Hugo Tito Cordeiro

Mestre

Reconhecimento de Patologias da Voz usando Técnicas de Processamento da Fala

Dissertação para obtenção do Grau de Doutor em
Engenharia Electrotécnica e de Computadores

Orientador: Doutor Carlos Meneses,
Prof. Coordenador, ISEL-IPL
Co-orientador: Doutor José Fonseca,
Prof. Auxiliar com Agregação, FCT-UNL

Júri:

Presidente: Prof. Doutor Luís Manuel Camarinha de Matos
Arguentes: Profª. Doutora Ana Paula de Brito Garcia Mendes
Prof. Doutor João Paulo Ramos Teixeira

Vogais: Prof. Doutor Carlos Eduardo de Meneses Ribeiro
Profª. Doutora Isabel Cristina Ramos Peixoto Guimarães
Prof. Doutor Pedro Miguel Dinis de Almeida

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Outubro de 2016

Reconhecimento de Patologias da Voz usando Técnicas de Processamento da Fala

Copyright © Hugo Cordeiro, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Agradecimentos

Em primeiro lugar gostaria de agradecer ao Prof. Carlos Meneses e ao Prof. José Fonseca por terem aceitado a orientação e a co-orientação desta tese. A vossa colaboração a todos os níveis, disponibilidade e acima de tudo a vossa motivação foram decisivos ao longo desta jornada.

Os meus agradecimentos à Prof.ª Isabel Guimarães, as suas contribuições e contactos foram determinantes para o desenvolvimento do trabalho.

Os meus sinceros agradecimentos à Prof.ª Ana Mendes que me disponibilizou a base de dados da MEEI, crucial no desenvolvimento desta tese.

À Prof.ª Maria Eugenia Dajer que disponibilizou a base de dados da Universidade de São Paulo, sendo esta uma mais-valia e uma motivação quando não possuía nenhuma base de dados para iniciar o trabalho.

Agradeço também ao ISEL e ao IPL pela dispensa de serviço de docente durante um semestre que me permitiu focar e realizar uma progressão mais célere na investigação.

Por fim agradeço à minha família, pois foi certamente o vosso apoio ao longo dos anos que me permitiu ter chegado aqui.

Resumo

O diagnóstico de patologias da voz envolve normalmente métodos invasivos que requerem esforços conjuntos de equipas multidisciplinares. A utilização de um método automático baseado em técnicas de processamento de fala, sendo não invasivo e rápido, pode ser um método de rastreio ou um diagnóstico preliminar ao realizado por especialistas.

Esta tese propõe soluções para a identificação de patologias da voz através do processamento do sinal de fala. Os métodos utilizados envolvem classificadores tipicamente usados em reconhecimento de orador, como por exemplo *support vector machines* e *Gaussian mixture models*. Os parâmetros que modelam a fonte do aparelho fonador não têm obtido resultados relevantes na distinção entre patologias. Contudo abordagens com parâmetros que modelam o trato vocal obtêm melhor sucesso nesta tarefa, assim como nos diagnósticos de vozes patológicas. Nesta linha, os parâmetros utilizados nesses classificadores têm como objectivo modelar o trato vocal, como por exemplo os *mel-frequency cepstral coefficients*, os *line spectral frequencies e mel-line spectral frequencies*.

É ainda proposto o uso de fala contínua como sinal para a identificação de patologias. Este sinal, ao exigir um maior esforço por parte do paciente e por ser mais rico em termos fonéticos, aliado ao facto de as patologias da voz produzirem alteração em todos os fonemas, permite melhorar os resultados no diagnóstico. Nesta abordagem foram realizados testes usando três classes: sujeitos saudáveis; sujeitos com patologias laríngeas fisiológicas (edemas e nódulos); e sujeitos com patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). A taxa de reconhecimento obtida foi de 84% para as três classes.

Esta tese propõe também soluções para o reconhecimento de vozes patológicas, com base na análise de formantes e na relação harmónicas-ruído. Neste sentido, foi efectuada a implementação de um algoritmo simples baseado em árvores de decisão que permitiu uma taxa de reconhecimento de 95%.

Termos Chave: vozes patológicas, patologias da voz, patologias laríngeas, reconhecimento, sinais de Fala.

Abstract

The diagnosis of voice pathology usually involves invasive methods that require efforts from multidisciplinary teams. The use of an automatic method based on speech processing techniques, which is non-invasive and fast, can be a method of screening or a preliminary diagnosis performed by an expert.

This thesis proposes solutions for pathological voice identification using speech signal processing. The methods use typically classifiers used in speaker recognition, such as support vector machines and Gaussian mixture models. The parameters that model the source of the vocal tract do not obtain relevant results in the diagnosis of pathologies. However approaches using parameters that model the vocal tract achieve better success in this task and in the diagnosis of pathological voices. Parameters that intended to model the vocal tract such as mel-frequency cepstral coefficients, line spectral frequencies and mel-line spectral frequencies are used in these classifiers.

Continuous speech signal is also proposed for voice pathologies identification. This signal, which requires a greater effort by the patient and is phonetically richer and the fact that voice pathologies affect all phonemes, improves the results in the diagnosis. In this approach tests were carried out using three classes: healthy subjects; subjects with physiological larynx pathologies (vocal fold nodules and oedemas); and subjects with neurological larynx pathologies (unilateral vocal fold paralysis). The recognition rate obtained was 84% for the three classes.

This thesis also proposes solutions to the pathological voice recognition based on formant analysis and harmonic-to-noise ratio. A simple algorithm based on decision trees allowed a 95% recognition rate.

Keywords: pathological voices, voice pathologies, laryngeal pathologies, recognition, speech signals.

Índice

Índice de Figuras

Índice de Tabelas

Acrónimos

ACC: *accuracy*

COPAS: *Dutch Corpus of Pathological and Normal Speech*

CGI: *glottal closure instant*

DMLSF: *differential mel-line spectral frequencies*

EER: *equal error rate*

FN: *false negative*

FP: *false positve*

GLDS: *generalized linear discriminant sequence kernel*

GMM: *gaussian mixture models*

GSL: *GMM supervector linear kernel*

HNR: *harmonic-to-noise-ratio*

LD: *linear discriminant*

LP: *linear prediction*

LSF: *line spectral frequencies*

MEEI: *Massachusetts Eye and Ear Infirmary*

MFCC: *mel-frequency cepstral coefficients*

MLPCC: *mel-linear prediction cepstral coefficients*

MLSF: mel-*line spectral frequencies*

PLP: *perceptual linear prediction*

PPV: *positive predictive value*

RPPC: *relative power of the periodic component*

SVM: *support vector machines*

TN: *true negative*

TP: *true positive*

TPR: *true positive rate*

UBM-GMM: *universal background model - gaussian mixture models*

1 Introdução

As patologias da voz [1] surgem devido a várias circunstâncias tais como o uso vocal exaustivo ou incorrecto, stress, inalação do fumo do tabaco, refluxo gástrico, perturbações neurológica ou problemas hormonais. Estas patologias afectam tipicamente as pregas vocais e são detectáveis através de laringoscopia directa, que consiste na visualização das pregas vocais com recurso a uma câmara. No entanto, este método é invasivo, desconfortável para o paciente e poderá nalguns casos, dependendo do equipamento utilizado, requerer uma anestesia local. No caso de o exame ser efectuado através de laringoscopia indirecta, é realizado através de um espelho que, apesar de menos incómodo, implica muitas vezes o uso de pequenas quantidades de sedativos locais.

A detecção de patologias da voz envolve frequentemente equipas multidisciplinares, existindo patologias que requerem mais atenção no diagnóstico, não estando dependentes apenas do exame de um médico otorrinolaringologista, mas também de neurologistas e de terapeutas da fala. Nestes casos, as patologias não estão unicamente relacionadas com as pregas vocais, mas com os músculos que as controlam ou que fazem parte do aparelho fonador.

1.1 Motivação

O uso de um método de avaliação de voz eficiente e não invasivo, que permita o fácil e rápido reconhecimento de patologias, pode ser útil numa primeira avaliação para rastreio ou como método complementar no diagnóstico das patologias da voz.

O processamento de fala encontra-se em constante progressão em áreas como o reconhecimento de fala, o reconhecimento de orador, a síntese de fala a partir de texto, a diarização e a codificação. Estas áreas são exemplos em que parâmetros extraídos do sinal de fala são usados com o intuito de reconhecer palavras, reconhecer pessoas, produzir palavras ou transmitir sinais, construindo aplicações de inegável utilidade.

Apesar de já existir desenvolvimento significativo no reconhecimento de vozes com patologias usando ferramentas de processamento de fala, o reconhecimento automático destas patologias ainda se encontra numa fase inicial. Quando não existe intenção de classificar qual a patologia está-se na presença de um sistema de reconhecimento de vozes patológicas. Neste

caso, o sistema consiste num classificador binário, onde o sinal de teste é classificado como tendo ou não patologia. Por outro lado o reconhecimento de patologias da voz consiste na identificação da patologia específica que afecta a voz de um paciente, sendo que esta área apresenta poucas contribuições encontrando-se ainda numa fase inicial de investigação.

1.2 Objectivo

As patologias da voz afectam a capacidade humana de comunicar. Actualmente, a detecção destas patologias é efectuada recorrendo a métodos invasivos e desconfortáveis para o paciente. A realização de diagnósticos por métodos não invasivos, de um modo rápido e eficiente, terá certamente um impacto positivo na detecção das patologias que afectam tipicamente o desempenho das pregas vocais, prejudicando a capacidade de produzir fala, principalmente dos sons vozeados. Consequentemente, o individuo realiza mais esforço vocal, o que na maior parte dos casos leva ao aumento progressivo dos sintomas da patologia.

No estado da arte pode-se constatar que existe bastante trabalho realizado no reconhecimento de vozes patológicas, sendo que nesta vertente não é reconhecida qual a patologia mas apenas se detecta se existe alguma. Por outro lado, a investigação no reconhecimento de patologias da voz, onde é reconhecida qual a patologia das pregas vocais, é escassa. Contudo, o reconhecimento das patologias da voz com recurso a métodos de reconhecimento de fala é viável, tendo em conta diversos trabalhos referidos no estado da arte desta tese.

O reconhecimento de vozes patológicas pode ser realizado com recurso a medidas que caracterizam o movimento das pregas vocais, tais como o *jitter do período fundamental* e o *shimmer*. No entanto, apesar das patologias afectarem tipicamente as pregas vocais, estas medidas não produzem resultados conclusivos no reconhecimento de patologias da voz. Nesta área, o recurso a modelos do filtro ainda não foi totalmente explorado.

O objectivo deste trabalho é estudar métodos normalmente usados em processamento de fala e aplicá-los ao reconhecimento de patologias da voz, com recurso a parâmetros que caracterizam o trato vocal. Como tal serão usados sistemas de classificação e técnicas de mineração de dados que permitam avaliar e analisar parâmetros dos sinais de fala de modo a desenvolver sistemas para reconhecer vozes patológicas e patologias da voz. Numa

abordagem inicial serão propostas soluções no âmbito do reconhecimento de vozes patológicas. Neste caso não existe identificação da patologia mas apenas é indicado se o paciente tem uma qualquer patologia. Posteriormente serão implementados métodos com o intuito de reconhecer as patologias da voz, onde se procederá à sua identificação. No final será apresentado um conjunto de sistemas de reconhecimento baseados em várias características do sinal de fala que têm como objectivo quer o reconhecimento de vozes patológicas quer o reconhecimento de patologias.

1.3 Questões de investigação

A elaboração das questões de investigação visa a adopção de hipóteses que poderão ser fundamentadas no decurso da investigação originando contribuições que representam soluções no reconhecimento de patologias da voz e no reconhecimento de vozes patológicas, através do processamento de sinais de fala. O ponto seguinte apresenta as questões de investigação abordadas nesta tese e as respectivas hipóteses.

1.3.1 Questões e hipóteses

Questão 1: O uso de parâmetros de modelação do trato vocal, usados em reconhecimento de fala/orador, serão fiáveis no reconhecimento de patologias da voz? Quais os parâmetros que proporcionam melhores resultados na discriminação de patologias?

Hipótese: Existem vários parâmetros que modelam o trato vocal, utilizados em várias aplicações de reconhecimento de fala, sendo os MFCC (*mel-frequency cepstral coefficients*) os parâmetros mais usados. Por outro lado, em aplicações de codificação, onde o principal objectivo é manter a percepção da fala, são usados parâmetros como os LSF (*line spectral frequencies*), sendo que estes têm fraco desempenho no reconhecimento. Os MLSF (*mel-line spectral frequencies*), derivados destes últimos, mas com informação perceptiva, proporcionam resultados semelhantes aos MFCC em aplicações de reconhecimento de orador. Deverá assim ser efectuada uma avaliação do desempenho destes parâmetros no reconhecimento de patologias da voz.

Questão 2: Será que o uso de vários fonemas, ou mesmo de fala contínua, no reconhecimento de patologias da voz pode aumentar a taxa de reconhecimento?

Hipótese: Assumindo que a patologia afecta as pregas vocais, ou seja a fonte, e que as perturbações na fonte têm posteriormente impacto no sinal que vai ser modelado no trato vocal, a perturbação da fonte vai ser transversal ao fonema, ou seja, será comum. Por outro lado, patologias diferentes deverão produzir perturbações diferentes, sendo provavelmente mais fácil caracterizar patologias usando fala contínua, assumindo que a patologia afecta de igual forma todos os fonemas vozeados. Esta hipótese é sustentada pelo facto de ser possível a um otorrinolaringologista ou terapeuta da fala experiente conseguir reconhecer patologias apenas ouvindo o paciente, visto que determinadas patologias produzem alterações típicas no sinal de fala.

1.3.2 Contribuições

De seguida são descritas de forma sucinta as questões de investigação que foram a base desta tese e as respectivas contribuições produzidas no decorrer da investigação para as resolver.

Questão 1

Um estudo da viabilidade do reconhecimento de patologias foi publicado em [2], [3]. Nesta contribuição foram avaliados os desempenhos dos parâmetros MFCC, LSF e MLSF, em vários sistemas de identificação com classes constituídas por sujeitos saudáveis, sujeitos com patologias laríngeas fisiológicas (edemas e nódulos) e sujeitos com patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). Os resultados obtidos demostram a viabilidade na utilização de parâmetros do sinal de fala relacionados com o trato vocal no reconhecimento destas patologias.

Numa primeira fase foi realizado um estudo da envolvente espectral através de predição linear (LPC – *linear predictive coding*), onde foi demostrado que, ao aumentar a ordem deste parâmetro, é encontrado um primeiro pico no espectro do sinal em vozes patológicas antes do típico primeiro formante da vogal /a/. Em [4] verificou-se que, usando um LPC de ordem 30, a frequência do primeiro pico e a sua largura de banda permitem a completa separação entre sujeitos saudáveis e sujeitos com patologias. Verificou-se ainda que este pico permite a separação, embora com erros, entre sujeitos com edemas de Reinke e sujeitos com nódulos.

Em [5], foram implementados os sistemas de identificação de patologias. Os sistemas avaliam sujeitos diagnosticados com nódulos e edema de Reinke. O objectivo principal desta contribuição é verificar se o primeiro pico do espectro LPC está presente noutras vogais, tendo sido encontrado também nas vogal /e/ e /i/. Para além dos sistemas avaliarem o desempenho do primeiro pico do espectro de LPC de ordem 30 também avaliaram o desempenho da frequência fundamental e do *jitter* do período fundamental nas três vogais /a/, /e/ e /i/.

Em [6], verificou-se que a frequência do primeiro pico de LPC de ordem 30 e a sua largura de banda permite classificar correctamente cerca de 80% dos sujeitos numa base de dados com maior número de sujeitos e um maior número de patologias. Este estudo foi realizado usando apenas a vogal /a/.

Questão 2

Em [8], foram implementados sistemas de identificação de patologias para a vogal /a/ e para fala contínua. Estes sistemas baseados em SVM (*support vector machines*) e GMM (*Gaussian mixture models*) usado MFCC realizaram identificação distinguindo entre 3 classes compostas por sujeitos saudáveis, sujeitos diagnosticados com patologias laríngeas fisiológicas (edemas e nódulos) e sujeitos diagnosticados com patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). Verificou-se que a taxa total de diagnostico nos sistemas que analisaram fala contínua superaram os resultados obtidos pelos sistemas que usaram apenas a vogal /a/. Estes resultados foram posteriormente confirmados em [2] usando outros parâmetros do sinal de fala como os parâmetros LSF e MLSF. Em [3] vários classificadores usando ambos os sinais de fala (vogal /a/ e fala contínua) e vários parâmetros foram combinados de modo optimizar os resultados.

1.4 Metodologia de investigação

O principal objectivo deste trabalho é implementar métodos que permitam o reconhecimento de vozes patológicas e o reconhecimento de patologias da voz através da análise de sinais de fala. Esta tese apresenta assim contribuições em duas áreas de reconhecimento de patologias, diferenciadas por se poder ou não determinar qual a patologia que afecta o paciente.

Neste contexto a selecção e análise de sinais é um ponto importante do ponto de vista das metodologias a adoptar, pois questões de desbalanceamento entre classes altamente

correlacionadas como sejam as classes de diferentes patologias da voz podem efectivamente ocorrer, nomeadamente em reconhecedores onde o processo de aprendizagem é dependente de todas as classes.

Após a selecção dos sinais de fala são extraídos parâmetros com base na análise de modelos matemáticos inerentes ao processamento de sinais de fala. A análise das características dos sinais de fala tem impacto nas taxas de reconhecimento dos sistemas e consequentemente na identificação de vozes patológicas e das respectivas patologias. Dependendo do grau de abstracção destas características, a sua análise pode ser feita por medidas quantitativas ao nível da taxa de reconhecimento. Por outro lado também é possível, em casos de características de mais baixo nível, o estudo de medidas estatísticas que permitem obter contribuições para esta tese, nomeadamente ao nível do reconhecimento de vozes patológicas. A utilização de sistemas de classificação de dados permite a avaliação dos parâmetros dos sinais de fala através da análise de resultados em termos de taxa de reconhecimento das classes envolvidas. Por outro lado, técnicas de mineração de dados permitem uma abordagem mais precisa na selecção de atributos, em particular de parâmetros do sinal de fala, que sejam considerados relevantes no reconhecimento de vozes patológicas e/ou na identificação das patologias.

Neste contexto a metodologia usada neste trabalho consiste no desenvolvimento das seguintes acções:

- Seleccionar sinais de fala de modo a primeiro maximizar a quantidade de dados a serem avaliados. No caso da identificação de patologias, maximizar a quantidade de sinais de dentro das classes das patologias de modo a balancear estas classes;

- Extrair parâmetros através de métodos matemáticos que permitam modelar as características dos sinais de fala;

- Analisar as características dos sinais de fala através da análise gráfica dos parâmetros extraídos ou através de métodos de classificação de baixo nível que permitam retirar conclusões ao nível da relevância das características, se estas forem propícias a esta abordagem;

- Se o nível de abstracção dos parâmetros for elevado, não sendo adequado a análise pelo método anterior, realizar-se-á a implementação de sistemas de reconhecimento baseados em modelos de aprendizagem automática para posterior análise dos resultados;

- Realizar optimizações com base nos resultados obtidos, quer sejam ao nível das características do sinal de fala ou ao nível dos parâmetros dos sistemas de reconhecimento. Em ambos os casos, serão adoptados métodos de validação cruzada;

- Retirar conclusões dos resultados obtidos e relacionar as características extraídas do sinal de fala com as características espectrais e perceptivas das patologias ou das vozes patológicas, comparativamente com as vozes saudáveis.

- Finalmente elaborar soluções e contribuições através da proposta de novas abordagens no reconhecimento de vozes patológicas e no reconhecimento de patologias da voz.

1.5 Estrutura da tese

O texto desta tese é constituído por seis capítulos. Neste primeiro capítulo é efectuada uma introdução onde é apresentada a motivação assim como o objectivo desta investigação. Seguidamente são apresentadas as questões de investigação e as contribuições para as resolver, assim como a metodologia de investigação que descreve as acções desenvolvidas para obter soluções e contribuições que satisfaçam os objectivos deste trabalho.

