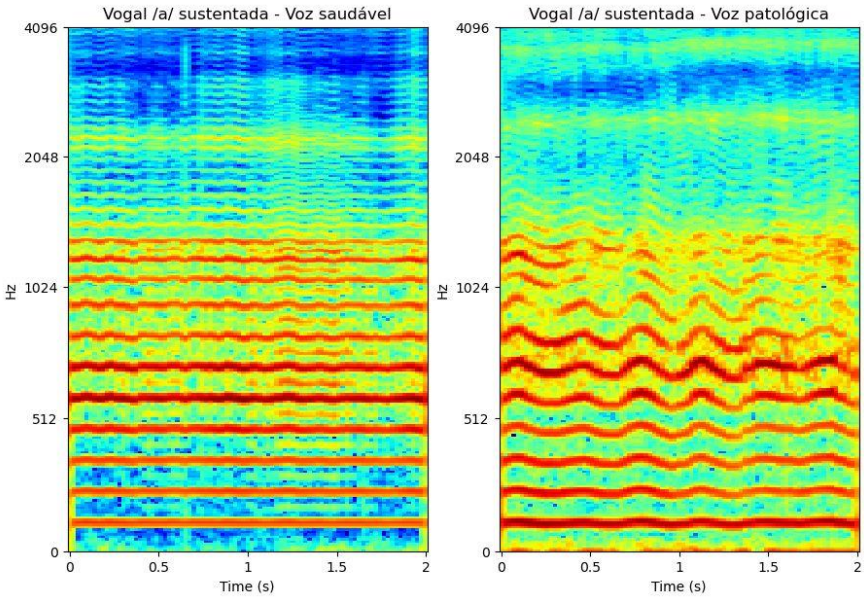


# Rastreio de Vozes Patológicas através de técnicas de Processamento de Fala

BRUNO SÉRGIO ANTUNES RODRIGUES

(Licenciado em Engenharia de Electrónica e Telecomunicações e de Computadores)

Dissertação para obtenção do grau de Mestre em Engenharia de Eletrónica e Telecomunicações, Perfil de Telecomunicações

Orientadores:
Doutor Hugo Tito Cordeiro
Doutor Gonçalo Marques

Júri:
Presidente:
Doutora Paula Maria Garcia Louro
Vogais:
Doutor Hugo Tito Cordeiro
Doutor Carlos Eduardo de Meneses Ribeiro

**Novembro de 2024**

# Rastreio de Vozes Patológicas através de técnicas de Processamento de Fala

BRUNO SÉRGIO ANTUNES RODRIGUES

(Licenciado em Engenharia de Electrónica e Telecomunicações e de Computadores)

Dissertação para obtenção do grau de Mestre em Engenharia de Eletrónica e Telecomunicações, Perfil de Telecomunicações

Orientadores:
Doutor Hugo Tito Cordeiro, ISEL-DEETC
Doutor Gonçalo Marques, ISEL-DEETC

Júri:
Presidente:
Doutora Paula Maria Garcia Louro, ISEL-DEETC
Vogais:
Doutor Hugo Tito Cordeiro, ISEL-DEETC
Doutor Carlos Eduardo de Meneses Ribeiro, ISEL-DEETC

**Novembro de 2024**

# AGRADECIMENTOS

Antes de tudo, agradeço à Divina Providência por me ter permitido realizar e concluir este trabalho. Estou ciente de que me foram proporcionadas a oportunidade e as condições para concluir mais esta etapa da minha vida, e estou muito grato por isso.

A primeira pessoa a quem quero agradecer é à Dora, pelo seu apoio incondicional ao longo deste trajeto. Sacrificou muitas noites, fins de semana e férias em prol deste, e de outros trabalhos meus no ISEL, sem uma única palavra de queixa. Não precisava de qualquer prova, mas certamente que esse apoio inabalável reforça a minha convicção de que foi a melhor pessoa que tive a felicidade de encontrar na vida. Sem ela, nem este trabalho, nem muitas outras coisas teriam sido possíveis.

Quero agradecer às duas pessoas fundamentais para a conclusão deste trabalho: os meus orientadores, Prof. Hugo Cordeiro e Prof. Gonçalo Marques. Agradeço pelo seu incentivo e chamada à razão quando me sentia frustrado com os resultados, o que acontecia sempre que não obtinha taxas de acerto de 90%. Agradeço por me terem posto, algumas vezes, o travão tão necessário quando me perdia a investigar caminhos em vez de documentar o que já tinha feito. Muito trabalho foi feito, e perdido, por não os ter ouvido mais vezes. Agradeço por todas as dúvidas esclarecidas, explicações dadas, indicações, sugestões e muita paciência gasta comigo. Mas principalmente, agradeço porque com eles, nunca senti que isto fosse um trabalho no sentido de ser uma tarefa, obrigação ou chatice. Sempre senti este trabalho (à parte da escrita do relatório, mas isso é outra história…) como uma atividade lúdica, que agora, com muita pena minha, chega ao seu fim.

Quero agradecer aos professores das disciplinas que frequentei, ou apenas assisti (que ainda foram bastantes, no tempo do Zoom). Todos, sem exceção, contribuíram para satisfazer a minha curiosidade e aumentar o meu conhecimento. Quero agradecer especialmente aos professores que me colocaram neste caminho, primeiro Processamento Digital de Sinais, e de Imagem e Biometria, depois *Machine Learning* e finalmente Processamento de Fala. Não era certamente um caminho que imaginasse seguir quando me inscrevi no Mestrado, e se o segui em detrimento de outros, foi devido ao contributo absolutamente decisivo dos professores destas disciplinas.

Quero também agradecer ao ISEL, como entidade. Nunca pensei que, tantos anos depois, o ISEL voltasse a ser um farol na minha vida. Espero, com enorme expectativa, que me tenha guiado a bom porto.

Quero agradecer aos meus pais, por muitas, muitas coisas, mas neste âmbito, quero agradecer à minha mãe pelo seu apoio. Não é fácil apoiar-se uma decisão que não se compreende, mas ela fê-lo sem reservas. Tal como aceitou, também sem reservas, que eu não estivesse tão presente como gostaríamos. Quero agradecer ao meu pai, que se estivesse ainda entre nós, certamente estaria orgulhoso deste trabalho.

Agradeço também à dona Glorinda, à Guida e ao Tó por todas as pequenas e grandes ações diárias, cuja importância foi imensurável para a conclusão deste trabalho.

# Declaração de integridade

Declaro que esta dissertação é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes listadas nas referências bibliográficas foram consultadas e estão devidamente mencionadas no texto. Mais declaro que todas as referências científicas e técnicas relevantes para o desenvolvimento do trabalho estão devidamente citadas e constam das referências bibliográficas.

O autor

Lisboa, 23 de setembro de 2024

“Compreender é uma forma de êxtase.”

Carl Sagan, O Cérebro de Broca

# RESUMO

Em 2012, um inquérito da *American Speech-Language-Hearing Association* (*ASHA*) revelou que um em cada treze norte-americanos sofre anualmente de distúrbios vocais. Extrapolando essa proporção para a população mundial, estima-se que mais de 600 milhões de pessoas possam ser afetadas anualmente. Estes distúrbios impactam negativamente a vida social e profissional, sendo habitualmente diagnosticados através de métodos invasivos e desconfortáveis.

Este trabalho propõe métodos não invasivos, baseados em processamento de sinais de fala, para discriminar entre oradores saudáveis e patológicos, e entre patologias. Utilizaram-se duas bases de dados contendo sinais de fala com a vogal /a/ sustentada, categorizados em quatro classes, sendo uma relativa a oradores saudáveis e as outras a oradores patológicos.

Os sinais de fala foram segmentados em tramas e os seus espectros foram decompostos em 20 bandas de energia. As médias e variações de cada banda ao longo do sinal foram usadas para discriminar entre classes, obtendo-se taxas de acerto médias entre 57,2%, numa das discriminações multiclasse, e 100%, na discriminação entre oradores saudáveis e com condições neurodegenerativas. Observou-se que as bandas correspondentes às primeiras harmónicas têm mais informação para as discriminações, seguidas das bandas relacionadas com o primeiro formante da vogal /a/.

Posteriormente, usaram-se parâmetros espectrais derivados das bandas mais relevantes, isoladamente e em conjunto com os parâmetros acústicos *shimmer* e *HNR*, para discriminar entre classes. A combinação dos parâmetros produziu melhorias estatisticamente significativas nas taxas de acerto médias em 7 das 16 discriminações consideradas. O ganho mais expressivo foi de 10,8%, numa das discriminações multiclasse, e a taxa de acerto média mais elevada foi de 96,8%, numa das discriminações entre oradores saudáveis e com patologias laríngeas fisiológicas. Globalmente, os resultados indicam que a combinação de parâmetros espectrais e acústicos é vantajosa para as discriminações analisadas.

**Palavras-chave:** Patologias de Voz, Parâmetros Espectrais, Parâmetros Acústicos, Discriminação, Sinais de Fala

# ABSTRACT

In 2012, a survey by the *American Speech-Language-Hearing Association* (*ASHA*) revealed that one in thirteen Americans suffers from voice disorders annually. Extrapolating this proportion to the global population, it is estimated that more than 600 million people may be affected annually. These disorders negatively impact social and professional life and are usually diagnosed through invasive and uncomfortable methods.

This study proposes non-invasive methods, based on speech signal processing, to distinguish between healthy and pathological speakers, as well as between different pathologies. Two datasets containing speech signals of the sustained vowel /a/ were used, categorized into four classes: one corresponding to healthy speakers and the others to pathological speakers.

The speech signals were segmented into frames and the frame spectra were decomposed into 20 energy bands. The average and variation values of each band across the signals were used to classify the samples, achieving average accuracy rates ranging from 57.2%, in one of the multiclass discriminations, to 100%, in the discrimination between healthy speakers and those with neurodegenerative conditions. Bands corresponding to the first harmonics were found to be the most informative for classification, followed by bands associated with the first formant of the vowel /a/.

Subsequently, spectral parameters derived from the most relevant bands were used both independently and combined with acoustic parameters *shimmer* and *HNR* to classify the samples. The combination of parameters led to statistically significant improvements in average accuracy rates in 7 out of the 16 classifications considered. The most notable gain was 10.8% in one of the multiclass discriminations, and the highest average accuracy rate was 96.8% in the discrimination between healthy speakers and those with physiological laryngeal pathologies. Overall, the results indicate that combining spectral and acoustic parameters is advantageous for the analyzed classifications.

**Keywords:** Voice Pathologies, Spectral Features, Acoustic Parameters, Discrimination, Speech Signals

# ÍNDICE

# LISTA DE FIGURAS

# LISTA DE TABELAS

# ACRÓNIMOS

| | |
|---|---|
| **ACC** | Accuracy |
| **ANN** | Artificial Neural Networks |
| **APQ** | Arithmetic Average of the Perturbation Quotients |
| **ASHA** | American Speech-Language-Hearing Association |
| **AUC** | Area Under the Curve |
| **bbLBST** | band-based Low Band Spectral Tilt |
| **bbLBSvT** | band-based Low Band Spectral variation Tilt |
| **BSTA** | Band Spectral Tilt Angle |
| **CAPE-V** | Consensus Auditory-Perceptual Evaluation of Voice |
| **CHVNGE** | Centro Hospitalar de Vila Nova de Gaia/Espinho |
| **CMED** | Collected and Multiple Existing Dataset |
| **ELA** | Esclerose Lateral Amiotrófica |
| **FBME** | First Band Maximum Energy |
| **FEM** | Fuzzy Entropy (Modified) |
| **FFT** | Fast Fourier Transform |
| **FN** | False Negative |
| **FP** | False Positive |
| **GMM** | Gaussian Mixture Models |
| **GNB** | Gaussian Naive Bayes |
| **GRBAS** | Grade, Roughness, Breathiness, Asthenia, Strain |
| **HBST** | High Band Spectral Tilt |
| **HC-FMUSP** | Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo |
| **HNR** | Harmonics-to-Noise Ratio |
| **HUPA** | Hospital Universitário Príncipe das Astúrias |
| **KNN** | K-Nearest Neighbors |
| **LBST** | Low Band Spectral Tilt |
| **LDA** | Linear Discriminant Analysis |
| **LPC** | Linear Predictive Coding |
| **MEEI** | Massachusetts Eye and Ear Infirmary |

| | |
|---|---|
| **MFCC** | Mel-Frequency Cepstral Coefficients |
| **NHIS** | National Health Interview Survey |
| **OvA** | One vs. All |
| **OvO** | One vs. One |
| **PCA** | Principal Component Analysis |
| **PCM** | Pulse Code Modulation |
| **PLP** | Perceptual Linear Predictive |
| **PPQ** | Pitch Perturbation Quotient |
| **PPV** | Predictive Positive Value |
| **QCP** | Quasi-Closed Phased |
| **RAP** | Relative Average Perturbation |
| **RBF** | Radial Basis Function |
| **RPPC** | Relative Power of the Periodic Component |
| **SBME** | Second Band Maximum Energy |
| **SGD** | Stochastic Gradient Descent |
| **SHDB** | Shimmer [dB] |
| **sMEEI** | subconjunto da base de dados do Massachusetts Eye and Ear Infirmary |
| **SNS** | Serviço Nacional de Saúde |
| **STFT** | Short Time Fourier Transform |
| **SVC** | Support Vector Classifier |
| **SVD** | Saarbrücken Voice Database |
| **SVM** | Support Vector Machines |
| **TBME** | Third Band Maximum Energy |
| **TN** | True Negative |
| **TP** | True Positive |
| **TPR** | True Positive Rate |
| **UFS** | Univariate Feature Selection |
| **USP** | Universidade de São Paulo |
| **UVFP** | Unilateral Vocal Fold Paralysis |
| **XGB** | Extreme Boosting Gradient |
| **ZCR** | Zero Crossing Rate |
| **ZFF** | Zero Frequency Filter |

# 1

# INTRODUÇÃO

A linguagem humana é um sistema de comunicação exclusivamente associado à nossa espécie, constituindo a base primordial da nossa capacidade comunicativa e uma das principais características que nos distinguem dos outros seres vivos. A sua importância é tão significativa que suscita a discussão sobre se o seu papel foi mais preponderante na evolução recente da nossa espécie, nos últimos 200.000 anos, do que o papel dos nossos genes [1]. A linguagem humana permite-nos transcender os limites da comunicação simples, capacitando-nos a partilhar as nossas ideias mais complexas, expressar os nossos pensamentos mais profundos e revelar os nossos sentimentos mais íntimos. Esta habilidade única possibilita-nos interagir com as pessoas à nossa volta de uma maneira singular e exclusiva, enriquecendo a experiência humana de forma inigualável.

A primeira forma de linguagem humana a emergir foi a fala, permanecendo como principal meio de comunicação, o mais utilizado e relevante até aos dias de hoje. Para além da simples transmissão e receção de sons, a fala é um processo sofisticado que depende de um conjunto complexo de mecanismos e das suas interações. Estes mecanismos incluem o aparelho fonador, que abrange a produção física dos sons da fala, o sistema auditivo/perceptual, que envolve a capacidade de ouvir e compreender os sons da fala, e o sistema nervoso central, responsável pelo processamento e interpretação das informações linguísticas [2].

A ferramenta fundamental para a fala é a voz e a sua importância ultrapassa a simples produção de sons. Para além da informação verbal relativa ao significado das palavras e frases, conhecida como informação linguística, a voz humana contém também informação paralinguística. Na voz, esta informação é obtida a partir de aspetos não verbais como o volume e o tom, e é tão importante para a comunicação oral como a informação linguística, pois complementa-a de uma forma crucial [3]. Além dessas informações, a voz também contém outra, intrínseca ao orador e, portanto, não controlada por ele: conhecida por informação extralinguística. Esta informação está presente na qualidade de voz, sendo possível, através da sua análise, identificar aspetos relacionados com a idade, género, estado emocional e até mesmo com o estado de saúde do orador [4].

A definição da qualidade de voz ainda não é consensual. Vários estudos tais como [5], [6] e [7] têm tentado estabelecer uma definição única e universalmente aceite, sem sucesso. Uma definição amplamente aceite e utilizada, com mais ou menos variações, define a qualidade de voz como a

percepção auditiva dos parâmetros vocais que caracterizam e identificam a voz de um orador [8]. Esta definição caracteriza a qualidade de voz como um atributo psicoacústico que depende tanto do sinal, como do ouvinte, ou seja, como um atributo subjetivo e não de uma medida física objetiva.

A qualidade de voz, ou a sua degradação, pode indiciar um distúrbio da voz como disfonia, rouquidão ou fadiga vocal e pode ter impactos significativos em várias áreas da vida [9]. Em termos de comunicação, os distúrbios da voz podem resultar em dificuldades de entendimento e interpretação podendo levar a mal-entendidos, falta de clareza e até mesmo frustração durante interações sociais, profissionais e pessoais. Pode também afetar a autoconfiança e autoestima, e aumentar os níveis de stress, ansiedade e depressão, especialmente se o orador se sentir incompreendido ou julgado com base na sua voz [10] [11]. No contexto profissional, os distúrbios da voz podem dificultar uma comunicação eficaz e persuasiva, prejudicando oportunidades de emprego, promoções e sucesso na carreira. Para profissionais que dependem da voz para trabalhar, como professores e cantores, esses distúrbios podem ameaçar seriamente a continuidade e o desenvolvimento das suas carreiras [12] [13].

Em casos mais graves, os distúrbios da voz podem ser causados por problemas de saúde como tumores malignos ou condições neurológicas, com consequências potencialmente fatais [14].

## 1.1 Motivações do Estudo

Vários estudos foram realizados com o objetivo de investigar a prevalência dos distúrbios de voz nas populações, abrangendo diferentes grupos etários, profissionais, geográficos e socioeconómicos. Entre os estudos efetuados em amostras representativas da população em geral, um destaca-se pelo seu resultado expressivo onde, entre 497 participantes adultos escolhidos aleatoriamente, 38,5% deles revelaram já ter sofrido um distúrbio vocal nalgum momento das suas vidas [15]. Alguns trabalhos foram feitos sobre os dados obtidos no inquérito americano *National Health Interview Survey* de 2012[1], de modo a aproveitar amostras mais significativas. As análises feitas aos resultados do inquérito revelaram que 7,6% dos inquiridos reportou ter tido um distúrbio de voz nos 12 meses anteriores ao inquérito [16], enquanto 4,0% dos inquiridos reportou ter tido um problema na voz, no mesmo período, com uma duração igual ou superior a uma semana [17].

Quando extrapolados para a população adulta residente em Portugal, estes resultados indicam uma estimativa de aproximadamente 650.000 pessoas afetadas por distúrbios vocais por ano, sendo que dessas, 350.000 apresentam distúrbios com duração igual ou superior a uma semana. Considerando a natureza funcional da condição, que influencia a autopercepção da mesma, estes valores aproximados podem pecar por defeito, pois indivíduos que padeçam de distúrbios de voz que não condicionem as suas vidas pessoais, profissionais e sociais podem não se considerar doentes [18]. Estas estimativas indiciam que os distúrbios da voz têm uma prevalência considerável na população.

[1] O *National Health Interview Survey* (NHIS) é um inquérito realizado anualmente nos Estados Unidos através de entrevista efetuada ao domicílio e aborda uma vasta gama de assuntos relacionados com a saúde. Na pesquisa de 2012 foram inquiridos 34525 adultos. Mais informação em https://www.cdc.gov/nchs/nhis/index.htm

### 1.1.1 Diagnóstico de distúrbios de voz

O diagnóstico de distúrbios de voz é um processo complexo que requer frequentemente uma análise multidimensional realizada por equipas multidisciplinares. A falta de um método único, abrangente e preciso para o diagnóstico torna necessário o trabalho conjunto de profissionais de diferentes áreas, como otorrinolaringologia, neurologia e terapia da fala. O diagnóstico pode envolver vários procedimentos, tais como entrevistas detalhadas, avaliações acústicas, perceptivas e da fisiologia laríngea. Este processo exige um gasto acrescido em termos de recursos humanos e tempo, bem como em termos da articulação entre os diferentes profissionais envolvidos [14].

A análise acústica consiste na quantificação do sinal de voz através da determinação dos parâmetros acústicos que o compõem. Os distúrbios de voz são geralmente caracterizados por instabilidade vocal e a análise acústica fornece medidas objetivas das caraterísticas do sinal de voz. No entanto, a sua avaliação clínica pode ser um desafio, pois esses parâmetros são muito relevantes na identificação de algumas patologias, mas não de todas [19].

A entrevista detalhada tem por objetivo avaliar os sintomas, histórico de saúde, hábitos vocais e outros factores, sociais, profissionais e emocionais, que possam ser relevantes para a perturbação na voz. Embora essas informações sejam importantes, não são geralmente suficientes para um diagnóstico completo. Já a análise perceptiva consiste na avaliação da voz por parte de um especialista comparando-a a referências adquiridas, durante a sua formação e prática profissional, de vozes categorizadas como normais, podendo até essa avaliação ser quantificada através de escalas padronizadas[2]. Embora este método seja muito utilizado em ambiente clínico por ser prático e acessível, tem algumas limitações devido à sua subjetividade, pois depende das referências do especialista, já que não existe uma definição universal de voz normal, e da sua qualificação, influenciada pela sua experiência [20].

Em suma, estes procedimentos, embora importantes, geralmente não permitem, por si só, um diagnóstico completo.

### 1.1.2 Diagnóstico de patologias laríngeas fisiológicas e neurológicas

Quando o distúrbio de voz é causado por uma patologia laríngea fisiológica, ou seja, por uma alteração fisiológica da laringe, o diagnóstico é realizado por um otorrinolaringologista através da avaliação da fisionomia laríngea. Esta avaliação envolve diferentes métodos, como a laringoscopia indireta, a endoscopia e a estroboscopia. Na laringoscopia indireta, a laringe é visualizada através da introdução de um espelho curvo na orofaringe do paciente, enquanto a língua é imobilizada para uma melhor visualização. A endoscopia rígida consiste na utilização de um tubo rígido com uma câmara

[2] São usadas, entre outras, as escalas GRBAS (*Grade, Roughness, Breathiness, Asthenia, Strain*) proposta por Hirano em 1981, ou CAPE-V (*Consensus Auditory-Perceptual Evaluation of Voice*) proposta pela *American Speech-Language-Hearing Association* (ASHA) em 2002. A utilização de diferentes escalas por especialistas constitui uma limitação da análise perceptiva, pois cada escala tem critérios de avaliação e pontuações diferentes, tornando difícil a comparação direta dos resultados entre diferentes avaliadores.

iluminada na extremidade (endoscópio) para visualização da laringe, sendo o endoscópio introduzido por via oral, enquanto a endoscopia flexível, ou nasofibrolaringoscopia, utiliza um endoscópio flexível introduzido por via nasal. A estroboscopia, que utiliza também a endoscopia, explora o fenómeno da persistência da visão, permitindo a observação da vibração das pregas vocais através de pulsos de luz sincronizados [21].

Apesar de eficazes para o diagnóstico, todos estes métodos são invasivos e desconfortáveis para o paciente, obrigando muitas vezes ao uso de anestesia para a sua realização, aumentando o tempo de recuperação e os riscos do procedimento.

A avaliação da fisionomia laríngea através dos métodos descritos é efetuada por pessoal médico especializado, com instrumentos e equipamentos próprios, em instalações médico-hospitalares. De acordo com dados disponibilizados pelo Centro Hospitalar de Vila Nova de Gaia/Espinho[3], em março de 2023 existiam 971 pacientes em lista de espera para uma primeira consulta de Otorrinolaringologia, estimando-se um tempo médio de espera de 128 dias. Este longo tempo de espera evidencia uma dificuldade mais geral relativa ao acesso a serviços de saúde. Este é um problema atual em Portugal, tendo sido o terceiro assunto mais discutido na última campanha eleitoral, demostrando a sua gravidade e a preocupação que gera na população [22].

Portanto, o acesso a estes procedimentos, tal como a outros serviços de saúde, não é imediato, podendo apresentar tempos de espera bastante longos.

Alguns distúrbios da voz têm origem em patologias laríngeas neurológicas, ou seja, são causados por doenças neurodegenerativas que afetam progressivamente o sistema nervoso, como a doença de Parkinson ou a esclerose lateral amiotrófica (*ELA*), entre outras. Nessas condições, as alterações neurológicas podem comprometer o controlo dos músculos da laringe responsáveis pela produção vocal, resultando em dificuldades na fonação, alterações na qualidade da voz e até mesmo na perda completa da capacidade vocal [23]. Os distúrbios de voz constituem frequentemente o primeiro sintoma de uma doença neurodegenerativa, e embora essas condições sejam incuráveis, existem tratamentos que podem ajudar a retardar a progressão da doença e melhorar a qualidade de vida do paciente [24].

Assim, um diagnóstico precoce dessas condições a partir de um distúrbio de voz pode ser de extrema importância para fornecer cuidados adequados e intervenções terapêuticas que possam ajudar a mitigar os sintomas e melhorar o bem-estar do paciente.

[3] O Centro Hospitalar de Vila Nova de Gaia/Espinho (CHVNGE) é uma das instituições do Serviço Nacional de Saúde (SNS) que disponibiliza consultas da especialidade de Otorrinolaringologia. Embora seja possível consultar na página do SNS os tempos de espera para consultas nas várias especialidades, incluindo Otorrinolaringologia, a obtenção e visualização desses dados não são intuitivas. O CHVNGE disponibilizou uma tabela, com dados referentes aos primeiros dois meses de 2023, utilizados aqui como exemplo. Esta tabela pode ser acedida em https://www.chvnge.min-saude.pt/files/share/FILE_20230316165328344254.pdf

### 1.1.3 Sistema de diagnóstico de patologia de voz

O desenvolvimento de um sistema automático de reconhecimento de patologias vocais, baseado em técnicas avançadas de processamento de fala e algoritmos de *Machine Learning*, tem o potencial de servir como uma ferramenta de triagem ou complementar ao diagnóstico médico. Este avanço poderia ter um impacto significativo na saúde pública e na prática clínica, ajudando a mitigar ou resolver os desafios previamente mencionados, nomeadamente:

- **Prevalência dos distúrbios de voz:** A existência de um sistema automático de reconhecimento de patologias vocais seria justificada pela prevalência desses distúrbios na população, que pode ser significativa conforme visto em [15], [16] e [17].
- **Necessidade de equipas multidisciplinares:** Um sistema capaz de identificar patologias vocais com precisão poderia dispensar a necessidade de equipas multidisciplinares para diagnóstico, o que simplificaria o processo de avaliação e permitiria uma abordagem mais direcionada para tratamento.
- **Necessidade de avaliações adicionais**: Um sistema suficientemente abrangente e preciso poderia dispensar algumas das avaliações tradicionalmente necessárias, poupando tempo e recursos, tanto para profissionais de saúde, como para pacientes.
- **Necessidade de avaliação da fisionomia laríngea:** Um sistema baseado em análises de sinais de fala que pudesse substituir avaliações da fisionomia laríngea traria benefícios para o paciente, pois seria não invasivo e inócuo, eliminando o desconforto físico, riscos e tempos de recuperação associados a procedimentos invasivos.
- **Acesso demorado:** Um sistema massificado, disponibilizado em computadores ou telemóveis, tornaria o seu acesso potencialmente universal, permitindo um diagnóstico imediato quando fosse necessário.
- **Condições neurológicas:** Um sistema que identifique distúrbios de voz como sintomas de condições neurológicas subjacentes permitiria tratamentos numa fase mais inicial, melhorando a qualidade de vida do paciente e atrasando a progressão da doença.

Em resumo, o desenvolvimento e a implementação de um sistema automático de reconhecimento de patologias vocais poderiam trazer uma série de benefícios significativos, desde a simplificação do processo de diagnóstico, até ao fornecimento de cuidados de saúde mais atempados.

## 1.2 Objetivos do estudo

Para o desenvolvimento de um sistema automático de rastreio de vozes patológicas e de identificação de patologias laríngeas, será crucial a aquisição de um grande conhecimento acerca dos sinais de fala e das suas caraterísticas distintivas. Essa compreensão profunda será fundamental para garantir a eficácia do sistema. Nesse contexto, este estudo propõe-se contribuir significativamente para a acumulação desses conhecimentos através da exploração das seguintes propostas:

- Investigar a contribuição, caso exista, das bandas de energia dos sinais de fala para diferenciar vozes saudáveis e patológicas e identificar patologias laríngeas.
- Avaliar o impacto da combinação da informação relativa às bandas de energia mais relevantes e de alguns parâmetros acústicos na distinção entre vozes saudáveis e patológicas, assim como entre diferentes patologias laríngeas.
- Consolidar a validade dos potenciais resultados obtidos através da utilização de dois conjuntos de dados distintos.

Espera-se que a exploração das metas propostas possa constituir mais um passo rumo ao objetivo de desenvolvimento do sistema acima idealizado.

## 1.3 Estrutura do trabalho

A estrutura do trabalho é apresentada de seguida, na Tabela 1-1, onde são referidos os capítulos que compõem o trabalho e as suas descrições resumidas.

Tabela 1-1 - Estrutura do trabalho

| **Capítulos** | **Descrição** |
|---|---|
| 1. Introdução | Apresentação das motivações para o estudo.<br>Objetivos propostos para o trabalho. |
| 2. Enquadramento e Estado da Arte | Apresentação dos temas abordados neste estudo.<br>Apresentação do trabalho relacionado com este estudo. |
| 3. Materiais e Métodos | Descrição dos conjuntos de dados, ferramentas e métricas utilizadas.<br>Descrição e fundamentação das metodologias aplicadas. |
| 4. Bandas de Energia | Investigação das bandas de energia nas discriminações entre classes.<br>Apresentação, interpretação e discussão dos resultados obtidos. |
| 5. Combinação com Parâmetros Acústicos | Investigação da introdução de parâmetros acústicos nas discriminações.<br>Apresentação, interpretação e discussão dos resultados obtidos. |
| 6. Conclusões | Conhecimentos adquiridos, limitações e propostas de trabalho futuro.<br>Divulgação realizada. |

A Tabela 1-1 apresenta a estrutura detalhada do relatório, organizando os capítulos e as suas respetivas descrições. Esta organização visa proporcionar uma visão clara e sequencial do estudo, desde a introdução das motivações e objetivos até à análise detalhada dos resultados e conclusões finais. Esta estrutura tenta facilitar a compreensão do desenvolvimento do trabalho e a integração dos diversos tópicos abordados.

# 2

# ENQUADRAMENTO E ESTADO DA ARTE

Neste capítulo apresentam-se os temas fundamentais nos quais se enquadra este estudo. Analisa-se o processo de produção de fala humana, nomeadamente o sistema fisiológico responsável por essa função, a importância das pregas vocais nesse processo, bem como algumas patologias que afetam esse órgão. São também analisados os sinais de fala produzidos, bem como os parâmetros acústicos e espectrais que os caracterizam. Finalmente é abordado e analisado o trabalho realizado com vista à diferenciação entre vozes saudáveis e patológicas baseado no estudo dos parâmetros espectrais e/ou acústicos dos sinais de fala.

## 2.1 Sistema Respiratório

O sistema, ou aparelho respiratório desempenha várias funções, destacando-se duas em particular. A primeira função consiste em fornecer a ventilação necessária para a troca gasosa entre o dióxido de carbono, um resíduo metabólico, e o oxigénio, fundamental para a sobrevivência e funcionamento saudável do organismo, destacando-se pela sua vital importância. A segunda função destaca-se pela sua relevância no contexto deste trabalho e consiste em adquirir, armazenar e fornecer o ar necessário para a produção dos sinais de fala, além de dar suporte aos sistemas, ou aparelhos responsáveis por essa produção.

De acordo com [20], o sistema respiratório humano é um conjunto de órgãos e estruturas que permitem a entrada e a saída de ar do corpo humano e é composto por duas partes principais: o trato respiratório e as estruturas musculoesqueléticas.

As estruturas musculoesqueléticas, tais como o diafragma e os músculos intercostais, oferecem suporte ao trato respiratório e garantem o movimento do ar durante a respiração. O diafragma é um músculo situado entre o tórax e o abdómen, que se contrai durante a inspiração, expandindo a cavidade torácica e criando uma pressão negativa relativamente ao ambiente exterior, que puxa o ar para dentro dos pulmões. Na expiração, o diafragma relaxa, deixando a cavidade torácica colapsar e comprimir-se, expulsando o ar dos pulmões. Os músculos intercostais situam-se entre as costelas e auxiliam o diafragma na respiração, expandindo e comprimindo a cavidade torácica [20].

### 2.1.1 Trato respiratório

O trato respiratório, representado na Figura 2-1, é um conjunto de vias aéreas que permite a circulação do ar desde o exterior até aos alvéolos pulmonares, situados nos pulmões, onde é efetuada a troca gasosa. O trato respiratório pode ser, e será conveniente fazê-lo no contexto deste trabalho, dividido em trato respiratório superior e inferior, conforme definidos em [20] e [25].

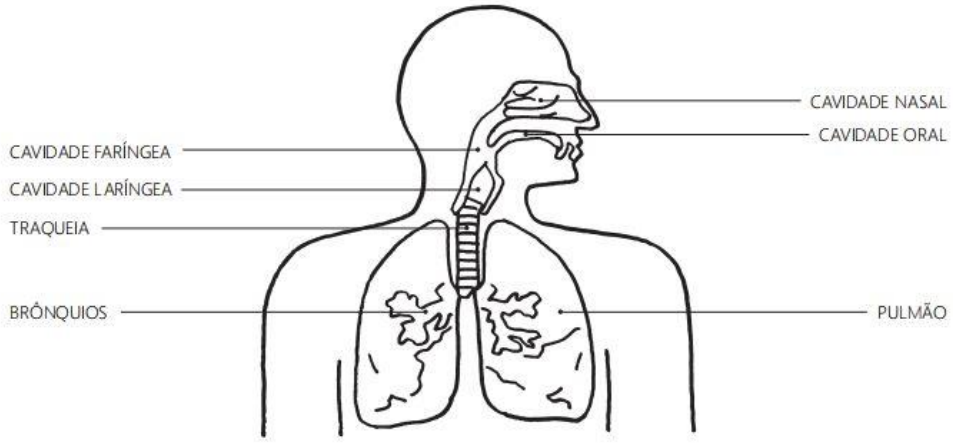


Figura 2-1 - Trato respiratório (retirado de [20])

O trato respiratório inferior, ou árvore traqueobrônquica, é constituído pela traqueia, pelos brônquios e pelos pulmões. A traqueia e os brônquios são formados por anéis cartilaginosos rígidos, de modo a oferecer suporte estrutural e impedir o colapso das vias respiratórias durante a inspiração. São anéis incompletos, em forma de 'C', e revestidos por músculo liso, oferecendo a flexibilidade necessária para a modificação do diâmetro das vias respiratórias, consoante as necessidades respiratórias. Os pulmões são órgãos moles e elásticos situados na cavidade torácica, responsáveis pela respiração. Não possuem músculos próprios, sendo auxiliados pelos músculos respiratórios, como o diafragma e os intercostais, na realização dos movimentos inspiratórios e expiratórios. Nos pulmões encontram-se os alvéolos pulmonares, pequenos sacos que se enchem de ar durante a inspiração e o expelem durante a expiração. Os alvéolos são os principais locais de troca gasosa no sistema respiratório, onde ocorre a captação de oxigénio e a libertação de dióxido de carbono.

O trato respiratório superior é constituído pelas cavidades nasais, oral, faríngea e laríngea, e desempenha funções fundamentais nos processos respiratório, de mastigação e deglutição, bem como nos processos de fonação, ressonância e articulação relacionados com a produção de sinais de fala. As cavidades nasais e oral são as vias de entrada e saída de ar no sistema respiratório, estando separadas pelo palato, duro e mole. No processo de respiração, as cavidades, ou fossas nasais têm a função de aquecer, humidificar e filtrar o ar inspirado, tornando-o mais adequado para o trato respiratório inferior. A cavidade faríngea é um tubo muscular em forma de funil que se estende das aberturas posteriores do nariz até à parte superior do esófago e da laringe, servindo como uma

passagem partilhada entre os aparelhos digestivo e respiratório. As paredes curvas superiores e posteriores estão ligadas ao esqueleto axial, mas as paredes laterais são flexíveis e musculares de modo a auxiliar no processo de deglutição [20] [25].

### 2.1.2 Laringe

A cavidade laríngea pode ser considerada parte do trato respiratório superior ou inferior, consoante o autor e o contexto. No âmbito deste trabalho, será considerada a fronteira entre os dois tratos respiratórios, fazendo a ligação entre eles.

A cavidade laríngea é um tubo cartilaginoso compreendido entre a faringe e a traqueia, sendo as cartilagens que a compõem interligadas por ligamentos e revestidas por membrana mucosa, fazendo parte da anatomia da laringe. A sua principal função é impedir que alimentos ou líquidos entrem no trato respiratório inferior, mas desempenha outras funções importantes, entre as quais se destaca a produção de voz, ou fonação.

A laringe é um conjunto de cartilagens, músculos, membranas e ligamentos, que lhe conferem a flexibilidade necessária para desempenhar as suas funções nos processos de deglutição, respiração e fonação. Uma representação desta estrutura pode ser visualizada na Figura 2-2, adaptada de [26].

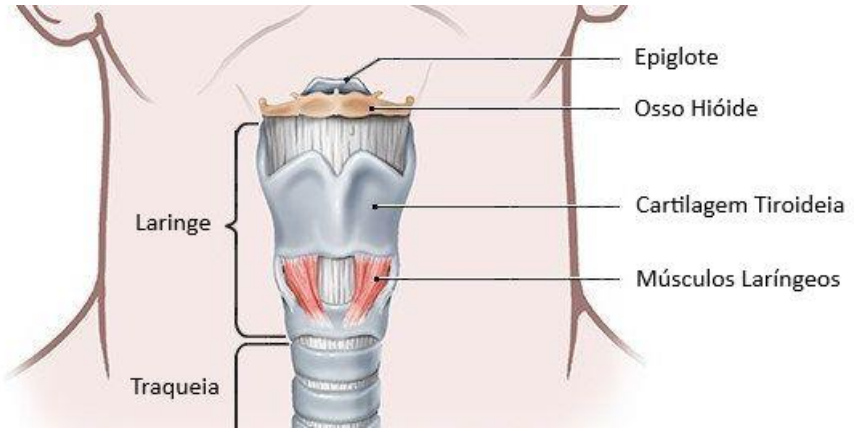


Figura 2-2 - Laringe (adaptado de [26])

A epiglote é uma pequena cartilagem em forma de folha, situada na parte superior da laringe, e tem como principal função a proteção das vias respiratórias contra a entrada de elementos estranhos. Em repouso, a epiglote mantém-se numa posição quase vertical, mas durante a deglutição, é empurrada para baixo pela língua, fechando a entrada da laringe e direcionando o alimento ou líquido para o esófago. O osso hióide é um osso em forma de 'U' situado na parte anterior do pescoço, entre a mandíbula e a laringe e desempenha funções importantes na deglutição e na fala. Durante esses processos, o osso hióide movimenta-se, ajudando a abrir o trato digestivo superior durante a deglutição, e alterando o formato do trato respiratório superior para a articulação dos sons da fala. Além disso, serve como ponto de ancoragem para, entre outros, os músculos da faringe e da laringe, sendo essencial para que estes músculos desempenhem as suas funções.

Pode ser visualizada na Figura 2-2 a cartilagem tiroideia, que é a maior de todas as cartilagens existentes na laringe. É uma estrutura em forma de 'V', composta por quatro placas cartilaginosas, duas planas, situadas lateralmente, e duas posteriores, em forma de haste. As duas placas posteriores servem como pontos de ligação da cartilagem, estando ligadas ao osso hióide em cima. As duas placas laterais juntam-se num ângulo que é tipicamente diferente consoante o género, sendo mais apertado nos homens. Essa diferença faz com que a junção, designada como proeminência laríngea ou tiroideia, seja mais saliente nos homens, tornando-a visível e vulgarmente conhecida como maçã de Adão. O ângulo da junção influencia também a voz, fazendo com que seja diferente em homens e mulheres.

Os músculos da laringe dividem-se em extrínsecos, se têm ligações na laringe e em estruturas externas, e intrínsecos, se têm ligações unicamente na laringe. Os músculos extrínsecos têm como função a suspensão e a movimentação da laringe, enquanto os intrínsecos desempenham funções tais como a contenção do ar abaixo da laringe, a prevenção da entrada de corpos estranhos no trato respiratório inferior e a produção de fonação. Estas funções são realizadas através da atuação dos músculos sobre as pregas vocais. Uma descrição mais detalhada da faringe pode ser encontrada em [20], [27] e [28], fontes onde foi baseada a descrição dada neste subcapítulo.

### 2.1.3 Pregas vocais

As pregas vocais são estruturas anatómicas fundamentais para o processo de produção de fala. Estão situadas no interior da laringe, estendendo-se horizontalmente ao longo desta e fixando-se anteriormente à face interna da cartilagem tiroideia, sendo essa a região de convergência das duas pregas vocais. Na Figura 2-3 está representada uma prega vocal vista transversalmente.

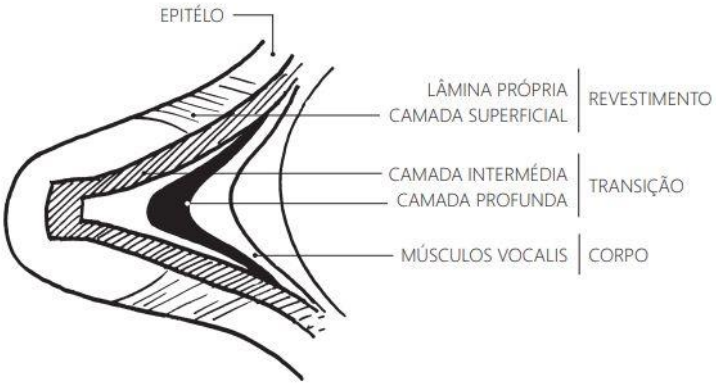


Figura 2-3 - Vista transversal da prega vocal (retirado de [20])

As pregas vocais são estruturas multilaminadas, com um formato que se assemelha a um triângulo quando vistas transversalmente e são constituídas essencialmente por músculo e membrana mucosa. A mucosa divide-se em epitélio e lâmina própria, sendo o epitélio uma camada resistente, não elástica e tendo como objetivo cobrir e manter a forma da prega vocal. A lâmina própria por sua vez divide-se em três camadas denominadas superficial, intermédia e profunda. A camada superficial da

lâmina própria, também designada como espaço de Reinke, é a camada mais solta e flexível das três, sendo constituída por uma rede de fibras de elastina e colagénio, assemelhando-se a uma gelatina. Esta camada é a que mais vibra durante o processo de fonação. Por baixo da lâmina própria está o músculo vocal, que constitui o corpo da prega vocal. As fibras musculares, de elastina e de colagénio estão dispostas paralelamente à borda livre da prega vocal.

As pregas vocais desempenham várias funções vitais no trato vocal e respiratório. Entre as principais, destacam-se a regulação do fluxo de ar durante os processos de respiração e fala, a proteção do trato respiratório inferior contra a entrada de corpos estranhos e o controlo da pressão através do encerramento da laringe, necessário em situações de esforço. No entanto, no contexto deste trabalho, o papel das pregas no processo de fonação é o mais relevante, pois é com a vibração das pregas vocais, impulsionada pelo fluxo de ar, que a voz é produzida. Na Figura 2-4, adaptada de [29], estão representadas as pregas vocais, abertas e fechadas, juntamente com algumas estruturas relevantes para o processo de fonação.

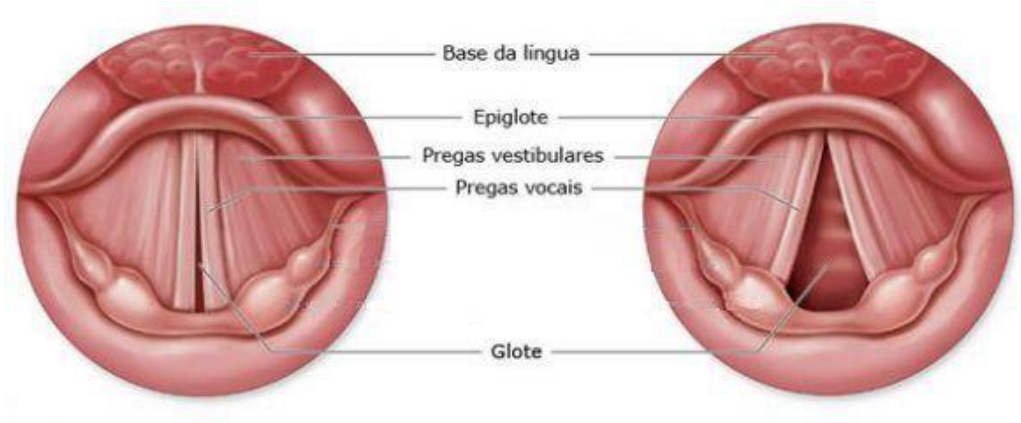


Figura 2-4 - Pregas vocais: fechadas (esq.) e abertas (dir.) (adaptado de [29])

As pregas vestibulares, também denominadas como pregas ventriculares ou falsas pregas vocais, estão localizadas acima das pregas vocais. São dois tecidos espessos e moles, constituídos por ligamentos e revestidas por mucosa. A sua principal função é a de auxiliar as pregas vocais tanto na proteção das vias respiratórias durante a deglutição, como a manter a pressão no trato respiratório inferior. Normalmente não participam no processo de fonação, ficando lateralizadas relativamente às pregas vocais durante este processo, mas podem desempenhar um papel indireto na modulação da voz em algumas situações específicas.

A glote é uma válvula formada pelo espaço horizontal entre as pregas vocais, sendo a sua abertura e fecho controlada por estas. Durante a respiração em repouso, a glote é uma abertura estreita, mas em situações de respiração forçada, ela torna-se numa abertura triangular devido ao afastamento das pregas vocais. Durante o processo de fonação, a glote abre e fecha como resultado da vibração das cordas vocais. A descrição dada das pregas vocais, da sua anatomia e funcionamento foi baseada em [20], [30] e [31], onde as pregas vocais são descritas com maior pormenor.

## 2.2 Fala

### 2.2.1 Produção de fala

No contexto da produção de fala, os órgãos e tecidos do aparelho respiratório, apresentados na subsecção anterior, são agora analisados em relação ao seu funcionamento no aparelho fonador humano. O aparelho fonador humano não é uma unidade anatómica, mas sim o conjunto de estruturas interligadas e coordenadas que, em conjunto, produzem a fala. Na Figura 2-5, adaptada de [32], representa-se o aparelho fonador humano, bem como os principais elementos que o compõem agrupados por sistemas de acordo com as suas funções.

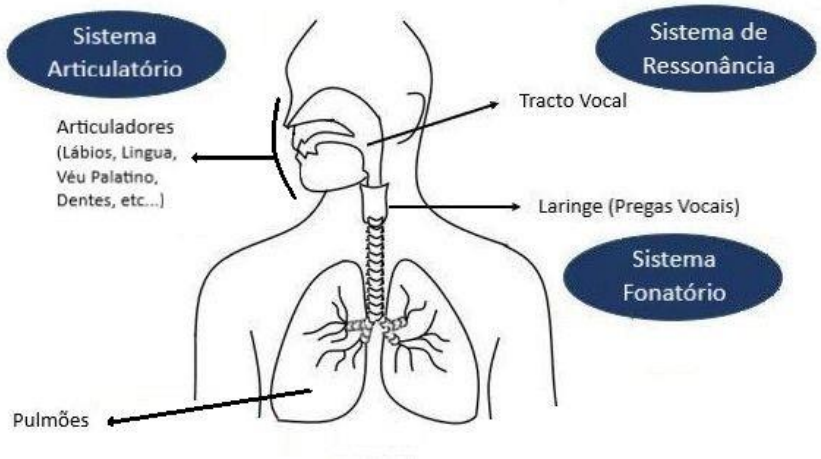


Figura 2-5 - Aparelho fonador humano (adaptado de [32])

O ar é o combustível para a produção de fala, traduzindo-se o seu volume, fluxo e pressão na energia para o processo. Após um ciclo inspiratório, durante o qual os pulmões são preenchidos com ar, o diafragma relaxa e os músculos intercostais, entre outros, contraem-se, comprimindo os pulmões e criando dentro destes uma pressão positiva relativa ao exterior. Essa diferença de pressão cria um fluxo de ar que percorre todo o trato respiratório inferior até chegar à laringe.

Na laringe, o fluxo de ar pode ou não ser restringido pela glote. Quando as pregas vocais se mantêm abertas e o ar flui livremente pela laringe, o sinal de fala gerado será não vozeado. Se houver restrição, ou seja, se as pregas vocais estiverem fechadas, a pressão abaixo da glote aumentará, forçando a abertura das pregas vocais e gerando um fluxo de ar. Esse fluxo de ar provoca uma redução da pressão abaixo da glote e, em conjunto com a elasticidade das pregas vocais e o efeito *Bernoulli*, faz com que estas voltem à sua posição inicial fechando-se. Este ciclo repete-se durante a fonação.

Os ciclos vibratórios das pregas vocais dão origem a alterações de pressão e do volume de ar que passa pela glote, criando uma onda sonora, com uma determinada frequência, que dará eventualmente origem a um sinal de fala vozeado. Essa frequência tem um valor igual ao número de ciclos vibratórios das pregas vocais por segundo e denomina-se frequência fundamental (F0) ou *pitch*.

O F0 é um parâmetro acústico do sinal de fala e define o tom, mais grave se F0 for mais baixo ou mais agudo se F0 for mais alto.

A frequência fundamental está diretamente associada às propriedades das pregas vocais, como o tamanho e a massa. Pregas vocais maiores, mais longas e espessas, tendem a vibrar mais lentamente devido à sua maior massa, o que resulta em menos ciclos de vibração por segundo, ou seja, numa frequência mais baixa. Portanto, quanto maiores forem as pregas vocais, menor será o valor da frequência fundamental, o que explica porque existe uma diferença típica no tom de voz entre os géneros, pois os homens geralmente têm pregas vocais maiores e mais espessas do que as mulheres. Além disso, mudanças na laringe relacionadas com envelhecimento, como a atrofia dos músculos intrínsecos da laringe ou a perda de elasticidade dos ligamentos, também têm impacto no tom de voz, o que explica porque se podem observar diferenças nas vozes entre oradores de idades diferentes.

