RAFAEL IRIYA

Análise de sinais de voz para reconhecimento de emoções

São Paulo
(2014)

RAFAEL IRIYA

Análise de sinais de voz para reconhecimento de emoções

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Mestre em
Ciências

Orientador: Prof. Dr. Miguel Arjona
Ramírez

São Paulo
(2014)

RAFAEL IRIYA

Análise de sinais de voz para reconhecimento de emoções

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Mestre em
Ciências

Área de Concentração:

Orientador: Prof. Dr. Miguel Arjona
Ramírez

São Paulo
(2014)

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 1^o de setembro de 2014.

Assinatura do autor

Assinatura do orientador

FICHA CATALOGRÁFICA

Iriya, Rafael. Análise de sinais de voz para reconhecimento de emoções / R. Iriya. – versão corr. -- São Paulo, 2014. 100 p. Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos. 1. Processamento de sinais 2. Reconhecimento de voz 3. Voz Falada 4. Reconhecimento de emoções I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II. t.

Resumo

Esta pesquisa é motivada pela crescente importância do reconhecimento automático de emoções, em especial através de sinais de voz, e suas aplicações em sistemas para interação homem-máquina. Neste contexto, são estudadas as emoções Felicidade, Medo, Nojo, Raiva, Tédio e Tristeza, além do estado Neutro, que são emoções geralmente consideradas como essenciais para um conjunto básico de emoções. São investigadas diversas questões relacionadas à análise de voz para reconhecimento de emoções, explorando vários parâmetros do sinal de voz, como por exemplo frequência fundamental (pitch), energia de curto prazo, formantes, coeficientes cepstrais e são testadas diferentes técnicas para a classificação, envolvendo reconhecimento de padrões e métodos estatísticos, como K-vizinhos mais próximos (KNN), Máquinas de Vetores de Suporte (SVM), Modelos de Misturas de Gaussianas (GMM) e Modelos Ocultos de Markov (HMM), destacando-se o uso de GMM como principal técnica utilizada por seu custo computacional e desempenho. Neste trabalho é desenvolvido um sistema de identificação em estágio único obtendo-se resultados superiores a diversos sistemas na literatura, com uma taxa de reconhecimento de até 74,86%. Além disso, recorre-se à psicologia e à teoria de emoções para incorporar-se a noção do espaço de emoções e suas dimensões a fim de desenvolver-se um sistema de classificação sequencial em três estágios, que passa por classificações nas dimensões Ativação, Avaliação e Domínio. Este sistema apresenta uma taxa de reconhecimento superior ao do sistema de estágio único, com até 82,41%, ao mesmo tempo em que é identificado um ponto de atenção no sistema de três estágios, que pode apresentar dificuldades na identificação de emoções que possuem baixo índice de reconhecimento em um dos estágios. Uma vez que existem poucos sistemas estado da arte que tratam o problema de verificação de emoções, um sistema também é desenvolvido para esta tarefa, obtendo-se um reconhecimento perfeito para as emoções Raiva, Neutro, Tédio e Tristeza. Por fim, é desenvolvido um sistema híbrido para tratar os problemas de verificação e de identificação em sequência, que tenta resolver o problema do classificador de três estágios e obtém uma taxa de reconhecimento de até 83%.

Palavras-chaves: Reconhecimento de Emoções; Voz; Espaço de emoções; Verificação de emoções; GMM;

Abstract

This work is motivated by the increase on the importance of automatic emotion recognition, especially through speech signals, and its applications in human-machine interaction systems. In this context, the emotions Happiness, Fear, Neutral, Disgust, Anger, Boredom and Sadness are selected for this study, which are usually considered essential for a basic set of emotions. Several topics related to emotion recognition through speech are investigated, including speech features, like pitch, energy, formants and MFCC as well as different classification algorithms that involve pattern recognition and stochastic modelling like K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), where GMM is selected as the main technique for its computational cost and performance. In this work, a single-stage identification system is developed, which outperforms several systems in the literature, with a recognition rate of up to 74.86%. Besides, the idea of emotional space dimensions from Psychology and Emotion Theory is reviewed for the development of a sequential classification system with 3 stages, that passes through classifications on the Activation, Evaluation and Dominance dimensions. This system outperforms the single-stage classifier with a recognition rate of up to 82.41%, at the same time as a point of attention is identified, as this kind of system may show difficulties on the identification of emotions that show low recognition rates in a specific stage. Since there are few state of the art systems that handle emotion verification, a system for this task is also developed in this work, showing itself to be a perfect recognizer for the Anger, Neutral, Boredom and Sadness emotions. Finally, a hybrid system is proposed to handle both the verification and the identification tasks sequentially, which tries to solve the 3-stage classifier problem and shows a recognition rate of up to 83%.

Key-words: Emotion recognition; Speech; Emotional space; Emotion verification; GMM;

Lista de ilustrações

Figura 1 – Diferentes percepções da Teoria de Emoções	28
Figura 2 – Um caso de exemplo segundo as 3 teorias	29
Figura 3 – Emoções Básicas segundo a Roda de Plutchik	33
Figura 4 – Distribuição de emoções no espaço de emoções	34
Figura 5 – Os problemas de identificação e de verificação de emoções	35
Figura 6 – O processo de classificação em 3 estágios	37
Figura 7 – Diagrama de blocos do processo de classificação	38
Figura 8 – Extração de parâmetros	43
Figura 9 – Filtros triangulares na escala mel	48
Figura 10 – Separação linear de amostras de duas classes	56
Figura 11 – Separação por hiperplanos com máxima margem	57
Figura 12 – Modelo de treliças para HMM. Cada nó representa um estado em um instante de tempo (janela) diferente, com os correspondentes vetores de observações	76
Figura 13 – Transição de nós na treliça(forward).	78
Figura 14 – O processo de classificação em 3 estágios	87
Figura 15 – Processo de classificação híbrido.	88

Lista de tabelas

Tabela 1 – Rótulos emocionais segundo Whissel e Plutchik.	31
Tabela 2 – Emoções básicas segundo diversos autores.	32
Tabela 3 – Resumo de alterações nos parâmetros da voz em diferentes emoções segundo Murray e Arnott	42
Tabela 4 – Frases da base EMO-DB	62
Tabela 5 – Número de amostras de áudio por locutor e por emoção na base EMO-DB.	63
Tabela 6 – Índice de Parâmetros	64
Tabela 7 – Exemplos de funções kernel	83
Tabela 8 – Grupos de emoções nas 3 dimensões do espaço de emoções	86
Tabela 9 – Estágios de classificação	86
Tabela 10 – Desempenho de GMM's como a porcentagem de emoções corretamente identificadas para diferentes números de gaussianas (M) em 3 experimentos, variando-se as condições iniciais.	90
Tabela 11 – Matriz de confusão das emoções para GMM, indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) nos casos de melhor desempenho geral.	91
Tabela 12 – Desempenho de HMM's com 8 gaussianas como a porcentagem de emoções corretamente identificadas para diferentes números de estados (Q) em 3 experimentos, com diferentes parâmetros iniciais.	92
Tabela 13 – Matriz de confusão das emoções para locutores de sexo masculino utilizando HMM's com 8 gaussianas, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). São apresentados os casos de melhores desempenhos em cada valor de Q (número de estados).	93
Tabela 14 – Matriz de confusão das emoções para locutores de sexo feminino utilizando HMM's com 8 gaussianas, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). São apresentados os casos de melhores desempenhos em cada valor de Q (número de estados).	94
Tabela 15 – Desempenho de KNN para diferentes números de vizinhos (K) e diferentes conjuntos de parâmetros, sendo L o número de parâmetros utilizados.	95

Tabela 16 – Matriz de confusão para locutores de sexo masculino utilizando KNN, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). Para cada experimento com número de parâmetros L, é apresentado o caso com o número de vizinhos que apresentou o melhor desempenho geral.	96
Tabela 17 – Matriz de confusão para locutores de sexo feminino utilizando KNN, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). Para cada experimento com número de parâmetros L, é apresentado o caso com o número de vizinhos que apresentou o melhor desempenho geral	97
Tabela 18 – Desempenho do SVM como porcentagens gerais de acertos para diferentes tipos de kernel e diferentes conjuntos de L parâmetros.	98
Tabela 19 – Matriz de confusão das emoções para locutores de sexo masculino utilizando SVM com kernel linear, tendo-se as porcentagens de amostras de voz classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) para cada experimento com número de parâmetros L.	99
Tabela 20 – Matriz de confusão das emoções para locutores de sexo feminino utilizando SVM com kernel linear, tendo-se as porcentagens de amostras de voz classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) para cada experimento com número de parâmetros L.	100
Tabela 21 – Parâmetros selecionados para a identificação em estágio único	100
Tabela 22 – Matrizes de confusão do classificador de único estágio indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas).	101
Tabela 23 – Parâmetros selecionados para cada estágio da identificação sequencial	102
Tabela 24 – Resultados gerais para cada estágio de classificação.	102
Tabela 25 – Matrizes de confusão do classificador sequencial de 3 estágios, indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas)	103
Tabela 26 – Parâmetros de voz selecionados para a verificação de cada emoção	104
Tabela 27 – Taxas de reconhecimento gerais para a verificação de cada emoção.	105
Tabela 28 – Porcentagem de reconhecimento para cada par Emoção Escolhida (linhas) $\times$ Emoção Real (colunas).	106
Tabela 29 – Matrizes de confusão para o sistema híbrido com verificação prévia da emoção Nojo.	107

Sumário

I	Introdução	15
1	Introdução	17
1.1	Motivação	17
1.2	Questões	18
1.2.1	Reconhecimento automático x reconhecimento humano	18
1.2.2	Condições para reconhecimento	19
1.2.3	Emoções para reconhecimento	20
1.2.4	Parâmetros de voz e técnicas de classificação	20
1.2.5	Bases de dados	21
1.3	Organização da dissertação	23
1.4	Contribuições do trabalho	24
II	Fundamentos Teóricos	25
2	Emoções e reconhecimento	27
2.1	A teoria de emoções	27
2.1.1	As teorias tradicionais	27
2.1.2	Emoções básicas	30
2.1.3	Modelo dimensional	32
2.2	Reconhecimento automático através da voz	34
2.2.1	Identificação x verificação	35
2.2.2	Identificação segundo o modelo de 3 dimensões	36
2.2.3	O processo de classificação	37
3	Técnicas	41
3.1	Parâmetros de voz	41
3.1.1	Extração de parâmetros	42
3.1.2	Tipos de parâmetros	43
3.1.3	Comparação entre parâmetros de voz	44
3.1.4	Parâmetros de voz populares	45
3.1.4.1	Pitch	45
3.1.4.2	Energia	45
3.1.4.3	Formantes	46
3.1.4.4	MFCC	46
3.1.4.5	LFPC	48
3.1.5	Seleção de parâmetros	48
3.2	Métodos de classificação	49

3.2.1	Classificação estática	50
3.2.2	Classificação dinâmica	50
3.2.3	Comparação entre métodos de classificação	51
3.3	Métodos de classificação populares	51
3.3.1	Gaussian Mixture Models	51
3.3.2	Hidden Markov Models	53
3.3.3	K-Nearest Neighbours	55
3.3.4	Support Vector Machines	56

III Parte Experimental 59

4 Implementação 61

4.1	Berlin Emotional Database of Speech	61
4.2	Parâmetros de voz	62
4.2.1	Parâmetros Seleccionados	62
4.2.2	Extração de parâmetros	63
4.2.3	Algoritmos para extração de parâmetros	65
4.2.3.1	Pitch	65
4.2.3.2	Energia de curto prazo	67
4.2.3.3	Formantes	68
4.2.3.4	MFCC	68
4.2.3.5	LFPC	69
4.2.4	Seleção de Parâmetros - Sequential Forward Selection	70
4.3	Métodos de classificação	70
4.3.1	Gaussian Mixture Models	71
4.3.1.1	Expectation Maximization	71
4.3.1.2	K-Means	72
4.3.1.3	Linde-Buzo-Gray	73
4.3.1.4	Split and Merge	74
4.3.2	Hidden Markov Models	76
4.3.2.1	Problema 1	77
4.3.2.2	Problema 3	78
4.3.3	K-Nearest Neighbours	81
4.3.4	Support Vector Machines	81
4.4	Modelos e estratégias de classificação	83
4.4.1	Identificação em estágio único	83
4.4.2	Parâmetros dependentes de classe	84
4.4.3	Identificação em três estágios/Modelo dimensional	85
4.4.4	Verificação de emoções	86
4.4.5	Classificação híbrida	87

4.5	Testes experimentais	88
5	Experimentos Práticos	89
5.1	Primeiros Experimentos	89
5.1.1	Testes com GMM	89
5.1.2	Testes com HMM	91
5.1.3	Testes com KNN	92
5.1.4	Testes com SVM	94
5.1.5	Conclusão dos Primeiros Testes	97
5.2	Identificação em Estágio Único	98
5.3	Identificação sequencial em 3 estágios	102
5.4	Verificação de emoções	104
5.5	Sistema Híbrido	105
5.6	Conclusão	107
	Referências	109

Parte I

Introdução

1 Introdução

1.1 Motivação

Desde a virada do milênio, o paradigma da comunicação homem-máquina vem se alterando para uma forma mais eficiente, onde não há a necessidade de periféricos como controles-remotos, mouses ou teclados. Em vez disso, espera-se que as máquinas comportem-se cada vez mais como seres humanos, no sentido de sentirem ou perceberem as ações e reações humanas e responderem de forma apropriada. Neste contexto, vem sendo muito desenvolvidas as áreas de sensores, processamento e reconhecimento de imagens e de áudio, inteligência artificial, entre muitos outros. Destacamos aqui o processamento de sinais de voz, principalmente por ser ainda o principal meio de comunicação dos homens e muitas vezes o único recurso disponível para o sistema, como por exemplo na análise de ligações telefônicas. Esta é uma área de estudo de longa data na pesquisa científica cuja importância é inquestionável. Há várias décadas tem se dado muito foco no reconhecimento de palavras e sentenças e também no reconhecimento de locutores para identificação biométrica, trabalho que resultou no desenvolvimento de técnicas com bom desempenho para ambas as tarefas e no fato de hoje em dia encontrarmos estas tecnologias presentes em diversos de telefones celulares, televisores e outros aparelhos domésticos. Entretanto, apesar de atualmente conseguirmos identificar com acurácia “quem” e “o que” foi dito, a questão de “como” foi dito ainda permanece um pouco mais distante. Este assunto tem obtido uma importância crescente na comunicação homem-máquina, pois pode ajudar a tomar decisões mais apropriadas baseadas no contexto do cenário. Não seria ótimo termos uma empregada robô, como vemos em diversos programas de televisão, capaz de interpretar nossos pedidos e reagir de maneira diferente de acordo com nosso estado emocional? É um sonho distante ainda, porém existem outras aplicações que poderiam se beneficiar desta tecnologia. Por exemplo, jogos interativos modernos já são capazes de reconhecer quem é o jogador que está falando e qual o comando que o mesmo executa por meio da voz. Se pudermos incorporar o elemento emotivo no sistema de recepção, este poderá ter uma resposta muito mais apropriada, aumentando a sensação de realidade e proporcionando um maior entretenimento. Além disso, podemos pensar em outras aplicações como ferramentas que possam descobrir se alguém está mentindo, se sua voz apresentar sintomas de ansiedade ou medo, ou acompanhar a evolução de um paciente em depressão ou com outra doença psicológica, verificando como sua voz se altera com o tempo. Um outro exemplo seriam os ambientes de contact-centers onde ferramentas capazes de identificar emoções como raiva ou ansiedade podem ajudar na avaliação dos operadores, pós-processando as ligações e encontrando situações de conflito, ou funcionar em tempo real para descobrir ligações

problemáticas e alertar um supervisor para intervir na conversa.

Neste projeto, estudamos os sinais de voz e alguns de seus parâmetros, como por exemplo frequência fundamental, energia, frequências formantes e algumas propriedades do seu espectro de frequência e investigamos diversas técnicas para se obter um bom reconhecimento de emoções.

1.2 Questões

O desafio de reconhecimento de emoções através da voz é extremamente complexo, necessitando estudos multidisciplinares nas áreas de Filosofia, Psicologia, Biologia, Engenharias e Computação e ainda está longe de ser resolvido, dado que várias questões ainda necessitam ser respondidas com mais certeza a fim de chegarmos mais próximo do objetivo. Nesta seção levantaremos algumas destas questões.

1.2.1 Reconhecimento automático x reconhecimento humano

A primeira coisa a se pensar quando estamos desenvolvendo um sistema para reconhecimento automático de emoções, no caso através da voz, é se realmente este sistema terá um desempenho aceitável, e se este desempenho é comparável ao desempenho humano. Parece claro que os humanos podem perceber e processar diferenças nas propriedades da voz e utilizar estas informações para identificar o estado emocional do locutor. Entretanto a verdade é que nem sempre conseguimos realizar esta análise perfeitamente, pois o reconhecimento de emoções é uma tarefa complexa que muitas vezes exige outros parâmetros como o conhecimento do contexto, do significado semântico da mensagem e da expressão facial da pessoa. Alguns estudos foram realizados mostrando que a habilidade humana para reconhecer emoções somente pela voz está entre 60% a 80% (1)(2) , (3)(4). Nestes testes, apenas os aspectos acústicos são levados em conta já que as frases utilizadas estão fora de contexto e não possuem nenhum significado semântico. Em (1), trechos de áudio de voz retirados de filmes contendo as emoções felicidade, raiva e tristeza são identificadas por quatro indivíduos, dois dos quais não tem familiaridade com a língua inglesa, utilizada nos trechos de áudio. Para os ouvintes familiarizados com o idioma inglês, o resultado geral apresentou uma acurácia geral de aproximadamente 85% enquanto os ouvintes não-familiarizados com a língua inglesa apresentaram uma taxa de aproximadamente 61%. Em (4) com as mesmas emoções, mais o estado neutro, o resultado foi de aproximadamente 78%, enquanto em (2), com seis emoções, Felicidade, Raiva, Tristeza, Medo, Nojo e Surpresa, o resultado foi de aproximadamente 62%, sugerindo que a acurácia tende a diminuir com o aumento do conjunto de emoções. Assim, estes resultados mostram que é teoricamente possível criarmos uma máquina capaz de reproduzir a habilidade humana, desde que conheçamos exatamente os parâmetros e os algoritmos que o ouvido e o cérebro humano

utilizam para realizar esta análise. Entretanto, se realmente formos na linha de reproduzir os mecanismos humanos, na prática não conhecemos de fato nem os parâmetros nem os algoritmos utilizados e, ao mesmo tempo, ficamos limitados a um desempenho imperfeito. De qualquer maneira, espera-se que selecionando adequadamente os parâmetros de voz e utilizando-se uma técnica adequada para a classificação, possamos criar sistemas com desempenho satisfatório. Diversos sistemas automáticos já foram propostos ao longo dos anos e já existem casos registrados onde a acurácia foi satisfatória, equiparando-se à humana, boa, superando a acurácia humana e excelente, beirando 100% de acertos, como exemplos (5) com 67,92%, (6) com 83,17% e (7) com 99,5%, respectivamente. Todavia, vale lembrar que o desempenho dos reconhecedores automáticos fica limitado às bases de dados utilizadas, ao número de emoções reconhecidas, ao número de locutores e às condições de independência de gênero e de locutor estabelecidas nos testes.

1.2.2 Condições para reconhecimento

Uma vez estabelecido que as emoções podem ser reconhecidas pela voz, precisamos entender o que determina a acurácia do processo. O reconhecimento é independente de texto, de idioma, de sexo ou de locutor? É de bom senso pensar que o conhecimento prévio destas condições facilita muito a tarefa. Porém seria realmente possível obter-se um bom desempenho em um cenário sem nenhum conhecimento prévio?

Em (6), a inclusão da independência de locutor diminui drasticamente a taxa de reconhecimento do sistema de 77% para 55% , mostrando que o cenário independente de locutor é extremamente mais difícil, como esperado. Já a análise semântica da mensagem tem grande utilidade prática e pode ajudar bastante no reconhecimento, sendo que já existem estudos que se utilizam do reconhecimento semântico de voz aliado ao reconhecimento de emoções (8)(9)(10). Porém, como comprovado pelos testes humanos, a análise semântica não é essencial para o reconhecimento. Quanto a questão de gênero, vários autores constatarem que os parâmetros da voz tem papéis diferentes para o reconhecimento de emoções em locutores do sexo masculino e feminino (11)(12)(13), o que faz sentido se pensarmos que a frequência fundamental das mulheres é geralmente mais alta do que a dos homens (11). Uma estratégia comum é realizar o reconhecimento para homens e para mulheres em sequências distintas, separando manualmente as amostras de homens e mulheres ou realizando um pré-processamento em um detector de gênero. Por fim a independência de idioma é um tanto complicada, pois Darwin (14) já afirmava que a expressão das emoções é fortemente biológica e resulta de hábitos e padrões que foram selecionados através da evolução, argumento que pode ser facilmente apoiado pelo fato de normalmente falarmos mais alto e mais agudo quando estamos com raiva, simplesmente porque o ouvido humano é mais sensível a estas alterações. Por outro lado há aqueles que consideram que as emoções possuem uma componente cultural (15), e assim diferentes povos podem expressar as emoções de maneira diferente, com comportamentos que foram sendo criados e modificados

ao longo dos anos, de acordo com a cultura do povo. Voltaremos a essa questão no capítulo 2, mas por ora destacamos que a maioria dos estudos costuma considerar a componente biológica mais importante e frequentemente utilizam bases de dados compostas por trechos de áudio em idiomas diferentes da língua nativa dos autores.

Por estes motivos, neste projeto trabalhamos em cenário independente de texto, locutor e idioma, porém com dependência de gênero, sendo as amostras de homens e mulheres separadas manualmente.

1.2.3 Emoções para reconhecimento

Espera-se que um sistema para reconhecimento automático de emoções tenha a capacidade de reconhecer o maior número de emoções possíveis, mas sabe-se que esta não é uma tarefa nada fácil. Quando pensamos em emoções, primeiramente nos vem à mente emoções como Felicidade, Tristeza, Raiva, porém, conforme começamos a ir mais fundo, a lista aumenta de forma assustadora, podendo chegar a milhares de palavras que remetem a emoções de alguma forma (16). Algumas das emoções, como por exemplo culpa ou vergonha, são difíceis até para os humanos as reconhecerem, sendo algumas vezes de conhecimento somente da pessoa que as experimenta. Portanto, que emoções devemos esperar que o sistema reconheça? A receita tradicional consiste em definir um conjunto de emoções básicas, composto das mesmas emoções que facilmente nos vêm à mente, que também sejam de fácil identificação, especialmente pela voz, pelo menos do ponto de vista humano, e principalmente cujo reconhecimento automático tenha alguma aplicação prática direta. Todavia, ainda não há um consenso de qual seria esse conjunto, embora muitas das emoções coincidam de estudo para estudo. Aprofundaremos melhor este assunto no capítulo 2.

1.2.4 Parâmetros de voz e técnicas de classificação

Quais são os parâmetros da voz que melhor caracterizam as emoções e quais as técnicas mais apropriadas para o reconhecimento? Estas são provavelmente as perguntas mais difíceis de se responder e talvez as mais importantes. Estas perguntas podem ser substituídas pela seguinte: como obter um modelo matemático que representa os mecanismos reais de reconhecimento de emoções? Se soubéssemos exatamente quais parâmetros da voz e que mecanismos o ouvido e o cérebro humano utilizam para realizar o discernimento entre emoções, ficaria muito fácil desenvolver um sistema para substituir um humano. Entretanto, normalmente é impossível traduzir fielmente todos os aspectos da realidade em modelos matemáticos. Muitos autores já chegaram a conclusões sobre os parâmetros mais importantes em seus estudos, porém estes ainda diferem bastante de estudo para estudo. A maioria dos autores exploram parâmetros prosódicos como energia e frequência fundamental, assim como propriedades espectrais da voz, como a distribuição

espectral e coeficientes cepstrais, sendo constatado que ambos os tipos possuem grande influência no reconhecimento de emoções. Para sabermos as técnicas e mecanismos a serem utilizados, precisamos primeiramente descobrir como modelar uma emoção. Pode uma emoção ser representada por um filtro, por um modelo probabilístico, ou podemos simplesmente considerar cada emoção como um conjunto de parâmetros, que pode ser classificada segundo algoritmos clássicos de reconhecimento de padrões? A literatura vem mostrando que tanto abordagens probabilísticas como determinísticas têm obtido bons resultados, principalmente utilizando algoritmos como Máquinas de Vetores de Suporte, Redes Neurais e modelos probabilísticos como Modelos de Misturas de Gaussianas ou Modelos Ocultos de Markov.

Enfim, a melhor solução ainda está atrelada a cada estudo individual. Todas estas questões são abordadas no capítulo 3.

1.2.5 Bases de dados

Como em qualquer projeto na área de processamento de voz, as bases de dados representam uma parte importante do projeto, pois fornecem a massa de dados necessária para a validação dos algoritmos e dos mecanismos desenvolvidos no estudo. Espera-se que boas bases de dados contenham milhares de amostras, centenas de locutores diferentes de ambos os sexos e o maior número possível de emoções, uma vez que os algoritmos de classificação estado da arte respondem melhor à dados similares aos que foram utilizados na geração dos modelos, sendo muitas vezes incapazes de se comportar corretamente quando deparados com dados muito diferentes dos utilizados no treinamento. Entretanto a criação de bases de dados para reconhecimento de emoções é uma tarefa complexa (17), e não é tão direta como a criação de outros tipos de bases de dados, como bases que representam diferentes tipos de ruídos sonoros ou bases para reconhecimento de palavras ou de locutores, uma vez que deve-se levar em conta a geração das emoções, que não é trivial.

Em geral, a manifestação das emoções pode se dar de 3 formas: atuada, induzida ou real. A manifestação real de uma emoção não pode ser obtida facilmente, simplesmente juntando um grupo de locutores em um estúdio e ordenando-os a manifestar emoções e gravar sentenças, repetidamente. Por este motivo é comum o uso de atuações, nas quais os atores gravam frases de conteúdo semântico desprezível, enquanto tentam mergulhar em uma determinada emoção durante a pronúncia de uma frase. Uma técnica comum utilizada por atores profissionais é forçar a lembrança de uma situação que desencadeia a emoção, por exemplo lembrar de um momento triste da vida para expressar tristeza. Emoções atuadas tendem a se diferenciar um pouco de manifestações reais, uma vez que as interpretações podem ser exageradas ou omitir detalhes importantes de uma manifestação real. Por outro lado, algumas pessoas argumentam que a utilização de emoções atuadas pode ser até melhor do que a de emoções reais, uma vez que elas são puras e livres de outros

fatores incontroláveis (18). Manifestações reais são em geral mais fracas e geralmente estão misturadas com outras emoções, o que torna ainda mais difícil obtê-las de forma natural. Exemplos de bases de dados atuadas bastante conhecidas são a Danish Emotional Speech database (DES)(19) e a Berlin Database of Emotional Speech (EMODB) (20).

A manifestação de emoções pode ser induzida de diversas formas. Por exemplo, podem ser apresentados textos ou cenas de filmes para os participantes e posteriormente este são convidados a comentarem sobre o que viram, de forma espontânea. Além disso, pode-se pedir diretamente para que os participantes comentem sobre alguma situação particular de suas vidas. Um outro método frequentemente usado é a montagem de um cenário fictício onde os participantes pensam que estão participando da avaliação de uma ferramenta, por exemplo um software para reconhecimento de voz e algumas situações são provocadas, para induzir uma reação emocional no participante, como a falha constante do software. A desvantagem da indução de emoções é que o resultado é imprevisível, e não se sabe o quanto o participante vai se entregar à tarefa, quais emoções serão demonstradas e qual o grau de manifestação. Como exemplos de bases induzidas temos a Belfast Induced Natural Emotion Database (21).