O Capítulo 2 é composto por quatro secções. Na primeira secção, uma vez que esta tese tem um carácter multidisciplinar e para facilitar a sua compreensão a um público alargado, são apresentados sucintamente os temas abordados no processamento de fala clássico, como seja a codificação de fala, síntese de fala, reconhecimento de fala e de orador. Numa segunda secção é descrita sucintamente a produção de fala e as patologias mais comuns da voz, incluindo todas as utilizadas nesta tese. Na terceira secção é apresentado o estado da arte no reconhecimento de vozes patológicas e na identificação de patologias da voz e métodos tipicamente usados em reconhecimento de orador. Este capítulo termina com as conclusões, que incluem considerações sobre os campos ainda a explorar, algumas das quais serão objecto de estudo nesta investigação.

No Capítulo 3 são apresentados os materiais e os métodos usados na investigação, sendo descritas as bases de dados utilizadas e os parâmetros do sinal de fala, assim como os reconhecedores e os métodos de classificação/mineração de dados utilizados no decorrer da investigação.

No Capítulo 4 é proposta a caracterização de vozes patológicas através da envolvente espectral e da análise de ruido através da potência relativa da componente periódica do sinal. São apresentadas soluções com base nesta característica e a implementação de um sistema de reconhecimento baseado em *support vector machines*, é implementado um sistema baseado em árvores de decisão para o reconhecimento de patologias da voz. O capítulo termina com a validação e conclusões.

O Capítulo 5 apresenta o estudo dos sistemas de identificação de patologias da voz baseados em fala contínua. É desenvolvido um conjunto de reconhecedores que comparam vários parâmetros dos sinais de fala extraídos da vogal /a/, tipicamente usada em reconhecimentos de patologias da voz e parâmetros extraídos da fala contínua. O sinal de fala contínua é proposto nesta tese para a identificação de patologias laríngeas fisiologias (edemas e nódulos) e patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). Por fim foi desenvolvido uma combinação dos sistemas de reconhecimento que permitiu optimizar os resultados obtidos. No final deste capítulo foi realizado um estudo para verificar o impacto da diminuição da largura de banda dos sinais de fala nos sistemas de reconhecimento. O capítulo termina com a validação e conclusões.

Esta tese termina com as conclusões, Capítulo 6, onde são apresentadas as conclusões, contribuições, artigos publicados e propostas de trabalho futuro.

2 Enquadramento e estado da arte

Atendendo à multidisciplinaridade do tema desta tese este capítulo contém uma primeira secção onde são apresentadas as áreas típicas de processamento da fala, que saem do âmbito normal da engenharia electrotécnica e de computadores. Na secção seguinte, também saindo do âmbito da engenharia electrotécnica e de computadores, são introduzidos conceitos sobre produção de fala e patologias da voz, em que se apresenta a fisiologia do aparelho vocal assim como a questões inerentes à produção vocal de modo a facilitar o entendimento do estado da arte em reconhecimento de patologias da voz e no reconhecimento de patologias da fala. São apresentadas também descrições das patologias da voz mais comuns, incluindo as usadas no âmbito desta tese.

Na terceira secção, o estado da arte, revê-se a literatura estando organizada em quatro partes: na primeira faz-se referência a trabalhos de reconhecimento de vozes patológicas; na segunda são descritos trabalhos sobre sistemas de reconhecimento de patologias da voz; na terceira parte apresentam-se trabalhos sobre sistemas de reconhecimento de orador; e na última apresentam-se as conclusões e as direcções a seguir.

2.1 Processamento digital de fala

O objecto do processamento digital de sinais de fala é a representação digital dos sinais de fala, a análise e extracção de características e o desenvolvimento de modelos de síntese. Todas estas ferramentas são cruciais na implementação de sistemas de comunicação falada, seja esta comunicação à distância ou comunicação homem-máquina. Tradicionalmente, estes sistemas estão divididos em sistemas de codificação, síntese, reconhecimento de fala e verificação e identificação do orador.

2.1.1 Codificação de fala

Uma das mais antigas aplicações do processamento digital de fala é a codificação [9]. Por codificação entende-se a representação eficiente do sinal (tentando diminuir o débito binário na representação do sinal) com vista à sua transmissão ou armazenamento, mas mantendo a qualidade acima da exigida pela aplicação. O esforço de investigação realizado nesta área tem conduzido a diversas normas vocacionadas para trabalhar na rede telefónica pública, e mais

recentemente têm sido desenvolvidas normas para as utilizar também em redes baseadas em protocolos transmitindo pacotes de informação nas redes sem fios [10].

No caso da transmissão de informação, o débito binário de codificação da fonte é um dos factores mais importantes na definição da largura de banda dos sistemas de comunicação. O armazenamento de grandes quantidades de informação para utilização posterior exige também a necessidade de reduzir o débito binário, já que este determina o espaço requerido na unidade de armazenamento. A necessidade de reduzir o débito binário mantém-se mesmo com o aumento da largura de banda dos canais de transmissão, possibilitando transmitir maior número de sinais no mesmo canal ou possibilitando lidar com canais ruidosos que careçam de códigos de detecção e correcção de erros de grande redundância. Assiste-se ainda ao emergir de serviços multimédia com integração de voz, imagens e dados, que necessitam de racionalizar a distribuição do débito binário total por cada uma das aplicações.

2.1.2 Síntese de fala

Nos sistemas de síntese de fala a partir de texto [11], [12], é gerado um sinal acústico de fala a partir de um texto fornecido pelo utilizador (eventualmente outro sistema automático). Uma das vantagens principais, senão mesmo a principal, dos sistemas de síntese de fala, prende-se com a necessidade dos sistemas com resposta automática através de voz diminuírem a quantidade de memória utilizada no armazenamento dos sinais respectivos. É ainda possível minimizar o esforço de adaptação do sistema a novas tarefas, não sendo necessário pré-gravar novas frases mas apenas inserir o texto correspondente. Uma das aplicações destes sistemas são os serviços automáticos via telefone, designadamente os de consulta a informação armazenada em bases de dados que se situem distantes dos utilizadores.

Os sistemas de síntese de fala a partir de texto são normalmente implementados dividindo o problema em dois: a conversão de texto em representação linguística (segmentos fonéticos, respectiva duração, contorno da energia e da frequência de vibração das cordas vocais); e a síntese da onda acústica a partir da representação linguística. No desenvolvimento destas duas tarefas, principalmente da primeira, são necessários conhecimentos profundos de linguística, resultando sempre sistemas dependentes da língua.

Numa era onde a modelação 3D está em grande desenvolvimento, a articulação de modelos 3D de faces com as expressões de produção de fala apresenta novos desafios nesta área [13],

aproximando a imagem e a fala e tornando mais naturais duas áreas que nem sempre seguiram ligadas.

2.1.3 Reconhecimento de fala

Os sistemas de reconhecimento de fala desempenham o papel do ouvinte, convertendo uma onda acústica numa mensagem escrita ou equivalente, ou identificando um comando falado [14]. A dimensão do léxico interpretado pelo reconhecedor é um parâmetro importante na sua complexidade e eficiência, podendo ser reconhecidas palavras de léxicos de pequenas dimensões (menos de 100 palavras), grandes dimensões (que podem chegar a 10000 palavras) e fala contínua, este último normalmente com constrições impostas pela tarefa a desempenhar e pela própria língua.

O reconhecimento envolve sempre uma fase de treino. Os sistemas treinados para serem utilizados com um orador específico têm um melhor desempenho que os sistemas para múltiplos oradores. Existem ainda sistemas capazes de se adaptar dinamicamente a um orador, melhorando o seu desempenho após a adaptação.

A conversão de fala para texto através do reconhecimento de fala contínua tem numerosas aplicações. É exemplo a legendagem automática em tempo real para ajuda a surdos, em programas televisivos ou mesmo no dia-a-dia. Estes sistemas podem ser previamente treinados para o locutor, melhorando o seu desempenho. Quando ligados em cadeia a um sistema de tradução e síntese de texto para fala, a conversão de fala para texto permite também a tradução automática entre línguas.

2.1.4 Reconhecimento do orador

Os sistemas de reconhecimento de oradores são implementados a partir do processamento de fala contínua sendo normalmente independentes do texto [15], isto é, o orador é reconhecido independentemente do conteúdo do sinal de fala.

O reconhecimento de orador poder ser divido em duas categorias: a verificação e a identificação. Um sistema de verificação do orador deve decidir se um orador é quem clama ser. Um sistema de identificação do orador deve decidir qual dos oradores, dentro de um conjunto de oradores, produziu determinada frase. Estes sistemas têm aplicações em

situações em que é necessário um controlo de acessos e a verificação de identidade, ou na ajuda de prova em alguns casos judiciais. A verificação de identidade do orador em sistemas acedidos através da linha telefónica (por exemplo os sistemas de *home banking*) complementa a segurança dada pela utilização de códigos numéricos digitados através do teclado do telefone, sem ser necessária a introdução de *hardware* adicional.

2.2 Produção de fala e patologias da voz

2.2.1 Aparelho fonador

A fala consiste na produção articulada de sons, os fonemas. A produção de um fonema é realizada pelo aparelho fonador (ver Figura 2-1) e consiste na passagem do ar que é expelido pelos plumões através das pregas vocais (ou cordas vocais) e pelo trato vocal e/ou pela cavidade nasal.

A produção de fonemas pelo aparelho fonador aparece vulgarmente modelada por um modelo fonte-filtro [16]. A fonte corresponde ao ar que sai dos pulmões quando comprimidos pelo diafragma e passa pelas pregas vocais provocando ou não a sua vibração. O filtro representa o trato vocal e a cavidade nasal que modelam o fluxo de ar proveniente da fonte.

No domínio da fonte existem dois tipos de sinais: sinais periódicos ou quase-periódicos e ruído branco. O primeiro ocorre quando existe vibração das pregas vocais sendo gerada uma forma de onda quase-periódica produzindo fonemas vozeados, ou seja, existe vozeamento provocado pela vibração das pregas vocais. Se, por outro lado, durante a passagem do ar as pregas vocais permanecerem completamente abertas é gerado um sinal de ruído branco e são produzidos fonemas onde não existe vozeamento, isto é, onde não existe vibração das pregas vocais.

A vibração das pregas vocais é portanto responsável pelo correto vozeamento dos fonemas. Estas estão situadas no interior da laringe e são constituídas por um tecido musculoso com duas pregas. As pregas são duas fibras elásticas que vibram com a passagem do ar e que podem modificar a sua forma e elasticidade através dos músculos da laringe, produzindo diferentes sons, consoante se quer por exemplo alterar o tom de voz ou cantar. Estes músculos são controlados pelo cérebro, que envia mensagens pelos nervos para controlar a

aproximação e a tensão das pregas vocais de modo a que estas vibrem quando o diafragma e os músculos empurram o ar para fora dos pulmões.

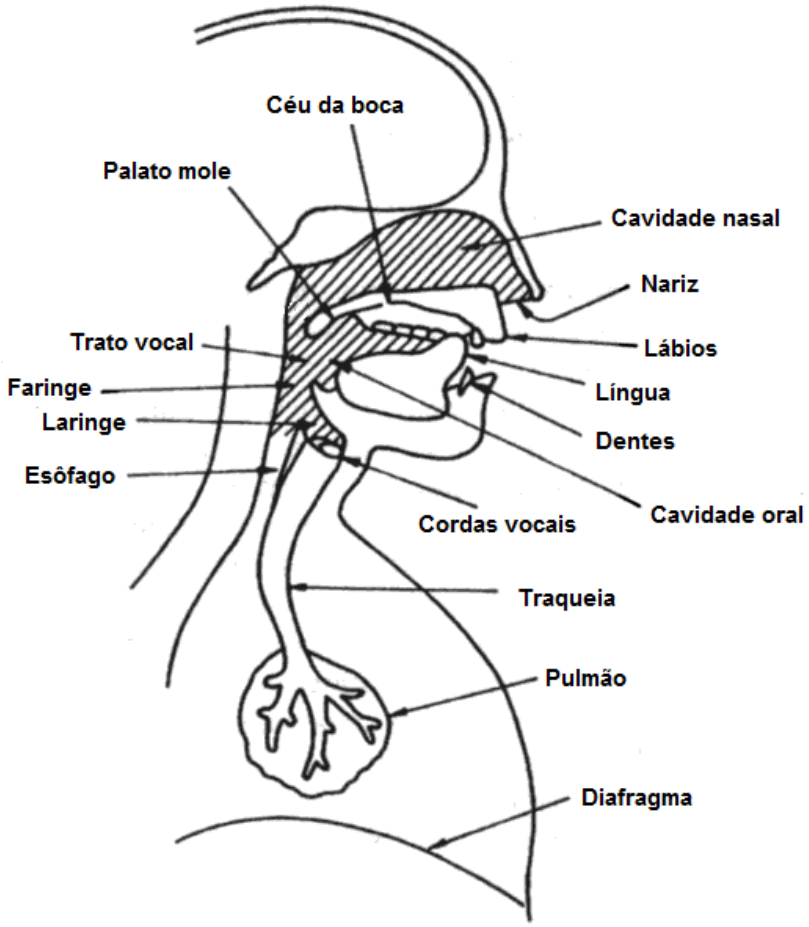

Figura 2-1 – Aparelho Fonador

A frequência de vibração das pregas vocais está associada à sua dimensão sendo tipicamente mais longas e robustas nos homens que nas mulheres, produzindo frequências mais baixas nos oradores masculinos ([80,160] Hz) do que nos femininos ([140,300] Hz). Nas crianças, as frequências fundamentais são ainda mais elevadas ([300,400] Hz).

O trato vocal funciona como um filtro do sinal proveniente da laringe, filtrando-o de modo a produzir o fonema desejado. A produção de um fonema pelo trato vocal é um processo bastante complexo, consistindo na adaptação de uma caixa-de-ressonância de um modo dinâmico e contínuo através da posição da língua, dos dentes e dos lábios. Se existir a produção de sons nasalados, o ar para além de passar pela boca passa também pelo nariz. Este

processo é comandado pelo cérebro que coordena os vários músculos envolvidos. A produção de fonemas de um modo contínuo é designada por fala contínua.

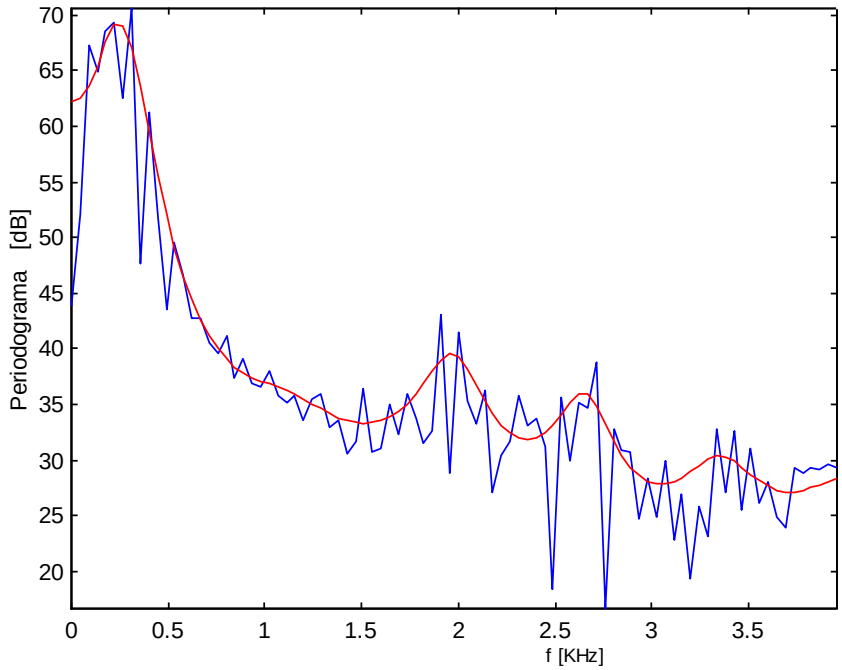

Figura 2-2 - Periodograma e respectiva envolvente espectral de uma trama (20 ms) de um segmento fonético correspondente a um /i/, produzido por um orador masculino. (F1=266 Hz, F2=2044Hz, F3=2711Hz, F4=3422Hz).

O processo de filtragem do sinal proveniente da laringe resulta num espectro caracterizado por picos, designados formantes, como ilustrado na Figura 2-2. Os formantes correspondem às ressonâncias do trato vocal e são caracterizados por uma frequência central e por uma largura de banda. A informação das frequências dos formantes (F1,F2,…) está directamente relacionada com o fonema pronunciado, estando bem caracterizado nas vogais, como mostra a Figura 2-3. Contudo, em processamento de fala, a análise do modelo do trato vocal está também relacionada com a caracterização do orador, ou seja, é com base nas características do trato vocal que se realiza o reconhecimento de orador.

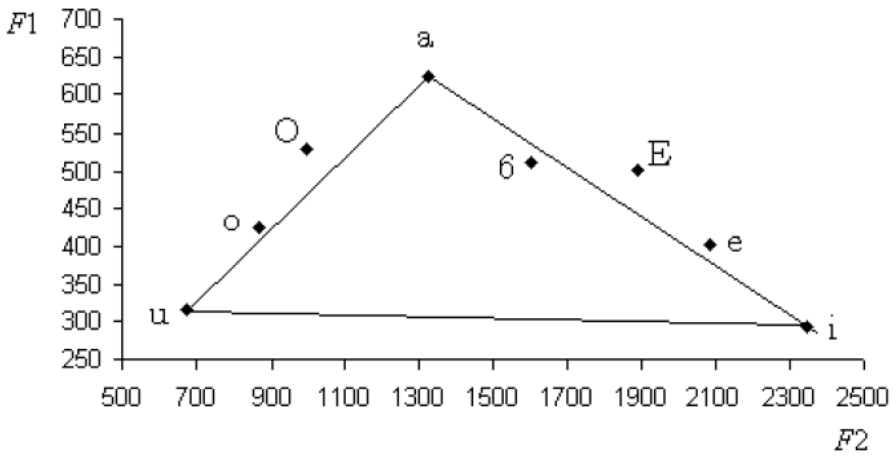

Figura 2-3 – Triângulo das vogais, frequências do primeiro e segundo formantes das vogais. Alfabeto SAMPA. Retirado de [17].

2.2.2 Patologias da Voz

As patologias que estão directamente ligadas à laringe são designadas por patologias da voz ou patologias laríngeas. Estas patologias podem ser causadas por lesões mínimas estruturais e/ou funcionais da laringe, lesões de massa localizada nas pregas vocais, alterações tecidulares da prega vocal, perturbações neurológicas e perturbações não orgânicas ou de tensão muscular [1].

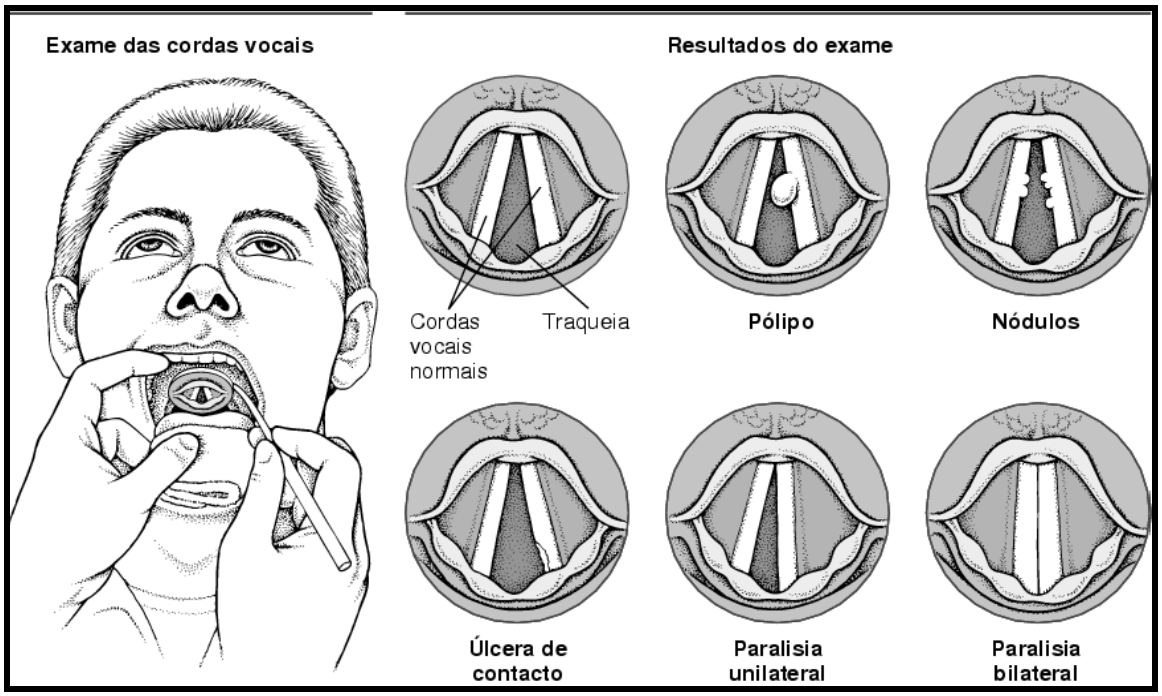

Figura 2-4 – Patologias da voz relacionadas com as pregas vocais.