A onda sonora formada na laringe atravessa de seguida o trato vocal, que atua como um sistema de ressonância. As cavidades faríngea, bucal e/ou nasal irão, através das suas formas, volumes, constituições e irregularidades, alterar a onda sonora original através de reflexões e absorções. Estas características irão moldar as harmónicas, ou seja, as ondas de frequências múltiplas da frequência fundamental, da onda resultante, podendo estas ser amplificadas ou atenuadas dependendo da configuração do trato vocal.

Para além da ressonância, também os elementos articuladores desempenham um papel fundamental na produção do sinal de fala. Estes podem ser estruturas móveis, como os lábios e a mandíbula, e fixas, como os dentes e o palato duro, que, através de variações nas suas configurações vão criar obstáculos à passagem do ar no trato vocal. A criação desses obstáculos dá origem ao processo de articulação que, combinado com os descritos anteriormente, resulta na produção da fala, especificamente, do fonema. O processo de produção de fala é descrito com maior pormenor em [20], [30] e [31], fontes onde se baseou a descrição dada neste subcapítulo.

### 2.2.2 Patologias da voz

De acordo com [20] e [33], quando existe uma perturbação, alteração ou dificuldade na produção da fala, comprometendo a sua qualidade de forma persistente está-se na presença de uma patologia da voz, ou vocal. Quando essa patologia está diretamente associada à laringe, denomina-se de patologia laríngea. Estas condições médicas podem dever-se a vários factores, incluindo lesões, doenças, abuso ou uso inadequado da voz, consumo de álcool ou tabaco, condições ambientais e predisposição genética. Neste trabalho, apenas quatro dessas patologias são estudadas, por limitação dos dados. Duas delas são patologias estruturais da laringe e as outras duas são patologias neurológicas que afetam a fala.

O edema de Reinke é uma alteração tecidular benigna e crónica na camada superficial da lâmina própria (espaço de Reinke). Caracteriza-se por um inchaço causado pela acumulação de fluido e é normalmente bilateral e assimétrico. Conforme o fluido se acumula, a elasticidade das pregas vocais é afetada e a sua massa aumenta, degradando a qualidade da voz, tornando-a mais grave e rouca. A

sua principal causa é o tabagismo, sendo uma patologia observada com maior frequência em oradores de meia-idade com longo historial de consumo tabágico.

Os nódulos vocais são lesões de massa benignas, bilaterais e geralmente simétricas. Habitualmente formam-se na zona de maior impacto entre as pregas vocais, tipicamente na zona de delimitação dos terços anterior e médio das pregas vocais. A sua formação deve-se geralmente a um trauma repetitivo nas cordas vocais, causado por mau uso vocal repetido e prolongado, como gritar ou falar demasiado. O aparecimento destas massas impede o fecho completo das pregas vocais, adicionando ruído à voz produzida e alterando o seu timbre, tornando-a rouca e áspera. São observados com maior frequência em oradores que fazem uso profissional da voz, como professores ou cantores.

A paralisia unilateral das pregas vocais é uma condição que afeta uma das pregas vocais, dificultando ou impedindo a sua vibração durante o processo de fonação, causando uma assimetria na função vocal. Essa paralisia pode ocorrer devido a danos nos nervos que controlam os músculos da laringe, impedindo o movimento adequado dos músculos afetados e, consequentemente, paralisando a prega vocal associada a esses músculos. Pode ser causada por lesões traumáticas, cirurgias na área do pescoço, tumores na região da laringe, infeções virais e distúrbios neurológicos. A paralisia pode afetar a prega vocal de várias maneiras, resultando numa variedade de possíveis sintomas, como rouquidão, dificuldade em projetar a voz ou mesmo dificuldade em engolir [20] [33].

Na Figura 2-6 estão representadas nas pregas vocais as três patologias descritas.

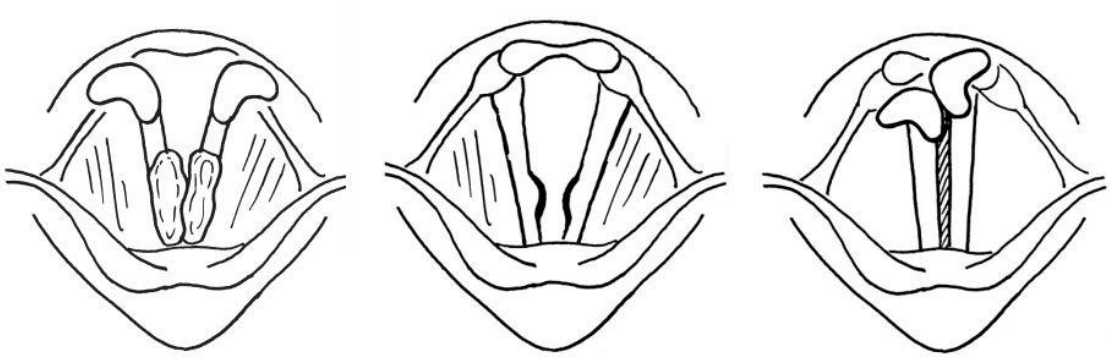

Figura 2-6 - Efeitos nas pregas vocais de edema de Reinke (esq.), nódulos vocais (centro) e paralisia unilateral das pregas vocais (dir) (adaptado de [20])

Patologias neurodegenerativas, como a doença de Huntington e a esclerose lateral amiotrófica, envolvem uma degradação dos neurónios responsáveis pelos movimentos musculares. Estas condições podem ter impactos significativos na voz devido aos seus efeitos sobre o sistema nervoso central, os nervos periféricos e os músculos envolvidos na produção vocal. Esta pode tornar-se fraca ou soprosa, devido à atrofia dos músculos responsáveis pela produção vocal, arrastada e imprecisa, devido à perda de capacidade de articulação ou distorcida e com interrupções devido a eventuais espasmos musculares provocados por estas condições neuromusculares [34], [35], [36].

### 2.2.3 Características da fala

De acordo com [37], o som, em termos físicos, é uma variação de pressão em relação à pressão média, ou estacionária, de um meio. Essa variação é gerada pelas vibrações mecânicas de um corpo, que provocam deslocamentos e oscilações nas moléculas do meio onde esse corpo está inserido. Essas oscilações geram regiões de maior ou menor densidade molecular, traduzindo-se em alterações na pressão local, que por sua vez se propagam pelo meio como ondas sonoras.

No contexto da fala, os sons são os fonemas produzidos com vibração das pregas vocais (se forem vozeados) ou pelas constrições na passagem do ar no trato vocal (se não vozeados), e moldados pelos sistemas ressonante e articulatório. Esses sons propagam-se pelo ar e podem ser captados por um ouvido humano, ou por um microfone, que converte pressão em tensão elétrica. Ao registar os valores dessa tensão elétrica ao longo do tempo, obtém-se uma representação física desses sons de fala, ou seja, um sinal de fala. Na Figura 2-7 pode ser visualizado o processo descrito, desde a propagação do fonema produzido até à obtenção do sinal de fala que o representa.

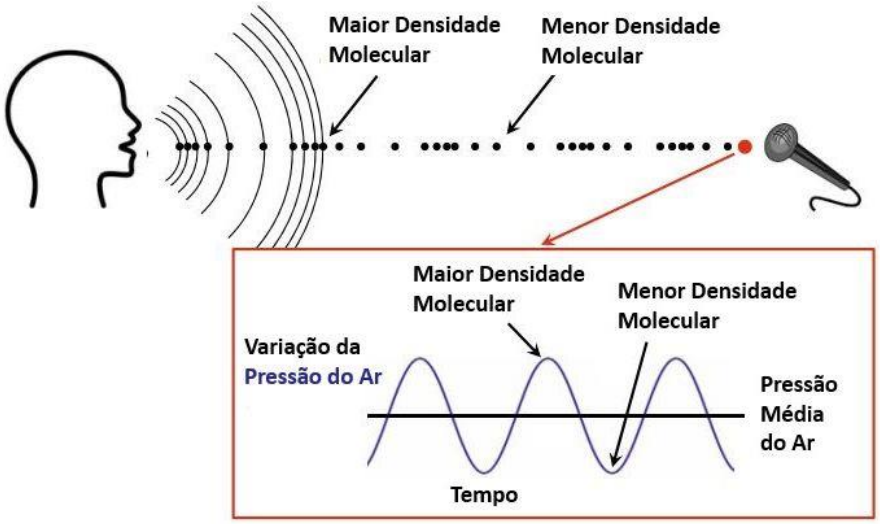


Figura 2-7 - Processo de obtenção de um sinal de fala (adaptado de [37])

Os sons, sendo oscilações, partilham características quantitativas com outras grandezas físicas que também oscilem periodicamente, como os campos elétrico e magnético, ou o nível do mar. Essas características incluem frequência, amplitude, energia e comprimento de onda. No entanto, essas características nem sempre são suficientes, ou mesmo úteis para um determinado objetivo. Nesses casos, torna-se necessário obter outros parâmetros, mais informativos para o propósito em questão. No contexto da fala, alguns parâmetros quantitativos podem ser obtidos através da evolução temporal do sinal de fala, ou seja, através da sua análise no domínio do tempo, designando-se estes por parâmetros acústicos. Existem outros parâmetros, também muito utilizados para caracterizar sinais de fala, obtidos num outro domínio, da frequência, designados por parâmetros espectrais.

### 2.2.3.1 Características espectrais

No século XIX, o matemático francês Jean Baptiste Joseph Fourier (1768-1830) demonstrou que qualquer sinal periódico e infinito no tempo pode ser decomposto numa soma de senos e cossenos com frequências múltiplas inteiras da frequência fundamental presente no sinal. Este conceito materializou-se na expressão conhecida como *Série de Fourier*, que se apresenta de seguida na Equação 2-1.

$$s(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n cos\left(\frac{2\pi nt}{T}\right) + b_n sen\left(\frac{2\pi nt}{T}\right) \right) \quad (2\text{-}1)$$

onde s(t) é o sinal periódico e infinito no tempo original, os coeficientes $a_0$, $a_n$ e $b_n$ são dependentes do sinal s(t) e T é o período fundamental do sinal s(t).

A partir desta ideia, Fourier desenvolveu uma ferramenta matemática, designada como *Transformada de Fourier*, que permite obter, a partir da representação temporal de um sinal, a distribuição dos sinais sinusoidais que o constituem, ou seja, a sua representação no domínio da frequência, tornando possível a análise do sinal nos dois domínios, como se mostra na Figura 2-8, adaptada de [38].

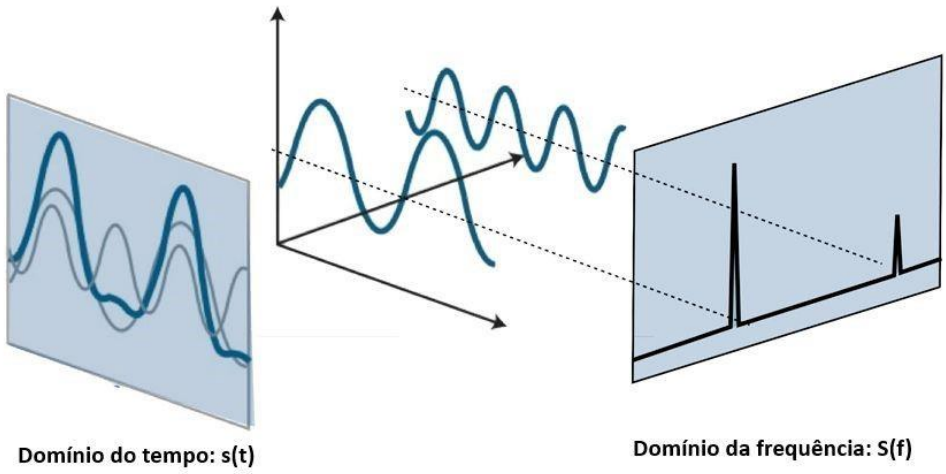


Figura 2-8 - Sinal s no domínio do tempo, s(t), e da frequência, S(f) (adaptado de [38])

No exemplo apresentado na Figura 2-8, fica evidente que a representação do sinal de fala no domínio da frequência, ou o seu espectro, facilita a distinção das suas diferentes componentes sinusoidais, nomeadamente as suas frequências e amplitudes. Essas componentes são, por si só, características muito significativas do sinal de fala, mas além disso, permitem que sejam derivadas, a partir delas, outras características potencialmente mais relevantes para algumas finalidades.

Contudo, os sinais de fala não são periódicos, pois são tipicamente compostos por uma sequência de fonemas, cada um com o seu próprio espectro distinto. Assim sendo, o espectro de um

sinal de fala será uma combinação dos espectros dos fonemas que o compõem, o que dificilmente será o resultado pretendido. Portanto, para se efetuar uma análise espectral, geralmente divide-se o sinal de fala em segmentos. Esses segmentos devem ter uma duração suficientemente longa para conter um período fundamental do sinal e suficientemente curta para maximizar a probabilidade de conter apenas um fonema, assumindo-se que o sinal é estacionário nesse segmento. Dessa forma, obtêm-se vários espectros relativos aos fonemas contidos no sinal de fala e não um único espectro relativo a todo o sinal de fala.

Os próprios fonemas não são verdadeiramente estacionários, mas quase estacionários, pois apresentam variações dentro de um mesmo segmento que resultam em alterações espectrais. No entanto, essas alterações são pouco significativas, sendo consideradas desprezáveis na maioria das aplicações e não sendo, por essa razão, impeditivas para a utilização da análise espectral. Neste trabalho, essas alterações são uma caraterística de interesse, como se discutirá adiante.

#### 2.2.3.2 Características acústicas

Os parâmetros acústicos abordados neste trabalho são descritos de forma pormenorizada em [20], [39], [40] e [41], fontes em que se baseai este subcapítulo. Os parâmetros acústicos são obtidos através de análise do sinal de fala no domínio do tempo, sendo um exemplo, a frequência fundamental dos pulsos glotais, já referida na Secção 2.2.1. A partir desta podem ser obtidas outras características quantificáveis, como a variação da frequência fundamental. Neste trabalho são consideradas três características acústicas quantificáveis, as quais se descrevem de seguida, com auxílio da Figura 2-9, onde podem ser visualizados os processos de extração de dois deles.

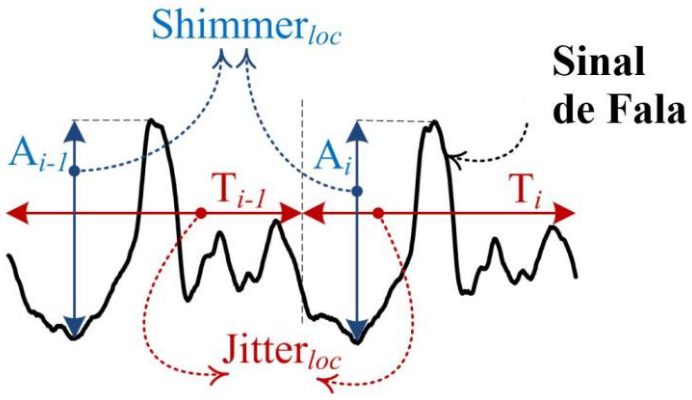


Figura 2-9 - Representação gráfica da obtenção dos parâmetros *shimmer* e *jitter* (adaptado de [39])

O *jitter* é um parâmetro que quantifica a perturbação rápida na frequência fundamental da voz, através da avaliação da variação do período glotal entre dois ciclos consecutivos. Devido à sua curta duração, essa variação não é considerada voluntária, tornando o *jitter* um indicador da estabilidade do sistema fonatório, pois depende do controlo da vibração das pregas vocais.

Matematicamente, o *jitter* pode ser definido como a diferença média entre períodos glotais consecutivos, dividida pelo período médio, conforme se expressa de seguida na Equação 2-2:

$$jitter = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \quad (2\text{-}2)$$

onde N é o número de períodos glotais considerados, T é um determinado período glotal, conforme ilustrado na Figura 2-9, e o resultado é adimensional, sendo normalmente expresso em percentagem.

Esta é uma definição de *jitter*, mas existem outras, que podem não ter em consideração o período médio (*absolute jitter*), podem considerar mais que um ciclo consecutivo (*RAP jitter* e *PPQ*), etc. Neste trabalho considerar-se-á apenas uma versão de *jitter*, como será explicado adiante.

O *shimmer* é um parâmetro que quantifica a perturbação rápida na intensidade da voz, avaliando a variação da amplitude do ciclo glotal entre dois ciclos consecutivos. Estas variações, pela sua curta duração, não são atribuídas a mudanças voluntárias. Dessa forma, o *shimmer* torna-se também um indicador para a estabilidade do sistema fonatório.

O *shimmer* pode ser definido como a diferença absoluta média entre a amplitude de períodos consecutivos, dividida pela amplitude média, como se observa na Equação 2-3:

$$shimmer = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \quad (2\text{-}3)$$

onde N é o número de períodos glotais considerados, A é a amplitude de um determinado período, conforme ilustrado na Figura 2-9, e o resultado é normalmente expresso em percentagem.

Tal como o *jitter*, também o *shimmer* tem variantes, podendo ser expresso em unidades logarítmicas, ou tendo em consideração mais ciclos adjacentes, mas também neste caso, apenas uma versão de *shimmer* será considerada neste trabalho.

O *Harmonics-to-Noise Ratio* (*HNR*) é um parâmetro que quantifica o ruído na voz, através da relação entre a potência contida nas componentes periódicas e a contida nas componentes aperiódicas, considerada ruído. Essa relação quantifica a eficiência do processo de fonação, pois quanto maior for o ruído glótico, menor será o valor de *HNR*. Os valores deste parâmetro podem abranger várias ordens de grandeza, sendo por essa razão medido em unidades logarítmicas, ou decibéis (dB), e a sua expressão apresenta-se na Equação 2-4:

$$HNR = 10 \times log_{10}\left(\frac{P_H}{P_R}\right) = 10 \times log_{10}\left(\frac{P_H}{P_T - P_H}\right) \quad (2\text{-}4)$$

onde $P_H$ é a potência das componentes periódicas, ou harmónicas, $P_N$ é a potência do ruído e $P_T$ é a potência total do sinal de fala.

Usualmente, para o cálculo do parâmetro *HNR*, utiliza-se a expressão mais à direita, pois é muito mais complexo medir diretamente a potência do ruído do que determiná-la subtraindo a potência harmónica à potência total do sinal. A determinação da potência harmónica pode envolver diferentes métodos e parâmetros, como parâmetros cepstrais, ou a autocorrelação do sinal de fala. Neste trabalho o *HNR*, tal como os restantes parâmetros acústicos, será obtido através de uma aplicação desenvolvida por terceiros.

## 2.3 Revisão da literatura

Esta subsecção apresenta uma abordagem teórica a algumas pesquisas desenvolvidas no contexto de discriminação entre vozes saudáveis e patológicas. Os trabalhos foram realizados com recurso a bases de dados contendo sinais de fala pertencentes a, pelo menos, estas duas categorias. Consistiram na extração de parâmetros dos sinais de fala e na utilização de métodos de processamento de sinal, de classificação automática e/ou de análise estatística com o objetivo de diferenciar vozes de oradores saudáveis e patológicos.

### 2.3.1 Estado da arte

A tendência atual em pesquisas recentes é a utilização de modelos de *deep learning* para classificar sinais de fala e categorizar oradores como saudáveis ou patológicos. Essa classificação pode ser feita com redes neuronais treinadas de raiz ou em redes pré-treinadas, nas quais a camada de decisão é substituída por uma camada treinada com os dados disponíveis, num processo chamado de *transfer learning*.

Esses estudos têm obtido resultados notáveis, frequentemente com taxas de acerto na ordem dos 98-100%, validando a utilidade dos modelos. No entanto, esses modelos são de alta complexidade e o seu funcionamento é pouco transparente. A análise do seu funcionamento, principalmente de quais as características do sinal de fala que contribuem para o seu desempenho, não é trivial e muito raramente é descrita nos estudos encontrados.

Por outro lado, quando se utiliza um modelo mais complexo que o exigido pelos dados, ou seja, com mais parâmetros que os fornecidos pelos dados, este tem tendência a adaptar-se excessivamente aos dados observados, num processo chamado de *overfitting* [42]. Esta é uma consequência a evitar, pois modelos com *overfitting* demonstram ineficácia a prever novos dados. Para evitar esse efeito, seria necessária uma quantidade de dados muito maior que a existente neste trabalho para treinar e testar uma rede neuronal.

Portanto, considerando que a utilização de redes neuronais torna a classificação muito suscetível a *overfitting*, e como este estudo visa não só discriminar de modo eficaz vozes saudáveis e patológicas, mas também compreender quais as características do sinal de fala que levam o modelo a efetuar a classificação, foi tomada a opção de não se considerar este tipo de modelo. Por essa razão, não são abordados estudos que utilizem apenas redes neuronais e quando estas são utilizadas, a ênfase é dada aos outros modelos usados nesses estudos.

#### 2.3.1.1 Parâmetros espectrais

Foram usados parâmetros espectrais chamados *Mel-Frequency Cepstral Coefficients* (*MFCC*) para discriminar vozes patológicas em [43]. Esses parâmetros são obtidos a partir do espectro do sinal de fala, convertendo as frequências para uma escala que reflete a percepção auditiva humana, conhecida como escala *Mel*. Após esta conversão de escala, aplica-se o logaritmo às amplitudes do espectro e, finalmente, a *transformada de cosseno*, uma técnica matemática semelhante à *Transformada de Fourier*, é aplicada à representação resultante do espectro. O estudo utilizou sinais de fala contendo a vogal /a/ sustentada, pertencentes à *Saarbrücken Voice Database* (*SVD*), uma base de dados com sinais de fala de mais de 2000 oradores, incluindo oradores patológicos com mais de 70 patologias diferentes. Essa característica torna a *SVD* uma referência comum em pesquisas de discriminação de vozes patológicas. O número de sinais utilizados variou ao longo do processo, começando com 160 sinais, metade de oradores saudáveis e metade de oradores patológicos, sendo incrementado em grupos de 40 sinais até atingir os 280, sempre com a mesma proporção de oradores saudáveis e patológicos. Para a classificação automática foi utilizado o modelo de máquinas de *Support Vector Machines* (*SVM*), sendo este treinado com 80% dos dados disponíveis e testado com os restantes 20% dos dados. O estudo teve como objetivos validar a combinação de *MFCC* e *SVM* para discriminar vozes patológicas e avaliar a influência do número de sinais de fala no desempenho do sistema. Os resultados indicaram que as métricas de desempenho do sistema, principalmente a taxa de acerto, apresentavam valores decrescentes com o aumento do número de sinais de fala. A taxa de acerto máxima foi de 84,37% com 160 sinais de fala, diminuindo gradualmente até 73,07% com 280 sinais. Com base nas taxas de acerto obtidas, concluiu-se que um sistema automático de discriminação de vozes patológicas utilizando *MFCC* como parâmetros de entrada e *SVM* como classificador é viável. No entanto, parece ser mais adequado para bases de dados de pequena dimensão, diminuindo a sua eficácia com o aumento da quantidade de sinais de fala disponíveis.

No estudo [44], os autores utilizaram um subconjunto da base de dados *Massachusetts Eye and Ear Infirmary Voice Disorders Database* (*MEEI*) contendo sinais de fala contínua categorizados em três classes: oradores saudáveis, patológicos com patologias laríngeas fisiológicas e patológicos com paralisia unilateral das pregas vocais (*Unilateral Vocal Fold Paralysis - UVFP*). Foram usados os parâmetros espectrais *MFCC*, *Line Spectral Frequencies* (*LSF*) e *Mel Line Spectral Frequencies* (*MLSF*). Os *LSF* contêm informação sobre as frequências e larguras de banda dos formantes, que representam o trato vocal. Os *MLSF* são baseados no segundo parâmetro, *LSF*, mas contêm informação perceptual devido ao uso de um banco de filtros definido na escala *Mel*. Os sinais de fala, cuja frequência de amostragem original era de 25 kHz, foram reamostrados para 4 kHz e classificados com dois modelos distintos: um baseado em *Gaussian Mixture Models* (*GMM*) e outro em *SVM*. Os resultados indicaram que a redução da largura de banda dos sinais de fala não prejudicou o desempenho dos sistemas, e em alguns casos o desempenho até melhorou, sugerindo que as baixas frequências contêm a informação suficiente para a discriminação de vozes saudáveis e patológicas.

Os parâmetros espectrais *MFCC* voltaram a ser usados num outro estudo [45], juntamente com alguns parâmetros temporais. A partir de sinais de fala da base de dados *SVD* contendo a vogal /a/ sustentada num tom baixo (*pitch* mais baixo que o normal), foram extraídos, para além do *MFCC*, os parâmetros potência, energia, entropia, e o *zero-crossing-rate* (*ZCR*), que quantifica as passagens por zero da amplitude do sinal num dado intervalo de tempo. Os valores destes parâmetros foram analisados para três categorias de sinais de fala, referentes a oradores com três patologias: laringite, leucoplasia e disfonia. Após análise dos valores médios e das suas variações para cada uma das categorias, concluiu-se que o parâmetro *MFCC*, mais concretamente os seus primeiros coeficientes, relativos às frequências mais baixas, é o mais adequado, entre os estudados, para ser usado num sistema de classificação automática.

Num estudo posterior [46], os mesmos autores investigaram a utilização de vários parâmetros temporais e espectrais para discriminar vozes patológicas. Neste estudo foram analisados os parâmetros *MFCC*, *delta MFCC*, frequência fundamental, relação entre o valor de pico da amplitude e o seu valor médio (*spectral crest*), entropia temporal e espectral, *spectral flatness*, que mede a uniformidade da distribuição de potência ao longo das frequências, *spectral flux,* que mede a variação do espectro do sinal em tramas sucessivas, extensão das pontas da distribuição de potência (*kurtosis*), ponto de *rol-off*, que é o ponto abaixo do qual está contida uma determinada quantidade da potência do sinal, *skewness*, que mede quão simétrica é a distribuição de potência do sinal, declive do espectro do sinal, *spread*, *HNR* e centroide espectral, que define um centro de massa, ou a frequência média do espectro. Foram utilizados 535 sinais de fala da base de dados *SVD*, contendo as vogais /a/, /i/ e /u/ sustentadas, categorizados como saudáveis e patológicos. Os valores dos parâmetros considerados foram obtidos e a sua variação entre as duas classes de sinais foi quantificada através de dois métodos: avaliação do primeiro valor próprio obtido através da aplicação de *Principal Component Analysis* (*PCA*), e avaliação do valor de um parâmetro denominado *Fuzzy Entropy* (*FEM*). Os resultados indicaram que os parâmetros *MFCC*, *delta MFCC* e centroide espectral contêm mais informação para a discriminação entre vozes saudáveis e patológicas de acordo com o método *FEM*. Já o *PCA* identificou os parâmetros *MFCC*, *delta MFCC* e entropia espectral como os mais influentes para esse processo.

Um novo parâmetro espectral denominado *Low Band Spectral Tilt* (*LBST*) foi proposto em [47] para auxiliar na discriminação entre vozes saudáveis e patológicas. O *LBST* é calculado a partir da diferença entre os valores de energia em dois máximos locais do espectro do sinal de fala. Um dos máximos está localizado na gama de frequência onde se situam as primeiras harmónicas, enquanto o outro máximo situa-se numa zona de frequências onde se espera que exista um máximo, correspondente ao primeiro formante da vogal /a/. Essa diferença é então dividida pela distância em frequência entre esses dois máximos, resultando num declive espectral. O desempenho do *LBST* foi avaliado em sinais de fala contendo a vogal /a/ sustentada, pertencentes a duas bases de dados: um subconjunto com 206 sinais da base de dados *MEEI* e um subconjunto de 46 sinais de fala da base de dados da *Universidade de São Paulo* (*USP*). Para comparação, foram utilizados dois outros parâmetros, o *First Spectral Peak – P1*, parâmetro espectral referente ao primeiro máximo local da envolvente espectral do sinal de fala, e o *Relative Power of the Periodic Component* (*RPPC*), parâmetro

temporal referente ao valor da autocorrelação normalizada do sinal de fala, com um atraso igual ao valor do seu período fundamental. Os três parâmetros foram avaliados individualmente com o objetivo de discriminar os sinais de oradores saudáveis e patológicos. Os resultados indicaram que a avaliação do parâmetro *LBST* obteve as melhores taxas de acerto, de 83,5% e 100% para as bases de dados pertencentes à *MEEI* e à *USP* respetivamente. Esses resultados validam o *LBST* como um parâmetro promissor para utilização num potencial sistema automático de discriminação de vozes saudáveis e patológicas.

No estudo [48] foi proposto um sistema automático para a deteção de indícios de disartria em sinais de fala, com o objetivo de auxiliar no diagnóstico precoce de acidentes vasculares cerebrais. Foram utilizados *ZCR*, frequência fundamental e energia do sinal de fala, que os autores denominam de parâmetros prosódicos, *jitter*, *shimmer* e *HNR*, denominados neste trabalho de parâmetros de qualidade, *MFCC*, denominados aqui de cepstrais e os centroides, *flux*, *slope* e entropia espectrais. Para a classificação dos sinais de fala em disártricos e não disártricos, foram utilizados quatro modelos de classificação automática: *SVM*, *Random Forest*, *K-Nearest Neighbors* (*KNN*) e *Naive Bayes*. Os resultados indicaram que a melhor combinação de parâmetros e classificador para a deteção de disartria foi a utilização dos parâmetros prosódicos e espectrais em conjunto com o classificador *Random Forest*, atingindo uma taxa de acerto de 94,33%. O classificador *SVM* obteve a segunda melhor taxa de acerto, de 87,13% quando usados os parâmetros de qualidade (*jitter*, *shimmer* e *HNR*) e os *MFCC* como parâmetros de entrada.

Quatro classificadores automáticos foram utilizados para distinguir entre vozes saudáveis e patológicas no trabalho [49]. Utilizaram-se sinais de fala da base de dados *SVD* contendo a vogal /a/ sustentada, categorizados como sinais de oradores saudáveis e patológicos. Esses sinais foram usados de duas formas: numa delas foram usados todos os sinais, 687 saudáveis e 1354 patológicos, e na outra alguns sinais foram descartados de modo a balancear as classes, tendo sido utilizados 685 sinais de oradores saudáveis e 685 de oradores patológicos. Vários parâmetros foram extraídos desses sinais, incluindo os coeficientes *MFCC* e alguns parâmetros estatísticos, não especificados pelos autores. Os parâmetros foram utilizados de duas formas: numa delas foram usados todos os parâmetros e na outra foi feita uma seleção de dez parâmetros recorrendo a um processo denominado *Gradient Boosting Machine*. Ou seja, no total foram exploradas quatro configurações diferentes de sinais e parâmetros. Essas configurações foram aplicadas aos classificadores automáticos *SVM*, *KNN*, *Decision Tree* e *Logistic Model Tree* e avaliados os seus desempenhos. O *SVM* atingiu a taxa de acerto mais elevada em três das quatro configurações, com o melhor desempenho global a ser obtido com os dados balanceados e avaliando todos os parâmetros com a taxa de acerto a atingir os 85,91%.

No estudo [50] foi utilizado o subconjunto de sinais de fala da base de dados *MEEI* já usado anteriormente em [44]. Os sinais foram categorizados em oradores saudáveis, patológicos com *UVFP* e patológicos com edema ou nódulos vocais (patologias laríngeas fisiológicas). No estudo foram utilizados sinais com a vogal /a/ sustentada e com fala contínua. Foram extraídos os parâmetros espectrais *LSF*, *MLSF* e *MFCC*, e os parâmetros diferenciais a estes associados. O *MFCC* foi agrupado com o *delta MFCC*, o *LSF* com o *DLSF*, que é a diferença entre os coeficientes *LSF* obtidos em tramas

consecutivas, e o *MLSF* com o *DMLSF*, similar ao *DLSF*. Os três conjuntos de parâmetros foram avaliados separadamente através de quatro classificadores automáticos: *GMM* multiclasse, *SVM One vs. One*, *Linear Discriminant Analysis* (*LDA*) multiclasse e *One vs. One*. Na discriminação entre vozes saudáveis e patológicas a combinação *LSF/SVM* obteve os melhores resultados, atingindo valores mais elevados nas métricas de avaliação. Já na discriminação entre as três classes, a combinação de *MLSF* e *GMM* obteve a taxa de acerto mais elevada. Em todas as combinações, os parâmetros extraídos dos sinais de fala contínua deram origem a melhores desempenhos, sugerindo que a fala contínua pode ser mais adequada para esta tarefa que a vogal /a/ sustentada.

Esta conclusão foi corroborada pelos resultados obtidos em [51], onde foi usado o mesmo subconjunto com sinais de fala contendo a vogal /a/ sustentada e fala contínua. Neste estudo foram testadas seis configurações diferentes de parâmetros e classificadores, bem como um classificador hierárquico composto por duas dessas configurações. Os modelos com parâmetros extraídos de sinais de fala contínua obtiveram o melhor desempenho em seis das sete métricas avaliadas, reforçando a ideia de que a fala contínua será a mais adequada para discriminação entre vozes saudáveis e patológicas. No entanto, na discriminação de patologias, os resultados não foram tão conclusivos. A dificuldade em detetar tramas vozeadas nos sinais de vozes patológicas, devido à potencial degradação da qualidade de voz, pode comprometer a extração das caraterísticas espectrais nestes sinais e consequentemente, o desempenho na discriminação entre patologias. O melhor desempenho foi atingido com uma configuração de classificação hierárquica, que combina *SVM* e *LDA* na classificação e parâmetros extraídos de sinais com a vogal /a/ sustentada e com fala contínua. Esta configuração atingiu uma taxa de acerto de 84,4% na classificação multiclasse. Quando usados sinais de fala contendo a vogal /a/ sustentada, os melhores desempenhos na classificação multiclasse foram obtidos com uma combinação de *MFCC* e *LDA* em [50], onde foi atingida uma taxa de acerto de 70,1%, e com uma combinação de *MLSF* e *LDA* em [51], com uma taxa de acerto de 68,2%.

Parâmetros relacionados com a fonte glotal são analisados em [52]. Sinais glotais foram estimados a partir de sinais de fala, categorizados como saudáveis e patológicos, de duas bases de dados, a *SVD* e o *corpus* do *Hospital Universitário Príncipe das Astúrias* (*HUPA*). A estimação dos sinais glotais foi efetuada através de dois métodos distintos: o *Quasi-Closed Phased* (*QCP*), que se baseia no modelo fonte-filtro para produção de sinais de fala, e o *Zero Frequency Filter* (*ZFF*). A partir dos sinais glotais estimados através do método *QCP* foram extraídos 12 parâmetros, denominados parâmetros glotais, sendo três deles espectrais e os restantes nove obtidos através de processamento dos sinais no domínio do tempo. A partir dos sinais glotais estimados através do método ZFF foram extraídos quatro parâmetros glotais. Outros quatro parâmetros glotais foram extraídos diretamente dos sinais de fala. Foram também extraídos valores para um parâmetro que os autores denominaram *MFCC glotais*, que corresponde aos parâmetros *MFCC* extraídos dos sinais glotais estimados. Todos esses parâmetros foram combinados em várias disposições diferentes e utilizados com um classificador *SVM* para discriminar oradores saudáveis e patológicos. Foram também extraídos os parâmetros *MFCC* (convencionais) e os coeficientes *Perceptual Linear Predictive* (*PLP*) a partir dos sinais de fala, e utilizados também com *SVM*, para que as taxas de acerto obtidas com os parâmetros convencionais

servissem de termo de comparação. O sistema implementado foi testado com cinco tipos diferentes de sinais de fala: sinais de fala da base de dados *SVD* contendo as vogais /a/, /i/ ou /u/ sustentadas, contendo fala contínua, e para sinais de fala da base de dados *HUPA* contendo a vogal /a/ sustentada. Apenas em dois dos cinco cenários, uma combinação de parâmetros glotais obteve uma taxa de acerto mais elevada que a obtida com os parâmetros espectrais convencionais. Em todos os cenários, a combinação dos parâmetros glotais e convencionais resultou na maior taxa de acerto, evidenciando que ambos os tipos de parâmetros fornecem informação útil e complementar para discriminar vozes saudáveis e patológicas, sendo vantajosa a sua combinação para essa tarefa.

#### 2.3.1.2 Parâmetros acústicos

Quatro parâmetros acústicos, nomeadamente frequência fundamental, *jitter*, *shimmer* e *HNR* foram estudados em [53]. Foram utilizados sinais de fala contendo as vogais /a/, /i/ e /u/ sustentadas, pertencentes a 120 oradores, sendo 80 saudáveis (40 de cada género) e os restantes 40 com disfonia. Os parâmetros foram extraídos utilizando a aplicação *Praat* e as médias foram analisadas para cada um de três grupos, saudáveis masculinos, saudáveis femininos e patológicos. Entre os grupos saudáveis, apenas a frequência fundamental apresentou uma diferença estatisticamente significativa, o que seria um resultado esperado, pois a frequência fundamental é tipicamente mais elevada em mulheres. Por essa razão, este parâmetro não foi usado na comparação entre oradores saudáveis e patológicos. Nessa comparação, verificou-se que o *jitter* e o *shimmer* apresentaram valores médios mais elevados no grupo de oradores patológicos, enquanto o *HNR* apresentou um valor médio mais baixo neste grupo. Esses resultados sugerem que o *jitter*, o *shimmer* e o *HNR* podem servir como indicadores para avaliar a presença de disfonia.

No estudo [54], foi utilizada uma base de dados massiva, que os autores denominaram por *Collected and Multiple Existing Dataset* (*CMED*), composta pela agregação de sinais de fala de várias bases de dados existentes, entre as quais a *SVD* e a *MEEI*. A base de dados *CMED* contém 8158 sinais de fala com as vogais /a/, /i/ ou /u/ sustentadas, que estão categorizados como saudáveis e patológicos. Os sinais de fala foram pré-processados através da divisão em tramas e da obtenção dos espectros das tramas através de *Short Time Fourier Transform* (*STFT*). Para cada espectro, foi estimado o ruído e subtraído ao espectro esse ruído, sendo o resultado convertido para o domínio do tempo através da *Transformada Inversa de Fourier*. Finalmente, o sinal de fala é reconstruído através das tramas processadas. Após esse pré-processamento, foram extraídos dez parâmetros acústicos dos sinais de fala. Quatro deles são relacionados com o *jitter* (*jitter local*, absoluto, *RAP* e *PPQ5*), quatro relacionados com o *shimmer* (absoluto, local, *APQ3* e *APQ5*) e os restantes são a frequência fundamental e a periodicidade. Aos parâmetros obtidos foram aplicadas três técnicas de seleção de caraterísticas para reduzir a dimensionalidade dos dados. As técnicas foram o *Univariate Feature Selection*, que os autores denominam de *Correlation Technique*, o *Information Gain Techinque* e o *PCA*. As quatro configurações, uma com todos os parâmetros e as outras após aplicação das técnicas referidas, são utilizadas em cinco classificadores automáticos: *SVM*, *Naive Bayes*, *Decision Trees*, *KNN* e *Random Forest* e avaliadas as taxas de acerto. A combinação que obteve a taxa de acerto mais

elevada, de 99,90%, foi a configuração *PCA* e *SVM*, validando esta combinação como adequada para um sistema de discriminação entre vozes saudáveis e patológicas.

Foi proposto, no estudo [55], um sistema para discriminar vozes patológicas, identificar a patologia e determinar a sua severidade. Foram utilizados 151 sinais de fala amostrados a 44,1 kHz, contendo a vogal sustentada /a/ com uma duração de 3 segundos. Os excertos iniciais e finais dos sinais de fala, com duração de um segundo, foram descartados, deixando para análise apenas o excerto intermédio. Foi aplicado um filtro *Butterworth* passa-banda de segunda ordem aos sinais de fala, de modo a manter a faixa de frequências esperadas para a frequência fundamental. Aos sinais resultantes foram extraídos os valores para os parâmetros acústicos *jitter*, *shimmer* e *HNR*. Os valores obtidos para os três parâmetros foram submetidos a um procedimento de classificação não supervisionada denominado *cobweb clustering*, resultando em 21 classes, ou *clusters*, e 12 folhas. Obtidos os *clusters*, foi identificado o *cluster* caracterizado pelos menores valores de *jitter* e *shimmer*, e pelos valores mais elevados de *HNR*. Esse *cluster* foi estabelecido como *cluster* de referência, representando as vozes saudáveis, e a partir do seu centroide, foram calculadas as distâncias e orientações para os centroides dos outros *clusters*. Com base na orientação, foram estimados os distúrbios vocais, sendo estimada a sua severidade com base na distância Euclidiana. Embora o método implementado tenha demonstrado limitações em termos de fiabilidade, principalmente, segundo os autores, devido à forma como o *cluster* de referência é determinado, o estudo apresentou resultados promissores, demonstrando o potencial dos parâmetros *jitter*, *shimmer* e *HNR* para o diagnostico de patologias vocais.

Outro estudo [56] utilizou sinais de fala da base de dados *SVD*, contendo as vogais sustentadas /a/, /i/ e /u/ em três tons diferentes, totalizando nove segmentos de fala por orador. Os sinais de fala foram divididos em três classes: oradores saudáveis, disfónicos e com paralisia das pregas vocais. Para cada segmento, foram extraídos valores de nove parâmetros acústicos: quatro parâmetros relacionados com *jitter* (*Jitta*, *Jitt*, *RAP* e *ppq5*), quatro parâmetros relacionados com *shimmer* (*Shim*, *SHDB*, *apq3* e *apq5*) e o parâmetro *HNR*, resultando em 81 valores por orador. Para reduzir a dimensionalidade dos dados, foram testados três métodos diferentes. Dois desses métodos consistem na seleção de características, especificamente *Agrupamento Hierárquico* e *Análise de Regressão Multilinear*, enquanto o terceiro método visa a redução da dimensão dos dados, utilizando *PCA*. Os conjuntos de dados resultantes foram classificados usando dois modelos de classificação automática diferentes, *Artificial Neural Networks* (*ANN*) e *SVM*. Os processos de classificação foram realizados separadamente para cada gênero, bem como para cada distúrbio. Os resultados obtidos foram contraditórios relativamente à melhor combinação de método de redução de dimensionalidade e classificador, pois para cada configuração de género/distúrbio, os resultados indicam uma combinação diferente como a mais adequada. Em relação à combinação *PCA/SVM*, foram alcançadas taxas de acerto de 83,3% para a combinação Feminino/Disfonia, 87,5% para a combinação Masculino/Disfonia, 76,3% para a combinação Feminino/Paralisia e 80,0% para a combinação Masculino/Paralisia. Apesar desta combinação não ter obtido o melhor desempenho em nenhuma das quatro configurações, as

taxas de acerto obtidas validam a utilização destes parâmetros acústicos em conjunto com *PCA* e *SVM* como uma solução viável para discriminar entre vozes saudáveis e patológicas.

#### 2.3.1.3 Conclusões

Os trabalhos mais recentes mostram a relevância dos parâmetros espectrais na discriminação de vozes saudáveis e patológicas, e de patologias, tendo sido usados em nove dos estudos abordados. Em todos os estudos, exceto em [47], foram usados parâmetros espectrais com informação perceptual (obtidos a partir da escala *Mel*). Quando foram usados parâmetros com e sem informação perceptual, os parâmetros espectrais obtidos na escala *Mel* obtiveram melhores desempenhos, ou fizeram parte de uma combinação com a mais elevada taxa de acerto, em sete estudos [44], [46], [48], [49], [50], [51] e [52]. Todos esses resultados demonstram ser vantajosa a obtenção e utilização de parâmetros espectrais na escala *Mel*.

Os estudos [44] e [45] indiciam que a informação para discriminação entre vozes saudáveis e patológicas poderá estar contida nas frequências mais baixas do espectro do sinal de fala. Em [45] foram usados alguns parâmetros e selecionadas as dez dimensões mais relevantes, sendo algumas delas os primeiros coeficientes do parâmetro *MFCC*, referentes às frequências mais baixas do sinal de fala. Em [44] foram comparados os desempenhos de sistemas de discriminação com sinais de fala amostrados a 25 kHz e a 4 kHz, onde se verificou que o desempenho obtido com os sinais amostrados a uma frequência mais baixa obtinham desempenhos iguais, ou ligeiramente superiores, validando assim a hipótese testada.

Quando usados classificadores automáticos, o modelo *SVM* destaca-se pela sua ampla utilização, pois é um dos classificadores testados em nove dos trabalhos abordados. Quando comparado o seu desempenho com outros classificadores, a utilização do *SVM* obteve o melhor, ou um dos melhores desempenhos, em quatro: [49], [50], [51] e [54], dos sete estudos onde essa comparação foi realizada, sugerindo a sua adequação para discriminação de vozes saudáveis e patológicas, e de patologias.

Verifica-se também uma preferência pela utilização de sinais de fala contendo a vogal /a/ sustentada nos trabalhos abordados, tendo sido usados em onze deles, quando apenas cinco trabalhos usaram sinais de fala contendo também as vogais /i/ e /u/ sustentadas, e quatro estudos usaram sinais de fala contínua. Embora os resultados de [50] e [51] indiciem que a fala contínua é o tipo de sinal de fala mais adequado para obtenção de parâmetros espectrais, a vogal /a/ sustentada demonstra também, pela sua frequente utilização e pelos resultados obtidos, ser adequada para sinais de fala neste contexto.

Os parâmetros acústicos *jitter*, *shimmer* e *HNR*, apesar de serem conhecidos e alvo de estudo há bastante tempo, ainda são utilizados em trabalhos recentes, como demonstram os cinco estudos abordados que se baseiam nos mesmos. Verifica-se que os três parâmetros são sempre usados em conjunto nos referidos estudos, o que sugere a existência de uma complementaridade entre eles. Os estudos [54] e [56] usaram estes parâmetros em conjunto com *PCA* e *SVM*, obtendo desempenhos

promissores e validando a sua utilidade para tarefas de discriminação de vozes saudáveis e patológicas.

Uma tendência que pode ser observada em trabalhos recentes é a combinação de parâmetros de diferentes tipos com o objetivo de explorar a sua potencial complementaridade e obter sistemas de discriminação com melhores desempenhos. Foram testadas combinações de parâmetros e comparados os desempenhos dos sistemas contra a utilização dos parâmetros isolados, em quatro estudos, [48], [49], [50] e [51]. Nesses estudos, os melhores desempenhos foram obtidos em situações onde existiu combinação de parâmetros, demonstrando a potencial utilidade dessa abordagem na implementação de sistemas automáticos de discriminação entre oradores saudáveis e com patologias vocais.

### 2.3.2 Trabalho relacionado

No trabalho [57] foi proposto um novo método para a estimação do parâmetro *LBST* com o objetivo de optimizar a eficiência computacional e mitigar problemas relacionados com a sua obtenção. Em [47], onde este parâmetro foi introduzido pela primeira vez, verificou-se uma degradação na taxa de acertos obtida com este parâmetro na discriminação entre patologias. Em alguns casos, onde a patologia estava num estado mais avançado, a qualidade da voz estava mais degradada, criando dificuldades na estimação da frequência fundamental, necessária para a determinação do parâmetro *LBST*. O método proposto neste estudo descarta a necessidade de determinar a frequência fundamental.

O parâmetro *LBST* relaciona a diferença entre dois máximos locais de energia, em zonas diferentes do espectro do sinal, e a diferença de frequências em que estes ocorrem, sendo expresso pela Equação 2-5:

$$LBST = \frac{SBME - FBME}{f_{SBME} - f_{FBME}} \tag{2-5}$$

onde *FBME* é o primeiro máximo local ou *First Band Maximum Energy*, situado na zona de baixas frequências do espectro, *SBME* é o segundo máximo, ou *Second Band Maximum Energy*, situado na zona de frequências intermédias do espectro, $f_{FBME}$ é a frequência onde ocorre o primeiro máximo e $f_{SBME}$ é a frequência onde ocorre o segundo máximo.

O procedimento proposto para a obtenção do parâmetro *LBST* neste estudo consiste em efetuar uma *STFT* aos sinais de fala, pertencentes ao subconjunto do *MEEI* usado anteriormente em [44], [50] e [51], dividindo-os em tramas de 30 ms, com um deslocamento de 10 ms, ou seja, com sobreposição entre tramas. Foram analisados apenas 500 ms de cada sinal, escolhidos das zonas mais estáveis dos mesmos. A gama de frequências foi dividida em duas, com um limiar de separação entre os 488,2 Hz e os 610,3 Hz, e os máximos em cada uma das duas bandas foram determinados, correspondendo aos *FBME* e *SBME*.

Já havia sido verificado em [47] que o declive *LBST* apresenta diferenças quando obtido a partir de sinais de oradores saudáveis ou patológicos. Um exemplo pode ser observado na Figura 2-10, onde se representam valores típicos para o parâmetro *LBST*: à esquerda, para um sinal de voz saudável, onde o *LBST* mostra um declive positivo, e à direita, para um sinal de voz patológico, onde se observa um declive negativo.

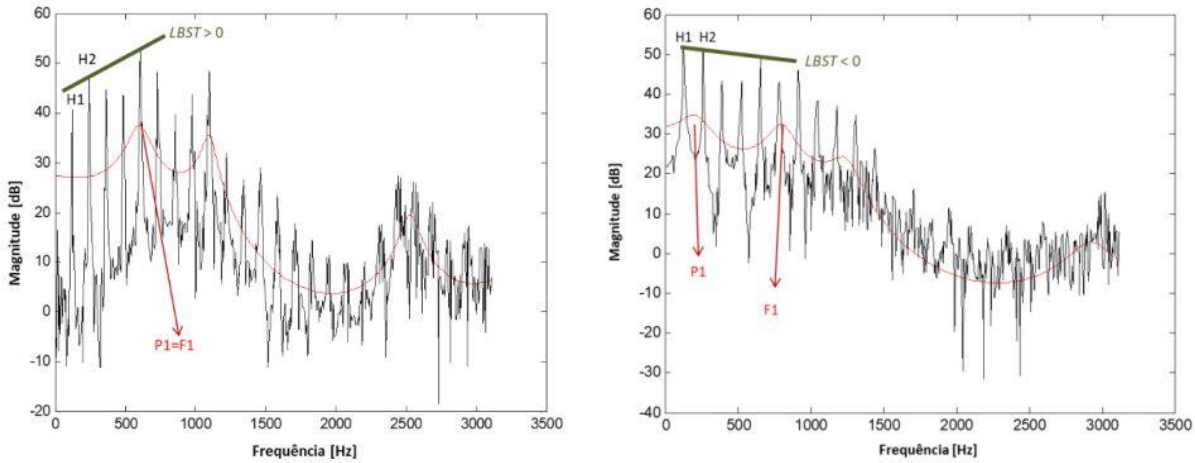


Figura 2-10 - *LBST* típico para orador saudável (esquerda) e com patologia de voz (direita) (retirado de [57])

No espectro dos sinais (a preto) podem ser verificadas as energias referentes às primeiras harmónicas *H1* e *H2*, usadas para determinar o *LBST* no método anterior. Podem também ser identificados os máximos locais na envolvente espectral (a vermelho) usados para determinar o *LBST* no método proposto neste trabalho.