Por fim, manifestações reais de emoções não podem ser gravadas facilmente em estúdio e portanto elas devem ser buscadas de outra maneira, em situações onde os locutores estão realmente interagindo de acordo com o contexto e com suas personalidades. Em princípio, estes tipos de amostras poderiam ser encontradas facilmente em programas de rádio, televisão ou filmes, porém existem algumas restrições que dificultam o trabalho. Primeiramente, necessita-se de autorização legal tanto dos participantes, quanto das emissoras e dos detentores dos direitos legais dos áudios. Segundo, é difícil obter amostras de áudio com condições boas de qualidade e inteligibilidade e alta relação sinal-ruído, uma vez que estas gravações não foram geradas com propósito de análise da voz. A Belfast Naturalistic Database (22) e a EmoTV Database (23) são exemplos de bases de dados contendo trechos de programas de tv.

1.3 Organização da dissertação

O restante deste trabalho está organizado como indicado abaixo:

- O capítulo 1 introduz o problema, levanta algumas das questões pendentes no problema de reconhecimento de emoções através da voz e termina por resumir as principais contribuições do trabalho.
- O capítulo 2 introduz alguns fundamentos da teoria de emoções segundo a Tradição Psicológica, onde é apresentado o conceito de emoções, as principais emoções básicas, o modelo dimensional do espaço de emoções e a distribuição das emoções segundo este modelo. Por fim é discutido o processo de reconhecimento automático de emoções através da voz.
- O capítulo 3 discute os principais parâmetros de voz e as principais técnicas para treinamento e classificação utilizados em sistemas estado da arte, entre aspectos teóricos e práticos. Também é comentada a importância do uso de algoritmos de seleção dos parâmetros e é descrito o algoritmo Sequential Forward Selection utilizado neste projeto.
- O capítulo 4 descreve detalhes da implementação dos sistemas propostos neste projeto, os algoritmos utilizados e apresenta os testes realizados.
- Finalmente, o capítulo 5 discute os experimentos práticos realizados e os resultados obtidos. Inicialmente é realizado um estudo comparativo entre diversos métodos de classificação utilizando apenas o Pitch e a energia de curto prazo. Depois, são inseridos mais parâmetros e utilizados algoritmos de seleção de parâmetros para encontrar o conjunto ótimo. São realizados testes para os problemas de verificação e de identificação, utilizando o modelo de 3 dimensões do espaço de emoções e é criado um algoritmo híbrido misturando as duas tarefas. Enfim, são apresentadas as conclusões do trabalho e sugestões para trabalhos futuros.

1.4 Contribuições do trabalho

- Desenvolvimento de sistema de classificação sequencial em 3 estágios, que se mostra superior ao sistema de estágio único (24).
- Identificação de ponto para melhoria no sistema de classificação sequencial de 3 estágios, trabalhando-se com a emoção Nojo (24).
- Utilização de algoritmo Linde-Buzo-Gray (LBG) modificado, com equação própria para separação de centros, como solução para a inicialização de gaussianas (24).
- Utilização conjunta de Log Frequency Power Coefficients (LFPC) e Mel Frequency Cepstral Coefficients (MFCC) e evidências da importância de ambos como parâmetros de voz para o reconhecimento de emoções (24).
- Investigação do problema de verificação de emoções, que é pouco investigado na literatura de reconhecimento de emoções.
- Desenvolvimento de sistema híbrido, combinando as tarefas de verificação e de identificação de emoções.
- Obtenção de conjuntos ótimos de parâmetros de voz com poucos parâmetros, o que pode indicar uma solução mais robusta (24).

Parte II

Fundamentos Teóricos

2 Emoções e reconhecimento

2.1 A teoria de emoções

Emoções fazem parte da vida diária dos seres humanos e todos conseguimos entender o que significa a palavra emoção, pelo menos superficialmente. Entretanto, desenvolver um sistema para reconhecimento automático de emoções é uma tarefa bastante complexa, que começa por um estudo mais profundo do que vem a ser uma emoção. Neste ponto, entra o estudo da teoria de emoções, que envolve diversas áreas como Filosofia, Biologia e principalmente a Psicologia, tradicionalmente agrupado num conjunto de pesquisas denominado a Tradição Psicológica (25). Apesar de a Tradição Psicológica não possuir conclusões firmes, ela fornece um amplo conjunto de idéias e técnicas que servem de base para o desenvolvimento de sistemas para reconhecimento automático de emoções.

Existe uma dificuldade em lidar com as palavras “emoção”, “sentimento” e “humor” que frequentemente são tratadas como sinônimos embora existam sutis diferenças. Há também a palavra da língua inglesa “affect” que não possui uma tradução fiel em português, sendo as vezes chamado de “estado afetivo”. Para entender estas diferenças, é necessário ter a idéia de que a “emoção” possui um significado amplo e que muitas vezes a emoção é confundida com as reações do corpo ao vivenciar o estado emocional, das quais a pessoa tem consciência e que podem ser observadas externamente. A emoção no sentido amplo pode ser traduzida na palavra “affect” ou “estado afetivo”. Já a diferença entre emoção e humor está principalmente na duração do estado emocional, sendo que a primeira está relacionada a um estado temporário, geralmente causado por alguma influência externa, enquanto o humor pode ter duração indeterminada, refletindo a personalidade da pessoa, alguma doença emocional, como por exemplo depressão, ou preferências como gostar de ou odiar algo ou alguém. Alguns autores (26) definem que “emoção” está relacionado a uma reação fisiológica, sendo portanto observável, enquanto “sentimento” está relacionado ao estado mental, sendo intrínseco à pessoa que vive o estado emocional. Ou seja, a emoção é orientada para o exterior, enquanto o sentimento é orientado para o interior, sendo a primeira de fato nosso objeto de estudo.

2.1.1 As teorias tradicionais

O fato de as emoções provocarem reações no corpo é o que possibilita o nosso estudo, uma vez que a tarefa de reconhecimento consiste em avaliar as alterações da voz do locutor para concluir qual seu estado emocional. Segundo a psicologia há várias teorias que explicam a manifestação da emoção (27), sendo uma das mais famosas a de James-Lang,

Figura 1 – Diferentes percepções da Teoria de Emoções

Fonte: Autoria própria

representada na Figura 1a. William James propôs que um indivíduo, após perceber um estímulo exterior, sofre alterações fisiológicas perturbadoras, como palpitações, falta de ar, etc, sendo precisamente o reconhecimento desses sintomas pelo cérebro o que gera a emoção. Uma teoria alternativa proposta alguns anos depois é a de Cannon-Bard (Figura 1b), que sugere que as reações fisiológicas e a experiência emocional ocorrem simultaneamente, uma vez que a interpretação do estímulo exterior ocorre simultaneamente em duas partes distintas do cérebro. Além destas teorias, existem as teorias cognitivistas, que afirmam que os processos cognitivos, como as percepções, recordações e aprendizagens, são fundamentais para se perceberem as emoções. Um exemplo é a Teoria Bifatorial de Schachter-Singer, representada na Figura 1c, que presume que nossa experiência da emoção cresce a partir de nossa consciência da excitação fisiológica. Segundo Schachter, uma experiência emocional exige dois fatores: uma alta excitação psicológica e uma interpretação consciente da excitação, a qual buscamos no mundo exterior. A figura 2 demonstra um exemplo em que um indivíduo avista uma cobra e a emoção Medo é gerada, segundo as três teorias. Apesar de ligeiramente diferentes, as três teorias tem em comum que as emoções possuem uma causa externa, que de certa forma reflete na natureza das emoções, e geram alterações fisiológicas que podem ser diretamente observadas. Estes dois fatos são extremamente importantes e servem como base para o desenvolvimento de um sistema para reconhecimento automático de emoções.

Figura 2 – Um caso de exemplo segundo as 3 teorias

Fonte: Autoria própria

Além destas teorias, temos a visão evolucionista do biólogo Charles Darwin (14). Darwin introduziu a idéia de que as emoções estão associadas a hábitos e padrões de ações selecionados naturalmente ao longo de anos e anos de evolução, por representarem vantagens para a sobrevivência. Como exemplos ele utiliza principalmente as alterações em expressões faciais e nos movimentos corporais, em decorrência de determinadas emoções para seres humanos e também animais. No caso da voz, podemos pensar que há uma tendência natural em falarmos mais alto e mais agudo quando estamos mais excitados, felizes ou com raiva e isto pode ser explicado pelo fato de nosso ouvido ser mais sensível nestas condições. Desta forma, a expressão das emoções deve manifestar-se da mesma maneira para toda a raça humana, independente da localização e da cultura do povo. Entretanto esta é uma questão que é desafiada por alguns autores, que afirmam que as emoções possuem uma componente cultural. Averill e Harre(15) afirmam que as emoções são modos institucionalizados de interpretar e responder a certos tipos de situações. Desta forma, povos de diferentes regiões possuem comportamentos típicos que foram sendo criados e modificados ao longo dos anos, por razões históricas e culturais. É fácil entendermos que algumas culturas podem ter uma maneira mais suave de falar e outras podem ser mais agressivas, mesmo em estado neutro. Em (1), testes de reconhecimento com ouvintes humanos mostrou uma diferença de mais de 20% de acurácia entre pessoas que tinham conhecimento do idioma daqueles que não possuíam. Todavia, o tradicional modelo independente de cultura é o mais utilizado, sendo a componente cultural considerada inexistente ou irrelevante, o que resulta em muitos autores utilizarem bases de dados em idioma diferente de suas línguas nativas em seus estudos. Em (28) é realizado um estudo com cinco idiomas diferentes existentes na Índia onde o sistema tenta reconhecer sete emoções: Raiva, Nojo, Medo, Felicidade, Tristeza, Surpresa e Neutro, tendo alcançado a surpreendente taxa de reconhecimento de 99,7%.

2.1.2 Emoções básicas

Conhecendo um pouco mais da teoria de emoções, devemos então pensar quais são as emoções que conhecemos e quais delas devem ser reconhecidas por um sistema automático, cujo objetivo é classificar estados emocionais com rótulos conhecidos. A princípio podemos pensar que estes são poucos, porém o homem utiliza um número imenso de rótulos para caracterizar estados emocionais. Em (16) é proposto um dicionário com mais de 4000 palavras descrevendo estados emocionais, das quais podemos ver alguns exemplos na tabela 1. Entretanto, não podemos classificar todas estas palavras como estados emocionais, já que muitas são praticamente sinônimos, como por exemplo irritado e aborrecido e algumas podem ser mais corretamente relacionadas a atitudes, como hostil, sarcástico, e curioso. Além disso, não podemos esperar que um sistema automático reconheça todos estes estados, uma vez que muitos deles são difíceis até mesmo para um ser humano reconhecer, sendo

alguns de conhecimento apenas da pessoa que vive o estado.

Tabela 1 – Rótulos emocionais segundo Whissel e Plutchik.

Receptivo	Aventureiro	Medroso
Corajoso	Agressivo	Impressionado
Irritado	Nervoso(raiva)	Furioso
Ansioso	Apático	Envergonhado
Frio	Entediado	Calmo
Atento	Cauteloso	Nauseado
Interessado	Desinteressado	Alegre
Preocupado	Desconfiado	Entusiasmado
Invejoso	Vazio	Generoso
Culpado	Feliz	Hesitante
Esperançoso	Confuso	Curioso
Desafiador	Deprimido	Desesperado
Nervoso(ansioso)	Paciente	Satisfeito
Instatisfeito	Orgulhoso	Hostil
Humilhado	Indeciso	Sarcástico
Ciumento	Possessivo	Solitário
Sereno	Tímido	Sociável
Arrependido	Teimoso	Submisso
Suspeito	Aterrorizado	Pronto
Rejeitado	Revoltado	Triste
Ressentido	Amedrontado	Tolerante
Confiante	Incerto	Infeliz
Cooperativo	Vingativo	Pensativo
Apreensivo	Maravilhado	Desobediente
Excitado	Empolgado	Expectante
Extático	Agradável	Desagradável
Disposto	Antagônico	Antecipatório
Intimidado	Atrevido	Exultante
Exasperado	Desamparado	Abandonado
Perdido	Sombrio	Ganancioso
Angustiado	Aborrecido	Desdenhoso
Contente	Crítico	Encantado
Exigente	Decepcionado	Desapontado
Animado	Desanimado	Desencorajado
Indignado	Apavorado	Perplexo
Impulsivo	Inquisitivo	Repugnante
Antecipatório	Detestável	Manso
Surpreso	Pacífico	Constrangido
Controlado	Incontrolável	Simpático
Intrigado	Imprudente	Despreocupado
Cordial	Alerta	Neutro

A maneira tradicional consiste em definir um conjunto de emoções básicas, composto de emoções que facilmente nos vêm à mente, e de ampla aceitação para a maioria das

peessoas. Espera-se também que estas emoções sejam de fácil identificação, pelo menos do ponto de vista humano e, principalmente, cujo reconhecimento automático tenha alguma aplicação prática direta. Na maioria dos casos, o conjunto de emoções básicas contém as emoções Alegria/Felicidade, Raiva, Medo/Ansiedade, Tristeza, Surpresa e Nojo. Ortony e Turner descrevem um resumo de emoções básicas selecionadas em diversos estudos, representado na tabela 2 (29).

Tabela 2 – Emoções básicas segundo diversos autores.

Autores	Emoções Básicas
Plutchik(30)	Aceitação, Antecipação, Alegria, Nojo, Medo, Raiva, Surpresa, Tristeza
Ekman et al(31)	Alegria, Nojo, Medo, Raiva, Surpresa, Tristeza
Oatley e Johnson-Laird(32)	Felicidade, Ansiedade, Nojo, Raiva, Tristeza
Weiner e Graham(33)	Felicidade, Tristeza

Um sistema famoso que descreve as emoções utilizando emoções básicas é a Roda de Plutchik, representada na Figura 3. Nela, 8 emoções básicas, Aceitação, Antecipação, Alegria, Nojo, Medo, Raiva, Surpresa e Tristeza são apresentadas em pares opostos (Aceitação x Nojo, Medo x Raiva, Antecipação x Surpresa, Alegria x Tristeza (30). Nesta representação, as demais emoções podem ser representadas por pontos no círculo, estando sempre entre duas das emoções básicas e sendo a intensidade das emoções definida pela distância ao centro da roda, que desta forma, representa o estado neutro. Este modelo estabelece que as emoções possuem naturezas bem distintas umas das outras e que são mutuamente exclusivas.

2.1.3 Modelo dimensional

Uma alternativa ao modelo de Plutchik é o modelo dimensional que descreve uma emoção como um ponto num espaço de emoções composto por duas ou três dimensões (34). A primeira dimensão é comumente denominada Ativação ou Excitação, que representa o grau de intensidade da emoção, isto é, o quanto esta emoção altera o estado emocional do indivíduo, em comparação com o estado neutro. Podemos facilmente pensar que Raiva e Alegria são emoções com alto nível de excitação. A segunda dimensão é chamada de Avaliação, Valência ou Prazer e têm relação com a nossa noção de julgamento das emoções em positivas e negativas, boas e más. Neste caso não há exemplo mais simples do que a Tristeza ser uma emoção claramente negativa e a Felicidade uma emoção claramente positiva. Por fim, a terceira dimensão, por vezes omitida, é denominada Domínio, Poder ou Controle, e tem a ver com o quanto a emoção têm influência em um indivíduo, ou

Figura 3 – Emoções Básicas segundo a Roda de Plutchik

Fonte: Autoria própria

seja, o quanto ela o controla ou quanto a pessoa tem controle sobre a emoção. A raiva e a tristeza são emoções que facilmente tomam o controle da pessoa, tornando-a vulnerável e irracional.

Embora a classificação nas 3 dimensões seja por vezes empírica ou baseada em senso comum, é possível que certos aspectos sejam refletidos em reações fisiológicas, que podem ser medidas e utilizadas na análise como parâmetros fisiológicos. Por exemplo, é de se esperar que alguém em alto estado de excitação, feliz ou com raiva, fale mais alto.

É óbvio que este modelo ainda não contempla todos os aspectos da emoção, afinal não podemos afirmar que as emoções são compostas por apenas duas ou três dimensões e que estas dimensões são de fato independentes ou ortogonais. Porém, ele ainda leva a consequências interessantes:

- Emoções podem ser agrupadas conforme suas posições em cada um dos eixos, podendo melhor caracterizar quais emoções são mais semelhantes ou mais distintas.
- Emoções próximas em uma ou mais dimensões devem ser mais difíceis de ser diferenciadas e devem apresentar parâmetros de voz semelhantes (34).
- Emoções próximas em uma dimensão porém distantes nas demais, podem ainda ser diferenciadas.

A figura 4 mostra a distribuição de algumas das emoções frequentemente consideradas como básicas, segundo o modelo de três dimensões com o espaço emocional composto

Figura 4 – Distribuição de emoções no espaço de emoções

Fonte: Autoria própria

pelos eixos Ativação, Avaliação e Domínio. Em comparação à Roda de Plutchik, algumas emoções podem ser opostas em um eixo porém localizar-se próximas em outro.

2.2 Reconhecimento automático através da voz

O reconhecimento automático de emoções através da voz consiste na tarefa de processar um sinal de áudio, identificar as alterações da voz e classificá-lo como pertencendo a alguma emoção conhecida. Podemos comparar este processo com a teoria de James-Lange, onde as alterações fisiológicas determinam a emoção.

Como já mencionado anteriormente, o problema de reconhecimento automático de emoções através da voz vem sendo tradicionalmente tratado como um problema estatístico ou de reconhecimento de padrões, análogo aos problemas de reconhecimento de voz e de locutor. Diversas amostras de voz cuja emoção é conhecida são utilizadas para gerar um modelo da emoção em questão, utilizando determinados parâmetros da voz, que devem ser extraídos do sinal de voz original. Posteriormente, amostras cujas emoções são desconhecidas são classificadas de acordo com a similaridade aos modelos gerados, utilizando os mesmos parâmetros de voz. A unidade de trabalho é geralmente uma frase completa, contendo uma manifestação clara de uma emoção, o que representa um cenário

parecido com o qual o cérebro humano trabalha para discernir emoções. Em outras palavras, as bases de dados utilizadas contêm sinais de áudio de poucos segundos cujo conteúdo deve fornecer informações suficientes para o reconhecimento. Os parâmetros de voz utilizados são escolhidos de acordo com idéias provindas da teoria de emoções e de resultados de estudos prévios envolvendo reconhecimento automático pela voz, e são melhor descritos no capítulo 3, assim como as técnicas para classificação.

2.2.1 Identificação x verificação

Frequentemente, um problema de reconhecimento, qualquer que seja ele, pode apresentar-se de duas maneiras: identificação ou verificação. Por identificação de emoções, entende-se o problema de classificar uma amostra como pertencendo a uma dentre um universo de E emoções (ver Figura 5a), cada qual treinada utilizando somente as amostras rotuladas como contendo a emoção em questão. Já a verificação de emoções envolve decidir se uma dada amostra pertence ou não a emoção em questão, existindo apenas dois modelos, o modelo da emoção e o modelo complementar (ver Figura 5b), treinado pelas amostras de todas as outras emoções do universo de treinamento.

Figura 5 – Os problemas de identificação e de verificação de emoções

Fonte: Autoria própria

Pode-se notar que ambos os problemas são limitados pelo tamanho do universo de emoções, que se for muito grande deve aumentar a complexidade do reconhecimento, e se for muito pequeno, torna o problema muito específico para as emoções escolhidas. Tradicionalmente, a solução é adotar um pequeno conjunto de emoções formado pelas

emoções consideradas básicas, o que já apresenta uma série de aplicações práticas. A decisão pela implementação de um sistema para identificação ou para verificação de emoções depende da aplicação em si. Por exemplo, um sistema de detecção de mentiras poderia estar mais interessado em detectar uma emoção como medo ou ansiedade e o custo de falsos positivos seria menor que o de falsos negativos, enquanto um jogo interativo estaria interessado em detectar o maior número de emoções possíveis.

2.2.2 Identificação segundo o modelo de 3 dimensões

O modelo dimensional do espaço de emoções pode ser utilizado para auxiliar o reconhecimento automático. Desta forma, o problema geral de reconhecimento de um conjunto grande de emoções pode ser dividido em problemas menores, se agruparmos as emoções conforme suas similaridades em cada dimensão, com o objetivo de se atingir um desempenho final superior ao da classificação direta. Assim, podemos ter três classificadores, um para cada dimensão do espaço emocional (Ativação, Avaliação e Domínio), cada qual especialista em uma classificação segundo um único aspecto. Um sistema seguindo este modelo pode ser do tipo sequencial, através de classificações sucessivas, como representado na figura 6, onde o conjunto de possíveis emoções vai afunilando após a classificação em cada uma das dimensões. A classificação se dá primeiramente na dimensão da Ativação, separando o conjunto total de emoções em alta excitação e baixa excitação. Depois, há uma classificação na dimensão da Avaliação, separando as emoções de caráter negativo das emoções de caráter positivo. Finalmente, a dimensão do Domínio dá a classificação final da emoção. A ordem da sequência de dimensões deve ser escolhida de acordo com a taxa de reconhecimento de cada classificador, quando selecionado um conjunto maior de emoções. A dimensão da Ativação é escolhida primeiro por ser intuitivamente mais fácil de ser reconhecida, tanto para os humanos quanto para as máquinas (35). É de conhecimento geral que parâmetros de voz simples como volume e frequência fundamental são de extrema importância para a dimensão da ativação (25) e é fácil para um ser humano notar a alteração nestes parâmetros quando uma pessoa está altamente emocionada.

Figura 6 – O processo de classificação em 3 estágios

Fonte: Autoria própria

Já existem estudos que fazem uso das dimensões do espaço de emoções para o reconhecimento através da voz. Em (36) o sistema utiliza parâmetros de voz de diferentes naturezas para as dimensões da Avaliação e da Ativação para identificar cinco emoções: Felicidade, Raiva, Tédio, Neutro e Tristeza. O experimento mostra que ao se utilizar um único conjunto de parâmetros de voz em um único classificador, as emoções mais próximas, principalmente na dimensão da Ativação, tendem a se confundir frequentemente. Lugger e Yang (37) utilizam seis emoções: Felicidade, Raiva, Tédio, Nojo, Medo e Tristeza, que são reconhecidas através de um sistema de classificação hierárquica com três estágios, passando pelas dimensões Ativação, Domínio e Avaliação. Os autores comparam o desempenho do sistema sequencial com o sistema de um único estágio, encontrando uma melhoria considerável de 14%. Xiao et al (38) apresentam um sistema hierárquico com as mesmas seis emoções, porém a teoria de emoções é levada em conta apenas para a dimensão da ativação. Os autores ainda utilizam um primeiro estágio adicional para separação prévia de gênero, aumentando o desempenho em cerca de 3% em relação ao cenário independente de gênero.

Ainda não possuímos conhecimento de um estudo que proponha um sistema não-sequencial, de modo a utilizar as classificações em cada dimensão do espaço de emoções de forma equivalente ou simultânea. Tal sistema poderia ter um desempenho ainda superior, uma vez que no sistema sequencial a classificação pode falhar logo nos primeiros estágios.

2.2.3 O processo de classificação

A Figura 7 descreve as duas etapas do processo de reconhecimento que pode ser entendido da mesma forma para os problemas de identificação e de verificação, assim como para o classificador de três estágios. Na etapa de treinamento, diversas amostras referentes

a uma mesma emoção são analisadas, sendo que estas podem pertencer a diversos locutores diferentes quando em um cenário independente de locutor. Tanto no treinamento quanto na classificação, uma amostra de sinal de voz passa primeiramente por um detector de atividade vocal (Voice Activity Detector - VAD), onde são selecionados os trechos em que o detector indica a presença de voz humana. Esta fragmentação do sinal é realizada para evitar tanto o uso de parâmetros de voz incorretos quanto o favorecimento de sinais com longos períodos de silêncio ou ruído. Destes trechos, são extraídos os parâmetros de voz relevantes para o processo e estes são usados para treinar o modelo da emoção (ou do conjunto de emoções). Um processo muito semelhante é realizado na fase de classificação, porém após os parâmetros de voz serem extraídos, o conjunto de parâmetros é confrontado com cada um dos modelos existentes e a classificação se dá pelo modelo que apresenta a melhor coerência.

Figura 7 – Diagrama de blocos do processo de classificação

Fonte: Autoria própria

Uma estratégia comum para gerar um cenário independente de locutor é a técnica de validação cruzada “deixe-um-locutor-fora” (em inglês “leave-one-speaker-out”). Esta estratégia consiste em criar todo o conjunto de modelos das emoções uma vez para cada um dos locutores, utilizando apenas as amostras que não pertencem a um dos locutores. Portanto se existem E modelos de emoções, estes E modelos são criados para cada um dos locutores e são diferentes. A fase de testes conta apenas com as amostras do locutor alvo e são confrontadas com os modelos criados para os seus testes. Em outras palavras, cada locutor tem a sua vez de ser deixado de fora do treinamento para que suas amostras não sejam utilizadas nenhuma vez durante o treinamento, de modo a garantir que os testes sejam independentes de locutor. Desta maneira, todas as amostras da base de dados são

utilizadas para teste, sendo que elas não participam do treinamento dos modelos utilizados para os seus testes.

3 Técnicas

3.1 Parâmetros de voz

Neste capítulo discutimos os parâmetros de voz, que são essenciais para o reconhecimento automático de emoções já que são eles os responsáveis por modelar o efeito das emoções na voz humana. Como já mencionado anteriormente, as emoções devem ser modeladas e reconhecidas pelas alterações fisiológicas da voz humana, quando um indivíduo vivencia um determinado estado emocional. Estas alterações podem envolver tanto a prosódia - o estudo do ritmo, entonação e demais atributos correlatos na fala - quanto a qualidade da voz, relacionada à inteligibilidade e à naturalidade da fala. Alterações na prosódia podem ser exemplificadas pela diferença entre as frases “Hoje acordei bem.” e “Hoje acordei bem!”, sendo que na primeira a pessoa fala com um tom mais neutro, enquanto na segunda a mesma demonstra uma entonação mais incisiva, demonstrando um nível mais alto de excitação. Embora ainda não exista um consenso de quais são exatamente os parâmetros que melhor caracterizam cada estado emocional, vários parâmetros são frequentemente encontrados em sistemas estado da arte, sendo muitos deles parâmetros utilizados comumente em outros sistemas envolvendo processamento de voz. Entre os principais parâmetros prosódicos estão a frequência fundamental (pitch), a energia e propriedades relacionadas à duração da voz e pausas, enquanto entre os parâmetros que descrevem a qualidade da voz encontram-se as frequências formantes, a distribuição espectral, representada por MFCC’s (Mel Frequency Cepstral Coefficients) ou LFPC’s (Log Frequency Power Coefficients), a relação harmônicos/ruído (Harmonic to Noise Ratio ou HNR) e o fluxo glotal.

Um estudo famoso realizado por Murray e Arnott (39) identifica alguns dos parâmetros prosódicos e qualitativos extensivamente estudados por pesquisas na área da psicologia. O resumo deste estudo pode ser visualizado na tabela 3, que associa as emoções básicas Alegria, Raiva, Tristeza, Medo e Nojo às suas respectivas alterações nos parâmetros da voz em comparação com o estado neutro.

Embora na prática estas alterações não ocorram como representado na tabela em 100% dos casos, fica claro que os parâmetros indicados têm influência nas emoções e a tabela fornece uma boa base para investigá-los. Discutimos individualmente alguns dos principais parâmetros mais adiante neste capítulo.

Tabela 3 – Resumo de alterações nos parâmetros da voz em diferentes emoções segundo Murray e Arnott

Emoção	Pitch	Intensidade	Velocidade de Fala	Qualidade da voz
Raiva	maior média, maior extensão, mudanças abruptas	mais alta	ligeiramente mais rápida	ofegante
Alegria	maior média, maior extensão	mais alta	mais rápida ou mais lenta	ofegante, estridente
Tristeza	menor média, menor extensão,	mais baixa	mais lenta	ressonante
Medo	maior média, maior extensão	normal	mais rápida	irregular
Nojo	menor média, maior extensão	mais baixa	mais lenta	resmungante

3.1.1 Extração de parâmetros

Até o momento tratamos um parâmetro de voz como algo um tanto abstrato, que simplesmente representa alguma característica da voz quando em um estado emocional. Entremos agora em uma definição mais precisa e matemática do que vem a ser um parâmetro de voz.