A Figura 2-4 mostra o resultado de um exame às pregas vocais e como se manifestam fisicamente algumas das patologias. Nos pontos seguintes vão ser descritas algumas das patologias que estão a ser testadas actualmente na implementação de sistemas de reconhecimento de patologias da voz. São apresentados vários tipos de patologias da voz [1]: lesões de massa localizadas; alterações tecidulares; lesões neurologicas periféricas e variações mínimas; estas por sua vez definem conjuntos de patologias que se manifestam na laringe.

2.2.2.1 Lesões de massa localizada

As lesões de massa localizada afectam a camada membranosa das pregas vocais. Estas lesões podem ser unilaterais ou bilaterais, podem ser benignas, pré-malignas ou malignas. Pólipos, nódulos, quistos e ulceras de contacto são algumas das lesões mais comuns.

2.2.2.1.1 Nódulos das pregas vocais

Os nódulos das pregas vocais são lesões benignas que podem ocorrer em ambas as pregas vocais, no local onde existe maior fricção. A sua localização é normalmente na união do terço anterior com os dois terços posteriores. Quando as pregas vocais vibram existem zonas de impacto sendo nessas zonas que aparecem as lesões que originam os nódulos. Esta patologia é comum em indivíduos que utilizam intensivamente a voz, tais como políticos, professores, cantores ou crianças que gritem frequentemente. Os nódulos têm a dimensão típica de uma cabeça de alfinete e podem aparecer em conjuntos ao longo de uma determinada área. Os seus sintomas são a disfonia, com a alteração de timbre da voz, através de uma rouquidão variável e podendo, em casos extremos, atingir a afonia impossibilitando por completo o vozeamento.

Os nódulos não permitem o fecho completo das pregas vocais adicionando assim ruído ao sinal de fala. Para compensar este efeito, o paciente tende a aumentar a tensão no músculo tirearitenóideo, aumentando ainda mais as forças de colisão das pregas vocais. Para além do ruído, esta patologia provoca irregularidades na frequência fundamental da voz (causando *jitter* no período fundamental) e variações na amplitude (*shimmer*).

2.2.2.1.2 Pólipos

Os pólipos surgem tipicamente isolados, com tamanho variável, sendo habitualmente maiores do que os nódulos. Este tumor benigno tende a sair fora da prega vocal e a dilatar-se, ficando

preso à superfície através de um pedúnculo (pólipo pediculado), ou então surge como uma massa distribuída na lâmina própria (pólipo séssil). Os pólipos surgem em indivíduos que usam exaustivamente a voz, apresentando normalmente sintomas idênticos aos nódulos, embora possam causar também tosse. Em casos extremos, o pólipo pode obstruir a laringe causando mesmo dificuldades respiratórias.

O grau de disfonia provocado por esta patologia é bastante variável, dependendo da posição, tamanho do pólipo e dimensão do pedúnculo (que lhe concede maior ou menor mobilidade).

2.2.2.1.3 Quistos

Os quistos são aglomerações de líquidos cuja causa típica é o bloqueio glandular. Encontram-se localizados na camada superficial da lâmina própria sendo tipicamente unilaterais. Esta patologia pode variar de cor e tamanho podendo ter ligeiras fendas com perda ocasional de líquido. O principal sintoma desta patologia é a disfonia acompanhada por um maior esforço vocal.

2.2.2.1.4 Úlceras de contacto

As úlceras são feridas nas pregas vocais que ocorrem devido ao esforço vocal, ao fumo do tabaco, à tosse persistente ou ao refluxo do ácido gástrico. Os sintomas são dores ligeiras e diferentes graus de rouquidão, dependendo da dimensão e da posição da úlcera. Embora estas lesões sejam tipicamente pequenas comprometem frequentemente o fecho da glote.

2.2.2.2 Alterações tecidulares

2.2.2.2.1 Edema de Reinke

O edema de Reinke surge tipicamente nas duas pregas vocais, sendo caracterizado pelo seu edema ('inchaço'), provocado pela acumulação dispersa de líquido no espaço de Reinke. À medida que o líquido se acumula, este espaço aumenta, fazendo com que as pregas vocais aumentem de espessura e se projetem para o interior da laringe. Em consequência, a voz fica mais rouca e com tonalidade mais grave uma vez que o edema provoca alterações na elasticidade das pregas vocais. Em casos extremos, o edema pode mesmo dificultar a passagem de ar. Em reacção a estas alterações o paciente tende a efectuar um maior esforço

vocal originando a abertura excessiva da glote e uma vibração assimétrica, irregular e aperiódica das pregas vocais.

Uma vez que esta patologia, cuja principal causa é o tabagismo, provoca uma voz mais grave, os seus sintomas são mais fáceis de detetar nos oradores femininos, uma vez que têm normalmente a voz mais aguda.

2.2.2.3 Lesões neurológicas periféricas

As patologias da voz como a disfonia espasmódica e a paralisia das pregas vocais têm tipicamente origem em lesões do sistema nervoso e consistem na incapacidade do paciente realizar o correcto controlo dos músculos do aparelho fonador.

2.2.2.3.1 Disfonia espasmódica

A disfonia espasmódica é uma patologia que consiste nos movimentos involuntários dos músculos do aparelho fonador, quer seja de forma permanente, quer seja esporádica. As causas desta patologia não estão determinadas. Existem, no entanto, indicadores de que esta patologia esteja relacionada com o sistema nervoso que não permite o controlo dos músculos. Existem também casos em que esta patologia ocorre em famílias, indiciando uma origem genética.

Esta patologia tem três vertentes: a disfonia espasmódica adutora, abdutora e mista. Na disfonia espasmódica adutora os movimentos involuntários dos músculos da laringe levam a que as pregas vocais fechem inibindo a sua vibração e, em caso extremos, a produção de fala. Neste caso a voz é produzida com muito esforço, como que estrangulada. A disfonia espasmódica abdutora traduz-se em movimentos inesperados que forçam as pregas vocais a abrir, não permitindo a sua vibração e passando o ar diretamente dos pulmões para o trato vocal, originado uma voz com bastante ruído e como que a sussurrar, sendo uma voz pouco audível e de baixa intensidade. Por último, a disfonia espasmódica mista envolve os dois tipos anteriores.

Associados a esta patologia, podem também ocorrer movimentos involuntários de outros músculos que não da laringe, como por exemplo dos músculos da boca e da língua, o piscar de

olhos, movimentos involuntários do pescoço e movimentos repetitivos dos braços e das pernas.

O diagnóstico desta patologia envolve tipicamente um médico otorrinolaringologista, um terapeuta da fala que analisa a qualidade vocal do paciente e um neurologista que avalia outros distúrbios do paciente.

2.2.2.3.2 Paralisia das pregas vocais

A tensão e posição das pregas vocais são controladas pelos músculos laríngeos, que por sua vez são controlados pelo sistema nervoso. A paralisia das pregas vocais surge quando estes músculos não conseguem executar a sua função. A paralisia é efetivamente no músculo sendo que as pregas vocais podem continuar a vibrar (no caso de parésia) embora de um modo não controlado. A paralisia pode ser provocada por vários fenómenos tais como a compressão dos músculos devido a infeções, tumores ou intoxicações. Em alguns casos, embora raros, é originada por traumatismos no sistema nervoso central causadas por acidentes vasculares ou como manifestação de uma doença neurológica.

A paralisia pode afetar apenas uma prega vocal fazendo com que as pregas vibrem com frequências diferentes. Nestes casos, a voz apresenta um som bitonal e o paciente não consegue falar alto, perdendo o poder de amplificação vocal. Se, por outro lado, a paralisia afetar as duas pregas, pode existir o risco da abertura da glote não ser total, provocando com isso dificuldades respiratórias assim como ruído à passagem do ar respirado.

2.2.2.4 Variações mínimas

As variações mínimas são desvios da anatomia laríngea e/ou fisiologia laríngea e podem aparecer ou não juntamente com outras patologias. Consistem em diferenças na posição, forma ou massa, tensão, elasticidade e viscosidade das pregas vocais. Dependendo do grau das perturbações podem ser clinicamente insignificantes. A desidratação das pregas vocais devido ao consumo de agentes desidratantes quer seja medicação ou álcool, as inflamações, a variz na prega vocal e o edema que consiste no aumento do volume das pregas vocais devido nomeadamente ao abuso vocal, são exemplos destas variações mínimas.

2.2.3 Efeitos das patologias da voz no sinal de fala

As patologias das pregas vocais e nomeadamente da laringe afectam a produção do sinal de fala. As vozes soprosas (*breathy*) e crepitantes (*creaky*) [18] são exemplos de perturbação no sinal de fala que podem ocorrer devido à existência de patologias da voz. O incorrecto funcionamento das pregas vocais ou da glote devido às patologias descritas no ponto anterior pode provocar que as pregas vocais permaneçam inadvertidamente abertas ou fechadas criado assim perturbações no sinal de fala.

As vozes soprosas são caracterizadas pela passagem indesejada de ar pelas pregas vocais sem que esta provoque a vibração adequada das mesmas. A voz, apesar de vozeada, apresenta algum ruído de origem soprosa devido às pregas vocais não terem capacidade para fechar adequadamente. No limite, o vozeamento é nulo, sendo que neste caso está-se na presença de uma voz sussurrada. Por outro lado, as pregas vocais permanecem fechadas com mais pressão e estão mais tensas que o normal, pelo que apenas a parte anterior das pregas vocais permite a passagem de ar. A parte posterior não vibra ou vibra com uma frequência diferente da parte anterior. Neste caso, está-se na presença de uma voz crepitante. As alterações na voz são perceptíveis através da qualidade vocal com alterações no timbre da voz, nomeadamente através de uma frequência fundamental muito baixa. No limite, a glote pode mesmo permanecer fechada sem que ocorra vibração nem passagem de ar pelas pregas vocais na produção de fonemas vozeados.

A variação da tensão das pregas vocais pode ainda provocar uma voz rouca, que resulta no aumento da tensão laríngea e na irregularidade das vibrações das pregas vocais. Este tipo de voz caracteriza-se pelas alterações na frequência fundamental, tornando-a mais grave e sendo patente uma maior tensão e esforço ao falar [19].

2.3 Estado da arte

Existem poucas publicações que podem ser consideradas relevantes no reconhecimento de patologias da voz. A investigação no reconhecimento de patologias é relativamente recente tendo os primeiros avanços ocorrido no ano de 2000 [20]. Desde então existem apenas cerca de uma dezena de publicações que se referem ao reconhecimento de patologias.

Existem, no entanto, muito mais publicações que se referem à discriminação de vozes que sofrem de patologias, sem que seja reconhecida qual a patologia. Trata-se neste caso de reconhecimento de vozes patológicas. Neste caso, o sistema consiste num classificador binário, onde o sinal de teste é classificado como tendo ou não patologia. Estas publicações começaram a surgir na década de 60, com trabalhos que visavam sobretudo procurar características no sinal de fala que permitissem distinguir vozes patológicas de vozes normais através da análise de espectrogramas de vogais sustentadas.

Um dos motivos que leva à lenta evolução desta área deve-se à não existência de bases de dados (*corpora*) normalizadas e disponíveis que permitam a divulgação de sinais de fala para realização de testes e comparação de resultados. Mesmo as bases de dados existentes não contêm, em muitos casos, um número significativo de sinais para teste, pelo que alguns dos trabalhos realizados apenas avaliam a discriminação nos sinais de treino.

Não estão também exactamente definidos os termos utilizados no reconhecimento de patologias da voz e no reconhecimento de vozes patológicas. Quando não existe intenção de classificar qual a patologia e se pretende apenas saber se está na presença de uma voz patológica está-se na presença de um sistema de reconhecimento de vozes patológicas.

Algumas vezes o leitor pode ser iludido pelos títulos das publicações. Assim, na descrição deste capítulo, assume-se que dentro do reconhecimento de patologias, onde se pretende classificar se a voz tem uma determina patologia, existem dois subsistemas: sistemas de identificação de patologias e sistemas de verificação de patologias.

Os sistemas de identificação têm como objectivo identificar uma patologia de entre um conjunto de patologias e têm tipicamente múltiplas saídas (sistema de classificação *M*-ário), isto é, quando para um sinal de teste colocado à entrada este vai ser classificado numa determinada patologia. No entanto existem sistemas de identificação em que são implementados *M* sistemas de decisão binária onde são realizados testes entre pares de patologias. Nestes sistemas a patologia que se pretende identificar é comparada com *M* outras patologias, resultando em *M* sistemas.

Os sistemas de verificação de patologias assumem à partida que o sujeito poderá ter uma determinada patologia. Neste caso, o resultado obtido à saída do reconhecedor é comparado com um limiar e é verificado se o paciente tem determinada patologia.

Actualmente pode afirmar-se que a base de dados de referência nesta área é a base de dados da MEEI (*Massachusetts Eye and Ear Infirmary*) [21], pelo facto de serem encontradas um grande número de publicações com referência a esta base de dados, nomeadamente no reconhecimento de vozes patológicas. Contudo, assiste-se também ao surgimento de trabalhos onde os autores usam as suas próprias bases de dados. Este facto é compreensível devido à base de dados da MEEI não ser gratuita. Por outro lado, este programa é usado por alguns médicos especialistas da área e contem ficheiros em quantidade e qualidade superior ao disponibilizado por outras bases de dados. No entanto, mesmo os trabalhos realizados com a base de dados da MEEI não usam um critério igual de selecção de sinais de treino e teste sendo difícil comparar os trabalhos publicados.

2.3.1 Reconhecimento de vozes patológicas

Num individuo saudável, quando é produzida uma vogal sustentada e portanto com vibração das pregas vocais, esta vibração tem uma periodicidade e uma intensidade quase-constantes. As patologias da voz manifestam-se tipicamente através da disfonia, ou seja, uma perturbação na vibração das pregas vocais. O efeito é audível e pode ser caracterizado por uma frequência de vibração irregular das cordas, sendo este fenómeno designado por *jitter do período fundamental*. Se a perturbação for ao nível da intensidade do pulso glotal, este fenómeno denomina-se por *shimmer*. Podem também ocorrer ambos os fenómenos em conjunto. Em casos extremos pode mesmo ocorrer ausência de vibração das pregas vocais, denominando-se então por afonia.

Os primeiros trabalhos realizados nas décadas de 60 e 70 [22], [23], são exemplos de trabalhos onde se pretende quantificar os níveis de *jitter do período fundamental* e *shimmer* para vozes saudáveis e não saudáveis. Posteriormente foram sendo desenvolvidos vários sistemas [24]–[26] que tentavam optimizar a estimação destes parâmetros de modo a conseguir resultados de classificação mais fidedignos.

Outro parâmetro tipicamente usado na detecção de patologias da voz é a relação harmónica-ruído, HNR (*harmonics-to-noise ratio*) [27], proposto por Yumoto, que caracteriza a quantidade de ruído presente no espectro do sinal. Esta característica tem como propósito determinar quantitativamente qual o grau de disfonia, pois quanto maior for o ruído no espectro do sinal de fala maior será a ausência de actividade vocal.

Em [28], Murphy classifica os sinais de fala segundo 4 tipos de perturbação: *jitter do período fundamental*; *shimmer*; presença de ruído ao longo do espectro (HNR); e perturbações na forma de onda. Neste trabalho verifica-se que existe relação na presença de *jitter do período fundamental* e *shimmer* e é desenvolvido um algoritmo através da análise das harmónicas para remover este efeito na estimação do HNR. Em [29] este tema é de novo abordado, sendo desenvolvido um algoritmo que remove a influência do *shimmer* na estimação do HNR.

Em 2002, Dibazar [30] propõe um sistema para reconhecimento de vozes patológicas onde são usados pela primeira vez MFCC. Neste trabalho são implementados vários classificadores mas é com um classificador HMM que se obtém o melhor desempenho, usando MFCC combinados com a frequência fundamental. A taxa de reconhecimento é de 98,3% usando a base de dados da MEEI. Os MFCC foram comparados com características extraídas pelo Multidimensional Voice Program (MDVP) [21], incluído na base de dados da MEEI sendo que estas são maioritariamente características retiradas da fonte. Em [31] encontra-se a descrição detalhada das 33 características calculadas pelo MDVP. Estas características foram retiradas da vogal /a/, ao passo que os MFCC foram estimados com base na vogal /a/ e em fala contínua, ou seja, de vários fonemas. Na vogal /a/ os MFCC obtêm resultados semelhantes aos parâmetros do MDVP, ou seja, 97% de taxa de reconhecimento (sempre o mesmo termo). Quando os MFCC são combinados com o valor da frequência fundamental, a taxa de reconhecimento sobe para os 98,3%.

Também em [30] a fala contínua é utilizada no reconhecimento de vozes patológicas. Utilizando os parâmetros MFCC obtêm-se 97,5% de taxa de reconhecimento, independentemente da combinação com o valor da frequência fundamental. Tendo os MFCC informação apenas sobre o trato vocal, ou seja do filtro, este trabalho demonstra que se consegue fazer reconhecimento de vozes patológicas com base na informação do trato vocal. Em [32], também usando a base de dados da MEEI, a fala contínua foi utilizada para classificação de vozes patológicas. Neste caso apenas as zonas vozeadas foram consideradas, não tendo os resultados atingido os valores obtidos pela vogal /a/. As medidas avaliadas foram o *jitter* do período fundamental, o *shimmer*, o declive espectral, o HNR e análise da envolvente espectral. Os autores relatam que a vogal /a/ obtém melhores resultados quando se usam características glotais mas quanto usadas características do LPC os resultados são idênticos, com 96% de taxa de reconhecimento.

Em [33], é realizado um estudo com 223 oradores saudáveis, 472 sujeitos diagnosticados com nódulos e 472 diagnosticados com paralisia das pregas vocais. Foram comparados parâmetros glotais como o *jitter* do período fundamental, o *shimmer* e o HNR com características do trato vocal, valores e desvios nas frequências dos primeiros e segundo formantes. Os resultados obtidos mostram que as características glotais obtêm melhores resultados que as características do tracto vocal. Se ambas forem combinadas existirá uma melhoria de resultados. Os autores concluem que os formantes contêm informação que permite a classificação de vozes patológicas, considerando que existem desvios nos formantes produzidos pelos sujeitos com patologias que são resultado de compensações realizadas no tracto vocal de modo a contra balançar o mau funcionamento da glote.

Em [34] Sáenz-Lechón efectua um estudo comparativo entre vários trabalhos realizados até 2006, onde é possível verificar que até então os classificadores implementados eram baseados em redes neuronais e discriminadores lineares, e que as características usadas iam desde o *jitter* do período fundamental e o *shimmer* até aos cepstra, MFCC e LPC. Neste estudo comparativo podem de facto encontrar-se sistemas muito diferentes, desde trabalhos que usam apenas características relacionadas com a fonte, outros que usam apenas características relacionadas com o filtro e outros que usam ambas. Embora as taxas de reconhecimento destes sistemas variem entre 82,6% e 98,3%, os resultados apresentados são provenientes de bases de dados diferentes e mesmo no conjunto de trabalhos que usa a base de dados da MEEI os dados não são usados de forma a poderem ser comparados.

Em 2007, Fonseca e Pereira [35] propõem um método baseado na energia das *wavelets* [36] para reconhecimento de vozes patológicas. É calculado o RMS (*Root Mean Square*) em 4 sub-bandas do sinal representativas das frequências mais altas, verificando-se que nestas bandas os valores de RMS são mais elevados em vozes patológicas do que em vozes normais. Utilizando um SVM (*support vector machine*) os resultados obtidos são de 91,6% de taxa de reconhecimento (em apenas 12 sinais de teste, o que representa 11 acertos em 12) se forem usados os valores de RMS das duas primeiras bandas das *wavelets*. Isto demonstra que é possível fazer reconhecimento de vozes patológicas a partir do ruído existente nas altas frequências do sinal.

