relacionada ao ritmo da fala, algo que varia muito entre estados emocionais. Foi feita a opção conservadora de utilizar apenas as *features* mais comuns no domínio.

Certamente, a principal lição que o autor deste trabalho levará de sua execução é que a pesquisa científica não deve ser conduzida com nenhum tipo de *bias*, isto é, inclinação a querer obter um resultado em particular. Desde o início deste projeto, o autor acreditava que o primeiro comitê de classificação projetado teria performance superior aos seus componentes individuais, o que não se provou verdadeiro. Por algum tempo, chegou a crer que este seria o resultado final do trabalho: uma constatação negativa. Finalmente, a substituição de dois dos modelos do comitê inicial por uma rede MLP provou ser o suficiente para que o comitê de classificação obtivesse um resultado ligeiramente superior aos seus componentes.

3.6. Considerações Finais

Neste capítulo discutiu-se o desenvolvimento deste trabalho, bem como os resultados obtidos. Foi proposto um modelo de comitê de classificação satisfatório, seu desempenho foi avaliado, suas limitações, expostas, e foram discutidas as principais dificuldades e limitações do trabalho desenvolvido. Destacou-se a importância de não se ter qualquer inclinação a um resultado em particular ao conduzir uma pesquisa científica. No capítulo seguinte, serão apresentadas as conclusões deste trabalho.

CAPÍTULO 4: CONCLUSÃO

4.1. Contribuições

Os resultados obtidos neste trabalho reafirmam que é possível utilizar a abordagem de comitê de classificação com sucesso para seriar instâncias de expressão oral de emoções.

O erro de projeto cometido no início do trabalho, que consistiu em montar um comitê de classificação com classificadores cujas performances eram muito diferentes, resultando em um *ensemble classifier* insatisfatório, poderá servir de exemplo para outros projetistas iniciantes de comitês de classificação. Este tipo de nuance de projeto é algo que não é documentado em trabalhos da área, mas, ainda assim, é muito relevante. A regra prática aprendida, que diz que performances não muito mais diferentes do que 10% em acurácia são necessárias para que modelos de classificação trabalhem bem em um comitê, é um conhecimento que é passado não através de artigos, mas de conversas de corredor em departamentos de computação. O autor carregará para sempre a importância de trocar ideias com seus colegas de laboratório.

Outra contribuição deste projeto é a de estender o trabalho que vem sendo desenvolvido no grupo de pesquisa do Prof. Dr. Jó Ueyama que, recentemente, desenvolveu um projeto empregando *ensemble classifiers* na seriação de expressões faciais. O futuro cruzamento da classificação de instâncias de expressão vocal com expressão facial servirá para criar um modelo de classificação, espera-se, ainda mais robusto do que ambos os trabalhos de maneira isolada.

Reafirma-se a importância de, ao embarcar em um projeto de pesquisa, ter a mente desprovida de quaisquer ideias de resultados "desejados". Ciência se faz justamente com o *teste* de hipóteses e não com uma ideia que se deseja *provar verdadeira*. O autor aprendeu que resultados negativos são tão importantes quanto, ou até mais importantes do que, resultados positivos, pois são exatamente os resultados negativos que fazem o indivíduo se questionar e entender o comportamento do sistema estudado de maneira muito mais profunda. O autor é muito grato por ter obtido um comitê de classificação com

performance insatisfatória antes de obter um bom comitê de classificação, pois *foi aquele resultado negativo que o fez pensar no verdadeiro significado de fazer ciência*.

4.2. Relacionamento entre o Curso e o Projeto

O autor deste trabalho não teve a oportunidade de cursar, durante a graduação, nenhuma disciplina sobre aprendizado de máquina, especificamente. O pouco contato com a área deu-se nas disciplinas Inteligência Artificial e Introdução a Redes Neurais (optativa), duas das disciplinas mais "empolgantes" do currículo, para o autor. Apesar da grande importância que a Inteligência Computacional vem adquirindo nos últimos tempos, a grade da Engenharia de Computação deixa a desejar nesse quesito.

Além dessas, as disciplinas mais úteis ao desenvolvimento deste projeto foram Processamento Digital de Sinais, Circuitos Eletrônicos I e Circuitos Eletrônicos II, que proveram ao autor o ferramental necessário à compreensão da natureza dos sinais de áudio e do processo de extração de *features* acústicas relacionadas à densidade espectral de energia do sinal, *que acontece totalmente no domínio da frequência*.