Observemos a Figura 8 que é um “close” da Figura 7 do capítulo 2. O sinal de áudio $s(n)$ é um sinal digital amostrado com uma determinada frequência de amostragem f_a contendo a voz de um dado locutor. Cada valor $s(n_1), s(n_2), s(n_3)...$ etc, representa uma amostra do sinal digital. Este sinal passa por um processo de segmentação, que resulta em diversas janelas $\mathbf{x}_j(n)|n = 1, 2, 3...N$, sendo a duração da janela da ordem de dezenas de milissegundos. N denota o comprimento da janela em amostras e j a j -ésima janela. As amostras que passam pelo teste de atividade vocal são então encaminhadas para os algoritmos de extração de parâmetros. Cada algoritmo de extração específico para um dado parâmetro analisa e processa o sinal contido na janela e gera então uma quantia $P(j, l)$, que definiremos agora como o l -ésimo parâmetro de curto prazo da janela j . Juntando todos os l -ésimos parâmetros de curto prazo gerados para cada janela, temos então o contorno do l -ésimo parâmetro, $\mathbf{P}(j, l)$, que representa a evolução do parâmetro ao longo do tempo, sendo que a definição em negrito denota uma matriz. A definição final do que vem a ser um parâmetro de voz é dada na próxima seção.

Figura 8 – Extração de parâmetros

Fonte: Autoria própria

3.1.2 Tipos de parâmetros

A partir daqui definiremos dois tipos de parâmetros: os parâmetros dinâmicos e os parâmetros globais, sendo que por muitas vezes chamaremos ambos apenas de parâmetros, uma vez que o contexto elimina a ambiguidade. Continuaremos a utilizar a nomenclatura abstrata “parâmetro” para falar que pitch, energia, formantes, etc são parâmetros, porém tenhamos em mente que a definição precisa de parâmetro é um dado numérico a ser utilizado como entrada para os algoritmos de classificação. Os parâmetros dinâmicos são apenas o próprio contorno dos parâmetros de curto prazo, de forma que unindo todos parâmetros temos uma matriz de parâmetros $\mathbf{P}_u(j)$ para cada instância de áudio u . A entrada final para os algoritmos de treinamento e de classificação, $\mathbf{P}(j)$, pode ser obtida concatenando-se a matrizes de parâmetros de cada instância de áudio do modelo:

$$\mathbf{P}(j) = [\mathbf{P}_1(j), \mathbf{P}_2(j), \dots, \mathbf{P}_U(j)] \quad (3.1)$$

Normalmente omitiremos o índice da instância de áudio u pois esta informação em geral não é importante, assim como por vezes omitiremos o índice da janela j .

Já os parâmetros globais são geralmente calculados a partir de propriedades estatísticas dos contornos de cada parâmetro, como média, desvio padrão, mediana, máximo, mínimo ou extensão de um determinado contorno de um parâmetro. O resultado é um vetor de parâmetros $\mathbf{p}_u$ para cada instância de áudio u , sendo que a matriz final $\mathbf{P}$, que é a entrada para os algoritmos de treinamento e classificação é dada por:

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_U] \quad (3.2)$$

O tipo de parâmetro escolhido depende diretamente do algoritmo de classificação utilizado. Geralmente abordagens estatísticas como GMM's ou HMM's utilizam parâmetros dinâmicos uma vez que todo o contorno pode ser utilizado para treinar as densidades de probabilidades. Por outro lado, técnicas determinísticas para reconhecimento de padrões geralmente utilizam parâmetros globais, já que cada amostra do contorno individualmente não representa um papel importante para uma classe e o tamanho do contorno pode variar

entre os arquivos de áudio.

Os parâmetros dinâmicos tem a vantagem de manter uma noção da informação temporal, principalmente quando utilizados em conjunto com as derivadas dos contornos, em especial a primeira e a segunda derivadas. A informação temporal é um aspecto importante das emoções, pois elas são fenômenos contínuos que evoluem com o tempo. A prática de utilizar-se as derivadas dos contornos tornou-se popular em estudos de reconhecimento de voz (40) porém recentemente vem sendo utilizada também para reconhecimento de emoções (3)(7).

3.1.3 Comparação entre parâmetros de voz

Como já mencionado, a questão dos parâmetros de voz já foi extensivamente estudada por diversos autores, porém ainda não existe um consenso em quais deles são os melhores ou mais representativos para o reconhecimento automático de emoções. Esta falta de consenso na literatura está relacionada ao uso de diferentes bases de dados, diferentes conjuntos de emoções e diferentes métodos de classificação em cada estudo. Na maioria dos casos, não é possível apontar se um estudo é melhor ou pior que o outro, apenas conclui-se que a tarefa de reconhecimento automático de emoções é extremamente complexa e depende de diversos fatores. De qualquer modo, alguns resultados interessantes foram encontrados e são citados nesta seção.

Em (41), os autores comparam o desempenho de coeficientes cepstrais de predição linear (Linear Prediction Cepstral Coefficients - LPCC), coeficientes cepstrais na escala mel (Mel Frequency Cepstral Coefficients - MFCC) - ambos populares em sistemas para reconhecimento de voz - e coeficientes de potência em escala logarítmica (Log Frequency Power Coefficients - LFPC) para o reconhecimento de seis emoções utilizando HMM's. Como resultado, os autores apontam taxas de reconhecimento de 77%, 56% e 51% para LFPC's, MFCC's e LPCC's respectivamente, indicando uma alta superioridade dos LFPC's. Possíveis causas da superioridade dos LFPC's são a alta dependência de texto e a filtragem do pitch durante a análise cepstral (42), de modo que os LFPC's preservam melhor a informação do pitch nos filtros de ordem baixa. A idéia da filtragem de ordem baixa é enfatizada por Oudeyer em (43). Oudeyer utiliza pitch, energia filtrada em filtros passa-baixas e passa-altas e MFCC's, obtendo parâmetros globais como média, máximo, mínimo, extensão e percentis, totalizando um conjunto de 200 parâmetros. Ele termina por concluir que parâmetros tradicionalmente considerados importantes como o pitch e a energia geral levam desvantagem em relação aos parâmetros relacionados à energia de baixa frequência, o que sugere que a energia do pitch tem mais importância do que os valores do pitch em si. Outros autores como Fernandez (44) e Schuller (45) também destacam MFCC's e a qualidade de voz em detrimento do pitch. Por outro lado, Lin (7) e Lugger(46) realizam experimentos para seleção de parâmetros e destacam o pitch como um parâmetro decisivo para o reconhecimento de emoções. No caso de (7), Lin verifica que não exatamente o

pitch mas as suas derivadas possuem um papel muito importante, de forma que os valores em si são menos importantes do que as alterações instantâneas. Isto bate com a intuição e é condizente com a maior melodização da voz para determinadas emoções, principalmente as de alta excitação como alegria e raiva, enquanto as emoções tédio ou neutro tendem a ser monotônicas.

Deste modo, verificamos a existência de algumas frentes que colocam um parâmetro na frente de outro e vice-versa, porém os resultados são altamente dependentes de cada estudo.

3.1.4 Parâmetros de voz populares

Nesta seção, introduzimos e definimos alguns parâmetros de voz populares em sistemas estado da arte e selecionados para este estudo. Detalhes sobre a implementação dos algoritmos para extração dos parâmetros são dados no capítulo 4.

3.1.4.1 Pitch

O termo pitch não possui uma tradução exata para o português porém pode ser entendido como a altura (frequência) percebida num sinal de áudio. Na maioria das vezes ele se iguala a frequência fundamental conhecida por f_0 , porém estes dois podem ser diferentes, uma vez que a frequência fundamental é uma propriedade acústica da onda sonora, enquanto o pitch é subjetivo. A subjetividade remete à área da psicoacústica e ao caráter não-linear do ouvido humano. Uma forma fácil de entender o pitch é pensar nas frequências emitidas ao cantarmos uma música. Embora existam várias frequências na voz humana, é o pitch que forma a melodia.

Intuitivamente, o pitch parece ter grande contribuição para a expressão das emoções e consequentemente para o reconhecimento. O estudo do pitch vem desde a psicologia, e está normalmente associado à dimensão da ativação, sendo que um pitch alto com muitas variações reflete emoções de alta excitação. Entretanto, estudos recentes afirmam que o papel do pitch para o reconhecimento de emoções é sombreado por outros parâmetros de voz.

O cálculo do pitch pode ser feito no domínio do tempo ou da frequência por diversas técnicas utilizando cruzamentos de zeros, autocorrelação ou coeficientes cepstrais. Um algoritmo popular é o algoritmo de Boersma (47) utilizado no software Praat (48), que calcula o pitch como o máximo de autocorrelação em uma janela de algumas dezenas de milissegundos.

3.1.4.2 Energia

A intensidade sonora percebida pelo ouvido humano é um parâmetro de voz bastante intuitivo. O nosso ouvido consegue distinguir a intensidade sonora em decibéis, entretanto é

difícil medir a intensidade sonora diretamente, na forma de um parâmetro de voz. Por este motivo, o que é usado normalmente é a energia do sinal, capturada através da amplitude das amostras. Geralmente, é calculada a energia de curto prazo, medindo a média da energia de uma janela de algumas dezenas de milisegundos. Todavia, a energia difere-se da intensidade sonora uma vez que a energia soma ruídos ao sinal puro de voz, por ser uma média que pode gerar o mesmo valor tanto com poucas amostras de alto valor, quanto com muitas amostras de valor mais baixo, além do fato de que muitas vezes o sinal de áudio gravado pode ter um ganho positivo ou negativo, alterando a percepção da intensidade sonora original.

A curva de energia depende de muitos fatores como os fonemas, o locutor, a cultura do locutor, o tipo de expressão (afirmativa, interrogativa, exclamativa), como também do estado emocional do locutor. Como o pitch, altos valores de energia e de suas variações estão normalmente correlacionados com emoções de alta excitação.

3.1.4.3 Formantes

As formantes representam picos na resposta em frequência do aparelho fonador humano, isto é, podem ser entendidas como frequências de ressonância do trato vocal humano. Grande parte da característica das vogais é decorrente das formantes, principalmente as duas primeiras formantes, comumente denominadas f_1 e f_2 , o que as tornam extremamente importantes para o estudo da fonética e da articulação. Como a articulação está bastante correlacionada com o estado emocional, é de se esperar que as formantes tenham influência no reconhecimento de emoções. Entretanto, poucos estudos selecionam as formantes como parâmetros decisivos para esta tarefa.

Para o cálculo das formantes é comum utilizar-se algoritmos que modelam o trato vocal, por exemplo através de codificação preditiva linear, sendo o algoritmo de Burg uma escolha habitual, como por exemplo no software Praat (48).

3.1.4.4 MFCC

Os coeficientes mel cepstrais (Mel Frequency Cepstral Coefficients) são uma representação paramétrica do espectro de frequências do sinal de voz. A idéia dos coeficientes mel cepstrais vem inicialmente do cepstro ou cepstrum cujo nome foi dado por ser uma inversão da palavra inglesa spectrum, o espectro do sinal. O cepstro de potência (Power Cepstrum) foi definido por Bogert (49) como

$$|\mathcal{F}^{-1}[\log(|\mathcal{F}[\mathbf{x}(n)]|^2)]|^2. \quad (3.3)$$

Sabendo-se que o espectro do sinal é obtido através de uma transformada de Fourier do sinal no domínio do tempo, o cepstro é obtido através da transformada inversa de Fourier do log espectro. O cepstro possui informações a respeito da taxa de variação em diferentes

bandas do espectro, uma vez que este pode ser visto como um sinal quase periódico (para sinais harmônicos), com uma espécie de modulação em amplitude. Enquanto o logaritmo reduz a faixa dinâmica, a transformada de Fourier separa as componentes deste sinal quase periódico, em um pico em volta da “frequência” do sinal e componentes de baixa frequência devidos à modulação em amplitude.

Uma outra forma de entender a importância do cepstro é lembrando que o sinal de voz $\mathbf{s}(n)$ pode ser modelado por um pulso glotal $\mathbf{u}(n)$ convolvido com a resposta impulsiva do trato vocal $\mathbf{v}(n)$,

$$\mathbf{s}(n) = \mathbf{u}(n) * \mathbf{v}(n). \quad (3.4)$$

Muitas vezes o que buscamos é o trato vocal $\mathbf{v}(n)$ e separar os sinais no domínio do tempo é difícil porque eles se encontram convolvidos. Entretanto, o cálculo do cepstro permite transformar a convolução em uma soma de duas componentes:

A transformada de Fourier transforma o sinal no tempo em

$$\mathbf{S}(e^{jw}) = \mathbf{U}(e^{jw})\mathbf{V}(e^{jw}) \quad (3.5)$$

e aplicando o log da magnitude temos

$$\log(|\mathbf{S}(e^{jw})|) = \log(|\mathbf{U}(e^{jw})|) + \log(|\mathbf{V}(e^{jw})|). \quad (3.6)$$

Finalmente, a aplicação da transformada de Fourier inversa neste log espectro obtém

$$\mathbf{c}(n) = \mathbf{c}_u(n) + \mathbf{c}_v(n). \quad (3.7)$$

A diferença entre o mel-cepstro e o cepstro é que para os coeficientes mel-cepstrais são aplicados filtros com bandas de frequências igualmente espaçadas na escala mel, como indicado na figura 9. O uso da escala mel advém de estudos da Psicoacústica, pois a escala mel se aproxima melhor do sistema auditivo humano, que possui um comportamento não-linear na frequência, perdendo cada vez mais sensibilidade em frequências superiores a 1 kHz. Para atenuar este problema, os filtros possuem comportamento linear até 1000 Hz e após isto a largura dos filtros aumenta de forma logarítmica.

Os MFCC's são bastante populares em sistemas de reconhecimento de voz e de identificação de locutor, mas a utilidade para sistemas para reconhecimento de emoções

Figura 9 – Filtros triangulares na escala mel

Fonte: Adaptada de (42)

tem sido demonstrada em alguns estudos, embora diversos autores argumentem que outras representações como LFPC's possuem melhores resultados (41).

3.1.4.5 LFPC

Os log coeficientes de potência (Log Frequency Power Coefficients ou LFPC) são coeficientes que refletem a distribuição espectral de energia do sinal. O cálculo destes coeficientes é bem direto, bastando passar o espectro do sinal por um banco de filtros que delimitam sub-bandas e calcular a média da energia nestas subbandas. Os centros das sub-bandas podem ter distribuição logarítmica para adequação ao modelo psicoacústico da audição humana, de modo a selecionar as frequências críticas.

LFPC's têm sido utilizados em reconhecimento de emoções, pois se adequam a conclusão de Oudeyer (43) de que a energia em baixas frequências tem papel importante nas emoções. Muitas vezes são preferíveis no lugar de MFCC's, pois melhor preservam a informação de pitch.

3.1.5 Seleção de parâmetros

Em sistemas de reconhecimento com um conjunto grande de parâmetros, é comum utilizar-se do recurso de seleção de parâmetros, que se baseia principalmente em assumir que o conjunto de observações contém parâmetros redundantes ou irrelevantes. Selecionando os parâmetros mais importantes, é possível diminuir a complexidade computacional tanto no algoritmo de classificação quanto na própria extração dos parâmetros e evitar problemas de overfitting, uma vez que um conjunto grande de parâmetros tende a tornar o sistema mais dependente dos dados de entrada, diminuindo o seu desempenho em situações genéricas. Além disso, embora na teoria um número maior de parâmetros não deva diminuir a acurácia

do sistema, isto pode não ser verdadeiro na prática já que o conjunto de dados de entrada é finito e pode ser tendencioso, de modo que algumas instâncias podem definir um padrão que não representa fielmente a classe. Nestes casos, o sistema pode apresentar melhor desempenho com um número menor de parâmetros ótimos.

O problema resume-se em: Dado um conjunto de parâmetros $\mathbf{P}_0 = [\mathbf{P}(l) | l = 1, 2, 3, \dots, L]$, encontrar um subconjunto $\mathbf{P}_{opt} = [\mathbf{P}(l) | l = 1, 2, 3, \dots, L_{opt}]$, com $L_{opt} < L$ que maximiza uma função objetiva $J(\mathbf{P})$:

$$\mathbf{P}_{opt} = \arg \max_{\mathbf{P}} J(\mathbf{P}) \quad (3.8)$$

Para resolver este problema, deve-se lembrar que pode existir dependência entre os parâmetros e não é possível avaliar os parâmetros independentemente. Portanto, uma primeira possibilidade é realizar uma busca exaustiva em todas as combinações de parâmetros e selecionar o subconjunto com melhor desempenho. Entretanto, existem métodos mais eficientes para executar esta tarefa, que buscam soluções ótimas ou subótimas com uma complexidade computacional bastante inferior.

A literatura mostra a utilização de diversos algoritmos diferentes. Lin e Wei (7) utilizam o algoritmo Sequential Forward Selection (SFS) (50) que funciona em conjunto com os algoritmos de classificação, atualizando o conjunto ótimo após o resultado da classificação de cada conjunto testado. Lugger e Yang (37) utilizam da mesma estratégia, porém com algoritmo Sequential Floating Forward Selection (SFFS) (50) que é uma modificação do algoritmo SFS. Para parâmetros globais é possível pré-selecionar o conjunto ótimo através de critérios objetivos como coeficientes de correlação (6) ou a taxa de Fisher (51). Alternativamente, Tato et al. (36) utilizam parâmetros que modificam um modelo de regressão linear.

3.2 Métodos de classificação

Novamente, lembramos que o reconhecimento de emoções através da voz tem sido resolvido com um problema estatístico ou de reconhecimento de padrões, onde um modelo é gerado através de instâncias cujas emoções são conhecidas e posteriormente uma instância cuja emoção é desconhecida é confrontada com os modelos existentes e o modelo mais adequado é escolhido. Nos modelos estatísticos, geralmente a emoção é modelada por uma densidade de probabilidade e a classificação se dá pela máxima verossimilhança. Na abordagem determinística, a classificação se dá otimizando alguma medida objetiva que compara os parâmetros de voz da instância com os parâmetros de voz do modelo. O tipo de parâmetros de voz, global ou dinâmico, dá origem a duas vertentes de classificação: estática e dinâmica.

3.2.1 Classificação estática

A diferença entre métodos de classificação estáticos e dinâmicos está no tipo de parâmetros de voz usados, que são dinâmicos (o contorno dos parâmetros de curto prazo) para os dinâmicos e globais para os estáticos. Os conceitos de classificação estática e dinâmica podem ser quase que confundidos com os de classificação estatística ou determinística, pois na maioria dos casos, métodos determinísticos utilizam parâmetros globais enquanto os métodos estatísticos utilizam parâmetros de curto prazo. Porém, nada impede que isto seja alterado, inclusive em (52) e (53) são utilizados parâmetros globais para treinar misturas de gaussianas enquanto Womack e Hansen (54) utilizam parâmetros de curto prazo para alimentar redes neurais artificiais. Os métodos estáticos utilizam parâmetros globais, que em geral são estatísticas calculadas a partir do contorno de parâmetros de curto prazo, como média, desvio padrão, máximo, mínimo, mediana, percentis, além de aspectos de duração e estatísticas em picos e vales. A necessidade de parâmetros globais advém dos parâmetros de curto prazo não apresentarem individualmente uma informação útil para o modelo, além de que o contorno pode ter comprimento variável, sendo que ter um número variável de parâmetros não é aceitável para muitos métodos, que podem por exemplo gerar matrizes ou vetores de tamanho fixo, de acordo com o número de parâmetros. Uma desvantagem da classificação estática é que como são geradas estatísticas globais do contorno, a informação temporal é perdida. Exemplos de métodos de classificação estática são o K-vizinhos mais próximos (K-Nearest Neighbours ou KNN), máquinas de vetores de suporte (Support Vector Machines ou SVM) e redes neurais artificiais (Artificial Neural Networks ou ANN).

3.2.2 Classificação dinâmica

Métodos de classificação dinâmica utilizam todo o contorno dos parâmetros de curto prazo tanto para treinar os modelos quanto para a classificação em si. A classificação dinâmica é mais comum para métodos estatísticos, uma vez que toda a sequência de parâmetros de curto prazo pode ser utilizada para alimentar o algoritmo, de forma a estimar a densidade de probabilidade para a ocorrência da sequência de observações ou calcular a verossimilhança desta sequência, utilizando um modelo específico. Assim, quanto mais pontos existirem, melhor é a estimativa da densidade de probabilidades e mais correto é o cálculo da verossimilhança. Exemplos comuns de métodos de classificação dinâmica são classificadores bayesianos simples, modelos de mistura de gaussianas (Gaussian Mixture Models ou GMM) e modelos ocultos de Markov (Hidden Markov Models ou HMM). A vantagem de classificadores dinâmicos é que eles mantêm a informação temporal nas sequências de observações, principalmente quando combinados com as derivadas do contorno. Entretanto, nos modelos estatísticos é necessário assumir a forma da distribuição probabilística da observação, que é na verdade imprevisível, e a forma escolhida pode não

ser um modelo realístico.

3.2.3 Comparação entre métodos de classificação

Como no caso dos parâmetros de voz, não é possível afirmar qual é o melhor método de classificação existente, uma vez que é difícil comparar resultados na literatura já que estes dependem da base de dados utilizada, das emoções, das condições de independência, do conjunto de parâmetros, entre outros fatores. Ou seja, as variáveis são muitas e não são independentes para isolarmos o problema do algoritmo de classificação. Comparações entre algoritmos só fazem sentido dentro de um mesmo estudo, e mesmo assim não são muito conclusivas devido às peculiaridades de cada estudo. Novamente, existem algumas frentes que recomendam um método em detrimento de outro e vice versa, sendo os algoritmos mais populares HMM, GMM e SVM. Oudeyer (43) e Schuller (45) mostram SVM's com desempenho superior a outros métodos como KNN e classificadores bayesianos. Comparações entre HMM's e SVM's foram realizadas por Fernandez e Picard (55), destacando a superioridade dos HMM's, e por Lin e Wei (7), mostrando o melhor desempenho dos HMM's sobre SVM's, GMM's e KNN's. Em (56), HMM's apresentam melhor desempenho em instâncias de curta duração enquanto SVM's são melhores em instâncias de tamanho variável. Já Fu (57), desenvolveu um sistema híbrido utilizando HMM's conjuntamente com SVM's que obteve melhores resultados do que os sistemas separados. Por outro lado, (52) e (53) mostram a superioridade de GMM's utilizando parâmetros globais em relação a HMM's utilizando parâmetros dinâmicos. Infelizmente a única conclusão sólida que pode ser inferida destes estudos é que todas estas técnicas apresentam alto potencial para serem utilizadas em sistemas para reconhecimento de emoções através da voz.

3.3 Métodos de classificação populares

Nesta seção definimos alguns dos métodos de classificação populares em sistemas estado da arte e selecionados para este estudo. A implementação e os algoritmos utilizados são descritos no capítulo 4.

3.3.1 Gaussian Mixture Models

Gaussian Mixture Models (GMM ou Modelos de Misturas de Gaussianas) são um tipo particular de modelos de misturas, cuja importância é indiscutível para a área de processamento de voz e principalmente para os temas de reconhecimento de voz e de locutor. Entretanto, este método vem se mostrando promissor também para o estudo de reconhecimento de emoções, tendo sido utilizado em diversos estudos (52)(53)(7)(28). Modelos de misturas são modelos estatísticos que assumem que uma população de dados pertence a uma distribuição em forma de mistura, ou seja, uma distribuição formada

por combinações lineares de diversas funções densidade de probabilidade. Os modelos de misturas mais conhecidos são os Modelos de Misturas Gaussianas (Gaussian Mixture Models ou GMM) que representam uma distribuição como a soma ponderada de diversas gaussianas de diferentes médias e matrizes de covariância:

$$\text{Prob}[\mathbf{X}|\lambda] = \sum_{m=1}^M c_m \mathcal{N}(\mathbf{X}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (3.9)$$

$$\mathcal{N}(\mathbf{X}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{\sqrt{2\pi^L |\boldsymbol{\Sigma}_m|}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{X} - \boldsymbol{\mu}_m)\right), \quad (3.10)$$

em que L é a dimensão de cada observação.

Deste modo, um modelo de misturas de gaussianas é definido pelos seguintes parâmetros:

- c_m : O peso ou probabilidade a priori de cada gaussiana m
- $\boldsymbol{\mu}_m$: A média de cada gaussiana m
- $\boldsymbol{\Sigma}_m$: A matriz de covariância de cada gaussiana m

Juntos, estes parâmetros representam o modelo $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = [\mathbf{c}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m], \quad 1 \leq m \leq M \quad (3.11)$$

GMM's são largamente utilizados para estimar funções densidade de probabilidade desconhecidas, onde a tarefa de estimação torna-se a de estimar os pesos de cada gaussiana e suas respectivas médias e matrizes de covariância a partir dos dados observados. Podem também ser utilizados como mecanismo de classificação bayesiana, onde é escolhido o modelo que apresenta maior verossimilhança em relação à sequência de observações. Para os GMM's, podem ser utilizados tanto parâmetros globais como parâmetros de curto prazo, sendo o segundo preferível na maioria dos casos, já que o modelo pode ser treinado por uma quantidade maior de dados.

Muitas vezes estes modelos podem ser considerados como um caso especial de HMM com apenas um estado e o algoritmo de Baum-Welch (58) pode então ser utilizado para treinar as gaussianas. Entretanto é mais comum a utilização do algoritmo tradicional Expectation-Maximization (EM) (59), o qual discutimos no capítulo 4.

Embora o algoritmo EM seja sempre convergente, não é garantido que ele convirja para o máximo global (59). Normalmente ele converge para um máximo local, que pode não ser muito significativo. Portanto, a inicialização dos parâmetros iniciais tem grande impacto

no resultado final do modelo e a escolha de um bom algoritmo de agrupamento como o K-Means (60) ou o LBG (61) é muito importante para o processo. Uma outra solução para este problema é a utilização de algoritmos que pós-processam o resultado final do EM tradicional ou permitem que o algoritmo seja utilizado recursivamente, de modo a tornar o algoritmo mais robusto. Como exemplo, temos o algoritmo Split and Merge (62).

3.3.2 Hidden Markov Models

Hidden Markov Models (HMM ou Modelos Ocultos de Markov) apresentam um grande potencial para modelar as emoções humanas, uma vez que possuem o poder de manter informações temporais da sequência de observações através do uso de estados. Eles têm sido objeto de escolha em diversos estudos (3)(7)(57)(52),(55)(7)(56), inclusive apontados como o melhor método de classificação em alguns casos (55)(7)(56).

HMM's são modelos estatísticos, associados ao Processo de Markov e às Cadeias de Markov. Neles, a ocorrência de uma sequência de eventos observáveis é decorrente de os eventos terem percorrido uma sequência de estados, sendo que cada estado emite um símbolo ou conjunto de símbolos observáveis, segundo alguma distribuição probabilística.

O Processo de Markov é um processo estocástico controlado por uma variável aleatória que representa o estado do sistema, cujo valor futuro depende dos estados passados. No processo de Markov de primeira ordem, o valor futuro depende apenas do estado atual, não da sequência de estados que o procederam, propriedade conhecida como Propriedade de Markov. Isto é, dado um conjunto de estados $\mathbf{e} = [e_1, e_2, e_3, \dots, e_Q]$ e que o estado atual é e_q , a probabilidade de a variável aleatória $z(t)$ assumir o estado e_r no instante t é dada por:

$$\text{Prob}[z(t) = e_r | z(t-1), z(t-2), z(t-3) \dots] = \text{Prob}[z(t) = e_r | z(t-1) = e_q] \quad (3.12)$$

Neste cenário, cada transição de estados de e_q para e_r pode ser colocada como um elemento a_{qr}

$$a_{qr} = \text{Prob}[z(t) = e_r | z(t-1) = e_q], \quad (3.13)$$

em uma matriz de transição de estados

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1Q} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2Q} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3Q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{Q1} & a_{Q2} & a_{Q3} & \dots & a_{QQ} \end{bmatrix}, \quad (3.14)$$

tal que

$$\sum_{r=1}^Q a_{qr} = 1, \quad (3.15)$$

$$a_{qr} \geq 0. \quad (3.16)$$

Os processos com número finito de estados são denominados Cadeias de Markov e são nosso objeto de estudos.