Foram testados vários valores para o limiar que separa as duas bandas de frequência, assim como para o número de pontos para a *STFT*. O melhor desempenho foi obtido para um limiar de separação situado em 585,9 Hz e um número de 1024 pontos para a *Transformada de Fourier*, onde se atingiu um valor de 88,4% para a taxa de acerto e de 0,932 para a métrica *Area Under the Curve* (*AUC*).

O desempenho da discriminação entre vozes saudáveis e patológicas, baseada no parâmetro *LBST* obtido com o método proposto, foi superior aos desempenhos obtidos com os parâmetros usados em [47], nomeadamente *RPPC*, *First Spectral Peak P1* e *LBST* obtido com o método anterior. Este resultado valida o parâmetro *LBST*, ou seja, a relação dos máximos locais da energia nas duas bandas de frequência, como indicador eficaz para discriminação entre vozes saudáveis e patológicas, demonstrando ser mais eficaz quando determinado pelo método proposto neste estudo.

O parâmetro *LBST* voltou a ser usado num outro estudo [58], juntamente com outros dois parâmetros espectrais relacionados, o *High Band Spectral Tilt* (*HBST*) e o *Band Spectral Tilt Angle* (*BSTA*). O *HBST* é um parâmetro muito similar ao *LBST*, mas para a determinação deste, o espectro do sinal é dividido em três bandas e não duas. Os máximos locais na segunda e terceira bandas, bem

como a diferença entre as frequências em que estes máximos estão localizados, definem o *HBST*, como se expressa na Equação 2-6:

$$HBST = \frac{TBME - SBME}{f_{TBME} - f_{SBME}} \tag{2-6}$$

onde *TBME* é o terceiro máximo local, ou *Third Band Maximum Energy* e $f_{TBME}$ é a frequência onde ocorre o este máximo e sendo *SBME* e $f_{SBME}$ os parâmetros vistos anteriormente na Equação 2-5.

O *BSTA* é o ângulo entre os declives *LBST* e *HBST*, conforme expresso na Equação 2-7:

$$\begin{aligned} BSTA &= 180° + arctan\left(\left|\frac{LBST - HBST}{1 + LBST \times HBST}\right|\right) \; se \; LBST > HBST \\ BSTA &= 180° - arctan\left(\left|\frac{LBST - HBST}{1 + LBST \times HBST}\right|\right) \; se \; LBST < HBST \end{aligned} \tag{2-7}$$

Podem ser visualizadas representações típicas destes parâmetros na Figura 2-11, para uma voz saudável à esquerda e para uma voz patológica à direita:

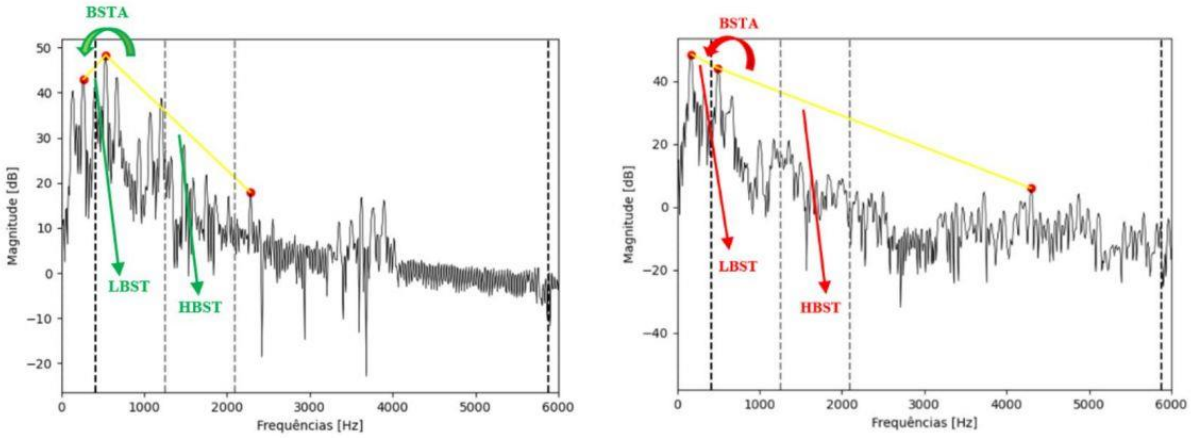


Figura 2-11 - *LBST, HBST* e *BSTA* típicos para orador saudável (esquerda) e com patologia de voz (direita) (retirado de [58])

Os máximos locais *FBME*, *SBME* e *TBME* podem ser identificados nos dois espectros, representados por três pontos vermelhos, um em cada faixa de frequências. Também os declives *LBST* e *HBST*, a amarelo, podem ser visualizados na figura, tal como o ângulo *BSTA* entre os dois declives. Observa-se que, no espectro da voz saudável, o *SBME* tem um valor mais elevado que o *FBME*, fazendo com que o declive *LBST* seja positivo e, consequentemente, que o ângulo *BSTA* seja claramente côncavo. Já no espectro da voz patológica, observa-se que o *FBME* é mais elevado que o SBME, tornando o declive *LBST* negativo e, neste caso, fazendo com que o ângulo *BSTA* seja aproximadamente raso. Verifica-se também que o declive *HBST* é negativo em ambos os espectros, embora no caso da voz saudável seja mais acentuado que no espectro da voz patológica.

No estudo foi utilizado um *corpus* com 54 sinais de fala contendo a vogal /a/ sustentada, sendo 39 desses sinais de oradores saudáveis e os restantes 15 de oradores diagnosticados com esclerose lateral amiotrófica (*ELA*) do tipo bulbar. No entanto, foram descartados os sinais de fala com oradores de idade inferior a 40, pois a *ELA* bulbar é tipicamente diagnosticada em faixas etárias superiores a 40 anos. Essa opção resultou em 18 sinais de fala de oradores saudáveis. Os sinais são divididos em tramas de 30 ms, obtidas com um deslocamento de 10 ms tendo sido aplicada a *Fast Fourier Transform* (*FFT*) a cada trama, de modo a obter o seu espectro.

Foram extraídos de cada espectro os valores dos três parâmetros *LBST*, *HBST* e *BSTA*, calculando-se o seu valor médio e desvio padrão para cada sinal de fala, sendo estes os valores utilizados no processo de discriminação. Para esse processo, são testados seis modelos diferentes de classificação automática, nomeadamente *Linear Support Vector Classifier* (*Linear SVC*), *Stochastic Gradient Descent* (*SGD*), *Gaussian Naive Bayes* (*GNB*), *Random Forest*, *LDA* e o método de *Regressão Logística*.

Os parâmetros foram utilizados tanto isoladamente como combinados, com cada um dos classificadores automáticos e os desempenhos foram comparados. O melhor desempenho foi obtido quando apenas o parâmetro *HBST* foi usado com o classificador *LDA*, tendo sido atingida uma taxa de acerto média de 75,0%. Para o parâmetro *LBST*, o melhor desempenho foi obtido com o classificador *Linear SVC*, tendo sido obtida uma taxa de acertos média de 64,5%. Estes resultados reforçam a conclusão de que a distribuição de energia no espectro do sinal de fala, bem como a análise da energia em diferentes zonas de frequência podem gerar parâmetros viáveis para sistemas de discriminação entre vozes saudáveis e patológicas.

# 3

# MATERIAIS E MÉTODOS

Neste capítulo apresentam-se os recursos utilizados neste estudo. Relativamente aos materiais, são descritas as bases de dados e as suas composições. Relativamente aos métodos, são descritas as técnicas usadas para processar e classificar os dados. Finalmente, apresentam-se as métricas utilizadas para aferir o desempenho ao longo das etapas do estudo.

## 3.1 Bases de Dados

Neste trabalho são usadas duas diferentes bases de dados, que contêm os sinais de fala objeto de estudo. A composição das bases de dados difere nas características dos oradores que das mesmas, em algumas patologias existentes e no conteúdo dos próprios sinais. No entanto, também partilham algumas características comuns, e por isso, relevantes para as análises em estudo.

As bases de dados são constituídas por ficheiros em formato *.wav*, monocanal e em todas existem segmentos de fala pertencentes a oradores saudáveis e a oradores patológicos. Em algumas dessas bases de dados estão presentes dezenas de patologias, no entanto, neste estudo, apenas serão consideradas as patologias abordadas na Secção 2.2.2. Assim, a composição das bases de dados, ou categorização, será a seguinte:

- ***control*:** Sinais de fala pertencentes a oradores saudáveis, cuja condição foi confirmada através de diagnóstico. O método de diagnóstico poderá ter sido diferente para cada base de dados. A estes sinais será atribuída a Classe 0.
- ***edema*:** Sinais de fala obtidos de oradores diagnosticados com edema de Reinke. A estes sinais será atribuída a Classe 1.
- ***neuro***: Sinais de fala de oradores diagnosticados com uma condição neurodegenerativa. Entre estas incluem-se doença de Huntington, doença de Parkinson ou esclerose lateral amiotrófica. A estes sinais será atribuída a Classe 2.
- ***nodulo*:** Sinais de fala obtidos de oradores diagnosticados com nódulos vocais. A estes sinais será atribuída a Classe 3.
- ***UVFP*:** Sinais de fala pertencentes a oradores diagnosticados com paralisia unilateral das pregas vocais. A estes sinais foi atribuída a Classe 4.

Os sinais de fala de oradores com edema de Reinke e nódulos vocais serão agrupados, dando origem à categoria:

- ***Patologias Laríngeas Fisiológicas* (*PhLP*):** Resulta da junção das classes 1 e 3, sendo atribuído a este conjunto a Classe 13.

As bases de dados não contêm as cinco categorias referidas, mas apenas quatro. Contêm as classes 0, 1 e 3, e a classe 2 ou 4.

Relativamente ao conteúdo dos sinais, os ficheiros podem conter uma vogal sustentada ou fala contínua. Nos estudos [50] e [51] verificou-se que sistemas baseados em parâmetros espectrais obtidos a partir de fala contínua produziram melhores resultados que os sistemas baseados numa vogal. Porém, verificou-se também que, quando usada fala contínua, a extração de parâmetros baseava-se na deteção de tramas vozeadas e que, em algumas vozes patológicas, devido à degradação da qualidade vocal, essa deteção poderia ser menos viável, comprometendo a extração dos parâmetros e, consequentemente, o desempenho do sistema. Essa foi uma das razões pela opção tomada de utilização de sinais de fala contendo uma vogal sustentada neste estudo.

Nem todas as bases de dados contêm sinais com fala contínua. A primeira base de dados utilizada, chamada base de dados da *Universidade de São Paulo*, que se descreverá um pouco mais adiante, apenas contém vogais sustentadas. Assim sendo, a opção por se usar uma vogal sustentada permitiu utilizar um maior número de dados, incluindo os da base de dados da *Universidade de São Paulo*.

As vogais sustentadas são estáveis, ou quase, e não contêm variações de entoação e efeitos de coarticulação. Para além disso, são desprovidas de outras características individuais do discurso do orador, como o débito de fala, ou o dialeto. Por essa razão, as vogais sustentadas são mais adequadas para a extração de parâmetros acústicos, sendo por essa razão as utilizadas para determinação do *jitter* e *shimmer*, por exemplo. Por outro lado, os parâmetros acústicos a ser utilizados neste trabalho serão mais significativos se forem extraídos a partir de sinais de fala estáveis. Por todas essas razões, optou-se pela utilização de sinais de fala contendo uma vogal sustentada [20] [25].

Quanto às vogais utilizadas, optou-se pela utilização da vogal /a/, sendo uma das razões, embora não a mais importante, a sua disponibilidade, pois tal como acontece com os sinais de fala contínua, nem todas as bases de dados contêm sinais de fala com todas as vogais. Voltando ao exemplo da base de dados da *Universidade de São Paulo*, os sinais de fala de oradores saudáveis apenas contêm a vogal /a/ sustentada. Assim sendo, a escolha de uma outra vogal impossibilitaria a utilização desta base de dados, pois não poderiam ser usados os sinais de fala de oradores saudáveis para controlo.

Uma outra razão prende-se com a representação na frequência. Cada vogal tem um espectro característico, com máximos locais, ou formantes, situados em gamas de frequência específicas, sendo essas gamas diferentes para cada vogal. Ao usar-se apenas uma vogal, não existirão diferenças significativas nas localizações dos formantes.

A escolha da vogal /a/ ficou a dever-se também à localização dos seus formantes. Pode ser visualizada na Figura 3-1 a representação das vogais não nasais em Português Europeu, obtidas a

partir de palavras lidas por nove oradores, através da localização dos seus dois primeiros formantes. A localização do primeiro formante é definida pela escala vertical, à direita, e a do segundo formante pela escala horizontal, em cima.

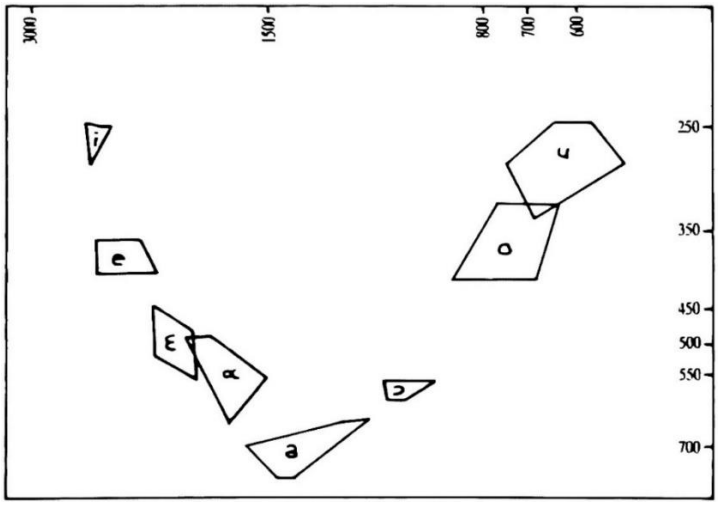


Figura 3-1 - Localização dos formantes das vogais em Português Europeu (retirado de [25])

Observa-se através da Figura 3-1 que a vogal /a/ tem o primeiro formante localizado numa frequência mais elevada, de aproximadamente 700 Hz. Esta é uma característica desejada, de modo a garantir uma distinção entre os máximos locais relativos ao primeiro formante, e às primeiras harmónicas. Além disso, a vogal /a/ é produzida com o trato vocal numa configuração amplamente aberta, permitindo que o som originado pela fonte glotal seja capturado com maior clareza. Por todas essas razões, optou-se pela utilização de sinais de fala contendo a vogal /a/ sustentada.

### 3.1.1 *Corpus* da Universidade de São Paulo (*USP*)

A primeira base de dados utilizada é a da *Universidade de São Paulo* (*USP*), também utilizada nos trabalhos [47] e [57] anteriormente abordados. A base de dados contém sinais de fala em ficheiro no formato *.wav*, adquiridos com frequência de amostragem de 22050 Hz, codificação em *Pulse Code Modulation* (*PCM*), quantificados com 16 bits e com uma duração mínima de dois segundos.

Os sinais de fala foram gravados no *Ambulatório de Voz do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo* (*HC-FMUSP*) com autorização do comitê de ética em pesquisa de seres humanos da *Universidade Federal de São Carlos*, protocolo 256/2010 [59]. Estes sinais foram gravados ao longo de dez anos e usados em vários estudos [60], [61] e [62].

Os sinais pertencentes a oradores saudáveis contêm a vogal /a/ sustentada, enquanto os referentes a oradores patológicos contêm as vogais /a/, /e/ e /i/ sustentadas. A distribuição dos oradores relativamente à condição, patologia, idade e género apresenta-se na Tabela 3-1.

Tabela 3-1 - Composição da Base de Dados *USP*

| Classe | Idade (Min – Max) | Idade (média ± d.p.) | Género (M – F) | Oradores |
|---|---|---|---|---|
| **control (0)** | 21 – 45 | 30,5 ± 9,0 | 10 – 5 | 15 |
| **edema (1)** | 28 – 48 | 38,5 ± 5,8 | 2 – 15 | 17 |
| **neuro (2)** | 22 – 90 | 58,5 ± 18,6 | 7 – 7 | 14 |
| **nodulo (3)** | 25 – 48 | 35,9 ± 7,5 | 2 – 13 | 15 |
| **Total** | 21 – 90 | 40,5 ± 14,9 | 21 – 40 | **61** |

Apresenta-se na Figura 3-2 a representação gráfica da distribuição de idades por classe, para uma melhor percepção.

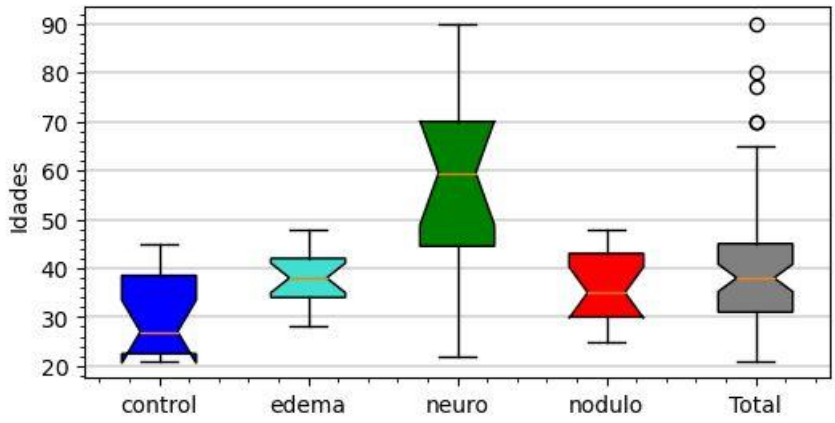


Figura 3-2 - Distribuição de idades por classe no *corpus USP*

Na Figura 3-3 apresenta-se a distribuição de oradores por género, dividida por classe.

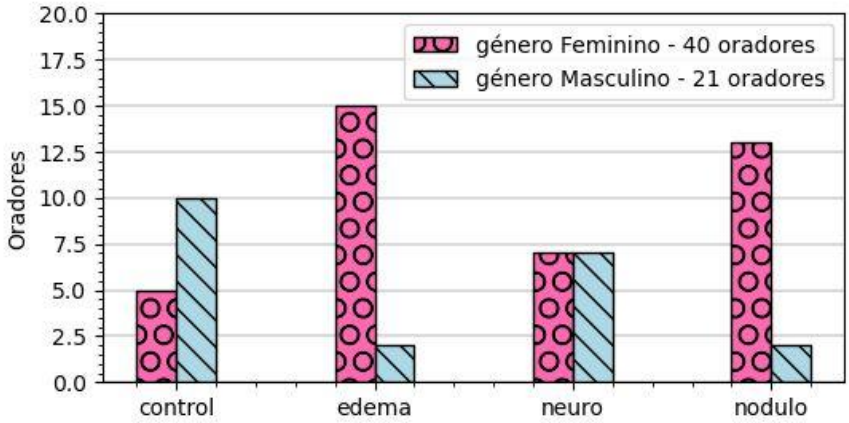


Figura 3-3 - Distribuição de oradores por género, por classe, no *corpus USP*

Verifica-se que a distribuição de idades varia significativamente entre as diferentes classes. A classe dos oradores saudáveis é composta predominantemente por oradores jovens, com cerca de metade dos oradores tendo menos de 25 anos. Em contraste, a classe dos oradores com patologias neuromusculares é composta por oradores mais velhos, com uma mediana próxima dos 60 anos. Essas diferenças nas distribuições etárias entre as classes não são ideais e podem influenciar os resultados, potencialmente aumentando as taxas de acerto.

Quanto à distribuição por género, verifica-se uma assimetria, com aproximadamente o dobro de oradores de género feminino em relação aos de género masculino. Essa proporção também não é ideal, especialmente porque a distribuição não é uniforme entre as classes. A classe dos oradores saudáveis tem uma proporção inversa, com o triplo de oradores de género masculino em comparação com os de género feminino. Esta distribuição por género não é a ideal e também pode aumentar as taxas de acerto, pois pode ser um factor de distinção entre as classes, influenciando os resultados.

No trabalho, foi utilizado um sinal de fala por orador, contendo a vogal /a/ sustentada.

### 3.1.2 *Corpus* do Massachusetts Eye and Ear Infirmary (*sMEEI*)

A segunda base de dados utilizada faz parte da *Massachusetts Eye and Ear Infirmary Voice Disorders Database* (*MEEI*) [63], já referida anteriormente. Esta é uma base de dados disponibilizada comercialmente pela *Kay Elemetrics Corporation*, sendo esta companhia, na condição de fornecedor, responsável pelas questões éticas, de proteção de dados, bem como pela categorização dos sinais.

A base de dados contém sinais de fala adquiridos com frequências de amostragem variáveis, entre os 10 e os 50 kHz e contendo a vogal /a/ sustentada ou fala contínua [50]. Os sinais de fala estão quantificados com 16 bits e os de oradores saudáveis têm uma duração de três segundos, enquanto os de oradores patológicos têm uma duração de um segundo.

Neste trabalho utiliza-se um subconjunto desta base de dados, usado nos trabalhos anteriormente descritos [44], [50] e [51]. Os sinais de fala deste subconjunto foram previamente reamostrados a 25 kHz, de modo a uniformizar as suas frequências de amostragem. Por razões que serão explicadas adiante, três sinais de fala foram retirados deste subconjunto, ficando o mesmo com ficheiros referentes a 151 oradores cuja distribuição relativamente à condição, patologia, idade e género apresenta-se na Tabela 3-2.

Tabela 3-2 - Composição do subconjunto usado da Base de Dados *MEEI*

| Classe | Idade (Min – Max) | Idade (média ± d.p.) | Género (M – F) | Oradores |
|---|---|---|---|---|
| **control (0)** | 22 – 52 | 35,2 ± 7,6 | 14 – 22 | 36 |
| **edema (1)** | 17 – 85 | 42,5 ± 14,7 | 8 – 29 | 37 |
| **nodulo (3)** | 13 – 48 | 29,4 ± 7,1 | 1 – 18 | 19 |
| **UVFP (4)** | 15 – 80 | 53,2 ± 17,1 | 30 – 29 | 59 |
| **Total** | 13 – 85 | 43,3 ± 16,5 | 53 – 98 | **151** |

Na Figura 3-4 apresenta-se a distribuição etária por classe.

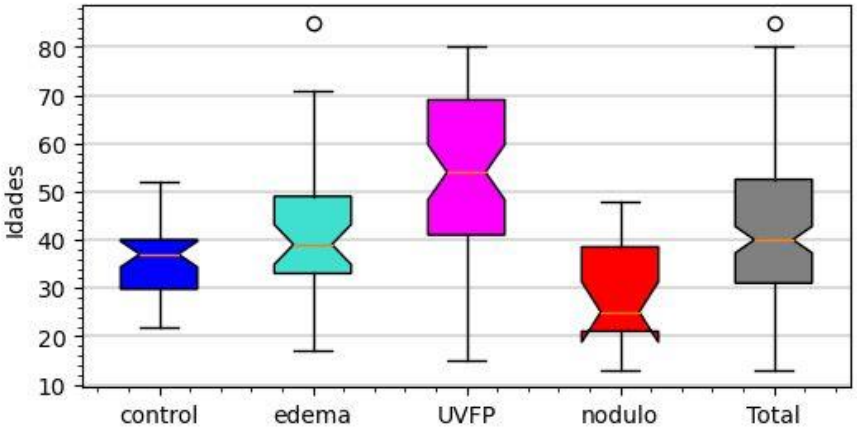


Figura 3-4 - Distribuição de idades por classe no *corpus sMEEI*

Verifica-se que, também nesta base de dados, a distribuição etária não é homogénea entre as classes. A classe dos oradores com nódulos vocais inclui maioritariamente oradores mais jovens, com menos de 25 anos, enquanto a classe *UVFP* é predominantemente composta por indivíduos mais idosos, com uma mediana de aproximadamente 55 anos. Também aqui estas diferenças podem potencialmente influenciar os resultados, uma vez que a idade é um factor importante na qualidade vocal. Essas variações etárias podem introduzir diferenças indesejadas nos sinais de fala das diferentes classes, tornando-se um factor de diferenciação entre elas.

De seguida, na Figura 3-5, apresenta-se a distribuição de oradores relativamente ao seu género, pelas diferentes classes.

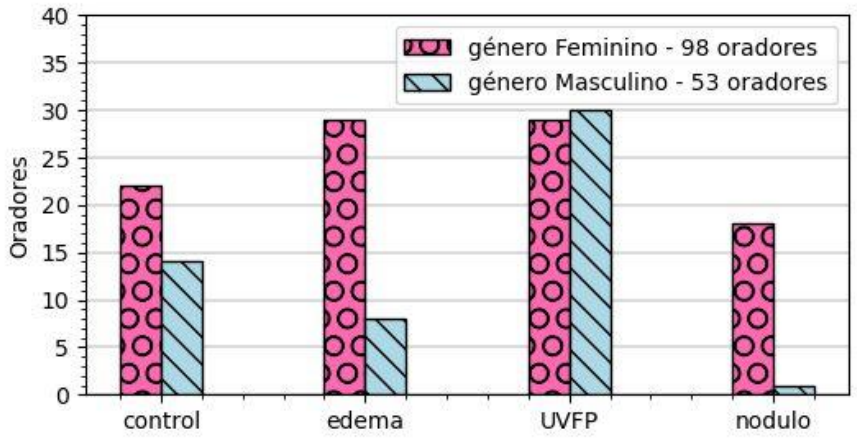


Figura 3-5 - Distribuição de oradores por género, por classe, no *corpus sMEEI*

Também nesta base de dados, a proporção entre oradores masculinos e femininos não é a ideal, pois observa-se que essa distribuição não é uniforme entre as classes.

Deste *corpus*, foi utilizado um sinal de fala por orador, contendo a vogal /a/ sustentada. O subconjunto utilizado neste trabalho será denominado *sMEEI*.

## 3.2 Métodos

Este estudo foi implementado em *Python* (versão 3.9.12), através da utilização de *Jupyter Notebooks* (versão 6.5.2) e os métodos utilizados estão, na sua grande maioria, disponíveis numa das suas bibliotecas, denominada *Scikit-learn* [64]. Esta biblioteca reúne um conjunto de funções que implementam algoritmos para diversas tarefas de *Machine Learning*, como classificação, regressão, *clustering* e redução de dimensionalidade, e é muito utilizada devido a esta variedade de métodos, bem como à facilidade de utilização dos mesmos. Por essas razões optou-se pela sua utilização neste estudo.

### 3.2.1 Divisão dos dados para classificação

Uma parte substancial do trabalho envolve a extração de parâmetros dos sinais de fala e a aplicação desses parâmetros num modelo de classificação. Esse modelo deverá, a partir dos dados, ajustar-se de modo a ser capaz de prever qual a classe, de um novo exemplo que venha a ser por ele avaliado. A capacidade de predição do modelo deve ser aferida, também a partir dos dados disponíveis, de modo a estimar-se a utilidade do mesmo. Ou seja, após a extração de parâmetros, os dados são divididos em duas partes, conforme se visualiza na Figura 3-6, adaptada de [65].

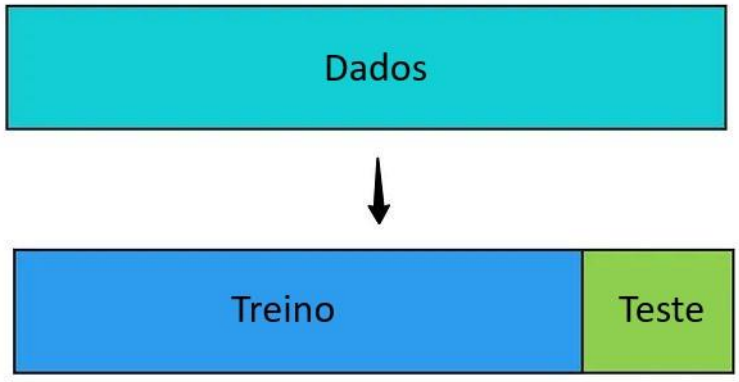


Figura 3-6 - Divisão dos dados nos conjuntos de treino e teste (adaptado de [65])

A primeira parte é utilizada para treinar o modelo de classificação. Durante essa fase, o modelo 'aprende' a identificar padrões nos dados que são característicos de diferentes classes. Esta aprendizagem é realizada através de algoritmos que ajustam os parâmetros internos do modelo para

minimizar erros na classificação. A este processo chama-se treino do modelo, e a parcela dos dados usada chama-se conjunto de treino.

A segunda parte dos dados é utilizada para avaliar o desempenho do modelo. A este processo chama-se teste do modelo e a parcela dos dados usada denomina-se conjunto de teste. O teste do classificador tem a função de aferir a capacidade de prever corretamente a classe dos dados, inclusive daqueles que o modelo não conhece. Por essa razão, é fundamental que o conjunto de teste seja mantido separado durante o treino do modelo.

Não existe uma proporção na divisão dos dados em conjuntos de treino e teste que garanta o melhor resultado em todos os casos. Existem abordagens que dão mais ênfase ao processo de teste, dividindo os dados em 70% para treino e 30% para teste, e existem outras que dão mais importância ao treino do modelo, utilizando 90% dos dados para treino e 10% para teste. Neste trabalho, optou-se por uma solução intermédia, usando 80% dos dados para treino e 20% para teste.

Optou-se também pela utilização de uma divisão estratificada, ou seja, mantendo nos subconjuntos a mesma proporção de exemplos de cada classe que existe no conjunto total dos dados. Dessa forma, assegura-se que os conjuntos de treino e teste são representativos da distribuição original das classes, ajudando a garantir que o modelo de classificação seja treinado e testado de maneira consistente em relação à distribuição real das classes.

Para maximizar o proveito obtido a partir dos dados, é utilizado um processo denominado *validação cruzada*. Este método implica a divisão dos dados em partes iguais, chamadas neste contexto de *folds*. Em cada iteração, um dos *folds* é utilizado para testar um modelo, enquanto os restantes são utilizados para treino do mesmo. O processo repete-se até que todos os *folds* tenham sido utilizados como conjunto de teste. Uma representação do processo, embora com uma divisão dos dados em três *folds* e não em cinco, como realizado neste trabalho, apresenta-se na Figura 3-7.

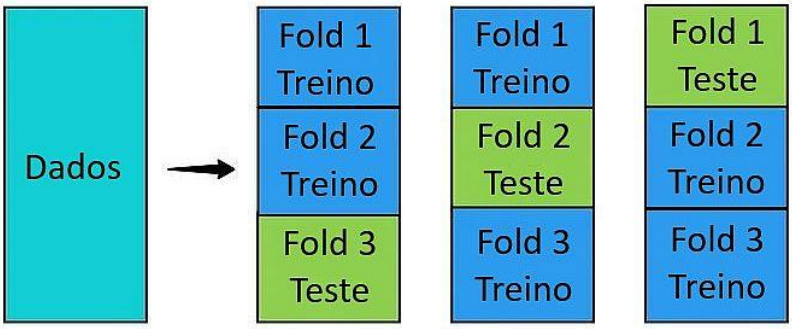


Figura 3-7 - Validação cruzada com 3 *folds* (adaptado de [65])

A validação cruzada, para além de maximizar a utilização dos dados disponíveis, pois usa todo o conjunto para treino e teste, oferece também uma avaliação mais robusta e fiável do modelo. Ao utilizar todos os dados para teste, a estimativa do desempenho do modelo é mais precisa. Neste trabalho optou-se pela utilização de validação cruzada com cinco *folds*, consistente com a divisão dos dados em 80% para treino e 20% para teste anteriormente referida.

### 3.2.2 Modelo de classificação: *Support Vector Machine* (*SVM*)

Vários classificadores foram testados neste trabalho, para além do modelo *SVM*, entre os quais se destacam os modelos *Random Forest, Extreme Gradient Boost* (*XGB*) e *Nearest Centroid*, pois obtiveram o melhor resultado em determinadas situações. No entanto, o modelo *SVM* obteve o melhor desempenho, ou muito próximo do melhor, em todas as situações. Por essa razão, a partir de certo ponto, optou-se pela utilização exclusiva do *SVM* e todos os resultados apresentados neste estudo são obtidos com esse modelo de classificação.

Quando dados de duas classes distintas podem ser separados por um plano, podem ser consideradas várias alternativas para a localização desse plano, sem que a opção escolhida comprometa, com base nos dados conhecidos, a eficácia da separação das classes. Um exemplo ilustrativo pode ser visualizado na Figura 3-8, onde os dados de duas classes perfeitamente separáveis são delimitados por quatro planos diferentes, todos capazes de realizar a separação com eficácia.

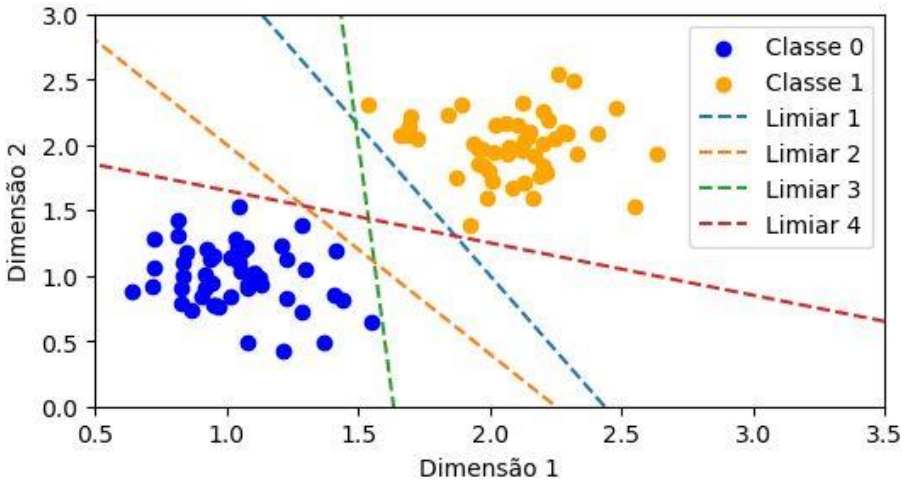


Figura 3-8 - Dados de duas classes separados por quatro possíveis planos

Verifica-se que qualquer um dos quatro limiares separa corretamente todos os dados conhecidos por classe. No entanto, na iminência do aparecimento de novos dados cujas classes não sejam conhecidas, logo, havendo a necessidade de os classificar com base no limiar de separação, a escolha de um plano menos adequado poderá comprometer o desempenho da classificação. Assim, qual o plano mais adequado, que garanta o melhor desempenho na classificação de dados futuros?

O método *SVM*, proposto em 1995 [66], não responde a esta questão, mas introduz um critério para a escolha do plano, o critério da máxima margem. O *SVM* é um método que procura determinar um hiperplano que, servindo como um limiar de separação, maximize a distância entre as duas classes.

Os dados de cada classe que estão mais próximos da outra classe são usados como vetores de suporte para a definição de dois hiperplanos, um para cada classe. Esses hiperplanos, conhecidos como margens, são posicionados de modo a estarem o mais afastados possível entre si, enquanto ainda tocam os pontos de dados mais próximos de ambas as classes. O hiperplano de separação entre

as duas classes é então definido como o plano que está equidistante dessas duas margens. Dessa forma, este hiperplano maximiza a margem, ou seja, a distância entre os hiperplanos das margens.

Porém, quando existe sobreposição de dados das duas classes, a definição das margens, e consequentemente, do hiperplano de separação, conforme com a descrição anterior, não é realizável. Nesses casos, utiliza-se uma técnica denominada *soft margin*, que introduz uma penalização associada a dados situados entre as duas margens ou do lado errado do limiar de decisão. Isso permite que alguns vetores de suporte estejam entre as duas margens. Esta técnica leva o *SVM* a procurar um equilíbrio entre maximizar a margem e minimizar a classificação errada dos dados. Na Figura 3-9 apresentam-se dois exemplos, para as duas situações descritas [67].

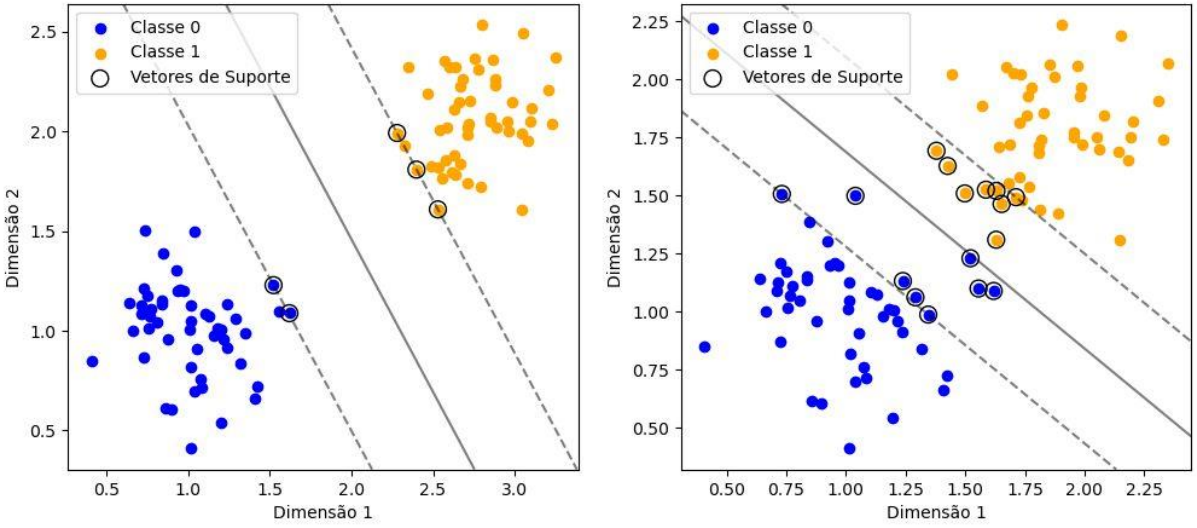


Figura 3-9 - Dois exemplos de separação de duas classes através de hiperplano obtido por *SVM*

Verifica-se, na figura à esquerda, que as margens, representadas por linhas tracejadas, são definidas por vectores de suporte, uma delas baseada em dois exemplos da classe *0* e a outra em três exemplos da classe *1*. Equidistante das duas margens, pode ser visualizado o hiperplano de separação. No exemplo à direita, observa-se a utilização da técnica de *soft margin*, onde, para permitir um maior afastamento entre as margens, são permitidos vectores de suporte no espaço entre as margens. O algoritmo de *SVM* utilizado neste trabalho usa esta técnica, permitindo a definição de um hiperplano de separação quando os dados das duas classes estão muito sobrepostos.

#### 3.2.2.1 *Kernels*

Nem sempre a melhor opção para separar os dados consoante a sua classe é um hiperplano reto, ou seja, linear. Em muitos casos, os dados poderão estar sobrepostos de tal forma que um hiperplano não linear, que permita a definição do limiar de separação entre as classes com uma maior flexibilidade, será mais adequado. Um hiperplano com essas características pode ser definido pelo *SVM*, desde que seja utilizado um *kernel* que o permita.

Os *kernels* no *SVM* são funções de transformação usadas para que, em casos onde as classes não sejam linearmente separáveis no espaço dimensional dos dados, possam sê-lo num espaço de maior dimensão. Os *kernels* transformam os dados para um espaço de alta, ou infinita dimensão, onde podem ser separados consoante a sua classe por um hiperplano linear. Contudo, se o *kernel* não for linear, o plano de separação no espaço dos dados também não será linear. Existem vários tipos comuns de *kernels* usados em *SVM*, cada um com características específicas que os tornam adequados para diferentes tipos de problemas. Na Figura 3-10 apresentam-se dois exemplos para a aplicação de *SVM*, com dois *kernels* diferentes, na separação dos dados consoante as suas classes.

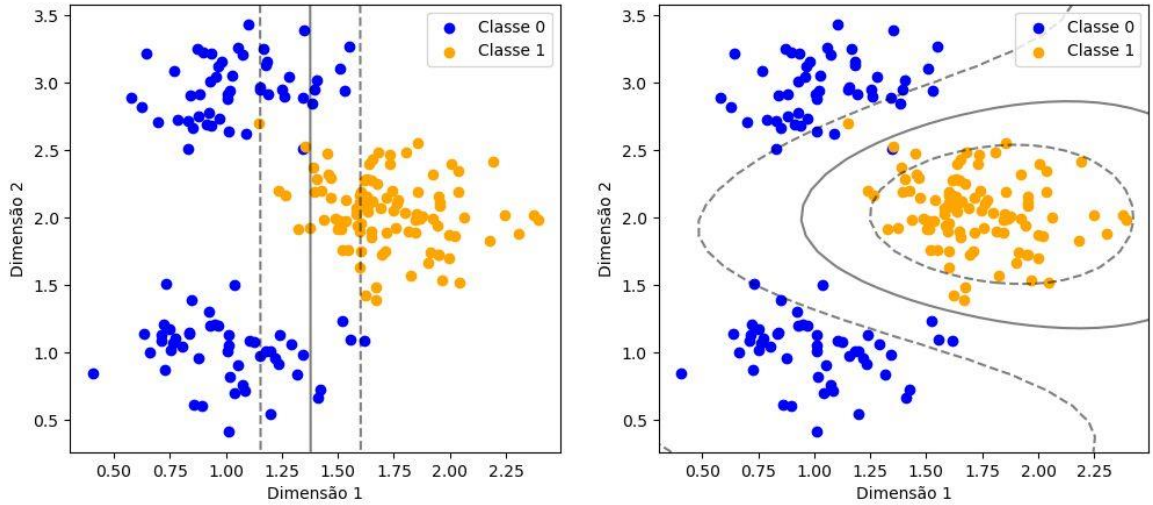


Figura 3-10 - Separação de duas classes através de *SVM* utilizando *kernel* linear (esq.) e gaussiano (dir.)

Observa-se na figura à esquerda, onde se usa um *kernel* linear, que o plano de separação não é adequado à disposição dos dados, verificando-se um número significativo de erros. Na figura à direita, por outro lado, onde se utiliza um *kernel* gaussiano denominado *Radial Basis Function* (*RBF*), a separação entre as classes é mais eficaz pois as margens, bem como o plano de separação, estão adaptados aos dados, verificando-se apenas dois exemplos erradamente classificados.

Neste trabalho foram testados os quatro *kernels* mais frequentemente utilizados em *SVM*, nomeadamente, o linear, polinomial, *RBF* e sigmoide. A utilização do *kernel RBF* produziu, na grande maioria dos casos, melhores desempenhos, e por essa razão, a partir de certo ponto, optou-se pela utilização exclusiva deste *kernel*. Portanto, os resultados apresentados neste trabalho foram obtidos utilizando *SVM* com *kernel RBF*, a menos que seja indicado explicitamente o uso de outro *kernel*.

#### 3.2.2.2 Hiperparâmetros

Os modos de funcionamento dos modelos de classificação podem, tipicamente, ser alterados com vista à sua optimização, através do ajuste de alguns parâmetros. Estes são denominados de hiperparâmetros, e existem também no modelo *SVM*. Este modelo tem dois hiperparâmetros cujos valores são usualmente ajustados para melhorar o desempenho do classificador.

Um deles é o parâmetro *gamma*, que define a influência de cada exemplo do conjunto de treino. O aumento do valor do parâmetro implica que o *SVM* tente ajustar-se mais a cada amostra do conjunto de treino. Embora se pretenda que o modelo aprenda o máximo com cada ponto do conjunto de dados, um valor demasiado elevado do parâmetro *gamma* levará o modelo a uma situação de *overfitting*.

O outro hiperparâmetro habitualmente ajustado é o parâmetro *C*, que define o equilíbrio entre a maximização do espaço entre as margens e a tolerância a dados incorretamente classificados. Um valor mais elevado para o parâmetro *C* aumenta a penalização pelos dados fora da zona da sua classe, fazendo com que o *SVM* defina margens mais próximas, enquanto um valor menor de *C* fará o *SVM* ser mais tolerante a erros e maximizar a distância entre as margens das classes. O efeito do parâmetro *C* na definição das margens e do hiperplano de separação pode ser visualizado na Figura 3-11.

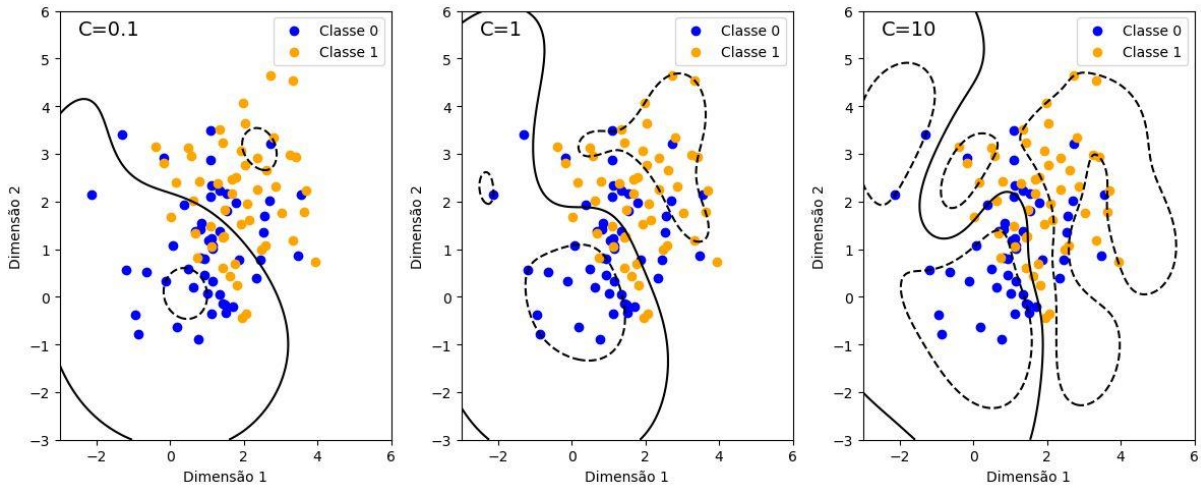


Figura 3-11 - Separação de duas classes através de hiperplano obtido por *SVM* com três diferentes valores de *C*

Observa-se, através da Figura 3-11, o efeito do valor do parâmetro *C* na definição das margens e, consequentemente, do plano de separação. À esquerda, para um valor menor de *C*, verifica-se uma maior distância entre as margens, bem como um maior número de pontos entre as duas margens, ou seja, fora da zona da sua classe. À direita, observa-se que, para um valor mais elevado de *C*, as margens estão mais próximas e, embora existam alguns erros, as zonas das classes contêm mais pontos do que nos outros dois exemplos.

Neste trabalho, porém, optou-se por não se ajustar os hiperparâmetros. Esse ajuste implica a divisão dos dados em treino, teste e validação, pois torna-se necessário uma terceira parcela de dados que verifique a progressão do desempenho do modelo até que este atinja um ponto ótimo. Este terceiro conjunto, de validação, é retirado do conjunto de treino, o que implica que o modelo tenha menos dados disponíveis para o seu treino. Na realização do trabalho verificou-se que a diminuição do conjunto de treino tinha, na maioria dos casos, mais peso no desempenho do sistema do que as potenciais melhorias obtidas com o ajuste de hiperparâmetros. Por essa razão, testes e resultados apresentados neste trabalho são obtidos com os valores de hiperparâmetros definidos por defeito no modelo *SVM*.

### 3.2.3 Redução de dimensionalidade

Ao contrário do que poderia ser intuitivo, um número maior de parâmetros não equivale a um melhor desempenho do classificador. Quando os dados possuem um número elevado de parâmetros, ou dimensões, um efeito indesejável chamado *Curse of Dimensionality*, ou *fenómeno de Hughes* [68], pode ocorrer. Este efeito consiste no seguinte: dado um número fixo de amostras, o desempenho esperado do classificador melhora com o aumento de dimensões até um determinado ponto, a partir do qual começa a degradar-se à medida que a dimensionalidade continua a aumentar. Este efeito pode ser visualizado no exemplo apresentado na Figura 3-12.

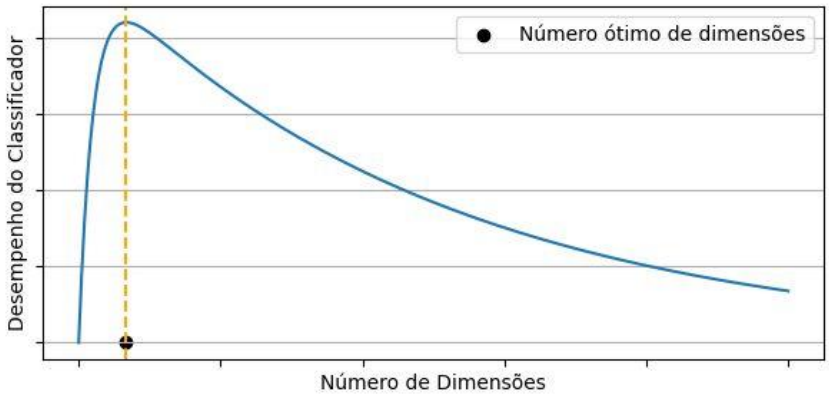


Figura 3-12 - Desempenho estimado de um classificador em função da dimensionalidade

No exemplo mostrado na figura, observa-se que o desempenho do classificador melhora com o aumento da dimensionalidade, mas apenas até um determinado número de dimensões, a partir do qual, o desempenho começa a piorar. Este efeito observa-se principalmente em classificadores baseados em distâncias, como é o caso do *SVM*.

Neste trabalho, existem situações onde o número de dimensões é muito elevado, atingindo 40 dimensões por amostra. Assim, torna-se necessário que a dimensionalidade dos dados seja reduzida.

#### 3.2.3.1 Principal Component Analysis (*PCA*)

A análise de componentes principais, ou *Principal Component Analysis* (*PCA*) é o método de redução de dimensionalidade utilizado neste estudo e consiste numa transformação linear que projeta os dados para um novo espaço dimensional onde as novas dimensões, denominadas de componentes principais, são combinações lineares das originais.

A primeira componente principal é a transformação linear que preserva a máxima variância possível dos dados originais, a segunda componente principal é a transformação linear, não correlacionada com a primeira, que contém a máxima variância, e assim sucessivamente [69]. Dessa forma, ao aplicar-se a transformação *PCA* aos dados, podem ser mantidas as primeiras componentes principais, consoante a dimensionalidade pretendida, e descartadas as restantes, garantindo-se que se preserva a máxima variância possível dos dados originais no processo.