O autor deste trabalho teve a oportunidade de passar um ano na República da Irlanda, como aluno visitante na *Trinity College Dublin*, onde participou de um estágio acadêmico de verão em reconhecimento de segmentos de risada no discurso humano utilizando SVM, sob a supervisão da Drª. Naomi Harte, uma especialista em reconhecimento de emoções no discurso humano. O conhecimento adquirido durante este período foi essencial para o desenvolvimento deste trabalho.

4.3. Considerações sobre o Curso de Graduação

O autor deste trabalho está no último ano do curso de Engenharia de Computação oferecido pela Universidade de São Paulo em São Carlos e tem uma visão, infelizmente, muito pessimista do corpo discente, do perfil dos egressos e da metodologia de ensino adotada por boa parte dos docentes responsáveis por disciplinas oferecidas ao curso. É comum que as discussões acerca de como melhorar o curso de Engenharia de Computação

se concentrem em torno da reestruturação curricular. No entanto, um aspecto muito mais importante do que a reestruturação curricular, ainda que conhecido por todos, é sempre ignorado.

Dentre os discentes do curso de Engenharia de Computação desta instituição, é extremamente comum a adoção de toda sorte de meios ilícitos para a aprovação em disciplinas. É entristecedor constatar que muitíssimos indivíduos, uma vez aprovados no processo seletivo de uma universidade pública, adotam uma postura completamente desonesta no cumprimento de suas obrigações acadêmicas. São práticas extremamente comuns: o plágio de trabalhos, a fraude em avaliações escritas, a forja de resultados e a declaração falaciosa da presença obrigatória em aulas, através da falsificação de firmas. É revoltante que o contribuinte paulista, muitas vezes gente simples, cujos filhos nem sequer sonham em estudar numa instituição como a Universidade de São Paulo, precise custear as despesas de indivíduos que agem de maneira tão desonesta para com seus professores, colegas, pais e, acima de tudo, para com a sociedade, que está custeando a formação de jovens que, uma vez egressos desta instituição, engenheiros, não serão capazes de retribuir o volumoso investimento público feito em seus estudos, durante vários anos, através de sua atuação como profissionais competentes.

Obviamente, nenhum docente é responsável pela desonestidade de seus alunos, mas é preciso repensar a maneira como se ensina engenharia e como se avalia o estudante de engenharia nesta instituição. Felizmente, temos docentes extremamente bem preparados, de currículos invejáveis. No entanto, boa parte ensina engenharia e avalia o estudante de engenharia de uma maneira inexplicavelmente parva, através de avaliações escritas e propostas de trabalho exatamente idênticas ano após ano, de forma que a aprovação na maior parte das disciplinas deste curso pode ser obtida por mera *repetição de respostas a problemas já amplamente conhecidos*. Parece não haver nenhuma preocupação com a capacidade de generalização do aprendizado por parte dos alunos, futuros engenheiros, e, portanto, os principais responsáveis pelo avanço tecnológico deste país. Se a condição atual for mantida, o Brasil jamais deixará de ser, majoritariamente, um importador de tecnologias alemã, francesa, americana, etc.

É preciso que a metodologia de ensino e o sistema de avaliação desta instituição sejam repensados, de forma a evitar que um indivíduo consiga obter da Universidade de

São Paulo o título de Engenheiro se valendo de desonestidade durante todo o decorrer de sua graduação.

4.4. Trabalhos Futuros

Este projeto foi iniciado tendo em mente que, caso bem-sucedido, como foi o caso, seria unido ao comitê de classificação baseado em expressões faciais proposto por Mano et al. (2015), integrante do mesmo grupo de pesquisa em que trabalha o autor deste documento. Assim, como trabalho futuro, propõe-se a elaboração de um classificador que una as duas propostas de comitê de classificação de emoções: para voz e para a face, resultando num sistema de classificação mais completo e mais robusto, por cruzar informações de mais de um tipo de expressão motora.

Propõe-se a extensão do conjunto de *features* utilizadas neste trabalho com a adição de *features* prosódicas. Espera-se que a adição de *features* prosódicas melhore significativamente os resultados obtidos com o comitê.

Finalmente, propõe-se o emprego da estratégia de comitê de classificação para expressões motoras gestuais, fechando a tríade de expressões motoras humanas.

REFERÊNCIAS

BHARGAVA'A, M.; POLZEHL, T. Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature. **Proceedings of the ICECIT-2012**, Elsevier, 2012. Disponível em <https://arxiv.org/pdf/1303.1761.pdf>. Acesso em: 31 mai. 2016, 00:27:44.