Também em 2007 Shama [36], avalia o HNR e a energia das sub-bandas do espectro do sinal para detectar vozes patológicas num subconjunto da base de dados da MEEI composto por 53

oradores saudáveis e 163 oradores com patologias laríngeas. O método é simples e usa metade dos dados para treino e outra metade para teste, sendo o classificador baseado no método do vizinho mais próximo. Os resultados obtidos rondam os 94% para o HNR e os 92% na análise das bandas do espectro do sinal. Os testes foram efectuados com a análise da vogal /a/.

Em publicações recentes [37], [38] o reconhecimento de vozes patológicas é realizado à custa de características espectrais, como por exemplo os MFCC, e os classificadores usados estão a migrar para modelos baseados em SVM e GMM. A principal vantagem dos parâmetros MFCC [39] relativamente às características da glote deve-se ao facto de não ser necessária a estimação da frequência fundamental. Devido às perturbações na voz das próprias patologias a estimação da frequência fundamental, assim como os parâmetros dele derivados, não é precisa. Em [31] é realizada uma comparação entre reconhecedores utilizando redes neuronais, vizinho mais próximo e SVM, conseguindo este último os melhores resultados.

2.3.2 Reconhecimento de patologias da voz

No ponto anterior, todos os trabalhos referidos consistiam no reconhecimento de vozes patológicas, sem existir intenção de determinar qual a patologia presente na voz do paciente, mas apenas se confirmando se este padece de uma qualquer patologia.

O primeiro trabalho a abordar o reconhecimento de patologias da voz foi publicado em 2000, onde Rosa e Pereira [20] realizaram um trabalho que consistiu na discriminação de patologias em que o resíduo do sinal de fala é calculado através do filtro de Kalman ou pelo filtro de Wiener. Do resíduo do sinal são extraídos sete parâmetros, entre os quais o *jitter* do período fundamental, que são usados no algoritmo Mann-Whitney para calcular medidas estatísticas. O reconhecimento é baseado num processo de discriminação entre pares de patologias, ou patologia/voz normal usando uma base de dados criada propositadamente para este estudo. Os parâmetros usados na discriminação são retirados de 3 vogais sustentadas, /a/, /e/ e /i/. Ao todo, foram testados 73 oradores, onde 25 são oradores saudáveis e 48 têm uma ou mais patologias, existindo no total 21 patologias diferentes. O processo de discriminação incide em 8 classes, sendo 6 patologias, uma classe para vozes normais e uma outra classe para as restantes patologias. O melhor resultado obtido foi uma discriminação de 54,79% para 231 combinações de pares de teste, com jitter do período fundamental como parâmetro de

discriminação, usando a vogal /e/. Apesar da taxa de reconhecimento ser relativamente baixa e a base de dados ter uma dimensão reduzida para o número de patologias que se pretende identificar, os autores assumem que o *jitter* do período fundamental pode discriminar patologias. Neste trabalho, o típico modelo de auto-regressão para calcular o sinal de resíduo não é usado porque este método é baseado num processo estacionário do trato vocal que não pode ser assumido numa vogal sustentada uma vez que existem perturbações, por exemplo, ao nível da língua. A verdade é que não comparam este método com outros que utilizam filtros de Wiener e de Kalman, pelo que não se sabe exactamente quais os benefícios desta solução. No entanto, o que acaba por limitar mais este trabalho é a dimensão da base de dados, pois é muito possível que com uma base de dados maior a taxa de discriminação fosse ainda mais baixa e o trabalho acabaria por não ter o mesmo impacto.

Em 2006, por Dibazar, o mesmo autor de [30], surge um trabalho [40] baseado em HMM e parâmetros MFCC que visa o reconhecimento de 5 patologias. Neste trabalho consegue-se uma taxa de identificação de cerca de 70% usando como princípio 5 classificadores, onde cada modelo HMM de cada patologia é comparado com um modelo HMM de todas as outras patologias, sendo o resultado final uma matriz de confusão, apenas para sinais da vogal sustentada /a/. Não está claro neste trabalho quantos sinais foram usados no treino e no teste e também não foi usada uma classe para classificação de vozes saudáveis, pelo que os resultados são validados assumindo que o indivíduo tem pelo menos uma das patologias. Contudo, pode-se concluir que é possível realizar a distinção de patologias com recurso aos parâmetros MFCC, tipicamente usados na caracterização do trato vocal em reconhecimento de fala e de orador.

Em 2009, Scalassara e Pereira [41] propõem a entropia como parâmetro para reconhecimento de patologias. Numa base de dados composta por 48 oradores, repartida de igual modo por 3 classes - ausência de patologia, nódulo das pregas vocais e edema de Reinke - a vogal sustentada /a/ foi adquirida para se realizarem os testes. A entropia é calculada em tramas de 50 ms, durante 1 segundo, tendo como objectivo caracterizar a incerteza do sinal. Este trabalho, descrito pelos autores como preliminar, não apresenta resultados mas apenas algumas conclusões. De facto, através da entropia é possível distinguir os sinais de oradores saudáveis dos sinais dos oradores com patologia, não sendo possível fazer a distinção entre diferentes patologias. A entropia apresenta valores superiores nos sinais de sujeitos com

patologias em relação aos sinais de sujeitos saudáveis, permitindo a total discriminação entre ambos. Os maiores valores de entropia nos sinais de sujeitos com patologias devem-se ao facto do sinal sofrer perturbações ao nível da frequência fundamental, como o *jitter* do período fundamental e o *shimmer*, que nos sinais provenientes de oradores não saudáveis têm valores mais elevados, contribuindo este facto para o aumento da incerteza dos sinais. Apesar dos autores caracterizarem as distribuições do *jitter* do período fundamental, do *shimmer* e ainda da amplitude do pulso glotal, estes não foram usados em qualquer tipo de classificação. No entanto, atendendo aos valores obtidos pelos autores, verifica-se que mais uma vez com estes três parâmetros talvez fosse possível realizar a discriminação entre alguns sinais de oradores saudáveis e sinais de oradores com patologias. Ainda assim, seria praticamente impossível realizar qualquer distinção entre patologias.

Também em 2008, Hosseini et al. [42] propõe o reconhecimento de 4 patologias (pólipos, Keratosis leukoplakia, disfonia espasmódica adutora e nódulos). São implementados 3 reconhecedores baseados em SVM onde os pólipos são treinados contra as restantes 3 patologias. As características usadas são obtidas a partir da entropia das *wavelets*, onde o sinal é separado em sub-bandas, sendo para cada uma calculada a entropia que é posteriormente usada na classificação. Contudo, não se conseguem tirar conclusões relevantes deste trabalho. Os autores não informam quantos oradores foram usados para teste, embora refiram que foram usados 75% para treino e o restante para teste, sendo conseguida uma taxa de reconhecimento de 87,5%, no melhor caso, o classificador de pólipo/nódulos. Nos outros dois casos, conseguem-se 82,5% e 81,81% para pólipos/adutor e pólipos/queratoses, respectivamente.

Em 2009, Fonseca e Pereira [43] foram usados *wavelets* no seguimento do trabalho [35], referenciado anteriormente. Neste caso, a abordagem é igual mas pretende-se classificar vozes não patológicas e duas patologias: nódulos e edema. O método usado em [35] consiste em calcular o RMS do ruído presente nas 4 sub-bandas mais altas do sinal. Os valores de RMS destas bandas são depois usados para treinar um SVM. São treinados vários SVM tendo como parâmetro o RMS de vozes normais, vozes com a patologia de nódulos e com edema. São realizados vários sistemas que avaliam vários valores de RMS, realizam a classificação de vozes sem patologias contra nódulos, contra edema e contra ambas e discriminam-se vozes com nódulos contra edemas. No total são implementados 13 sistemas de reconhecimento, mas

apenas 4 realizam a distinção entre nódulos e edemas. Baseado no trabalho anterior, verifica-se que é possível distinguir vozes normais de vozes com patologias. O problema surge quando se pretende distinguir entre patologias. Neste caso verifica-se que o valor de RMS das 4 sub-bandas usadas não permite fazer essa distinção pois os valores são semelhantes. Assim, os autores conseguiram fazer a distinção entre patologias com base no *jitter* do período fundamental dos sinais de mais baixa frequência filtrados nas *wavelets*, obtendo uma taxa de reconhecimento entre as duas patologias de 82,4%, para cerca de 12 testes. Num trabalho anteriormente referenciado [41], os mesmos autores não tinham conseguido usar o *jitter* do período fundamental para fazer distinção das mesmas patologias, sendo que neste caso o uso de *wavelets* e o cálculo do *jitter* do período fundamental em sub-bandas do sinal levou a bons resultados.

As *wavelets* foram usadas também em 2011 [44], onde um sinal de 1 segundo é decomposto em tramas de 24 ms sendo cada trama dividida em 4 sub-tramas, para garantir maior estabilidade do sinal. De seguida é aplicada uma *wavelet* a cada sub-trama e calculada a energia para cada banda da *wavelet*, que posteriormente é usada numa rede neuronal multicamada. Os melhores resultados são obtidos com a decomposição *wavelet* de 6ª ordem, obtendo-se uma taxa de reconhecimento de cerca de 90% entre vozes não patológicas, voz com edema de Reinke e nódulos. Uma quarta classe foi ainda usada para teste e treino referente a doentes com várias patologias da voz de origem neuronal, tendo a taxa de reconhecimento descido para 87%. Neste caso foram usadas *wavelets* até à 7ª ordem. Foram usados apenas 3 sinais de cada classe para treino, ou seja um total de 12 sinais de teste. Apesar do número muito pequeno de sinais, este é o primeiro trabalho onde foi efetivamente realizada a identificação de várias patologias, num total de 4 classes incluindo vozes sem patologias.

Em 2009 [45], Markaki usa a modulação espectral [46] e SVM na identificação de pólipos utilizando a vogal /a/ da base de dados da MEEI. O sistema consegue uma taxa de reconhecimento de patologias de 88,6%, entre 5 classes incluindo sujeitos saudáveis. As características são obtidas através de modulação espectral onde o espectro discreto do sinal é modelado em sub-bandas. Verifica-se que nos sinais obtidos por pacientes com pólipos existe maior energia na banda da frequência fundamental do sinal de fala. Depois da extracção de características é usado um algoritmo de redução de dimensionalidade [46] baseado em

decomposição em componentes principais. Por último, é aplicado um algoritmo de selecção de parâmetros baseado na informação mútua [47], onde são usadas para classificação apenas as características mais relevantes, sendo que este valor depende do classificador. Os autores referem que foram usados 83 casos com patologias, sendo 75% dos casos usados para treinar três classificadores que fazem discriminação entre pacientes com pólipos contra outras três patologias. Um outro classificador é usado para separar sujeitos saudáveis de sujeitos diagnosticados com pólipos.

Em 2011 [48], os mesmos autores comparam os resultados obtidos no trabalho anterior [45] com os obtidos tendo como parâmetro os MFCC. Os resultados obtidos por estes parâmetros são cerca de 25% inferiores aos resultados obtidos pela modulação espectral. De notar que nos MFCC não foi aplicado nenhum algoritmo de análise de componentes principais, uma vez que estes não são apresentados de forma matricial. São adicionadas mais duas patologias ao trabalho anterior, que são treinadas com SVM, perfazendo um total de sete sistemas com taxas de reconhecimento a variar entre 76% e 95%.

No mesmo ano é apresentado um trabalho que visa o reconhecimento de 5 patologias [49], não existindo uma classe para vozes não patológicas. São usadas duas vogais arábicas que equivalem às vogais /a/ e /i/, retiradas duma base de dados constituída por números de 1 a 10, ditos por 72 pacientes, gerando assim 720 sinais de fala. De cada vogal são extraídos o valor do 1º e 2º formante apenas de 4 ou 5 tramas da parte central do sinal da vogal (onde este é mais estável). Com estas quatro características e usando uma rede neuronal são conseguidas taxas de reconhecimento de 67,8% para pacientes masculinos e 52,5% para pacientes femininos num total de 720 sinais onde 80% são usados para treino e 20% usados para teste. Existem vários sinais com a mesma vogal por orador, uma vez que no total, a base de dados contém 51 oradores masculinos e 21 femininos. Não fica, no entanto, claro como foram repartidos os sinais de teste e treino. Nomeadamente, os autores afirmam que os 20% dos sinais de teste foram escolhidos aleatoriamente, sem revelarem se estes oradores foram incluídos no treino. Foram também realizados testes com um classificador baseado em quantificação vectorial mas os resultados foram inferiores aos obtidos pela rede neuronal. No entanto, os autores demonstram que nalgumas patologias existem desvios nos valores médios dos formantes em relação ao normal que aparentemente permitem a discriminação de patologias. Mas um estudo baseado apenas em médias não é muito conclusivo, pois é também relevante perceber

qual a variância dos dados, neste caso das frequências dos formantes. De facto, os valores médios dos formantes em algumas patologias estão muito próximos podendo este facto ser considerado também como uma característica do orador. Note-se que, se os sinais de treino e teste usados não são de oradores diferentes, pode-se estar acima de tudo a realizar reconhecimento de orador. Os autores também usaram um classificador vectorial com o qual obtiveram um resultado de 35% para os oradores masculinos e de 28% para os oradores femininos, o que representa cerca de metade da qualidade dos resultados obtidos pela rede neuronal.

Neste último trabalho, apesar das baixas taxas de reconhecimento, comparadas por exemplo com os dois trabalhos anteriores [45] e [48], foi implementado um único classificador que fez a discriminação de várias patologias. Em grande parte dos trabalhos anteriores, à excepção do trabalho [44], também implementado com redes neuronais, todos os outros realizam reconhecimento através de vários classificadores treinados com pares de patologias. Uma vez que nem sempre são realizadas todas as combinações entre patologias, os resultados obtidos não têm normalmente em conta todas as combinações possíveis.

Comparando diretamente os trabalhos [44] e [49] verifica-se que o primeiro tem uma taxa de reconhecimento de 90% e o segundo 60%. Contudo, o primeiro usa 12 sinais de teste e o segundo usa 144, pelo que nem estes resultados são diretamente comparáveis.

No decorrer desta investigação, já quando esta está num estado avançado são publicados mais dois trabalhos, [50], [51] no reconhecimento de patologias da voz. Estes trabalhos, dado avançado estado da investigação, não foram tidos em conta nas direcções a seguir na investigação que serão expostas nas conclusões do último capítulo. Contudo os seus resultados e métodos são de seguida descritos.

Em Setembro de 2014 é publicado um trabalho [50] que foca o reconhecimento de patologias com base nas características de áudio da norma MPEG-7 [52]. Neste trabalho os autores usam os mesmos dados que no trabalho [48], ou seja o sinal de fala é a vogal /a/, sendo escolhidos 53 sujeitos saudáveis e 173 sujeitos com patologias, dividido por cinco patologias, treinadas aos pares, resultando sete sistemas SVM diferentes. As características retiradas da norma MPEG-7 são o *Audio Spectrum Spread*, *Audio Spectral Flatness*, *Audio Spectral Centroid* e o *Harmonic Spectral Spread*. Os autores clamam taxas de 99.9% nos sete reconhecedores, e

comparam os seus resultados com os apresentados em [48]. Apesar dos bons resultados note-se que os autores usam 10% dos dados para teste, e no trabalho anterior são usados 25% dos dados para teste. Por este motivo devem existir reservas sobre as efectivas melhorias dos métodos apresentados.

O último trabalho conhecido no reconhecimento de patologias da voz foi apresentado em 2015. Em [51], os autores apresentam um sistema baseado em GMM, com uma base de dados semelhante a [50]. Os parâmetros do sinal de fala são o *Auditory processed spectrum* e *ALL-pole model based cepstral coefficients* [51]. O sistema apresentado permite distinguir cinco patologias usando fala contínua. As taxas de reconhecimento variam entre 82% e 94% para este sinal de fala, sendo em qualquer patologia e em qualquer um dos reconhecedores superiores às obtidas pela vogal /a/. Os autores referem também que existe pouca investigação no reconhecimento de patologias da voz, não sendo encontrado nessa publicação nenhum trabalho relevante que não esteja descrito neste estado da arte. De referir que em [51] é atribuída a [8], uma contribuição desta tese, a primeira implementação de um sistema baseado em fala continua na investigação de reconhecimento de identificação de patologias da voz através do processamento de sinais de fala.

2.3.3 Sistemas de reconhecimento de orador

Neste ponto é realizada uma breve descrição de sistemas de classificação utilizados no reconhecimento de orador, pois verifica-se que muitos dos sistemas implementados nesta área migram para outras, nomeadamente para o reconhecimento de vozes patológicas e para o reconhecimento de patologias da voz.

Desde a publicação de Reynolds [53] em 2000 que os sistemas utilizando UBM-GMM se encontram na linha da frente do reconhecimento de orador. Até então, já eram usados modelos de misturas gaussianas no reconhecimento de orador [54], mas estes eram gerados independentemente do modelo do mundo. Nesta nova abordagem, os modelos são gerados através da adaptação do modelo do mundo por um processo bayesiano [55] provocando uma melhoria no reconhecimento, assim como a utilização de um método computacionalmente menos exigente na avaliação [56].

Paralelamente, também em 2000, Campbell propôs a utilização de SVM na área de reconhecimento de orador [57]. Os SVM estavam então a ganhar popularidade nas mais

diversas aplicações. Sendo classificadores que discriminam duas classes, podem ser aplicados na verificação de orador para separar o modelo do orador dos impostores. No entanto, só posteriormente os sistemas baseados em SVM começaram a ganhar expressão [58]–[60].

Em [60], Campbell introduziu o '*generalized linear discriminant sequence kernel*' (kernel GLDS) e explicou que para usar um '*sequence Kernel*' são necessárias duas frases: uma para treinar o modelo e outra com que será realizado o teste, o qual traduzirá as medidas de semelhança entre as frases.

Em [61] são desenvolvidas métricas para os GMM, que são usadas posteriormente no reconhecimento com SVM [58]. Nesta publicação, Campbell usa a divergência Kullback-Liebler para definir um novo '*sequence kernel*', o 'GMM *supervector linear kernel*' (kernel GSL), baseado em super-vectores obtidos através de sistemas UBM-GMM. Estes super-vectores lineares são obtidos através das adaptações das médias de uma frase de treino, no UBM-GMM.

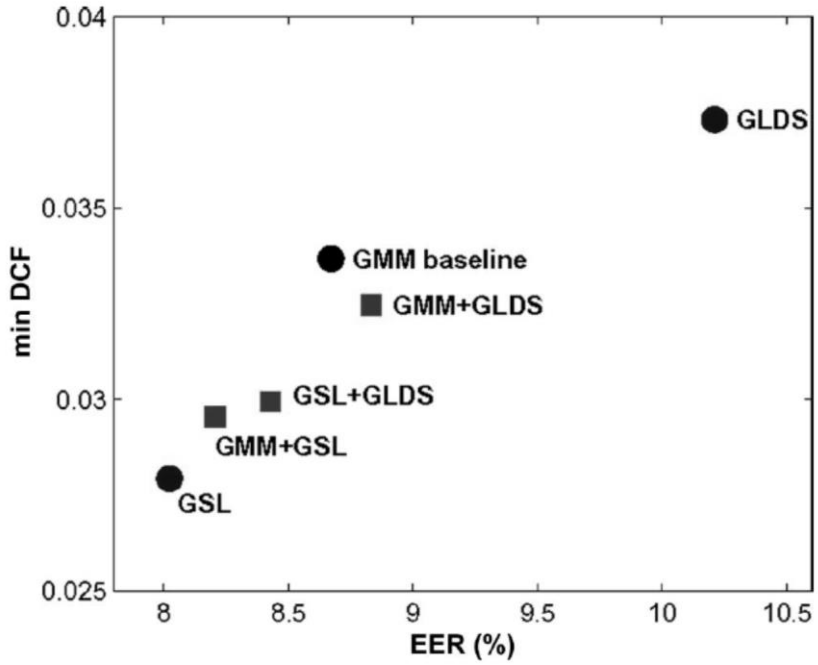

Figura 2-5 Comparação de vários sistemas de reconhecimento.
A base de dados usada foi a NIST 2006 SRE (Retirado de [62]).

Em [62], foram comparados três sistemas: um sistema clássico UBM-GMM que serviu como referência de base; um sistema SVM GLDS e um sistema SVM GLS. Os testes foram conduzidos sob as mesmas condições na plataforma ALIZE [61] e foram também efectuadas fusões entre

os vários sistemas. Os resultados são mostrados na Figura 2-5, onde se verifica que o sistema SVM GSL, que é treinado com os vectores dos GMM, obtidos através da adaptação do UBM, obtém os melhores resultados ao nível do EER (*Equal Error Rate*).