No entanto, a variância dos dados é apenas um critério para a escolha das dimensões, não sendo garantido que seja o mais adequado. Apresenta-se, na Figura 3-13, um conjunto de amostras pertencentes a duas classes.

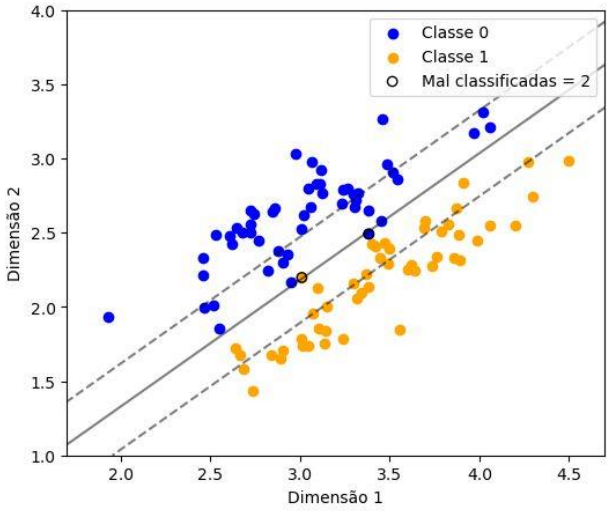


Figura 3-13 - Dados pertencentes a duas classes, com aplicação de classificador *SVM*

Após aplicação do *PCA* aos dados, estes podem ser visualizados representados nas duas componentes principais resultantes, na Figura 3-14. À esquerda, observa-se a aplicação de um classificador *SVM* sobre a primeira componente principal. À direita, o *SVM* foi aplicado sobre a segunda componente principal.

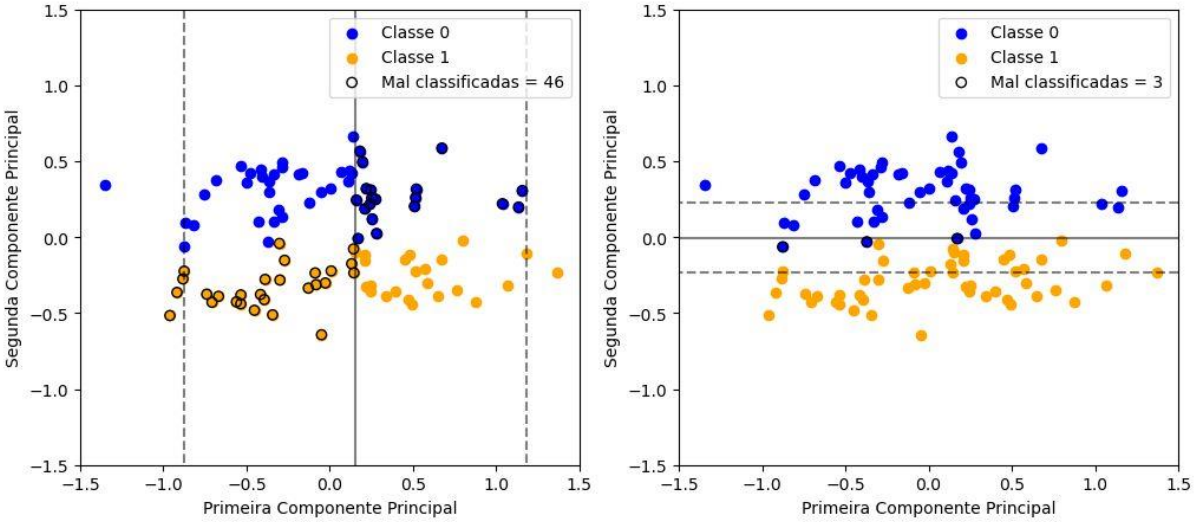


Figura 3-14 - Dados representados nas suas componentes principais, com aplicação de *SVM* sobre a componente principal (esq.) e sobre a segunda componente principal (dir.)

Observa-se, pela figura da esquerda, que a primeira componente principal, embora contenha a maior variância dos dados, não contém a maior informação para a discriminação entre as classes, pois teve 46 erros. Por outro lado, a segunda componente principal, embora não seja a que contém a maior variância dos dados, contém mais informação para a discriminação entre as classes, pois teve apenas 3 erros, como se observa na figura da direita.

Mesmo a discriminação sobre a segunda componente principal obteve um resultado pior do que a discriminação efetuada sobre os dados nas suas dimensões originais, pois aí a discriminação entre as classes teve apenas 2 erros.

Portanto, apesar do *PCA* ser um método útil, e por vezes necessário, para a redução de dimensionalidade, sendo por essa razão aplicado neste trabalho, como se verá adiante, deve-se ter presente que a sua utilização acarreta o risco de se perder informação importante.

#### 3.2.3.2 Feature Selection

O termo *Feature Selection* não se refere a um método, mas a um conjunto de métodos para redução de dimensionalidade. Ao contrário do *PCA*, onde existe uma transformação dos dados para que a redução de dimensionalidade mantenha a máxima variância possível, as técnicas de *Feature Selection* consistem na escolha de algumas das dimensões existentes de acordo com um determinado critério, sem qualquer transformação dos dados.

Existem várias técnicas e critérios para a seleção de dimensões a manter, sendo uma delas a denominada *Univariate Feature Selection* (*UFS*), que consiste na análise isolada de cada uma das dimensões dos dados, de acordo com um critério ou métrica definidos. Neste trabalho, é treinado um classificador *SVM* para cada uma das dimensões e avaliados os seus desempenhos. Esses desempenhos são o critério para a escolha das dimensões a manter, sendo descartadas as dimensões que deram origem a classificadores com piores desempenhos.

### 3.2.4 Normalização dos dados: *Standard Scaler*

Um factor que pode ter uma influência indesejável no desempenho de um classificador é a diferença de escalonamentos. É frequente que dados multidimensionais tenham diferentes gamas de valores para cada dimensão, sendo um exemplo típico, a caracterização de uma pessoa pela sua idade, peso e altura. Neste caso, a amostra terá três dimensões, cada uma delas com a sua própria gama de valores.

Esta heterogeneidade nas escalas das dimensões pode influenciar negativamente o desempenho de alguns classificadores, especialmente dos classificadores baseados em distâncias, como o *SVM*. Neste modelo, essa influência é maior se for usado um *kernel* não linear, como acontece neste trabalho, onde se utiliza o *kernel RBF*. Isto acontece porque quando as dimensões dos dados têm escalas muito diferentes, a dimensão com a maior escala pode dominar a distância entre os pontos. O *SVM* utiliza essas distâncias para encontrar o hiperplano de separação ótimo entre as classes. Se

uma dimensão domina a distância, o modelo pode não considerar convenientemente as outras dimensões, podendo levar a um modelo mal ajustado.

Nestes casos, torna-se conveniente a utilização de uma técnica de reescalonamento, que transforme os dados de modo que todas as suas dimensões tenham a mesma escala. Neste trabalho, utiliza-se uma ferramenta chamada *Standard Scaler*. Esta técnica consiste na determinação, para cada dimensão dos dados, da média e desvio padrão e posterior utilização desses valores para transformar os dados de modo que fiquem com média nula e variância unitárias, conforme se expressa na Equação 3-1:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3-1}$$

onde $z_{ij}$ é a i-ésima amostra da dimensão j após transformação, $x_{ij}$ é a i-ésima amostra original da dimensão j, $\mu_j$ é a média dos valores da dimensão j, e $\sigma_j$ é o desvio padrão dos valores da dimensão j.

Um exemplo de aplicação de *Standard Scaler*, e do seu efeito no desempenho do *SVM* pode ser observado na Figura 3-15.

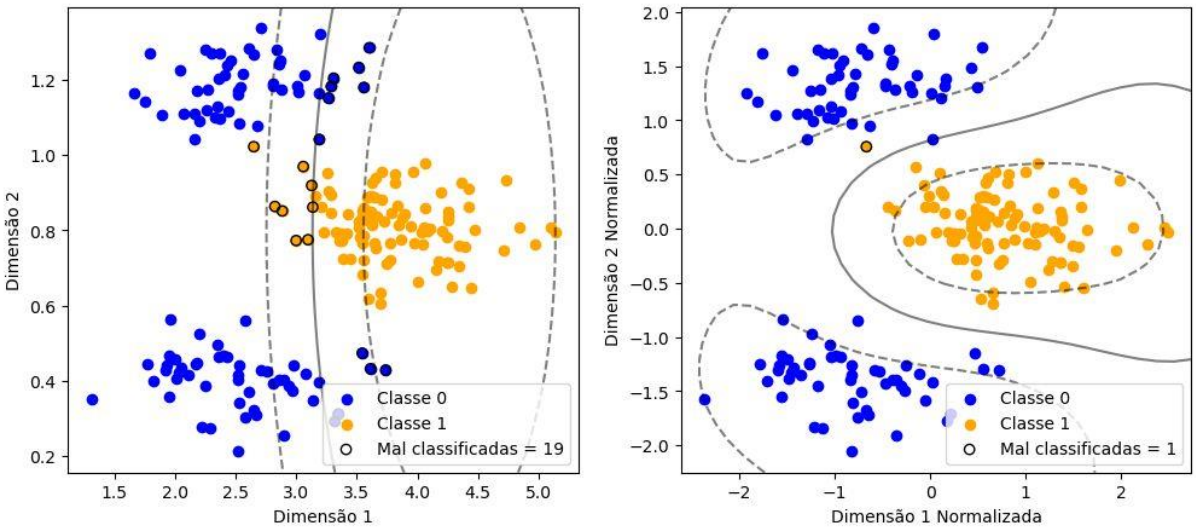


Figura 3-15 - Dados pertencentes a duas classes; à esquerda, originais; à direita, após normalização

Observa-se, na figura à esquerda, que a gama de valores da dimensão 1 é aproximadamente quatro vezes superior à gama da dimensão 2. Esta diferença de escalas afetou o desempenho do *SVM*, verificando-se que as margens e o hiperplano de separação não se ajustaram devidamente aos dados. Na figura à direita, por outro lado, os dados transformados com escalas semelhantes nas duas dimensões permitiram que o *SVM* definisse as margens e um hiperplano de separação convenientemente adaptados aos dados. O exemplo apresentado demonstra claramente a utilidade da normalização dos dados e o seu efeito no desempenho do *SVM*.

## 3.3 Métricas de avaliação de desempenho

Os exemplos com classificadores vistos até este ponto serviram apenas como demonstrações complementares às descrições dos métodos utilizados. Por essa razão, a avaliação do desempenho dos classificadores foi realizada de forma percetiva, baseada na inspeção visual. No entanto, é necessária uma forma mais rigorosa de avaliar os desempenhos dos diferentes sistemas, bem como dos testes efetuados. Com esse fim serão utilizadas métricas de avaliação que permitirão quantificar o desempenho dos classificadores, validar os seus resultados e fundamentar as conclusões obtidas.

### 3.3.1 Classificação multiclasse

Para o cálculo das métricas de desempenho, será útil a obtenção de uma matriz de confusão, pois esta permite a determinação das métricas a partir dos seus valores. A matriz de confusão é uma ferramenta de visualização que simplifica a análise do desempenho do classificador. Consiste numa tabela que distribui as predições realizadas na fase de teste através das linhas, correspondentes às verdadeiras classes das amostras, e das colunas, que correspondem às classes atribuídas no teste.

Na Tabela 3-3 observa-se uma matriz de confusão para a classificação entre 3 classes, onde se verifica a distribuição, pelas linhas, das verdadeiras classes das amostras (*Real* $C_i$) e pelas colunas, das classes atribuídas na classificação (*Predicted* $C_i$). Quanto aos elementos da tabela, os elementos $T_i$ são as amostras da classe *i* corretamente classificadas como pertencendo à classe *i*, e os elementos F*ij* são os elementos da classe *i* incorretamente classificados como pertencendo à classe *j*.

Tabela 3-3 - Matriz de confusão para classificação entre 3 classes

| | **Predicted $C_0$** | **Predicted $C_1$** | **Predicted $C_2$** |
|---|---|---|---|
| **Real $C_0$** | $T_0$ | $F_{01}$ | $F_{02}$ |
| **Real $C_1$** | $F_{10}$ | $T_1$ | $F_{12}$ |
| **Real $C_2$** | $F_{20}$ | $F_{21}$ | $T_2$ |

A partir da matriz de confusão poderão ser calculadas as métricas de avaliação. A primeira, e mais importante, métrica utilizada é a taxa de acerto (*Accuracy* – *Acc*). Esta métrica representa a proporção de previsões corretas, relativas a todas as classes, entre todas as previsões realizadas, e pode ser expressa pela Equação 3-2:

$$Acc = \frac{T_0 + T_1 + T_2}{T_0 + F_{01} + F_{02} + F_{10} + T_1 + F_{12} + F_{20} + F_{21} + T_2} \tag{3-2}$$

Esta é a única métrica relativa a classificação multiclasse utilizada neste trabalho. No entanto, existem situações em que, a partir de uma classificação multiclasse, se pretende obter métricas relativas a uma classe específica. Nesse caso, a classificação passa a denominar-se de ‘um contra

todos' (*One vs. All – OvA*), e assumindo que a classe de interesse é a classe *0*, as outras duas classes, *1* e *2*, são agregadas numa única classe. Nessa situação, a taxa de acertos relativa a essa classe será dada pela Equação 3-3:

$$Acc_0 = \frac{T_0 + T_1 + T_2 + F_{12} + F_{21}}{T_0 + F_{01} + F_{02} + F_{10} + T_1 + F_{12} + F_{20} + F_{21} + T_2} \quad (3\text{-}3)$$

Outras métricas, apesar de serem binárias, podem ser calculadas a partir da matriz de confusão obtida com classificação multiclasse, quando se considera uma classificação *OvA*. Uma delas é o *Precision*, ou *Predictive Positive Value* (*PPV*). Esta métrica mede a proporção de amostras classificadas como pertencentes a uma determinada classe que realmente são dessa classe. O *PPV* da classe *0* pode ser expresso pela Equação 3-4:

$$Precision_0 = PPV_0 = \frac{T_0}{T_0 + F_{10} + F_{20}} \quad (3\text{-}4)$$

O $PPV_0$ avalia a fiabilidade do modelo ao classificar as amostras como sendo da classe *0*.

Uma outra métrica, também binária, complementar ao *PPV*, é o *Recall*, ou *True Positive Rate* (*TPR*). Esta métrica também pode ser calculada a partir dos valores da matriz de confusão multiclasse e mede a proporção de amostras verdadeiramente pertencentes a uma determinada classe que foram corretamente identificados como sendo dessa classe. A determinação do *TPR* relativo à classe *0* pode ser expressa através da Equação 3-5:

$$Recall_0 = TPR_0 = \frac{T_0}{T_0 + F_{01} + F_{02}} \quad (3\text{-}5)$$

O *Recall* avalia a capacidade, por parte do modelo, de identificar todas as amostras de uma determinada classe como pertencendo a essa classe.

### 3.3.2 Classificação binária

Quando a classificação é efetuada entre duas classes, ou seja, binária, o cálculo das métricas de avaliação simplifica-se, pois existem menos variáveis.

Na Tabela 3-4 observa-se uma matriz de confusão para a classificação entre duas classes. Neste caso, as classes são tipicamente chamadas de *Positivas* ou *Negativas*. Verifica-se, tal como na classificação multiclasse, a distribuição, pelas linhas, das verdadeiras classes das amostras de treino (*Real* $C_i$) e pelas colunas, das classes atribuídas na classificação (*Predicted* $C_i$). Quanto aos elementos da tabela, *TP*, ou *True Positives*, refere-se às amostras Positivas classificadas como Positivas, *TN*, ou *True Negatives*, refere-se às amostras Negativas classificadas como Negativas, *FP*, ou *False Positives*, refere-se às amostras Negativas classificadas como Positivas e *FN*, ou *False Negatives*, às amostras Positivas classificadas como amostras Negativas.

Tabela 3-4 - Matriz de confusão para classificação binária

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| **Real Positive** | TP | FN |
| **Real Negative** | FP | TN |

Tal como na classificação multiclasse, as métricas podem ser obtidas a partir da matriz de confusão. Assim, a taxa de acerto é dada pela Equação 3-6:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{3-6}$$

Quanto ao *PPV*, pode ser determinado pela Equação 3-7:

$$Precision = PPV = \frac{TP}{TP + FP} \tag{3-7}$$

Finalmente, o *TPR* pode ser calculado pela Equação 3-8:

$$Recall = TPR = \frac{TP}{TP + FN} \tag{3-8}$$

Conforme anteriormente mencionado, as métricas *TPR* e *PPV* são complementares. Ambas podem ser ajustadas pelo limiar de decisão, que define a atribuição das classes às amostras. No entanto, ao melhorar o valor de uma métrica através desse ajuste, o valor da outra tende a piorar. Portanto, um potencial ajuste do limiar de decisão será uma solução de compromisso, pois o resultado pretendido será a maximização simultânea dos valores de ambas as métricas.

Uma forma de aferir que ambas as métricas têm valores elevados será através da utilização de uma outra métrica, chamada *F1-Score*. A métrica *F1-Score* consiste numa média harmónica entre as métricas *Precision* e *Recall*, sendo obtida a partir destas através da Equação 3-9:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3-9}$$

Esta métrica pode também ser representada, tal como as anteriormente referidas, em função de *TP*, *TN*, *FP* e *FN*, como se expressa na Equação 3-10:

$$F1 = \frac{2 \times TP}{2 \times TP + FN + FP} \tag{3-10}$$

A métrica *F1-Score* é particularmente adequada para avaliar o desempenho de sistemas de classificação em conjuntos de dados em que as classes não sejam balanceadas, como é o caso das bases de dados usadas neste estudo. Nessas situações, o classificador pode não classificar corretamente as amostras e ter tendência a atribuir mais amostras de teste à classe com mais exemplos, devido a um viés causado pela diferença de amostras entre as classes. O *F1-Score* avalia esse problema, pois tem em conta tanto o *TPR* como o *PPV*. Sendo estas duas métricas complementares, caso exista um viés na classificação, irá baixar significativamente o valor de uma delas e isso repercutir-se-á no valor da métrica *F1-Score*.

### 3.3.3 Estatística

A obtenção de resultados através da divisão dos dados em treino e teste, conforme descrito na Secção 3.2.1 apresenta uma desvantagem que o uso da validação cruzada não consegue mitigar: os resultados são muito dependentes da divisão inicial dos dados. Para mitigar de forma eficaz essa dependência, a divisão dos dados é repetida várias vezes, sempre com obtenção de *folds* diferentes.

Ao repetir o processo várias vezes, registando os resultados em cada iteração, não se obtém apenas um resultado por métrica, mas sim tantos resultados quantas as repetições. O objetivo não é analisar cada um desses resultados individualmente, mas sim calcular medidas estatísticas a partir do conjunto de resultados. Dessa forma, é possível calcular os valores médios para cada métrica, que são mais fiáveis do que os valores obtidos numa única iteração.

Também é possível determinar os desvios padrão dos valores de cada métrica, que representam a variação dos resultados. Essa variação é fundamental para a determinação de uma margem de confiança, que reflete a precisão com que se pode estimar o valor médio de uma métrica.

Embora os diferentes conjuntos de *folds* obtidos não sejam independentes, pois provêm do mesmo conjunto de dados, são suficientemente distintos para evidenciar a variação dos resultados. Essa variação permite a estimativa da incerteza associada aos valores das métricas, realizada através do cálculo das margens de confiança.

Partindo do pressuposto que os valores obtidos pelas repetições seguem, aproximadamente, uma distribuição normal, a margem de confiança pode ser calculada a partir do desvio padrão. Neste trabalho utiliza-se a margem de confiança de 95%, que pode ser obtida através da Equação 3-11:

$$mC_{95\%} = \pm 1{,}96 \times \sigma \tag{3-11}$$

onde σ é o desvio padrão dos valores da métrica obtidas nas repetições.

Apesar de conterem um erro associado, pois os valores das métricas não seguem exactamente uma distribuição normal, as margens de confiança obtidas através do desvio padrão são utilizadas neste trabalho, pois fornecem uma medida prática e intuitiva da variação esperada para o valor médio de uma determinada métrica. Ao utilizar uma margem de confiança de 95%, espera-se que qualquer novo valor para uma determinada métrica caia dentro desse intervalo em 95% das vezes.

# 4

# BANDAS DE ENERGIA

Neste capítulo apresenta-se o trabalho realizado relativo às bandas de energia, sendo detalhados todos os passos para determinar quais as bandas de energia e de variação de energia que permitem discriminar os sinais de fala entre saudáveis ou patológicos, assim como, nos patológicos, identificar qual a patologia existente.

## 4.1 Pré-processamento dos sinais de fala

Antes de extrair quaisquer parâmetros ou realizar qualquer análise aos sinais de fala, alguns passos de pré-processamento foram aplicados aos mesmos no sentido de torná-los mais uniformes entre si. Dessa forma, pretende-se minimizar possíveis influências de características inerentes aos sinais, mas consideradas sem interesse, como a duração ou o volume da fonação, nos resultados obtidos.

Após a extração dos sinais de fala dos ficheiros *.wav* para o ambiente *Python*, as suas amplitudes foram normalizadas para que a gama de amplitudes para todos os sinais ficasse contida entre os valores -1 e 1. Esta normalização foi efetuada através da determinação do valor máximo absoluto em cada sinal de fala e posterior divisão de todas as suas amostras por esse valor, conforme expresso na Equação 4-1:

$$x_n[n] = \frac{x[n]}{max|x[n]|} \tag{4-1}$$

Desta forma garante-se a uniformidade de todos os sinais de fala relativamente à sua amplitude, mais concretamente, à gama de valores que esta pode tomar.

O passo seguinte consiste na remoção de silêncios no início e final dos sinais de fala, para focar a análise na porção vozeada. A remoção de silêncios é efetuada através de um processo baseado em [70]. Os sinais são divididos em tramas de 15 ms, conforme proposto no método original, sem sobreposição. As tramas são percorridas da esquerda para a direita e as suas potências são calculadas e comparadas com um limiar k1. Quando a potência de uma trama excede esse limiar, a potência da

trama seguinte é comparada com um outro limiar k2 e, se o exceder, a primeira trama é considerada o início do sinal de fala e as tramas anteriores são descartadas. Se, por outro lado, a trama seguinte tiver uma potência inferior a k1, o processo recomeça com a procura de uma trama cuja potência seja superior ao limiar k1. Para a deteção do limite final da porção vozeada, o processo é simétrico, sendo as potências das tramas comparadas com o limiar k1, mas percorridas da direita para a esquerda.

Apresenta-se, a título de exemplo, na Figura 4-1, o efeito da remoção de silêncios num sinal de fala da base de dados *USP*.

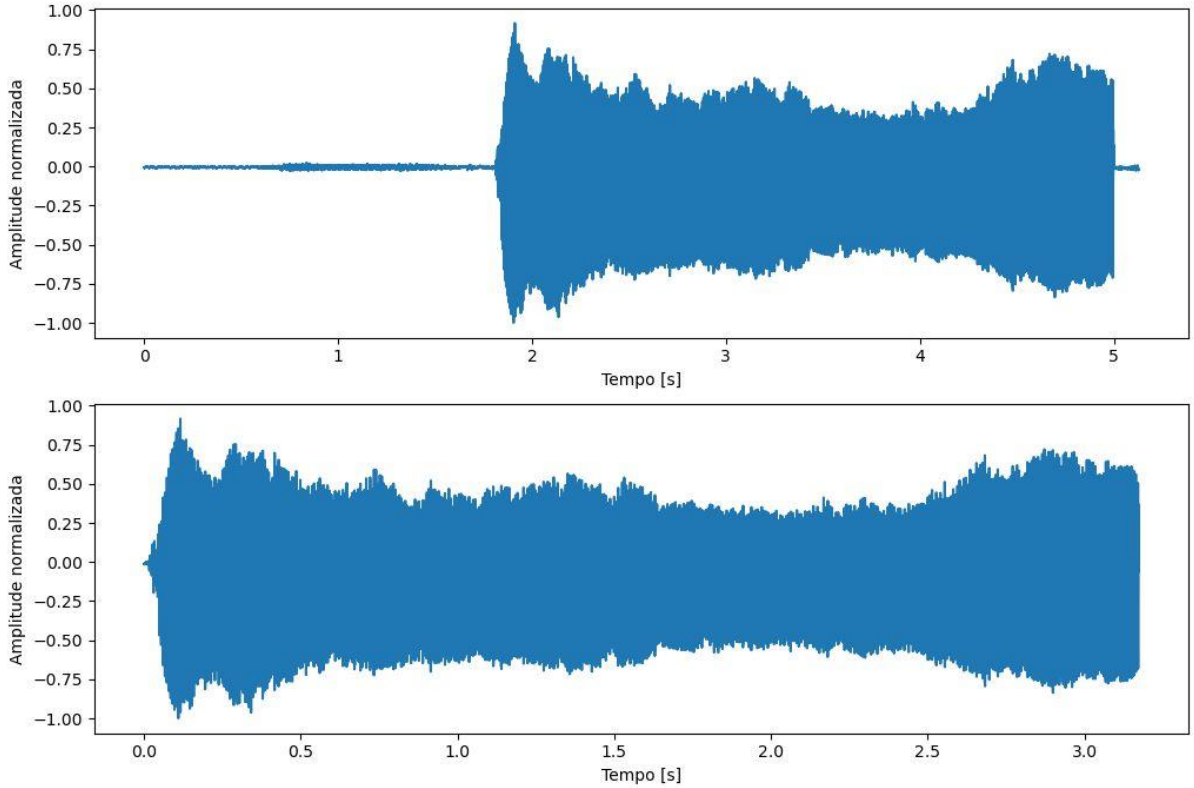


Figura 4-1 - Sinal de fala antes (em cima) e após (em baixo) da remoção de silêncios

No exemplo da Figura 4-1, observa-se que a aplicação do método de remoção de silêncios descarta cerca de 40% do sinal de fala, pois o sinal original tinha uma duração superior a cinco segundos e o sinal resultante tem uma duração um pouco superior a três segundos. Esta remoção evita o processamento desnecessário de partes que não contêm fala, evitando também potenciais efeitos na análise e nos resultados obtidos.

O passo seguinte no pré-processamento dos sinais de fala é a uniformização dos mesmos relativamente às suas durações. Para cada base de dados, foi determinada a menor duração entre todos os sinais de fala, ajustando-se todos os sinais para essa duração, mantendo a parte inicial dos mesmos e descartando a duração excedente.

Na Figura 4-2, podem ser visualizadas as durações dos sinais de fala de cada base de dados, identificados quanto à sua classe:

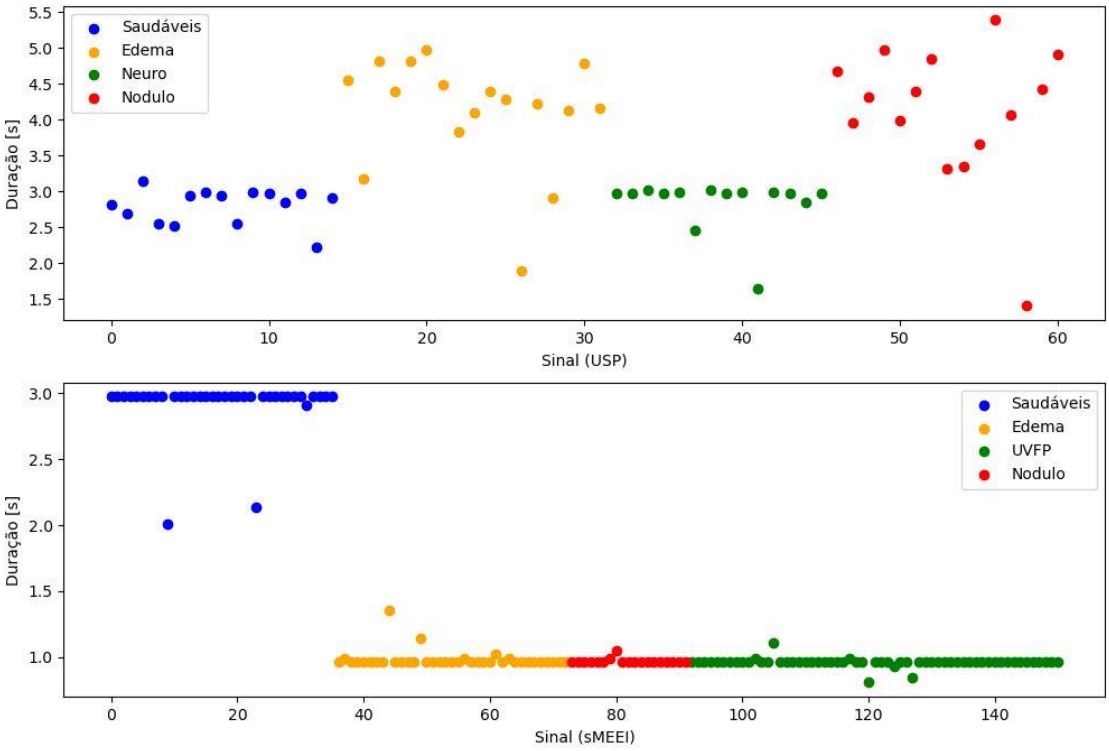


Figura 4-2 - Duração dos sinais de fala da base de dados *USP* (em cima) e *sMEEI* (em baixo)

Os sinais de fala apresentam diferentes durações e, principalmente, essas durações variam consoante a classe, o que é mais evidente na base de dados *sMEEI*. Nesta base de dados, os sinais de oradores saudáveis têm o triplo da duração dos sinais de oradores com patologias da fala. Essas diferenças de duração, especialmente as correlações entre duração e classes, poderiam afetar os resultados.

Para mitigar esse efeito, optou-se por limitar a duração dos sinais à menor duração em cada base de dados. Embora isso resulte na exclusão de partes potencialmente úteis dos sinais, essa abordagem foi considerada a mais adequada. Dessa forma, os sinais de fala da base de dados *USP* foram limitados aos primeiros 1,406 segundos da parte vozeada e os sinais de fala da base de dados *sMEEI* aos primeiros 0,810 segundos da parte vozeada.

Após a remoção de silêncios e a limitação das durações, a amplitude dos sinais de fala foi novamente normalizada, através da aplicação da Equação 4-1, para garantir que as gamas de amplitude dos sinais de fala se mantenham homogéneas para todos os oradores.

## 4.2 Obtenção das bandas de energia

Após o pré-processamento dos sinais de fala, o passo seguinte foi a obtenção das energias por banda. Essas energias foram calculadas ao longo dos sinais pois, para além das energias em si, pretende-se também obter informações acerca da evolução das mesmas ao longo do tempo.

Para isso, os sinais de fala foram decompostos em tramas de 30 ms, obtidas com um andamento de 10 ms. A duração das tramas foi definida com o objetivo de conter pelo menos dois períodos glotais completos. Considerou-se que um período glotal terá, no máximo, 30 ms, logo, assumiu-se que a frequência fundamental mínima seria de, aproximadamente, 67 Hz. Esta estimativa pareceu adequada, pois não se esperam sinais de fala com frequência fundamental abaixo desse valor.

As tramas têm uma sobreposição de 20 ms para mitigar os efeitos da potencial fuga espectral, que ocorre quando o conteúdo de frequência de um sinal não se alinha perfeitamente com os limites das tramas. A sobreposição também aumenta a robustez contra ruído, permitindo melhorar a qualidade da extração de características e obter uma representação mais precisa do conteúdo espectral da vogal /a/. Por outro lado, a obtenção das tramas com um andamento menor que a duração da trama aumenta a resolução temporal, permitindo uma análise mais precisa das mudanças rápidas no sinal de fala ao longo do tempo.

As três primeiras e as três últimas tramas de cada sinal foram descartadas para mitigar potenciais instabilidades no início e fim dos sinais de fala. Dessa forma, os sinais de fala do *corpus USP* deram origem a 133 tramas utilizáveis e os da base de dados *sMEEI* a 73 tramas utilizáveis.

As energias por banda de cada trama são obtidas por filtragem, tendo sido esta efetuada no domínio da frequência, pois desta forma, tanto a definição do banco dos filtros, como o próprio processo de filtragem é mais simples do que efetuado no domínio do tempo. Assim, a cada uma das tramas foi aplicada a *FFT* com 4096 pontos, permitindo uma resolução de aproximadamente 5,4 Hz para os sinais da base de dados *USP*, que têm uma frequência de amostragem de 22050 Hz, e de aproximadamente 6,1 Hz para os sinais do *corpus sMEEI*, que têm uma frequência de amostragem de 25 kHz.

As energias por banda de frequência foram calculadas para os espectros de todas as tramas utilizadas. Para esse efeito, foi configurado um banco de filtros idêntico ao usado em [58], mas constituído por 20 filtros para a obter maior resolução nas bandas de energia. A gama de frequências coberta pelo banco de filtros estende-se dos 0 Hz (exclusive) até um pouco acima dos 4 kHz. Embora seja de esperar que os sinais de fala não contenham componente contínua, não se pretende que, caso exista, essa componente influencie o valor da energia da primeira banda de frequências. O limite superior do banco de filtros foi definido com base nos resultados obtidos em [44], que demonstraram ser possível discriminar oradores saudáveis e patológicos, bem como oradores com patologias laríngeas de origem fisiológica e neuromuscular, apenas com as componentes de baixas frequências dos sinais de fala. A utilização dessas componentes permite uma poupança em termos de complexidade computacional, pois apenas uma parte do espectro das tramas do sinal é processada.

Para definir as larguras de banda dos filtros, foram testadas três diferentes escalas, designadamente a escala linear, e as escalas perceptuais *Mel* [71] e *Bark*. A escala *Mel* apresentou de forma consistente os melhores resultados nos testes iniciais, levando, por essa razão, ao abandono dos filtros definidos nas outras escalas. Assim, os resultados apresentados neste trabalho referem-se apenas ao banco de filtros definido na escala *Mel*.

A escala *Mel* é uma escala de frequências baseada na percepção humana das frequências sonoras, daí ser chamada de escala perceptual. Alguns estudos psicofísicos mostraram que os

humanos percebem as mudanças de tom de maneira não linear. Por exemplo, a diferença perceptual entre 500 Hz e 1000 Hz é diferente da diferença entre 5000 Hz e 5500 Hz, mesmo que ambas as diferenças absolutas sejam de 500 Hz. A escala *Mel*, ao ter-se revelado a mais adequada para definição dos filtros e, consequentemente, das bandas de energia, indica que a informação perceptual é relevante para as discriminações efetuadas neste trabalho, reforçando o observado na Secção 2.3.1.3.

A escala *Mel* pode ser obtida através da Equação 4-2, a partir da escala linear de frequências, representada por *f* na expressão:

$$Mel = 2595 \times log_{10}\left(1 + \frac{f}{700}\right) \tag{4-2}$$

Após definição dos filtros na escala *Mel*, estes podem ser convertidos de volta à escala linear de frequências através da Equação 4-3:

$$f = 700 \times \left(10^{\frac{Mel}{2595}} - 1\right) \tag{4-3}$$

Foram definidos, na escala *Mel*, filtros triangulares com resposta em frequência de área unitária. Esses filtros são uniformemente espaçados e têm larguras de banda iguais, na escala *Mel*. Essa configuração garante que o banco de filtros resultante cubra toda a faixa de frequências considerada com resolução perceptual constante. As larguras de banda dos filtros sobrepõem-se às dos filtros adjacentes para garantir que todas as frequências tenham aproximadamente a mesma importância na contabilização das energias nessa banda.

O banco de filtros utilizado pode ser visualizado na Figura 4-3.

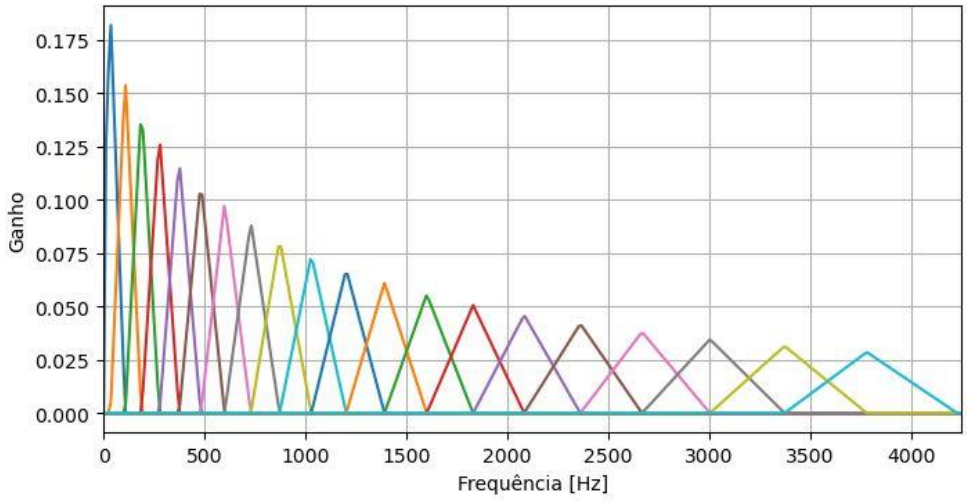


Figura 4-3 - Banco de filtros utilizado

Os limites inferiores e superiores de cada banda definida pelos filtros são apresentados de seguida na Tabela 4-1. Existem ligeiras diferenças entre os filtros das duas bases de dados devido às diferentes frequências de amostragem. Essas diferenças fazem com que os pontos da *FFT* estejam situados em frequências diferentes, resultando em pequenas diferenças nos limites das bandas.

Tabela 4-1 - Limites inferiores e superiores das bandas de energia definidas pelos filtros

| | **Corpus USP** | | **Corpus sMEEI** | |
|---|---|---|---|---|
| | **Limite Inferior [Hz]** | **Limite Superior [Hz]** | **Limite Inferior [Hz]** | **Limite Superior [Hz]** |
| **Banda 1** | 5 | 102 | 6 | 104 |
| **Banda 2** | 38 | 183 | 37 | 183 |
| **Banda 3** | 108 | 275 | 110 | 275 |
| **Banda 4** | 188 | 371 | 189 | 372 |
| **Banda 5** | 280 | 479 | 281 | 476 |
| **Banda 6** | 377 | 598 | 378 | 598 |
| **Banda 7** | 484 | 727 | 482 | 726 |
| **Banda 8** | 603 | 872 | 604 | 867 |
| **Banda 9** | 732 | 1028 | 732 | 1025 |
| **Banda 10** | 877 | 1200 | 873 | 1196 |
| **Banda 11** | 1034 | 1389 | 1032 | 1392 |
| **Banda 12** | 1206 | 1599 | 1202 | 1599 |
| **Banda 13** | 1394 | 1830 | 1398 | 1831 |
| **Banda 14** | 1604 | 2083 | 1605 | 2081 |
| **Banda 15** | 1836 | 2358 | 1837 | 2362 |
| **Banda 16** | 2089 | 2665 | 2087 | 2667 |
| **Banda 17** | 2363 | 3004 | 2368 | 3003 |
| **Banda 18** | 2670 | 3370 | 2673 | 3369 |
| **Banda 19** | 3009 | 3779 | 3009 | 3778 |
| **Banda 20** | 3375 | 4226 | 3375 | 4223 |

O banco de filtros definido é aplicado a todas as tramas para obter as energias dos sinais de fala em cada uma das bandas definidas pelos filtros. A filtragem origina um conjunto de 20 por N valores por cada sinal de fala, onde 20 é o número de filtros e bandas, e N é o número de tramas analisadas em cada sinal. Para os sinais do *corpus USP*, o resultado traduz-se num conjunto de 20 por 133 valores, enquanto para o *corpus sMEEI* resulta num conjunto de 20 por 73 valores.

Depois de obtidas as energias por banda para cada trama, foi calculada a média da distribuição de energia pelas bandas de cada sinal de fala. Assim, para cada sinal de fala, foi obtida a média das energias de todas as tramas em cada banda. Esse cálculo deu origem a um conjunto de 20 valores por sinal de fala. A variação da energia em cada banda ao longo de cada sinal de fala foi também avaliada através do cálculo do desvio padrão das energias, em cada banda, gerando outro conjunto de 20 valores por sinal de fala.

A Figura 4-4 apresenta uma representação dos passos efetuados.

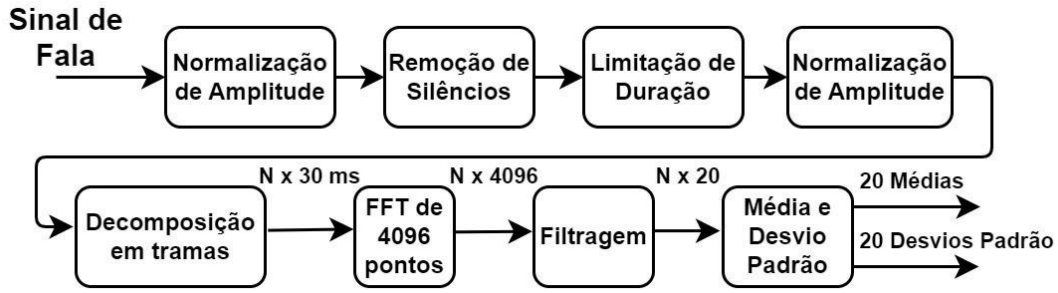


Figura 4-4 - Obtenção de médias e desvios padrão das energias, por banda

Após a obtenção das 20 médias e 20 desvios padrão, totalizando 40 valores por sinal de fala, esses valores são avaliados para determinar quais as bandas de energia que melhor discriminam os sinais entre saudáveis e patológicos, bem como, quando existente, entre patologias.

## 4.3 Proposta de dois novos parâmetros: bbLBST e bbLBSvT

A partir dos 20 valores referentes às médias de energia por banda, em cada sinal de fala, pode ser obtido um parâmetro equivalente ao parâmetro *LBST*, proposto nos estudos [47] e [57]. Em ambos os estudos, o parâmetro *LBST* é obtido a partir de uma divisão da faixa de frequências em duas, consistindo numa razão entre a diferença de energia em dois máximos locais no espectro, um em cada parcela da faixa de frequências, e a diferença em frequência entre esses dois máximos.

O mesmo princípio foi aplicado neste caso, com recurso às médias das energias por banda. Em [57], verificou-se que o limiar ótimo para a divisão entre as duas faixas de frequência seria 585,9 Hz, a que corresponde, neste trabalho, à parte final da banda 6. Assim, as bandas de energia foram divididas nas faixas das bandas 1 a 6 e das bandas 7 a 20, onde se procurarão os máximos locais.

As médias das energias são normalizadas para que o seu valor máximo seja unitário e procuram-se as bandas que contêm a maior energia em cada uma das faixas. O valor mais alto de energia na faixa das bandas 1 a 6 denomina-se, neste caso, de e1, a banda onde está situada denomina-se b1, enquanto o valor mais alto de energia na outra faixa denomina-se de e2 e a banda onde se situa denomina-se b2. O parâmetro resultante, o qual se denominou *band-based Low Band Spectral Tilt* (*bbLBST*), pode ser calculado através da Equação 4-4:

$$bbLBST = \frac{e2 - e1}{b2 - b1} \tag{4-4}$$

Um parâmetro equivalente pode ser obtido de forma similar ao anterior, através dos 20 desvios padrão das energias por banda, que será referente à variação de energia ao longo do sinal, em cada banda. Os 20 desvios padrão sofrem o mesmo reescalonamento que as médias, de modo a preservarem a sua relação com estas. As bandas são divididas em duas faixas, de 1 a 6 e de 7 a 20, procurando-se o máximo em cada faixa. Ao valor do desvio padrão mais elevado nas bandas 1 a 6 denomina-se ed1, a banda onde está situado denomina-se bd1, o máximo local na outra faixa denomina-se ed2 e a banda onde está situado denomina-se bd2. O parâmetro obtido a partir destes quatro valores foi designado como *band-based Low Band Spectral variation Tilt* (*bbLBSvT*) e a sua determinação pode ser efetuada através da Equação 4-5:

$$bbLBSvT = \frac{ed2 - ed1}{bd2 - bd1} \tag{4-5}$$

Estes dois parâmetros foram calculados para todos os sinais de fala das duas bases de dados, os seus valores analisados e usados para realizar as discriminações consideradas neste estudo.

## 4.4 Determinação das bandas mais relevantes

Para determinar as bandas de energia mais relevantes para os diferentes processos de discriminação considerados, cada banda foi analisada isoladamente. Desta forma, considera-se que cada sinal de fala é representado por 40 parâmetros unidimensionais diferentes, analisando-se e comparando os resultados obtidos para cada parâmetro. Os dados foram divididos em dois grupos, de treino e teste. Os dados de treino foram usados para definir o valor de um limiar que maximizasse a taxa de acertos na discriminação. Os dados de teste foram usados para avaliar a taxa de acertos obtida através desse limiar. A taxa de acertos obtida com os dados de teste foi a utilizada na análise.

Embora fosse possível utilizar todos os dados para definir um limiar e avaliar a taxa de acertos obtida, optou-se pela divisão de treino e teste para evitar que o resultado fosse excessivamente influenciado pelos dados específicos utilizados numa única iteração. Ao dividir os dados em treino e teste, conferiu-se uma maior fiabilidade aos resultados, pois estes foram obtidos em dados não vistos durante o treino, obtendo-se conclusões mais fiáveis e aplicáveis a novos dados.

Foi utilizada validação cruzada de 5 *folds*, ou seja, a divisão de dados em treino e teste teve sempre uma proporção de 80/20. A opção pela validação cruzada foi tomada com o objetivo de que fossem aproveitados todos os dados disponíveis para teste, conforme explicado na Secção 3.2.1. A divisão dos dados em *folds* foi sempre efetuada após um embaralhamento pseudo-aleatório (dependente de uma *seed*) dos dados, para garantir que não estivessem ordenados por classe.

O processo foi repetido 100 vezes, sempre com *seed* diferente, assegurando diferentes divisões dos dados e que o resultado não fosse dependente de uma única divisão dos dados. Ao repetir a

divisão, obtendo *folds* diferentes e calculando a taxa de acerto em cada iteração, obtém-se também uma distribuição dessa métrica, permitindo uma melhor compreensão dos resultados, conforme descrito na Secção 3.3.3.

Obtidas as médias das taxas de acerto para cada uma das dimensões, estas serão comparadas entre si para se verificar quais as mais elevadas, que indicarão quais as dimensões mais relevantes para as discriminações consideradas.

## 4.5 Resultados e discussão

Após calcular as médias e desvios padrão das energias em cada banda, para cada sinal de fala, estes valores foram agrupados de acordo com a classe dos sinais. De seguida, foram calculadas as médias dos 40 valores, 20 médias e 20 desvios padrão, para cada classe de modo a obter-se um valor médio para cada banda, por classe. Desta forma podem ser analisados os valores típicos dos 40 valores para cada classe, conforme se visualiza na Figura 4-5.

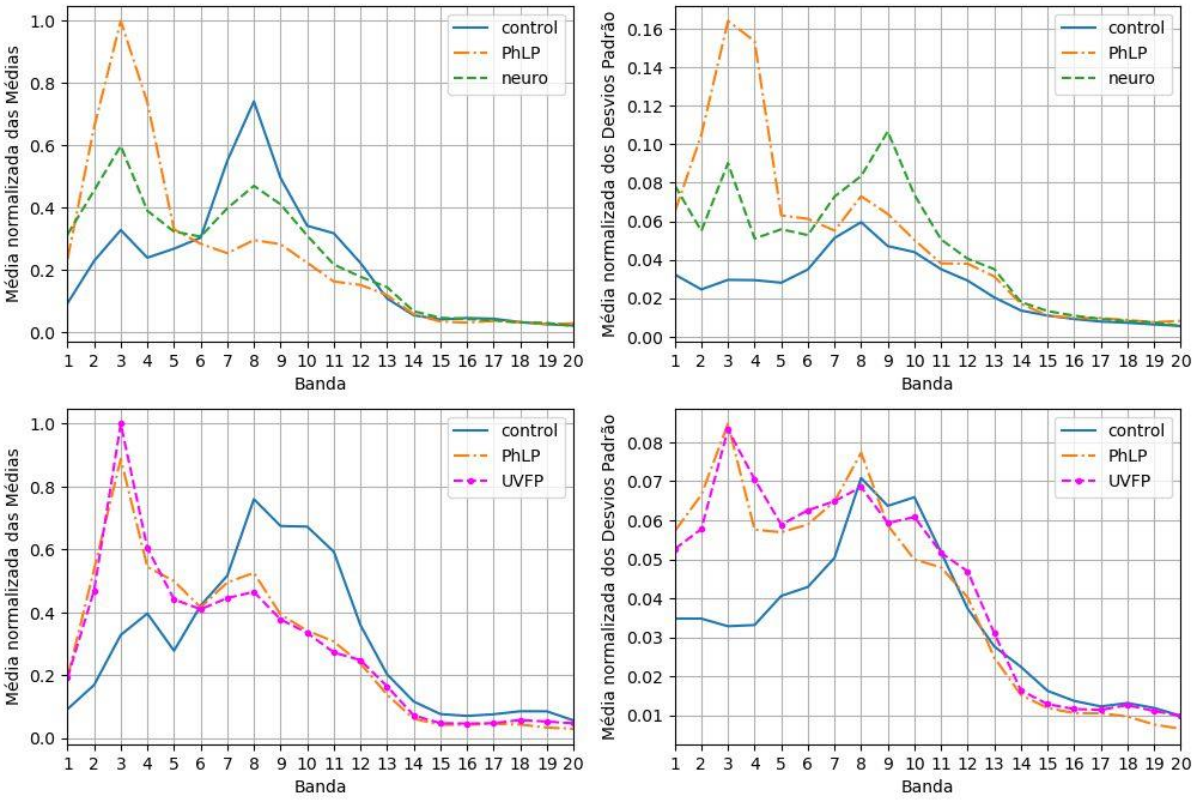


Figura 4-5 - Média normalizada das médias, à esquerda, e desvios padrão, à direita, das energias por banda, referentes à base de dados *USP*, em cima, e *sMEEI*, em baixo

Relativamente aos valores das bandas, as médias das médias das energias, nos gráficos à esquerda, foram normalizadas para que o valor máximo seja unitário. Esta normalização foi efetuada separadamente para cada base de dados. Quanto às médias dos desvios padrão das energias, nos

gráficos à direita, usam a escala normalizada das médias das médias, para manterem a relação direta com estas.

A partir dos gráficos da esquerda, observa-se que as distribuições de energia diferem entre as três classes no *corpus USP*. Os sinais de fala dos oradores saudáveis (*control*) têm maior energia na banda 8, que corresponde aproximadamente à faixa de frequências entre 603 e 872 Hz. Esta é a faixa de frequências onde geralmente se encontra o primeiro formante da vogal /a/ sustentada, conforme previamente observado na Figura 3-1.