BURKHARDT, F.; AESCHKE, A.; ROLFES, M.; SENDLMEIER, W. F.; WEISS, B. A Database of German Emotional Speech. **Proc. Interspeech 2005**, Lisboa, 2005.

CANUTO, A. M. d. P. **Combining neural networks and fuzzy logic for applications in character recognition**. Tese (Doutorado) — University of Kent at Canterbury, 2001.

CHANEL, G.; KIERKELS, J. J.; SOLEYMANI, M.; PUN, T. Short-term emotion assessment in a recall paradigm. **International Journal of Human-Computer Studies**, Elsevier, v. 67, 8 ed., 2009. pp. 607–627.

CHAVHAN, Y.; DHORE, M. L.; YESAWARE, P. Speech Emotion Recognition Using Support Vector Machine. **International Journal of Computer Applications**, v.1, 20 ed., 2010. pp. 6–9. Disponível em: <https://goo.gl/IiJ8BS>. Acesso em 31 mai. 2016, 00:08:31.

CLARK, P.; NIBLETT, T. The cn2 induction algorithm. **Machine learning**. Springer, v. 3, 4 ed. , 1989. pp. 261–283.

CONSTANTINI, G.; IADAROLA, I.; PAOLONI, A.; TODISCO, M. **EMOVO Corpus: an Italian Emotional Speech Database**. **Proceedings of the International Conference on Language Resources and Evaluation**, 2014. pp. 3501–3504.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods.** Cambridge University Press, 2000.

DESMET, P. A multilayered model of product emotions. **The Design Journal**. Bloomsbury Journals (formerly Berg Journals), v. 6, 2 ed., 2003. pp. 4–13.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. John Wiley & Sons, 2012.

DUDA, P. E.; RICHARD, O. H., **Pattern Classification and Scene Analysis**. John Wiley and Sons, New York, 1973.

DUPUIS, K.; PICHORA-FULLER, M. K. **Toronto emotional speech set (TESS) Collection**, Toronto, 2010. Disponível em: <https://goo.gl/yiUH04>. Acesso em 30 mai. 2016, 15:17:56.

EKMAN, P. **Cross-cultural studies of facial expression**. New York: Academic, 1973.

EKMAN, P. **Darwin and facial expression: A century of research in review**. Ishk, 2006.

EYBEN, F.; WENINGER, F.; GROSS, F.; SCHULLER, B. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, **Proc. ACM Multimedia (MM)**, Barcelona, 2013. pp. 835-838.

FACELI, K. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen-LTC, 2011.

FONTAINE, J. R.; POORTINGA, Y. H.; SETIADI, B.; MARKAM, S. S. Cognitive structure of emotion terms in Indonesia and the Netherlands. **Cognition & Emotion**. Taylor & Francis, v. 16, 1 ed., 2002. pp. 61–86.

FURUI, S. **Digital Speech Processing, Synthesis and Recognition**, 2 ed., Taylor & Francis, 2000. ISBN: 978-0-82-470452-0.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN I. H. The WEKA Data Mining Software: An Update, **SIGKDD Explorations**. v. 11, 1 ed., 2009.

HASSAM, A.; DAMPER, R. I. Multi-class and hierarchical SVMs for emotion recognition. **Conference: INTERSPEECH 2010, 11th Annual Conference of The International Speech Communication Association**, Makuhari, 2010. Disponível em <https://goo.gl/wmekfu>. Acesso em: 31 mai. 2016, 00:14:46.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition**, 2 ed., Pearson Education, 2009. pp 295–302. ISBN 978-0-13-187321-6.

KHANCHANDANI, K. B.; HUSSAIN, M. A. Emotion recognition using multilayer perceptron and generalized feed forward neural network. **Journal of Scientific and Industrial Research**, v. 68, 2009. pp. 367–371.

LEVENTHAL, H. A perceptual-motor theory of emotion. **Advances in experimental social psychology**, v. 17, 1984. pp. 117–182.

LICHTENSTEIN, A.; OEHME, A.; KUPSCHICK, S.; JÜRGENSOHN, T. Comparing two emotion models for deriving affective states from physiological data. **Affect and Emotion in Human-Computer Interaction**. Springer, 2008. pp. 35–50.

LITTLEWORT, G.; WHITEHILL, J.; WU, T.; FASEL, I.; FRANK, M.; MOVELLAN, J.; BARTLETT, M. The computer expression recognition toolbox (cert). IEEE. **Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on**, 2011. pp. 298–305.