Além dos estados, os processos podem emitir um símbolo ou conjunto de símbolos observáveis em cada estado (no nosso caso, o conjunto de parâmetros) com certa distribuição de probabilidade. É comum a utilização de misturas de gaussianas, onde a distribuição de probabilidades da ocorrência de um símbolo $\mathbf{x}(t)$ em um determinado estado $z(t)$ é dada por:

$$b[\mathbf{x}(t)] = \sum_{m=1}^M c_m \cdot \frac{1}{\sqrt{2\pi^L |\boldsymbol{\Sigma}_m|}} \exp \left(-\frac{1}{2} (\mathbf{x}(t) - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1} (\mathbf{x}(t) - \boldsymbol{\mu}_m) \right), \quad (3.17)$$

em que L é a dimensão de cada observação.

Os processos em que apenas a sequência de observações é conhecida, mas não a sequência de estados (que são “ocultos”) são chamados de Modelos Ocultos de Markov (HMM).

Sendo assim, um HMM é definido pelo seguinte conjunto de parâmetros:

- Q : O número de estados do modelo. Lembrando que os estados são ocultos, muitas vezes o número de estados também é desconhecido, sendo que tanto o significado dos estados quanto o valor de Q devem ser estimados pelas características do processo.
- L : O número de símbolos observáveis distintos em cada estado. Os símbolos referem-se às saídas físicas do processo, no nosso caso, os parâmetros de voz.

- $\mathbf{A}$: A matriz de transição de estados.
- $\boldsymbol{\pi}$: O vetor de estados iniciais. Define as probabilidades para cada um dos estados do processo no tempo $t = 0$.
- $b_q(\mathbf{x}(t))$: A distribuição de probabilidades em cada estado. Este parâmetro na verdade contempla quatro itens: o número de gaussianas M , a distribuição a priori das gaussianas c , a média $\boldsymbol{\mu}_m$ e a matriz de covariâncias $\boldsymbol{\Sigma}_m$ de cada gaussiana.

Consideremos este conjunto de parâmetros que representa o modelo como $\boldsymbol{\lambda}$.

Neste contexto, problemas clássicos surgem:

1. Dada uma sequência de observações, como calcular eficientemente a probabilidade de a observação ter sido gerada pelo modelo?
2. Dada uma sequência de observações, qual é a sequência de estados percorrida?
3. Dada uma sequência de observações, qual o modelo que melhor a descreve? Em outras palavras, quais são os valores dos parâmetros do modelo que devem ser utilizados para maximizar a probabilidade da observação ter sido gerada pelo mesmo?

Apenas os problemas 1 e 3 são necessários para o reconhecimento de emoções e suas soluções são descritas no capítulo 4.

3.3.3 K-Nearest Neighbours

O KNN (K-Nearest Neighbours ou K-Vizinhos Mais Próximos) é um algoritmo não-paramétrico utilizado em reconhecimento de padrões com parâmetros globais. É um dos algoritmos mais simples utilizados em aprendizagem de máquina, no qual não são gerados modelos propriamente ditos. Em vez disso, a fase de treinamento consiste simplesmente em armazenar o vetor de parâmetros das amostras de treinamento e suas conhecidas classes. Uma amostra de teste cuja classe é desconhecida é classificada de acordo com sua proximidade às amostras de treinamento, os seus chamados vizinhos. Para avaliar esta proximidade podem ser definidos vários critérios, como por exemplo a distância euclidiana entre a amostra e os vizinhos. Os K vizinhos mais próximos, por exemplo aqueles com menor distância euclidiana, definem a classificação da amostra, sendo a classe escolhida aquela que se repete mais vezes dentre os vizinhos mais próximos. Em caso de empates, podem ser escolhidos mais vizinhos até o desempate. O valor K pode ser escolhido arbitrariamente e tipicamente são escolhidos valores acima de 10. Em geral, valores altos de K são mais precisos pois diminuem o ruído geral, porém existe um compromisso ao se aumentar demais o valor de K pois as fronteiras entre as classes no espaço de parâmetros

tornam-se menos definidas (63).

O KNN já foi utilizado algumas vezes na literatura de reconhecimento de emoções pela voz (7), (64) (45). Apesar de ser uma técnica muito simples, foram encontrados resultados na faixa do reconhecimento humano, acima de 60%, sendo que Lin(7) apresentou uma taxa de reconhecimento de cerca de 83% para quatro emoções.

Entretanto, o KNN é um algoritmo muito simples e está longe de ser a melhor solução para um problema de reconhecimento de padrões. Além de a escolha do valor ótimo K ser complicada, este algoritmo é altamente prejudicado pela existência de parâmetros irrelevantes ou ruidosos, sendo necessário um grande esforço para escalá-los apropriadamente.

3.3.4 Support Vector Machines

Support Vector Machines (SVM ou Máquinas de Vetores de Suporte) é uma técnica utilizada em reconhecimento de padrões para decisões binárias, que tem sido largamente utilizada para o reconhecimento de emoções através da voz (7)(65)(53)(57)(43)(45) (55) (56). A origem das SVM está fortemente ligada à teoria de minimização de riscos e à dimensão de Vapnik Chervonenkis (VC) (66), que mede a capacidade da máquina de aprendizagem, isto é, sua habilidade de lidar com amostras que não foram usadas para o treinamento. Uma das vantagens das SVM é não sofrer da Maldição da Dimensionalidade (66)(67).

As SVM fornecem uma solução ótima para o problema de separação de duas classes linearmente separáveis. Como podemos observar na figura 10, podemos pensar em diversas retas que separam as amostras para o caso de até duas dimensões, ou hiperplanos para dimensões maiores.

Figura 10 – Separação linear de amostras de duas classes

Fonte: Adaptada de (67)

A escolha do hiperplano ótimo depende das amostras, onde alguns pontos podem ser mais importantes que outros. Neste contexto, são definidos os Vetores de Suporte que nada mais são do que as amostras mais próximas à fronteira de decisão que alterariam o hiperplano ótimo caso fossem removidas. O algoritmo de treinamento das SVM procura pelo hiperplano que separa as classes com a maior margem, definida pela distância entre as amostras de cada classe mais próximas ao hiperplano, conforme mostrado na figura 11.

Figura 11 – Separação por hiperplanos com máxima margem

Fonte: Adaptada de (67)

Uma desvantagem das SVM's é que elas são ideais para resolver decisões binárias entre duas classes, porém em um problema multiclases a solução exige diversas decisões binárias e um mecanismo de decisão final.

Parte III

Parte Experimental

4 Implementação

Até o momento, descrevemos conceitos básicos associados ao estudo de reconhecimento de emoções através da voz e algumas convenções e resultados encontrados na literatura. Neste capítulo, introduzimos o que foi de fato desenvolvido neste trabalho, começando por apresentar algumas escolhas e definições a respeito da base de dados utilizada, os parâmetros e os métodos de classificação selecionados, além da descrição dos algoritmos envolvidos em cada processo. Terminamos por dar detalhes de todos modelos e estratégias de classificação adotados e apresentamos os testes executados.

4.1 Berlin Emotional Database of Speech

A base de dados Berlin Emotional Database of Speech (EMO-DB) (20) foi escolhida para este estudo. Apesar de ser uma base de dados composta por emoções atuadas, neste momento optamos por uma base estável e popular, que já foi utilizada em diversos estudos (37)(38)(64) (6)(68)(45)(65). Uma vez que a comparação entre resultados na literatura é difícil já que existem muitas variáveis envolvidas, a utilização de uma base conhecida e bastante estudada facilita a comparação de resultados, pois elimina algumas destas variáveis, como as emoções escolhidas, o número de emoções, além das próprias amostras de áudio. Por este motivo, comparações de resultados de nosso estudo são realizadas somente com outros estudos que utilizam a base EMO-DB.

A base de dados EMO-DB é composta de sete emoções: Felicidade (F), Medo (M), Neutro (Ne) , Nojo (No), Raiva (R), Tédio (Te) e Tristeza (Tr). 10 locutores, sendo 5 do sexo masculino e 5 do sexo feminino gravaram um total de 535 amostras de áudio contendo frases de curta duração, de menos de cinco segundos. Dez frases diferentes foram gravadas em todas as emoções, sendo o conteúdo das frases emocionalmente neutro. A tabela 4 apresenta as frases originais em alemão e a tradução para português.

A tabela 5 mostra a distribuição de amostras de áudio com os rótulos dos locutores dados da mesma maneira que eles são identificados na base EMO-DB (ex: Locutor03). Podemos ver que os dados são desbalanceados, no sentido que algumas emoções possuem mais amostras que outras e que dentro de uma mesma emoção alguns locutores possuem mais amostras que outros. Dentro de uma mesma emoção, o número de amostras varia de 0 a 16. Para todos os locutores, tanto homens quanto mulheres, a emoção Raiva possui mais amostras que as demais emoções e no caso masculino, a emoção Nojo possui apenas uma ou duas amostras para a maioria dos locutores. Esta disparidade pode ter grande impacto nas taxas de reconhecimento das diferentes emoções, pois os algoritmos de treinamento podem acabar por favorecer emoções com mais amostras, principalmente na criação de

Tabela 4 – Frases da base EMO-DB

Frase	Alemão	Português
1	Der Lappen liegt auf dem Eisschrank.	A toalha de mesa está sobre a geladeira.
2	Das will sie am Mittwoch abgeben.	Ela vai entregá-lo na quarta feira.
3	Heute abend könnte ich es ihm sagen.	Hoje a noite eu poderia contar a ele.
4	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	A folha de papel preta está localizada lá em cima, ao lado do pedaço de madeira .
5	In sieben Stunden wird es soweit sein.	Em sete horas será.
6	Was sind denn das für Tüten, die da unter dem Tisch stehen?	E quanto às sacolas que estão embaixo da mesa?
7	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	Eles acabaram de carregá-lo para o andar de cima e agora eles estão descendo novamente.
8	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Atualmente, nos finais de semana eu sempre ia para casa e via a Agnes.
9	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	Eu simplesmente irei descartar isto e então irei tomar uma bebida com o Karl.
10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	Estará no lugar onde nós sempre o guardamos.

modelos que combinam dados de várias emoções.

Outros motivos para a escolha da base EMO-DB foram o fato de ela possuir um número razoável de emoções e de locutores e que as emoções selecionadas são consideradas como básicas em diversos estudos e estão bem distribuídas no espaço de emoções, segundo o modelo dimensional. Além disso as amostras foram cuidadosamente escolhidas, uma vez que a base passou por testes humanos imediatamente após o término das gravações. 20 indivíduos participaram, onde as frases com taxas de reconhecimento de emoções abaixo de 80% ou naturalidade abaixo de 60% foram eliminadas. O conjunto original possuía 800 amostras, pelo menos uma para cada tripla de locutor, emoção e frase, com algumas duplicatas.

4.2 Parâmetros de voz

4.2.1 Parâmetros Selecionados

Selecionamos para este estudo um conjunto de 96 parâmetros de curto prazo, incluindo Pitch, Energia, as 5 primeiras Formantes, os 13 primeiros MFCC e os 12 primeiros

Tabela 5 – Número de amostras de áudio por locutor e por emoção na base EMO-DB.

	Locutor03	Locutor10	Locutor11	Locutor12	Locutor15
F	7	4	8	2	6
M	4	8	10	6	8
Ne	11	4	9	4	11
No	1	1	2	2	5
R	14	10	11	12	13
Te	5	8	8	5	9
Tr	7	3	7	4	4

(a) Locutores de Sexo Masculino

	Locutor08	Locutor09	Locutor13	Locutor14	Locutor16
F	11	4	10	8	11
M	6	1	7	12	7
Ne	10	9	9	7	5
No	0	8	8	8	11
R	12	13	12	16	14
Te	10	4	10	8	14
Tr	9	4	5	10	9

(b) Locutores de Sexo Feminino

LFPC, além das primeiras e segundas derivadas de cada parâmetro, que neste estudo são consideradas como parâmetros independentes do contorno. A lista completa de parâmetros pode ser visualizada na tabela 6.

4.2.2 Extração de parâmetros

Todos os parâmetros são extraídos em janelas de 40 ms, devido à necessidade de a janela ser duas a três vezes maior que o período de pitch máximo para extração do pitch (47). As janelas possuem sobreposição de 75%, sendo calculadas em passos de 10 ms. Não utilizamos um algoritmo para detecção de atividade vocal (VAD), uma vez que nos aproveitamos da decisão de janelas vozeadas e não-vozeadas realizada pelo algoritmo estimador de pitch. Desta forma, o pitch é sempre o primeiro parâmetro extraído e as janelas que possuem pitch nulo são eliminadas.

Além disso, aplicamos a subtração de média

$$p(j) \leftarrow p(j) - \mu_p, \quad (4.1)$$

Tabela 6 – Índice de Parâmetros

Índice	Parâmetro
1	Pitch
2	Pitch - 1ª Derivada
3	Pitch - 2ª Derivada
4	LogEnergy
5	LogEnergy - 1ª Derivada
6	LogEnergy - 2ª Derivada
7	1º LFPC
8	2º LFPC
9	3º LFPC
10	4º LFPC
11	5º LFPC
12	6º LFPC
13	7º LFPC
14	8º LFPC
15	9º LFPC
16	10º LFPC
17	11º LFPC
18	12º LFPC
19	1º LFPC - 1ª Derivada
20	2º LFPC - 1ª Derivada
21	3º LFPC - 1ª Derivada
22	4º LFPC - 1ª Derivada
23	5º LFPC - 1ª Derivada
24	6º LFPC - 1ª Derivada
25	7º LFPC - 1ª Derivada
26	8º LFPC - 1ª Derivada
27	9º LFPC - 1ª Derivada
28	10º LFPC - 1ª Derivada
29	11º LFPC - 1ª Derivada
30	12º LFPC - 1ª Derivada
31	1º LFPC - 2ª Derivada
32	2º LFPC - 2ª Derivada
33	3º LFPC - 2ª Derivada
34	4º LFPC - 2ª Derivada
35	5º LFPC - 2ª Derivada
36	6º LFPC - 2ª Derivada
37	7º LFPC - 2ª Derivada
38	8º LFPC - 2ª Derivada
39	9º LFPC - 2ª Derivada
40	10º LFPC - 2ª Derivada
41	11º LFPC - 2ª Derivada
42	12º LFPC - 2ª Derivada
43	1º Formante
44	2º Formante
45	3º Formante
46	4º Formante
47	5º Formante
48	1º Formante - 1ª Derivada
49	2º Formante - 1ª Derivada
50	3º Formante - 1ª Derivada

(a) Parâmetros de 1 a 50

Índice	Parâmetro
51	4º Formante - 1ª Derivada
52	5º Formante - 1ª Derivada
53	1º Formante - 2ª Derivada
54	2º Formante - 2ª Derivada
55	3º Formante - 2ª Derivada
56	4º Formante - 2ª Derivada
57	5º Formante - 2ª Derivada
58	1º MFCC
59	2º MFCC
60	3º MFCC
61	4º MFCC
62	5º MFCC
63	6º MFCC
64	7º MFCC
65	8º MFCC
66	9º MFCC
67	10º MFCC
68	11º MFCC
69	12º MFCC
70	13º MFCC
71	1º MFCC - 1ª Derivada
72	2º MFCC - 1ª Derivada
73	3º MFCC - 1ª Derivada
74	4º MFCC - 1ª Derivada
75	5º MFCC - 1ª Derivada
76	6º MFCC - 1ª Derivada
77	7º MFCC - 1ª Derivada
78	8º MFCC - 1ª Derivada
79	9º MFCC - 1ª Derivada
80	10º MFCC - 1ª Derivada
81	11º MFCC - 1ª Derivada
82	12º MFCC - 1ª Derivada
83	13º MFCC - 1ª Derivada
84	1º MFCC - 2ª Derivada
85	2º MFCC - 2ª Derivada
86	3º MFCC - 2ª Derivada
87	4º MFCC - 2ª Derivada
88	5º MFCC - 2ª Derivada
89	6º MFCC - 2ª Derivada
90	7º MFCC - 2ª Derivada
91	8º MFCC - 2ª Derivada
92	9º MFCC - 2ª Derivada
93	10º MFCC - 2ª Derivada
94	11º MFCC - 2ª Derivada
95	12º MFCC - 2ª Derivada
96	13º MFCC - 2ª Derivada

(b) Parâmetros de 51 a 96

$$\mu_p = \frac{1}{J} \sum_{j=1}^J p(j) \quad (4.2)$$

para todos os parâmetros p , para normalização de todas as instâncias de áudio e para garantir a independência de locutor.

O conjunto de 96 parâmetros nunca é utilizado por completo, sendo nos primeiros testes selecionados apenas o pitch e a energia e nos demais, executada a seleção de parâmetros.

4.2.3 Algoritmos para extração de parâmetros

Nesta seção descrevemos os algoritmos para extração de parâmetros de voz. Lembremos que todos os parâmetros são extraídos de janelas $\mathbf{x}_j(n)$ com comprimento J , das quais optaremos por omitir o índice da janela j para maior clareza do texto.

4.2.3.1 Pitch

Para a extração do Pitch foi utilizado o algoritmo desenvolvido por Paul Boersma em (47), incorporado no software Praat (48). Foi escolhido por ser um algoritmo relativamente simples, com boa eficiência e ser extensivamente utilizado por pesquisadores de diversas áreas. É um algoritmo baseado em autocorrelação, que na verdade estima o período de Pitch, obtendo o número de amostras entre os máximos de correlação do sinal, ou seja, o atraso τ . Para dar mais robustez ao algoritmo, este conta também com mecanismo para decisão de janelas vozeadas e não-vozeadas, que faz as vezes de um detector de atividade vocal (VAD) e mantém uma máquina de estados para modelar a continuidade da voz e força algumas janelas a ter pitch nulo, mesmo que a autocorrelação acuse a existência de pitch entre os valores mínimo e máximo estipulados. Além disso, como a autocorrelação pode indicar máximos inadequados, principalmente em oitavas da frequência fundamental, o algoritmo ainda conta com mecanismos para evitar a seleção destes valores, através de uma máquina de estados e funções de custo para penalizar “pulos” de oitavas entre janelas.

Algoritmo

1. Realizar um upsampling para remover o lóbulo lateral da transformada de Fourier da janela de Hann em componentes próximas à frequência de amostragem. Para isto, realizamos uma FFT de todo o sinal e filtramos linearmente até 0 as frequências de 95% a 100% da frequência de Nyquist, finalizando o processo com uma IFFT de uma ordem superior do que a FFT original.
2. Calcular a média global do sinal.
3. Para cada janela:
 - a) Obter a média local μ_x e subtraí-la de todas as amostras da janela.

b) Multiplicar $\mathbf{x}(n)$ pela função da janela de Hann $\mathbf{w}(n)$ dada pela expressão

$$w(n) = \frac{1}{2} - \frac{1}{2} \cos \frac{2\pi n}{N}, \quad 0 \leq n < N \quad (4.3)$$

para obter o sinal janelado

$$a(n) = [x(n_{mid} - \frac{1}{2}N + n) - \mu_x]w(n), \quad 0 \leq n < N \quad (4.4)$$

onde n_{mid} é a amostra correspondente ao instante central da janela (20 ms).

- c) Realizar a FFT do sinal e obter a densidade espectral de potência $|\mathbf{X}(k)|^2$.
d) Realizar a IFFT da densidade espectral de potência do sinal janelado que por definição é a função de autocorrelação do sinal no domínio do atraso $\mathbf{r}_a(\tau)$.
e) Dividir pela autocorrelação da janela

$$r_w(\tau) = \left(1 - \frac{|\tau|}{N}\right) \left(\frac{2}{3} + \frac{1}{3} \cos \frac{2\pi n}{N}\right) + \frac{1}{2\pi} \sin \frac{2\pi n}{N}, \quad (4.5)$$

que resulta na versão discreta da função de autocorrelação do sinal

$$r_x(\tau) = \frac{r_a(\tau)}{r_w(\tau)} \quad (4.6)$$

f) Calcular a versão contínua da função de autocorrelação $\mathbf{r}(\tau)$ pela expressão

$$r(\tau) = \sum_{i=1}^I r_{n_r-i} \frac{\sin \pi(\phi_l + i - 1)}{\pi(\phi_l + m - 1)} \left(\frac{1}{2} + \frac{1}{2} \cos \frac{\pi(\phi_l + i - 1)}{\phi_l + N} \right) + \sum_{i=1}^I r_{n_l+i} \frac{\sin \pi(\phi_r + i - 1)}{\pi(\phi_r + m - 1)} \left(\frac{1}{2} + \frac{1}{2} \cos \frac{\pi(\phi_r + i - 1)}{\phi_r + N} \right), \quad (4.7)$$

$$n_l = \left\lfloor \frac{\tau}{\delta\tau} \right\rfloor, \quad (4.8)$$

$$n_r = n_l + 1, \quad (4.9)$$

$$\phi_l = \frac{\tau}{\delta\tau} - n_l, \quad (4.10)$$

$$\phi_r = 1 - \phi_l, \quad (4.11)$$

$$I = \min(500, I'), \quad (4.12)$$

$$I' = \arg \max_{I''} \left[(n_l + I'') \delta \tau < \frac{N}{2} \right], \quad (4.13)$$

que é basicamente uma interpolação do tipo $\text{sen}(x)/x$ e obter os valores máximos e seus respectivos atrasos pelo algoritmo de Brent (69).

- g) Selecionar a componente sem voz (pitch nulo) e os 3 atrasos com maiores valores de autocorrelação como possíveis candidatos ao período de pitch e calcular suas relevâncias R pelas expressões

$$R = \text{VoicingThreshold} + \max \left(0.2 - \frac{(\text{localAbsolutePeak}/\text{globalAbsolutePeak})}{\text{SilenceThreshold}/(1 + \text{VoicingThreshold})} \right), \quad (4.14)$$

onde

VoicingThreshold (Limiar de voz) = 0,4,

SilenceThreshold (Limiar de silêncio) = 0,05,

OctaveCost = 0.1,

MinimumPitch = 75 Hz,

MaximumPitch = 600 Hz,

para a componente sem voz e

$$R = r(\tau_{max}) - \text{OctaveCost}^2 \cdot \log(\text{MinimumPitch} \cdot \tau_{max}), \quad (4.15)$$

para os demais. OctaveCost é um parâmetro que favorece f_0 's maiores. Como teoricamente todas as oitavas da frequência fundamental tem iguais altos picos de correlação, que podem sofrer alterações devido ao ruído aditivo, devemos escolher com cuidado a frequência correta, que a princípio é a mais baixa. Entretanto, devemos levar em conta a diferença entre a frequência fundamental acústica e o Pitch percebido. (47),

4. Utilizar o algoritmo de Viterbi (70) para selecionar a sequência de períodos de pitch com maior verossimilhança minimizando o custo dado pela função de custo (4.15).

4.2.3.2 Energia de curto prazo

Para a extração da energia de curto Prazo, é calculado o log da média da potência do sinal a cada janela, de acordo com a expressão

$$E_j = \log \frac{\sum_{n=1}^N x(n)^2}{N}, \quad (4.16)$$

onde E_j é a energia de curto prazo na janela j , $x(n)$ é a n -ésima amostra do sinal de voz dentro da janela e N é o comprimento da janela em amostras.

4.2.3.3 Formantes

Para a extração das frequências formantes, utilizamos novamente o software Praat, que conta com um algoritmo baseado em Codificação Preditiva Linear (Linear Predictive Coding - LPC), numa versão conhecida como Algoritmo de Burg (71). Por LPC, obtemos um filtro IIR All-Pole que modela o trato vocal, cujas fases dos pares de polos complexos correspondem às frequências de ressonância do sistema (72).

Algoritmo

1. Calcular os coeficientes de predição linear α_k de um filtro de ordem m por LPC.
2. Calcular os polos

$$p_k = r_k \cdot e^{j\phi_k} \quad (4.17)$$

(4.17)

,onde r_k é o módulo do k -ésimo polo e ϕ_k é a fase do k -ésimo polo do polinômio da função de transferência do filtro

$$H(z) = \frac{1}{1 - \sum_{k=1}^m \alpha_k z^{-1}} = \frac{1}{\prod_{k=1}^m (1 - p_k z^{-1})}. \quad (4.18)$$

3. Calcular as formantes F_k pela expressão aproximada

$$F_k = \frac{f_a}{2\pi} \phi_k, \quad (4.19)$$

onde f_a é a frequência de amostragem.

4.2.3.4 MFCC

Para o cálculo dos coeficientes mel cepstrais, utilizamos o algoritmo tradicional descrito por Davis e Mermel(42). Os 13 primeiros coeficientes são calculados a partir dos filtros triangulares na escala mel.

Algoritmo

1. Para cada janela, obter o sinal janelado $\mathbf{x}(n)$ com uma janela de Hamming.
2. Realizar a FFT do sinal janelado e obter a magnitude do espectro $|\mathbf{X}(k)|$.

3. Obter o log espectro do sinal

$$X_{\log}(k) = \log |X(k)|^2, \quad 0 \leq k < K. \quad (4.20)$$

4. Para os 13 primeiros filtros triangulares, que formam a matriz $\Phi(l)$, obter a saída

$$\mathbf{X}_f(l) = \Phi(l) \mathbf{X}_{\log}(k) \quad 1 \leq l \leq 13. \quad (4.21)$$

5. Calcular os coeficientes MFCC(l) realizando uma DCT-II de $\mathbf{X}_f(l)$ pela expressão:

$$\mathbf{MFCC}(l) = \text{DCT}[\mathbf{X}_f(l)] \quad (4.22)$$

4.2.3.5 LFPC

Calculamos os LFPC utilizando o algoritmo detalhado em (41). 12 faixas de frequências são selecionadas, com centros distribuídos logaritmicamente.

Algoritmo

1. Para cada janela, obter o sinal janelado $\mathbf{x}(n)$ com uma janela de Hamming.
2. Realizar a FFT do sinal janelado e obter a magnitude do espectro $|\mathbf{X}(k)|$.
3. Para cada sub-banda l , com largura de banda b_l e frequência central f_l definidos pelas expressões

$$b_l = \alpha b_{l-1} \quad (4.23)$$

e

$$f_l = f_1 + \sum_{j=1}^{l-1} b_j + \frac{b_l - b_1}{2}, \quad (4.24)$$

calcular o filtro

$$U_l(k) = \begin{cases} 1 & \text{se } l_L \leq k \leq l_H \\ 0 & \text{caso contrário,} \end{cases} \quad (4.25)$$

sendo b_1 a banda do primeiro filtro = 54 Hz,

f_1 a frequência central do primeiro filtro = 127 Hz,

α o fator de crescimento logarítmico = 1,4,

l_L o limite inferior (menor frequência discreta) da faixa de passagem l e

l_H o limite superior (maior frequência discreta) da faixa de passagem l .

4. Calcular para cada um dos filtros, a saída

$$\text{LFPC}_l(k) = |X(k)|U_l(k) \quad (4.26)$$

4.2.4 Seleção de Parâmetros - Sequential Forward Selection

Este trabalho emprega a seleção de parâmetros para GMM's que utilizam parâmetros dinâmicos, isto é, o contorno dos parâmetros de curto prazo. Como neste cenário é difícil utilizar métodos analíticos para pré-seleção de parâmetros antes da classificação, adotamos o algoritmo Sequential Forward Selection (SFS) que atua em conjunto com a GMM, atualizando o conjunto ótimo de parâmetros de voz após testar cada modelo gerado por um dado conjunto de parâmetros, verificando a classificação de cada amostra de voz da base de dados. O algoritmo SFS foi escolhido por sua simplicidade e por priorizar conjuntos pequenos de parâmetros de voz.