Os oradores com patologias laríngeas fisiológicas (*PhLP*) e neurodegenerativas (*neuro*) também têm máximos locais na banda 8, mas a maior concentração de energia ocorre na banda 3, que cobre aproximadamente a faixa de frequências entre 110 e 275 Hz. Esta faixa inclui as primeiras harmónicas, podendo, no entanto, não abranger a frequência fundamental. Entre as duas classes de oradores patológicos, a concentração de energia nestas duas bandas é diferente. Na classe *PhLP*, o máximo na banda 3 é pronunciado enquanto na classe *neuro*, os máximos nas bandas 3 e 8 são semelhantes.

No *corpus sMEEI*, os sinais de oradores saudáveis apresentam uma distribuição de energia similar aos da base de dados *USP*, com um pico pronunciado na banda 8. No entanto, os sinais do *corpus sMEEI* também mostram uma grande concentração de energia nas bandas 9, 10 e 11, que abrangem aproximadamente a faixa de frequências entre 732 e 1390 Hz. Esta diferença pode ser atribuída à pronúncia da vogal, já que a base de dados *USP* contém sinais de oradores brasileiros, enquanto a base de dados *sMEEI* é constituída por oradores norte-americanos, sendo por essa razão de esperar que existam diferenças na pronúncia da vogal /a/ e, consequentemente, na localização dos seus formantes, neste caso, dos primeiro e segundo.

Para as duas classes patológicas do *corpus sMEEI*, as distribuições de energia mostram uma concordância quase perfeita. Embora as patologias sejam de naturezas diferentes, uma afetando a fisiologia das pregas vocais e a outra o controlo destas, essa diferença não se reflete na distribuição de energia.

Quanto às variações das energias por banda, observa-se que em ambas as bases de dados, os sinais de oradores saudáveis têm maior variação de energia na banda 8, onde também se situa a maior concentração de energia. Observa-se também que não existem grandes variações de energia nas bandas mais baixas, indiciando que a frequência fundamental mantém a sua energia durante a fonação, o que pode ser interpretado como um sinal de estabilidade vocal, esperada em oradores saudáveis.

Relativamente às duas classes patológicas do *corpus sMEEI*, tal como acontece com a distribuição de energia, também a variação de energia é muito similar nas duas classes. Ambas têm o seu máximo, com valores muito aproximados, na banda 3, tendo também valores muito aproximados em quase todas as restantes bandas. Observam-se algumas ligeiras diferenças na banda 8, bem como nas bandas 2 e 4. Estas duas últimas podem estar relacionadas com a frequência fundamental média em cada classe, pois conforme se mostrou na Tabela 3-2, as classes de oradores com patologias fisiológicas têm uma proporção significativamente maior de oradores de género feminino, o que eleva o *pitch* médio da classe em comparação com a de oradores com paralisia unilateral das pregas vocais.

A classe *PhLP* da base de dados *USP* também apresenta um máximo na banda 3, semelhante ao observado no *corpus sMEEI*, indiciando que os oradores com estas patologias podem ter alguma dificuldade em manter o tom e/ou a intensidade durante a fonação. No entanto, o máximo nesta classe, no *corpus USP* é muito mais pronunciado, indiciando que os oradores com estas patologias do *corpus USP* apresentam uma maior instabilidade vocal relativamente aos da base de dados *sMEEI*.

Quanto aos oradores com patologias de ordem neuromuscular, verifica-se que a variação de energia atinge o seu máximo na banda 9, que compreende aproximadamente a faixa de frequências entre 732 e 1028 Hz, onde ainda se espera que possa estar situado o primeiro formante. Esta maior variação de energia nesta banda de frequências poderá dever-se a uma dificuldade em produzir e manter o fonema, possivelmente devido a uma potencial degradação do controlo neuromuscular.

### 4.5.1 Análise dos parâmetros bbLBST e bbLBSvT

Obtidas as médias das energias por banda para cada classe, os valores do parâmetro *bbLBST* por classe podem ser determinados e visualizados na Figura 4-6:

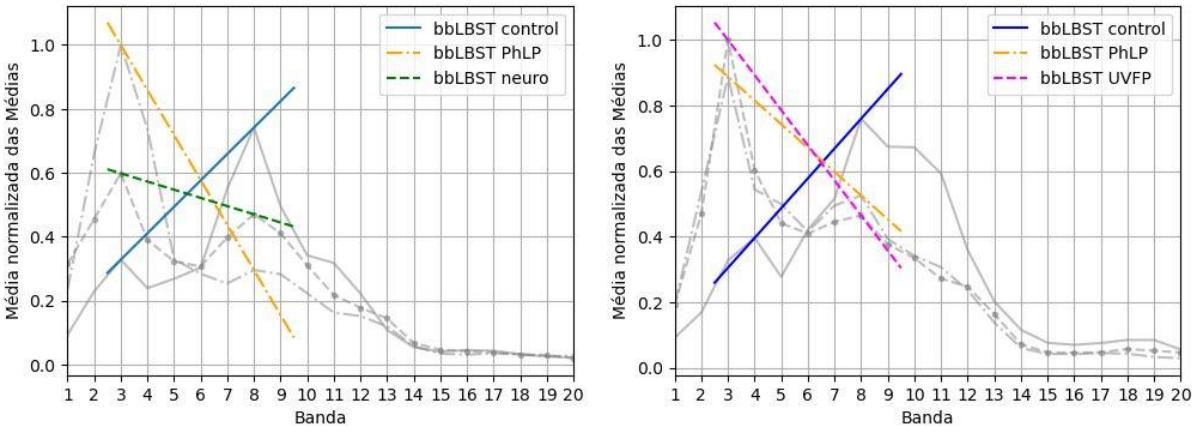


Figura 4-6 - Representação do parâmetro *bbLBST* por classe, para as bases de dados *USP* (esq.) e *sMEEI* (dir.)

Verifica-se que o parâmetro *bbLBST*, que representa um declive, é positivo para a classe dos oradores saudáveis em ambas as bases de dados, e é negativo para todas as outras classes. Observa-se que para a classe *PhLP*, o declive é negativo e bastante acentuado em ambas as bases de dados. O mesmo acontece com a classe *UVFP*, o que era esperado, dado que, conforme visto anteriormente, a distribuição das energias entre as classes *PhLP* e *UVFP* é idêntica. Para a classe dos oradores com patologias neurodegenerativas, o declive é também negativo, mas menos acentuado.

A constatação de que o declive representado pelo parâmetro *bbLBST* é positivo para os oradores saudáveis e negativo para os oradores patológicos confirma as conclusões dos estudos [47] e [57], bem como o observado nas Figuras 2-10 e 2-11.

De seguida, na Figura 4-7, são apresentados os valores médios por classe do parâmetro *bbLBSvT*, obtidos a partir dos valores médios por classe da variação de energia por banda.

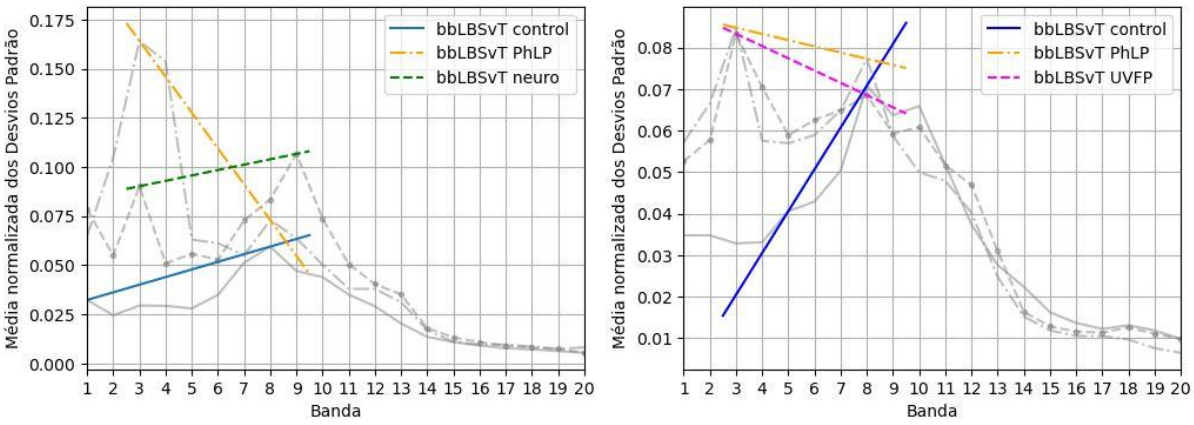


Figura 4-7 - Representação do parâmetro *bbLBSvT* por classe, para as bases de dados *USP* (esq.) e *sMEEI* (dir.)

Verifica-se na classe *PhLP* nas duas bases de dados, e na classe *UVFP*, presente no *corpus sMEEI*, que o valor típico é negativo, ou seja, a média da variação de energia por banda imita a média da distribuição de energia, reforçando o observado na Figura 4-5. Relativamente à classe dos oradores saudáveis, para as duas bases de dados, verifica-se que o declive correspondente ao parâmetro *bbLBSvT* é positivo, imitando também o comportamento do parâmetro *bbLBST*. Quanto à classe *neuro*, verifica-se que o declive é também positivo, o que indicia que, para esta classe, embora a banda com maior energia seja a correspondente à faixa de frequências das primeiras harmónicas, a maior variação de energia ocorre na banda referente à faixa de frequências esperada para o primeiro formante.

A distribuição dos oradores de acordo com seus valores dos parâmetros *bbLBST* e *bbLBSvT* pode ser visualizada de seguida, na Figura 4-8.

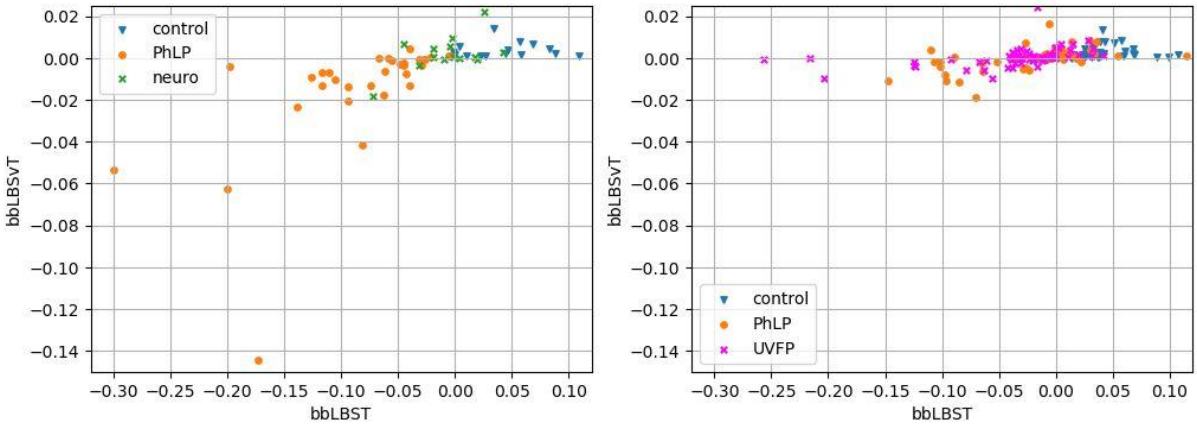


Figura 4-8 - Distribuição das amostras das bases de dados *USP* (esq.) e *sMEEI* (dir.) de acordo com os parâmetros *bbLBST* e *bbLBSvT*

Na distribuição dos oradores do *corpus USP*, as três classes mostram padrões distintos. A classe *PhLP* tem a maioria das suas amostras na região dos valores negativos (mais à esquerda e abaixo).

Os oradores saudáveis concentram-se na região dos valores positivos. Já os oradores com patologias neurodegenerativas situam-se numa zona intermediária entre as outras duas classes. O gráfico à esquerda demonstra que esses dois parâmetros permitem discriminar as três classes, pois é possível observar e diferenciar os três *clusters* correspondentes, embora nem todos os elementos de cada classe estejam agrupados no *cluster* da sua respetiva classe.

A distribuição das amostras do *corpus sMEEI* reforça a conclusão de que é possível diferenciar as classes *control* e *PhLP* através dos parâmetros *bbLBST* e *bbLBSvT*, pois também no gráfico à direita é possível observar os *clusters* referentes às duas classes. Quanto à classe *UVFP*, tal como observado anteriormente para os valores médios, a distribuição dos seus oradores é bastante similar à dos oradores com patologias laríngeas fisiológicas.

A zona onde se situam os oradores saudáveis pode ser visualizada com mais detalhe na Figura 4-9, que se apresenta de seguida.

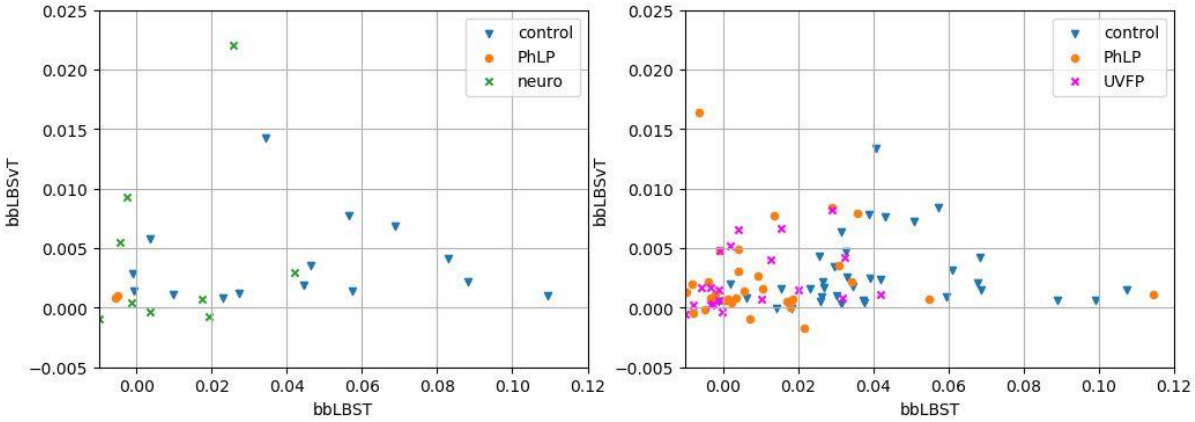


Figura 4-9 - Ampliação da distribuição das amostras das bases de dados *USP* (esq.) e *sMEEI* (dir.) de acordo com os parâmetros *bbLBST* e *bbLBSvT*

A Figura 4-9 mostra que os oradores saudáveis estão, quase na totalidade, contidos na zona dos valores positivos. Mesmo os poucos valores negativos estão muito próximos de zero. Estes resultados indicam que estes dois parâmetros têm o potencial de identificar oradores patológicos quando os valores destes parâmetros são significativamente negativos.

Para avaliar a capacidade dos parâmetros *bbLBST* e *bbLBSvT* de discriminar os sinais das diferentes classes, estes parâmetros foram usados para treinar e testar um classificador *SVM*. Foi utilizada validação cruzada com 5 *folds*, fazendo com que a proporção entre os grupos de treino e teste fosse de 80/20. O processo foi repetido 1000 vezes, com divisão em *folds* diferentes em todas as iterações.

Os resultados obtidos são apresentados de seguida na Tabela 4-2, onde as taxas de acerto referem-se ao valor médio obtido nas 1000 repetições com uma margem de confiança de 95%, expressas em percentagem. Da mesma forma são apresentados os valores da métrica *F1-Score*, usada para validar os resultados.

Tabela 4-2 - Resultados obtidos com os parâmetros *bbLBST* e *bbLBSvT*

| Discriminação | Taxa de acerto [%] | F1-Score [%] | Corpus |
|---|---|---|---|
| **Control vs. Neuro** | 71,3 ± 6,8 | 69,4 ± 8,8 | USP |
| **Control vs. UFVP** | 89,8 ± 1,7 | 91,5 ± 1,5 | sMEEI |
| **Control vs. PhLP** | 96,0 ± 4,2 | 97,2 ± 3,1 | USP |
| **Control vs. PhLP** | 83,5 ± 2,6 | 85,8 ± 2,3 | sMEEI |
| **Control vs. Patológicas** | 88,0 ± 3,0 | 92,2 ± 2,1 | USP |
| **Control vs. Patológicas** | 88,2 ± 1,7 | 92,1 ± 1,1 | sMEEI |
| **PhLP vs. Neuro** | 84,3 ± 2,3 | 89,2 ± 1,6 | USP |
| **PhLP vs. UVFP** | 60,4 ± 5,0 | 58,5 ± 7,2 | sMEEI |
| **Multiclasse** | 75,6 ± 3,1 | n/a | USP |
| **Multiclasse** | 60,0 ± 3,8 | n/a | sMEEI |

Os resultados indicam que estes parâmetros têm potencial para discriminar entre oradores saudáveis e com patologias fisiológicas, com taxas de acerto médias de 95,9% na base de dados *USP* e de 83,6% na base de dados *sMEEI*. Os parâmetros propostos aparentam também ser úteis para discriminar entre as classes *control* e *UVFP*, entre as classes *PhLP* e *neuro*, e entre as vozes saudáveis e patológicas, com taxas de acerto médias na casa do 80%.

Verificou-se anteriormente que os valores médios por classe da distribuição e variação de energia por banda eram bastante similares para as classes *PhLP* e *UVFP*. Essa similaridade reflete-se também no desempenho dos parâmetros *bbLBST* e *bbLBSvT*, que mostram ser pouco úteis na discriminação entre estas duas classes, com uma taxa de acerto de 60,0% e uma taxa de acerto de 60,5% na classificação multiclasse que envolve estas duas classes. Este resultado era esperado, pois estes dois parâmetros baseiam-se nas bandas de energia e, conforme verificado anteriormente, as bandas de energia são quase perfeitamente concordantes entre estas duas classes.

Os valores da métrica *F1-Score* são suficientemente elevados e aproximados aos valores das taxas de acerto, corroborando que estes últimos são suficientemente independentes do desbalanceamento entre algumas classes. Esta validação é especialmente importante na discriminação entre oradores saudáveis e patológicos, onde a diferença de amostras entre as duas classes é expressiva.

### 4.5.2 Bandas mais relevantes: Discriminação Saudáveis vs. Patológicas

Para a discriminação de vozes saudáveis e patológicas, existem quatro discriminações diferentes: entre saudáveis e com paralisia unilateral das pregas vocais, entre saudáveis e com patologias neurodegenerativas, entre saudáveis e com patologias laríngeas fisiológicas e entre saudáveis e patológicas, com as patologias presentes em cada uma das bases de dados.

Alguns resultados são agrupados, mas não devem ser comparados, pois as patologias são diferentes, sendo de esperar que tenham diferentes distribuições e variações de energia por banda.

Os resultados consistem em taxas de acerto obtidas através da aplicação de um classificador *SVM* no *fold* de teste. Este classificador foi treinado com os *folds* de treino, e utiliza um *kernel* linear, adequado para dados unidimensionais. Em duas situações não foi possível treinar o classificador em todas as dimensões devido ao desbalanceamento das classes utilizadas. Uma dessas situações ocorreu na discriminação entre as patologias laríngeas fisiológicas e neurodegenerativas, na base de dados *USP*, onde a proporção de amostras das duas classes é de 32/14. A outra situação ocorreu na discriminação entre vozes saudáveis e patológicas no *corpus sMEEI*, onde a proporção é de 115/36. Nesses casos, foram testados limiares que excluíram os três valores mais baixos e os três valores mais elevados dos dados de treino, para evitar que a maior taxa de acerto fosse obtida considerando todos os dados como pertencentes à classe maioritária. O limiar foi tomando os outros valores dos dados de treino e testado com estes dados. O limiar que obteve a maior taxa de acerto foi escolhido e usado nos dados de teste para obter a taxa de acertos utilizada nos resultados.

Os resultados expressam, para as bandas de energia e para as bandas de variação de energia, as duas maiores taxas de acerto obtidas em cada discriminação, expressas em média e margem de confiança de 95%. Não se apresentam as taxas de acerto para todas as bandas, pois seriam tabelas demasiado extensas e com muita informação de relevância reduzida. Apresentam-se as duas melhores, e não apenas a melhor, para mostrar se a diferença nas taxas de acerto entre as duas bandas mais relevantes é reduzida ou expressiva.

Os resultados para as discriminações entre oradores saudáveis e com patologias de ordem neuromuscular do *corpus USP*, e entre oradores saudáveis e patológicos com paralisia unilateral do *corpus sMEEI*, são apresentados na Tabela 4-3.

Tabela 4-3 - Bandas mais relevantes nas discriminações *control* vs. *neuro* e *control* vs. *UVFP*

| Discriminação | Banda | Faixa de frequências [Hz] | Taxa de acerto [%] |
|---|---|---|---|
| **Control vs. Neuro (USP)** | Energia - 1 | 5 - 102 | 100,0 ± 0,0 |
| | Energia - 4 | 188 - 371 | 85,3 ± 25,0 |
| | Variação de Energia - 3 | 108 - 275 | 88,8 ± 23,8 |
| | Variação de Energia - 2 | 38 - 183 | 86,4 ± 23,2 |
| **Control vs. UVFP (sMEEI)** | Energia - 3 | 110 - 275 | 87,4 ± 13,1 |
| | Energia - 1 | 6 - 104 | 85,0 ± 15,6 |
| | Variação de Energia - 3 | 110 - 275 | 82,1 ± 15,7 |
| | Variação de Energia - 4 | 189 - 372 | 79,4 ± 16,0 |

Verifica-se que as margens de confiança são bastantes expressivas (excetuando um caso, que será abordado de seguida), na ordem dos 15% para a base de dados *sMEEI* e dos 23% para o *corpus*

*USP*, relativamente à diferença nas taxas de acerto médias entre a banda mais relevante e a segunda mais relevante, que é de cerca de 2-3%, para cada uma das bandas. Verifica-se também que os valores para a margem de confiança são maiores para as taxas de acerto do *corpus USP*, o que implica uma menor confiança nos resultados obtidos para esta base de dados.

Analisando os resultados obtidos, verifica-se que as primeiras quatro bandas, que englobam a gama de frequências onde se situam a frequência fundamental e primeiras harmónicas, são as mais relevantes para as duas discriminações consideradas. Observa-se também que, para a discriminação entre as classes *control* e *neuro*, foi obtida uma taxa de acerto de 100%, em todas as classificações conforme revelado pela margem de confiança. Este resultado destaca-se dos restantes, pois é um resultado bastante promissor.

De seguida, na figura 4-10, apresentam-se os sinais de fala das classes *control* e *neuro*, à esquerda, e das classes *control* e *UVFP* do *corpus sMEEI*, à direita, representados pela banda mais relevante de energia e de variação de energia para cada uma das bases de dados.

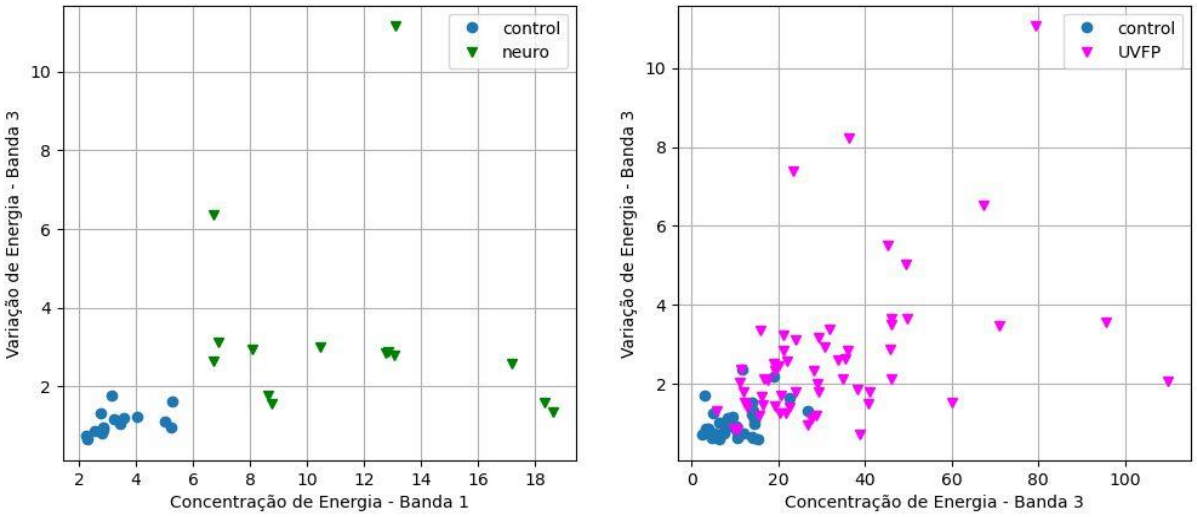


Figura 4-10 - Sinais de fala das classes *control* e *neuro* (esq.) e *control* e *UVFP* (dir.) representados nas bandas mais relevantes para a sua discriminação

Através da observação da Figura 4-10 verifica-se que as bandas mais relevantes permitem discriminar os oradores das classes consideradas neste caso. Essa discriminação é perfeita para os oradores saudáveis e com patologias neurodegenerativas do *corpus USP*, à esquerda, corroborando a taxa de acerto de 100% obtida na determinação das bandas mais relevantes. A discriminação é também visivelmente eficiente para os oradores saudáveis e com paralisia unilateral das pregas vocais da base de dados *sMEEI*, à direita, observando-se uma dispersão acentuada dos oradores patológicos.

De seguida, na Tabela 4-4, apresentam-se os resultados para as discriminações entre oradores saudáveis e com patologias laríngeas fisiológicas, de ambas as bases de dados. Estes resultados têm especial importância pois são os únicos que podem ser diretamente comparados, já que se referem à mesma patologia, ou seja, às mesmas classes.

Tabela 4-4 - Bandas mais relevantes na discriminação *control* vs. *PhLP* para ambas as bases de dados

| Discriminação / BD | Banda | Faixa de frequências [Hz] | Taxa de acerto [%] |
|---|---|---|---|
| **Control vs. PhLP (USP)** | Energia - 4 | 188 - 371 | 93,3 ± 14,3 |
| | Energia - 8 | 603 - 872 | 89,9 ± 16,4 |
| | Variação de Energia - 3 | 108 - 275 | 96,7 ± 9,7 |
| | Variação de Energia - 4 | 188 - 371 | 92,5 ± 16,8 |
| **Control vs. PhLP (sMEEI)** | Energia - 1 | 6 - 104 | 79,9 ± 16,2 |
| | Energia - 9 | 732- 1025 | 79,5 ± 16,9 |
| | Variação de Energia - 1 | 6 - 104 | 70,6 ± 17,6 |
| | Variação de Energia - 4 | 189 - 372 | 69,8 ± 18,1 |

As taxas de acerto médias obtidas permitem determinar que as bandas 1, 3 e 4 são as mais relevantes para a discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas. No entanto, verifica-se que as bandas 8 e 9, referentes à gama de frequências do primeiro formante da vogal /a/, também são relevantes para esta discriminação, aparecendo como as segundas bandas de concentração de energia com maior taxa de acerto média obtida. Tendo em conta as margens de confiança, que voltam a ser expressivas relativamente à diferença na taxa de acerto média entre as bandas mais relevantes e as segundas bandas mais relevantes, estes resultados indicam que as bandas referentes ao primeiro formante também são relevantes para esta discriminação.

De seguida, na Figura 4-11, apresentam-se os sinais de fala dessas duas classes, representados nas bandas mais relevantes para a base de dados *USP*.

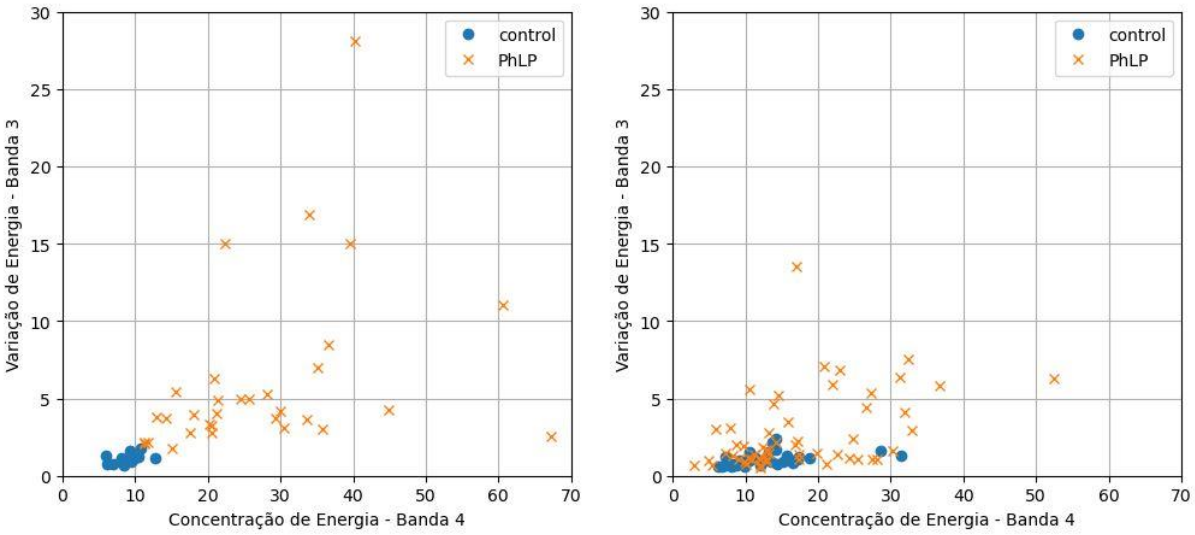


Figura 4-11 - Sinais de fala das classes *control* e *PhLP* do *corpus USP* (esq.) e *sMEEI* (dir.) representados nas bandas mais relevantes para a sua discriminação na base de dados *USP*

Para os sinais da base de dados *USP*, à esquerda, verifica-se que estas duas bandas permitem discriminar de forma muito eficiente os oradores destas duas classes. Na base de dados *sMEEI*, cujas amostras se podem visualizar à direita, essa discriminação não é tão eficiente, com muitas amostras das duas classes aglomeradas numa zona. No entanto, verifica-se uma dispersão de algumas amostras da classe *PhLP*, o que não ocorre com os oradores saudáveis, permitindo uma discriminação, embora menos eficiente, dos sinais de fala dessas duas classes.

Na Figura 4-12 apresentam-se os oradores destas duas classes representados pelos seus valores nas bandas mais relevantes para a classe *sMEEI*.

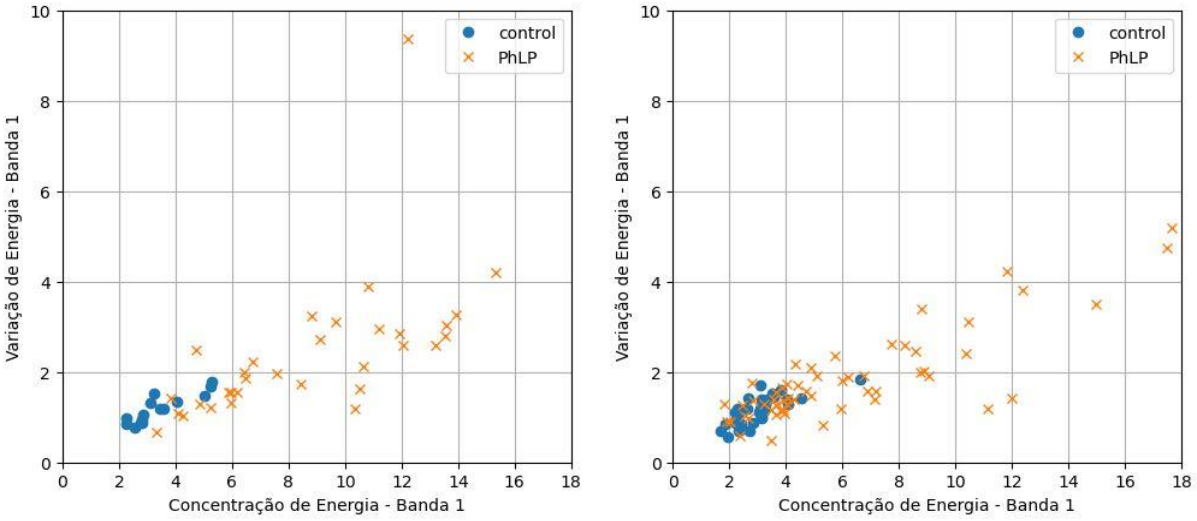


Figura 4-12 - Sinais de fala das classes *control* e *PhLP* do *corpus USP* (esq.) e *sMEEI* (dir.) representadas nas bandas mais relevantes para a sua discriminação na base de dados *sMEEI*

Observa-se que estas bandas permitem uma discriminação entre oradores saudáveis e da classe *PhLP*, nas duas bases de dados. Na zona onde se encontram os oradores saudáveis também se encontram alguns oradores patológicos. No entanto, estes têm tendência a dispersar-se, podendo muitos ser identificados a partir destas duas bandas. Portanto, estas bandas podem ser consideradas relevantes para a discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas.

Sendo as duas classes iguais nas duas bases de dados, algumas outras análises podem ser feitas. Verifica-se que as taxas de acerto médias obtidas no *corpus USP* são consistentemente mais elevadas que as obtidas na base de dados *sMEEI*. Isto poderá dever-se ao estado potencialmente mais avançado da patologia nos pacientes de uma das bases de dados, resultando numa degradação de fala diferente entre os oradores das duas bases de dados. Poderá também dever-se às condições de aquisição dos sinais, como o ambiente ou o equipamento, que podem influenciar o espectro dos sinais, ou poderá dever-se simplesmente à variabilidade dos dados. Não é possível tirar conclusões sobre estas diferenças, pois não se conhecem informações acerca das condições que poderiam causá-las. No entanto, constata-se que a discriminação dos sinais de fala pelos valores nas bandas mais relevantes aparenta, de forma consistente, obter melhores resultados na base de dados *USP*.

Para os resultados seguintes, as classes patológicas de cada base de dados foram agrupadas numa única classe chamada *Patológicas*. Para a classe *sMEEI* ocorreu uma das duas situações em que o *SVM* não conseguiu determinar um limiar válido de decisão para todas as dimensões, resultando em taxas de acerto de 76,16%. Este valor representa a percentagem de oradores patológicos, 115 de 151 amostras, na base de dados, indicando que o modelo classificou todos os oradores como patológicos. Por essa razão, foi necessária a utilização do método descrito no início desta secção.

Apesar de ambas as classes serem designadas como classe *Patológicas*, os resultados não são comparáveis, pois as classes englobam diferentes patologias em cada base de dados.

Os resultados para a discriminação entre as duas classes estão apresentados na Tabela 4-5.

Tabela 4-5 - Bandas mais relevantes na discriminação Saudáveis vs. Patológicas para ambas as bases de dados

| Discriminação / BD | Banda | Faixa de frequências [Hz] | Taxa de acerto [%] |
|---|---|---|---|
| **Control vs. Patológicas (USP)** | Energia - 4 | 188 - 371 | 91,2 ± 13,3 |
| | Energia - 1 | 5 - 102 | 88,7 ± 16,0 |
| | Variação de Energia - 3 | 108 - 275 | 91,8 ± 13,3 |
| | Variação de Energia - 2 | 38 - 183 | 89,7 ± 15,9 |
| **Control vs. Patológicas (sMEEI)** | Energia - 10 | 873- 1196 | 85,6 ± 11,3 |
| | Energia - 9 | 732- 1025 | 81,3 ± 11,6 |
| | Variação de Energia - 4 | 189 - 372 | 84,6 ± 10,6 |
| | Variação de Energia - 6 | 378 - 598 | 78,9 ± 10,3 |

Relativamente às bandas obtidas para a base de dados *USP*, verifica-se que a banda de energia mais relevante para a discriminação entre oradores saudáveis e patológicos é a banda 4, onde tipicamente se encontram as primeiras harmónicas do sinal de voz. Este resultado é consistente com os verificados anteriormente na discriminação entre as classes *control* e *PhLP*, onde também é identificada a banda 4 como a banda de energia mais relevante. Quanto à banda de variação de energia, o resultado é também consistente com as análises anteriores, sendo a banda 3 sempre considerada a banda mais relevante para a discriminação entre vozes saudáveis e patológicas.

No *corpus sMEEI*, a banda de variação de energia mais relevante para esta discriminação foi a banda 4, o que também está alinhado com os resultados anteriores, já que a banda 4 foi a segunda banda de variação de energia mais relevante nas discriminações anteriores. Quanto às bandas de energia consideradas mais relevantes, foram as bandas 9 e 10, que cobrem a zona de frequências do primeiro e potencialmente, também do segundo formante. Este resultado é interessante, visto que nas discriminações anteriores as bandas mais relevantes estavam associadas às primeiras harmónicas, mas demonstra que as bandas associadas ao primeiro formante e, eventualmente, também ao segundo formante, também são importantes para a discriminação entre vozes saudáveis e patológicas.

Os sinais de fala representados nas bandas mais relevantes apresentam-se na Figura 4-13.

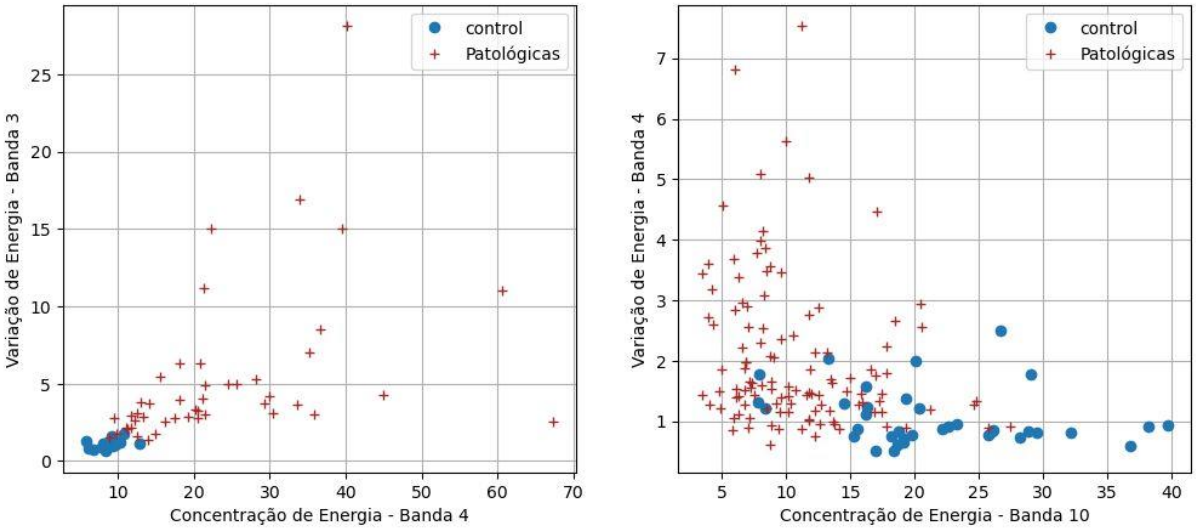


Figura 4-13 - Sinais de fala de oradores saudáveis e patológicos do *corpus USP* (esq.) e *sMEEI* (dir.) representados nas bandas mais relevantes para a sua discriminação em cada base de dados

A Figura 4-13 demonstra que as bandas de energia e de variação de energia consideradas mais relevantes permitem uma discriminação entre oradores saudáveis e patológicas, embora não seja perfeita. Esta discriminação é mais eficiente na base de dados *USP*, onde os oradores saudáveis estão aglomerados numa pequena área, enquanto os patológicos estão dispersos. No entanto, a discriminação também é perceptível no *corpus sMEEI*, onde os oradores das duas classes estão distribuídos de maneira diferente ao longo das duas dimensões.

### 4.5.3 Bandas mais relevantes: Discriminação entre patologias e multiclasse

Apresentam-se de seguida, na Tabela 4-6, os resultados para as discriminações entre patologias, nas duas bases de dados:

Tabela 4-6 - Bandas mais relevantes nas discriminações entre patologias

| Discriminação | Banda | Faixa de frequências [Hz] | Taxa de acerto [%] |
|---|---|---|---|
| **PhLP vs. Neuro (USP)** | Energia - 8 | 603 - 872 | 82,1 ± 19,9 |
| | Energia - 10 | 877 - 1200 | 77,9 ± 19,9 |
| | Variação de Energia - 3 | 108 - 275 | 71,4 ± 20,5 |
| | Variação de Energia - 7 | 484 - 727 | 71,3 ± 24,9 |
| **PhLP vs. UVFP (sMEEI)** | Energia - 8 | 604 - 867 | 57,2 ± 17,4 |
| | Energia - 19 | 3009 - 3778 | 56,9 ± 17,3 |
| | Variação de Energia - 8 | 604 - 867 | 54,5 ± 14,0 |
| | Variação de Energia - 5 | 281 - 476 | 54,0 ± 15,2 |

A discriminação entre patologias na base de dados *USP* foi a segunda situação onde o *SVM* não conseguiu determinar os limiares adequados para a classificação. Para muitas dimensões, a taxa de acertos obtida foi de 69,6%, que é a percentagem de oradores da classe *PhLP*, 32 de 46 amostras entre as classes consideradas. Assim, também para essa discriminação, foi usado o método descrito no início da Secção 4.5.2.

Verifica-se que a banda 3 de variação de energia continua a ser a mais relevante na base de dados *USP*, com uma diferença na taxa de acerto média de 0,1% em relação à segunda banda mais relevante. Esta diferença é pouco significativa, especialmente quando comparada com o valor de 20,5% para a margem de confiança de 95%. No entanto, o facto de que esta banda seja repetidamente a mais relevante é significativo e demonstra que é uma banda de variação de energia especialmente relevante para as discriminações analisadas.

A banda 8 foi a mais relevante tanto em energia nas duas bases de dados como em variação de energia no *corpus sMEEI* para a discriminação entre patologias, reforçando a importância das bandas que cobrem as faixas de frequência do primeiro formante nas discriminações, especialmente na discriminação entre patologias, ou seja, não envolvendo oradores saudáveis.

Um facto curioso prende-se com a segunda banda de energia mais relevante para a discriminação entre as classes *PhLP* e *UVFP*, que é a banda 19. Conforme mostrado na Figura 4-5, as bandas acima da Banda 13 aparentavam não contribuir significativamente para diferenciar oradores das diferentes classes, pois os seus valores eram muito semelhantes. Até este ponto, nenhuma banda acima da banda 11 se tinha revelado uma das mais relevantes. Adiante neste trabalho tentar-se-á explicar este resultado, destacando, por agora, que se trata de um resultado inesperado.

Na Figura 4-14, apresentada de seguida, são visualizados os oradores representados pelos valores nas bandas mais relevantes para cada uma das discriminações analisadas.

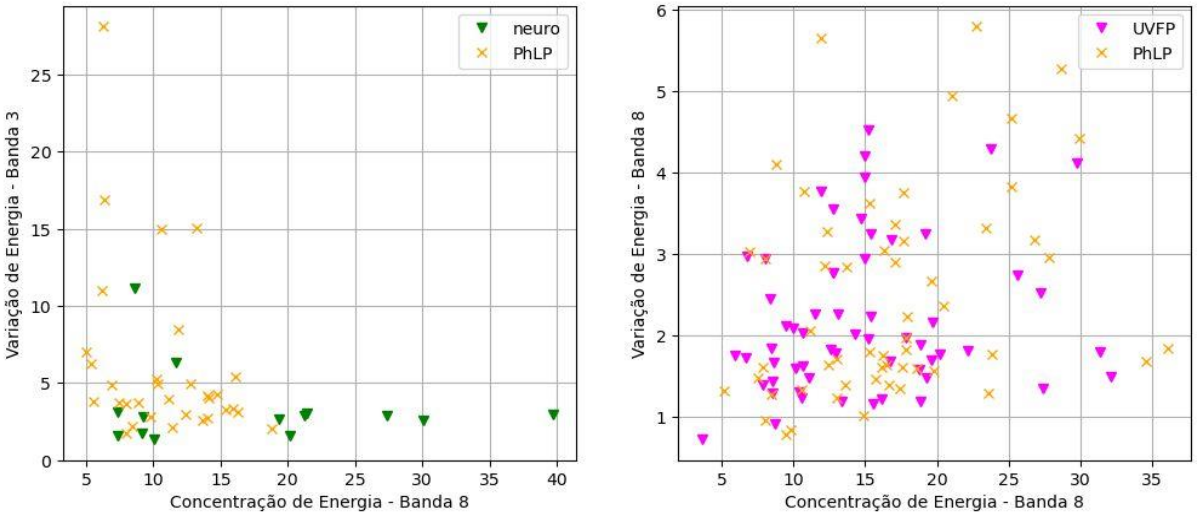


Figura 4-14 - Sinais de fala das classes *PhLP* e *neuro* (esq.) e *PhLP* e *UVFP* (dir.) representados nas bandas mais relevantes para a sua discriminação

Na base de dados *USP*, a discriminação entre as duas patologias pelas bandas mais relevantes, embora não seja a mais eficiente, é visível, o que não se verifica na base de dados *sMEEI*. Nesta, observa-se que as amostras das duas classes estão quase perfeitamente misturadas. Este resultado, em conjunto com a taxa de acerto média de apenas 57,2% para a banda mais relevante, indica que as bandas de energia não são muito adequadas para discriminar entre patologias laríngeas fisiológicas e paralisia unilateral das pregas vocais. Este resultado não é surpreendente, pois na Figura 4-5 já se tinha observado que os valores médios por classe para essas duas patologias eram bastante similares, sugerindo que a discriminação entre os oradores com estas patologias poderia ser menos conseguida.

De seguida, na Tabela 4-7, são apresentados os resultados obtidos para a discriminação multiclasse nas duas bases de dados. É importante salientar que esses resultados não são diretamente comparáveis, pois uma das patologias difere entre as duas bases de dados.

Tabela 4-7 - Bandas mais relevantes nas discriminações multiclasse

| **Discriminação / BD** | **Banda** | **Faixa de frequências [Hz]** | **Taxa de acerto [%]** |
|---|---|---|---|
| **Multiclasse (USP)** | Energia - 4 | 188 - 371 | 74,6 ± 21,7 |
| | Energia - 8 | 603 - 872 | 69,1 ± 14,9 |
| | Variação de Energia - 3 | 108 - 275 | 73,5 ± 11,8 |
| | Variação de Energia - 4 | 188 - 371 | 69,0 ± 17,4 |
| **Multiclasse (sMEEI)** | Energia - 1 | 6 - 104 | 52,0 ± 13,4 |
| | Energia - 2 | 37- 183 | 51,3 ± 11,9 |
| | Variação de Energia - 3 | 110 - 275 | 49,9 ± 12,7 |
| | Variação de Energia - 1 | 6 - 104 | 48,7 ± 12,9 |

Observa-se que, para ambas as bases de dados, a banda de variação de energia mais relevante é a banda 3. Esta banda já tinha mostrado ser a banda de variação de energia mais relevante noutras análises, especialmente no *corpus USP*, destacando a importância desta banda em específico para um potencial sistema automático de discriminação entre vozes saudáveis e patológicas, e entre patologias.

A banda 4, no *corpus USP*, e a banda 1, no *corpus sMEEI* são as bandas de energia mais relevantes para a discriminação multiclasse, sugerindo que as concentrações de energia nas baixas frequências, que incluem a frequência fundamental e as primeiras harmónicas do sinal de fala são as que, potencialmente, contêm mais informação útil para as discriminações consideradas.

A banda 8, como segunda banda de energia mais relevante, em conjunto com os resultados obtidos noutras discriminações onde esta banda foi considerada a mais relevante, indica que, embora não tão crucial como a energia nas baixas frequências, a energia do sinal de fala na gama de frequências do primeiro formante tem também uma relevância significativa nas discriminações.

Na Figura 4-15, podem ser visualizados os oradores representados pelos valores nas bandas mais relevantes na discriminação multiclasse para cada base de dados.

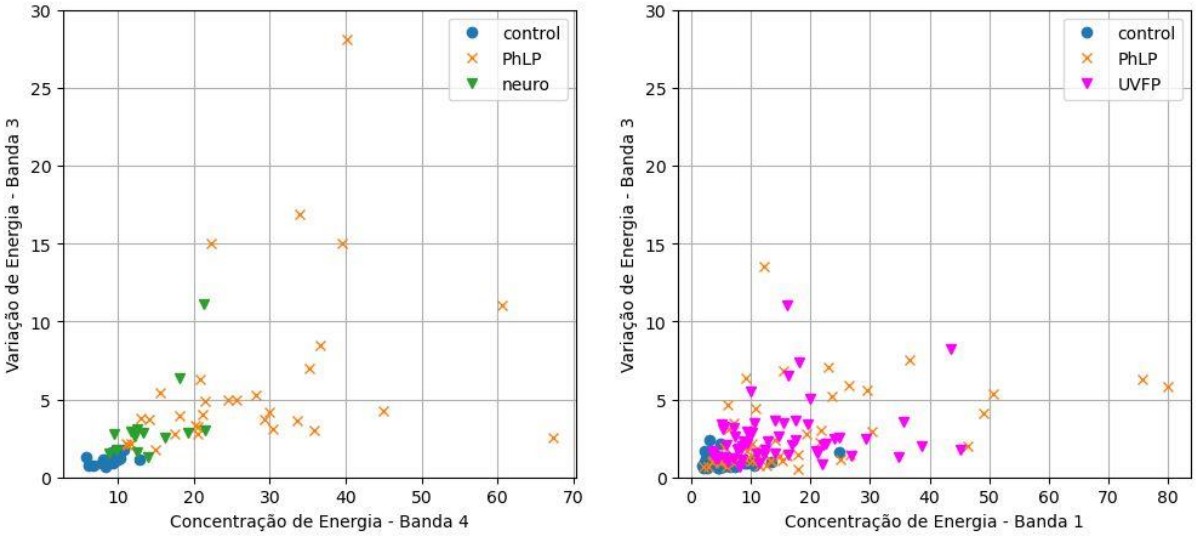


Figura 4-15 - Sinais de fala do *corpus USP* (esq.) e *sMEEI* (dir.) representados nas bandas mais relevantes para a discriminação multiclasse em cada base de dados

Verifica-se que, em ambas as bases de dados, os oradores saudáveis estão concentrados numa área específica, enquanto os oradores patológicos estão mais dispersos. A Figura 4-16 apresenta ampliações das zonas onde estão localizados os oradores saudáveis, para uma melhor análise.