MAHLKE, S.; MINGE, M. Consideration of multiple components of emotions in human-technology interaction. **Affect and emotion in human-computer interaction**. Springer, 2008. pp. 51–62.

MANO, L. Y.; GIANCRISTOFARO, G. T.; FAIÇAL, B. S.; LIBRALON, G. L.; PESSIN, G.; GOMES, P. H. Exploiting the use of ensemble classifiers to enhance the precision of

user's emotion classification. **ACM. Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)**. 2015. p. 5.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, 2003. p. 1.

MORRISON, D.; WANG, R., DE SILVA, L. C. Ensemble methods for spoken emotion recognition in call-centers. **Speech Communication**, v. 49, 2 ed., Elsevier, 2007. pp. 98–112. Disponível em: <http://goo.gl/cXaBRm>. Acesso em: 30 mai. 2016, 11:56:24.

ØHRN, A.; ROWLAND, T. Rough sets: a knowledge discovery technique for multifactorial medical outcomes. **American journal of physical medicine & rehabilitation**, LWW, v. 79, 1 ed., 2000. pp. 100–108.

PAO, T.; CHEN, Y.; YEH, J.; LU, J. Detecting Emotions in Mandarin Speech. **Computational Linguistics and Chinese Language Processing**, v. 10., 3 ed., 2005. pp. 347– 362.

PANKSEPP, J. Toward a general psychobiological theory of emotions. **Behavioral and Brain Sciences**, v. 5, 3 ed., Cambridge University Press, 1982. pp. 407–422. Disponível em: <http://goo.gl/Mnqklm>. Acesso em: 29 mai. 2016, 17:14:32.

PETER, C.; URBAN, B. Emotion in human-computer interaction. **Expanding the Frontiers of Visual Analytics and Visualization**. Springer, 2012. pp. 239–262.

PICARD, R. W. **Affective Computing**. 2 ed. MIT Press, 1998. cap. 6, pp. 165–184. ISBN 0-262-16170-2.

RAMAKRISHNAN, S.; EMARY, I. M. E. Speech emotion recognition approaches in human computer interaction. **Telecommunication Systems**, Springer, v. 52, 3 ed., 2013. pp. 1467–1478.

RUSSELL, J. A. A circumplex model of affect. **Journal of personality and social psychology**, American Psychological Association, v. 39, 6 ed., 1980. p. 1161.

SCHERER, K. R. What are emotions? and how can they be measured? **Social science information**, v. 44, 4 ed., Sage Publications, 2005. pp. 695–729.

SCHULLER, B.; REITER, S.; MULLER, R.; AL-HAMES, M.; LANG, M.; RIGOLL, G. Speaker independent speech emotion recognition by ensemble classification. **IEEE. Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on**. 2005. pp. 864–867.

SCHULLER, B.; BATLINER, A. **Computational Paralinguistics: Emotion, affect and personality in speech and language processing**, 1 ed., Wiley, 2014. ISBN: 978-1-119-97136-8.

PEIPEI, S.; ZHOU, C.; XIONG, C. Automatic Speech Emotion Recognition using Support Vector Machine. **Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on**, v. 2, IEEE, 2011. pp. 621–625.

STEMMLER, G. Methodological considerations in the psychophysiological study of emotion. **Handbook of affective sciences**, 2003. pp. 225–255.

UEYAMA, J.; VILLAS, L. A.; PINTO, A. R.; GONÇALVES, V. P.; PESSIN, G.; PAZZI, R. W.; BRAUN, T. et al. Nodepm: A remote monitoring alert system for energy consumption using probabilistic techniques. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 14, 1 ed., 2014. pp. 848–867.

VERVERIDIS, K.; KOTROPOULOS, C. A Review of Emotional Speech Databases. **Proceedings of the 9th Panhellenic Conference in Informatics**, 2003. pp. 560–575.

XAVIER, R. A. C.; GARCIA, F. E.; NERIS, V. P. de A. Decisões de design de interfaces ruins e o impacto delas na interação: um estudo preliminar considerando o estado emocional de idosos. BRAZILIAN COMPUTER SOCIETY. **Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems**, 2012. pp. 127–136.

ZHOU, F.; QU, X.; HELANDER, M. G.; JIAO, J. R. Affect prediction from physiological measures via visual stimuli. **International Journal of Human-Computer Studies**, v. 69, 12 ed., Elsevier, 2011. pp. 801–819.