2.4 Conclusões

No início deste capítulo apresentou-se uma descrição geral das áreas de processamento da fala. Na secção seguinte apresentou-se o aparelho vocal assim como as patologias da voz mais comuns que podem ser encontradas na bibliografia. Na secção anterior apresentou-se o estado da arte descrevendo as publicações consideradas relevantes no reconhecimento de vozes patológicas e de patologias da fala. Apresenta-se agora uma análise das limitações e lacunas encontradas no estado da arte e definem-se os caminhos seguidos na investigação.

O único ponto em comum em todos os trabalhos apresentados sobre reconhecimento de vozes patológicas e de patologias é o facto do sinal de fala usado nos vários sistemas ser a vogal sustentada /a/. Em [63], é realizado um estudo onde se verifica que a vogal /a/ é aquela que tem, no domínio do tempo, os picos de maior amplitude e de menor abertura, o que faz com que nesta vogal exista uma boa correlação entre os parâmetros retirados de uma eletroglotografia laríngea e os parâmetros acústicos.

Apenas em [20] e [41] foram realizados sistemas utilizando características da fonte, como por exemplo o *jitter* do período fundamental, no reconhecimento de patologias. Em [20], o trabalho foi mais ambicioso pois poderia realizar a discriminação de 8 classes, entre as quais 6 patologias, mas não existiu uma base de dados consistente que permitisse validar os resultados. Em [41], fazer o reconhecimento de patologias através de *jitter* do período fundamental não era o principal objectivo pois não são publicados resultados, mas é evidente através dos valores médios do *jitter* do período fundamental e da sua variância que não é possível distinguir as duas patologias em estudo.

Os restantes trabalhos propõem-se realizar reconhecimento de patologias através de características que modelam o trato vocal (MFCC, modulação espectral e formantes). Nestes e nomeadamente em [34] e [37], dos mesmos autores, foram obtidos resultados acima de 85% na detecção de pólipos contra outras três patologias. Em [48] os resultados obtidos pela

modulação espectral são comparados com os obtidos com parâmetros MFCC, concluindo-se que os MFCC obtêm pior desempenho. Os autores, ao usarem modulação espectral, procuram correlacionar o espectro do sinal com a frequência fundamental. Os MFCC de baixa ordem apenas contêm informação do trato vocal, mas ainda assim é possível realizar reconhecimento de patologias. Por outro lado, em [49], são usados apenas os 1º e 2º formantes. Os autores clamam que conseguem identificar 5 patologias, usando uma base de dados que pode ser considerada relevante dada o contexto, pois permite a realização de 144 testes.

O reconhecimento de vozes patológicas começa assim por ser efectuado com características que estão estritamente ligadas às pregas vocais, como sejam o *jitter* do período fundamental e o *shimmer*, embora recentemente as características ligadas ao modelo do filtro, ou seja do trato vocal, apareçam com resultados equivalentes. A combinação de vários parâmetros é também usual. Ao nível dos classificadores são visíveis que os progressos efectuados noutras áreas, como por exemplo no reconhecimento de orador, estão também a optimizar o reconhecimento de vozes patológicas, contribuindo para melhorar os resultados.

A investigação no reconhecimento de patologias da voz está portanto numa fase inicial. Os pressupostos que são usados no reconhecimento de vozes patológicas, que usam características da fonte, não mostram resultados conclusivos no reconhecimento de patologias da voz. Os sistemas que usam características que modelam o trato vocal conseguem alguns resultados relevantes. Em alguns casos os autores não relacionam nem discutem os resultados obtidos e não procuram encontrar características no sinal de fala que estejam relacionadas com determinada patologia. Este facto é de extrema importância, pois a discussão de resultados permite, muitas vezes, perspectivar outras soluções.

Os resultados obtidos, na grande parte dos trabalhos, não são comparáveis pois não existe uma base de dados de referência e, mesmo nos casos em que é usada a mesma base de dados, não são usados exactamente os mesmos sinais e as mesmas patologias.

Em todos os trabalhos de investigação encontrados é usada sempre a vogal /a/, ou semelhante, nos sistemas de reconhecimento de patologias da voz. Não existem, no entanto, até à dada, sistemas de reconhecimento de patologias da voz que utilizem fala contínua. Existe contudo um trabalho de reconhecimento de vozes patológicas em que a fala contínua obteve uma taxa de reconhecimento semelhante à da vogal /a/.

O facto de os resultados não serem comparáveis torna difícil estabelecer um ponto de partida que possa ser tomado como referencia, nomeadamente ao nível de reconhecimento de patologias da voz. No entanto, através da revisão do estado da arte, podem ser encontradas algumas lacunas que ainda não foram devidamente exploradas. É exemplo a utilização de fala continua no reconhecimento de patologias da voz. Neste caso, a única base de dados conhecida que contém sinais que permitam esta investigação é a base de dados da MEEI. Uma vez que os resultados obtidos no reconhecimento de vozes patológicas foram semelhantes à vogal /a/ é espectável que o sinal de fala continua contenha informação relevante no reconhecimento de patologias.

Novas abordagens no estudo de vozes patológicas também devem continuar a ser investigadas para que possam transmitir métodos simples, mas acima de tudo eficazes no reconhecimento de vozes patológicas. Verifica-se que novos métodos desenvolvidos no reconhecimento de vozes patológicas contribuem para o posterior desenvolvimento no reconhecimento de patologias da voz.

3 Materiais e métodos

Neste capítulo vão ser descritas as bases de dados usadas para estimar as características dos sinais de fala que permitirão desenvolver os classificadores usados na implementação dos sistemas de reconhecimento. São também descritos os parâmetros que vão ser usados quer na detecção de vozes com patologias quer na detecção das próprias patologias, assim como os classificadores usados na análise dos parâmetros dos sinais de fala e na implementação dos sistemas de reconhecimento.

3.1 Bases de dados

No desenvolvimento deste trabalho foram usadas duas bases de dados: a base de dados da Universidade de São Paulo e a MEEI. Ainda numa fase inicial do trabalho foi utilizada a base de dados COPAS, que contém principalmente sujeitos com patologias associadas à produção de fala, como por exemplo fenda do palato, disartria, desordens articulares entre outras e um número reduzido de sujeitos diagnosticados com patologias da voz.

É comum, em aplicações de processamento de fala, que a frequência de amostragem dos sinais seja 8 kHz, denominada banda telefónica. No reconhecimento de vozes patológicas, contudo, os sinais são amostrados com frequências que podem ir até 50 kHz. O aumento da frequência de amostragem permite uma maior resolução na frequência e a estimação mais precisa de parâmetros como a frequência fundamental, o *jitter* do período fundamental e o *shimmer*.

3.1.1 Base de dados COPAS

A *Dutch Corpus of Pathological and Normal Speech* (Copas) foi criada para o desenvolvimento de algoritmos associados a patologias da fala. Esta base de dados é composta por 319 sujeitos, dos quais 122 são saudáveis. Dos sujeitos não saudáveis, apenas 7 contêm uma patologia da voz, a disfonia espasmódica. Os restantes são afectados com patologias da fala.

Os ficheiros desta base de dados foram adquiridos com uma frequência de amostragem de 16 kHz em PCM (*Pulse Code Modulation*) uniforme com 16 bits por amostra. Nos sinais de fala adquiridos, constam vogais sustentadas, fala contínua espontânea, semi-espontânea e lida.

Esta base de dados serviu essencialmente para implementação, numa fase inicial do trabalho, de algoritmos de estimação de formantes, frequência fundamental e *jitter* do período fundamental, usados posteriormente em diversas fases do trabalho.

3.1.2 Base de dados da Universidade de São Paulo

Os sinais desta base de dados são parte da base de dados do Grupo de Bioengenharia da Escola de Engenharia da Universidade de São Paulo, Brasil. A Base de dados é constituída por 47 sujeitos, repartidos por 3 classes: saudáveis (16), diagnosticados com edema de Reinke (16) e diagnosticados com nódulos (15). A base de dados contém apenas sujeitos diagnosticados com uma patologia, tendo o diagnóstico sido realizado através de laringoscopia.

Os sinais desta base de dados foram adquiridos com uma frequência de amostragem de 22050 Hz em PCM uniforme e quantificados com 16 bits por amostra. Os oradores produziram a vogal /a/ com um nível confortável de amplitude, durante 5 segundos. Foram adquiridos vários sinais e seleccionados os mais estáveis (com menos variação vocal). Para os sujeitos diagnosticados com patologias também foram adquiridas as vogais /e/ e /i/.

3.1.3 Base de dados MEEI

A base de dados da *Massachusetts Eye and Ear Infirmary* (MEEI) é uma base de dados de referência na investigação de vozes patológicas. Contém sinais de fala de 53 sujeitos saudáveis e 724 sujeitos com patologias da voz. Dos sujeitos com patologias apenas 477 têm informação das patologias diagnosticadas através de laringoscopia. Muitos dos sujeitos são diagnosticados com mais que uma patologia sendo frequente encontrar sujeitos com 5 patologias.

Os sinais foram adquiridos com frequências de amostragem que vão desde os 10 kHz aos 50 kHz e todos quantificados com 16 bits. Os sinais gravados foram a vogal /a/ e a passagem ”rainbow” com fala contínua lida.

3.2 Parâmetros do sinal de Fala

A extracção de parâmetros de um sinal tem como objectivo uma nova representação, noutro espaço, normalmente de menor dimensão. Esta transformação pode ter várias motivações do ponto de vista da aplicação, mas em reconhecimento deverá permitir a separação dos sinais nas classes pretendidas.

As características retiradas do sinal de fala podem modelar a fonte (glote) ou o filtro (tracto vocal). Na modulação da glote são usadas tipicamente a frequência fundamental, que traduz a frequência fundamental de vibração das pregas vocais, o *jitter* do período fundamental, que traduz os desvios de frequência das pregas vocais em períodos consecutivos e a relação harmónica-ruído (HNR) que mede a relação sinal ruído no sinal periódico produzido pela glote.

Na modulação do tracto vocal, as representações mais comuns são por os parâmetros MFCC (*mel-frequency cepstrum coefficient*), PLP (*perceptual linear prediction*), MLPCC (*mel-frequencies linear prediction cepstrum coefficient*), LSF (line spectral frequencies) e coeficientes LP (*linear prediction*). Em particular, nas aplicações de reconhecimento de fala são usados tipicamente os parâmetros MFCC, MLPCC e PLP. Qualquer um destes parâmetros tem em conta informação perceptiva. No caso dos parâmetros MFCC e MLPCC através da introdução de um filtro de escala Mel; no caso dos PLP, através de um filtro de escala de Bark. Os filtros têm como objectivo a modelação do mecanismo do ouvido humano, nomeadamente da membrana basilar.

3.2.1 Parâmetros da glote (fonte)

Neste ponto vão ser descritos alguns dos parâmetros usados nesta tese que estão relacionados com medidas glotais. Serão apresentadas algumas equações que são consideradas relevantes uma vez que foram implementadas e usadas ao logo do trabalho.

3.2.1.1 Frequência fundamental e Jitter do período fundamental

A frequência fundamental mede a frequência de vibração das pregas vocais e consequentemente o período fundamental. O *jitter* do período fundamental mede desvios na frequência entre períodos consecutivos. Estas duas características permitem caracterizar a vibração das pregas vocais.

Existem vários métodos para estimação da frequência fundamental. O mais comum é a estimação do primeiro pico da autocorrelação normalizada do sinal ou mesmo do resíduo de predição. No entanto, a estimação do *jitter* do período fundamental requer uma estimação mais precisa de período glotal fundamental. O método vulgarmente utilizado é baseado no algoritmo DSYPA (dynamic programming project phase-slope algorithm)[64]. Este algoritmo tem como finalidade a estimação dos instantes de abertura e fecho da glote (GCI - *glottal*

closure instant). No caso concreto do *jitter* do período fundamental este é estimado através das variações do período glotal fundamental calculado nos instantes de fecho da glote. Estes instantes são mais fáceis de identificar do que os instantes de abertura, uma vez que o fecho das pregas vocais é abrupto e provoca uma descontinuidade na passagem do ar pela glote [26]. De facto, através destes instantes, também pode ser estimado o período de frequência fundamental.

$$Pitch\ \ Jitter_{local} = 100 * \frac{\sum_{i=2}^{N} \left| P_i - P_{i-1} \right| / (N-1)}{\sum_{i=1}^{N} \left| P_i \right| / (N)} \quad (3.1)$$

A equação 3.1. mostra a estimação do Jitter do período fundamental, sendo *N* o número de períodos fundamentais nos quais foi estimado o jitter do período fundamental e P_i o período da frequência fundamental sendo este estimado com os valores CGI.

3.2.1.2 HNR

A relação harmónica-ruído HNR (*harmonics-to-noise-Ratio*) dá informação sobre a quantidade de ruído presente num sinal harmónico. Quando é gerado um som vozeado, o sinal contem as harmónicas da frequência fundamental com energia da componente periódica do sinal. Entre estas harmónicas, idealmente no caso de uma voz saudável, não deveria existir sinal relevante, mas no caso de uma voz soporosa típica de vozes patológicas vai existir energia proveniente do ruído causado pelo desempenho deficiente das pregas vocais, nomeadamente pela incapacidade destas fecharem completamente e consequentemente permitirem a passagem de ar. Assim, em vozes patologias o valor do HNR será tipicamente mais baixo que em vozes saudáveis.

Existem dois métodos típicos para estimar o HNR. O primeiro [65] assume que o ruído é aditivo, e envolve análise espectral que estima a energia das harmónicas e o ruído no intervalo entre harmónicas. Este método tem um custo computacional elevado uma vez que tem de se estimar o espectro do sinal no longo termo para que o espectro contenha resolução suficiente para permitir a estimação do ruído entre as harmónicas. Este processo é repetido de modo a obter medidas de HNR mais exactas. A equação 3.2 mostra como é estimado o HNR pelo

método da análise espectral, sendo o numerador a energia das harmónicas dada por $E_{s(k)}$ e o denominador a energia do ruído dada por $E_{w(k)}$.

$$HNR_{dB} = 10log_{10}\frac{E_{s(k)}}{E_{w(k)}} \quad (3.2)$$

O segundo método [66] apresenta uma solução baseada na autocorrelação. Neste caso a estimação do ruído tem em conta o valor do primeiro pico da autocorrelação normalizada $r'_x(\tau_{max})$, equação 3.3. No domínio do tempo, o valor deste 1º pico é vulgarmente usado para estimação da frequência fundamental. Como se pode verificar pela equação 3.4, o HNR é estimado pelo quociente entre o valor do primeiro pico da autocorrelação normalizada e a sua diferença para 1, que estima o ruído.

$$r'_x(\tau_{max}) = \frac{r(T_0)}{r(0)} \quad (3.3)$$

$$HNR_{dB} = 10log_{10}\frac{r'_x(\tau_{max})}{1-r'_x(\tau_{max})} \quad (3.4)$$

$$RPPC = r'_x(\tau_{max}) \quad (3.5)$$

Assume-se portanto que quanto mais próximo este valor for de 1 maior é o valor de HNR. Na verdade se este valor for 1 significará que o sinal é periódico e valor do ruído será zero. Os autores afirmam que este método é mais eficiente que o método de análise espectral. No entanto, o valor máximo do pico da autocorrelação também poderá variar com a periodicidade do sinal, sendo que este primeiro pico será tão mais elevado quanto menor for a aperiodicidade do sinal. Assim, existe também uma medida indirecta relativa ao *jitter* do período fundamental.

Como se pode verificar na equação 3.4, o ruído estimado no denominador é o complemento da componente periódica do sinal. Deste modo, faz sentido que apenas se possa usar o valor da componente periódica do sinal. Este valor, designado por RPPC (*Relative Power of the Periodic Component*), foi introduzido em [6] e demonstrou obter resultados relevantes na detecção de vozes patológicas, sendo dado pelo numerador da expressão 3.4.

3.2.2 Parâmetros de trato vocal (filtro)

Existem vários parâmetros que modelam o tracto vocal. Estes são por exemplo os MFCC, usados tipicamente em aplicações de reconhecimento de fala, e os LSF tipicamente usados em aplicação de codificação de fala. Os MLSF são coeficientes obtidos através dos LSF através da introdução da informação perceptual, tendo sido estes parâmetros propostos em reconhecimento de orador. Estes três parâmetros foram usados como características dos sinais de fala em vários sistemas implementados nesta tese. Nos próximos pontos são descritas as suas principais características.

3.2.2.1 Parâmetros MFCC

Os coeficientes cepstra são obtidos através da transformada inversa de Fourier do logaritmo do espectro do sinal. Os parâmetros cepstra têm como particularidade conseguirem separar a excitação do trato vocal, ou seja, a fonte do filtro. É precisamente no filtro (trato vocal), correspondente aos primeiros coeficientes cepstra, que existem mais particularidades que permitem a distinção entre oradores.

Os coeficientes mel-cepstra (MFCC) [67] são derivados dos coeficientes cepstra e surgem com a introdução de informação perceptiva, através da filtragem do espectro do sinal com um banco de filtros de escala Mel. Este banco de filtros tem como objectivo a modelação das bandas críticas da membrana basilar. A sensibilidade da membrana basilar é maior nas baixas frequências, onde até cerca de 1 kHz as bandas críticas estão espaçadas linearmente e o ouvido humano consegue ter uma maior resolução de frequências. A partir de 1 kHz o espaçamento das bandas críticas é aproximadamente logarítmico e começa a ser perdida a resolução na frequência.

Ao nível do espectro, o banco de filtros de escala Mel provoca uma suavização do espectro original onde prevalecem os principais detalhes que devem ser considerados relevantes no reconhecimento. A outra diferença reside no facto de usar a transformada inversa do cosseno (IDCT) em vez da transformada inversa de Fourier. A transformada de Fourier produz coeficientes complexos, mas a IDCT apenas utiliza números reais, o que torna o algoritmo de extracção dos parâmetros MFCC computacionalmente mais eficiente.

Os MFCC são actualmente os parâmetros mais usados em todo o tipo de aplicações de reconhecimento de fala e orador.

3.2.2.2 Parâmetros LSF

A transformação dos coeficientes de predição linear em parâmetros LSF foi introduzida por Itakura em 1975 [68], tendo as suas propriedades sido mais tarde estudadas por Soong e Juang [69]. Por definição, os parâmetros LSF são as frequências correspondentes às raízes de dois polinómios de ordem p+1, $P(z)$ e $Q(z)$, derivados do filtro inverso de predição linear $A(z)$, de ordem p (sendo p a ordem de predição). $P(z)$ corresponde ao trato vocal com a fonte glotal completamente fechada (coeficiente de reflexão k_{p+1}=1) e $Q(z)$ representa o trato vocal com a fonte glotal completamente aberta (coeficiente de reflexão k_{p+1}=-1).

A conversão de coeficientes de predição linear em parâmetros LSF converte cada raiz de $A(z)$ num par de raízes complexas conjugadas no círculo unitário. Se o espectro de entrada for plano, os parâmetros LSF estão separados uniformemente entre 0 e $F_s/2$. Se uma raiz de $A(z)$ apresentar um valor do seu módulo perto da unidade, a largura de banda é estreita sendo muito provável que este pólo corresponda a um formante. Pelo contrário, se uma raiz de $A(z)$ apresentar um valor de módulo baixo, será grande a largura de banda correspondente e a sua contribuição traduzir-se-á apenas na inclinação espectral, estando as raízes correspondentes de $P(z)$ e $Q(z)$ afastadas. Deste modo é possível relacionar a distância entre duas raízes consecutivas, $P(z)$ e $Q(z)$, com a largura de banda dos formantes, sendo também esta uma característica do orador.

3.2.2.3 Parâmetros MLSF

Os parâmetros LSF são tipicamente usados em codificação de sinais de fala [70]. Meneses e Trancoso [71] demonstraram que estes parâmetros podem também ter um bom desempenho na adaptação ao orador no contexto da codificação fonética.

Os parâmetros LSF não são utilizados em aplicações de reconhecimento de orador pois uma vez testados não demonstraram um desempenho convincente [72]. Estes parâmetros não têm em conta a informação perceptiva e este facto é uma possível razão para o seu menor desempenho. Em [73], Cordeiro e Meneses introduziram nos parâmetros LSF a informação

perceptiva, originando parâmetros Mel-LSF, ou MLSF, obtendo-se resultados semelhantes aos MFCC em sistemas de reconhecimento de orador.