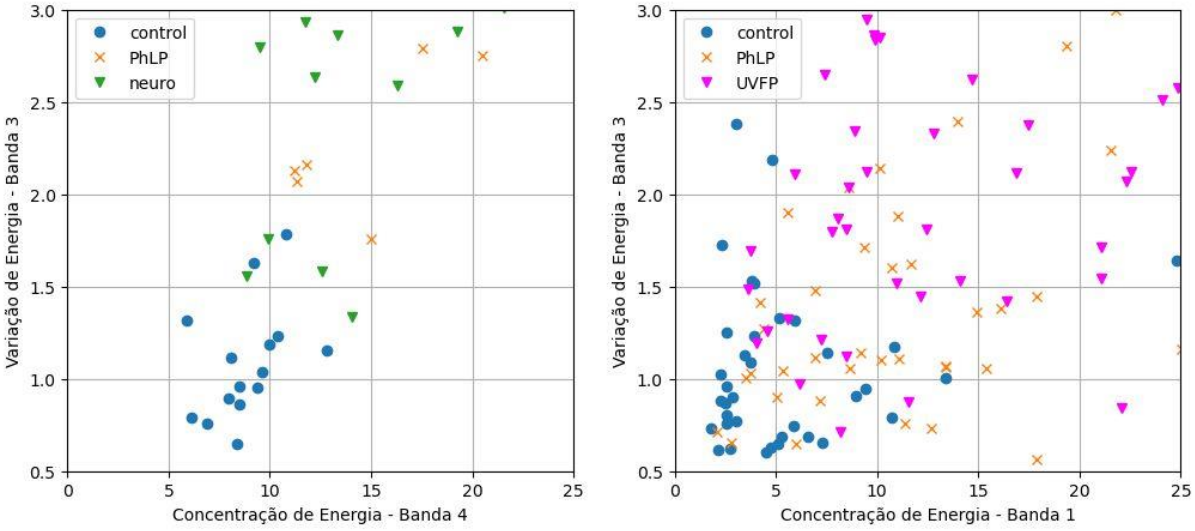


Figura 4-16 - Sinais de fala do *corpus USP* (esq.) e *sMEEI* (dir.) representados nas bandas mais relevantes para a discriminação multiclasse em cada base de dados - Ampliação da zona dos oradores saudáveis

Para o *corpus USP*, à esquerda, observa-se que os oradores saudáveis estão contidos numa pequena zona, como já tinha sido referido, enquanto os oradores da classe *PhLP* estão bastante dispersos. Os oradores com patologias de ordem neuromuscular situam-se numa posição intermédia,

permitindo a discriminação entre as três classes, embora não de forma perfeita. No *corpus sMEEI*, a discriminação é mais difícil, seguindo duas tendências vistas anteriormente, nomeadamente, que os resultados obtidos no *corpus USP* são, tipicamente, melhores que os obtidos no *sMEEI*, e que a discriminação entre as classes *PhLP* e *UVFP* através das bandas de energia não apresenta bons resultados. Portanto, a discriminação entre as três classes no *corpus sMEEI* é menos bem-sucedida.

### 4.5.4 Outros resultados: Edema de Reinke e nódulos vocais

Embora não tenha sido um objetivo inicial deste trabalho, foi realizada uma análise aos sinais de fala de oradores com patologias laríngeas fisiológicas, através da discriminação entre as duas classes que os constituem. Assim, foram obtidos os valores médios por classe para a energia e para a variação de energia por banda, como ilustrado na Figura 4-17. Tal como na Figura 4-5, os gráficos superiores correspondem ao *corpus USP*, os inferiores ao *sMEEI*, os da esquerda à concentração de energia por banda, os da direita referem-se à variação de energia por banda, e os valores estão normalizados para que a energia máxima em cada base de dados seja unitária.

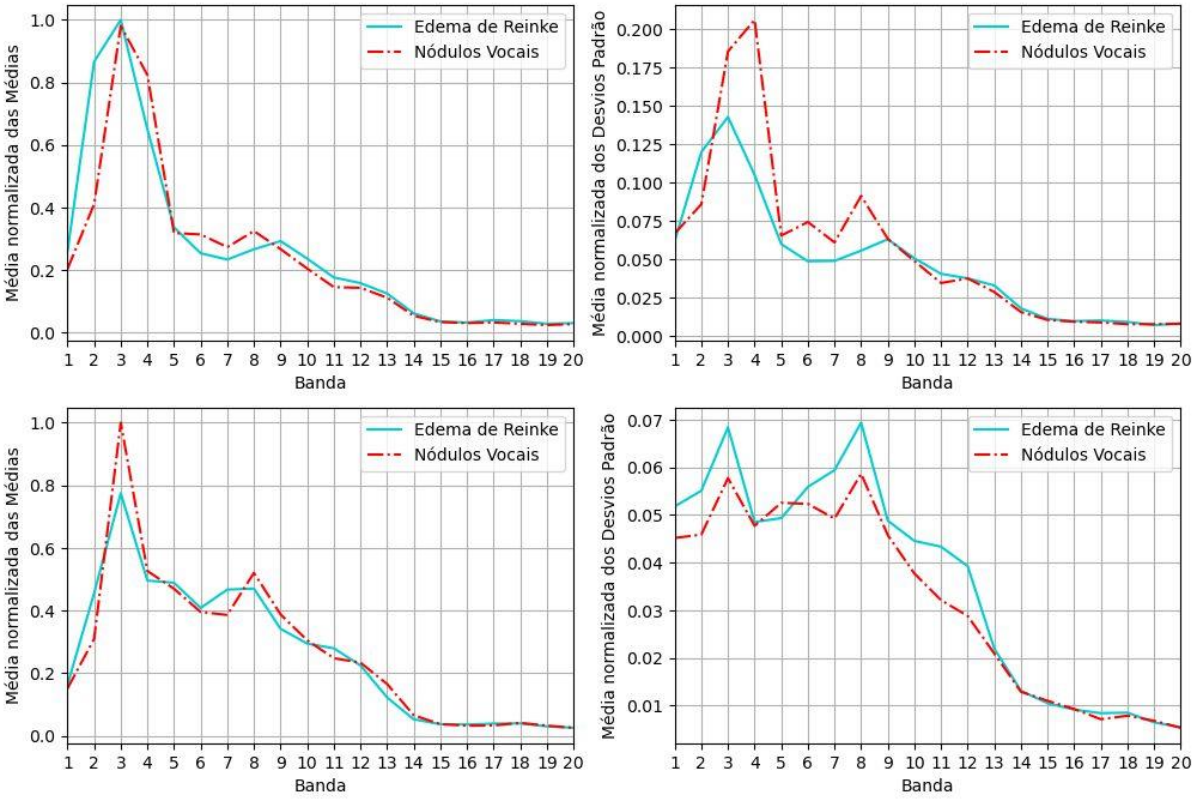


Figura 4-17 - Média normalizada das médias (esq.) e desvios padrão (dir.) das energias por banda para oradores patológicos com edema de Reinke e nódulos vocais do corpus *USP* (em cima) e *sMEE*I (em baixo)

Analisando os valores médios por classe, verifica-se que a distribuição de energia por bandas é bastante semelhante entre as duas classes, com algumas diferenças nas bandas 2 e 4 na base de

dados *USP*. Essas diferenças podem estar mais relacionadas com a diferença da frequência fundamental média por classe do que propriamente com o efeito específico da patologia no espectro dos sinais de fala. No *corpus sMEEI*, também existe uma diferença na energia da banda 3, que pode ser relevante para uma possível discriminação entre os oradores das duas classes.

Relativamente às bandas de variação de energia, nos gráficos à direita, as diferenças observadas são contraditórias. Na base de dados *USP* verifica-se uma maior variação na classe dos pacientes com nódulos vocais, enquanto no *corpus sMEEI* a variação é maior para os oradores com edema de Reinke.

Na Figura 4-18, podem ser visualizados os oradores representados pelos valores dos parâmetros *bbLBST* e *bbLBSvT*.

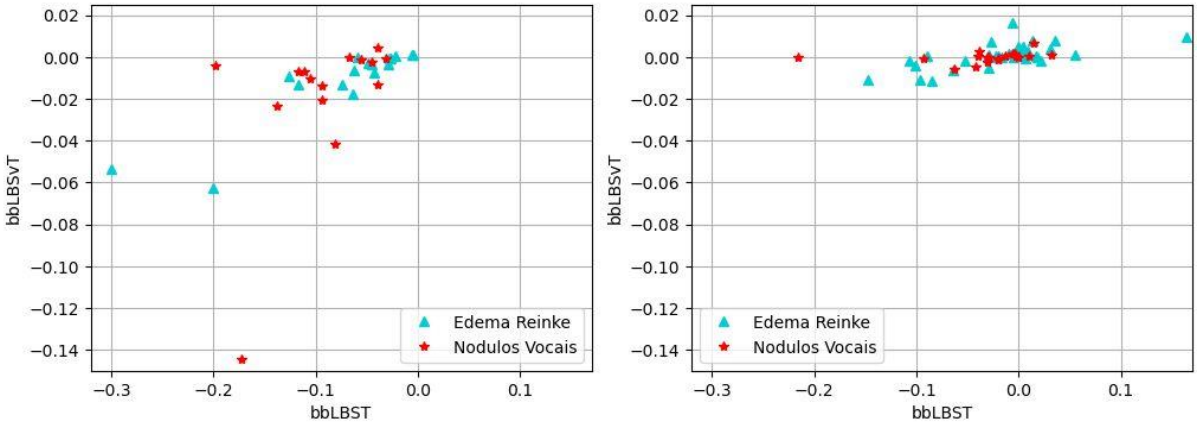


Figura 4-18 - Distribuição dos oradores patológicos das classes *edema* e *nodulo* das bases de dados *USP* (esq.) e *sMEEI* (dir.) de acordo com os parâmetros *bbLBST* e *bbLBSvT*

Observa-se que os oradores estão misturados, não sendo possível discriminar os oradores das duas classes com base nestes parâmetros. Para confirmar esta conclusão, foi treinado e testado um classificador *SVM* com estes parâmetros, tendo sido usada validação cruzada e repetindo-se o processo 100 vezes, tal como efetuado anteriormente. Os resultados apresentam-se na Tabela 4-8.

Tabela 4-8 - Taxas de acerto obtidas com os parâmetros *bbLBST* e *bbLBSvT* para oradores patológicos com edema de Reinke e nódulos vocais

| **Discriminação** | **Taxa de acerto [%]** | **F1-Score [%]** | **Corpus** |
|---|---|---|---|
| **Edema vs. Nodulo** | 57,9 ± 9,8 | 49,7 ± 15,5 | **USP** |
| **Edema vs. Nodulo** | 66,1 ± 1,1 (*) | 0,0 ± 0,4 | **sMEEI** |

(*) resultado inválido

A taxa de acerto média obtida no *corpus USP*, com base nos parâmetros *bbLBST* e *bbLBSvT*, foi de 57,9%, confirmando que estes parâmetros não são os mais adequados para essa discriminação. Na base de dados *sMEEI* o classificador não conseguiu definir um limiar válido para a classificação, como indica a métrica *F1-Score*. A taxa de acerto média obtida foi de 66,1% refletindo a proporção de oradores com edema de Reinke, 37 no total de 56 amostras. Ou seja, o *SVM* classificou todos os oradores como pertencentes à classe maioritária.

De seguida, foram determinadas as bandas mais relevantes para a discriminação entre estas duas classes, cujos resultados se apresentam na Tabela 4-9.

Tabela 4-9 - Bandas mais relevantes na discriminação *edema* vs. *nodulo* para ambas as bases de dados

| Discriminação / BD | Banda | Faixa de frequências [Hz] | Taxa de acerto [%] |
|---|---|---|---|
| **Edema Reinke vs. Nódulos Vocais (USP)** | Energia - 8 | 603 - 872 | 64,8 ± 32,8 |
| | Energia - 17 | 2363 - 3004 | 60,8 ± 33,7 |
| | Variação de Energia - 6 | 377 - 598 | 58,4 ± 28,2 |
| | Variação de Energia - 14 | 1604 - 2083 | 57,2 ± 28,9 |
| **Edema Reinke vs. Nódulos Vocais (sMEEI)** | Energia - 5 | 281 - 476 | 66,1 ± 6,9 (*) |
| | Energia - 12 | 1202- 1599 | 66,1 ± 6,9 (*) |
| | Variação de Energia - 1 | 6 - 104 | 66,1 ± 6,9 (*) |
| | Variação de Energia - 16 | 2087 - 2667 | 66,1 ± 6,9 (*) |

(*) resultados inválidos

Os resultados referentes à base de dados *sMEEI* são apresentados por uma questão de uniformização, mas não são válidos. Nem o *SVM*, nem o método implementado anteriormente para encontrar um limiar, conseguiram obter resultados válidos. As taxas de acerto médias foram obtidas considerando todos os oradores como pertencendo à classe *edema*. No *corpus USP*, as taxas de acerto obtidas indicam que, tal como para a análise das classes *PhLP* e *UVFP*, a discriminação entre as patologias edema de Reinke e nódulos vocais, existindo, não é eficiente.

Além disso, tal como na discriminação entre as patologias *PhLP* e *UVFP*, voltam a aparecer bandas acima da banda 11 como mais relevantes. Estes resultados sugerem que, embora as bandas acima da banda 11 contenham pouca informação para a discriminação entre patologias, essas bandas contêm alguma utilidade. Quando as bandas associadas à frequência fundamental e às primeiras harmónicas, bem como à gama de frequências dos dois primeiros formantes, apresentam desempenhos mais baixos, as bandas de frequências mais elevadas acabam por revelar-se úteis.

As bandas mais elevadas apresentam uma grande atenuação, pois o espectro do sinal de fala tem comportamento idêntico à resposta em frequência de um filtro passa-baixo. Portanto, surge a dúvida se a energia nestas bandas mais altas realmente representa a energia do sinal de fala ou se, devido à atenuação, outros efeitos, como o ruído ambiental, se tornam significativos, influenciando os

resultados obtidos. Não existe informação acerca do ambiente em que os sinais de fala foram recolhidos, e se, neste caso, os sinais de fala de pacientes com edema de Reinke foram recolhidos num ambiente diferente do ambiente onde foram recolhidos os de pacientes com nódulos vocais, essa diferença de ambiente e de ruído de fundo, entre outras variáveis, pode ter influenciado este resultado.

O aparecimento das bandas 14 e 17 como segundas bandas mais relevantes poderá justificar uma investigação futura dessas bandas mais altas, onde estarão situados os terceiro e quarto formantes da vogal /a/, que podem ter uma especial importância. Potencialmente, a atenuação dessas bandas pode ser mitigada através da utilização, por exemplo, de um filtro de pré-enfase.

A Figura 4-19, apresenta os oradores da base de dados *USP*, representados pelas bandas mais relevantes para a discriminação das patologias. Os sinais do *corpus sMEEI* não são apresentados, pois a determinação das bandas mais relevantes revelou-se inválida.

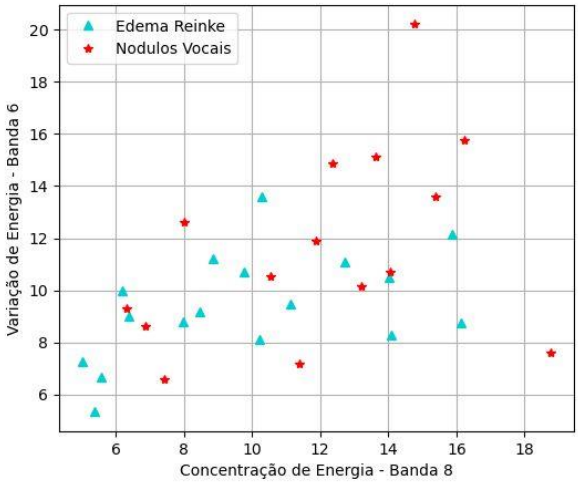


Figura 4-19 - Sinais de fala das classes *edema* e *nodulo* do *corpus USP* representados nas bandas mais relevantes para a sua discriminação

A visualização da representação dos sinais de fala nas bandas mais relevantes permite concluir que a discriminação entre estas duas classes, através da energia e da variação de energia por banda, não produz resultados satisfatórios. Esta conclusão confirma como adequada a opção de agrupar as classes referentes aos oradores com edema de Reinke e com nódulos vocais na classe *PhLP*. A junção foi feita devido à semelhança na natureza das duas patologias, e agora verifica-se que a combinação das classes eliminou uma abordagem que, provavelmente, não teria sucesso na discriminação entre elas através da análise das bandas de energia.

### 4.5.5 Resumo das bandas mais relevantes

Não considerando a última discriminação analisada, entre oradores com edema de Reinke e nódulos vocais, foram efetuadas, no total, dez discriminações diferentes, tendo-se obtido uma diversidade considerável de bandas consideradas mais relevantes para as diferentes discriminações.

Na Tabela 4-10, apresenta-se uma compilação das bandas que se destacaram nessas discriminações.

Tabela 4-10 - Bandas mais relevantes nas diferentes discriminações (compilação)

| **Discriminação (Corpus)** | **Bandas mais relevantes de energia** | | **Bandas mais relevantes de variação de energia** | |
|---|---|---|---|---|
| | **Primeira** | **Segunda** | **Primeira** | **Segunda** |
| **Control vs. Neuro (USP)** | Banda 1 | Banda 4 | Banda 3 | Banda 2 |
| **Control vs. UFVP (sMEEI)** | Banda 3 | Banda 1 | Banda 3 | Banda 4 |
| **Control vs. PhLP (USP)** | Banda 4 | Banda 8 | Banda 3 | Banda 4 |
| **Control vs. PhLP (sMEEI)** | Banda 1 | Banda 9 | Banda 1 | Banda 4 |
| **Control vs. Patológicas (USP)** | Banda 4 | Banda 1 | Banda 3 | Banda 2 |
| **Control vs. Patológicas (sMEEI)** | Banda 10 | Banda 9 | Banda 4 | Banda 6 |
| **PhLP vs. Neuro (USP)** | Banda 8 | Banda 10 | Banda 3 | Banda 7 |
| **PhLP vs. UVFP (sMEEI)** | Banda 8 | Banda 19 | Banda 8 | Banda 5 |
| **Multiclasse (USP)** | Banda 4 | Banda 8 | Banda 3 | Banda 4 |
| **Multiclasse (sMEEI)** | Banda 1 | Banda 2 | Banda 3 | Banda 1 |

Observa-se que oito diferentes bandas de energia e oito diferentes bandas de variação de energia foram consideradas como uma das mais relevantes em pelo menos uma das discriminações analisadas. Todas as bandas de 1 a 10 foram, em alguma discriminação, consideradas como uma das duas mais relevantes, demonstrando a sua importância nas análises consideradas. Apenas por uma vez, uma banda entre 11 e 20 foi identificada como uma das duas bandas mais relevantes, um resultado considerado inesperado cujas possíveis justificações foram discutidas na Secção 4.5.4.

As bandas de energia mais frequentemente identificadas como mais relevantes foram as bandas 1, identificada cinco vezes, e as bandas 4 e 8, identificadas quatro vezes cada. As bandas de variação de energia mais frequentemente identificadas como mais relevantes foram a banda 3, identificada sete vezes, e a banda 4, identificada quatro vezes. Este resultado, somado ao facto de que as bandas entre 1 e 4, relacionadas com a frequência fundamental e primeiras harmónicas, foram identificadas 27 vezes como uma das duas mais relevantes, demonstra que estas frequências contêm o maior potencial para as discriminações consideradas neste estudo.

As bandas 7, 8 e 9, associadas ao primeiro formante da vogal /a/, foram identificadas oito vezes como uma das duas bandas mais relevantes, demonstrando terem também estas frequências um potencial significativo nas discriminações entre oradores saudáveis e patológicos, e entre patologias.

# 5

# COMBINAÇÃO COM PARÂMETROS ACÚSTICOS

Neste capítulo, apresenta-se o trabalho realizado relativo à obtenção de parâmetros espectrais a partir das bandas de energia e de variação de energia mais relevantes para as discriminações de sinais de fala entre saudáveis e patológicos, assim como para a identificação de qual a patologia presente nos sinais de fala patológicos. Também será estudada a combinação desses parâmetros com parâmetros acústicos para avaliar se essa combinação proporciona melhorias na eficácia das discriminações consideradas. Serão também apresentados e discutidos os resultados obtidos com esse processo.

## 5.1 Parâmetros espectrais

As bandas de energia e de variação de energia totalizam 40 valores, ou dimensões. Este número é demasiado elevado, comparativamente ao número de amostras disponíveis, para que as bandas sejam utilizadas como parâmetros espectrais para um sistema automático de discriminação. Conforme descrito na Secção 3.2.3, um maior número de dimensões não implica necessariamente um melhor desempenho do classificador, devido ao efeito conhecido como *Curse of Dimensionality*. Embora não haja uma regra específica para esta questão, algumas fontes informais recomendam um rácio mínimo de cinco a dez amostras por dimensão. Comparando o número de bandas, 40, com o número de amostras do *corpus USP*, que é, recorde-se, 61, verifica-se que a dimensionalidade é excessiva. Especialmente porque algumas discriminações não usam todas as amostras pois não consideram uma das classes, com o caso limite da discriminação entre oradores saudáveis e com patologias neurodegenerativas onde apenas 29 amostras são consideradas. Portanto, é necessário reduzir a dimensionalidade dos dados.

Na ideia inicial, as bandas mais relevantes para as diferentes discriminações seriam utilizadas como parâmetros espectrais. No entanto, foram identificadas 16 bandas diferentes como uma das duas bandas mais relevantes em pelo menos uma discriminação. Para além disso, as margens de confiança de 95% obtidas nessas discriminações são demasiado amplas para garantir que essas 16 bandas são realmente as mais relevantes. Portanto, essa possibilidade foi descartada, pois ainda levaria a uma

dimensionalidade elevada, sem uma grande confiança de que não seriam descartadas algumas bandas importantes.

Outra possibilidade é a utilização dos parâmetros *bbLBST* e *bbLBSvT*, propostos neste trabalho, que demonstraram ser úteis nas discriminações testadas. Como estes parâmetros são unidimensionais, o problema da dimensionalidade não se aplica com a sua utilização. No entanto, estes parâmetros apenas refletem o comportamento das duas bandas com os valores de energia ou de variação de energia mais elevados, sem considerar o comportamento das restantes bandas. Portanto, a utilização dos parâmetros *bbLBST* e *bbLBSvT* implica que a informação presente em outras bandas não seja considerada.

Assim, optou-se pela definição de novos parâmetros espectrais a partir das bandas de energia e de variação de energia. Estes parâmetros devem ter uma dimensionalidade reduzida e devem refletir a informação das bandas mais relevantes para as discriminações a realizar.

A obtenção desses parâmetros começou pela seleção das bandas que contribuem para a sua definição. Na Figura 4-5 observou-se que as bandas acima da banda 14, ou em alguns casos, mesmo abaixo dessa, como sucedeu com as bandas de energia do *corpus USP*, apresentam pouca ou nenhuma diferença nos valores médios por classe, sugerindo que podem conter pouca informação relevante para a discriminação entre as classes. Os resultados das Secções 4.5.2 e 4.5.3 corroboram essa conclusão, pois apenas numa discriminação uma banda acima da banda 10 foi considerada uma das mais relevantes, tendo sido apresentadas algumas possíveis justificações para esse resultado na Secção 4.5.4.

Assim, foram mantidas as primeiras 12 bandas de energia bem como as primeiras 12 bandas de variação de energia, descartando-se as bandas de 13 a 20. A banda 12 foi definida como limite superior, pois cobre a faixa de frequências esperadas para o segundo formante da vogal /a/. Na Secção 4.5.5 observou-se que as frequências relacionadas com as frequências do primeiro formante revelaram-se importantes. Portanto, supõe-se que as frequências relacionadas com o segundo formante também tenham alguma relevância.

As 12 bandas mantidas, como demonstrado no capítulo anterior, não têm todas a mesma relevância nas discriminações. As bandas associadas à frequência fundamental e às primeiras harmónicas foram identificadas como uma das duas bandas mais relevantes 29 vezes, enquanto as bandas relacionadas com o primeiro formante foram-no 10 vezes. Portanto, as bandas foram divididas em dois grupos, sendo o primeiro grupo constituído pelas bandas de 1 a 6 e o segundo pelas bandas de 7 a 12. Assim, são considerados quatro grupos de bandas, as primeiras 6 bandas de energia, as primeiras 6 bandas de variação de energia, as bandas de 7 a 12 de energia e as bandas de 7 a 12 de variação de energia.

As bandas mantidas totalizam 24 dimensões, um número ainda bastante elevado relativamente ao número de amostras. Para reduzir a dimensionalidade, aplicou-se o *PCA* separadamente a cada um dos 4 grupos e as componentes obtidas são os parâmetros espectrais a ser testados. Utilizam-se as duas primeiras componentes principais relativas a cada um dos grupos das bandas de 1 a 6, pois

estas revelaram-se mais relevantes, e utiliza-se a primeira componente principal relativa a cada um dos grupos das bandas de 7 a 12, totalizando 6 dimensões.

Apresenta-se, na Figura 5-1, uma representação do processo de obtenção dos parâmetros espectrais a partir das bandas de concentração e de variação de energia.

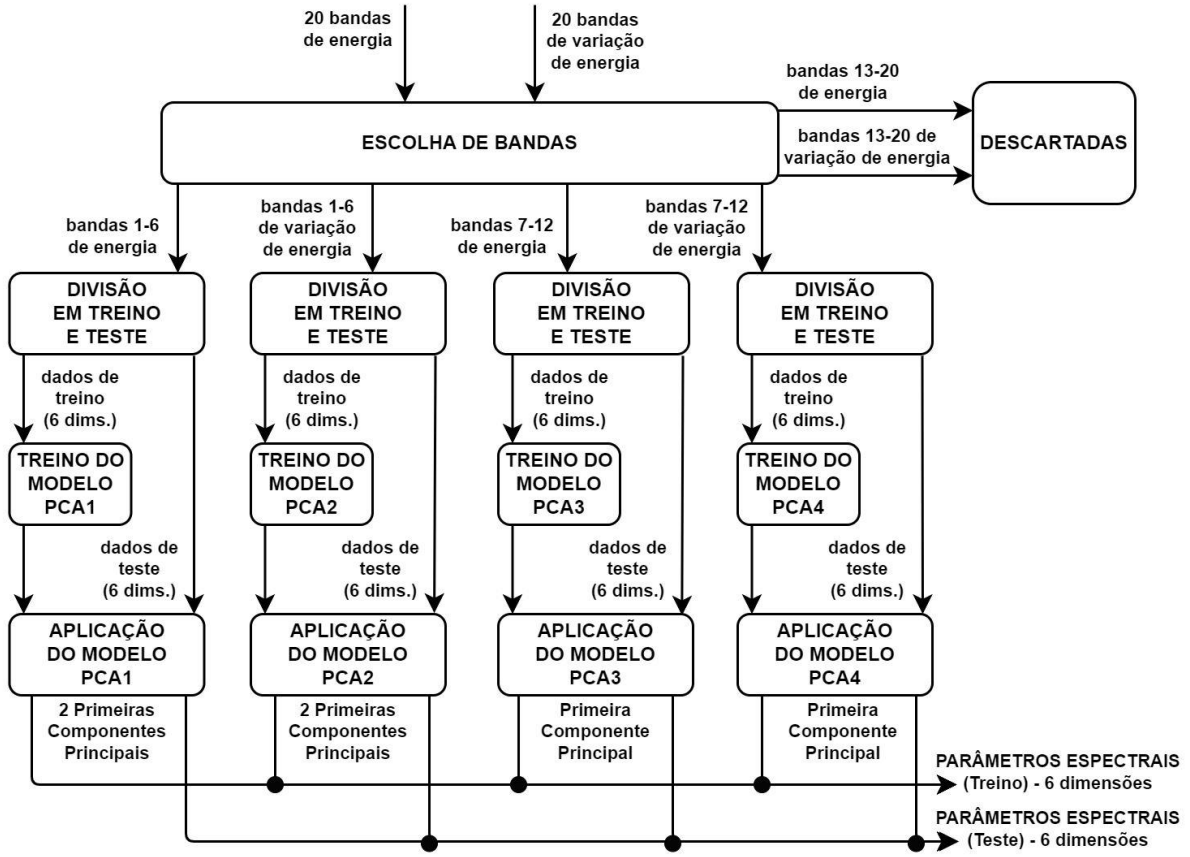


Figura 5-1 - Obtenção de parâmetros espectrais

Como pode ser observado na Figura 5-1, a aplicação do *PCA* ocorre durante o processo de divisão dos dados em conjuntos de treino e teste. O modelo *PCA* é ajustado apenas com os dados de treino e, de seguida, aplicado tanto aos dados de treino como aos de teste. Embora esta abordagem aumente a complexidade do processo, ela é crucial para garantir a validade dos resultados, assegurando que os dados de teste não sejam vistos em nenhuma etapa anterior.

Estes parâmetros espectrais incorporam informações de todas as bandas envolvidas no processo e, por essa razão, espera-se que as discriminações realizadas com estes parâmetros sejam eficazes. Para confirmar ou refutar esta expectativa, o desempenho destes parâmetros é avaliado e comparado com o desempenho dos parâmetros *bbLBST* e *bbLBSvT*. Aqueles que obtêm os melhores resultados são combinados com parâmetros acústicos e utilizados nas discriminações consideradas neste trabalho, para avaliar se a combinação destes parâmetros pode melhorar as suas eficácias.

## 5.2 Parâmetros acústicos

Os parâmetros acústicos foram extraídos dos sinais de fala através da aplicação *Praat* [72], embora esta ferramenta tenha sido utilizada de forma indirecta. O *Praat* é uma ferramenta de análise e síntese de fala, que permite a análise acústica, manipulação e visualização de sinais de fala, sendo amplamente utilizado em pesquisas linguísticas, fonéticas e de fonoaudiologia. O *Praat* foi usado com recurso a uma biblioteca de *Python* chamada *Parselmouth* [73]. Esta biblioteca foi desenvolvida para integrar as funcionalidades do *Praat* no ambiente *Python*, disponibilizando alguns métodos para esse efeito. Um desses métodos foi adaptado para permitir a extração dos parâmetros *jitter, shimmer* e *HNR* dos sinais de fala e foram estes os parâmetros posteriormente testados em combinação com os parâmetros espectrais.

A utilização de ferramentas e bibliotecas desenvolvidas por terceiros pode ser vantajosa, e no caso deste trabalho foi, pois poupou o trabalho e principalmente, o tempo necessário para implementar as mesmas funcionalidades. No entanto, também a sua utilização traz também algumas desvantagens, sendo uma delas a de que a sua operação não é transparente para o utilizador. Assim, quando existe algum problema na sua utilização, a resolução é muito difícil, se não impossível. A extração de parâmetros acústicos neste trabalho falhou em três sinais de fala do subconjunto utilizado nos estudos [44], [50] e [51], e por essa razão, esses três sinais de fala não foram utilizados neste estudo, como mencionado na Secção 3.1.2.

Também é difícil adaptar ferramentas criadas por terceiros para necessidades específicas, estando o utilizador limitado às opções disponibilizadas pelos recursos. O *Praat* acede diretamente aos ficheiros dos sinais de fala, impossibilitando a realização do pré-processamento descrito na Secção 4.1. Analisando as possíveis consequências desta impossibilidade, não se espera que as diferenças de amplitude entre os sinais influenciem os parâmetros acústicos em questão, mas desconhece-se se o *Praat* tem capacidade de reconhecer silêncios e não os processar e contabilizar na estimação dos parâmetros acústicos. Quanto às diferenças de duração entre os sinais de fala, estas podem potencialmente influenciar os valores dos parâmetros acústicos e, consequentemente, os resultados.

Conforme descrito na Secção 2.2.3.2 e observado, por exemplo no estudo [56], existem diversas variantes dos parâmetros *jitter* e *shimmer*. O *Praat* permite a extração de várias dessas variantes e, inicialmente, todas foram consideradas. No entanto, ao calcular a correlação entre as diferentes variantes de *jitter* e de *shimmer* para se aferir as suas redundâncias, verificou-se que as correlações eram bastante elevadas, com valores entre 0,99 e 1,0, exceto para o *shimmer* medido em dB. Esse resultado levou a uma ponderação sobre se seria vantajoso utilizar todas essas variantes, aumentando a dimensionalidade dos dados, especialmente quando se optou anteriormente por descartar informação importante das bandas de energia devido à necessidade de reduzir a dimensionalidade.

Assim, foi extraída e utilizada apenas uma variante de cada um dos dois parâmetros, sendo essa a variante global. Ou seja, foram calculados os valores para todos os ciclos glotais e, de seguida, calculado o valor médio para todo o sinal. Estes cálculos foram já descritos nas Equações 2-2 e 2-3, sendo o *jitter* e *shimmer* quantificados em percentagem. Quanto ao *HNR*, apresenta-se em decibéis,

pois, como explicado na Secção 2.2.3.2, a sua gama de valores é bastante alargada, sendo conveniente expressar este parâmetro numa escala logarítmica.

Na Secção 3.2.4 foi detalhado e demonstrado, como dimensões com diferentes escalonamentos podem comprometer o desempenho do classificador. Os três parâmetros acústicos têm escalas muito diferentes entre si e estas são também potencialmente diferentes das escalas dos parâmetros espectrais. Assim, existe a necessidade de se reescalonar os dados para que tenham as mesmas gamas de valores. O reescalonamento consistiu na normalização, para que todas as dimensões tenham média nula e variância unitária, e foi realizado através do *Standard Scaler*, disponível na biblioteca *Scikit-Learn*. Este recurso é utilizado através do treino de um modelo, seguido da aplicação desse modelo aos dados. Assim, os parâmetros espectrais e acústicos foram divididos em grupos de treino e teste, sendo os dados de treino usados para configurar o modelo *Standard Scaler*, que foi então aplicado aos dados de treino e teste. Tal como com o *PCA*, utilizado anteriormente, esta abordagem permite que os dados de teste se mantenham não vistos por nenhum processo anterior ao teste do classificador, garantindo a validade dos resultados.

A Figura 5-2 apresenta uma representação gráfica do processo de combinação dos parâmetros espectrais e acústicos.

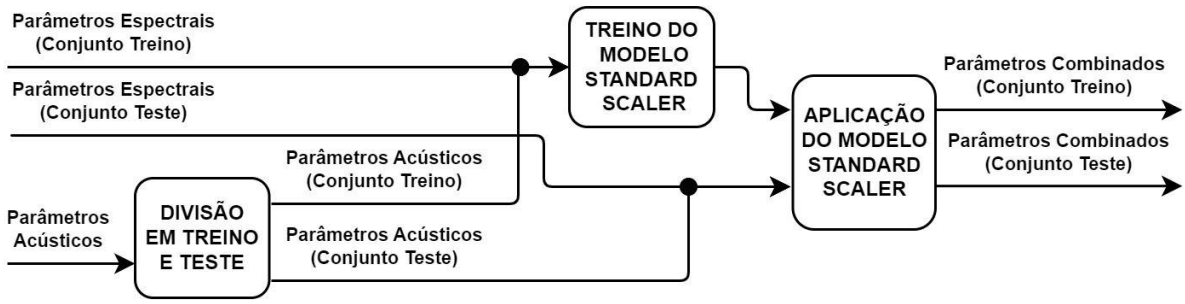


Figura 5-2 - Obtenção de parâmetros combinados

Os parâmetros acústicos foram alvo de uma análise prévia à sua utilização para determinar se deveriam todos ser utilizados e, em caso negativo, qual ou quais parâmetros que deveriam ser descartados. Para a escolha dos parâmetros acústicos a combinar com os parâmetros espectrais, foram analisadas as distribuições dos seus valores por classe para se avaliar se existiria uma maior sobreposição entre os valores em algum deles. Foram também realizadas todas as discriminações consideradas neste trabalho utilizando cada um dos parâmetros isoladamente, de modo a avaliar o potencial de cada um dos parâmetros para estas discriminações. Algumas dessas discriminações não produziram resultados válidos, identificados pelos valores da métrica *F1-Score*. Isto ocorreu devido ao desbalanceamento das classes, bastante acentuados em algumas discriminações, mas apesar dessas situações, os resultados válidos obtidos permitiram uma conclusão acerca de quais os parâmetros acústicos mais adequados para combinar com os espectrais.

## 5.3 Resultados e discussão

### 5.3.1 Escolha dos parâmetros espectrais

Os parâmetros espectrais derivados das bandas de energia e de variação de energia foram escolhidos entre as duas opções consideradas. A primeira opção são os parâmetros *bbLBST* e *bbLBSvT*, e a segunda opção são os parâmetros obtidos através de seleção de bandas e aplicação de *PCA*, conforme descrito na Secção 4.2.1. Para estes últimos parâmetros, foram obtidas as taxas de acerto médias e os valores médios da métrica *F1-Score*, bem como as respetivas margens de confiança a 95%, utilizando validação cruzada e com 1000 repetições. Os resultados obtidos foram comparados com os referentes aos parâmetros *bbLBST* e *bbLBSvT*, anteriormente obtidos e apresentados na Tabela 4-2. Ambos podem ser observados e comparados de seguida, na Tabela 5-1.

Tabela 5-1 - Comparação de resultados obtidos com os dois potenciais grupos de parâmetros espectrais

| Discriminação | Taxa de acerto [%] Seleção e PCA | F1-Score [%] Seleção e PCA | Taxa de acerto [%] bbLBST e bbLBSvT |
|---|---|---|---|
| **Control vs. Neuro (USP)** | **92,5 ± 2,8** | 93,1 ± 2,8 | 71,3 ± 6,8 |
| **Control vs. UFVP (sMEEI)** | 85,3 ± 3,1 | 80,9 ± 4,4 | **89,8 ± 1,7** |
| **Control vs. PhLP (USP)** | 96,0 ± 3,8 | 93,6 ± 6,2 | 96,0 ± 4,2 |
| **Control vs. PhLP (sMEEI)** | **85,7 ± 3,5** | 80,6 ± 5,2 | 83,5 ± 2,6 |
| **Control vs. Patológicas (USP)** | **94,5 ± 1,7** | 89,1 ± 3,9 | 88,0 ± 3,0 |
| **Control vs. Patológicas (sMEEI)** | **89,8 ± 2,2** | 77,7 ± 4,8 | 88,2 ± 1,7 |
| **PhLP vs. Neuro (USP)** | **86,8 ± 6,3** | 91,1 ± 4,1 | 84,3 ± 2,3 |
| **PhLP vs. UVFP (sMEEI)** | **64,0 ± 5,0** | 64,5 ± 5,1 | 60,4 ± 5,0 |
| **Multiclasse (USP)** | **85,2 ± 4,2** | n/a | 75,6 ± 3,1 |
| **Multiclasse (sMEEI)** | **63,7 ± 2,0** | n/a | 60,0 ± 3,8 |

Observa-se que apenas na discriminação entre as classes *control* e *UFVP*, os parâmetros *bbLBST* e *bbLBSvT* obtiveram o melhor desempenho, enquanto os desempenhos foram iguais apenas na discriminação entre as classes *control* e *PhLP*, na base de dados *USP*. Em todas as outras situações, o desempenho com os parâmetros obtidos através de seleção de bandas e aplicação de *PCA* obtiveram taxas de acerto médias mais elevadas, sendo as diferenças estatisticamente significativas na discriminação entre as classes *control* e *neuro*, e na discriminação multiclasse no *corpus USP*. Estes resultados demonstram ser mais adequada a utilização de parâmetros que reflitam todas as bandas entre 1 e 12, do que apenas as bandas com maiores valores de concentração e de variação de energia. Assim, os parâmetros obtidos através de seleção de bandas e aplicação de *PCA*, conforme definido na Secção 5.1, foram escolhidos como parâmetros espectrais.

### 5.3.2 Escolha dos parâmetros acústicos

Após a extração dos parâmetros acústicos através da biblioteca *Parselmouth*, e antes da sua combinação com os parâmetros espectrais determinados na secção anterior, é importante analisar a distribuição dos valores dos parâmetros acústicos por classe.

Na Figura 5-3 é apresentada a distribuição dos valores do *jitter* por classe. Devido às diferenças nas gamas de valores por classe, os gráficos são exibidos numa escala logarítmica.

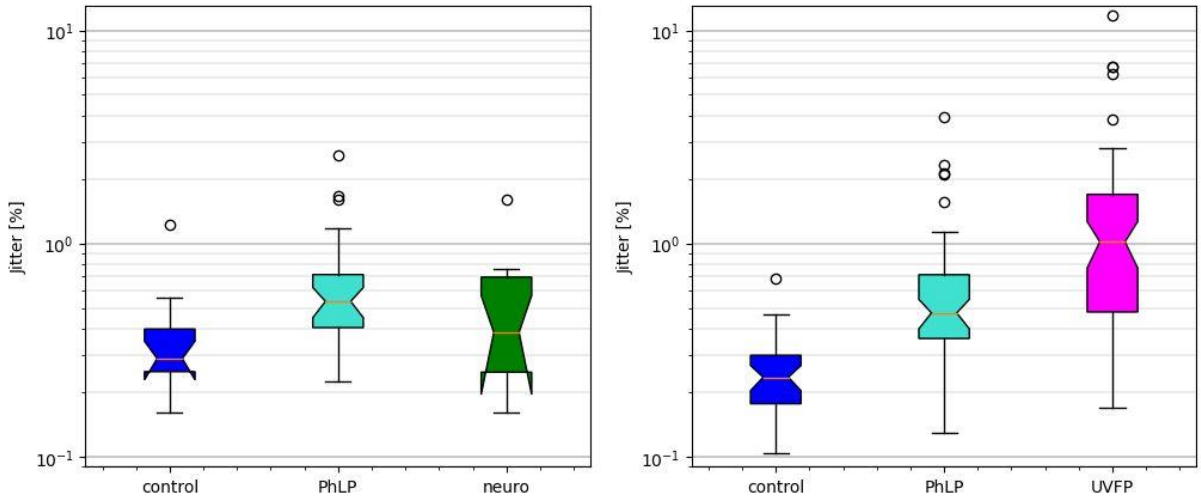


Figura 5-3 - Distribuição dos valores do parâmetro *jitter* por classe, na base de dados *USP* (esq.) e *sMEEI* (dir.)

Na Figura 5-4 pode ser visualizada a distribuição dos valores do parâmetro *shimmer*, por classe, para as duas bases de dados. Também neste caso, para melhor visualização, optou-se pela utilização de uma escala logarítmica.

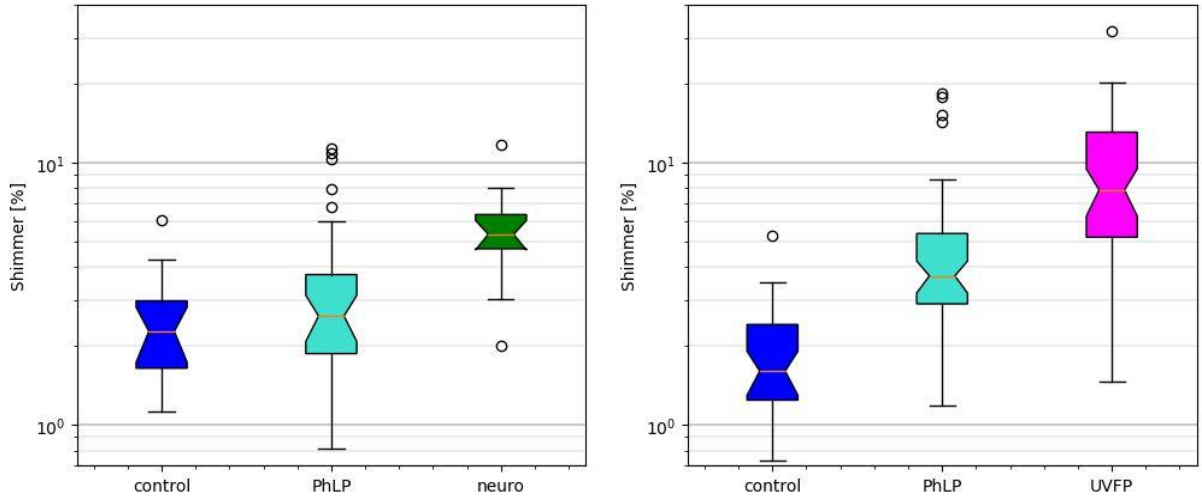


Figura 5-4 - Distribuição dos valores do *shimmer* por classe, na base de dados *USP* (esq.) e *sMEEI* (dir.)

Analisando as distribuições de *jitter* e de *shimmer*, verifica-se que a classe *control* apresenta sempre a mediana mais baixa, indicando valores mais baixos dos dois parâmetros para os oradores desta classe. Verifica-se também que os valores destes parâmetros, na classe *control*, variam numa faixa mais estreita, indicando uma maior estabilidade comparativamente aos oradores das classes patológicas. Estes comportamentos eram esperados, pois o *jitter* e *shimmer* são indicadores de instabilidade vocal e espera-se que esta seja mais pronunciada em oradores patológicos.

Entre as classes patológicas, os gráficos sugerem que a paralisia unilateral das pregas vocais é a condição que tem o maior impacto no *jitter* e no *shimmer*, pois tanto a mediana como a faixa de valores desses dois parâmetros são mais elevadas nesta classe. Verifica-se também que as condições neurodegenerativas afetam significativamente o *shimmer*, mas não o *jitter*. Este resultado é curioso, pois estas condições afetam o controlo das pregas vocais, e seria de esperar que essa influência se traduzisse num aumento dos valores de ambos os parâmetros, e não de apenas um.

Verifica-se também que existe, nos dois parâmetros, maior sobreposição entre os valores por classe no corpus *USP*. Não se conhecendo informações que possam justificar esta diferença, como o estado das condições patológicas, mais inicial ou mais avançado, esta diferença pode ser explicada pela variabilidade dos dados. No entanto, espera-se que essa maior sobreposição dos valores dos dois parâmetros possa ter um impacto negativo nos resultados referentes ao corpus *USP*.

De seguida, na Figura 5-5, apresenta-se a distribuição dos valores do parâmetro HNR por classe. Sendo este um parâmetro logarítmico, uma escala linear permite uma melhor visualização.

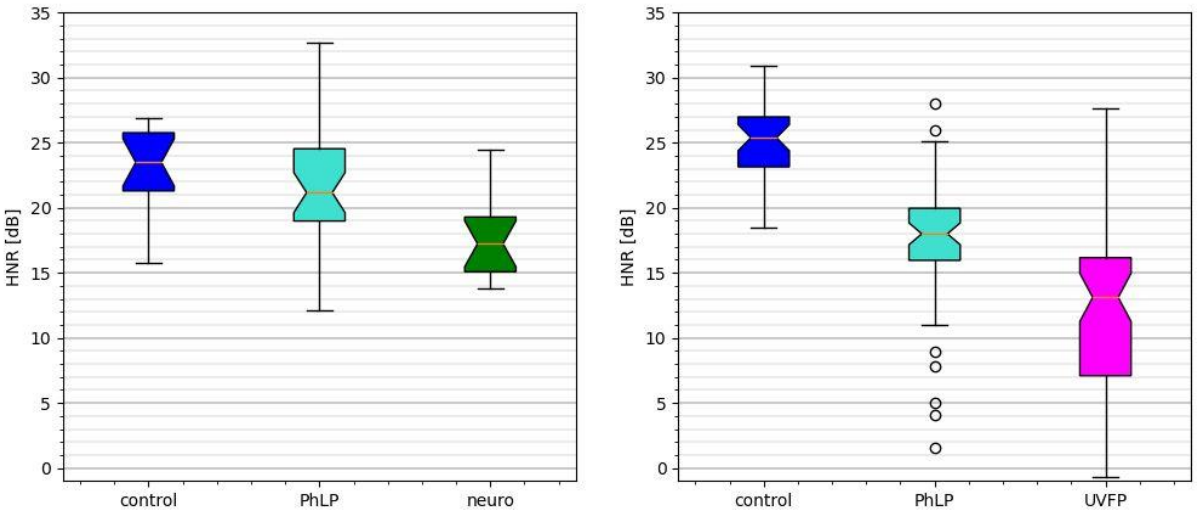


Figura 5-5 - Distribuição dos valores do parâmetro *HNR* por classe, na base de dados *USP* (esq.) e *sMEEI* (dir.)

Observa-se que a classe *control* apresenta, nas duas bases de dados, a mediana mais elevada para o parâmetro *HNR*, indicando valores mais altos nesta classe. Apresenta também a faixa de valores mais estreita, indicando menor variabilidade de valores para o parâmetro *HNR*. Estes comportamentos eram também esperados pois o *HNR* quantifica a qualidade vocal e espera-se que esta tenha valores mais elevados e consistentes em oradores saudáveis relativamente a oradores patológicos.

A distribuição dos valores de *HNR* por classe mostra que a paralisia unilateral das pregas vocais, assim como observado para o *jitter* e *shimmer*, é a condição que produz o maior impacto também no parâmetro *HNR*. Isso é evidenciado pela mediana mais baixa e pela maior variabilidade dos seus valores. Tal como verificado com os outros dois parâmetros, também se observa uma maior sobreposição nos valores do *HNR* por classe na base de dados *USP* em comparação com a *sMEEI*.

Comparando as distribuições dos três parâmetros por classe, não é evidente que algum deles tenha maior ou menor sobreposição. Portanto, esta análise não permite concluir acerca de qual, ou quais parâmetros são mais, ou menos adequados para combinar com os parâmetros espectrais.

Embora algumas discriminações não tenham produzido resultados válidos, impossibilitando uma análise mais completa, apresentam-se de seguida, na Tabela 5-2, os valores médios e margens de confiança de 95% para as taxas de acerto (*ACC*) bem como para a métrica *F1-Score* (*F1-S*) obtidas com cada um dos parâmetros acústicos utilizados isoladamente. Estes resultados foram obtidos com validação cruzada, utilizando 5 *folds* e 1000 repetições.