O algoritmo SFS é um algoritmo simples de busca gananciosa, ideal para situações onde deseja-se minimizar o número de parâmetros. Este algoritmo inicia com um conjunto vazio e iterativamente adiciona novos parâmetros ao conjunto ótimo de modo a maximizar uma determinada função objetiva $J(\mathbf{P})$. A implementação do SFS é bastante simples e segue os passos abaixo:

1. Iniciar com um conjunto vazio $\mathbf{P} = \emptyset$.
2. Calcular $J(\mathbf{P} \cup \{\mathbf{p}_l\})$ para cada parâmetro $\mathbf{p}_l$.
3. Selecionar o próximo melhor parâmetro

$$\mathbf{p}_+ = \arg \max_{\mathbf{P} \cup \{\mathbf{p}_l\}} J(\mathbf{P} \cup \{\mathbf{p}_l\}). \quad (4.27)$$

4. Se o melhor parâmetro encontrado não aumentar $J(\mathbf{P})$, encerrar. Se não continuar para o passo 5.
5. Atualizar $\mathbf{P} \leftarrow \mathbf{P} \cup \{\mathbf{p}_+\}$.
6. Voltar ao passo 2.

4.3 Métodos de classificação

Nesta seção são apresentados os métodos de classificação utilizados no projeto. Na maior parte do estudo utilizamos GMM's, porém inicialmente realizamos testes com outros métodos frequentes na literatura incluindo HMM, KNN e SVM, para ter uma idéia do desempenho de cada método. Apresentaremos a seguir os detalhes de implementação de cada método. Lembramos que os algoritmos possuem como entrada as matrizes de parâmetros $\mathbf{P}(j)$, onde j denota o índice da janela e por vezes omitiremos o índice j , para indicar toda a matriz de parâmetros.

4.3.1 Gaussian Mixture Models

O tradicional algoritmo de Expectation Maximization (59) foi utilizado para estimar os parâmetros das gaussianas. Para a classificação, não é necessário nenhum algoritmo em especial, bastando apenas calcular a probabilidade da sequência de observações segundo cada modelo, e escolher o modelo com maior verossimilhança.

Para a inicialização das gaussianas foram testados os algoritmos K-Means e Linde-Buzo-Gray (LBG) (61), sendo escolhido o LBG por ser mais estável. O algoritmo Split-and-Merge para refinamento da solução foi também testado, porém foi abandonado devido a resultados inexpressivos. De qualquer forma, consideramos importante descrever cada um dos algoritmos testados, inclusive os abandonados, uma vez que a implementação pode ter impacto nos resultados e consequentemente, na decisão pelos algoritmos.

4.3.1.1 Expectation Maximization

Expectation Maximization é um algoritmo iterativo comumente usado para problemas de maximização de verossimilhança (Maximum Likelihood Estimation - MLE). É composto de duas etapas, a etapa de Esperança (Passo E), na qual é calculada a verossimilhança e a etapa de Maximização (Passo M), na qual os parâmetros das componentes gaussianas são estimados e o modelo é modificado de modo a maximizar a verossimilhança. O algoritmo é sempre convergente, de modo que a cada iteração, a verossimilhança não decresce, isto é,

$$L(\mathbf{P}, \boldsymbol{\lambda}) \leq L(\mathbf{P}, \bar{\boldsymbol{\lambda}}), \quad (4.28)$$

$$L(\mathbf{P}, \boldsymbol{\lambda}) = \sum_{j=1}^J \ln \left\{ \sum_{m=1}^M c_m \mathcal{N}(\mathbf{P}(j), \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right\}. \quad (4.29)$$

Algoritmo

1. Inicializar as médias e matrizes de covariância de todas as componentes gaussianas por meio de algum algoritmo de agrupamento como o K-Means ou LBG.
2. Passo E: Para cada gaussiana, calcular a probabilidade a posteriori

$$\gamma_m(j) = \frac{c_m \mathcal{N}(\mathbf{P}(j), \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^M c_i \mathcal{N}(\mathbf{P}(j), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}, \quad (4.30)$$

$$\sum_{m=1}^M \gamma_m(j) = 1, \forall j, \quad (4.31)$$

que representa o quanto a gaussiana m é responsável pela observação $\mathbf{P}(j)$, assim como a frequência total da gaussiana m

$$N_m = \sum_{j=1}^J \gamma_m(j). \quad (4.32)$$

3. Passo M: Para cada gaussiana, calcular os novos valores de c_m , $\boldsymbol{\mu}_m$ e $\boldsymbol{\Sigma}_m$ pelas expressões

$$\bar{c}_m = \frac{N_m}{J}, \quad (4.33)$$

$$\bar{\boldsymbol{\mu}}_m = \frac{1}{N_m} \sum_{j=1}^J \gamma_m(j) \mathbf{P}(j), \quad (4.34)$$

$$\bar{\boldsymbol{\Sigma}}_m = \frac{1}{N_m} \sum_{j=1}^J \gamma_m(j) (\mathbf{P}(j) - \boldsymbol{\mu}_m)(\mathbf{P}(j) - \boldsymbol{\mu}_m)^T. \quad (4.35)$$

4. Calcular a nova verossimilhança $\bar{L}(\mathbf{P}, \boldsymbol{\lambda})$ pela expressão (4.29). Se a verossimilhança convergir, isto é

$$\left| \frac{\bar{L}(\mathbf{P}, \boldsymbol{\lambda})}{L(\mathbf{P}, \boldsymbol{\lambda})} - 1 \right| < 10^{-10}, \quad (4.36)$$

encerrar. Se não, voltar ao passo 2.

4.3.1.2 K-Means

O K-Means (60) é um algoritmo para agrupamento não-supervisionado, isto é, uma tentativa de agrupar em conjuntos uma sequência de dados inicialmente não relacionados, sob algum critério de semelhanças e diferenças. Basicamente o K-means define centróides e agrupa os conjuntos pela proximidade dos dados aos centróides, de acordo com algum critério, que pode ser por exemplo a distância euclidiana

$$\text{Dist}(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{\sum_{l=1}^L [x(l) - \mu(l)]^2}, \quad (4.37)$$

onde x é uma determinada observação, μ é o centróide em questão e L é a dimensão dos vetores.

O algoritmo pode ser utilizado para inicializar misturas de gaussianas ou como método de treinamento e classificação para quantização vetorial.

Algoritmo

1. Inicializar M centróides escolhendo aleatoriamente M pontos da sequência de dados.
2. Indexar cada ponto da sequência de dados ao centróide mais próximo, calculando a distância euclidiana dos pontos aos centróides.
3. Calcular a média e o desvio padrão de cada grupo, utilizando os pontos que pertencem a cada grupo. As novas médias definem novos centróides a cada iteração.
4. Se a diferença dos novos centros aos antigos for suficientemente pequena, encerrar. Se não, voltar ao passo 2.

Devido à escolha aleatória dos centróides iniciais, se rodarmos o K-means inúmeras vezes, o resultado final pode variar significativamente, de modo que o resultado é sempre sub-ótimo, uma vez que a minimização das diferenças dos centróides é um processo local. Deste modo, muitas vezes o resultado final pode não ser significativo devido a uma má escolha inicial.

4.3.1.3 Linde-Buzo-Gray

O algoritmo de Linde-Buzo-Gray (LBG) (61) é uma alternativa ao K-Means, como método de agrupamento não-supervisionado. O algoritmo segue praticamente todos os passos do K-means, exceto que a escolha inicial dos centróides não é aleatória: o primeiro centróide escolhido é sempre a média do conjunto completo dos dados e a partir daí, o centro segue sendo dividido em dois, com a adição de pequenas perturbações aos centros. Este processo é repetido para cada centro até atingir-se o número final de centros.

Algoritmo

1. Inicializar o primeiro centróide μ_0 como a média de todas as amostras.
2. Dividir o centróide em dois, μ_1 e μ_2 através de pequenas perturbações ϵ_1 e ϵ_2 , pela expressão

$$\begin{aligned}\mu_1 &= \mu_0 + \epsilon_1 \\ \mu_2 &= \mu_0 + \epsilon_2.\end{aligned}\tag{4.38}$$

3. Calcular a média e o desvio padrão de cada grupo, utilizando os pontos que pertencem a cada grupo. As novas médias definem novos centróides a cada iteração.

4. Dividir novamente os centróides gerados e repetir o passo 3 até atingir o número final de centros.
5. Continuar o algoritmo exatamente como o K-Means.

Algumas escolhas típicas para as perturbações são:

$$\begin{aligned} \epsilon_1 &= \mathbf{0}, & \epsilon_2 &= \boldsymbol{\mu}_0 \epsilon, & 0 < \epsilon << 1 \\ \epsilon_1 &= \boldsymbol{\mu}_0 \epsilon, & \epsilon_2 &= -\boldsymbol{\mu}_0 \epsilon, & 0 < \epsilon << 1 \end{aligned} \quad (4.39)$$

Porém, para este trabalho propomos uma modificação, utilizando as perturbações

$$\epsilon_1 = \boldsymbol{\sigma}_0 \epsilon, \quad \epsilon_2 = -\boldsymbol{\sigma}_0 \epsilon, \quad 0 < \epsilon < 1, \quad (4.40)$$

onde $\boldsymbol{\sigma}_0$ é um vetor de desvios padrão do centróide a ser dividido.

Uma desvantagem do LBG é que como as perturbações adicionadas aos centros são arbitrárias, não há garantia de que o resultado final seja ótimo. O resultado final é sempre o mesmo e pode ser melhor, pior ou igual ao K-Means. Entretanto, adotamos o LBG quando utilizamos GMM's com seleção de parâmetros pois a utilização de um algoritmo imprevisível neste contexto pode comprometer o conjunto ótimo.

4.3.1.4 Split and Merge

O objetivo do algoritmo Split and Merge é fornecer uma solução para o problema dos máximos locais, refinando a distribuição de modo a aumentar a verossimilhança. A idéia é refinar a solução e manter constante o número de gaussianas, simultaneamente unindo duas gaussianas em uma região pouco povoada e separando gaussianas em uma região bastante povoada, de modo a caracterizar com mais detalhes esta região.

Os candidatos para a junção são selecionados a partir da similaridade de suas probabilidades a posteriori, por uma espécie de correlação cruzada definida por

$$J_{merge}(m_1, m_2, \boldsymbol{\lambda}) = \mathbf{Prob}_{m_1}(\mathbf{P}, \boldsymbol{\lambda})^T \mathbf{Prob}_{m_2}(\mathbf{P}, \boldsymbol{\lambda}), \quad (4.41)$$

$$\mathbf{Prob}_m(\mathbf{P}, \boldsymbol{\lambda}) = [\text{Prob}(m|\mathbf{P}(1)), \text{Prob}(m|\mathbf{P}(2)), \dots, \text{Prob}(m|\mathbf{P}(J))]. \quad (4.42)$$

Desta forma, as gaussianas com maior J_{merge} devem ser unidas. Já os candidatos para a separação podem ser escolhidos pela divergência local de Kullback-Leibler definida por

$$J_{split}(m, \boldsymbol{\lambda}) = \int \Phi_m(\mathbf{P}, \boldsymbol{\lambda}) \log \frac{\Phi_m(\mathbf{P}, \boldsymbol{\lambda})}{\text{Prob}(\mathbf{P}(j)|m, \boldsymbol{\lambda})}, \quad (4.43)$$

$$\Phi_m(\mathbf{P}, \boldsymbol{\lambda}) = \frac{\sum_{j=1}^J \delta(\mathbf{P} - \mathbf{P}(j)) \text{Prob}(m|\mathbf{P}(j), \boldsymbol{\lambda})}{\sum_{j=1}^J \text{Prob}(m|\mathbf{P}(j), \boldsymbol{\lambda})}, \quad (4.44)$$

onde o candidato com maior $J_{split}(m, \boldsymbol{\lambda})$ deve ser separado.

Algoritmo

1. Realizar o algoritmo EM tradicional, obtendo um modelo inicial $\boldsymbol{\lambda}$.
2. Calcular as métricas J_{merge} e $J_{split}(m, \boldsymbol{\lambda})$ e selecionar as gaussianas que serão modificadas.
3. Atualizar as gaussianas a serem unidas pelas equações

$$\bar{c}_m = c_{m_1} + c_{m_2}, \quad (4.45)$$

$$\bar{\boldsymbol{\mu}}_m = \frac{\boldsymbol{\mu}_{m_1} \sum_{j=1}^J \text{Prob}(m_1|\mathbf{P}(j), \boldsymbol{\lambda}) + \boldsymbol{\mu}_{m_2} \sum_{j=1}^J \text{Prob}(m_2|\mathbf{P}(j), \boldsymbol{\lambda})}{\sum_{j=1}^J \text{Prob}(m_1|\mathbf{P}(j), \boldsymbol{\lambda}) + \sum_{j=1}^J \text{Prob}(m_2|\mathbf{P}(j), \boldsymbol{\lambda})}, \quad (4.46)$$

$$\bar{\boldsymbol{\Sigma}}_m = \frac{\boldsymbol{\Sigma}_{m_1} \sum_{j=1}^J P(m_1|\mathbf{P}(j), \boldsymbol{\lambda}) + \boldsymbol{\Sigma}_{m_2} \sum_{j=1}^J P(m_2|\mathbf{P}(j), \boldsymbol{\lambda})}{\sum_{j=1}^J P(m_1|\mathbf{P}(j), \boldsymbol{\lambda}) + \sum_{j=1}^J P(m_2|\mathbf{P}(j), \boldsymbol{\lambda})}. \quad (4.47)$$

4. Atualizar as gaussianas a serem separadas pelas equações

$$\bar{c}_{m_1} = \bar{c}_{m_2} = \frac{c_m}{2}, \quad (4.48)$$

$$\begin{aligned} \bar{\boldsymbol{\mu}}_{m_1} &= \boldsymbol{\mu}_m(1 + \epsilon), & 0 < \epsilon \ll 1 \\ \bar{\boldsymbol{\mu}}_{m_2} &= \boldsymbol{\mu}_m(1 - \epsilon), & 0 < \epsilon \ll 1, \end{aligned} \quad (4.49)$$

$$\bar{\boldsymbol{\Sigma}}_{m_1} = \bar{\boldsymbol{\Sigma}}_{m_2} = \det(\boldsymbol{\Sigma}_m)^{1/d} \mathbf{I}_d. \quad (4.50)$$

5. Realizar novamente o algoritmo EM tradicional obtendo um novo modelo $\bar{\boldsymbol{\lambda}}$ e a nova verossimilhança $\bar{L}(x, \boldsymbol{\lambda})$.
6. Se $\bar{L}(x, \boldsymbol{\lambda}) > L(x, \boldsymbol{\lambda})$, encerrar o algoritmo. Se não, voltar ao passo 2.

4.3.2 Hidden Markov Models

Como já mencionado no capítulo 3, existem três problemas clássicos envolvendo HMM:

1. Dada uma sequência de observações, como calcular eficientemente a probabilidade de a observação ter sido gerada pelo modelo?
2. Dada uma sequência de observações, qual é a sequência de estados percorrida?
3. Dada uma sequência de observações, qual o modelo que melhor a descreve? Em outras palavras, quais são os valores dos parâmetros do modelo que devem ser utilizados para maximizar a probabilidade da observação ter sido gerada pelo mesmo?

As soluções para estes problemas são descritas em (58), seguindo um modelo de treliças, representado na figura 12. Neste modelo, cada nó representa uma possibilidade de um estado ter sido atingido em um instante de tempo diferente, com diversos caminhos de chegada e de saída e exibindo sempre o vetor de observações correspondente a esse instante de tempo. Ou seja, cada nó pode ser representado por uma tripla [estado, instante, observação].

Figura 12 – Modelo de treliças para HMM. Cada nó representa um estado em um instante de tempo (janela) diferente, com os correspondentes vetores de observações

Fonte: Adaptada de (58)

Os problemas 1 e 3 são de interesse para o reconhecimento automático de emoções e são descritos a seguir.

4.3.2.1 Problema 1

Para resolver o problema 1, calcular a probabilidade da observação dado o modelo $\text{Prob}[\mathbf{P}|\boldsymbol{\lambda}]$, devemos levar em conta todos os caminhos do diagrama de treliças. Uma primeira solução seria enumerar cada uma das possíveis sequências de estados e somar todas as probabilidades. Obviamente, esta solução não é a mais recomendada, dado que acabamos por refazer muitas contas já que vários caminhos compartilham trechos, e que existem outras muito mais eficientes, como por exemplo o algoritmo Forward-Backward. O nome deste algoritmo advém de sua constituição em duas partes, forward (para frente) e backward (para trás). Para a resolução do problema 1, necessitamos da parte forward do algoritmo descrita a seguir:

1. Definir a variável forward

$$\alpha_j(q) = \text{Prob}[\mathbf{P}(1)\mathbf{P}(2)\dots\mathbf{P}(j), z(j) = e_q | \boldsymbol{\lambda}], \quad (4.51)$$

isto é, a probabilidade de estar no nó da treliça com estado e_q no instante j , tendo gerado a sequência de observações $P(1), P(2), \dots, P(j)$.

2. Calcular

$$\alpha_1(q) = \pi b_q[\mathbf{P}(1)], \quad 1 \leq q \leq Q \quad (4.52)$$

para os nós iniciais.

3. Calcular

$$\alpha_{j+1}(q) = \sum_{r=1}^Q \alpha_j(q) a_{qr} b_r[\mathbf{P}(j+1)] \quad 1 \leq q \leq Q, \quad 1 \leq j \leq J-1 \quad (4.53)$$

para os demais nós.

4. Calcular a probabilidade final

$$\text{Prob}[\mathbf{P}(j) | \boldsymbol{\lambda}] = \sum_{q=1}^Q \alpha_j(q) \quad (4.54)$$

A figura 13 demonstra o detalhe entre transições de nós na treliça na operação forward.

Figura 13 – Transição de nós na treliça(forward).

Fonte: Adaptada de (58)

4.3.2.2 Problema 3

Este é o problema mais difícil, uma vez que não existe solução analítica para encontrar o conjunto de parâmetros que maximiza $\text{Prob}[\mathbf{P}|\boldsymbol{\lambda}]$. Entretanto, soluções iterativas são possíveis para encontrar máximos locais, como o algoritmo de Baum-Welch e derivações do algoritmo de EM, técnicas de gradiente descendente, entre outros. Aqui discutimos um algoritmo baseado no trabalho de Baum e Welch que é basicamente uma implementação alternativa do algoritmo de EM.

Antes de prosseguirmos no algoritmo precisamos definir outras variáveis:

- $\beta_j(q)$: representa a probabilidade de estar no nó da treliça com estado e_q no instante j e gerar as observações $\mathbf{P}(j+1), \mathbf{P}(j+2) \dots \mathbf{P}(J)$. Isto é:

$$\beta_j(q) = \text{Prob}[\mathbf{P}(j+1)\mathbf{P}(j+2) \dots \mathbf{P}(J), z(j) = e_q | \boldsymbol{\lambda}] \quad (4.55)$$

- $\gamma_j(q)$: representa a probabilidade de estar no nó da treliça com estado e_q no instante j dada toda a sequência de observações. Isto é:

$$\gamma_j(q) = \text{Prob}[z(j) = e_q | \mathbf{P}(j), \boldsymbol{\lambda}] \quad (4.56)$$

- $\xi_j(q, r)$: representa a probabilidade de transitar entre dois nós da treliça, do estado e_q para o estado e_r , entre os instantes j e $j+1$ dada a sequência de observações. Isto é:

$$\xi_j(q, r) = \text{Prob}[z(j) = e_q, \quad z(j+1) = e_r | \mathbf{P}(j), \boldsymbol{\lambda}] \quad (4.57)$$

- $E_\gamma(q)$: representa o número esperado de transições partindo do estado e_q .
- $E_\xi(q)$: representa o número esperado de transições do estado e_q para o estado e_r , dada toda a sequência de observações.
- $\gamma_j(q, m)$: representa a probabilidade de estar no nó da treliça com estado e_q no instante j , utilizando a componente gaussiana m , para a sequência de observações. Isto é:

$$\gamma_j(q, m) = \text{Prob}[z(j) = e_q, m | \mathbf{P}(j), \boldsymbol{\lambda}] \quad (4.58)$$

Todos os passos do algoritmo podem ser derivados maximizando-se a função de custo de Baum:

$$L(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = \sum_{\mathbf{e}} \text{Prob}[\mathbf{e} | \mathbf{P}, \boldsymbol{\lambda}] \log(\text{Prob}[\mathbf{e} | \mathbf{P}, \bar{\boldsymbol{\lambda}}]), \quad (4.59)$$

onde $\bar{\boldsymbol{\lambda}}$ representa o novo modelo atualizado, e $\mathbf{e}$, uma particular sequência de estados.

Foi provado que este procedimento tem verossimilhança crescente, isto é, a cada iteração, $P[x | \bar{\boldsymbol{\lambda}}] \geq P[x | \boldsymbol{\lambda}]$, ou seja, o novo modelo sempre possui verossimilhança maior ou igual ao modelo anterior (58).

Algoritmo

1. Inicializar as médias e matrizes de covariância de todas as combinações de estados e componentes gaussianas por meio de algum algoritmo de agrupamento como o K-Means.
2. Calcular

$$\beta_J(q) = 1, \quad 1 \leq q \leq Q, \quad (4.60)$$

$$\beta_j(q) = \sum_{r=1}^Q a_{qr} b_r[\mathbf{P}(j+1)] \beta_{j+1}(r), \quad 1 \leq q \leq Q, \quad 1 \leq j \leq J-1, \quad (4.61)$$

para todos os nós.

3. Calcular

$$\gamma_j(q) = \frac{\alpha_j(q) \beta_j(q)}{\sum_{k=1}^Q \alpha_j(k) \beta_j(k)} \quad (4.62)$$

para todos os nós.

4. Calcular

$$\xi_j(q, r) = \frac{\alpha_j(q) a_{qr} b_r [\mathbf{P}(j+1)] \beta_{j+1}(r)}{\sum_{k=1}^Q \alpha_j(k) \beta_j(k)} \quad (4.63)$$

para todos os nós.

5. Calcular

$$E_\gamma(q) = \sum_{j=1}^{J-1} \gamma_j(q) \quad (4.64)$$

para todos os nós.

6. Calcular

$$E_\xi(q, r) = \sum_{j=1}^{J-1} \xi_j(q, r) \quad (4.65)$$

para todos os nós.

7. Calcular

$$\bar{\pi}_q = \gamma_1(q) \quad (4.66)$$

para todos os estados, como o número esperado de transições partindo do estado e_q no instante 1.

8. Calcular

$$\bar{a}_{qr}(q, r) = \frac{E_\xi(q, r)}{E_\gamma(q)} \quad (4.67)$$

para todos os estados, como uma razão entre o número esperado de transições do estado e_q para o estado e_r pelo número esperado de transições partindo do estado e_q .

9. Calcular

$$\gamma_j(q, m) = \gamma_j(q) \frac{c_{qm} \mathcal{N}(\mathbf{P}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{k=1}^M c_{qk} \mathcal{N}(\mathbf{P}, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)} \quad (4.68)$$

para todos os nós e componentes.

10. Calcular

$$\bar{c}_{qm} = \frac{\sum_{j=1}^J \gamma_j(q, m)}{\sum_{j=1}^J \sum_{k=1}^M \gamma_j(q, k)} \quad (4.69)$$

para cada estado e_q e componente gaussiana m como a razão do número esperado de vezes no estado e_q , utilizando a componente gaussiana m , pelo número esperado de vezes no estado e_q .

11. Calcular

$$\bar{\boldsymbol{\mu}}_{qm} = \frac{\sum_{j=1}^J \gamma_j(q, m) \cdot \mathbf{P}(j)}{\sum_{j=1}^J \gamma_j(q, m)} \quad (4.70)$$

para cada estado e_q e componente gaussiana m como a razão do número esperado de vezes no estado e_q , utilizando a componente gaussiana m , porém ponderado pela sequencia de observações, pelo número esperado de vezes no estado e_q , utilizando a componente gaussiana m .

12. Calcular

$$\bar{\Sigma}_{qm} = \frac{\sum_{j=1}^J \gamma_j(q, m) \cdot [\mathbf{P}(j) - \boldsymbol{\mu}_{qm}][\mathbf{P}(j) - \boldsymbol{\mu}_{qm}]^\top}{\sum_{j=1}^J \gamma_j(q, m)} \quad (4.71)$$

para cada estado e_q e componente gaussiana m como a razão do número esperado de vezes no estado e_q , utilizando a componente gaussiana m , porém ponderado por uma estimativa da matriz de correlação dada pela sequência de observações e a média atual, pelo número esperado de vezes no estado e_q , utilizando a componente gaussiana m .

13. Calcular a verossimilhança

$$P[\mathbf{P}|\bar{\boldsymbol{\lambda}}] = \sum_{q=1}^Q \alpha_j(q) \beta_j(q) \quad (4.72)$$

com o novo modelo $\bar{\boldsymbol{\lambda}}$ e se suficientemente maior que a verossimilhança na iteração anterior, adotar o novo modelo.

4.3.3 K-Nearest Neighbours

A implementação do KNN (60) é bastante direta, bastando definir o critério de distâncias, o valor de K e o mecanismo de desempates, em caso de duas ou mais classes apresentarem o mesmo número de vizinhos.

Algoritmo

1. Rotular cada amostra de treinamento e guardar seu vetor de parâmetros
2. Normalizar os parâmetros de voz para que contribuam igualmente.
3. Calcular a distância Euclidiana da amostra alvo para cada amostra de treinamento.
4. Selecionar as K amostras com menor distância euclidiana
5. Escolher a classe que mais se repete dentre os K vizinhos.
6. Em caso de empates, escolher mais vizinhos até o desempate.

4.3.4 Support Vector Machines

Para a implementação de SVM foi utilizada a biblioteca libSVM (73) que já conta com uma solução para problemas multi-classes. Entretanto, acabamos por desenvolver um mecanismo de classificação próprio baseado em votações, já que este apresentou desempenho superior à solução da libSVM. Em caso de empates, a escolha segue um caminho eliminatório entre as emoções candidatas, utilizando os modelos que confrontam as emoções candidatas diretamente, de modo a selecionar a vencedora final como a classe

escolhida definitivamente.

Para entender o algoritmo que gera o hiperplano separador ótimo, é necessário voltar um pouco para o problema de separação de duas classes por um hiperplano linear. Considerando um conjunto de amostras $\mathbf{P}$, cada amostra $\mathbf{p}(u) \in R^L$ - sendo u o índice da amostra e L o número de parâmetros de voz - deve ser mapeada para uma das duas classes como $y(u) \in [-1, 1]$. Os pontos $\mathbf{p}(u)$ que se encontram no hiperplano separador satisfazem

$$\mathbf{w} \cdot \mathbf{p}(u) + b = 0, \quad (4.73)$$

onde $\mathbf{w}$ é normal ao hiperplano e $|b|/\|\mathbf{w}\|$ é a distância medida ao longo da reta perpendicular do hiperplano à origem. Considerando-se agora que as amostras de treinamento satisfazem

$$\mathbf{w} \cdot \mathbf{p}(u) + b \geq +1, \quad y(u) = +1, \quad (4.74)$$

$$\mathbf{w} \cdot \mathbf{p}(u) + b \leq -1, \quad y(u) = -1, \quad (4.75)$$

ou combinando as equações (4.74) e (4.75),

$$y(u)[\mathbf{w} \cdot \mathbf{p}(u) + b] - 1 \geq 0, \quad (4.76)$$

os pontos para os quais vale a igualdade nas equações (4.74) e (4.75) encontram-se nos respectivos hiperplanos, com distâncias à origem iguais a $|1 - b|/\|\mathbf{w}\|$ e $|-1 - b|/\|\mathbf{w}\|$, resultando numa margem igual a $2/\|\mathbf{w}\|$.

Portanto, para encontrar a solução ótima basta resolver:

Minimizar

$$\|\mathbf{w}\|^2, \quad (4.77)$$

sujeito a (4.76).