Como foi referido no ponto anterior, os LSF são obtidos através da transformação dos coeficientes do filtro de predição linear, podendo estes ser obtidos através da autocorrelação do sinal. Para introduzir o filtro perceptivo, a autocorrelação é calculada através da transformada inversa do espectro de Mel, como apresentado no fluxograma da Figura 3-1, obtendo-se assim os coeficientes MLSF.

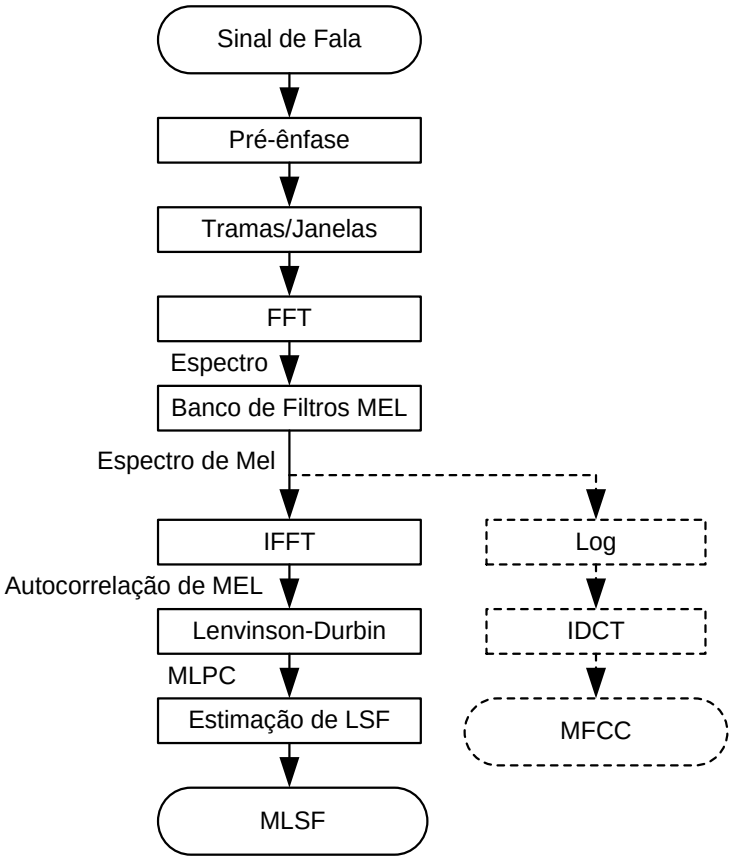

Figura 3-1 Estimação dos parâmetros MLSF *vs*. MFCC

Os parâmetros MLSF têm as mesmas características, ao nível da representação, dos LSF. Como apresentado na Figura 3-2 estão dispostos sobre o círculo unitário. Esta propriedade permite usufruir das vantagens oferecidas na quantificação dos LSF de modo a reconhecer oradores remotamente.

Também as diferenças entre os pólos dos LSF (ou MLSF) em tramas consecutivas contêm informação do orador, pois estão relacionadas com variações ao longo do tempo na largura de

banda dos formantes. A partir destas diferenças, foram criados outros parâmetros, os DMLSF [73] e foi desenvolvido trabalho no sentido de avaliar o seu potencial em reconhecimento do orador, verificando-se que nesta área o seu desempenho foi melhor que os LSF obtendo resultados semelhantes aos MFCC.

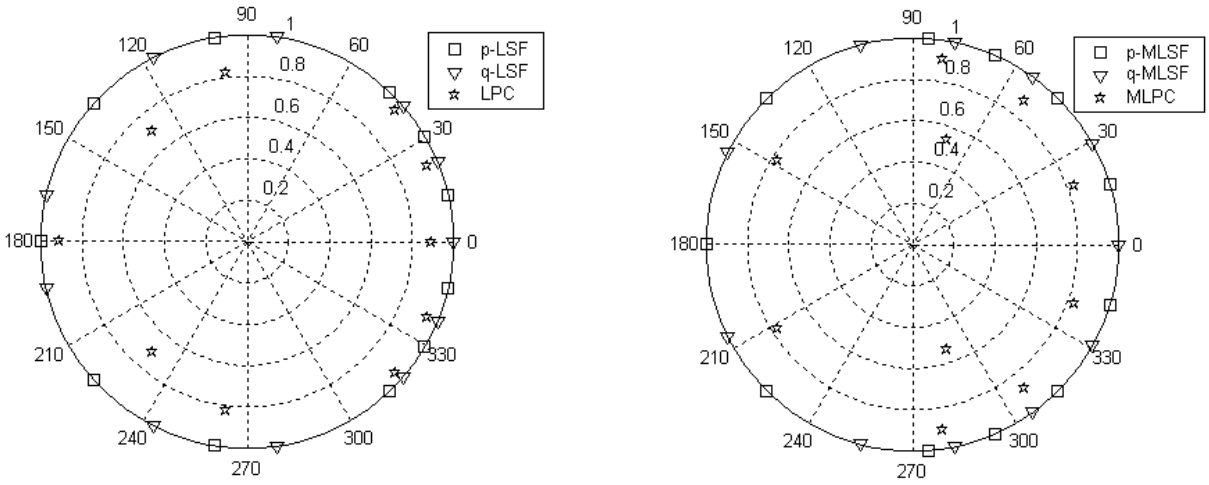

Figura 3-2 Parâmetros LSF *vs*. MLSF. À esquerda, coeficientes LSF e respectivos pólos do filtro LPC. À direita coeficientes MLSF e respectivos pólos do filtro MLPC. Os resultados correspondem à mesma trama do sinal.

Os parâmetros MLSF são assim uma derivação dos parâmetros LSF, que contêm informação directa do tracto vocal ao nível dos formantes e têm em conta a informação perceptiva, representativa das propriedades do ouvido humano, através da introdução dos filtros da escala MEL.

3.3 Classificadores

Nesta secção apresenta-se uma descrição dos classificadores usados nesta investigação: *Support Vector Machines* (SVM); *Discriminant Analysis* (DA); *Gaussian Mixture Models* (GMM); e árvores de decisão. Estes classificadores vão ser usados na classificação de sinais de fala, quer seja no reconhecimento de vozes patológicas quer seja no reconhecimento de patologias da fala.

3.3.1 *Support vector machines*

Os SVM são classificadores binários, isto é, apenas podem classificar dados numa de duas classes, C_1 ou C_2. Para um conjunto de treino *T* tem-se que:

$$\mathrm{T} = \{(X_1, y_1), (X_2, y_2), \ldots, (X_m, y_m)\}, \text{ onde}$$

$$y_k = 1 \; se \; X_k \in C_1, y_k = -1 \; se \; X_k \in C_2 , \qquad (3.6)$$

ou seja, os dados das classes C_i são mapeados num novo espaço, onde os dados estão explicitamente classificados numa das classes através do valor de *Y*. Assumindo também que estas classes são separáveis, cabe ao SVM encontrar um hiperplano que as separe. Existem vários tipos de classificadores lineares que permitem encontrar diferentes planos de separação para o mesmo conjunto de treino, como mostra a Figura 3-3. Pelo facto de existirem diferentes soluções, a escolha de uma que não seja a mais adequada pode comprometer o desempenho do classificador na fase de teste. Neste contexto, o SVM surge com uma forma de encontrar um hiperplano baseado no conceito de máxima margem, sendo escolhido o plano entre os dados que maximize a distância entre as duas classes.

O classificador de máxima margem é usado tipicamente quando num espaço de ordem *N* um classificador linear não tem uma generalização razoável, isto é, apesar de classificar bem o conjunto de treino depois não consegue generalizar eficientemente para o conjunto de teste.

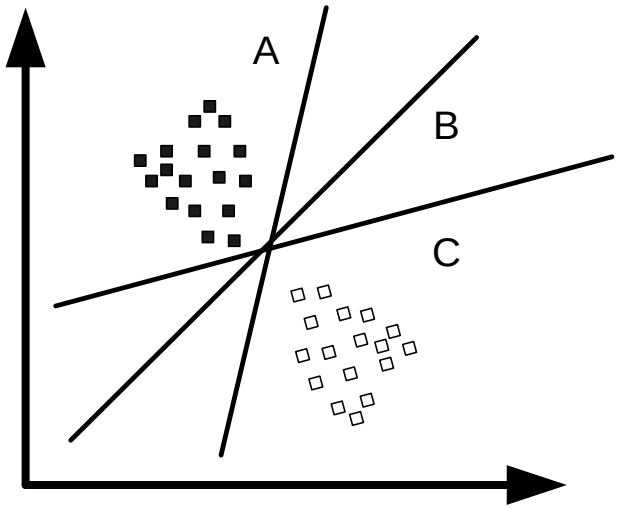

Figura 3-3 Exemplos da separação de duas classes por classificadores lineares.

3.3.1.1 Margens

O modo como é obtida a margem de classificação consiste em encontrar um ou mais pontos de cada classe que forneçam o suporte para os hiperplanos, sendo esta a origem do nome *suport vectors*. Os vectores de suporte estão colocados nas fronteiras de cada classe (Figura 3-4) e são representados pelas seguintes equações:

$$\begin{aligned} w\cdot x+b&=-1\\ w\cdot x+b&=+1 \end{aligned}. \quad (3.7)$$

Equidistante dos hiperplanos que definem as fronteiras é definido o hiperplano de decisão, descrito pela seguinte equação:

$$w\cdot x+b=0. \quad (3.8)$$

A classificação na classe C_1 é considerada correcta, se:

$$y_i\left(w\cdot x+b\right)>0. \quad (3.9)$$

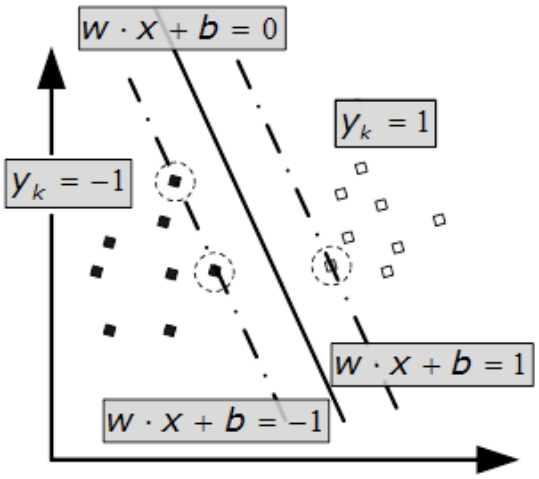

Figura 3-4 - Exemplo da separação de duas classes por um classificador de máxima margem.

A maximização da margem é conseguida através da utilização dos multiplicadores de Lagrange α_i. Verifica-se em [74], que após a minimização em ordem a w, se obtém:

$$w = \sum_{i=1}^{m} \alpha_i y_i X_i \,. \qquad (3.11)$$

A equação de classificação é dada por:

$$\sum_{i=1}^{m} y_i \alpha_i \left(X_i \cdot X \right) + b > 0 \,, \qquad (3.12)$$

onde α_i são 0, à excepção dos vectores de suporte X_i. Basicamente, apenas os vectores de suporte, ou seja, os pontos na fronteira das duas classes, têm influência na definição da fronteira de classificação, sendo necessário no mínimo um ponto de cada classe. É garantido deste modo que, de facto, a margem é máxima.

3.3.1.2 Kernels

Raramente são encontrados problemas onde as classes são separáveis, pois os dados das classes estão normalmente sobrepostos. Nestes casos, os classificadores lineares encontram uma solução baseada na minimização do erro, como por exemplo o método dos mínimos quadrados. Quando os dados não são linearmente separáveis, a divisão óptima das classes pode estar comprometida. Nestes casos podem-se mapear os dados para outro espaço de

características de maior dimensionalidade onde os dados serão, supostamente, separáveis linearmente e então nesse espaço usar um classificador linear. Contudo, este método tem dois problemas: a quantidade de dados pode não ser suficiente para se obter uma representação significativa numa dimensão superior, problema usualmente apelidado da 'maldição da dimensionalidade'; o facto de se realizar cálculos em dimensões elevadas é computacionalmente dispendioso, devido ao aumento da dimensão dos vectores de parâmetros.

Com a aplicação do *'Kernel trick'* a classificadores de máxima margem [75] o produto interno da equação 3.12 é substituído por uma função kernel, pelo que a equação de classificação resulta:

$$\sum_{i=1}^{m} y_i \alpha_i K(X_i, X) + b > 0. \qquad (3.13)$$

O uso de *kernels* permite que os dados sejam separados numa dimensão *N*, mas a classificação dos dados continua a ser efectuada na dimensão original, o que é computacionalmente mais eficiente. Na nova dimensão, onde os dados são mapeados, estes são separáveis linearmente, sendo usado um hiperplano de máxima margem no caso dos SVM. Contudo, visto que esta transformação é não linear, também na dimensão original os dados poderão ficar separados por uma função não linear. São normalmente usados dois tipos de *Kernels*: polinomial (3.14) e gaussiano (3.15).

$$K(X_1, X_2) = (X_1 \cdot X_2 + 1)^d \qquad (3.14)$$

$$K(X_1, X_2) = \exp\left(- \frac{\|X_1 - X_2\|^2}{2\sigma^2} \right) \qquad (3.15)$$

3.3.2 *Discriminant analysis*

A discriminação linear e a discriminação quadrática são dois classificadores clássicos de métodos da análise discriminante (*Discriminant analysis*) [76] que pretendem separar classes com recurso a superfícies de decisão lineares e quadráticas. Apesar de não serem

classificadores de máxima margem como os SVM têm como principais vantagens a pouca exigência computacional e a possibilidade de classificar múltiplas classes.

A classificação de uma amostra numa classe é dada por:

$$\hat{y} = \arg\min_{y=1,\ldots,K} \sum_{k=1}^{K} \hat{P}(k \mid x) C(y \mid k). \quad (3.16)$$

Onde $\hat{y}$ é a amostra predita, *K* é o número de classes, $\hat{P}(k|x)$ é a probabilidade a posteriori e $\mathrm{C(y|k)}$ o custo de classificação da amostra *y* quando a classe verdadeira é *k*.

A probabilidade a posteriori é definida por:

$$\hat{P}(k \mid x) = \frac{P(x \mid k)P(k)}{P(x)}. \quad 3.17$$

Sendo que $P(k)$ é a probabilidade a prior e $P(x|k)$ é modelado por um distribuição normal multivariada dada por:

$$P(x \mid k) = \frac{1}{\left(2\pi|\Sigma_k|\right)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) \right). \quad 3.18$$

Note-se que, para a estimação da discriminação linear, se assume que as matrizes de covariância Σ_k são iguais para todas as classes. Não fazendo essa assunção, a discriminação é quadrática permitindo uma adaptação não linear do discriminador aos dados.

3.3.3 *Gaussian mixture models*

O sistema de reconhecimento baseado em GMM consiste na caracterização de uma classe através de um modelo de misturas gaussianas. Num sistema GMM, uma classe λ é representada por uma distribuição densidade de probabilidade, dada por:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\vec{x}). \quad (3.19)$$

Na equação anterior $p_i(\vec{x})$ é uma distribuição normal, com um vector de médias μ e uma matriz de covariância Σ. Estes valores são estimados pelo algoritmo '*Expectation-Maximization*' (EM) [77], assim como, os pesos w_i de cada uma das gaussianas. O modelo da classe é então caracterizado com $\lambda=(w_i,\mu_i,\sigma_i^2)$, $i=(1,...,M)$ onde M é o número de gaussianas da mistura. As misturas vão modelar as amostras sendo caracterizadas pelas suas médias e variância, o que permite uma caracterização distributiva dos dados.

Tipicamente, no reconhecimento de orador e aplicações semelhantes, são usadas misturas de gaussianas com matrizes de covariância diagonais. Esta técnica tem principalmente duas razões: do ponto de vista computacional é menos exigente; tem sido demonstrado empiricamente que com esta particularidade as taxas de erro são mais baixas [78].

A avaliação de um sinal de fala Y na classe λ tem como resultado um valor de probabilidade, ou verosimilhança. Assumindo que os vectores das N amostras de Y são independentes, a verosimilhança é o produto de todas as probabilidades de cada vector, pelo que os valores obtidos são muito pequenos. Para evitar este facto, é normalmente usada a verosimilhança-logarítmica normalizada. Assim, se Y for constituído por N amostras, o valor final da verosimilhança é obtido pela expressão seguinte:

$$\log\ p(Y \mid \lambda) = \frac{1}{N}\sum_{t=1}^{N} \log\ p(\vec{y}_t \mid \lambda). \qquad (3.20)$$

A verosimilhança é uma medida de semelhança, sendo que uma amostra é classificada na classe em que tiver maior verosimilhança.

3.3.4 Árvores de decisão

As árvores de decisão são bastante comuns em processos de mineração de dados [79]. As árvores de decisão são modelos representativos de tabelas de classificação através de uma estrutura hierárquica que se traduz numa árvore invertida da raiz para as folhas. Cada nível da árvore é traduzido por um nó onde, com base na análise de um atributo, é realizada uma decisão (Figura 3-5).

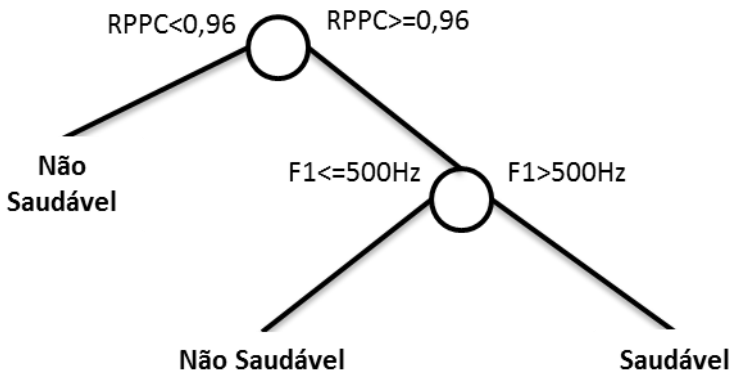

Figura 3-5 – Exemplo de uma árvore de decisão com dois nós, para classificação de sujeitos saudáveis.

A análise de dados por árvores de decisão pode ser bastante útil na determinação de valores de limiares de atributos, nomeadamente quando a quantidade de atributos é baixa. Caso contrário, a árvore pode tornar-se demasiadamente complexa para que se consiga retirar informação útil acerca dos dados. Árvores muito complexas e extensas tipicamente ficam demasiado ajustadas ao conjunto de treino, não generalizando devidamente os dados. Consequentemente, podem não obter resultados relevantes no conjunto de teste. Nestes casos é comum o uso de técnicas de poda de modo a aumentar a capacidade de generalização da árvore.

A construção de uma árvore de decisão assenta da determinação de medidas de impureza ou incerteza, sendo estas medidas dadas, por exemplo, pelo índex de Gini ou pela entropia.

$$Gini(t) = 1 - \sum_{i=0}^{C} [p(i \mid t)]^2 . \qquad (3.21)$$

$$Entropia(t) = -\sum_{i=0}^{C} p(i \mid t) \log_2 p(i \mid t) . \quad (3.22)$$

Onde *C* é o número de classes e $p(i \mid t)$ é a fracção de registos da classe *i* dado um nó *t*. Os nós que minimizam a impureza são os que contêm maior ganho informação, medida que se pretende maximizar na hierarquização dos nós. O ganho de informação é dado genericamente por:

$$\Delta = I(pai) - \sum_{j=1}^{n} \frac{N(v_j)}{N} I(v_j). \quad (3.23)$$

Onde $I(.)$ é a medida de impureza, *n* é o número de atributos, *N* e o número total de objectos do nó-pai e $N(v_j)$ é o número de exemplos associados ao nó-filho v_j .

Com estas medidas consegue-se construir uma árvore de decisão onde no primeiro nó se encontra o nó uma maior informação, sendo os nós da árvore organizados com base na informação de cada atributo. No caso de atributos com valores contínuos, estes métodos permitem encontrar também os seus limiares óptimos de decisão com base na impureza. Nestes casos, os valores dos atributos são comparados com um limiar através de algoritmos iterativos sendo escolhido para limiar o valor que minimiza a impureza.