Tabela 5-2 - Resultados obtidos com parâmetros acústicos utilizados isoladamente

| | jitter | | shimmer | | HNR | | |
|---|---|---|---|---|---|---|---|
| **Discriminação** | **ACC [%]** | **F1-S [%]** | **ACC [%]** | **F1-S [%]** | **ACC [%]** | **F1-S [%]** | |
| **Control vs. Neuro (USP)** | 58,5 ± 5,1 | 68,5 ± 3,7 | **83,8 ± 3,4** | 84,7 ± 2,8 | 78,1 ± 4,8 | 79,3 ± 4,3 | |
| **Control vs. UFVP (sMEEI)** | 80,3 ± 2,1 | 78,8 ± 2,0 | 88,8 ± 1,3 | 86,5 ± 1,7 | **91,3 ± 1,9** | 89,3 ± 2,4 | |
| **Control vs. PhLP (USP)** | 66,9 ± 6,0 | n/e | 68,1 ± 0,3 | n/e | 68,0 ± 1,5 | n/e | (*) |
| **Control vs. PhLP (sMEEI)** | 80,7 ± 1,6 | 76,5 ± 1,9 | 80,9 ± 1,8 | 74,8 ± 2,6 | **85,1 ± 2,2** | 81,2 ± 3,0 | |
| **Control vs. Patol. (USP)** | 75,4 ± 0,4 | n/e | 75,4 ± 0,0 | n/e | 75,4 ± 0,5 | n/e | (*) |
| **Control vs. Patol. (sMEEI)** | 76,7 ± 1,7 | n/e | 87,7 ± 1,1 | 73,0 ± 1,7 | 87,1 ± 1,0 | 73,6 ± 2,0 | (*) |
| **PhLP vs. Neuro (USP)** | 68,9 ± 3,0 | 81,6 ± 2,2 | **80,4 ± 4,5** | 86,1 ± 3,5 | 69,6 ± 7,2 | 79,9 ± 4,7 | |
| **PhLP vs. UVFP (sMEEI)** | 66,6 ± 2,5 | 70,8 ± 1,9 | 69,6 ± 1,8 | 71,6 ± 1,6 | **70,7 ± 1,7** | 71,6 ± 2,0 | |
| **Multiclasse (USP)** | 51,2 ± 5,3 | n/a | **60,6 ± 3,5** | n/a | 53,4 ± 5,5 | n/a | |
| **Multiclasse (sMEEI)** | 61,2 ± 4,4 | n/a | 65,4 ± 1,8 | n/a | **68,7 ± 1,6** | n/a | |

(*) resultados inválidos

Observa-se que o parâmetro *jitter* obtém o pior desempenho em todas as discriminações, chegando a ter uma taxa de acerto média 20% inferior às obtidas pelos outros dois parâmetros acústicos num caso. Estes resultados sugerem que o *jitter* será, dos três parâmetros acústicos considerados, aquele que terá menor potencial para as discriminações entre as classes.

O parâmetro *shimmer* demonstra um desempenho globalmente melhor nas discriminações da base de dados *USP*, enquanto o *HNR* apresenta um melhor desempenho global no *corpus sMEEI*. Embora os resultados das duas bases de dados não sejam diretamente comparáveis, observa-se que as discriminações efetuadas na base de dados *sMEEI* obtêm taxas de acerto médias mais elevadas.

A discriminação entre as classes *control* e *PhLP*, que permitiria uma comparação direta entre o desempenho dos parâmetros nas duas bases de dados, não produziu resultados válidos numa das bases de dados, impossibilitando uma comparação direta entre os resultados. Apesar disso, considerando que o *jitter* demonstrou consistentemente ser o parâmetro acústico com pior desempenho, que o *shimmer* demonstrou ser o parâmetro com maior potencial nas discriminações do *corpus USP* e o *HNR* nas da base de dados *sMEEI*, optou-se por descartar o parâmetro *jitter*.

Apresentam-se de seguida, na Figura 5-6, os oradores da base de dados *USP*, à esquerda, e do *corpus sMEEI*, à direita, representados nos seus valores dos parâmetros *shimmer* e *HNR*.

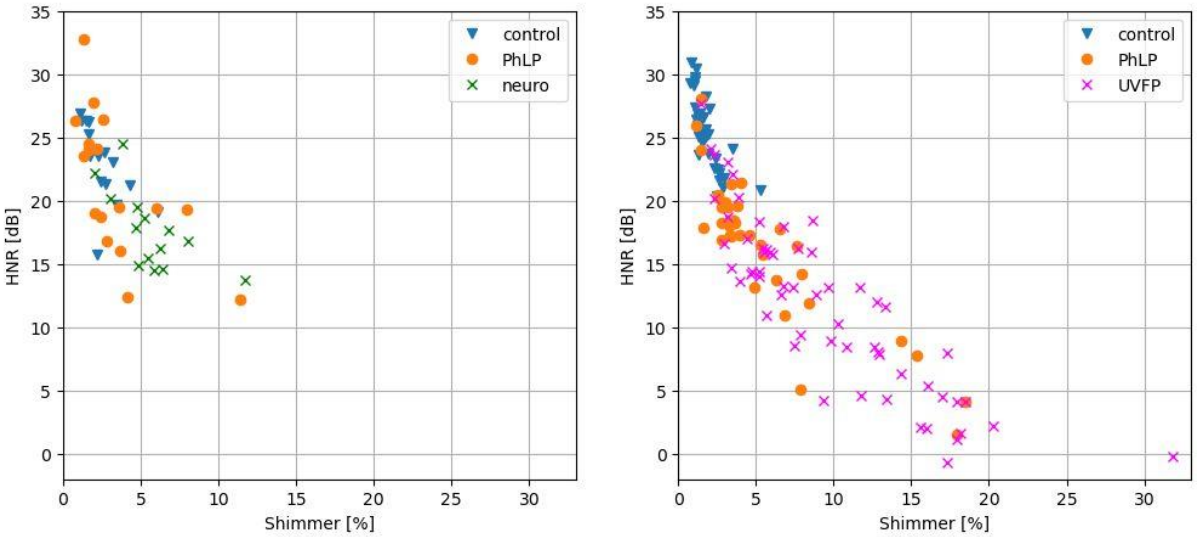


Figura 5-6 - Distribuição dos sinais de fala das bases de dados *USP* (esq.) e *sMEEI* (dir.) de acordo com os valores dos parâmetros *shimmer* e *HNR*

No *corpus USP*, observa-se que todos os oradores saudáveis, exceto um, apresentam valores de *HNR* superiores a 20 dB e de *shimmer* inferiores a 5%, que são valores típicos para oradores saudáveis adultos. No entanto, observam-se vários oradores das classes patológicas com valores similares, ocupando a mesma região que os oradores saudáveis.

Na base de dados *sMEEI*, existe uma distinção mais definida entre as classes. A classe dos oradores saudáveis também apresenta valores de *shimmer* abaixo de 5% e de *HNR* acima de 20 dB, tal como acontece no *corpus USP*. A classe *UVFP* destaca-se das restantes por apresentar os valores mais altos de *shimmer* e mais baixos de *HNR*, reforçando o observado nas Figuras 5-4 e 5-5, onde se concluiu que esta patologia tem o maior impacto nestes dois parâmetros. As classes *control* e *PhLP* mostram alguma sobreposição, mas pode ser observada uma tendência onde os oradores saudáveis têm valores mais altos de *HNR* e mais baixos de *shimmer*.

Observa-se uma maior sobreposição entre as classes no *corpus USP*, em linha com o observado e discutido anteriormente nesta secção. Esses resultados sugerem que os parâmetros acústicos podem ser menos eficazes para discriminar as classes nesse *corpus* em comparação com o *sMEEI*.

### 5.3.3 Análise das discriminações entre classes através de parâmetros combinados

São agora analisadas as nove diferentes discriminações entre classes realizadas com recurso à combinação de parâmetros espectrais e acústicos. Recorde-se que os parâmetros espectrais utilizados são os obtidos por seleção de bandas e posterior redução de dimensionalidade através de *PCA*, conforme ilustrado na Figura 5-1, resultando em 6 dimensões. Os parâmetros acústicos utilizados são o *shimmer* e o *HNR*. A combinação destes parâmetros resulta em 8 dimensões.

Os resultados são obtidos, tal como anteriormente, com utilização de validação cruzada, com 5 *folds* e 1000 repetições. As taxas de acerto e a métrica *F1-Score* são apresentadas no seu valor médio e margem de confiança de 95%. As matrizes de confusão apresentam valores médios, razão pela qual contêm valores decimais. São também apresentados os valores normalizados, entre parenteses, expressos em percentagem, para permitir uma análise mais completa.

#### 5.3.3.1 Discriminação entre oradores saudáveis e com patologias neurodegenerativas

Apresentam-se na Tabela 5-3, os resultados para a discriminação entre oradores saudáveis e patológicos, com patologias neurodegenerativas, realizada na base de dados *USP*.

Tabela 5-3 - Resultados na discriminação entre oradores saudáveis e com patologias neurodegenerativas (base de dados *USP*)

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 77,5 ± 7,1 | 92,5 ± 2,8 | **93,0 ± 1,3** |
| **F1-Score [%]** | 78.8 ± 6,8 | 93,1 ± 2,8 | 93,6 ± 1,3 |

E podem ser visualizadas as taxas de acerto, na Figura 5-7, para melhor análise.

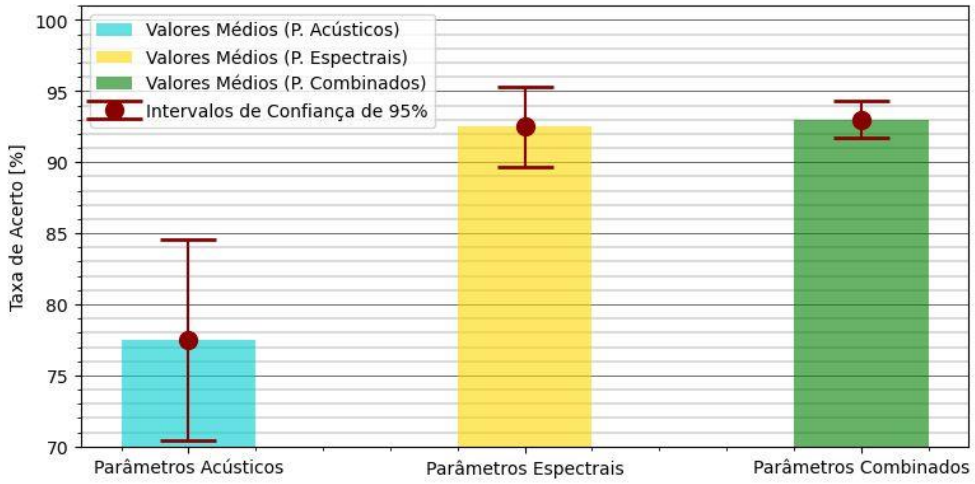


Figura 5-7 - Taxas de acerto na discriminação entre oradores saudáveis e com patologias neurodegenerativas

Os resultados indiciam que a combinação de parâmetros acústicos e espectrais melhora o desempenho da discriminação entre oradores saudáveis e com patologias neurodegenerativas em comparação com a utilização de apenas parâmetros espectrais. Embora essa melhoria não seja estatisticamente significativa, pois está dentro das margens de confiança de 95%, é importante destacar que a introdução dos parâmetros *shimmer* e *HNR* melhorou o desempenho da discriminação. Especialmente porque os parâmetros acústicos apresentaram uma taxa de acerto média significativamente inferior, em 15 pontos percentuais, à obtida com os parâmetros espectrais, mas apesar desta diferença no desempenho, a introdução dos parâmetros acústicos mostrou-se vantajosa nesta discriminação.

A seguir, são apresentadas, na Tabela 5-4, as matrizes de confusão obtidas com parâmetros espectrais, em cima, e com parâmetros combinados, em baixo, nesta discriminação.

Tabela 5-4 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores saudáveis e com patologias neurodegenerativas, com valores normalizados, em percentagem, entre parênteses

| | | **Control - predicted** | **Neuro - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 14,8 (98,97 %) | 0,2 (1,03 %) |
| | **Neuro - real** | 2,0 (14,47 %) | 12,0 (85,53 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 15,0 (99,75 %) | 0,0 (0,25 %) |
| | **Neuro - real** | 2,0 (14,29 %) | 12,0 (85,71 %) |

Observa-se que a introdução de parâmetros acústicos na discriminação melhorou a identificação de oradores em ambas as classes, embora a melhoria na identificação de oradores patológicos seja residual, e apenas perceptível através da análise dos valores normalizados. Estes resultados, mais concretamente as melhorias em todos os campos da matriz de confusão, sugerem que a combinação de parâmetros poderá ser vantajosa nesta discriminação.

### 5.3.3.2 Discriminação entre oradores saudáveis e patológicos com UVFP

Apresentam-se na Tabela 5-5 os resultados para a discriminação entre oradores saudáveis e patológicos, com paralisia unilateral das pregas vocais, na base de dados *sMEEI*.

Tabela 5-5 - Resultados na discriminação entre oradores saudáveis e com *UVFP*, efetuada na base de dados *sMEEI*

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 90,7 ± 1,1 | 85,3 ± 3,1 | **91,3 ± 1,9** |
| **F1-Score [%]** | 88,5 ± 1,3 | 80,9 ± 4,4 | 89,2 ± 2,4 |

Representam-se graficamente, na Figura 5-8, as taxas de acerto para esta discriminação.

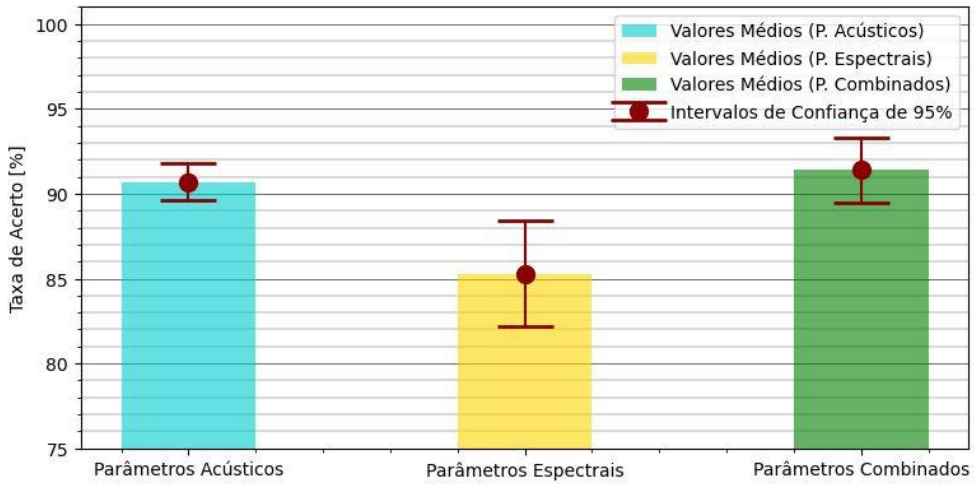


Figura 5-8 - Taxas de acerto na discriminação entre oradores saudáveis e com *UVFP*

Verifica-se que a introdução dos parâmetros acústicos na discriminação entre oradores saudáveis e patológicos com *UVFP* melhora de forma muito significativa o desempenho. Conforme verificado na Secção 5.3.2, a paralisia unilateral das pregas vocais é a patologia que mais impacta os valores dos parâmetros acústicos. Essa conclusão é reforçada pelos resultados agora obtidos, que mostram que os parâmetros acústicos apresentam um desempenho superior ao dos parâmetros espectrais. Portanto, nesta discriminação, a vantagem de incluir parâmetros acústicos é evidente.

Na Tabela 5-6, são apresentadas as matrizes de confusão obtidas para esta discriminação.

Tabela 5-6 - Matrizes de confusão médias obtidas com parâmetros espectrais (em cima), com parâmetros combinados (em baixo) na discriminação entre oradores saudáveis e com paralisia unilateral das pregas vocais, com valores normalizados entre parênteses

| | | **Control - predicted** | **UVFP - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 29,6 (82,30 %) | 6,4 (17,70 %) |
| | **UVFP - real** | 7,6 (12,86 %) | 51,4 (87,14 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 34,3 (95,36 %) | 1,7 (4,64 %) |
| | **UVFP - real** | 6,6 (11,19 %) | 52,4 (88,81 %) |

Comparando as matrizes de confusão obtidas, verifica-se que a introdução de parâmetros acústicos na discriminação introduz melhorias globais, subindo as taxas de identificação de oradores de ambas as classes. Essa essa melhoria é mais acentuada nos oradores saudáveis, com uma subida

de quase 13 pontos percentuais. Estes resultados demonstram de forma clara ser vantajosa a inclusão de parâmetros acústicos na discriminação entre oradores das classes *control* e *UVFP*.

#### 5.3.3.3 Discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas

Esta é a única discriminação entre classes iguais nas duas bases de dados. Assim, os resultados podem ser diretamente comparados e, por essa razão, serão agrupados.

Os resultados referentes a esta discriminação, nas duas bases de dados, apresentam-se na Tabela 5-7.

Tabela 5-7 - Resultados na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas

| Métrica | Corpus | Parâmetros acústicos | Parâmetros espectrais | Parâmetros combinados |
|---|---|---|---|---|
| **Taxa de acertos [%]** | USP | 68,1 ± 0,1 (*) | 96,0 ± 3,8 | **96,8 ± 2,8** |
| **F1-Score [%]** | USP | n/e | 93,6 ± 6,2 | 95,1 ± 4,4 |
| **Taxa de acertos [%]** | sMEEI | 84,6 ± 1,9 | 85,7 ± 3,5 | **91,9 ± 2,5** |
| **F1-Score [%]** | sMEEI | 80,3 ± 2,8 | 80,6 ± 5,2 | 89,6 ± 3,4 |

(*) resultado inválido

Podem ser visualizadas de seguida, na Figura 5-9, as taxas de acerto para esta discriminação, na base de dados *USP*, à esquerda, e para a base de dados *sMEEI*, à direita.

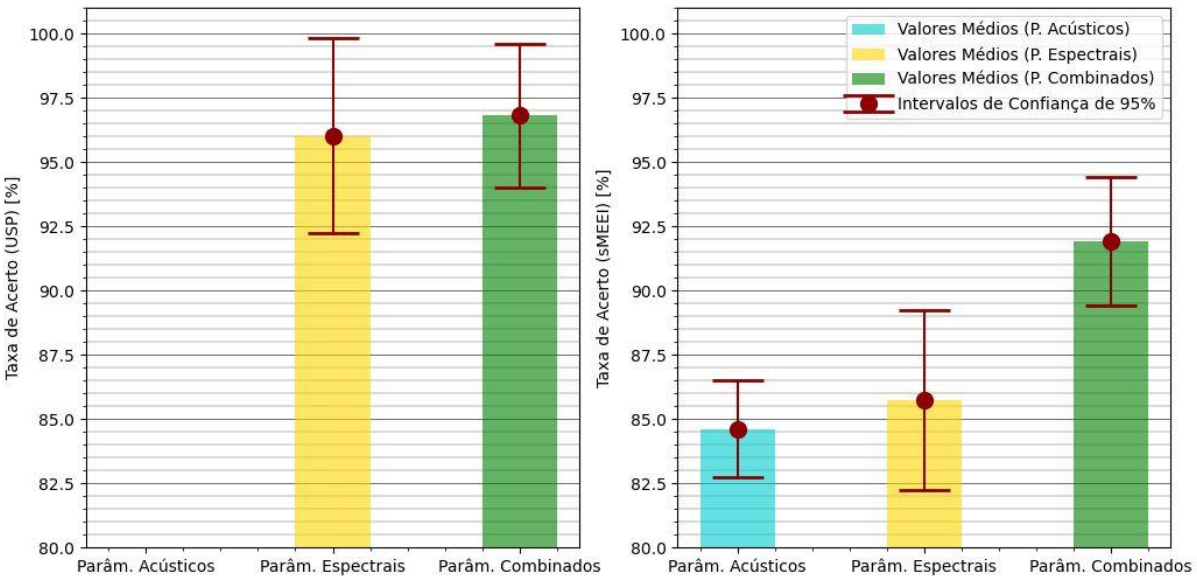


Figura 5-9 - Taxas de acerto na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas na base de dados *USP* (esq.) e *sMEEI* (dir.)

Esta discriminação não obteve resultados válidos ao utilizar apenas parâmetros acústicos no *corpus USP*, devido ao desbalanceamento das classes. Apesar disso, a introdução desses parâmetros na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas demonstrou ser vantajosa nas duas bases de dados, sendo essa melhoria mais evidente numa delas. Na base de dados *sMEEI*, as taxas de acerto médias obtidas com parâmetros acústicos e com parâmetros espectrais são semelhantes, com 1,1% de diferença. No entanto, verifica-se que a combinação dos parâmetros introduz uma melhoria notável na discriminação, de 6,2%, superior à soma das margens de confiança de 95%, que será de 6,0%, tornando a melhoria estatisticamente significativa. No corpus *USP*, a taxa de acerto média subiu 0,8% com a introdução de parâmetros acústicos. Embora esta melhoria não seja significativa quando comparada com as margens de confiança de 95%, é relevante quando analisada em conjunto com o resultado obtido no corpus *sMEEI*, pois demonstra a vantagem de combinar parâmetros acústicos e espectrais na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas.

De seguida, na Tabela 5-8, apresentam-se as matrizes de confusão obtidas nesta discriminação no *corpus USP*, e na Tabela 5-9 são apresentadas as matrizes de confusão para o *corpus sMEEI*.

Tabela 5-8 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas no corpus *USP,* com valores normalizados entre parênteses

| | | **Control - predicted** | **PhLP - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 13,6 (90,86 %) | 1,4 (9,14 %) |
| | **PhLP - real** | 0,5 (1,55 %) | 31,5 (98,45 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 14,6 (97,17 %) | 0,4 (2,83 %) |
| | **PhLP - real** | 1,1 (3,31 %) | 30,9 (96,69 %) |

Tabela 5-9 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas no corpus *sMEEI*, com valores normalizados entre parênteses

| | | **Control - predicted** | **PhLP - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 27,5 (76,31 %) | 8,5 (23,69 %) |
| | **PhLP - real** | 4,6 (8,27 %) | 51,4 (91,73 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 32,3 (89,67 %) | 3,7 (10,33 %) |
| | **PhLP - real** | 3,7 (6,64 %) | 52,3 (93,36 %) |

As matrizes de confusão obtidas na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas permitem concluir que a introdução de parâmetros acústicos nesta discriminação melhora a identificação de oradores saudáveis em ambas as bases de dados. Estas melhorias são significativas, com valores de 6,31% na base de dados *USP* e de 13,36% na identificação de oradores saudáveis na base de dados sMEEI.

No entanto, em relação à identificação de oradores patológicos, os resultados obtidos nas duas bases de dados são divergentes, pois na base de dados *USP* houve uma diminuição de 1,76% enquanto no *corpus sMEEI* houve uma melhoria de 1,63% nessa identificação. Considerando todos os resultados obtidos e avaliando os ganhos e perdas que estes demonstram, conclui-se que a inclusão dos parâmetros acústicos foi vantajosa também nesta discriminação, sendo essa vantagem mais evidente na base de dados *sMEEI*.

#### 5.3.3.4 Discriminação entre oradores saudáveis e patológicos (corpus USP)

Na Tabela 5-10, apresentada de seguida, podem ser visualizados os resultados obtidos para a discriminação entre oradores saudáveis e patológicos, realizada na base de dados *USP*.

Tabela 5-10 - Resultados na discriminação entre oradores saudáveis e patológicos na base de dados *USP*

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 75,4 ± 0,0 (*) | **94,5 ± 1,7** | 93,5 ± 2,7 |
| **F1-Score [%]** | n/e | 89,1 ± 3,9 | 87,2 ± 5,6 |

resultado inválido (*)

E apresentam-se as taxas de acerto obtidas nesta discriminação, na Figura 5-10.

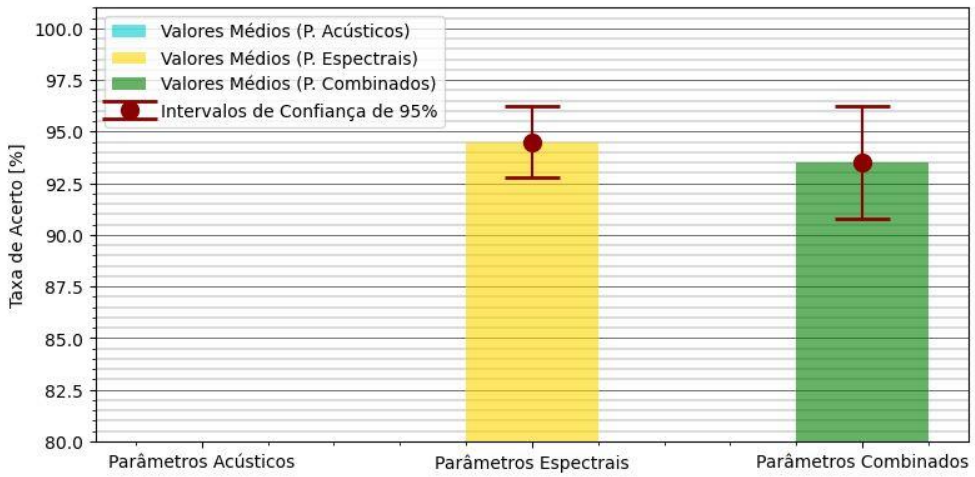


Figura 5-10 - Taxas de acerto na discriminação entre oradores saudáveis e patológicos no corpus *USP*

Esta discriminação também não obteve resultados válidos ao utilizar apenas parâmetros acústicos, devido ao desbalanceamento das classes. Isto impossibilita uma análise mais completa ao contributo dos parâmetros acústicos quando combinados com os espectrais, pois o seu desempenho isolado não pode ser avaliado. Pode-se observar, no entanto, que a combinação dos parâmetros acústicos e espectrais resulta numa ligeira degradação do desempenho na discriminação, baixando um ponto percentual na taxa de acerto média.

As matrizes de confusão obtidas na discriminação entre oradores saudáveis e patológicos apresentam-se na Tabela 5-11.

Tabela 5-11 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores saudáveis e patológicos no corpus *USP*, com valores normalizados, expressos em percentagem, entre parênteses

| | | **Control - predicted** | **Patológicas - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 13,7 (91,08 %) | 1,3 (8,92 %) |
| | **Patol. - real** | 2,0 (4,36 %) | 44,0 (95,64 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 13,5 (89,83 %) | 1,5 (10,17 %) |
| | **Patol. - real** | 2,4 (5,25 %) | 43,6 (94,75 %) |

As matrizes de confusão indicam que a introdução de parâmetros acústicos degrada a identificação tanto de oradores saudáveis como de patológicos. Verificando-se uma degradação, obviamente não se pode concluir que a introdução de parâmetros acústicos seja vantajosa nesta discriminação. No entanto, não sendo uma degradação muito acentuada, inferior a 1% tanto na identificação de oradores saudáveis como na de oradores patológicos, também não se pode concluir que a introdução de parâmetros acústicos nesta discriminação seja claramente desvantajosa.

#### 5.3.3.5 Discriminação entre oradores saudáveis e patológicos (corpus sMEEI)

Apresentam-se na Tabela 5-12 os resultados para a discriminação entre oradores saudáveis e patológicos, na base de dados *sMEEI*.

Tabela 5-12 - Resultados na discriminação entre oradores saudáveis e patológicos na base de dados *sMEEI*

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 87,4 ± 1,1 | 89,8 ± 2,2 | **90,9 ± 1,8** |
| **F1-Score [%]** | 73,9 ± 2,4 | 77,7 ± 4,8 | 81,2 ± 4,0 |

E apresentam-se, na Figura 5-11, as taxas de acerto obtidas nesta discriminação.

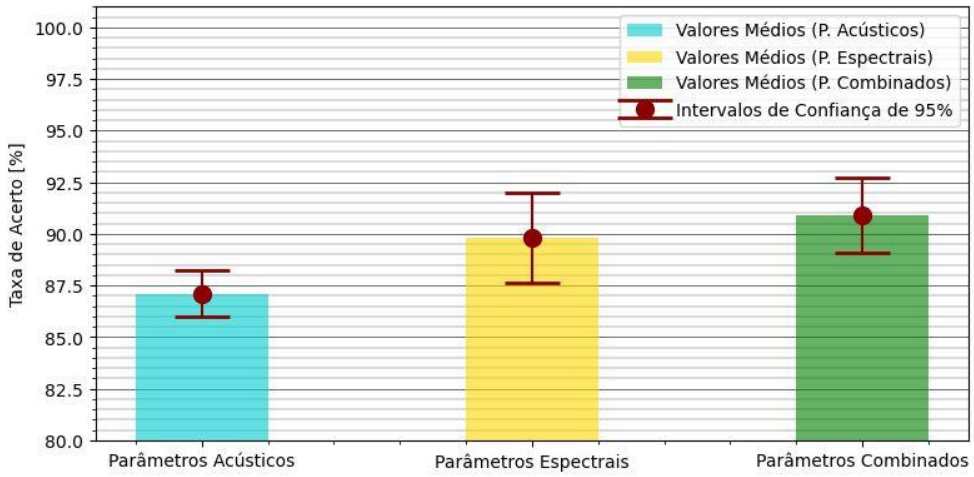


Figura 5-11 - Taxas de acerto na discriminação entre oradores saudáveis e patológicos no corpus *sMEEI*

A introdução de parâmetros acústicos na discriminação entre oradores saudáveis e patológicos melhora o desempenho nesta discriminação efetuada no corpus *sMEEI.* Este resultado era esperado pois a inclusão dos parâmetros acústicos tinha já melhorado as discriminações entre vozes saudáveis e cada uma das patologias presentes nesta base de dados. No entanto, ao contrário das discriminações anteriores, onde a melhoria tinha sido estatisticamente significativa, sendo superior às margens de confiança de 95%, isso não acontece neste caso onde as classes patológicas estão agrupadas. Estes resultados sugerem que não existe uma vantagem na agregação das classes *PhLP* e *UVFP* numa só.

Apresentam-se de seguida, na Tabela 5-13, as matrizes de confusão obtidas nesta discriminação, na base de dados *sMEEI*.

Tabela 5-13 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores saudáveis e patológicos no corpus *sMEEI*, com valores normalizados entre parênteses

| | | **Control - predicted** | **Patológicas - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 26,8 (74,57 %) | 9,2 (25,43 %) |
| | **Patol. - real** | 6,2 (5,42 %) | 108,8 (94,58 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 29,8 (82,79 %) | 6,2 (17,21 %) |
| | **Patol. - real** | 7,6 (6,60 %) | 107,4 (93,40 %) |

As matrizes de confusão indicam que a inclusão dos parâmetros acústicos na discriminação entre oradores saudáveis e patológicos melhora bastante a identificação de oradores saudáveis, com uma melhoria superior a 8 pontos percentuais, correspondente a mais três oradores saudáveis corretamente identificados. Por outro lado, a inclusão destes parâmetros degrada ligeiramente a

identificação de oradores patológicos, baixando o valor em 1,18%. Estes resultados, em conjunto com as taxas de acerto, sugerem ser vantajosa a combinação de parâmetros espectrais e acústicos nesta discriminação.

No entanto, pode-se concluir de forma mais categórica que é mais vantajoso não agrupar os oradores com patologias laríngeas fisiológicas e com paralisia unilateral das pregas vocais numa única classe. Nas discriminações efetuadas entre oradores saudáveis e com cada uma dessas patologias, tanto a identificação de oradores saudáveis como a de oradores patológicos melhoraram com a inclusão de parâmetros acústicos na discriminação, conforme observado nas Tabelas 5-6 e 5-9. Isso contrasta com o que ocorre quando as patologias estão agrupadas numa única classe, onde essa identificação diminui, embora de forma ligeira.

#### 5.3.3.6 Discriminação entre patologias laríngeas fisiológicas e neurodegenerativas

A Tabela 5-14, apresentada de seguida, mostra os resultados obtidos para a discriminação entre as patologias presentes na base de dados *USP*.

Tabela 5-14 - Resultados na discriminação entre oradores com patologias laríngeas fisiológicas e com patologias neurodegenerativas (base de dados *USP*)

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 76,8 ± 5,6 | **86,8 ± 6,3** | 82,7 ± 7,1 |
| **F1-Score [%]** | 83,8 ± 3,9 | 91,1 ± 4,1 | 88,0 ± 4,8 |

Podem ser visualizadas de seguida, na Figura 5-12, as taxas de acerto para a discriminação entre as patologias laríngeas fisiológicas e neurodegenerativas.

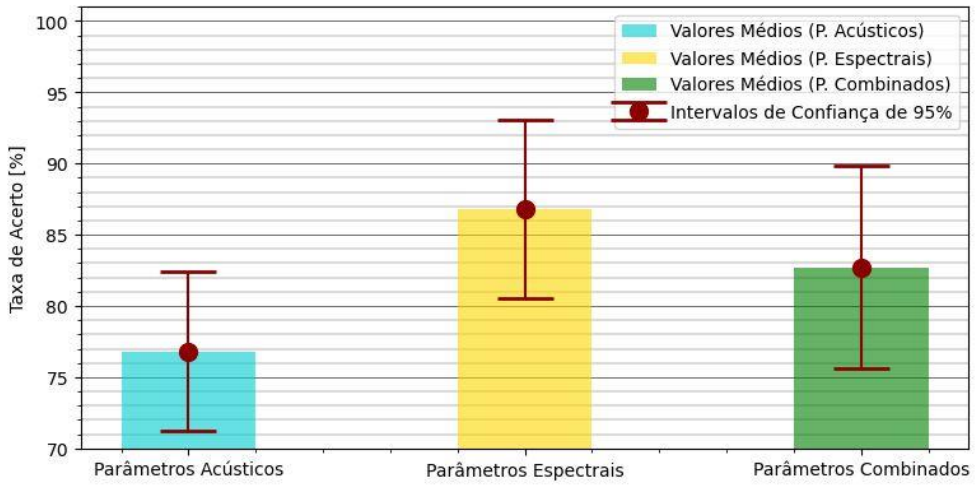


Figura 5-12 - Taxas de acerto na discriminação entre oradores com patologias laríngeas fisiológicas e com patologias neurodegenerativas

Na discriminação entre as duas patologias presentes na base de dados USP observa-se uma degradação no desempenho, refletida pela redução da taxa de acerto média em 4,1%. Embora essa diminuição não seja estatisticamente significativa, pois é menor que ambas as margens de confiança de 95%, o valor é expressivo, não se podendo considerar como uma degradação ligeira.

Apresentam-se de seguida, na Tabela 5-15, as matrizes de confusão obtidas na discriminação entre patologias.

Tabela 5-15 - Matrizes de confusão médias obtidas com parâmetros espectrais (em cima), com parâmetros combinados (em baixo) na discriminação entre oradores com patologias laríngeas fisiológicas e com patologias neurodegenerativas, com valores normalizados entre parênteses

| | | **PhLP - predicted** | **Neuro - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **PhLP - real** | 31,2 (97,48 %) | 0,8 (2,52 %) |
| | **Neuro - real** | 5,3 (37,60 %) | 8,7 (62,40 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **PhLP - real** | 29,2 (91,27 %) | 2,8 (8,73 %) |
| | **Neuro - real** | 5,2 (36,81 %) | 8,8 (63,19 %) |

As matrizes de confusão mostram que a identificação de oradores da classe PhLP piora de uma forma acentuada com a inclusão de parâmetros acústicos na discriminação, apresentando uma descida de mais de 6 pontos percentuais. Por outro lado, a identificação de oradores com patologias de ordem neuromuscular melhora ligeiramente, com um aumento de 0,79%. Estes resultados, em conjunto com as taxas de acerto obtidas, sugerem que a introdução de parâmetros acústicos na discriminação entre estas duas patologias não traz benefícios.

### 5.3.3.7 Discriminação entre patologias laríngeas fisiológicas e UVFP

Apresentam-se na Tabela 5-16, os resultados para a discriminação entre as patologias presentes na base de dados *sMEEI*.

Tabela 5-16 - Resultados na discriminação entre oradores com patologias laríngeas fisiológicas e com *UVFP* (base de dados *sMEEI*)

| | **Parâmetros acústicos** | **Parâmetros espectrais** | **Parâmetros combinados** |
|---|---|---|---|
| **Taxa de acertos [%]** | 70,4 ± 2,8 | 64,0 ± 5,0 | **72,7 ± 4,3** |
| **F1-Score [%]** | 71,3 ± 2,5 | 64,5 ± 5,1 | 72,9 ± 4,5 |

E, na Figura 5-13, podem ser visualizadas as taxas de acerto referentes à discriminação entre patologias laríngeas fisiológicas e paralisia unilateral das pregas vocais.

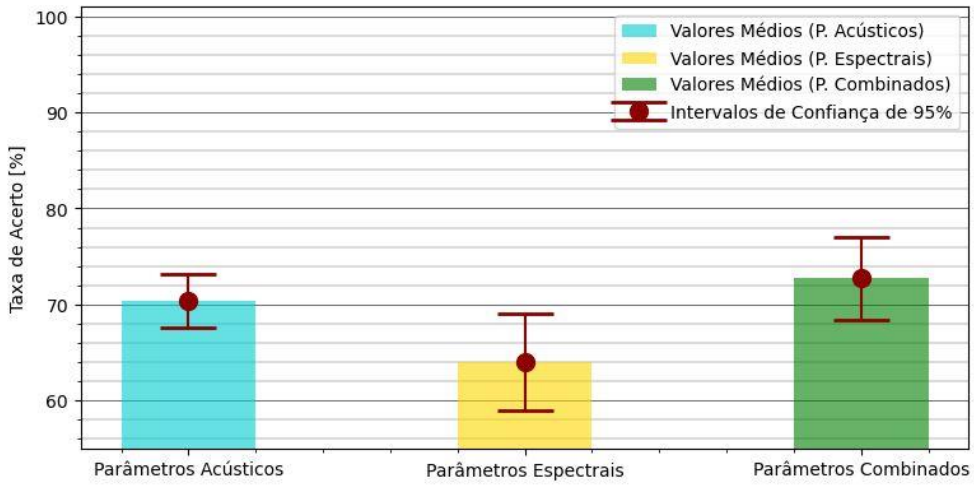


Figura 5-13 - Taxas de acerto na discriminação entre oradores com patologias laríngeas fisiológicas e com *UVFP*

Mais uma vez, os parâmetros acústicos demonstram ser úteis numa discriminação envolvendo a classe *UVFP*, apresentando uma taxa de acerto média mais elevada que a obtida com parâmeros espectrais. A paralisia unilateral das pregas vocais tem um grande impacto no *shimmer* e no *HNR*, que se reflete também nesta discriminação, melhorando a taxa de acerto quando estes dois parâmetros acústicos são combinados com os espectrais.

As matrizes de confusão obtidas na discriminação entre as patologias presentes no *corpus sMEEI* são apresentadas na Tabela 5-17.

Tabela 5-17 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre oradores com patologias laríngeas fisiológicas e com paralisia unilateral das pregas vocais, com valores normalizados entre parênteses

| | | **Control - predicted** | **UVFP - predicted** |
|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 37,6 (67,20 %) | 18,4 (32,80 %) |
| | **UVFP - real** | 23,1 (39,07 %) | 35,9 (60,93 %) |

| | | | |
|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 42,4 (75,66 %) | 13,6 (24,34 %) |
| | **UVFP - real** | 17,8 (30,16 %) | 41,2 (69,84 %) |

Observa-se, nas matrizes de confusão, que a identificação de oradores patológicos melhora com a inclusão de parâmetros acústicos na discriminação, para ambas as patologias. Na identificação de oradores com patologias laríngeas fisiológicas, a melhoria é de 8,46% enquanto na identificação de oradores cm paralisia unilateral das pregas vocais, a melhoria atinge 8,91%. Estes resultados indicam que a combinação de parâmetros acústicos e espectrais é vantajosa para esta discriminação.

### 5.3.3.8 Discriminação multiclasse (USP)

Nesta discriminação, para além da taxa de acerto global, é também possível obter a taxa de acerto nas discriminações de cada uma das classes contra as restantes, utilizando a abordagem *One vs. All* (*OvA*) conforme descrita na Equação 3-3. Por outro lado, a métrica *F1-Score* não pode ser aplicada diretamente a classificações multiclasse, pois é específica para classificações binárias. Assim, embora seja possível calcular o *F1-Score* para as classificações *OvA*, esta métrica não se apresenta aqui, sendo verificada a validade dos resultados através de inspeção às matrizes de confusão.

Apresentam-se as taxas de acerto obtidas nesta discriminação de seguida, na Tabela 5-18.

Tabela 5-18 - Taxas de acerto, em percentagem, na discriminação multiclasse, e nas discriminações *OvA* derivadas, no *corpus USP*

| | **Control vs. All** | **PhLP vs. All** | **Neuro vs. All** | **Multiclasse** |
|---|---|---|---|---|
| **Parâmetros Acústicos** | 73,8 ± 4,3 | 59,2 ± 5,6 | 80,1 ± 4,0 | 56,5 ± 5,4 |
| **Parâmetros Espectrais** | **94,5 ± 2,0** | **90,4 ± 4,0** | **85,5 ± 3,9** | **85,1 ± 4,2** |
| **Parâmetros Combinados** | 94,2 ± 2,7 | 85,9 ± 4,9 | 84,2 ± 4,6 | 82,2 ± 5,1 |

Na Figura 5-14 apresentam-se as representações gráficas das taxas de acerto obtidas.

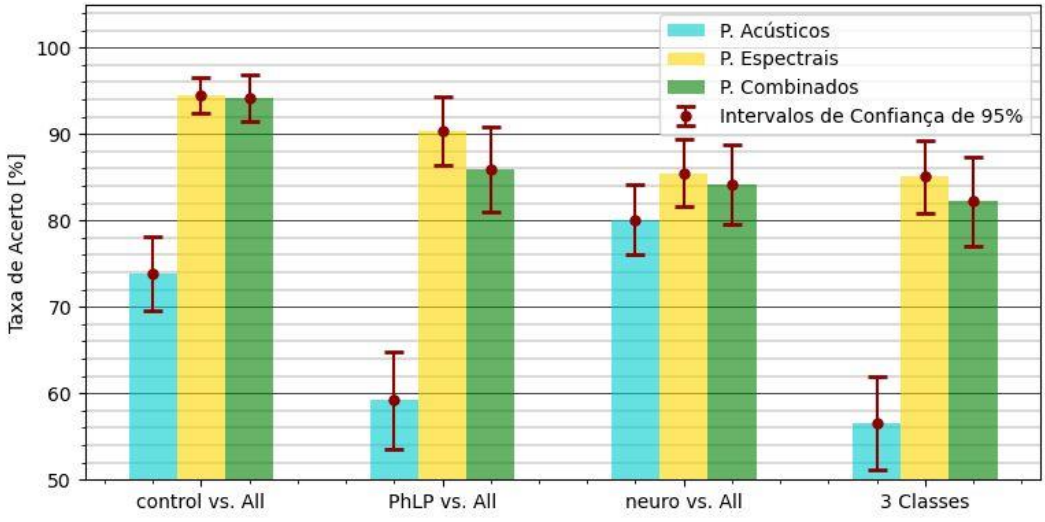


Figura 5-14 - Taxas de acerto na discriminação multiclasse, e nas discriminações OvA derivadas, no corpus USP

Observa-se que a introdução de parâmetros acústicos na discriminação entre três classes reduz a taxa de acertos tanto nesta discriminação, como nas três discriminações *OvA* derivadas desta. Esse efeito é mais ligeiro na discriminação de oradores saudáveis, com uma degradação de apenas 0,3%, e mais acentuada na discriminação de oradores com patologias laríngeas fisiológicas, com uma degradação de 4,5%. Embora o desempenho piore em todas as discriminações, a degradação é

sempre inferior à margem de confiança de 95%, não sendo considerada, por essa razão, estatisticamente significativa em nenhum dos casos. Observa-se também que a degradação é mais acentuada nas discriminações onde o desempenho dos parâmetros acústicos é pior, nas discriminações *PhLP* vs. *All* e entre as três classes, onde as taxas de acerto médias obtidas com parâmetros acústicos situaram-se na ordem dos 50%.

As matrizes de confusão obtidas nesta discriminação apresentam-se na Tabela 5-19.

Tabela 5-19 - Matrizes de confusão médias obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre três classes na base de dados *USP*, com valores normalizados, expressos em percentagem, entre parênteses

| | | **Control - pred.** | **PhLP - pred.** | **Neuro - pred.** |
|---|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 13,7 (91,59 %) | 0,1 (0,66 %) | 1,2 (7,75 %) |
| | **PhLP - real** | 0,1 (0,28 %) | 31,0 (96,87 %) | 0,9 (2,85 %) |
| | **Neuro - real** | 2,0 (14,36 %) | 4,8 (34,17 %) | 7,2 (51,46%) |

| | | | | |
|---|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 14,2 (94,85 %) | 0,6 (3,91 %) | 0,2 (1,24 %) |
| | **PhLP - real** | 0,7 (2,13 %) | 28,1 (87,98 %) | 3,2 (9,89 %) |
| | **Neuro - real** | 2,1 (14,75 %) | 4,2 (30,02 %) | 7,7 (55,23 %) |

As matrizes de confusão revelam que a introdução de parâmetros acústicos nesta discriminação traz algumas melhorias ao desempenho, ao contrário do que indicam as taxas de acerto médias. A identificação de oradores saudáveis melhora, tal como a identificação de oradores com patologias neurodegenerativas. No entanto, a identificação de oradores com patologias laríngeas fisiológicas piora, sendo essa degradação tão acentuada que impacta o desempenho global da discriminação.

### 5.3.3.9 Discriminação multiclasse (sMEEI)

A Tabela 5-20, apresentada de seguida, mostra os resultados obtidos para a discriminação entre três classes, bem como para as três discriminações *OvA*, na base de dados *sMEEI*.

Tabela 5-20 - Taxas de acerto, em percentagem, na discriminação multiclasse, e nas discriminações *OvA* derivadas, no corpus *sMEEI*

| | **Control vs. All** | **PhLP vs. All** | **UVFP vs. All** | **Multiclasse** |
|---|---|---|---|---|
| **Parâmetros Acústicos** | 88,3 ± 1,1 | 69,5 ± 2,2 | 76,0 ± 1,8 | 66,9 ± 2,0 |
| **Parâmetros Espectrais** | 88,8 ± 2,0 | 69,9 ± 4,6 | 68,8 ± 4,5 | 63,7 ± 4,4 |
| **Parâmetros Combinados** | **91,7 ± 1,6** | **77,8 ± 3,4** | **79,4 ± 3,2** | **74,5 ± 3,3** |

Podem ser visualizadas, na Figura 5-15, as taxas de acerto para estas discriminações.

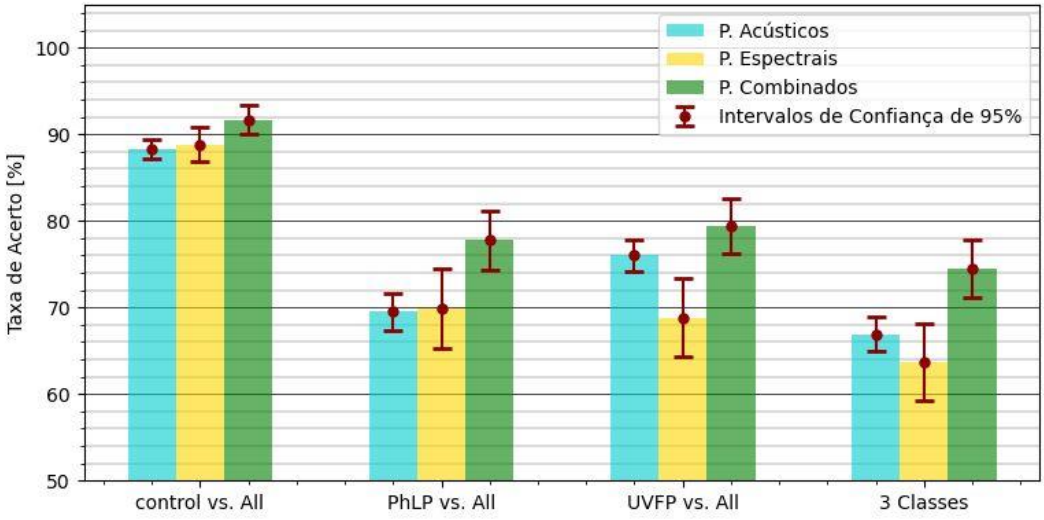


Figura 5-15 - Taxas de acerto na discriminação multiclasse, e nas discriminações *OvA* derivadas, no *corpus sMEEI*

No *corpus sMEEI,* observa-se uma melhoria generalizada na discriminação entre as três classes, assim como nas discriminações *OvA* derivadas desta. Em duas dessas discriminações, multiclasse e *UVFP vs. All*, a melhoria é superior às margens de confiança de 95%, tornando-a estatisticamente significativa. Estes resultados estão relacionados com o facto de que, nestas duas discriminações, as taxas de acerto obtidas com parâmetros acústicos são superiores às obtidas com espectrais.

De seguida, na Tabela 5-21, apresentam-se as matrizes de confusão obtidas neste *corpus*.

Tabela 5-21 - Matrizes de confusão média obtidas com parâmetros espectrais (cima), com parâmetros combinados (baixo) na discriminação entre três classes na base de dados *sMEEI*, com valores normalizados entre parênteses

| | | **Control - pred.** | **PhLP - pred.** | **UVFP - pred.** |
|---|---|---|---|---|
| **Parâmetros Espectrais** | **Control - real** | 28,4 (79,0 %) | 4,4 (12,1 %) | 3,2 (8,9 %) |
| | **PhLP - real** | 3,3 (5,9 %) | 35,4 (63,2 %) | 17,3 (30,9 %) |
| | **UVFP - real** | 6,1 (10,4 %) | 20,5 (34,7 %) | 32,4 (54,9 %) |

| | | | | |
|---|---|---|---|---|
| **Parâmetros Combinados** | **Control - real** | 32,0 (89,0 %) | 3,5 (9,6 %) | 0,5 (1,4 %) |
| | **PhLP - real** | 4,0 (7,2 %) | 39,7 (70,8 %) | 12,3 (22,0 %) |
| | **UVFP - real** | 4,5 (7,6 %) | 13,7 (23,2 %) | 40,8 (69,2 %) |

Observa-se que as melhorias obtidas com a inclusão de parâmetros acústicos são generalizadas, melhorando a identificação das três classes de forma bastante acentuada. Pode-se concluir que a combinação de parâmetros espectrais e acústicos é claramente vantajosa na discriminação entre as três classes presentes na base de dados *sMEEI*.

### 5.3.4 Compilação das taxas de acerto obtidas com parâmetros combinados

Apresentam-se de seguida as taxas de acerto obtidas em todas as discriminações analisadas, compiladas na Tabela 5-22.