Para resolver este problema de minimização sujeito a restrições, recorre-se aos multiplicadores de Lagrange (74). O problema pode ser substituído pelo lagrangeano

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{u=1}^U \alpha_u y[\mathbf{w} \cdot \mathbf{p}(u) + b] + \sum_{t=1}^T \alpha_t, \quad (4.78)$$

Este é um problema de otimização convexa quadrática que pode ser substituído pela sua forma dual:

Maximizar

$$L_D = \sum_u \alpha_u - \sum_{u,u'} \alpha_u \alpha_{u'} y(u) y(u') \mathbf{p}(u) \mathbf{p}(u') \quad (4.79)$$

sujeito a

$$\sum_u \alpha_u y(u) = 0 \quad (4.80)$$

Uma vez que a única forma que as amostras aparecem na fórmula (4.79) é na forma de um produto interno $\mathbf{p}(u) \mathbf{p}(u')$, pode-se facilmente estender a implementação para o caso não-linear, de modo que o produto interno $\mathbf{p}(u) \mathbf{p}(u')$ pode ser substituído pelo produto interno $\phi[\mathbf{p}(u)] \phi[\mathbf{p}(u')]$, onde ϕ é a transformação de $\mathbf{p}(u)$ para um espaço aumentado. Ao pensar-se que existe uma função kernel

$$K(\mathbf{p}(u) \mathbf{p}(u')) = \phi[\mathbf{p}(u)] \phi[\mathbf{p}(u')], \quad (4.81)$$

não é necessário sequer saber qual é a transformada ϕ e é possível aplicar diretamente a expressão (4.81) em (4.79). Algumas funções kernel comumente utilizadas são apresentadas na tabela 7:

Tabela 7 – Exemplos de funções kernel

Tipo	Expressão	Parâmetros
Polinomial	$[\mathbf{p}(u) \mathbf{p}(u') + 1]^k$	k
Base Radial	$e^{-\frac{\ \mathbf{p}(u) - \mathbf{p}(u')\ ^2}{2\sigma^2}}$	σ
Sigmoidal	$\tanh[\beta \mathbf{p}(u) \mathbf{p}(u') - \sigma]$	β, σ

4.4 Modelos e estratégias de classificação

Nesta seção, explicamos as abordagens adotadas para cada cenário investigado neste estudo.

4.4.1 Identificação em estágio único

Este é o cenário modelado na maioria dos estudos (2)(3)(4)(6)(7)(41)(64)(45). Neste cenário, cada emoção possui seu modelo λ que utiliza sempre os mesmos parâmetros de

voz, e a diferença de seus valores entre as classes é o que determina a emoção.

Neste estudo, consideramos este cenário para os testes iniciais utilizando somente pitch e energia, tanto para classificação dinâmica como estática e depois realizamos testes com GMM's, com seleção de parâmetros dinâmicos pelo algoritmo SFS, que atua em conjunto com o GMM, atualizando o conjunto ótimo de parâmetros após testar cada conjunto específico. Neste caso, o algoritmo SFS é implementado de modo a maximizar a seguinte função objetiva:

$$\text{rr}(\mathbf{P}) = \frac{\sum_{e=1}^E \text{rr}_e(\mathbf{P})}{E}, \quad (4.82)$$

onde E é o número de emoções, $\text{rr}(\mathbf{P})$ é a taxa de reconhecimento geral utilizando o conjunto de parâmetros $\mathbf{P}$ e $\text{rr}_e(\mathbf{P})$ denota a taxa de reconhecimento individual de cada emoção. O conjunto ótimo rr_{opt} é dado por

$$\begin{aligned} \text{rr}_{opt} &= \text{rr}(\mathbf{P}_{opt}), \\ \mathbf{P}_{opt} &= \arg \max_{\mathbf{P}} [\text{rr}(\mathbf{P})], \end{aligned} \quad (4.83)$$

Em outras palavras, o critério de desempenho para o SFS é acurácia geral dada pela média da diagonal principal da matriz de confusões. A média é preferível em vez da proporção entre acertos totais/esperados pois a base de dados EMO-DB utilizada possui amostras desbalanceadas entre as emoções. Utilizando-se a média entre os acertos de cada emoção não se privilegia nenhuma emoção. Para cada conjunto de parâmetros testado, são criados todos os modelos em contexto deixe-um-locutor-fora e é realizada a classificação de todas as amostras de voz. O conjunto é fechado e considerado ótimo quando a adição de qualquer parâmetro não é capaz de incrementar a função objetiva.

O maior problema com este modelo é que ele espera que todas as emoções sejam representadas pelos mesmos parâmetros de voz, o que é uma suposição ousada. Mesmo utilizando técnicas para seleção de parâmetros, alguns destes parâmetros podem não ser relevantes para todas as emoções, podendo até mesmo prejudicar a classificação, já que eles podem direcionar a classificação para uma emoção em que estes parâmetros podem ser cruciais. Um modelo que tenta contornar este problema é o modelo de parâmetros dependentes de classe.

4.4.2 Parâmetros dependentes de classe

Levando-se e conta que certos parâmetros são importantes para algumas emoções mas não para outras, pode-se pensar em estratégias para contornar este problema. Assim, uma alternativa é adotar modelos com parâmetros dependentes de classe, onde o modelo

de cada emoção pode ser decomposto em sub-modelos, cada um com um conjunto diferente de parâmetros que dependem de a emoção pertencer a um determinado grupo de emoções. Desta forma um modelo λ pode ser expresso por

$$\lambda = [\lambda_1 \lambda_2, \lambda_3, \dots \lambda_n], \quad (4.84)$$

onde cada sub-modelo $\lambda_1, \lambda_2, \lambda_3$ representa um subgrupo de emoções. Este modelo tem sido utilizado em alguns trabalhos na literatura (37), (38), (36), (75). Liu (50) utiliza combinações de pares de emoções selecionando o melhor conjunto de parâmetros em cada classificação com o algoritmo SFS. Entretanto, a maioria dos estudos organiza os grupos de emoções utilizando o modelo dimensional do espaço de emoções e separa as emoções de acordo com suas proximidades nas dimensões Ativação, Avaliação e Domínio.

4.4.3 Identificação em três estágios/Modelo dimensional

Uma forma de implementar modelos com parâmetros dependentes de classes é utilizando a noção de um espaço de emoções composto das dimensões Avaliação, Ativação e Domínio. Desta forma, podem-se agrupar as emoções segundo suas proximidades nos três eixos do espaço de emoções, supondo-se que cada dimensão pode ser representada por um diferente conjunto de parâmetros de voz, de modo que as emoções que se localizam mais próximas em um dos eixos do espaço apresentam valores também mais próximos para este conjunto de parâmetros. Da mesma maneira, todas as emoções individualmente também possuem conjuntos de parâmetros de voz específicos.

Sendo assim, as emoções utilizadas neste estudo, Felicidade, Medo, Neutro, Nojo, Raiva, Tédio e Tristeza podem ser agrupadas segundo a tabela 8. A base para a geração dos grupos são as convenções encontradas na literatura (25), (38)(36), (37). Entretanto, a escolha de Raiva como uma emoção positiva é uma fuga da teoria de emoções com o objetivo de aumentar-se a taxa de reconhecimento, para a classificação sequencial. Como Felicidade e Raiva são emoções que frequentemente se confundem (25) (36)(64)(38)(37) é preferível deixar esta distinção para o último estágio. Agrupamentos de emoções segundo a taxa de reconhecimento é uma técnica que já foi utilizada em (38) e (75).

Assim, é possível utilizar um sistema de classificação sequencial em 3 estágios, passando por classificações nas dimensões Ativação, Avaliação e Domínio. A figura 9 mostra a distribuição das emoções em grupos para cada estágio de classificação. O primeiro estágio separa as emoções de alta ativação (Felicidade, Medo, Nojo, Raiva) das de baixa ativação (Neutro, Tédio, Tristeza). Se classificada como alta ativação, o segundo estágio separa Felicidade e Raiva (Positivas) de Medo e Nojo (Negativas) e em caso contrário são separados Neutro (Positiva) de Tédio e Tristeza. O último estágio dá a classificação final, separando Felicidade de Raiva, Medo de Nojo ou Tédio de Tristeza.

Tabela 8 – Grupos de emoções nas 3 dimensões do espaço de emoções

	Ativação	Avaliação	Domínio
Felicidade	Alta	Positiva	Alto
Medo	Alta	Negativa	Baixo
Neutro	Baixa	Positiva	Alto
Nojo	Alta	Negativa	Alto
Raiva	Alta	Positiva	Baixo
Tédio	Baixa	Negativa	Alto
Tristeza	Baixa	Negativa	Baixo

Tabela 9 – Estágios de classificação

Estágio	Alta / Positiva	Baixa / Negativa
Estágio 1: Ativação	Felicidade, Medo, Nojo, Raiva	Neutro, Tédio, Tristeza
Estágio 2(1): Avaliação	Felicidade, Raiva	Medo, Nojo
Estágio 2(2): Avaliação	Neutro	Tédio, Tristeza
Estágio 3(1): Domínio	Felicidade	Raiva
Estágio 3(2): Domínio	Medo	Nojo
Estágio 3(3): Domínio	Tédio	Tristeza

Para melhor ilustrar o processo, a Figura 14 repete a figura apresentada no capítulo 2 com o diagrama de blocos da classificação sequencial, porém com as emoções utilizadas de fato.

Para cada estágio, a seleção de parâmetros pode ser realizada do mesmo modo que na classificação em estágio único, com a diferença de que a taxa de reconhecimento de cada emoção é dada agora pela porcentagem de acertos da emoção em relação ao seu grupo correto.

4.4.4 Verificação de emoções

Modelos dependentes de classe fazem ainda mais sentido quando se trata do problema de verificação de emoções. Como o interesse neste caso é detectar uma única emoção, é possível criar modelos com parâmetros específicos para separar a emoção alvo de todas as demais, sem a necessidade de se preocupar com quais são as outras emoções. Portanto, o sistema de classificação deve se comportar como um identificador de estágio único onde existem apenas dois modelos, o modelo da emoção alvo e o modelo complementar, formado

Figura 14 – O processo de classificação em 3 estágios

Fonte: Autoria Própria

por todas as demais emoções.

A seleção de parâmetros é realizada como nos casos anteriores, sendo as taxas de reconhecimento das demais emoções dadas pela porcentagem de vezes em que elas não são classificadas como a emoção alvo. A função objetiva pode ser alterada para:

$$\text{rr}(\mathbf{P}) = 0.5 \left(\text{rr}_a(\mathbf{P}) + \frac{\sum_{e=1}^{E-1} \text{rr}_e(\mathbf{P})}{E-1} \right), \quad (4.85)$$

onde $\text{rr}(\mathbf{P})$ é a taxa de reconhecimento geral para o conjunto de parâmetros $\mathbf{P}$, $\text{rr}_a(\mathbf{P})$ é a taxa de reconhecimento da emoção alvo e $\text{rr}_e(\mathbf{P})$ é a taxa de reconhecimento para cada uma das emoções restantes. Desta forma, o sistema torna-se um especialista na emoção alvo, onde é priorizado o custo de falsos negativos sobre o custo de falsos positivos.

4.4.5 Classificação híbrida

É possível também pensar num sistema híbrido que inicialmente verifica uma determinada emoção, e depois identifica as demais, de modo a priorizar a emoção alvo, mas ainda assim realizar a identificação completa. A idéia do processo é ilustrada na Figura 15.

Além disso, uma vez que o classificador sequencial pode apresentar dificuldades em classificar emoções que não possuem altas taxas de reconhecimento nos primeiros estágios, o classificador híbrido pode resolver alguns destes problemas, já que ele é projetado para ser um especialista na emoção problemática. Com isso, pode-se até aumentar a taxa de reconhecimento geral do sistema.

Figura 15 – Processo de classificação híbrido.

Fonte: Autoria Própria

4.5 Testes experimentais

Todos os experimentos foram realizados em cenário independente de locutor, utilizando a técnica de validação cruzada “deixe-um-locutor-fora” (em inglês “leave-one-speaker-out”). Ou seja, todos os experimentos contemplam toda a base de dados, sendo as amostras de voz de cada locutor confrontadas com modelos criados especificamente para eles. Entretanto, é considerado um cenário dependente de gênero, sendo as amostras de homens e mulheres previamente separadas e os testes para homens e mulheres realizados em processos distintos. Embora o sistema pudesse apresentar uma etapa inicial de detecção de gênero como em (38), este processo pode ser considerado fora do escopo do trabalho e portanto não foi implementado. Além disso, elevadas taxas de detecção de gênero, próximas a 100%, foram encontradas em (76), o que comprova a viabilidade do processo.

Desta forma os seguintes experimentos foram realizados:

- Testes iniciais de identificação utilizando os parâmetros Pitch e Energia, comparando os métodos GMM, HMM, KNN e SVM.
- Testes de identificação em um único estágio utilizando GMM com seleção de parâmetros.
- Testes de identificação em três estágios, utilizando GMM’s com seleção de parâmetros para cada estágio.
- Testes de verificação de cada emoção da base EMO-DB.
- Testes de sistema híbrido de verificação e identificação, utilizando a emoção Nojo.

No capítulo seguinte são apresentados mais detalhes dos testes e os respectivos resultados.

5 Experimentos Práticos

5.1 Primeiros Experimentos

Para os primeiros experimentos, foram selecionados os parâmetros de voz pitch e energia, que embora não tenham passado por nenhum algoritmo para esta seleção, possuem uma explicação bastante razoável para suas escolhas:

- Intuitivamente possuem papel importante principalmente para classificar emoções de alta excitação como raiva e alegria, pois parece ter grande relevância o volume (ou variação do volume) e a forma como as palavras e frases são entonadas. Por exemplo, é fácil notar que uma pessoa irritada tende a falar em volume mais alto e em tons mais agudos. Além disso, esta análise não depende muito de quem está falando, ou seja do locutor.
- Já foram estudados extensivamente pela área da Psicologia e ficaram famosos pelo estudo de Murray e Arnott (39).
- São parâmetros usados em diversos trabalhos na literatura, embora alguns autores afirmem que existam parâmetros mais relevantes. Em (7) a primeira e segunda derivadas temporais do pitch foram identificadas como os principais parâmetros para o reconhecimento de emoções e em (6), de 133 parâmetros incluindo estatísticas de pitch, energia, MFCC's e formantes, foram escolhidas 35, entre eles a média do pitch e 5 estatísticas da energia.

5.1.1 Testes com GMM

Foram realizados diversos experimentos alterando-se o número M de Gaussianas da mistura. Para a inicialização dos valores iniciais, dispomos de um sorteio aleatório com distribuição uniforme para gerar o vetor de probabilidades de cada componente da mistura e utilizamos o método K-means para gerar a distribuição inicial das gaussianas (vetor de médias e matriz de covariância). É importante lembrar que devido às inicializações aleatórias das probabilidades das componentes e dos centros iniciais das gaussianas, o resultado do modelo inicial nem sempre é o mesmo e por este motivo realizamos 3 experimentos para cada vez que alteramos o número de gaussianas. Utilizamos matrizes de covariância completas com dimensão $L = 6$ (pitch, energia e suas primeiras e segundas derivadas temporais), e limite máximo de 100 iterações para a re-estimação dos parâmetros do GMM. A Tabela 10 mostra os resultados dos experimentos realizados para o conjunto de locutores

Tabela 10 – Desempenho de GMM's como a porcentagem de emoções corretamente identificadas para diferentes números de gaussianas (M) em 3 experimentos, variando-se as condições iniciais.

	% de Acertos (1)	% de Acertos (2)	% de Acertos (3)	Média
M=1	50,64	50,64	50,64	50,64
M=2	51,07	50,64	50,21	50,64
M=4	48,93	51,07	52,36	50,79
M=8	53,65	52,79	54,08	53,51
M=16	51,93	54,94	54,08	53,65
M=32	56,65	54,08	53,65	54,79
M=64	56,65	54,51	52,36	54,51

(a) Locutores De Sexo Masculino

	% de Acertos (1)	% de Acertos (2)	% de Acertos (3)	Média
M=1	40,07	40,07	40,07	40,07
M=2	50,33	50,00	49,67	50,00
M=4	52,32	51,66	51,66	51,21
M=8	58,61	56,29	56,62	57,17
M=16	59,60	57,28	57,62	58,17
M=32	56,95	57,95	57,95	57,62
M=64	58,61	58,61	56,95	58,06

(b) Locutores de Sexo Feminino

de sexo masculino e feminino. Para cada valor de M que representa o número de gaussianas da mistura, variando de 1 a 32, temos a porcentagem da quantidade de emoções reconhecidas corretamente em 3 experimentos, com diferentes condições iniciais. Podemos perceber que os testes com mulheres obtiveram desempenho ligeiramente superior para todos os valores de M maiores do que 1 e que para ambos os sexos, a partir de 8 gaussianas o resultado varia pouco, tendo o valor inicial dos parâmetros assumido papel cada vez mais importante no desempenho.

A Tabela 11 mostra a matriz de confusão para os melhores resultados gerais para homens (52,79%, M = 32) e mulheres (59,10%, M = 16). Mostramos aqui apenas os melhores casos, pois a distribuição do desempenho das emoções seguiu o mesmo perfil para homens e mulheres nos demais casos. As linhas da matriz indicam a emoção correta das amostras de voz, e as colunas, as emoções em que as amostras foram classificadas. Assim, para cada emoção representando as escolhas, e cada emoção real das amostras, temos a porcentagem de amostras classificadas sob este par. Observando estas matrizes, percebemos que há muito menos confusões nos testes dos locutores de sexo feminino, com o principal erro sendo a escolha de raiva no lugar de felicidade, inclusive mais intensas do que no caso masculino. Já para os testes com locutores de sexo masculino, podemos notar muitas confusões de raiva por felicidade, nojo por neutro, medo por nojo e neutro por tédio. Além disso, é possível perceber que embora em alguns casos o desempenho geral seja melhor, podem ocorrer mais erros grotescos de confusão, priorizando a escolha de uma

Tabela 11 – Matriz de confusão das emoções para GMM, indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) nos casos de melhor desempenho geral.

	F	M	Ne	No	R	Te	Tr
F	40,74	11,11	0,00	0,00	48,15	0,00	0,00
M	13,89	63,89	2,78	8,33	11,11	0,00	0,00
Ne	0,00	5,13	33,33	33,33	0,00	15,38	12,82
No	0,00	45,45	0,00	45,45	0,00	0,00	9,09
R	11,67	3,33	0,00	0,00	85,00	0,00	0,00
Te	0,00	8,57	34,29	8,57	0,00	37,14	11,43
Tr	0,00	0,00	28,00	0,00	0,00	8,00	64,00

(a) Locutores de Sexo Masculino (Desempenho Geral: 56.65%,
Número de gaussianas = 32)

	F	M	Ne	No	R	Te	Tr
F	18,18	9,09	0,00	13,64	59,09	0,00	0,00
M	0,00	51,52	9,09	18,18	21,21	0,00	0,00
Ne	0,00	0,00	57,50	15,00	0,00	10,00	17,50
No	0,00	0,00	14,29	74,29	8,57	2,86	0,00
R	10,45	8,96	1,49	4,48	74,63	0,00	0,00
Te	0,00	0,00	21,74	6,52	0,00	56,52	15,22
Tr	0,00	0,00	2,70	5,41	0,00	10,81	81,08

(b) Locutores de Sexo Feminino (Desempenho Geral: 59,60%,
Número de gaussianas = 16)

ou mais classes incorretas. Como exemplo o caso feminino apresentou maiores confusões de felicidade por raiva, embora o desempenho geral tenha sido melhor.

5.1.2 Testes com HMM

Realizamos diversos testes com diferentes números de gaussianas, variando também o número de estados Q . Para a inicialização dos parâmetros iniciais, o K-Means foi novamente utilizado com sorteios aleatórios com distribuição uniforme para o vetor de estados iniciais, matriz de transições de estados e matriz de probabilidades das componentes gaussianas.

A Tabela 12 mostra o desempenho geral para HMM's com 8 gaussianas e números de estados Q variando de 2 a 4. Testes com mais gaussianas apresentaram piores resultados, ao mesmo tempo em que aumentaram muito o custo computacional. Comparado com o uso do GMM, verificamos que a inclusão de mais estados não surtiu grandes efeitos no desempenho geral, até mesmo baixando o resultado geral em alguns casos para locutores do sexo feminino, o que pode ser explicado principalmente pelo fator estocástico da seleção dos valores iniciais. Mesmo assim, nos melhores casos foi possível aumentar tanto o desempenho masculino (59,23 %) quanto o feminino (61,26 %).

As Tabelas 13 e 14 mostram as matrizes de confusão dos experimentos com melhor desempenho para cada valor de Q . Podemos perceber que o perfil da distribuição é parecido

Tabela 12 – Desempenho de HMM's com 8 gaussianas como a porcentagem de emoções corretamente identificadas para diferentes números de estados (Q) em 3 experimentos, com diferentes parâmetros iniciais.

	% de Acertos (1)	% de Acertos (2)	% de Acertos (3)	Média
Q = 2	54,08	51,07	54,08	53,08
Q = 3	55,79	57,51	51,50	54,93
Q = 4	53,22	52,79	59,23	55,08

(a) Locutores de Sexo Masculino

	% de Acertos (1)	% de Acertos (2)	% de Acertos (3)	Média
Q = 2	52,98	55,97	56,95	55,30
Q = 3	57,62	56,62	57,28	57,17
Q = 4	57,95	61,26	61,26	60,16

(b) Locutores de Sexo Feminino

com o dos testes com GMM e que as mesmas confusões são notadas, porém, cada hora uma emoção diferente é priorizada. Entendemos que novamente este fato tem muito mais a ver com a escolha inicial dos parâmetros do HMM do que com o número de estados, uma vez que para todos os valores de Q houve flutuações na distribuição das confusões.

5.1.3 Testes com KNN

Tratemos agora de métodos de classificação estáticos com parâmetros globais. Iniciamos esta etapa com um dos algoritmos existentes mais simples, o K-Vizinhos mais Próximos (KNN), como tentativa de avaliar o tamanho do problema e o grau de separação dos parâmetros nas diferentes classes.

Considerando os conjuntos de parâmetros, realizamos três ensaios:

- No primeiro, utilizamos somente média e desvio padrão dos contornos de pitch e energia e de suas primeiras e segundas derivadas temporais, obtendo uma matriz de $L=12$ parâmetros globais.
- No segundo, utilizamos também os 2,5º, 25º, 50º, 75º e 97,5º percentis, para os contornos e as derivadas temporais como forma de descrever melhor as densidades de probabilidade das classes, totalizando uma matriz de $L=42$ parâmetros globais.
- No terceiro, utilizamos também um parâmetro determinado pela máxima duração (número de janelas consecutivas) em cada região delimitada pelos percentis, como tentativa de modelar melhor a dinâmica das emoções. Neste último conjunto, a matriz de parâmetros totaliza $L=78$ parâmetros globais.

Realizamos diversos experimentos alterando-se o número de vizinhos de 1 a 10. A Tabela 15 mostra os resultados do desempenho geral do KNN para os testes com os três

Tabela 13 – Matriz de confusão das emoções para locutores de sexo masculino utilizando HMM's com 8 gaussianas, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). São apresentados os casos de melhores desempenhos em cada valor de Q (número de estados).

	F	M	Ne	No	R	Te	Tr
F	48,14	7,41	0,00	0,00	44,44	0,00	0,00
M	22,22	47,22	8,33	8,33	13,89	0,00	0,00
Ne	0,00	7,69	41,03	25,64	0,00	7,69	17,95
No	9,09	27,27	0,00	63,64	0,00	0,00	0,00
R	20,00	6,67	0,00	0,00	73,33	0,00	0,00
Te	0,00	8,57	37,14	8,57	0,00	34,29	11,43
Tr	0,00	0,00	16,00	8,00	0,00	8,00	68,00

(a) Locutores de Sexo Masculino (Desempenho Geral: 54,08%,
M = 8, Q = 2)

	F	M	Ne	No	R	Te	Tr
F	37,03	14,81	0,00	0,00	48,19	0,00	0,00
M	8,33	66,67	0,00	13,89	11,11	0,00	0,00
Ne	0,00	7,69	35,89	28,21	0,00	10,26	17,95
No	9,09	36,36	0,00	45,45	0,00	0,00	9,09
R	10,00	5,00	0,00	0,00	85,00	0,00	0,00
Te	0,00	8,57	37,14	8,57	0,00	37,14	8,57
Tr	0,00	0,00	20,00	4,00	0,00	8,00	68,00

(b) Locutores de Sexo Masculino (Desempenho Geral: 57,51%,
M = 8, Q = 3)

	F	M	Ne	No	R	Te	Tr
F	40,74	7,41	0,00	0,00	51,85	0,00	0,00
M	19,44	55,56	11,11	0,00	13,89	0,00	0,00
Ne	0,00	7,69	43,59	25,64	0,00	12,82	10,26
No	9,09	45,45	0,00	36,36	0,00	0,00	9,09
R	5,00	5,00	0,00	0,00	90,00	0,00	0,00
Te	0,00	11,43	37,14	2,86	0,00	40,00	8,57
Tr	0,00	0,00	20,00	4,00	0,00	4,00	72,00

(c) Locutores de Sexo Masculino (Desempenho Geral: 59,23%,
M = 8, Q = 4)

conjuntos de parâmetros ($L = 12$, $L = 42$ e $L = 78$), variando o número de vizinhos K de 1 a 10. Como esperado, não existe um valor claro de K que otimiza o desempenho e podemos perceber que tanto a inclusão dos percentis quanto de suas durações máximas puderam elevar levemente o desempenho na maioria dos casos. Entretanto, o desempenho geral foi bastante inferior aos obtidos com GMM e HMM.

Os resultados máximos de 50,64% (masculino) e 48,01% (feminino) foram obtidos, o que reflete a simplicidade do algoritmo.

As Tabelas 16 e 17 mostram a matriz de confusão nos melhores casos para cada conjunto de parâmetros. Como podemos ver, o desempenho individual de muitas emoções também piorou em relação ao GMM e ao HMM e a inclusão de novos parâmetros alterou pouco o perfil da distribuição. No caso masculino, tivemos problemas com Felicidade

Tabela 14 – Matriz de confusão das emoções para locutores de sexo feminino utilizando HMM's com 8 gaussianas, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). São apresentados os casos de melhores desempenhos em cada valor de Q (número de estados).

	F	M	Ne	No	R	Te	Tr
F	15,91	6,82	2,27	11,36	63,64	0,00	0,00
M	0,00	45,45	6,06	24,24	24,24	0,00	0,00
Ne	0,00	0,00	65,00	10,00	0,00	5,00	20,00
No	0,00	5,71	17,14	62,86	8,57	0,00	5,71
R	5,97	4,48	1,49	5,97	82,09	0,00	0,00
Te	2,17	0,00	39,13	6,52	0,00	43,48	8,69
Tr	0,00	0,00	8,11	10,81	0,00	8,11	72,97

(a) Locutores de Sexo Feminino (Desempenho Geral: 56,95%,
M = 8, Q = 2)

	F	M	Ne	No	R	Te	Tr
F	22,73	11,36	2,27	9,09	54,55	0,00	0,00
M	0,00	51,52	6,06	18,18	24,24	0,00	0,00
Ne	0,00	0,00	55,00	17,50	0,00	12,50	15,00
No	2,86	2,86	22,86	62,86	5,71	0,00	2,86
R	8,96	5,97	0,00	4,48	80,59	0,00	0,00
Te	0,00	0,00	23,91	8,69	0,00	47,83	19,57
Tr	0,00	0,00	10,81	5,41	0,00	10,81	72,97

(b) Locutores de Sexo Feminino (Desempenho Geral: 57,62%,
M = 8, Q = 3)

	F	M	Ne	No	R	Te	Tr
F	20,45	6,82	2,27	9,09	61,36	0,00	0,00
M	3,03	39,39	9,09	27,27	21,21	0,00	0,00
Ne	0,00	0,00	45,00	20,00	0,00	12,50	22,50
No	5,71	2,86	17,14	68,57	2,86	2,86	0,00
R	4,47	1,49	0,00	5,97	88,06	0,00	0,00
Te	4,35	0,00	17,39	4,35	0,00	69,57	4,35
Tr	0,00	0,00	8,11	5,41	0,00	5,45	81,08

(c) Locutores de Sexo Feminino (Desempenho Geral: 61,26%,
M = 8, Q = 4)

(confundido principalmente com Raiva), Nojo (obteve 0% de acertos e foi confundido principalmente com Medo), Tédio (confundido principalmente com Neutro) e Tristeza (confundido principalmente com Neutro). No caso feminino, tivemos problemas com Felicidade (confundido principalmente com Raiva), Medo (confundido principalmente com Raiva), Nojo (teve uma distribuição bastante esparsa) e Raiva (confundido principalmente com Felicidade). Entretanto a emoção Neutra obteve bom desempenho em ambos os casos.