4 Caracterização de vozes patológicas através da envolvente espectral

A caracterização de vozes patológicas através da envolvente espectral permite assumir que uma característica tipicamente usada na modelação do tracto vocal poder ser usada na descriminação de vozes patológicas. Este capítulo vai apresentar as metodologias, e as consequentes contribuições, que tiveram como ponto de partida o estudo da envolvente espectral. Os métodos apresentados na estimação e análise da envolvente espectral permitem distinguir vozes patológicas de vozes saudáveis. Em situações pontuais os algoritmos apresentados neste capítulo permitem mesmo o reconhecimento de patologias.

Juntamente com a envolvente espectral são também avaliadas outras medidas como a frequência fundamental, o *Jitter* e o RPPC. Os métodos utilizados visam basicamente a visualização de medidas estatísticas e a interpretação dos resultados. São ainda apresentadas soluções para reconhecimento de vozes patológicas com árvores de decisão que permitem a aproximação simples em termos de modelos de classificação. No final do capítulo são apresentadas a validação dos métodos propostos e as conclusões.

4.1 Primeiro pico da envolvente espectral

O espectro do sinal pode ser caracterizado pela sua envolvente espectral (ver Figura 4-1) sendo esta estimada através da resposta impulsiva do filtro LPC, que remove a informação harmónica. No caso do sinal de fala ser vozeado, o espectro é representado por harmónicas múltiplas da frequência fundamental, existindo em zonas de ressonância do tracto vocal (formantes ou não) um aumento da energia das harmónicas. Essa energia pode ser modelada por picos da envolvente espectral em que cada pico é modelado por um par de polos complexos do filtro LPC. É comum em aplicações de processamento de fala usar filtros LPC com ordem entre 8 e 16, sendo que tipicamente são encontrados até 4 formantes para frequências até 4 kHz.

Uma vez que se dispõe de uma maior largura de banda para analisar colocou-se a hipótese de procurar mais informação na envolvente espectral que pudesse conter informação que permitisse a caracterização de vozes patológicas. Sendo a largura de banda dos sinais

analisados de 25 kHz, aumentou-se sucessivamente a ordem do filtro LPC, realizando-se análise visual de modo a encontrar padrões entre as vozes patológicas.

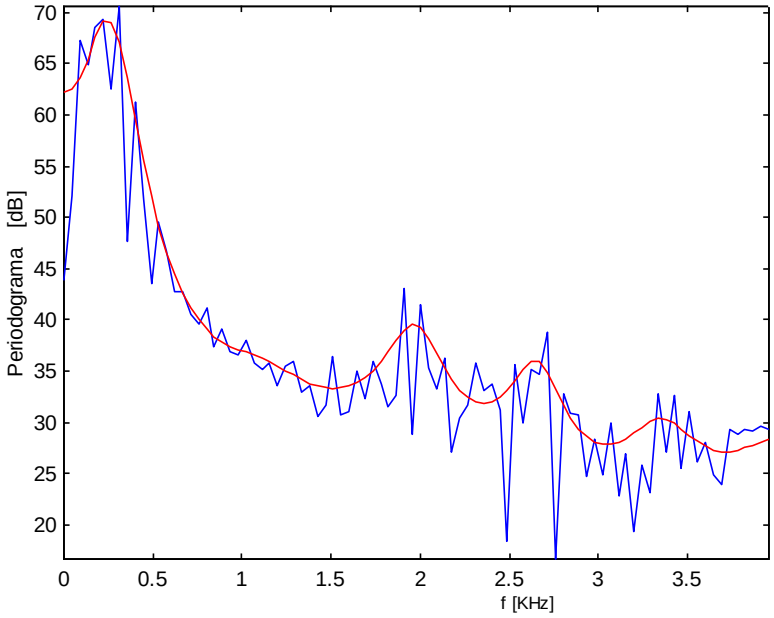

Figura 4-1 – Espectro e Envolvente de uma trama de um sinal de fala

Tabela 4-1 – Base de Dados da Universidade de São Paulo.

Patologia	**Género Masculino**	**Género Feminino**	**Total**
Edema de Reinke	2	14	16
Nódulos	2	13	15
Saudáveis	11	5	16

A base de dados da Universidade de S. Paulo foi usada para um primeiro estudo, Tabela 4-1. Esta base de dados, devido à sua dimensão, permitiu a análise individual de cada ficheiro. Verificou-se que com ordem 30 do filtro LPC, nas vozes patológicas a envolvente espectral exibe um pico antes do pico do primeiro formante que modela as primeiras harmónicas. Nos oradores saudáveis, este pico ou não existe ou se existir aparece com uma maior largura de banda (ver Figura 4-2). Verifica-se também que se não for usado o filtro de pré-enfase, tipicamente aplicado em aplicações de processamento de fala, a energia das baixas frequências não é atenuada, o que contribui para uma melhor detecção deste pico.

Em [5] foram apresentados os primeiros resultados do estudo deste pico da envolvente espectral. O filtro LPC usado foi de ordem 30, para tramas com um andamento de 10 ms e com dimensão de 20 ms. Os coeficientes do filtro LPC foram calculados para todas as tramas do sinal e foram realizadas para cada ficheiro médias para o valor da frequência do primeiro pico da envolvente espectral e para a sua largura de banda.

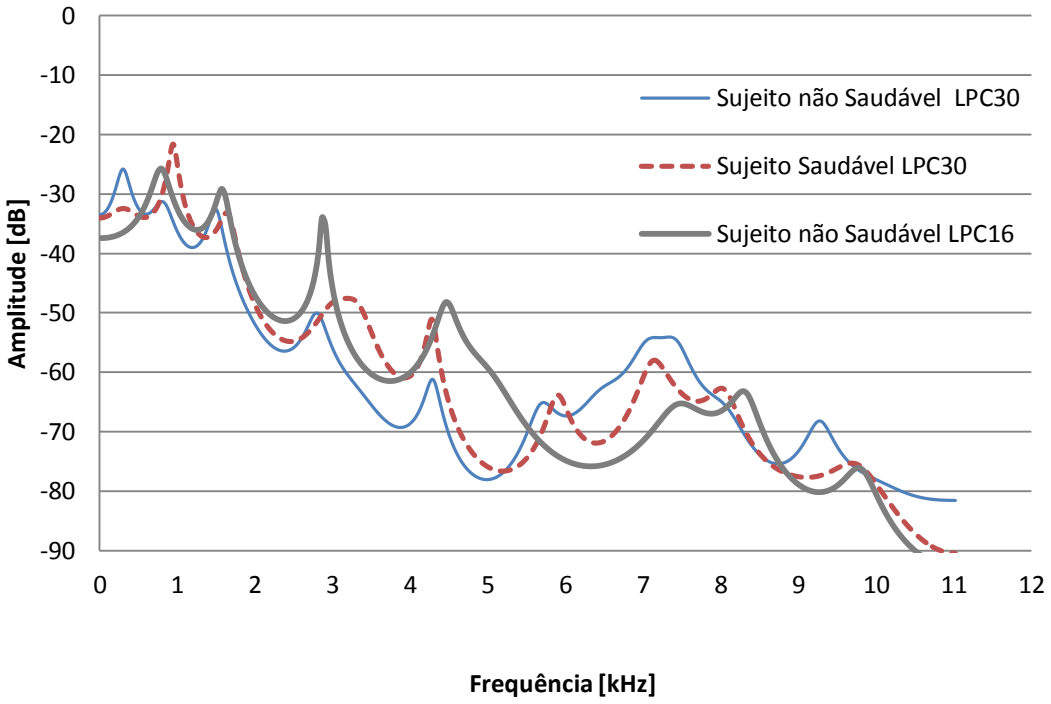

Figura 4-2 – Envolvente espectral de sujeitos saudaveis e não saudaveis para diversas ordens. Nos sujeitos não saudaveis está representada a mesma trama.

Ao serem analisados todos os oradores verifica-se que os oradores com patologias podem ser discriminados dos saudáveis com base na frequência do 1º pico da envolvente espectral e na sua largura de banda, Figura 4-3. Apenas com estes dois parâmetros estas classes ficam totalmente separadas. Parece, de facto, existir uma correlação entre estes dois parâmetros, uma vez que em ambas as classes foram encontrados picos com frequências entre 300 e 400 Hz. No entanto, nos oradores com patologias este pico tem uma largura de banda inferior comparado com os oradores saudáveis. Este facto permite que não exista sobreposição entre classes.

Após uma análise detalhada ao nível das patologias verifica-se que os pacientes com edema de Reinke ou nodulo têm um padrão que permite distinguir estas duas classes, Figura 4-3.

Verifica-se que os sujeitos com edema de Reinke têm um valor de frequência do primeiro pico mais baixo que os oradores com nódulos. Por outro lado, os sujeitos com nódulos têm uma largura de banda mais estável. Assim, por inspecção dos dados, verifica-se que se consegue separar estas duas classes com apenas 4 erros entre 31 oradores, o que representa uma taxa de reconhecimento de 87%.

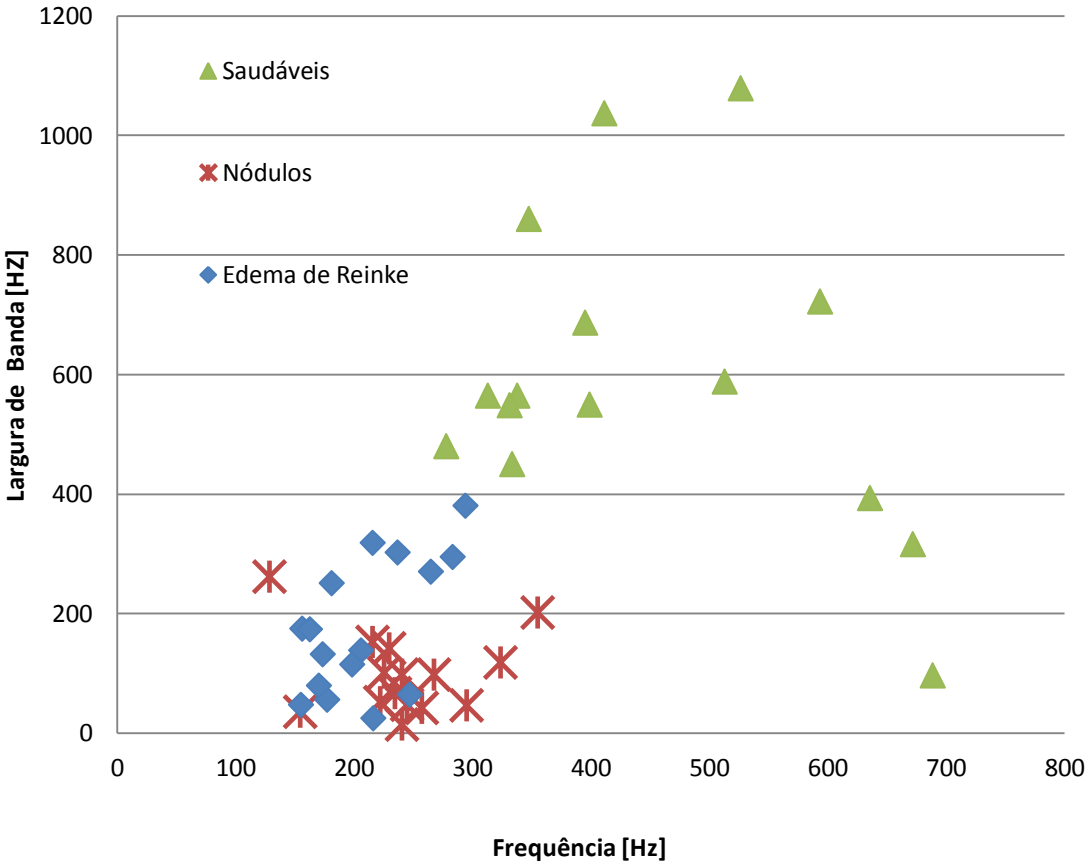

Figura 4-3 – Valor médio da frequência e largura de banda do primeiro pico da envolvente espectral estimada com LPC de ordem 30.

O primeiro pico da envolvente espectral está possivelmente relacionado com o valor das primeiras harmónicas, nomeadamente com a primeira harmónica que representa a frequência fundamental, sendo este aspecto discutido no final deste capítulo. É também uma característica dos oradores com edema de Reinke uma redução da frequência fundamental devido à perda de elasticidade das pregas vocais. Nos oradores femininos, que por norma têm a frequência fundamental mais elevada, esta característica é mais evidente. Dois dos quatro erros que ocorrem são precisamente devido aos dois oradores masculinos que são

diagnosticados com nódulos, pois naturalmente os oradores masculinos já têm uma frequência fundamental menor que os femininos.

Considerando apenas os oradores femininos com patologias desta base de dados conseguir-se-ia diagnosticar correctamente 25 dos 27 oradores, o que representa uma taxa de reconhecimento de 93% entre estas duas patologias da voz. Estes valores, juntamente com o facto de se conseguir separar totalmente os oradores saudáveis dos diagnosticados com patologias, revelam que o primeiro pico da envolvente espectral tem potencial na detecção de vozes com patologias e que também poderá ter alguma informação acerca das patologias.

4.2 Primeiro pico da envolvente espectral nas vogais /e/ e /i/

No ponto anterior verificou-se que a envolvente espectral consegue modelar um pico no espectro antes do típico primeiro formante da vogal /a/, o que permite distinguir oradores saudáveis de oradores com edema de Reinke ou nódulos. Colocou-se então a questão se este primeiro pico está presente apenas na vogal /a/ ou se também está presente noutros fonemas. Uma vez que a base de dados da Universidade de São Paulo contém, apenas para os oradores não saudáveis, os sinais das vogais /e/ e /i/, procedeu-se à inspecção das envolventes espectrais destes sinais para procurar o mesmo primeiro pico na envolvente espectral [5].

Note-se que esta base de dados também foi usada nesta tarefa nos trabalhos [43], [44] onde aos autores conseguiram taxas de acerto de 82% e 87%, usando 80% dos dados para treino e 20% dos dados para teste, apenas para a vogal sustentada /a/.

Mais uma vez realizou-se o cálculo dos coeficientes de predição linear para uma ordem de 30 apenas para os oradores com patologias. Numa primeira análise verifica-se que nos gráficos obtidos não só a envolvente espectral apresenta os valores de um pico abaixo do primeiro formante como também parece existir um padrão que permite realizar classificações das patologias. Assim sendo, foram treinados 3 classificadores SVM, um para cada vogal, de modo avaliar a taxa de reconhecimento destes parâmetros na identificação destas duas patologias. Para os oradores com patologias os resultados obtidos para a largura de banda e a frequência são apresentados nas figuras seguintes.

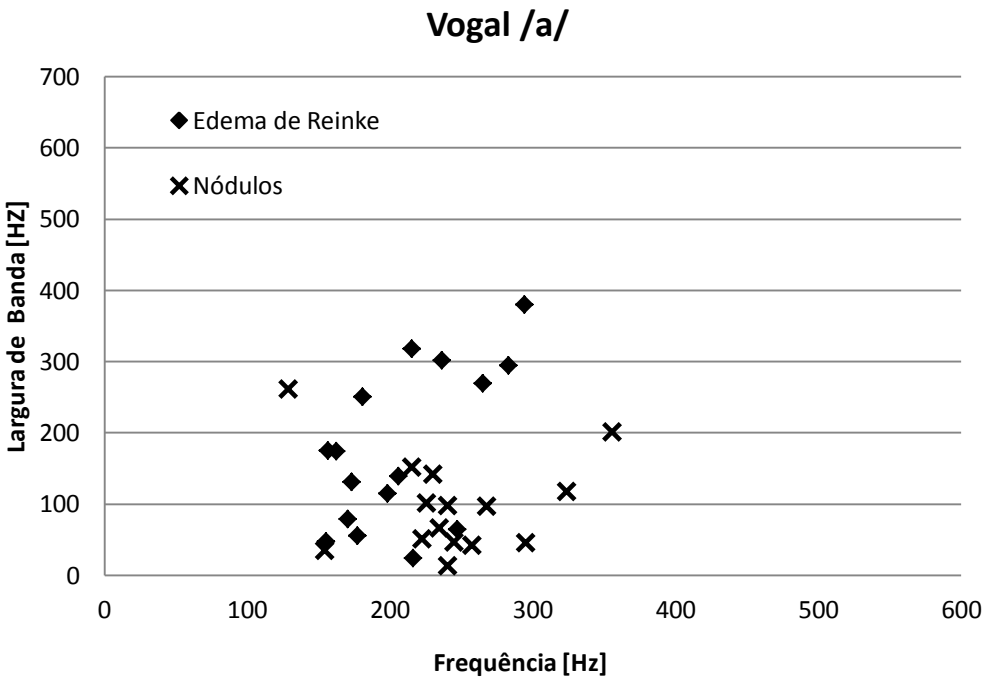

Figura 4-4 - Valor médio da frequência e largura de banda do primeiro pico da envolvente espectral estimada com LPC de ordem 30 na vogal /a/, nos sujeitos diagnosticados com patologias.

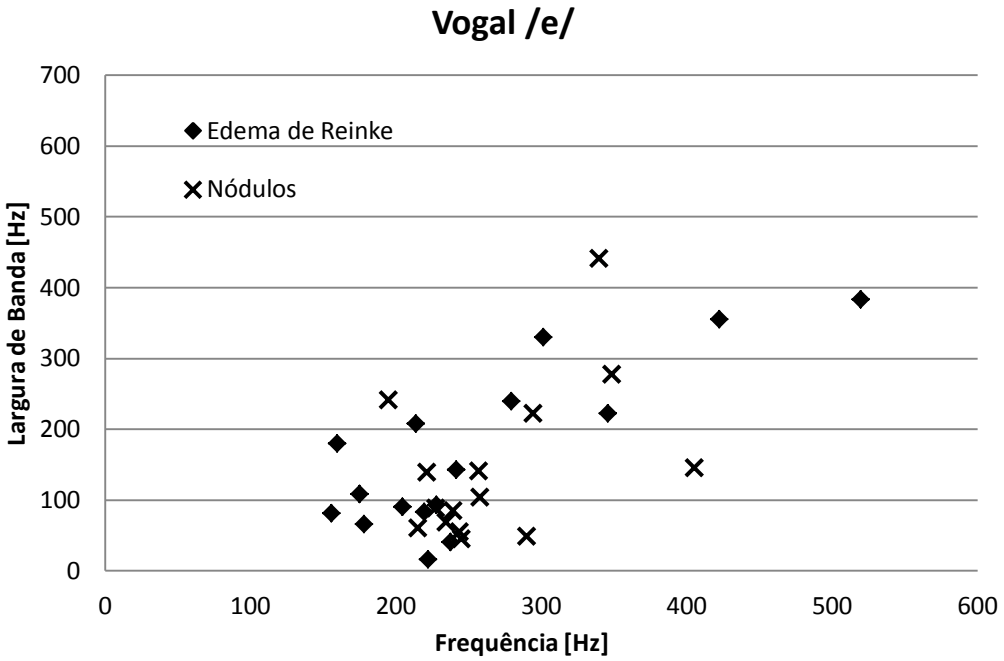

Figura 4-5 - Valor médio da frequência e largura de banda do primeiro pico da envolvente espectral estimada com LPC de ordem 30 na vogal /e/, nos sujeitos diagnosticados com patologias.

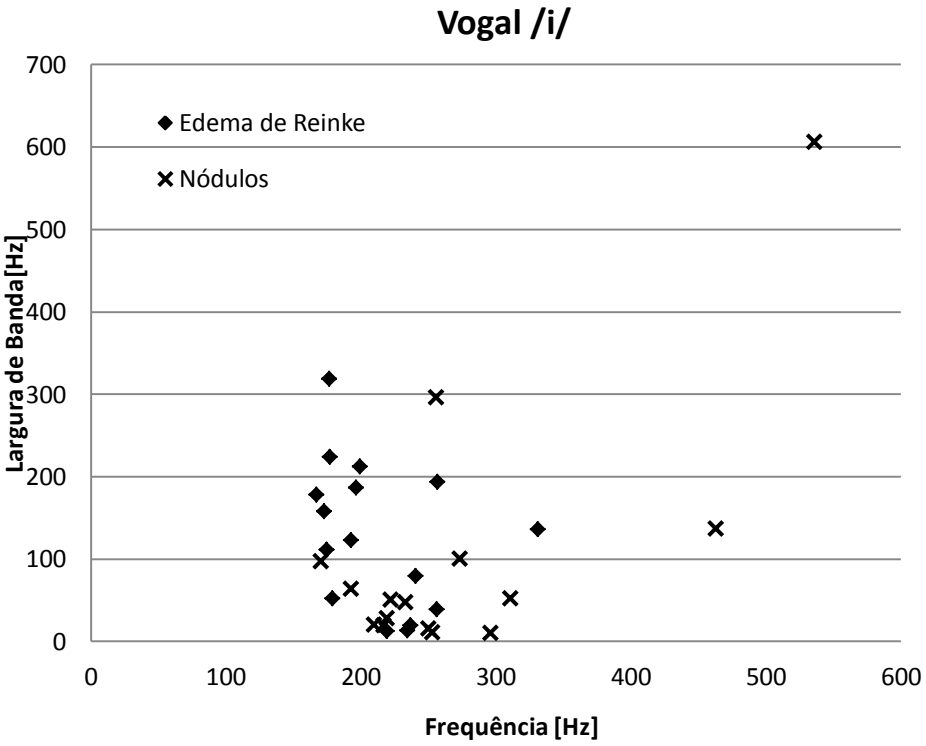

Figura 4-6 - Valor médio da frequência e largura de banda do primeiro pico da envolvente espectral estimada com LPC de ordem 30 na vogal /i/, nos sujeitos diagnosticados com patologias.