Tabela 5-22 - Compilação de taxas de acerto, em percentagem, obtidas em todas as discriminações

| Discriminação | P. Acústicos | P. Espectrais | P. Combinados |
|---|---|---|---|
| **Control vs. Neuro (USP)** | 77,5 ± 7,1 | 92,5 ± 2,8 | **93,0 ± 1,3** |
| **Control vs. UFVP (sMEEI)** | 90,7 ± 1,1 | 85,3 ± 3,1 | **91,3 ± 1,9** |
| **Control vs. PhLP (USP)** | 68,1 ± 0,1 (*) | 96,0 ± 3,8 | **96,8 ± 2,8** |
| **Control vs. PhLP (sMEEI)** | 84,6 ± 1,9 | 85,7 ± 3,5 | **91,9 ± 2,5** |
| **Control vs. Patol. (USP)** | 75,4 ± 0,0 (*) | **94,5 ± 1,7** | 93,5 ± 2,7 |
| **Control vs. Patol. (sMEEI)** | 87,4 ± 1,1 | 89,8 ± 2,2 | **90,9 ± 1,8** |
| **PhLP vs. Neuro (USP)** | 76,8 ± 5,6 | **86,8 ± 6,3** | 82,7 ± 7,1 |
| **PhLP vs. UVFP (sMEEI)** | 70,4 ± 2,8 | 64,0 ± 5,0 | **72,7 ± 4,3** |
| **Multiclasse (USP)** | 56,5 ± 5,4 | **85,1 ± 4,2** | 82,2 ± 5,1 |
| **Control vs. All (USP)** | 73,8 ± 4,3 | **94,5 ± 2,0** | 94,2 ± 2,7 |
| **PhLP vs. All (USP)** | 59,2 ± 5,6 | **90,4 ± 4,0** | 85,9 ± 4,9 |
| **Neuro vs. All (USP)** | 80,1 ± 4,0 | **85,5 ± 3,9** | 84,2 ± 4,6 |
| **Multiclasse (sMEEI)** | 66,9 ± 2,0 | 63,7 ± 4,4 | **74,5 ± 3,3** |
| **Control vs. All (sMEEI)** | 88,3 ± 1,1 | 88,8 ± 2,0 | **91,7 ± 1,6** |
| **PhLP vs. All (sMEEI)** | 69,5 ± 2,2 | 69,9 ± 4,6 | **77,8 ± 3,4** |
| **UVFP vs. All (sMEEI)** | 76,0 ± 1,8 | 68,8 ± 4,5 | **79,4 ± 3,2** |

Resultados inválidos (*)

Analisando todas as taxas de acerto obtidas, verifica-se que a introdução de parâmetros acústicos melhorou o desempenho em 10 discriminações, e piorou em 6 delas. Nas 6 discriminações em que a taxa de acerto média diminuiu, a diferença não foi considerada significativa, pois foi menor que a soma das margens de confiança de 95%. Por outro lado, em 5 discriminações houve uma melhoria estatisticamente significativa, pois a diferença foi superior à soma das margens de confiança.

Nas 6 discriminações em que a inclusão de parâmetros acústicos degradou o desempenho, verifica-se que as taxas de acerto obtidas com parâmetros acústicos foram acentuadamente inferiores às obtidas com parâmetros espectrais. Já quando os desempenhos obtidos com parâmetros acústicos e espectrais foi semelhante, como aconteceu em quatro discriminações onde a diferença nas taxas de acerto foi inferior a 1%, observou-se sempre uma melhoria na taxa de acertos média quando se usam os parâmetros combinados comparativamente à obtida apenas com parâmetros espectrais.

A análise global dos resultados permite concluir que a combinação de parâmetros acústicos e espectrais é vantajosa para os desempenhos das discriminações consideradas neste trabalho. A combinação resultou num aumento da taxa de acerto média em mais casos (10), alguns deles com uma melhoria estatisticamente significativa, do que os casos em que resultou numa degradação (6). Quando envolvidos oradores com paralisia unilateral das pregas vocais, a combinação de parâmetros é claramente benéfica para o desempenho da discriminação, pois provocou sempre melhorias significativas na taxa de acerto média.

Relativamente às outras classes, é mais relevante analisar o desempenho dos parâmetros acústicos isoladamente do que focar a análise nas classes em si. Nos casos em que os parâmetros acústicos apresentaram um desempenho muito inferior ao obtido pelos espectrais, a combinação dos parâmetros piorou o desempenho na discriminação. Quando a utilização de parâmetros acústicos e espectrais isoladamente obteve taxas de acerto médias aproximadas, a combinação destes parâmetros mostrou ser benéfica para o desempenho das discriminações. Em resumo, o benefício da combinação de parâmetros espectrais e acústicos, bem como se o benefício será significativo, depende do desempenho obtido pelos parâmetros acústicos na discriminação em questão. Se os parâmetros acústicos não tiverem uma boa capacidade de discriminar as classes, a sua inclusão tende a degradar o desempenho nessa discriminação.

Outra conclusão que sai reforçada pelos resultados está relacionada com a combinação das classes patológicas numa única classe. Observa-se que as discriminações *control vs. All*, nas duas bases de dados, apresentaram taxas de acerto médias mais elevadas, de 94,2% e 91,7%, do que nas discriminações entre oradores saudáveis e patológicos, onde se agregaram as classes patológicas numa só, e onde se obtiveram taxas de acerto médias de 93,5% e 90,0%. Estes resultados corroboram o que tinha sido discutido na Secção 5.3.3.5, indicando não ser benéfica a agregação das classes patológicas numa só, sendo mais vantajoso efetuar a discriminação entre todas as classes, e posteriormente derivar os resultados para a discriminação entre oradores saudáveis e patológicos.

### 5.3.5 Comparação com resultados obtidos noutros estudos

Esta subsecção compara os resultados deste estudo com os de outros trabalhos que utilizaram as mesmas bases de dados. Conforme descrito nas Secções 2.3.1.1 e 3.1.1, os estudos [47] e [57] foram realizados utilizando a base de dados *USP*, também usada neste trabalho. A métrica comparável entre os estudos é a taxa de acerto na discriminação entre oradores saudáveis e patológicos. Esses estudos atingiram uma taxa de acerto de 100%, o que já serve como referência, pois qualquer resultado em percentagem, como os apresentados aqui, é implicitamente comparado com o valor máximo.

Os estudos também usaram um subconjunto da base de dados *MEEI* semelhante ao deste trabalho, mas contendo alguns sinais de fala de oradores com patologias que não estão presentes neste trabalho. Assim, optou-se por não se comparar esses resultados.

Conforme descrito nas Secções 2.3.1.1 e 3.1.2, um subconjunto da *MEEI* foi utilizado nos trabalhos [44], [50] e [51]. Esse subconjunto difere apenas em três sinais de fala, que foram descartados

aqui pois não foi possível extrair alguns parâmetros acústicos. Apesar dessa diferença, optou-se por comparar os resultados de modo a validar os métodos usados neste trabalho.

Os estudos [50] e [51] utilizaram sinais de fala contínua, e com a vogal /a/ sustentada. Para garantir uma comparação mais precisa, foram considerados apenas os resultados obtidos com a vogal /a/ sustentada, pois dessa forma os resultados referem-se a sinais de fala com o mesmo conteúdo. Por essa razão, os resultados do estudo [44], que utiliza exclusivamente fala contínua, não foram considerados.

Os estudos [50] e [51] fornecem resultados para as métricas *TPR* e *PPV*, descritas na Secção 3.3.1. Assim, essas métricas foram calculadas a partir das matrizes de confusão deste estudo, com recurso às equações 3.4 e 3.5, também abordadas na Secção 3.3.1. Os estudos apresentam também a taxa de acerto (*ACC*) na discriminação entre as três classes. Os trabalhos comparados empregam várias combinações de parâmetros e classificadores.

Para a comparação, foi utilizado o melhor resultado de cada métrica. No entanto, essa poderá não ser a forma mais adequada de apresentar as métricas *PPV* e *TPR*, pois estas estão relacionadas, e quando uma apresenta um valor elevado, a outra geralmente tende a ser mais baixa. Assim, para simplificar a análise e evitar a discussão sobre o significado de 'melhor' valor para essas métricas, optou-se por utilizar os valores mais elevados de cada uma delas, independentemente de pertencerem, ou não, ao mesmo teste. Isso será suficiente para validar dos resultados obtidos neste estudo.

Os resultados apresentam-se de seguida, na Tabela 5-23.

Tabela 5-23 - Comparação de resultados, em percentagem, deste trabalho e dos estudos [50] e [51]

| Métrica | P. Espectrais | P. Combinados | Resultados [50] | Resultados [51] |
|---|---|---|---|---|
| PPV Classe control | 75,2 ± 4,8 | 79,0 ± 2,8 | **80,8** | **80,8** |
| TPR Classe control | 79,0 ± 5,6 | 89,0 ± 5,5 | **91,6** | 58,3 |
| PPV Classe PhLP | 58,8 ± 6,0 | **69,8 ± 5,1** | 69,2 | 64,1 |
| TPR Classe PhLP | 63,2 ± 8,1 | **70,8 ± 6,0** | 69,5 | 69,5 |
| PPV Classe UVFP | 61,3 ± 6,7 | 76,1 ± 5,3 | **76,5** | 67,2 |
| TPR Classe UVFP | 54,9 ± 7,6 | 69,1 ± 4,8 | **78,0** | **78,0** |
| ACC Multiclasse | 63,7 ± 4,4 | **74,5 ± 3,3** | 70,1 | 68,2 |

Comparando as sete métricas analisadas, observa-se que este trabalho obteve valores mais elevados em três delas relativamente ao trabalho [50], e em cinco dessas métricas relativamente ao estudo [51]. Em relação ao estudo [50], cinco das diferenças de valores são inferiores à margem de confiança de 95%, ou seja, não são estatisticamente significativas. Já em relação ao estudo [51] apenas duas métricas não apresentaram diferenças significativas.

Verifica-se que este trabalho obteve uma taxa de acerto média mais elevada na discriminação entre as três classes, com uma diferença superior à margem de confiança de 95%, sugerindo um melhor desempenho na discriminação com os parâmetros utilizados neste trabalho.

A única métrica em que este estudo obteve um valor inferior, com uma diferença superior à margem de confiança, foi no *TPR* para a classe *UVFP*, em comparação com ambos os estudos [50] e [51] (muito provavelmente ambos os estudos realizaram o mesmo teste). Essa diferença ocorreu quando os estudos utilizaram um classificador *LDA* na discriminação entre as três classes. A Tabela 5-24 apresenta a comparação dos valores das métricas *TPR* e *PPV* obtidas nesse caso, bem como os valores da métrica *F1-Score* calculadas a partir destas, para uma melhor análise comparativa.

Tabela 5-24 - Comparação parcial de resultados, em percentagem, deste trabalho e dos estudos [50] e [51]

| | TPR UVFP [%] | PPV UVFP [%] | F1-Score UVFP [%] |
|---|---|---|---|
| **Estudos [50] e [51]** | 78,0 | 65,7 | 71,3 |
| **Este estudo** | 69,1 | 76,1 | **72,4** |

Observa-se que a diferença no *TPR* não se deve a um desempenho inferior deste trabalho, mas sim a uma diferença no limiar de decisão entre as classes. Nos trabalhos comparados, o limiar de decisão privilegia o *TPR*, dando mais enfase à identificação de todos os oradores com esta patologia, enquanto neste trabalho, o limiar de decisão privilegia o *PPV*, dando mais enfase à precisão da identificação e minimização dos falsos positivos. A métrica *F1-Score*, usada aqui para avaliar as métricas *PPV* e *TPR* em conjunto, obteve um valor superior neste trabalho, contrariando a ideia de que este trabalho tenha tido um desempenho inferior relativamente ao teste com *LDA*.

A análise dos resultados confirma a relevância dos valores obtidos, embora essa conclusão se limite à combinação dos parâmetros espectrais desenvolvidos neste estudo com parâmetros acústicos na discriminação entre três classes efetuada na base de dados *sMEEI*, pois foram os resultados comparados com os obtidos nos estudos [50] e [51].

Comparando os resultados obtidos apenas com parâmetros espectrais aos dos estudos [50] e [51], observa-se que os valores deste estudo foram consistentemente inferiores, muitas vezes com diferenças estatisticamente significativas. Tendo em conta que os estudos comparados utilizaram parâmetros espectrais, estes resultados sugerem que a utilização de parâmetros espectrais diferentes, com maior poder discriminatório, em combinação com parâmetros acústicos, poderá melhorar os resultados nas discriminações realizadas relativamente aos obtidos nestes estudos. No entanto, é possível que tais parâmetros espectrais possam ter um desempenho bastante superior aos parâmetros acústicos, podendo a combinação deles reduzir o desempenho, como aconteceu em alguns casos na base de dados *USP.* Certamente será uma hipótese que merece ser investigada, em trabalhos futuros.

# 6

# CONCLUSÕES

Nesta secção apresentam-se as conclusões obtidas e o conhecimento adquirido através da realização deste trabalho. Estas fornecem indicações importantes e sugerem linhas de investigação a ser retomadas, espera-se, em trabalhos futuros.

Neste trabalho, foram utilizadas duas bases de dados, cada uma com amostras de oradores saudáveis e com três patologias da fala, sendo duas delas comuns a ambas. Isso permitiu um maior número de amostras, totalizando 212 oradores, e possibilitou o estudo de quatro patologias ao invés das três existentes em cada *corpus*. As patologias comuns, edema de Reinke e nódulos vocais, foram agrupadas numa única classe chamada patologias laríngeas fisiológicas (*PhLP*), devido à sua semelhança na alteração da fisiologia das pregas vocais, conforme descrito na Secção 2.2.2. Na Secção 4.5.4, estas patologias foram estudadas separadamente, obtendo-se uma taxa de acerto média de 57,9% na discriminação entre elas no *corpus USP*, enquanto no *corpus sMEEI*, não foi possível obter uma taxa de acerto média válida, devido ao desbalanceamento entre as duas classes, demonstrando que a utilização de bandas de energia não é adequada para a discriminação entre estas duas patologias.

A presença de três classes em comum entre as bases de dados permitiu uma comparação direta de alguns resultados, identificando potenciais vieses específicos e funcionando como validação externa simultânea. No entanto, o uso de duas bases de dados trouxe desvantagens, como diferenças desconhecidas nos sinais de fala devido a condições de aquisição diferentes, tais como possíveis variações nos níveis de reverberação, ruído de fundo e características do equipamento utilizado. Além disso, as diferenças na distribuição de género e idade dos oradores, bem como no grau de severidade das patologias, podem ter tido influência nos resultados, e por essa razão introduzem alguma incerteza nas conclusões. Apesar disso, optou-se por apresentar estas comparações e conclusões, reconhecendo essas limitações.

## 6.1 Bandas de energia

Antes de serem apresentados resultados, já algumas conclusões puderam ser formuladas a partir de testes preliminares. Conforme discutido na Secção 4.2, o banco de filtros e as bandas de

energia definidas numa escala perceptual (*Mel*) obtiveram melhores resultados, nesses testes, do que as bandas linearmente definidas. Estes resultados demonstram a importância de incluir informação perceptual nos parâmetros espectrais para discriminar entre vozes saudáveis e patológicas, bem como para identificar patologias. A relevância da informação perceptual, já evidenciada em trabalhos anteriores, conforme discutido na Secção 2.3.1.3., foi confirmada pelos resultados neste trabalho.

O modelo de classificação utilizado também fornece informações valiosas sobre os dados. Entre os modelos testados, o *SVM* com *kernel RBF* apresentou os melhores resultados, especialmente quando comparados com os obtidos com o *kernel* linear, sugerindo que as patologias afetam de maneira complexa as bandas de energia e de variação de energia dos sinais de fala. Se as patologias tivessem um efeito linear, aumentando ou diminuindo valores de bandas específicas conforme a severidade, um *kernel* linear ter-se-ia revelado o mais adequado para as discriminações testadas. No entanto, os resultados sugerem que os oradores saudáveis, quando representados pelas bandas, estão agrupados num *cluster*, e quando desenvolvem patologias, afastam-se desse *cluster*.

Esse afastamento, porém, não ocorre sempre na mesma direção, ou seja, através da alteração consistente dos valores de determinadas dimensões. Embora os resultados tenham mostrado que existem direções, ou bandas, onde as patologias exercem maior impacto, essas direções, bem como os impactos que sofrem variam, não seguindo um padrão fixo que poderia ser capturado por um limiar linear. Os melhores resultados obtidos com *SVM* com *kernel RBF*, bem como o padrão de maior dispersão dos oradores patológicos em relação aos saudáveis conforme observado na Secção 4.5.2 confirmam essa interpretação. Além disso, os bons resultados, embora inferiores aos do *SVM*, obtidos com o classificador *Nearest Centroid*, também sugerem que uma abordagem baseada em *clusters* pode ser eficaz.

As bandas de energia dos sinais de fala foram utilizadas para calcular uma variante do parâmetro *LBST*, chamada *bbLBST*, e um parâmetro equivalente, *bbLBSvT*, foi calculado a partir das bandas de variação de energia. Estes parâmetros foram testados para discriminar entre as diferentes classes. As taxas de acerto médias nas discriminações binárias variaram entre 60,4%, na discriminação entre a classe *PhLP* e os oradores com paralisia unilateral das pregas vocais (*UVFP*), e 96,0%, na discriminação entre a classe dos oradores saudáveis e a *PhLP* no *corpus USP*, com seis das oito discriminações a atingir taxas de acerto médias acima dos 80%.

Estes resultados corroboram estudos anteriores, indicando que a relação entre a diferença dos picos locais nas zonas de baixas e médias frequências do espectro, e a diferença na frequência em que ocorrem, é um parâmetro eficaz para discriminar as classes consideradas neste estudo. Os resultados também demonstram que essa relação, quando calculada a partir da variação de energia ao longo do sinal, é também útil para essas discriminações. A distribuição das amostras em função desses parâmetros, observada nas Figuras 4-8 e 4-9, mostra que as amostras de oradores saudáveis tendem a ter valores positivos, ou quando negativos, muito baixos. Isso sugere que estes parâmetros podem ser úteis num sistema de rastreio de patologias, pois valores negativos destes parâmetros identificam oradores patológicos, sem falsos positivos.

As bandas de energia e de variação de energia mais relevantes para cada discriminação foram identificadas com margens de confiança bastante amplas, entre 10% e 25%, quando as bandas mais relevantes foram determinadas, por vezes, por diferenças inferiores a 1%. Assim, são mais relevantes os resultados relativos à frequência com que determinadas bandas são identificadas, do que a identificação da banda mais relevante em cada discriminação. Apesar dessas limitações, considera-se que algumas dessas taxas de acerto merecem ser destacadas.

Na discriminação entre oradores saudáveis e com patologias neurodegenerativas, foi obtida uma taxa de acerto de 100% com a banda de energia mais relevante, a banda 1, que cobre a faixa entre 5 e 102 Hz. No entanto, alguns factores podem questionar a fiabilidade desse resultado. A faixa de frequências utilizada pode não conter a frequência fundamental em alguns sinais, o que sugere que a discriminação pode estar mais relacionada com o ruído de fundo do que com o conteúdo da fala. Além disso, essa faixa pode incluir ruído de 50 ou 60 Hz, gerado pela rede elétrica, especialmente se os sinais foram captados em condições diferentes para os oradores saudáveis e patológicos. Outro ponto a considerar é que esta discriminação foi feita com o menor número de amostras, apenas 29 no total entre as duas classes, o que limita a confiança nos resultados. No entanto, se confirmado em estudos futuros com outras bases de dados, esse resultado pode ser promissor para distinguir entre oradores saudáveis e com patologias neurodegenerativas.

Um outro resultado que merece destaque é a taxa de acerto média de 57,2% na discriminação entre as classes *PhLP* e *UVFP*, que foi o pior desempenho obtido, com uma diferença para a segunda taxa mais baixa superior a 13%. A Figura 4-5 mostrou que os valores das bandas nos sinais de fala para essas duas patologias eram quase idênticos. A partir destes resultados conclui-se que estas patologias, embora de naturezas diferentes, pois uma afeta a fisiologia das pregas vocais e a outra afeta o seu controlo, influenciam o espectro dos sinais de fala de forma semelhante. Assim, as bandas de energia e de variação de energia não são adequadas para a discriminação entre estas patologias.

No geral, as taxas de acerto médias para determinar as bandas mais relevantes para as discriminações foram significativamente mais elevadas no *corpus USP*. Um dos factores que contribui para esta diferença é a presença da classe *UVFP* que, como mostrado, é pouco distinguível da classe *PhLP*, baixando por essa razão as taxas de acerto no *corpus sMEEI*. Outras potenciais razões foram já referidas no início desta secção, destacando-se a distribuição de oradores por género. No *corpus USP*, existem mais oradores masculinos que femininos na classe dos oradores saudáveis, ao contrário do que acontece nas restantes classes. O género do orador afeta as bandas, pois uma faixa que corresponde à frequência fundamental num orador feminino, pode corresponder a uma harmónica num orador masculino, influenciando a discriminação.

Um dos objetivos deste trabalho consiste na investigação da contribuição, caso exista, das bandas de energia na diferenciação entre vozes saudáveis e patológicas e na identificação de patologias laríngeas. Os resultados demonstram que essa contribuição existe, pois, as discriminações foram efetuadas com base nas bandas, com maior ou menor eficácia. Apenas a discriminação entre as classes *PhLP* e *UVFP* obteve uma taxa de acerto média próxima dos 50%, que seria o resultado esperado para uma classificação aleatória.

As bandas de energia identificadas como tendo maior poder discriminatório foram as bandas 1, que abrange aproximadamente as frequências entre os 6 e os 102 Hz, relativas à frequência fundamental, as bandas 3 e 4, que compreendem a faixa entre os 110 e o 371 Hz, relativa às primeiras harmónicas, e a banda 8, que abrange aproximadamente a faixa entre os 604 e os 867 Hz, relativa à frequência do primeiro formante da vogal /a/. Outras bandas, relacionadas com a frequência fundamental e primeiras harmónicas (banda 2), bem como com o primeiro formante da vogal /a/ (bandas 9 e 10) também se revelaram relevantes. Estes resultados indicam que estas patologias afetam o espectro da vogal /a/, especialmente nas baixas frequências e, consequentemente, oradores saudáveis e patológicos podem ser discriminados com base nas bandas de frequência identificadas.

Na discriminação entre patologias, a banda 8 destacou-se como a mais relevante, sugerindo que as patologias influenciam a pronúncia da vogal /a/ de forma diferente. Esta informação pode ser útil para a implementação de um classificador hierárquico, que faça uma discriminação entre oradores saudáveis e patológicos, seguida da identificação da patologia.

Entre as bandas de variação de energia, a banda 3, que corresponde à faixa entre 110 e 275 Hz, associada às primeiras harmónicas, destacou-se como sendo a mais relevante, identificada como tal em sete das dez discriminações testadas. Este resultado reflete a instabilidade vocal presente em oradores patológicos, manifestada pela dificuldade em manter a estabilidade da frequência fundamental e, consequentemente, das suas harmónicas. Essa dificuldade resulta numa variação significativa do espectro nesta faixa de frequências ao longo do sinal de fala.

## 6.2 Combinação com parâmetros acústicos

Foram testados dois parâmetros espectrais diferentes, obtidos a partir das bandas de baixas e médias frequências, antes de serem combinados com parâmetros acústicos: os parâmetros bbLBST e bbLBSvT, que contêm informação da banda com maior amplitude em cada zona de frequência, e um parâmetro obtido por *PCA* a partir das bandas, contendo informação de todas elas. De um modo geral, o segundo parâmetro apresentou um melhor desempenho nas discriminações, o que indica que existe informação útil em várias bandas nas baixas e médias frequências, e que é importante que essa informação esteja presente nos parâmetros espectrais utilizados.

Esta conclusão foi corroborada pelos resultados obtidos pelos classificadores *Random Forest* e *XGB* em testes preliminares, que mostraram um bom desempenho quando todas as bandas foram utilizadas na classificação dos oradores. Estes classificadores têm a vantagem conhecida de serem adequados para dados com elevada dimensionalidade, e o bom desempenho observado indica que, ao reduzir a dimensionalidade dos dados antes da combinação com os parâmetros acústicos, alguma informação importante foi perdida. Embora, neste trabalho, estes classificadores não se tenham revelado ideais, a sua utilização poderá ser adequada em trabalhos futuros, com bases de dados onde o rácio entre amostras e dimensões seja suficiente para evitar a necessidade de redução da dimensionalidade.

A análise dos valores dos parâmetros *shimmer*, *jitter* e *HNR* nas classes dos oradores saudáveis e *PhLP* mostrou uma maior sobreposição de valores no *corpus USP*. Esta diferença sugere que os oradores patológicos nessa base de dados estão num estado mais inicial da patologia, pois é improvável que a progressão da patologia afete os valores destes parâmetros, piorando-os, para depois voltar a melhorá-los. Esta conclusão é importante para entender as diferenças nos resultados entre as duas bases de dados. Também foi observado que, testando as discriminações com estes parâmetros isoladamente, o *shimmer* apresentou um desempenho globalmente melhor no *corpus USP*, enquanto o *HNR* teve um desempenho melhor no *corpus sMEEI*, com indivíduos num estado mais avançado da patologia. Estes resultados levaram à opção de que apenas estes dois parâmetros seriam combinados com os espectrais.

Os resultados obtidos no corpus *USP* apresentam algumas contradições. Nas discriminações entre oradores saudáveis e com cada uma das patologias, a combinação de parâmetros espectrais e acústicos melhorou as taxas de acerto médias. No entanto, na discriminação entre as duas patologias, e entre as três classes, as taxas de acerto médias diminuíram. Estas diferenças não foram estatisticamente significativas, não se podendo concluir, apenas com base nestes resultados, se a combinação dos parâmetros espectrais e acústicos é benéfica ou prejudicial para as discriminações consideradas.

Na base de dados *sMEEI*, a inclusão dos parâmetros acústicos melhorou as taxas de acerto médias em todas as discriminações, variando essas melhorias entre 6,0%, na discriminação entre oradores saudáveis e com paralisia unilateral das pregas vocais, e 10,8% na discriminação entre as três classes. Todas estas melhorias são estatisticamente significativas, permitindo concluir que a combinação de parâmetros espectrais e acústicos é benéfica para as discriminações consideradas nesta base de dados. Esses resultados evidenciam a complementaridade destes parâmetros e a vantagem de incluir ambos em sistemas automáticos de discriminação entre oradores saudáveis, com patologias laríngeas fisiológicas e com paralisia unilateral das pregas vocais.

Os resultados observados na discriminação entre oradores saudáveis e com patologias laríngeas fisiológicas, considerando que o *corpus sMEEI* contém oradores patológicos em estados mais avançados das patologias, permitem uma análise um pouco mais aprofundada. Os resultados obtidos no *corpus USP*, onde se observam taxas de acerto médias mais elevadas com a utilização de parâmetros espectrais, sugere que estas patologias, no estado inicial, têm um maior impacto no espectro do sinal, tanto na distribuição de energia pelas baixas e médias frequências, como pela variação da mesma ao longo do tempo. Conclui-se também que, à medida que essas patologias avançam, esse impacto nos parâmetros espectrais diminui enquanto o efeito nos parâmetros acústicos aumenta. Dessa forma, a combinação dos parâmetros espectrais e acústicos torna-se mais vantajosa para discriminar entre oradores saudáveis e patológicos à medida que as patologias progridem.

## 6.3 Divulgação

"*Multi pertransibunt et augebitur scientia.*" Esta citação do *Livro de Daniel*, que se traduz como "muitos passarão e o conhecimento será aumentado", foi escolhida por Francis Bacon como epígrafe de sua ambiciosa obra *Novum Organum*, publicada em 1620. Bacon considerou essa citação apropriada para refletir o tema central do livro, que se propunha reformar e ampliar o conhecimento humano. Hoje, mais de 400 anos depois, num tempo em que os estudos e trabalhos de investigação são mais numerosos do que nunca, essa citação faz ainda mais sentido, pois tão importante como os trabalhos de investigação é a divulgação dos resultados e conclusões obtidas.

Divulgar os resultados de teses e dissertações através de artigos científicos é crucial para o avanço do conhecimento. Esses artigos permitem que a comunidade científica tenha acesso a novos dados, métodos e conclusões, permitindo a troca de ideias e a colaboração entre investigadores. A publicação em revistas científicas e apresentação em conferências, após revisão por pares, também garante a credibilidade e a qualidade do trabalho, validando os resultados e aumentando a sua visibilidade. Além disso, a disseminação de resultados contribui para o desenvolvimento de soluções práticas e inovadoras, abrindo caminho para futuras investigações no campo estudado.

Reconhecendo a importância da divulgação de resultados e da disseminação de conhecimento, foram escritos dois artigos científicos ao longo deste trabalho para refletir o trabalho realizado. Esses artigos documentam as etapas e resultados alcançados até a data da sua escrita, descrevendo de forma fiel a linha condutora do trabalho. Embora o estudo tenha avançado além do que está descrito nos artigos, eles asseguram que os avanços obtidos sejam partilhados com a comunidade científica.

A Tabela 6-1 apresenta informações detalhadas sobre esses artigos.

Tabela 6-1 - Detalhes acerca de artigos científicos escritos ao longo do trabalho

| | |
|---|---|
| **Título** | **Bandas espectrais de energia para discriminação de patologias laríngeas em sinais de fala: Discriminação entre vozes saudáveis e não saudáveis, e entre patologias** |
| **Autores** | Bruno Rodrigues, Hugo Cordeiro, Gonçalo Marques |
| **Conferência** | CISTI 2023 - 18th Iberian Conference on Information Systems and Technologies |
| **Local e Data** | Aveiro – 20 a 24 de junho 2023 |
| **Título** | **Discriminating Voice Pathologies Through a Combination of Spectral and Acoustic Features** |
| **Autores** | Bruno Rodrigues, Hugo Cordeiro, Gonçalo Marques |
| **Conferência** | HCist 2024 - Health and Social Care Information Systems and Technologies |
| **Local e Data** | Funchal – 13 a 15 de novembro 2024 |

O primeiro artigo, *Bandas espectrais* (…) documenta o trabalho relacionado com o Capítulo 4 deste estudo, realizado até à data da sua escrita. O artigo foca-se na determinação de quais as bandas

de energia e de variação de energia que permitem uma discriminação entre vozes saudáveis e patológicas, e entre diferentes patologias. Esta análise foi efetuada apenas na base de dados *USP*, e foi realizada com uma única iteração sobre todos os dados. Por essa razão existem discrepâncias nos resultados entre o artigo e este trabalho, no entanto, os resultados apontam a conclusões semelhantes em ambos os estudos.

O segundo artigo científico, *Discriminating Voice Pathologies (…)*, foca-se na combinação de parâmetros espectrais e acústicos para discriminação entre vozes saudáveis e patológicas, e entre patologias, abordada no Capítulo 5 deste estudo. Os resultados apresentados no artigo referem-se apenas à base de dados *sMEEI*, e apresentam também algumas discrepâncias em comparação com os resultados obtidos neste trabalho, devido à ausência de normalização da duração dos sinais de fala, e à utilização do *jitter* no artigo, ao contrário do realizado neste trabalho. Apesar dessas diferenças nos resultados, as conclusões de ambos os estudos foram coincidentes.

## 6.4 Trabalho futuro

Alguns resultados obtidos neste estudo, mais do que esclarecer, levantaram novas questões que este trabalho acabou por não abordar. Potenciais investigações a essas questões traduzir-se-iam, na maioria dos casos, em desvios consideráveis ao trabalho, que tentou manter o foco numa linha coerente. No entanto, dado que essas questões surgiram, acredita-se que merecem ser exploradas em estudos futuros.

As bandas de energia nas frequências mais baixas podem requerer uma outra definição, ajustada à frequência fundamental. Observou-se que estas bandas são as mais importantes nas discriminações analisadas e, por essa razão, precisam de uma maior precisão na sua definição e clareza no seu significado. A banda 3, que mostrou ter uma especial importância, abrange aproximadamente a faixa de frequências entre 110 e 275 Hz. Esta faixa de frequências pode corresponder, em oradores de género feminino, à frequência fundamental, enquanto pode corresponder às segunda e terceira harmónicas em oradores masculinos. Portanto, propõe-se, para trabalhos futuros, a definição das bandas abaixo do primeiro formante em função da frequência fundamental.

Propõe-se também um estudo futuro das bandas de energia em função do género do orador. A base de dados *USP* não é adequada para este estudo, pois contém apenas 21 oradores masculinos, dos quais 10 são saudáveis e, por outro lado, os oradores de género feminino são 40, dos quais apenas 5 são saudáveis. Porém, a base de dados *MEEI* pode ser mais apropriada pois contém um número maior de oradores de ambos os géneros, com distribuições por classe mais homogéneas. Alguns testes realizados neste trabalho tendo em conta o género dos oradores, mostraram resultados interessantes, mas pouco conclusivos. Conforme mencionado anteriormente, as bandas de energia nas baixas frequências podem ter significados diferentes para oradores masculinos e femininos. Portanto, para uma investigação futura que tenha em conta o género dos oradores, será essencial definir as bandas de energia nas baixas frequências em função da frequência fundamental.

Durante a determinação das bandas de energia e de variação de energia mais relevantes, observou-se que, nas discriminações entre as classes *PhLP* e *UVFP*, e entre as patologias edema de Reinke e nódulos vocais, com taxas de acerto excecionalmente baixas, bandas acima da banda 10 surgiram entre as mais relevantes. Estes resultados sugerem que bandas situadas em frequências mais elevadas podem ter um potencial discriminatório que, embora menor que o das bandas mais baixas, poderá não ser desprezável. Este estudo focou-se desde o início, influenciado pelos resultados dos trabalhos [44] e [45], na análise da informação espectral nas baixas frequências, razão pela qual as bandas acima da banda 10 não foram investigadas de forma mais aprofundada, apesar dos resultados obtidos nas duas discriminações mencionadas. No entanto, estes resultados levantam a questão do potencial das bandas mais altas, que se considera merecer ser investigada em estudos futuros, com uma potencial utilização de um filtro de pré-enfase para mitigar a atenuação nas frequências mais elevadas.

Quando comparadas, na Secção 5.3.5, algumas métricas obtidas neste estudo com as obtidas nos estudos [50] e [51], constatou-se que os desempenhos obtidos com parâmetros combinados estavam em linha com os dos estudos comparados. No entanto, os desempenhos obtidos apenas com parâmetros espectrais foram inferiores aos dos referidos estudos, que também utilizaram apenas parâmetros espectrais. Estes resultados sugerem que a combinação de parâmetros espectrais com maior poder discriminatório entre as classes e parâmetros acústicos pode melhorar os desempenhos obtidos neste estudo. Além disso, os estudos comparados indicaram que a extração de parâmetros espectrais a partir de fala contínua pode ser mais eficaz que o uso da vogal /a/ sustentada, como foi feito neste trabalho. Propõe-se para um potencial trabalho futuro a combinação de parâmetros espectrais extraídos de fala contínua com parâmetros acústicos, a fim de investigar o seu potencial na discriminação entre oradores saudáveis e patológicos, e entre diferentes patologias.

Os resultados indicam que as bandas mais relevantes para uma discriminação específica não são as mesmas para todas as discriminações. Observou-se também que os parâmetros acústicos podem ter maior poder discriminatório nas identificações que envolvem a paralisia unilateral das pregas vocais, em comparação com os parâmetros espectrais. Portanto, propõe-se também para trabalho futuro o desenvolvimento de um modelo hierárquico que utilize e combine diferentes parâmetros de forma mais adequada consoante a discriminação a efetuar. Esse modelo pode ser utilizado numa potencial aplicação de rastreio de patologias baseada em sinais de fala, cuja implementação seria o objetivo derradeiro deste e de outros estudos relacionados, e cujas vantagens foram discutidas na Secção 1.1.3.

# BIBLIOGRAFIA


[1] M. Pagel, "Q&A: What is human language, when did it evolve and why should we care?," BMC Biol, vol. 15, no. 1, p. 64, 2017.

[2] W. T. Fitch, "The Biology and Evolution of Speech: A Comparative Analysis," Annual Review of Linguistics, vol. 4, no. 1, pp. 255-279, 2018.

[3] M. Ephratt, "Linguistic, paralinguistic and extralinguistic speech and silence," Journal of Pragmatics, vol. 43, no. 9, pp. 2286-2307, 2011.

[4] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, & R. Zäske, "Speaker perception," Wiley Interdiscip Rev Cogn Sci, vol. 5, no. 1, pp. 15-25, jan. 2014.

[5] J. Kreiman, D. Van Lancker Sidtis, & B. Gerratt, "Defining and measuring voice quality," in Voqual'03, 2003.

[6] J. Kreiman, Y. Lee, M. Garellek, R. Samlan, & B. R. Gerratt, "Validating a psychoacoustic model of voice quality," in J. Acoust. Soc. Am., vol. 149, no. 1, pp. 457-465, jan. 2021,

[7] J. Kreiman, "Information conveyed by voice quality," J. Acoust. Soc. Am., vol. 155, no. 2, pp. 1264-1271, fev. 2024.

[8] M. Behlau, G. Madazio, D. Feijó, & P. Pontes, "Avaliação de voz," in Voz: o livro do especialista, vol. 1, pp. 85-245, 2001.

[9] L. Lee, J. Stemple, L. Glaze, & L. Kelchner, "Quick Screen for Voice and Supplementary Documents for Identifying Pediatric Voice Disorders," *Language, Speech, and Hearing Services in Schools*, vol. 35, pp. 308-319, 2004.

[10] M. Dietrich, K. Verdolini Abbott, J. Gartner-Schmidt, & C. A. Rosen, "The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice," J. Voice, vol. 22, no. 4, pp. 472-488, jul. 2008.

[11] S. Misono, C. B. Peterson, L. Meredith, K. Banks, D. Bandyopadhyay, B. Yueh, & P. A. Frazier, "Psychosocial distress in patients presenting with voice concerns," J. Voice, vol. 28, no. 6, pp. 753-761, 2014.

[12] S. H. Chen, S. C. Chiang, Y. M. Chung, L. C. Hsiao, & T. Y. Hsiao, "Risk factors and effects of voice problems for teachers," J. Voice, vol. 24, no. 2, pp. 183-190, mar. 2010.

[13] C. Munier, R. Kinsella, "The prevalence and impact of voice problems in primary school teachers," Occupational Medicine, vol. 58, no. 1, pp. 74-76, 2007.

[14] K. Omori, "Diagnosis of voice disorders," Japan Medical Association Journal, vol. 54, no. 4, pp. 248-253, 2011.

[15] N. Spantideas, E. Drosou, A. Karatsis, & D. Assimakopoulos, "Voice Disorders in the General Greek Population and in Patients With Laryngopharyngeal Reflux. Prevalence and Risk Factors," Journal of Voice, vol. 29, no. 3, pp. 389.e27-389.e32, 2015.

[16] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," The Laryngoscope, vol. 124, no. 10, pp. 2359-2362, 2014.

[17] H. J. Hoffman, C.-M. Li, K. Losonczy, M. S. Chiu, J. B. Lucas, and K. O. St. Louis., "Voice, speech, and language disorders in the U.S. population: The 2012 National Health Interview Survey (NHIS)," in Abstracts of the 47th Annual Meeting of the Society for Epidemiologic Research, p. 156, Abstract No. 648, Seattle, WA, Jun. 24-27, 2014.

[18] S. P. P. Giannini, M. do R. D. de O. Latorre, & L. P. Ferreira, "Distúrbio de voz: definição de caso em estudos epidemiológicos", Distúrb Comun, vol. 28, nº 4, jan. 2017.

[19] A. E. Morris, S. A. Norris, J. S. Perlmutter, & J. W. Mink, "Quantitative, clinically relevant acoustic measurements of focal embouchure dystonia," Movement Disorders, vol. 33, no. 3, pp. 449–458, fev. 2018.

[20] I. Guimarães, "A Ciência e a Arte da Voz Humana", Escola Superior de Saúde de Alcoitão, Alcabideche, pp. 1-124. 2007.

[21] F. M. Ramos, T. Órfão, J. Laranjeiro, M. G. Ribeiro & M. Santos, "Sensibilidade e especificidade da laringoscopia indirecta e nasofibroscopia laríngea na detecção de lesões malignas e pré-malignas da laringe," Revista Portuguesa De Otorrinolaringologia E Cirurgia De Cabeça E Pescoço, vol. 50, no. 3, set. 2012.

[22] M. Eça & C. Fernández-Sesma, "Eleições Legislativas 2024: Análise da Conversação Social", Llorente y Cuenca (LLYC), mar. 2024.

[23] G. Woodson, "Management of neurologic disorders of the larynx," The Annals of Otology, Rhinology, and Laryngology, vol. 117, no. 5, pp. 317–326, 2008.

[24] A. L. Merati, Y. D. Heman-Ackah, M. Abaza, K. W. Altman, L. Sulica, & S. Belamowicz, "Common movement disorders affecting the larynx: a report from the neurolaryngology committee of the AAO-HNS," Otolaryngol Head Neck Surg, vol. 133, no. 5, pp. 654–665, 2005.

[25] M. R. D. Martins, "Ouvir Falar – Introdução à Fonética do Português", Editorial Caminho, Lisboa, pp. 12-95. 1988.

[26] Institute for Quality and Efficiency in Health Care (IQWiG, Germany), "How does the larynx work?," 27 mar. 2024, [online]. Disponível: https://www.informedhealth.org/how-does-the-larynx-work.html. [Acedido: 9 abr. 2024].

[27] K. L. McCullagh, R. N. Shah, & B. Y. Huang, "Anatomy of the Larynx and Cervical Trachea," Neuroimaging Clinics of North America, vol. 32, no.. 4, pp. 809–829, 2022.

[28] H. M. Tucker, "The Larynx", 2nd ed. Theme Medical Publlishers Inc., New York, pp. 1-34. 1993.

[29] Central da Fonoaudiologia, "Nódulos de Pregas Vocais – Central da Fonoaudiologia," [online]. Disponível: https://www.centraldafonoaudiologia.com.br/tratamentos/fonoaudiologia-nodulos-de-pregas-vocais/. [Acedido: 11 abr. 2024].

[30] M. Behlau, "Voz, O Livro do Especialista", Revinter, Rio de Janeiro, pp. 1-37. 2001.

[31] S. M. Rua, "Estudo Morfológico-Dinâmico do Tracto Vocal Humano," Dissertação de Mestrado, Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, set. 2006.

[32] MCC - MacroCosmos Creations Private Limited, "The Human Voice, Voice Disorders and Voice Therapy," 29 fev. 2023, [online]. Disponível: https://medium.com/@creationsmacrocosmos/the-human-voice-voice-disorders-and-voice-therapy-7162d0740a10. [Acedido: 16 abr. 2024]

[33] K. Verdolini, C. A. Rosen, & R. C. Branski, "Classification Manual for Voice Disorders-I," Psychology Press, pp. 19-88, 2005.

[34] P. Podoll, P. Caspary, H. W. Lange & J. Noth, “Language functions in Huntington’s disease,”, Brain, vol. 111, pp. 1475-1503, 1988.

[35] S. Sapir, “Multiple factors are involved in the dysarthria associated with Parkinson’s disease: A review with implications for clinical practice and research,” J. Speech, Lang., Hearing Res., vol. 57, no. 4, pp. 1330–1343, ago. 2014.

[36] T. Makkonen, H. Ruottinen, R. Puhto, M. Helminen & J. Palmio, "Speech deterioration in amyotrophic lateral sclerosis (ALS) after manifestation of bulbar symptoms," Int. J. Lang. Commun. Disord., vol. 53, pp. 385-392, mar. 2018.

[37] M. Müller, "Fundamentals of Music Processing", Springer, London, pp. 18-110. 2015.

[38] I2tutorials, “What do you mean by Fourier Transforms?,” 27 set. 2019, [online]. Disponível: https://www.i2tutorials.com/what-do-you-mean-by-fourier-transforms-how-can-we-use-in-machine-learning/. [Acedido: 23 abr. 2024]

[39] M. Vashkevich, A. Petrovsky, & Y. Rushkevich, "Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato," Signal Process. Algorithms, Archit. Arrange. Appl. (SPA), pp. 267-272, set. 2019.

[40] R. J. Baken & R. F. Orlikoff, “Clinical measurement of speech and voice”, 2nd Edition. Singular Thomson Learning, 2000.

[41] J. Fernandes, F. Teixeira, V. Guedes, A. Junior, & J. P. Teixeira, "Harmonic to Noise Ratio Measurement - Selection of Window and Length," Procedia Computer Science, vol. 138, pp. 280-285, 2018.

[42] J. V. Guttag, "Introduction to Computation and Programming using Python", 2a ed., The MIT Press, Cambridge, Massachussets, cap. 17-19. 2016

[43] F. T. AL-Dhief, N. M. A. Latiff, M. M. Baki, N. N. N. A. Malik, N. Sabri & M. A. A. Albadr, "Voice Pathology Detection Using Support Vector Machine Based on Different Number of Voice Signals," 2021 26th IEEE Asia-Pacific Conference on Communications (APCC), Kuala Lumpur, Malaysia, pp. 1-6, 2021.

[44] H. Cordeiro, C. Meneses, “Low band continuous speech system for voice pathologies identification,” 2018 Signal Process. Algorithms, Archit. Arrange. Appl., pp. 315–320, set. 2018.

[45] S. Gayathri & E. Priya, "Identification of voice pathology from temporal and cepstral features for vowel a low intonation," in 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), pp. 345-350, 2022.

[46] S. Gayathri & E. Priya, "Spectro-Temporal Based Feature Extraction for Identification of Pathological Voice Signals from Normal," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, pp. 1-6, 2023.

[47] H. Cordeiro, C. Meneses, “Parâmetros espectrais de vozes saudáveis e patológicas,” in 14th Iberian Conference on Information Systems and Technologies (CISTI), jun. 2019.

[48] Y. Yang et al., "Speech Feature-Based Machine Learning Model and Smart Devices for Stroke Early Recognition," 2023 International Conference on Advanced Robotics and Mechatronics (ICARM), Sanya, China, 2023, pp. 354-359, 2023.

[49] H. M. A. Mohammed, A. Omeroglu, M. Polat, E. Oral, & I. Ozbek, "Voice Pathology Classification Using Machine Learning," in Proc. 2nd International Symposium on Applied Science and Engineering (ISASE 2021), pp. 354-357, mai. 2021.

[50] H. Cordeiro, J. Fonseca, I. Guimarães & C. Meneses, "Voice pathologies identification speech signals, features and classifiers evaluation," 2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, pp. 81-86, 2015.

[51] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Hierarchical Classification and System Combination for Automatically Identifying Physiological and Neuromuscular Laryngeal

Pathologies," Journal of voice: official journal of the Voice Foundation, vol. 31, no. 3, pp. 384.e9-384.e14, May 2017.

[52] S. R. Kadiri & P. Alku, "Analysis and Detection of Pathological Voice Using Glottal Source Features," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 367-379, fev. 2020

[53] P. Hippargekar, S. Bhise, S. Kothule, & S. Shelke, "Acoustic Voice Analysis of Normal and Pathological Voices in Indian Population Using Praat Software," Indian Journal of Otolaryngology and Head & Neck Surgery, vol. 74, no. Suppl 3, pp. 5069-5074, dez. 2022,

[54] M. Rehman, A. Shafique, Q.-U.-A. Azhar, S. S. Jamal, Y. Gheraibia, & A. Usman, "Voice disorder detection using machine learning algorithms: An application in speech and language pathology," Engineering Applications of Artificial Intelligence, fev. 2024.

[55] K. Lee, C. Moon, & Y. Nam, "Diagnosing Vocal Disorders using Cobweb Clustering of the Jitter, Shimmer, and Harmonics-to-Noise Ratio," KSII Transactions on Internet and Information Systems, vol. 12, pp. 5541-5554, 2018.

[56] J. Teixeira, N. Alves, & P. Fernandes, "Vocal Acoustic Analysis: ANN Versus SVM in Classification of Dysphonic Voices and Vocal Cord Paralysis," International Journal of E-Health and Medical Communications, vol. 11, pp. 37-51, 2020.

[57] H. Cordeiro, C. Meneses, “An improved algorithm for the Low Band Spectral Tilt estimation for pathological voice detection,” 2019 Signal Process. Algorithms, Archit. Arrange. Appl., pp. 202-207, set. 2019.

[58] S. Cotter, H. Cordeiro and G. Marques, “Classificação de Pacientes Diagnosticados com ELA através de Parâmetros Espectrais,” in 17th Iberian Conference on Information Systems and Technologies (CISTI), pp. 25-30, jun. 2022.

[59] G. A. R. Silva, M. A. R. Alves, B. C. Bispo, M. E. Dajer, and P. M. Rodrigues, "Diferenciação entre edema de reinke e nódulos vocais através de parâmetros não-lineares da voz," in XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, pp. 1-5. Sociedade Brasileira de Telecomunicações, 2021.

[60] M. O. Rosa et al., "Adaptive estimation of residue signal for voice pathology diagnosis," IEEE Trans. Biol. Eng., vol. 47, pp. 96–104, 2000.

[61] P. R. Scalassara et al., “Autoregressive decomposition and pole tracking applied to vocal fold nodule signals,” Pattern Recognit. Lett. 28 (11), pp 1360–1367, 2007.

[62] P. R. Scalassara, M. E. Dajer, C. D. Maciel, R. C. Guido, & J. C. Pereira, “Relative entropy measures applied to healthy and pathological voice characterization,” Appl. Math. Comput., vol. 207, no. 1, pp. 95–108, 2009.

[63] Eye M, Infirmary E. Voice Disorders Database, (Version 1.03 CD-ROM). Kay Elemetrics Corp., Lincoln Park, NJ; 1994.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, & E. Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[65] A. R. Lemos, “Cross-Validation: What is it and why use it?,” Towards Data Science, 14 abr. 2022, [online]. Disponível: https://towardsdatascience.com/cross-validation-705644663568. [Acedido: 28 mai. 2024]

[66] C. Cortes & V. N. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, pp. 273-297, 1995.

[67] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data", Cambridge University Press, Cambridge, pp. 211-218. 2017.

[68] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Transactions on Information Theory, vol. 14, no. 1, pp. 55-63, jan. 1968.

[69] I. Jolliffe, "Principal component analysis: A beginner's guide - I. Introduction and application," Weather, vol. 45, pp. 375-382, 1990.

[70] L. Lamel, L. Rabiner, A. Rosenberg & J. Wilpon, “An improved endpoint detector for isolated word recognition,” in IEEE Transactionson Acoustics, Speech, and Signal Processing, vol. 29, no. 4, pp. 777-785, ago. 1981.

[71] H. Stevens, S. S. Volkmann, & E. B. Newman, “A Scale for the Measurement of the Psychological Magnitude Pitch,” J. Acoust. Soc. Am., vol. 8, no. 3, pp. 185–190, 1937.

[72] P. Boersma & D. Weenink, "Praat: Doing phonetics by computer," Phonetic Sciences, University of Amsterdam. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[73] Y. Jadoul, B. Thompson, B. de Boer, "Introducing Parselmouth: A Python interface to Praat," Journal of Phonetics, vol. 71, pp. 1-15, 2018.