5.1.4 Testes com SVM

A biblioteca libSVM (73) foi utilizada para a realização dos testes com SVM em duas abordagens: Na primeira utilizamos diretamente a solução multiclases implementada no pacote libSVM. Na segunda utilizamos máquinas para emoções duas a duas, onde cada

Tabela 15 – Desempenho de KNN para diferentes números de vizinhos (K) e diferentes conjuntos de parâmetros, sendo L o número de parâmetros utilizados.

	% de Acertos (L = 12)	% de Acertos (L = 42)	% de Acertos (L = 78)
K=1	37,77	44,64	46,35
K=2	36,91	44,21	45,49
K=3	44,21	45,49	48,49
K=4	46,35	47,21	45,49
K=5	43,35	44,64	49,35
K=6	46,35	45,92	48,07
K=7	44,64	46,35	47,21
K=8	45,92	47,21	48,49
K=9	45,49	48,49	50,64
K=10	47,21	49,79	48,49

(a) Locutores de Sexo Masculino

	% de Acertos (L = 12)	% de Acertos (L = 42)	% de Acertos (L = 78)
K=1	39,07	45,03	46,03
K=2	38,08	42,38	42,05
K=3	41,72	47,68	47,35
K=4	42,05	44,37	45,03
K=5	41,06	44,37	45,36
K=6	43,71	41,72	44,37
K=7	45,03	43,38	45,03
K=8	44,37	44,70	46,36
K=9	48,01	46,03	47,02
K=10	45,03	45,03	45,36

(b) Locutores de Sexo Feminino

máquina seleciona a mais provável entre as duas e criamos um esquema de votação que seleciona a emoção escolhida mais vezes. Além disso, criamos um mecanismo de desempates: No caso de empates, a escolha segue um caminho eliminatório entre as emoções candidatas, utilizando as máquinas que confrontam as emoções candidatas diretamente, de modo a selecionar a vencedora final como a classe escolhida definitivamente.

Os conjuntos de parâmetros utilizados foram os mesmos dos testes com KNN, e realizamos testes com kernel linear, polinomial de graus 2,3 e 4, funções de base radial (rbf) gaussiana e sigmoidal. Entretanto, na solução multiclases da biblioteca libSVM somente o kernel polinomial gerou resultados satisfatórios, sendo que os outros kernels geraram desempenho próximo a 0%. Da mesma maneira, na solução por votações o kernel rbf gaussiano e o kernel sigmoidal apresentaram o mesmo problema. Acreditamos que nestes casos ocorreu overfitting nas amostras de treinamento, e por terem apresentado resultado tão insatisfatório não são apresentados.

A Tabela 18 mostra os resultados para o desempenho geral do SVM para kernel linear e kernel polinomial com graus 2, 3 e 4, para os dois sistemas testados, variando novamente os 3 conjuntos de parâmetros. Podemos perceber que o desempenho geral do método de votação com kernel linear foi superior e que a inclusão dos percentis e de suas durações máximas conseguiu elevar o desempenho da classificação. O resultado também foi superior

Tabela 16 – Matriz de confusão para locutores de sexo masculino utilizando KNN, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). Para cada experimento com número de parâmetros L , é apresentado o caso com o número de vizinhos que apresentou o melhor desempenho geral.

	F	M	Ne	No	R	Te	Tr
F	3,70	33,33	0,00	0,00	62,96	0,00	0,00
M	2,78	52,78	16,67	0,00	27,78	0,00	0,00
Ne	0,00	20,51	58,97	2,56	0,00	12,82	5,13
No	18,18	54,55	27,27	0,00	0,00	0,00	0,00
R	10,00	20,00	0,00	0,00	70,00	0,00	0,00
Te	0,00	5,71	45,71	2,86	0,00	40,00	5,71
Tr	0,00	4,00	32,00	0,00	0,00	20,00	44,00

(a) Locutores de Sexo Masculino (Desempenho Geral: 47,21%,
 $K = 10$, $L = 12$)

	F	M	Ne	No	R	Te	Tr
F	25,93	11,11	0,00	0,00	62,96	0,00	0,00
M	8,33	41,67	19,44	0,00	30,56	0,00	0,00
Ne	0,00	15,38	58,97	2,56	0,00	12,82	10,26
No	18,18	54,55	9,09	0,00	0,00	18,18	0,00
Ra	6,67	8,33	0,00	0,00	85,00	0,00	0,00
Te	0,00	11,43	34,29	2,86	2,86	40,00	8,57
Tr	4,00	20,00	48,00	0,00	0,00	4,00	24,00

(b) Locutores de Sexo Masculino (Desempenho Geral: 49,79%,
 $K = 10$, $L = 42$)

	F	M	Ne	No	R	Te	Tr
F	18,52	14,81	0,00	0,00	66,67	0,00	0,00
M	8,33	38,89	19,44	0,00	33,33	0,00	0,00
Ne	2,56	12,82	66,67	0,00	0,00	10,26	7,69
No	18,18	54,55	9,09	0,00	0,00	18,18	0,00
R	5,00	10,00	0,00	0,00	85,00	0,00	0,00
Te	2,86	11,43	28,57	2,86	0,00	45,71	8,57
Tr	4,00	20,00	52,00	0,00	0,00	0,00	24,00

(c) Locutores de Sexo Masculino (Desempenho Geral: 50,64%,
 $K = 9$, $L = 78$)

ao método KNN, porém inferior ao GMM e ao HMM, sendo a taxa de reconhecimento máxima obtida de 57,08% para homens e de 53,31% para mulheres.

As Tabelas 19 e 20 mostram as matrizes de confusão dos experimentos para kernel linear. No caso masculino as principais confusões foram de Felicidade por Raiva, Medo por Raiva e Nojo beirando 0% de acurácia sendo confundido principalmente com Felicidade, Neutro e Medo. Notamos que a inclusão de novos parâmetros pôde melhorar o desempenho individual de praticamente todas as emoções. No caso feminino, as principais confusões ocorrem entre Felicidade e Raiva, e Neutro e Tédio. A inclusão de novos parâmetros não alterou significativamente o perfil da distribuição, porém, conseguiu elevar o desempenho da emoção Felicidade (kernel linear), embora estes aumentos tenham sido acompanhados da diminuição do desempenho de outras emoções.

Tabela 17 – Matriz de confusão para locutores de sexo feminino utilizando KNN, tendo-se as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas). Para cada experimento com número de parâmetros L , é apresentado o caso com o número de vizinhos que apresentou o melhor desempenho geral

	F	M	Ne	No	R	Te	Tr
F	22,73	2,27	0,00	11,36	63,64	0,00	0,00
M	0,00	30,30	3,03	27,27	39,39	0,00	0,00
Ne	0,00	0,00	70,00	10,00	0,00	15,00	5,00
No	17,14	8,57	31,43	34,29	5,71	2,86	0,00
R	16,42	13,43	1,49	2,99	65,67	0,00	0,00
Te	2,17	0,00	28,26	8,70	0,00	52,17	8,70
Tr	0,00	0,00	37,84	2,70	0,00	13,51	45,95

(a) Locutores de Sexo Feminino (Desempenho Geral: 48,01%,
K = 9, L = 12)

	F	M	Ne	No	R	Te	Tr
F	43,18	4,55	0,00	4,55	47,73	0,00	0,00
M	15,15	42,42	0,00	18,18	24,24	0,00	0,00
Ne	5,00	2,50	67,50	2,50	0,00	22,50	0,00
No	25,71	22,86	20,00	20,00	8,57	0,00	2,86
R	40,30	8,96	1,49	0,00	49,25	0,00	0,00
Te	6,52	0,00	32,61	2,17	0,00	50,00	8,70
Tr	0,00	2,70	18,92	0,00	0,00	21,62	56,76

(b) Locutores de Sexo Feminino (Desempenho Geral: 47,68%,
K = 3, L = 42)

	F	M	Ne	No	R	Te	Tr
F	40,91	4,55	0,00	4,55	50,00	0,00	0,00
M	15,15	45,45	0,00	15,15	24,24	0,00	0,00
Ne	5,00	2,50	62,50	7,50	0,00	20,00	2,50
No	22,86	22,86	20,00	22,86	8,57	0,00	2,86
R	40,30	10,45	1,49	0,00	47,76	0,00	0,00
Te	8,70	0,00	30,43	0,00	0,00	52,17	8,70
Tr	0,00	2,70	24,32	0,00	0,00	16,22	56,76

(c) Locutores de Sexo Feminino (Desempenho Geral: 47,35%,
K = 3, L = 78)

5.1.5 Conclusão dos Primeiros Testes

Concluimos que HMM e GMM obtiveram os melhores desempenhos gerais entre os métodos testados. Mesmo utilizando apenas pitch e energia de curto prazo conseguimos chegar próximo (59,23% para homens e 61,26% para mulheres) da acurácia humana que na maioria dos testes não ultrapassa 70%. Concluimos também que um melhor desempenho geral não necessariamente acompanha um melhor desempenho individual de cada emoção. Neste contexto os testes com HMM's e GMM's produziram matrizes de confusão mais estáveis, principalmente no caso feminino, onde somente a emoção felicidade obteve acurácia inferior a 50%. Assim, confirmamos que o reconhecimento entre homens e mulheres têm perfis diferentes e que os mesmos parâmetros de fato não podem ser utilizados da mesma forma para ambos.

Tabela 18 – Desempenho do SVM como porcentagens gerais de acertos para diferentes tipos de kernel e diferentes conjuntos de L parâmetros.

Kernel	L = 12	L = 42	L = 78)
Linear(Votação)	45,92	54,08	57,08
Polinomial Grau 2(Votação)	50,64	50,21	51,50
Polinomial Grau 3(Votação)	51,93	50,64	51,93
Polinomial Grau 4(Votação)	50,21	47,64	53,22
Polinomial Grau 2(Multiclasses)	50,21	49,36	53,22
Polinomial Grau 3(Multiclasses)	51,07	47,21	52,36
Polinomial Grau 4(Multiclasses)	50,64	45,49	51,93

(a) Locutores de Sexo Masculino

Kernel	L = 12	L = 42	L = 78
Linear(Votação)	51,99	52,98	50,33
Polinomial Grau 2(Votação)	51,99	48,34	53,31
Polinomial Grau 3(Votação)	45,69	48,34	52,31
Polinomial Grau 4(Votação)	45,36	48,68	51,99
Polinomial Grau 2(Multiclasses)	48,34	49,01	51,99
Polinomial Grau 3(Multiclasses)	42,05	45,36	50,00
Polinomial Grau 4(Multiclasses)	42,72	47,02	50,99

(b) Locutores de Sexo Feminino

Embora nossos testes tenham sido bastante simples, com poucos parâmetros e de voz e sem a utilização de algoritmos para seleção de parâmetros, os resultados apoiam a superioridade de HMM's, indicada por alguns autores (55)(7)(56). Entretanto, o desempenho de HMM's e GMM's foram semelhantes tanto no geral como nas matrizes de confusão, sendo o HMM ligeiramente superior em troca de um custo computacional muito maior. Além disso, a superioridade dos HMM's sobre os GMM's ocorre apenas em alguns casos, dependendo da inicialização aleatória dos valores iniciais. Assim, optamos por utilizar GMM's nos demais testes, que seguem com o objetivo de melhorar os resultados obtidos, principalmente maximizando as diagonais principais das matrizes de confusão, e eliminando confusões grotescas.

5.2 Identificação em Estágio Único

Uma vez que os primeiros testes utilizando Pitch e Energia levaram a resultados ainda um pouco insatisfatórios, seguimos com novos testes utilizando novos parâmetros com os objetivos de obter melhores resultados e eliminar as principais confusões. Lembrando que Pitch e Energia atingem principalmente o nível da prosódia, adicionamos agora parâmetros que tentam revelar um pouco sobre a qualidade do áudio, e remetem a uma descrição do trato vocal do locutor durante a produção das emoções. Para este propósito, introduzimos as frequências formantes, coeficientes de distribuição espectral de energia e os coeficientes cepstrais. Para todos os novos testes, utilizamos GMM's de 16 gaussianas com matrizes de covariância diagonais, que representa um compromisso entre

Tabela 19 – Matriz de confusão das emoções para locutores de sexo masculino utilizando SVM com kernel linear, tendo-se as porcentagens de amostras de voz classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) para cada experimento com número de parâmetros L.

	F	M	Ne	No	R	Te	Tr
F	25,93	7,41	0,00	7,41	55,56	3,70	0,00
M	11,11	33,33	11,11	2,78	36,11	5,56	0,00
Ne	0,00	5,13	46,15	17,95	0,00	15,38	15,38
No	45,45	0,00	45,45	0,00	0,00	9,09	0,00
R	21,67	10,00	0,00	0,00	68,33	0,00	0,00
Te	2,86	5,71	25,71	8,57	0,00	45,71	11,43
Tr	0,00	4,00	16,00	8,00	0,00	20,00	52,00

(a) Locutores de Sexo Masculino (Desempenho Geral: 45,92%, Kernel Linear, L = 12)

	F	M	Ne	No	R	Te	Tr
F	29,63	14,81	3,70	11,11	40,74	0,00	0,00
M	13,89	44,44	5,56	8,33	19,44	5,56	2,78
Ne	0,00	10,26	48,72	7,69	0,00	15,38	17,95
No	36,36	18,18	36,36	9,09	0,00	0,00	0,00
R	16,67	8,33	0,00	0,00	75,00	0,00	0,00
Te	0,00	11,43	11,43	8,57	0,00	57,14	11,43
Tr	0,00	0,00	16,00	0,00	0,00	16,00	68,00

(b) Locutores de Sexo Masculino (Desempenho Geral: 54,08%, Kernel Linear, L = 42)

	F	M	Ne	No	R	Te	Tr
F	33,33	14,81	3,70	11,11	37,04	0,00	0,00
M	5,56	55,56	11,11	5,56	22,22	0,00	0,00
Ne	0,00	12,82	56,41	5,13	0,00	12,82	12,82
No	27,27	36,36	0,00	9,09	9,09	9,09	9,09
R	21,67	6,67	0,00	1,67	70,00	0,00	0,00
Te	2,86	5,71	22,86	0,00	0,00	62,86	5,71
Tr	0,00	0,00	16,00	8,00	0,00	8,00	68,00

(c) Locutores de Sexo Masculino (Desempenho Geral: 57,08%, Kernel Linear, L = 78)

desempenho e custo computacional, para os resultados dos testes anteriores. O algoritmo SFS foi aplicado para a seleção de parâmetros, trazendo à tona um problema no uso do K-Means: a divergência dos resultados produzidos a cada vez que os modelos de GMM são treinados novamente, devido à inicialização aleatória dos modelos no K-Means. Assim, existe a possibilidade de em alguma rodada do SFS um parâmetro mais relevante (que na maioria das vezes produz melhores resultados) produzir um resultado pior do que outro menos relevante e vice versa, prejudicando a seleção, e gerando de diferenças de até 5% na taxa de reconhecimento. Para resolver este problema, testamos os algoritmos Split and Merge e LBG. O Split and Merge em poucas vezes conseguiu superar a performance do algoritmo padrão de EM e ainda assim não conseguiu resolver o problema das inicializações, ao mesmo tempo em que aumentou muito a complexidade computacional pelo fato de re-executar o algoritmo de EM diversas vezes. Assim, adotamos o algoritmo

Tabela 20 – Matriz de confusão das emoções para locutores de sexo feminino utilizando SVM com kernel linear, tendo-se as porcentagens de amostras de voz classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas) para cada experimento com número de parâmetros L.

	F	M	Ne	No	R	Te	Tr
F	29,55	2,27	0,00	13,64	52,27	2,27	0,00
M	6,06	57,58	0,00	12,12	24,24	0,00	0,00
Ne	0,00	0,00	57,50	10,00	0,00	15,00	17,50
No	8,57	8,57	14,29	48,57	8,57	5,71	5,71
Ra	22,39	7,46	0,00	4,48	65,67	0,00	0,00
Te	4,35	0,00	19,57	6,52	0,00	56,52	13,04
Tr	0,00	0,00	16,22	16,22	0,00	27,03	40,54

(a) Locutores de Sexo Feminino (Desempenho Geral: 51,99%, Kernel Linear, L = 12)

	F	M	Ne	No	R	Te	Tr
F	36,36	4,55	0,00	9,09	45,45	4,55	0,00
M	9,09	48,48	0,00	9,09	27,27	3,03	3,03
Ne	0,00	0,00	47,50	2,50	2,50	30,00	17,50
No	8,57	11,43	11,43	54,29	5,71	5,71	2,86
R	26,87	5,97	1,49	1,49	64,18	0,00	0,00
Te	4,35	2,17	17,39	4,35	0,00	67,39	4,35
Tr	0,00	18,92	27,03	0,00	5,41	5,41	43,24

(b) Locutores de Sexo Feminino (Desempenho Geral: 52,98%, Kernel Linear, L = 42)

	F	M	Ne	No	R	Te	Tr
F	43,18	2,27	0,00	4,55	47,73	2,27	0,00
M	15,15	51,52	0,00	12,12	18,18	0,00	3,03
Ne	0,00	0,00	50,00	5,00	5,00	25,00	15,00
No	5,71	14,29	17,14	45,71	8,57	5,71	2,86
R	31,34	10,45	2,99	1,49	53,73	0,00	0,00
Te	2,17	0,00	19,57	4,35	4,35	65,22	4,35
Tr	0,00	10,81	21,62	8,11	10,81	10,81	37,84

(c) Locutores de Sexo Feminino (Desempenho Geral: 50,33%, Kernel Linear, L = 78)

LBG, que apesar de por vezes apresentar taxas de reconhecimento inferiores ao K-Means, obtém sempre o mesmo resultado, que se mostrou similar à média de resultados do K-Means.

Tabela 21 – Parâmetros selecionados para a identificação em estágio único

Homens	Mulheres
2º LFPC, Energia, δ 1ª Formante, δ Pitch, 5º MFCC, 8º MFCC, δ 8º LFPC, δ 4º LFPC, δ 2º MFCC	δ^2 3º LFPC, δ^2 5º LFPC, δ^2 7º LFPC, 2º MFCC, δ 3º MFCC, δ 7º MFCC

A Tabela 21 apresenta os parâmetros selecionados para a classificação em único

estágio, com os parâmetros em ordem decrescente de relevância e sendo δ e δ^2 , a primeira e segunda derivadas temporais dos parâmetros, respectivamente. Pode-se ver que houve uma distinta diferença entre os parâmetros selecionados para homens e mulheres. No caso masculino, energia e a primeira derivada do pitch aparecem como uns dos parâmetros mais importantes, porém, sequer aparecem no caso feminino, no qual perdem espaço para coeficientes LFPC de baixa ordem, o que apoia os resultados de Oudeyer (43), de que a energia em baixas frequências pode ser mais significativa do que o pitch e a energia total. Coeficientes LFPC's aparecem com maior relevância do que MFCC's, apoiando as conclusões de Nwe (41).

Tabela 22 – Matrizes de confusão do classificador de único estágio indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas).

	F	M	Ne	No	R	Te	Tr
F	77,78	7,41	0,00	0,00	14,81	0,00	0,00
M	11,11	75,00	0,00	0,00	5,56	5,56	2,78
Ne	0,00	2,56	56,41	2,56	0,00	23,08	15,38
No	0,00	9,09	0,00	72,73	0,00	18,18	0,00
R	8,33	5,00	0,00	0,00	86,67	0,00	0,00
Te	0,00	0,00	11,43	2,86	0,00	71,43	14,29
Tr	0,00	4,00	0,00	0,00	0,00	12,00	84,00

(a) Locutores de sexo masculino - Resultado geral: 74,86%

	F	M	Ne	No	R	Te	Tr
F	40,91	2,27	0,00	15,91	40,91	0,00	0,00
M	9,09	72,73	6,06	6,06	6,06	0,00	0,00
Ne	0,00	0,00	65,00	15,00	0,00	15,00	5,00
No	5,71	5,71	5,71	80,00	0,00	0,00	2,86
R	14,93	7,46	0,00	1,49	76,12	0,00	0,00
Te	0,00	0,00	10,87	13,04	0,00	76,09	0,00
Tr	0,00	0,00	2,70	5,41	0,00	0,00	91,89

(b) Locutores de sexo feminino - Resultado geral: 71,82%

Os resultados do classificador de único estágio, utilizando os parâmetros selecionados na tabela 21, são apresentados na tabela 22. Podemos ver que os resultados de 74,86% masculino e 71,82% feminino foram bastante superiores aos dos primeiros testes, estando o resultado geral no 70º percentil, assim como a acurácia geral da maioria das emoções em ambos os casos masculino e feminino. Entretanto, temos problemas com a emoção Felicidade no caso feminino e Neutro no caso masculino. Comparando com resultados da literatura para a base EMO-DB em cenário independente de locutor e utilizando apenas parâmetros acústicos de voz, o resultado geral é bastante superior ao encontrado em (6), que obteve uma taxa de reconhecimento geral de 55% utilizando Perceptrons Multicamadas e também um pouco acima dos 70,8% obtidos em (64), utilizando KNN e dos 71,07% obtidos em (45). Em compensação, está um pouco abaixo dos 75,5% obtidos com o uso de SVM por Shami e Verhelst (68), porém, devemos destacar que estes utilizaram

apenas 493 das 535 instâncias de áudio da base EMO-DB, não havendo informações sobre quais as instâncias deixadas de lado.

5.3 Identificação sequencial em 3 estágios

Tabela 23 – Parâmetros selecionados para cada estágio da identificação sequencial

Estágio	Homens	Mulheres
Estágio 1 : Ativação	δ 2º LFPC, δ 3º LFPC, δ^2 5º LFPC	δ^2 3º LFPC, δ 5º MFCC, 3º MFCC, 6º MFCC, δ^2 5º MFCC
Estágio 2(1): Avaliação	Pitch, 10º MFCC, 8º LFPC, Energia, δ 5ª Formante	Pitch, δ 2º MFCC, 2º MFCC, δ 3º MFCC
Estágio 2(2): Avaliação	δ^2 2º MFCC, δ 8º LFPC, δ^2 3º MFCC, δ 10º LFPC	δ 8º LFPC, δ^2 6º LFPC, δ 7º MFCC, δ^2 Energia, δ 4º MFCC, δ 2º MFCC, δ^2 12º LFPC
Estágio 3(1): Domínio	δ^2 2º LFPC, δ 6º LFPC, 7º LFPC, 11º MFCC	δ^2 8º LFPC, δ 8º MFCC
Estágio 3(2): Domínio	δ 1ª Formante, δ 2º MFCC	δ^2 7º LFPC, δ^2 5º LFPC, δ 3º MFCC, δ^2 8º LFPC
Estágio 3(3): Domínio	δ^2 9º LFPC, δ Energia	δ 1ª Formante, δ^2 13º MFCC

A tabela 23 apresenta os parâmetros de voz selecionados para cada estágio do sistema de identificação sequencial. Podemos ver que os resultados para os casos masculino e feminino foram novamente bastante diferentes, assim como para cada estágio individual. Observamos novamente uma predominância dos LFPC's, embora muitos MFCC's também tenham sido selecionados. Um fato interessante é que, contrariamente à idéia tradicional de que Pitch e Energia são importantes para a dimensão da Ativação, verificamos a seleção destes parâmetros somente nos estágios que consideram as dimensões da Avaliação e do Domínio. Uma outra curiosidade é que no caso feminino, todos os parâmetros selecionados para o sistema de estágio único aparecem em algum momento no sistema de 3 estágios, porém acrescidos de outros parâmetros, mostrando que a classificação sequencial é de certa forma mais completa e permite a seleção de parâmetros extras.

Tabela 24 – Resultados gerais para cada estágio de classificação.

Estágio	Resultado Geral Masculino	Resultado Geral Feminino
Estágio 1: Ativação	93,77	96,43
Estágio 2(1): Avaliação	94,67	91,22
Estágio 2(2): Avaliação	94,77	93,88
Estágio 3(1): Domínio	92,96	77,37
Estágio 3(2): Domínio	94,07	95,54
Estágio 3(3): Domínio	94,29	98,91

Tabela 25 – Matrizes de confusão do classificador sequencial de 3 estágios, indicando as porcentagens de amostras classificadas sob os pares Emoção Real (linhas) $\times$ Emoção Escolhida (colunas)

	F	M	Ne	No	R	Te	Tr
F	92,59	0,00	0,00	0,00	7,41	0,00	0,00
M	2,78	86,11	0,00	0,00	2,78	0,00	8,33
Ne	0,00	0,00	76,92	5,13	0,00	12,82	5,13
No	9,09	9,09	9,09	54,55	0,00	18,18	0,00
R	6,67	5,00	0,00	1,67	86,67	0,00	0,00
Te	0,00	0,00	5,71	2,86	0,00	80,00	11,43
Tr	0,00	0,00	0,00	0,00	0,00	0,00	100,00

(a) Locutores de sexo masculino - Geral: 82,41%

	F	M	Ne	No	R	Te	Tr
F	63,64	2,27	0,00	4,55	29,55	0,00	0,00
M	6,06	81,82	0,00	6,06	6,06	0,00	0,00
Ne	2,50	0,00	87,50	2,50	0,00	7,50	0,00
No	5,71	2,86	5,71	71,43	0,00	8,57	5,71
R	11,94	7,46	0,00	2,99	77,61	0,00	0,00
Te	0,00	0,00	10,87	0,00	0,00	86,96	2,17
Tr	0,00	0,00	0,00	0,00	0,00	0,00	100,00

(b) Locutores de sexo feminino - Geral: 81,28%

Os resultados gerais para cada estágio, utilizando os parâmetros selecionados na tabela 23 são apresentados na tabela 24. As matrizes de confusão do resultado final são mostrados na tabela 25. Podemos ver que, em comparação com o desempenho do classificador de único estágio, o resultado geral foi melhor tanto no caso masculino, que subiu de 74,86% para 82,41 %, quanto no caso feminino, que subiu de 71,82% para 81,28%. Em ambos os casos masculino e feminino, Tristeza teve desempenho perfeito, enquanto Nojo foi a única emoção que teve as taxas de reconhecimento diminuídas em relação ao classificador de estágio único. Isso ocorreu porque, embora a taxa geral no estágio da Ativação tenha sido 93%, a emoção Nojo teve uma taxa menor neste estágio (72,7% masculino e 80% feminino) e como a classificação é sequencial, a taxa de reconhecimento final não pode ser maior do que a do primeiro estágio. Este é um resultado importante que mostra um ponto de atenção para este método. No caso masculino, esse efeito provavelmente foi acentuado devido à falta de amostras de voz para a emoção Nojo e provavelmente o modelo para emoções de alta ativação foi criado favorecendo as emoções que possuem mais amostras, como por exemplo, a Raiva. De qualquer modo, os resultados mostram que o desempenho do sistema sequencial pode superar o do sistema de único estágio, chegando mais próximo ao reconhecimento humano.

Comparando-se este resultado com a literatura, o resultado foi superior aos 79,47% obtidos por Xiao et Al (38), que emprega um sistema de classificação sequencial utilizando apenas a dimensão da Ativação e não considera a emoção Nojo, que no nosso caso apresentou-se como a emoção de reconhecimento mais difícil. Por outro lado, o resultado está um

pouco abaixo dos 88,8% encontrados por Lugger e Yang (37), o que pode ser explicado principalmente pelo fato de Lugger e Yang também desconsiderarem a emoção Nojo, que apresentou uma grande diminuição na taxa de reconhecimento geral pela sua baixa taxa de reconhecimento individual, além do fato de que a existência de uma emoção a mais sempre tende a diminuir a taxa de reconhecimento das demais emoções.

5.4 Verificação de emoções

Tabela 26 – Parâmetros de voz selecionados para a verificação de cada emoção

Estágio	Homens	Mulheres
Felicidade	Pitch, δ 5ª Formante, 1º MFCC, 8º LFPC, 1º LFPC, 10º MFCC, 7º MFCC	δ Pitch, Pitch, δ 11º LFPC
Medo	δ 2ª Formante, δ 7º LFPC, 7º LFPC	3º LFPC, 12º LFPC, 11º MFCC, 7º LFPC, δ^2 12º MFCC, δ^2 7º MFCC
Neutro	δ Pitch, δ 1º MFCC, 5º LFPC, 1º LFPC, 8º MFCC, δ 13º MFCC	δ^2 4º LFPC, δ^2 6º LFPC, δ^2 12º LFPC, 4º MFCC, δ 5º MFCC, δ^2 Energia, 2º LFPC, δ^2 9º MFCC, 6º LFPC, 12º LFPC, 1ª Formante, 3ª Formante, δ 3º MFCC
Nojo	δ 2ª Formante, 8º MFCC, 9º MFCC, 12º LFPC, 2º LFPC, δ 1º MFCC, δ^2 12º MFCC	δ^2 7º LFPC, δ 3º LFPC, δ 4º LFPC, δ^2 3ª Formante, δ 2º LFPC
Raiva	δ 1º LFPC, δ 2º MFCC, δ^2 6º LFPC, δ 6º MFCC, δ^2 6º MFCC	δ 2º MFCC, 7º MFCC, 5º MFCC, δ 8º MFCC, δ^2 10º LFPC
Tédio	δ 12º MFCC, δ^2 9º MFCC, δ 5º LFPC, δ 9º MFCC, 8º LFPC, δ^2 5º LFPC, δ 5ª Formante, δ^2 3º MFCC, δ 13º MFCC	δ^2 Pitch, δ^2 7º LFPC, 6º MFCC, δ 8º LFPC, δ 7º MFCC, 5º MFCC, δ 1ª Formante
Tristeza	2º LFPC, δ^2 1ª Formante, δ 1º MFCC, 12º LFPC, δ 5ª Formante, δ 7º MFCC	δ 1ª Formante, 8º MFCC, δ Energia, δ^2 7º LFPC, δ^2 11º LFPC, δ^2 2º LFPC, δ 5º MFCC, δ^2 Energia

Os parâmetros de voz selecionados para a verificação de cada uma das emoções da base EMO-DB estão descritos na Tabela 26. Novamente os parâmetros para homens e mulheres, assim como para cada emoção, apresentam comportamentos distintos. Podemos observar novamente a importância dos LFPC's e para muitas emoções também dos MFCC's. Vemos que o Pitch aparece com extrema relevância para a emoção Felicidade, o que confere com a visão tradicional deste parâmetro e ao mesmo tempo, nota-se uma maior relevância das formantes para algumas emoções, sendo que estes parâmetros não costumam apresentar relevância na literatura.

Tabela 27 – Taxas de reconhecimento gerais para a verificação de cada emoção.

	Emoção alvo	Demais emoções	Geral
F	92,59	89,13	90,86
M	80,56	89,28	84,92
Ne	94,87	87,35	91,11
No	94,87	87,35	91,11
R	100,00	88,31	94,16
Te	97,14	88,74	92,94
Tr	92,00	90,48	91,24

(a) Locutores de sexo masculino

	Emoção alvo	Demais emoções	Geral
F	93,18	82,85	88,02
M	90,91	90,14	90,53
Ne	100,00	80,23	90,12
No	94,29	87,84	91,07
R	95,52	90,49	93,01
Te	100,00	86,41	93,21
Tr	100,00	98,63	99,32

(b) Locutores de sexo feminino

Os resultados para a verificação de cada emoção utilizando os parâmetros da Tabela 26 são apresentados na Tabela 27, representando a porcentagem geral de acertos para as instâncias da emoção alvo e do restante das instâncias. O resultado individual para cada emoção é mostrado na Tabela 28, onde $\tilde{E}$ refere-se à porcentagem de amostras que foram consideradas como não pertencentes à emoção alvo. Para melhor visualização, na Tabela 28 deliberadamente invertimos as linhas com as colunas, sendo as colunas as emoções reais e as linhas os modelos selecionados na classificação. Para a maioria das emoções, a taxa de reconhecimento geral foi maior que na tarefa de identificação, embora tenha ocorrido um aumento do número de falsos positivos. Isto ocorreu principalmente devido à função objetiva utilizada, que atribui um peso maior à taxa de reconhecimento da emoção alvo, enquanto para as outras é calculada uma média. Deste modo, torna-se claro o compromisso entre aumentar a taxa de reconhecimento de uma emoção e aumentar a taxa de falsos positivos de emoções que são de alguma forma similares. Mesmo assim, este algoritmo fornece uma boa solução quando o custo de falsos positivos é menor que o custo de falsos negativos.

5.5 Sistema Híbrido

Como exemplo para o sistema híbrido, utilizamos a emoção Nojo para o estágio de verificação, uma vez que esta emoção apresentou baixa taxa de reconhecimento no classificador sequencial de 3 estágios, principalmente no caso masculino. Os resultados são mostrados na Tabela 29. Podemos ver que a emoção Nojo teve seus resultados altamente elevados em ambos os casos masculino e feminino, em troca de uma penalidade na acurácia

Tabela 28 – Porcentagem de reconhecimento para cada par Emoção Escolhida (linhas) $\times$ Emoção Real (colunas).

	F	M	Ne	No	R	Te	Tr
F	92,59	2,78	0,00	9,09	53,33	0,00	0,00
$\tilde{F}$	7,41	97,22	100,00	90,91	46,67	100,00	100,00
M	18,52	80,56	5,13	0,00	16,67	0,00	24,00
$\tilde{M}$	81,48	19,44	94,87	100,00	83,33	100,00	76,00
Ne	0,00	2,78	94,87	0,00	0,00	57,14	16,00
$\tilde{N}e$	100,00	97,22	5,13	100,00	100,00	42,86	84,00
No	3,70	5,56	7,69	90,91	1,67	11,43	0,00
$\tilde{N}o$	96,30	94,44	92,31	9,09	98,33	88,57	100,00
R	48,15	19,44	2,56	0,00	100,00	0,00	0,00
$\tilde{R}$	51,85	80,56	97,44	100,00	0,00	100,00	100,00
Te	0,00	0,00	38,46	9,09	0,00	97,14	20,00
$\tilde{T}e$	100,00	100,00	61,54	90,91	100,00	2,86	80,00
Tr	0,00	8,33	23,08	0,00	0,00	25,71	92,00
$\tilde{T}r$	100,00	91,67	76,92	100,00	100,00	74,29	8,00

(a) Locutores de sexo masculino

	F	M	Ne	No	R	Te	Tr
F	93,18	21,21	0,00	8,57	73,13	0,00	0,00
$\tilde{F}$	6,82	78,79	100,00	91,43	26,87	100,00	100,00
M	2,27	90,91	5,00	17,14	10,45	0,00	24,32
$\tilde{M}$	97,73	9,09	95,00	82,86	89,55	100,00	75,68
Ne	2,27	6,06	100,00	25,71	1,49	69,57	13,51
$\tilde{N}e$	97,73	93,94	0,00	74,29	98,51	30,43	86,49
No	6,82	21,21	25,00	94,29	7,46	4,35	8,11
$\tilde{N}o$	93,18	78,79	75,00	5,71	92,54	95,65	91,89
R	54,55	0,00	2,50	0,00	95,52	0,00	0,00
$\tilde{R}$	45,45	100,00	97,50	100,00	4,48	100,00	100,00
Te	0,00	0,00	65,00	5,71	0,00	100,00	10,81
$\tilde{T}e$	100,00	100,00	35,00	94,29	100,00	0,00	89,19
Tr	0,00	6,06	0,00	0,00	0,00	2,17	100,00
$\tilde{T}r$	100,00	93,94	100,00	100,00	100,00	97,83	0,00

(b) Locutores de sexo feminino

de algumas emoções, devido ao aumento de falsos positivos no estágio de verificação. No caso masculino, o resultado geral foi ainda superior ao do classificador sequencial de 3 estágios com 83,91% contra 82,41% enquanto no caso feminino o resultado decaiu drasticamente de 81,28% para 73,79% uma vez que o aumento da taxa da emoção Nojo não foi suficiente para compensar a diminuição ocorrida nas demais emoções. Ainda assim, este método forneceu resultados melhores que o classificador de estágio único, mostrando-se uma boa alternativa em cenários em que é necessário realizar uma identificação completa, mas deve-se priorizar uma determinada emoção. Um outro possível cenário seria o cascadeamento de diversos estágios de verificação para emoções diferentes.

Tabela 29 – Matrizes de confusão para o sistema híbrido com verificação prévia da emoção Nojo.

	F	M	Ne	No	R	Te	Tr
F	88,89	0,00	0,00	3,70	7,41	0,00	0,00
M	2,78	80,56	0,00	5,56	2,78	0,00	8,33
Ne	0,00	5,13	71,79	7,69	0,00	10,26	5,13
No	0,00	0,00	9,09	90,91	0,00	0,00	0,00
R	5,00	6,67	0,00	1,67	86,67	0,00	0,00
Te	0,00	2,86	5,71	11,43	0,00	68,57	11,43
Tr	0,00	0,00	0,00	0,00	0,00	0,00	100,00

(a) Locutores de sexo masculino - Geral: 83,91%

	F	M	Ne	No	R	Te	Tr
F	59,09	2,27	0,00	6,82	31,82	0,00	0,00
M	6,06	54,55	0,00	21,21	12,12	3,03	3,03
Ne	2,50	2,50	62,50	25,00	0,00	7,50	0,00
No	5,71	0,00	0,00	94,29	0,00	0,00	0,00
R	11,94	8,96	0,00	7,46	71,64	0,00	0,00
Te	0,00	0,00	10,87	4,35	0,00	82,61	2,17
Tr	0,00	0,00	0,00	8,11	0,00	0,00	91,89

(b) Locutores de sexo feminino - Geral: 73,79%

5.6 Conclusão

Neste trabalho, analisamos os problemas de verificação e de identificação de emoções através da voz. Foram utilizados modelos dependentes de classes que se mostraram úteis para ambos os problemas. Em ambos os casos, foram alcançadas taxas de reconhecimento elevadas e comparáveis a sistemas estado da arte, sendo que os resultados encontrados qualificam o sistema proposto para diversas aplicações. No melhor caso, a emoção Tristeza pôde ser identificada com 100% de acurácia para ambos os casos feminino e masculino, em cenário independente de locutor. Além disso, as emoções Raiva no caso masculino e Neutro e Tédio no caso feminino obtiveram 100% de acurácia para o cenário de verificação, que costuma ser ignorado pelos autores na literatura.

O sistema de identificação em estágio único apresentou resultados superiores a diversos sistemas encontrados na literatura, que foram ainda ultrapassados pelo também proposto sistema de classificação sequencial em 3 estágios. Neste sistema, identificamos um ponto de atenção, que pode apresentar dificuldades em classificar emoções que possuem baixa taxa de reconhecimento em um dos estados, principalmente se um dos estágios iniciais. No nosso caso, especialmente a emoção Nojo apresentou-se como problema para o classificador sequencial. Como alternativa ao sistema de 3 estágios, propusemos um sistema híbrido utilizando verificação e identificação em 3 estágios, para impulsionar o desempenho de determinadas emoções. Este sistema apresentou melhor desempenho que o classificador de estágio único, e no caso masculino pôde ultrapassar até mesmo o classificador sequencial de 3 estágios, apresentando-se como uma boa alternativa em determinados cenários.

Além disso, utilizando GMM's com parâmetros dinâmicos e com um algoritmo LBG modificado, conjuntos ótimos de parâmetros puderam ser obtidos com baixo número de parâmetros, o que torna os sistemas menos vulneráveis a problemas de overfitting.

Por fim, desenvolvemos uma infraestrutura que pode ser reutilizada em trabalhos futuros, para o estudo de outras emoções, outras bases de dados e outros parâmetros de voz. Possíveis trabalhos futuros envolvem a investigação de mais parâmetros para uma melhor classificação entre Raiva e Felicidade no caso feminino e a validação dos algoritmos propostos em bases de dados com emoções reais. Além disso, pode-se pensar no desenvolvimento de um sistema que utiliza as projeções no espaço de emoções de forma simultânea, em vez de sequencial, para contornar erros de classificação em qualquer um dos estágios.

Referências

- [1] SHARMA, R. N. U.; KIM, C. Emotion recognition in spontaneous emotional utterances from movie sequences. In: *WSEAS International Conference on Electronics, Control & Signal Processing*. [S.l.: s.n.], 2002. Citado 2 vezes nas páginas 18 e 30.
- [2] RAZAK, A. A. et al. Comparison between fuzzy and NN method for speech emotion recognition. In: IEEE. *Third International Conference on Information Technology and Applications, 2005*. [S.l.], 2005. v. 1, p. 297–302. Citado 2 vezes nas páginas 18 e 83.
- [3] TÓTH, S. L.; SZTAHÓ, D.; VICSI, K. Speech emotion perception by human and machine. In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. [S.l.]: Springer, 2008. p. 213–224. Citado 4 vezes nas páginas 18, 44, 53 e 83.
- [4] HYUN, K.-H.; KIM, E.-H.; KWAK, Y.-K. Robust speech emotion recognition using log frequency power ratio. In: IEEE. *SICE-ICASE, 2006. International Joint Conference*. [S.l.], 2006. p. 2586–2589. Citado 2 vezes nas páginas 18 e 83.
- [5] HYUN, K. H.; KIM, E. H.; KWAK, Y. K. Improvement of emotion recognition by Bayesian classifier using non-zero-pitch concept. In: IEEE. *IEEE International Workshop on Robot and Human Interactive Communication, 2005*. [S.l.], 2005. p. 312–316. Citado na página 19.
- [6] ILIOU, T.; ANAGNOSTOPOULOS, C.-N. Comparison of different classifiers for emotion recognition. In: IEEE. *13th Panhellenic Conference on Informatics, 2009*. [S.l.], 2009. p. 102–106. Citado 6 vezes nas páginas 19, 49, 61, 83, 89 e 101.
- [7] LIN, Y.-L.; WEI, G. Speech emotion recognition based on HMM and SVM. In: IEEE. *Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005*. [S.l.], 2005. v. 8, p. 4898–4901. Citado 9 vezes nas páginas 19, 44, 49, 51, 53, 56, 83, 89 e 98.
- [8] CAO, M.; FANG, G.; REN, F. EEG-based emotion recognition in Chinese emotional words. In: IEEE. *IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 2011*. [S.l.], 2011. p. 452–456. Citado na página 19.
- [9] CHUANG, Z.-J.; WU, C.-H. Emotion recognition using acoustic features and textual content. In: IEEE. *IEEE International Conference on Multimedia and Expo, 2004*. [S.l.], 2004. v. 1, p. 53–56. Citado na página 19.
- [10] WU, C.-H.; LIANG, W.-B. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, IEEE, v. 2, n. 1, p. 10–21, 2011. Citado na página 19.
- [11] VOGT, T.; ANDRÉ, E. Improving automatic emotion recognition from speech via gender differentiation. In: CITESEER. *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. [S.l.], 2006. Citado na página 19.
- [12] NAKATSU, R.; NICHOLSON, J.; TOSA, N. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-Based Systems*, Elsevier, v. 13, n. 7, p. 497–504, 2000. Citado na página 19.

- [13] SCHULLER, B. et al. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In: IEEE. *IEEE International Conference on Multimedia and Expo, 2008*. [S.l.], 2008. p. 1333–1336. Citado na página 19.
- [14] DARWIN, C. *The expression of the emotions in man and animals*. Oxford: Oxford University, 1998. Citado 2 vezes nas páginas 19 e 30.
- [15] AVERILL, J. R. The acquisition of emotions during adulthood. *The social construction of emotions*, Basil Blackwell Oxford,, England, p. 98–118, 1986. Citado 2 vezes nas páginas 19 e 30.
- [16] WHISSELL, C. et al. A dictionary of affect in language: IV. reliability, validity, and applications. *Perceptual and Motor Skills*, Ammons Scientific, v. 62, n. 3, p. 875–888, 1986. Citado 2 vezes nas páginas 20 e 30.
- [17] DOUGLAS-COWIE, E. et al. Multimodal databases of everyday emotion: facing up to complexity. In: *INTERSPEECH*. [S.l.: s.n.], 2005. p. 813–816. Citado na página 21.
- [18] BÄNZIGER, T.; SCHERER, K. R. Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In: *Affective computing and intelligent interaction*. [S.l.]: Springer, 2007. p. 476–487. Citado na página 22.
- [19] ENGBERG, I. S.; HANSEN, A. V. Documentation of the Danish Emotional Speech Database DES. *Internal AAU report, Center for Person Kommunikation, Denmark*, 1996. Citado na página 22.
- [20] BURKHARDT, F. et al. A database of German emotional speech. In: *Interspeech*. [S.l.: s.n.], 2005. p. 1517–1520. Citado 2 vezes nas páginas 22 e 61.
- [21] SNEDDON, I. et al. The Belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, IEEE, v. 3, n. 1, p. 32–41, 2012. Citado na página 22.
- [22] DOUGLAS-COWIE, E. et al. Emotional speech: Towards a new generation of databases. *Speech communication*, Elsevier, v. 40, n. 1, p. 33–60, 2003. Citado na página 22.
- [23] DEVILLERS, L. et al. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. In: *5th international conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. [S.l.: s.n.], 2006. Citado na página 22.
- [24] IRIYA, R.; RAMÍREZ, M. A. Gaussian mixture models with class-dependent features for speech emotion recognition. In: IEEE. *Statistical Signal Processing Workshop, 2014*. [S.l.], 2014. p. 480–483. Citado na página 24.
- [25] COWIE, R. et al. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, IEEE, v. 18, n. 1, p. 32–80, 2001. Citado 3 vezes nas páginas 27, 36 e 85.
- [26] DAMASIO, A. R. *Looking for Spinoza: Joy, sorrow, and the feeling brain*. New York: Random House, 2004. Citado na página 27.

- [27] HOUWER, J. D.; HERMANS, D. *Cognition and emotion: Reviews of current research and theories*. [S.l.]: New York: Psychology, 2010. Citado na página 27.
- [28] KANDALI, A. B.; ROUTRAY, A.; BASU, T. K. Comparison of features based on MFCCs and eigen values of autocorrelation matrix for cross-lingual vocal emotion recognition in five languages of Assam. In: IEEE. *India Conference (INDICON), 2009 Annual IEEE*. [S.l.], 2009. p. 1–4. Citado 2 vezes nas páginas 30 e 51.
- [29] ORTONY, A.; TURNER, T. J. What's basic about basic emotions? *Psychological review*, American Psychological Association, v. 97, n. 3, p. 315, 1990. Citado na página 32.
- [30] PLUTCHIK, R. *The emotions: Facts, theories, and a new model*. New York: New Random House, 1962. Citado na página 32.
- [31] EKMAN, P.; FRIESEN, W. V.; ELLSWORTH, P. Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press, 1972. Citado na página 32.
- [32] OATLEY, K.; JOHNSON-LAIRD, P. N. Towards a cognitive theory of emotions. *Cognition and emotion*, Taylor & Francis, v. 1, n. 1, p. 29–50, 1987. Citado na página 32.
- [33] WEINER, B.; GRAHAM, S. An attributional approach to emotional development. *Emotions, cognition, and behavior*, p. 167–191, 1984. Citado na página 32.
- [34] PEREIRA, C. Dimensions of emotional meaning in speech. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. [S.l.: s.n.], 2000. Citado 2 vezes nas páginas 32 e 33.
- [35] WU, S.; FALK, T.; CHAN, W.-Y. Automatic recognition of speech emotion using long-term spectro-temporal features. In: IEEE. *16th International Conference on Digital Signal Processing, 2009*. [S.l.], 2009. p. 1–6. Citado na página 36.
- [36] TATO, R. et al. Emotional space improves emotion recognition. In: *INTERSPEECH*. [S.l.: s.n.], 2002. Citado 3 vezes nas páginas 37, 49 e 85.
- [37] LUGGER, M.; YANG, B. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: IEEE. *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008*. [S.l.], 2008. p. 4945–4948. Citado 5 vezes nas páginas 37, 49, 61, 85 e 104.
- [38] XIAO, Z. et al. Hierarchical classification of emotional speech. *IEEE Transactions on Multimedia*, 2007. Citado 5 vezes nas páginas 37, 61, 85, 88 e 103.
- [39] MURRAY, I. R.; ARNOTT, J. L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 93, n. 2, p. 1097–1108, 1993. Citado 2 vezes nas páginas 41 e 89.
- [40] FURUI, S. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In: IEEE. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1986*. [S.l.], 1986. v. 11, p. 1991–1994. Citado na página 44.

- [41] NWE, T. L.; FOO, S. W.; SILVA, L. C. D. Speech emotion recognition using hidden Markov models. *Speech communication*, Elsevier, v. 41, n. 4, p. 603–623, 2003. Citado 5 vezes nas páginas 44, 48, 69, 83 e 101.
- [42] DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, IEEE, v. 28, n. 4, p. 357–366, 1980. Citado 3 vezes nas páginas 44, 48 e 68.
- [43] PIERRE-YVES, O. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, Elsevier, v. 59, n. 1, p. 157–183, 2003. Citado 5 vezes nas páginas 44, 48, 51, 56 e 101.
- [44] FERNANDEZ, R.; PICARD, R. W. Classical and novel discriminant features for affect recognition from speech. In: *Interspeech*. [S.l.: s.n.], 2005. p. 473–476. Citado na página 44.
- [45] SCHULLER, B. et al. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: *INTERSPEECH*. [S.l.: s.n.], 2005. p. 805–808. Citado 6 vezes nas páginas 44, 51, 56, 61, 83 e 101.
- [46] LUGGER, M.; YANG, B. An incremental analysis of different feature groups in speaker independent emotion recognition. *ICPhS, Saarbrücken*, 2007. Citado na página 44.
- [47] BOERSMA, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: AMSTERDAM. *Proceedings of the institute of phonetic sciences*. [S.l.], 1993. v. 17, n. 1193, p. 97–110. Citado 4 vezes nas páginas 45, 63, 65 e 67.
- [48] BOERSMA, P. Praat, a system for doing phonetics by computer. *Glott international*, v. 5, n. 9/10, p. 341–345, 2002. Citado 3 vezes nas páginas 45, 46 e 65.
- [49] BOGERT, B. P.; HEALY, M. J.; TUKEY, J. W. The quefrency alanalysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In: CHAPTER. *Proceedings of the symposium on time series analysis*. [S.l.], 1963. v. 15, p. 209–243. Citado na página 46.
- [50] LIU, H.; MOTODA, H. *Computational methods of feature selection*. Cleveland: CRC, 2007. Citado 2 vezes nas páginas 49 e 85.
- [51] XIAO, Z. et al. Features extraction and selection for emotional speech classification. In: IEEE. *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*. [S.l.], 2005. p. 411–416. Citado na página 49.
- [52] SCHULLER, B.; RIGOLL, G.; LANG, M. Hidden Markov model-based speech emotion recognition. In: IEEE. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003*. [S.l.], 2003. v. 2, p. II–1. Citado 3 vezes nas páginas 50, 51 e 53.
- [53] KIM, E. H. et al. Improved emotion recognition with a novel speaker-independent feature. *IEEE/ASME Transactions on Mechatronics*, IEEE, v. 14, n. 3, p. 317–325, 2009. Citado 3 vezes nas páginas 50, 51 e 56.

- [54] WOMACK, B. D.; HANSEN, J. H. Classification of speech under stress using target driven features. *Speech Communication*, Elsevier, v. 20, n. 1, p. 131–150, 1996. Citado na página 50.
- [55] FERNANDEZ, R.; PICARD, R. W. Modeling drivers speech under stress. *Speech Communication*, Elsevier, v. 40, n. 1, p. 145–159, 2003. Citado 4 vezes nas páginas 51, 53, 56 e 98.
- [56] KWON, O.-W. et al. Emotion recognition by speech signals. In: *INTERSPEECH*. [S.l.: s.n.], 2003. Citado 4 vezes nas páginas 51, 53, 56 e 98.
- [57] FU, L.; MAO, X.; CHEN, L. Speaker independent emotion recognition based on SVM/HMMs fusion system. In: IEEE. *International Conference on Audio, Language and Image Processing, 2008*. [S.l.], 2008. p. 61–65. Citado 3 vezes nas páginas 51, 53 e 56.
- [58] RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, IEEE, v. 77, n. 2, p. 257–286, 1989. Citado 4 vezes nas páginas 52, 76, 78 e 79.
- [59] MOON, T. K. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, IEEE, v. 13, n. 6, p. 47–60, 1996. Citado 2 vezes nas páginas 52 e 71.
- [60] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. Hoboken: John Wiley & Sons, 2012. Citado 3 vezes nas páginas 53, 72 e 81.
- [61] LINDE, Y.; BUZO, A.; GRAY, R. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, IEEE, v. 28, n. 1, p. 84–95, 1980. Citado 3 vezes nas páginas 53, 71 e 73.
- [62] UEDA, N. et al. Split and merge EM algorithm for improving Gaussian mixture density estimates. *Journal of VLSI signal processing systems for signal, image and video technology*, Springer, v. 26, n. 1-2, p. 133–140, 2000. Citado na página 53.
- [63] EVERITT, B. S. et al. Miscellaneous clustering methods. *Cluster Analysis, 5th Edition*, Wiley Online Library, p. 215–255, 2011. Citado na página 56.
- [64] HASSAN, A.; DAMPER, R. I. Emotion recognition from speech using extended feature selection and a simple classifier. In: *INTERSPEECH*. [S.l.: s.n.], 2009. p. 2043–2046. Citado 5 vezes nas páginas 56, 61, 83, 85 e 101.
- [65] KOTTI, M.; PATERNO, F.; KOTROPOULOS, C. Speaker-independent negative emotion recognition. In: IEEE. *2nd International Workshop on Cognitive Information Processing (CIP), 2010*. [S.l.], 2010. p. 417–422. Citado 2 vezes nas páginas 56 e 61.
- [66] BURGESS, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 121–167, 1998. Citado na página 56.
- [67] BERWICK, R. An idiot’s guide to support vector machines (SVMs). *Retrieved on October*, v. 21, p. 2011, 2003. Citado 2 vezes nas páginas 56 e 57.
- [68] SHAMI, M.; VERHELST, W. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, Elsevier, v. 49, n. 3, p. 201–212, 2007. Citado 2 vezes nas páginas 61 e 101.

- [69] PRESS, W. H. et al. *Numerical Recipes in C: The Art of Scientific Computing* (; Cambridge. [S.l.]: Cambridge Univ. Press, 1992. Citado na página 67.
- [70] JR, G. D. F. The Viterbi algorithm. *Proceedings of the IEEE*, IEEE, v. 61, n. 3, p. 268–278, 1973. Citado na página 67.
- [71] CHILDERS, D. G.; KESLER, S. B. *Modern spectrum analysis*. New York: IEEE, 1978. Citado na página 68.
- [72] BÖHM, T.; NÉMETH, G. Algorithm for formant tracking, modification and synthesis. *GYULA SALLAI–president, Scientific Association for Infocommunications ÁKOS DETREKÖI–president, National Council of Hungary for Information and Communications Technology*, p. 15, 2007. Citado na página 68.
- [73] CHANG, C.-C.; LIN, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 2, n. 3, p. 27, 2011. Citado 2 vezes nas páginas 81 e 94.
- [74] STEWART, J. *Multivariable calculus*. Stamford: Cengage Learning, 2011. Citado na página 82.
- [75] LIU, J. et al. Speech emotion recognition based on a fusion of all-class and pairwise-class feature selection. In: *Computational Science–ICCS 2007*. [S.l.]: Springer, 2007. p. 168–175. Citado na página 85.
- [76] KOTTI, M.; KOTROPOULOS, C. Gender classification in two emotional speech databases. In: IEEE. *19th International Conference on Pattern Recognition, 2008*. [S.l.], 2008. p. 1–4. Citado na página 88.