Uma vez que nesta base de dados existem apenas quatro oradores masculinos com patologias, dois por patologia, este foram ignorados, sendo os sistemas criados apenas para oradores femininos. Os classificadores SVM foram treinados com 50% dos dados para teste e 50% dos dados para treino. Para cada vogal foi estimada a largura de banda e a frequência do primeiro pico do espectro LPC, em tramas de 20 ms com um andamento de 10 ms, durante 4 troços de 250 ms. Nestes 4 troços foram calculadas as médias dos parâmetros usados na classificação, originando 4 segmentos de classificação ou teste por orador. Para todas as vogais foram usados os mesmos oradores para teste e treino.

Uma vez que um dos sintomas nos oradores femininos com edema de Reinke é a diminuição da frequência fundamental, este parâmetro também foi estimado nos moldes anteriores para verificar o impacto no reconhecimento destas patologias. Do mesmo modo foram estimados valores médios do *jitter* do período fundamental local para perceber se existe alguma correlação deste parâmetro com estas patologias. Os resultados são apresentados nas tabelas seguintes.

Tabela 4-2 - Matrizes de confusão para a frequência do 1º pico do LPC e largura de banda.

	Vogal /a/		**Taxa de acerto na patologia**	**Vogal /e/**		**Taxa de acerto na patologia**	**Vogal /i/**		**Taxa de acerto na patologia**
Edema de Reinke	24	4	85,7%	16	12	57,1%	19	9	67,8 %
Nódulos	4	20	83,3%	5	19	79,2%	7	17	70,8 %
Taxa de reconhecimento total	**84,6**%			67,3%			69,2%		

Tabela 4-3 - Matrizes de confusão para a frequência do 1º pico do LPC, largura de banda e *jitter*.

	Vogal /a/		**Taxa de acerto na patologia**	**Vogal /e/**		**Taxa de acerto na patologia**	**Vogal /i/**		**Taxa de acerto na patologia**
Edema de Reinke	24	4	85,7%	15	13	53,1%	19	9	67,8 %
Nódulos	4	20	83,3%	5	19	79,2%	7	17	70,8 %
Taxa de reconhecimento total	**84,6%**			65,4%			69,2%		

Tabela 4-4 - Matrizes de confusão para a frequência do 1º pico do LPC, largura de banda e frequência fundamental.

	Vogal /a/		**Taxa de acerto na patologia**	**Vogal /e/**		**Taxa de acerto na patologia**	**Vogal /i/**		**Taxa de acerto na patologia**
Edema de Reinke	24	4	85,7%	22	6	78,6%	13	15	46,4 %
Nódulos	4	20	83,3%	8	16	66,7%	1	23	95,8 %
Taxa de reconhecimento total	**84,6%**			73,1%			69,2%		

Tabela 4-5 - Matrizes de confusão para a frequência do 1º pico do LPC, largura de banda, frequência fundamental e *jitter*.

	Vogal /a/		**Taxa de acerto na patologia**	**Vogal /e/**		**Taxa de acerto na patologia**	**Vogal /i/**		**Taxa de acerto na patologia**
Edema de Reinke	24	4	85,7%	26	2	98,2%	16	12	57,1 %
Nódulos	4	20	83,3%	5	19	79,2%	4	20	83,3 %
Taxa de reconhecimento total	84,6%			**86,5%**			69,2%		

Analisando os resultados obtidos nas tabelas anteriores, verifica-se que na vogal /a/ os resultados são sempre consistentes com as várias características usadas. Para os 4 sistemas implementados é obtido sempre 84,6% de taxa de reconhecimento, sendo que o uso da frequência fundamental ou *jitter* do período fundamental não contribuem para alteração da taxa de reconhecimento de patologias. Para a vogal /i/ os resultados da taxa total de acertos também permanecem iguais nos quatro sistemas, independentemente das características usadas. No entanto, analisando o resultado das taxas de acerto, verifica-se que a taxa de reconhecimento dentro dos nódulos aumenta mas a taxa de acerto no edema de Reinke diminui. Podemos portanto constatar que o facto de usar mais características apenas aumenta o desbalanceamento das classes.

Nos sistemas implementados com a vogal /e/ consegue-se melhorar os resultados usando a frequência fundamental e o *jitter* do período fundamental. Usando apenas a frequência do primeiro pico do LPC e a largura de banda obtêm-se uma taxa de reconhecimento de 67,3%, mas utilizando também a frequência fundamental e o *jitter* do período fundamental consegue-se uma taxa de reconhecimento de 86,5%. Ao nível das patologias, este valor representa uma taxa de reconhecimento de 98,2% nos segmentos com edema de Reinke e 79,2% nos nódulos, sendo que para a vogal /a/ são obtidos 85,7% e 83,3%, respectivamente.

Verifica-se portanto que a vogal /a/ apresenta a taxa de reconhecimento mais consistente, mas que nas outras vogais analisadas, /e/ e /i/, o primeiro pico do LPC também contem informação da patologia. Nestas duas vogais apenas com a frequência do primeiro pico do LPC e da sua largura de banda consegue obter taxas de acerto que rondam os 70%, resultados

esses que, no caso da vogal /e/, são optimizados usando a frequência fundamental e o *jitter* do período fundamental.

Outro dos objectivos do trabalho consiste em averiguar se o primeiro pico da envolvente espectral está presente noutras vogais que não o /a/. Pelos resultados mostrados na Tabela 4-4 e Tabela 4-5 pode-se confirmar esta hipótese. A Tabela 4-6 mostra o valor médio e o desvio padrão de todos os parâmetros usados no trabalho.

Tabela 4-6 – Valores médios dos parâmetros do sinal de fala usados nos reconhecedores para cada vogal e patatologia da voz.

		1º Pico da Envolvente Espectral				***Jitter* do período fundamental**		**Frequência fundamental**	
	Vogal	μ Frequência [Hz]	σ Frequência [Hz]	μ Largura de Banda [Hz]	σ Largura de Banda [Hz]	μ [%]	σ [%]	μ [Hz]	σ [Hz]
Edema de Reinke	/a/	214	44,4	170,9	119	2,8	2,5	162,5	42,2
	/e/	265	117,6	165,1	131,7	5,7	6,5	183,8	44,3
	/i/	218	65,8	108,2	88,0	5	3,9	187,1	40,5
Nódulos	/a/	258	41,4	90,4	52,1	5,4	4,8	226,2	31,8
	/e/	271,2	65,5	130	120,9	6,6	5,5	225,1	28,2
	/i/	282,7	101,8	89,6	164,2	6,4	5,6	227,2	32,7

Pode-se assim concluir que, para vozes com estas duas patologias, em todas as vogais existe a presença de um primeiro pico antes do primeiro formante. Contudo, pelos resultados obtidos ao nível dos reconhecedores implementados, dos valores apresentados na tabela anterior e pela análise das figuras apresentadas no inicio deste ponto, o cálculo da frequência do primeiro pico do LPC e da sua largura de banda tem menor variância para a vogal /a/. Este facto deve-se ao primeiro formante desta vogal se encontrar em valores médios de 800 Hz para este conjunto de oradores. Por outro lado, para a vogal /e/ e /i/, os valores médios do primeiro formante são 520 Hz e 515 Hz, respectivamente. Existe assim uma maior proximidade entre o primeiro pico do LPC e o real primeiro formante nestas duas vogais, o que pode tornar a detecção do primeiro pico do LPC menos eficiente.

Do ponto de vista da identificação de vozes patológicas acredita-se que estes desvios na detecção do primeiro pico do LPC não deverão ter grande impacto. No ponto anterior podemos verificar que o sujeito saudável onde foi detectada a frequência mais baixa do 1º pico do LPC, ronda os 277 Hz mas este pico tem uma largura de banda 480 Hz, sendo este valor acima dos valores encontrados nas vogais /e/ e /i/, nos sujeitos com patologias.

4.3 Primeiro pico da envolvente espectral e RPPC na base de dados da MEEI

Nos dois pontos anteriores verificou-se, na base de dados da Universidade de S. Paulo, a ocorrência de um primeiro pico na envolvente espectral antes do formante correspondente à vogal, que permite distinguir entre sujeitos saudáveis e sujeitos com patologias. Esse primeiro pico encontra-se aparentemente em todos os fonemas analisados em vozes patológicas. As patologias analisadas foram o edema de Reinke e os nódulos, num total de 32 sujeitos com patologias. Apesar ser uma base de dados de relativa pequena dimensão permitiu realizar trabalho exploratório e chegar a conclusões importantes.

De modo a verificar a existência do primeiro pico do espectro do LPC numa base de dados de maior dimensão, foram efectuados testes na base de dados da MEEI. Desta base de dados foram usados 53 sujeitos saudáveis e 153 sujeitos com várias patologias, identificadas na Tabela 4-7, tendo sido usados oradores femininos e masculinos. Oradores com múltiplas patologias foram apenas uma vez. O sinal de fala utilizado neste trabalho [6] foi a vogal sustentada /a/, sendo os dados semelhantes aos usados em [36].

Tabela 4-7 – Descrição da base de dados.

Patologia	**Número de sujeitos**
Saudaveis	53
Nódulos	15
Edema	37
Paralisia das Pregas Vocais	65
Pólipos	15
Keratosis\Leukoplakia	21

Com este conjunto de dados vai ser realizado um estudo da envolvente espectral de modo a verificar se também nesta existe um primeiro pico no espectro LPC antes do típico primeiro formante. Além disso, pretende-se verificar se este primeiro pico está presente noutras patologias das pregas vocais. Se tal facto se verificar, podemos afirmar com mais segurança que o primeiro pico da envolvente espectral de ordem elevada contém informação relevante na caracterização de vozes patológicas. Paralelamente vão ser analisados outros elementos da envolvente espectral, como sejam as frequências e larguras de banda dos outros formantes da envolvente espectral e verificar se existem outros elementos que possam ajudar na caracterização de vozes patológicas.

Paralelamente foi avaliado o impacto do RPPC (*Relative Power of the Periodic Component*). Esta medida consiste no cálculo do numerador da equação 3.4 sendo esta equação que estima o HNR com base na autocorrelação. Uma vez que o RPPC representa a potência relativa da componente periódica, também é verdade que sendo este parâmetro o valor no primeiro máximo da autocorrelação normalizada e que este pode ser usado para avaliar e decidir a presença de vozeamento, contém informação sobre vozes patológicas. Sendo a autocorrelação uma medida de semelhança do sinal com ele próprio ao longo do tempo, é espectável que sujeitos saudáveis tenham este valor mais elevado que sujeitos com patologias da voz. Em sujeitos não saudáveis, onde os problemas no vozeamento se reflectem na variação periodicidade e na variação intensidade do sinal de fala, é expectável que o valor RPPC seja tipicamente inferior aos sujeitos saudáveis.

Outro dos objectivos desta tese é desenvolver um reconhecedor de fácil interpretação, que permitisse a classificação de vozes patológicas. Como tal, o objectivo é implementar uma árvore de decisão em que os valores da classificação dos nós serão obtidos através da geração de árvores de decisão automática. As árvores de decisão são construídas através de relevância das características. A sua geração automática tende a criar demasiados nós sendo frequente a realização de podas de modo a não adaptar a árvore exclusivamente aos dados de treino e generalizar para os dados de teste. Contudo os valores de decisão estimados pela árvore para as características analisadas podem indicar o caminho para a construção manual de uma árvore de decisão.

4.3.1 Análise das características

O sinal de fala usado no desenvolvimento foi a vogal /a/, pois esta também é a única vogal sustentada presente na base de dados da MEEI. Foram extraídos os valores médios da frequência de primeiro pico do LPC assim como a sua largura de banda. A figura seguinte mostra o resultado destes valores para todos os oradores da base de dados.

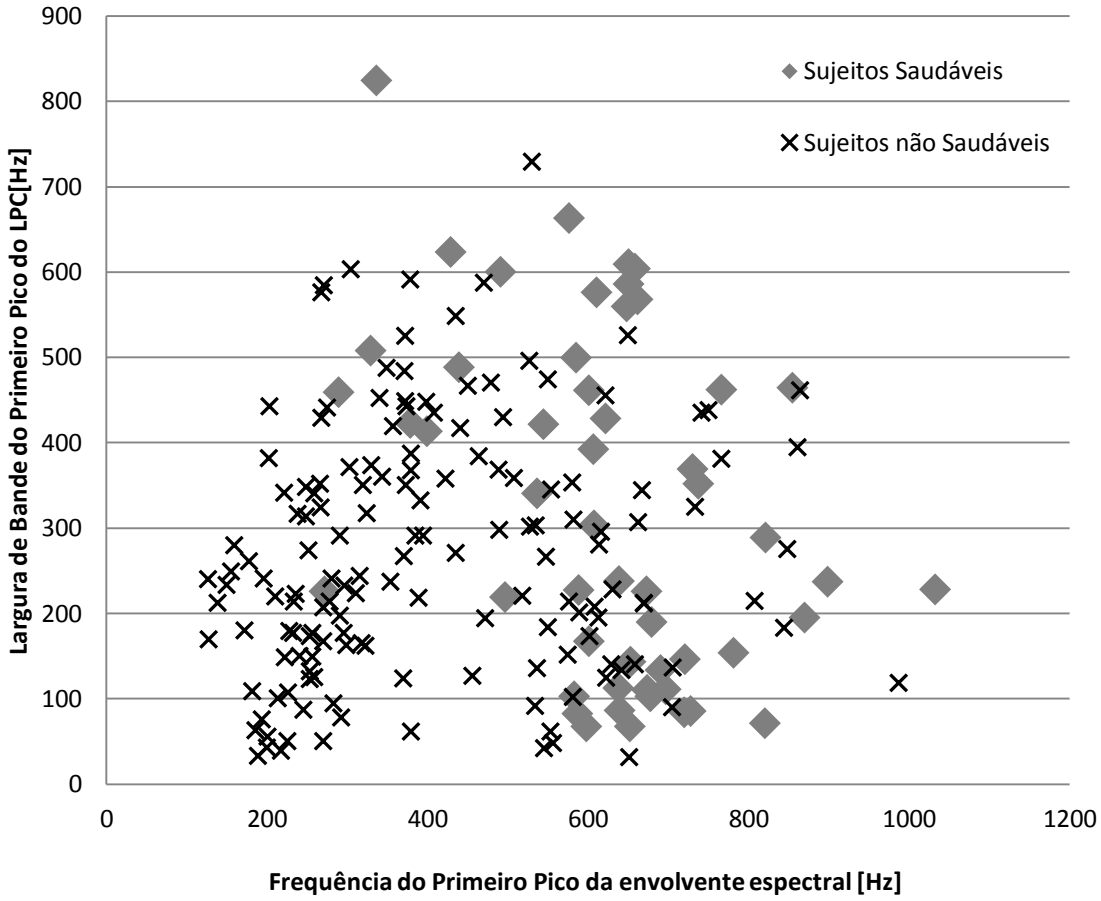

Figura 4-7 - Valor médio da frequência e largura de banda do primeiro pico da envolvente espectral estimada com LPC de ordem 30, sujeitos saudáveis e não saudaveis na base de dados da MEEI.

Como se pode verificar, continua a existir uma grande quantidade de sujeitos diagnosticados com patologias da voz que no sinal de fala analisado contêm um primeiro pico no espectro antes do primeiro formante, sendo que nesta base de dados o valor médio do primeiro formante da Vogal /a/ estimado é de 660 Hz. Verifica-se também na Figura 4-7 que apenas um dos 53 oradores saudáveis está numa zona que é efectivamente ocupada por sujeitos com patologias. Neste caso, é estimado como sujeito saudável com uma frequência do primeiro

pico média de 271 Hz com uma largura de banda máxima de 225 Hz. Todos os outros oradores apresentam semelhanças com os estudos anteriores onde o valor da frequência do primeiro pico é a frequência do primeiro formante, ou se for estimada um valor de frequência mais baixo, então esse pico tem uma largura de banda elevada. Por outro lado, existem 153 sujeitos diagnosticados com várias patologias das pregas vocais e 40 sujeitos que têm a frequência do primeiro formante acima de 550 Hz, sendo que nestes não existe a predominância do primeiro pico da envolvente espectral. Ainda assim em 74% dos sujeitos diagnosticados com patologias observa-se um pico de frequência inferior ao valor típico do primeiro formante.

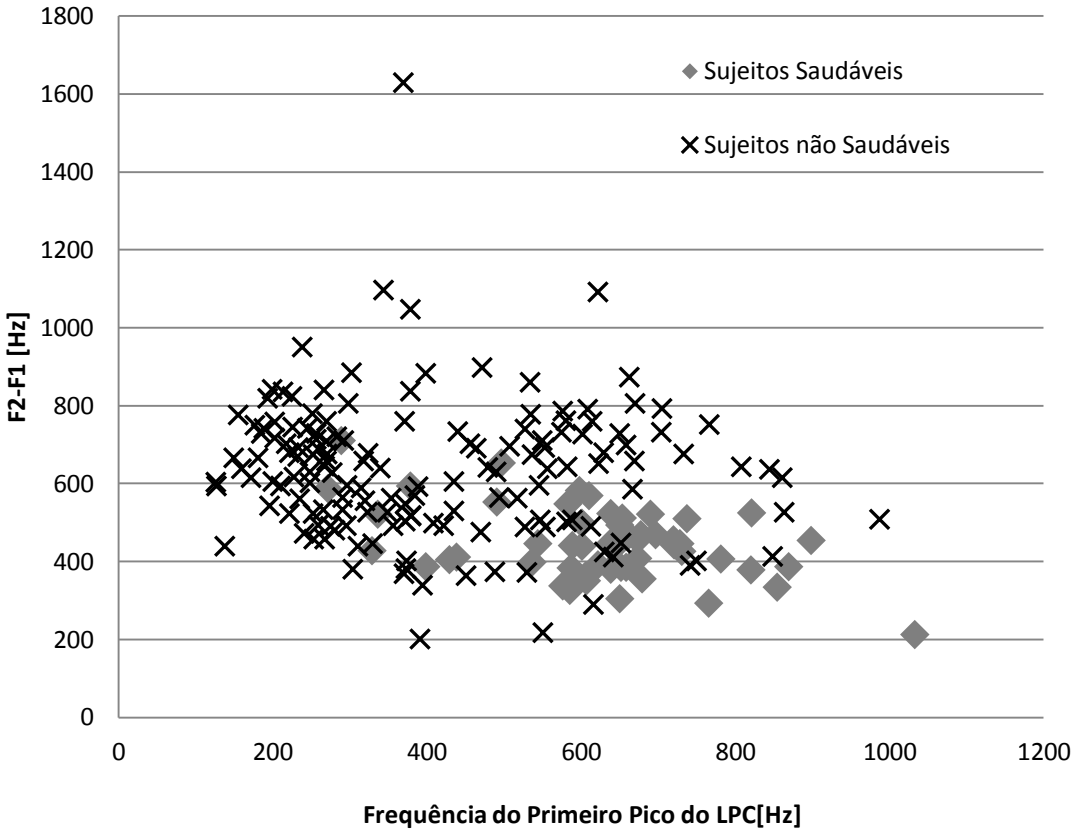

Figura 4-8 - Valor médio da frequência do primeiro pico da envolvente espectral estimada com LPC de ordem 30 e a diferença entre os dois primeiros formantes, sujeitos saudáveis e não saudaveis na base de dados da MEEI.

Observando os dados dos valores médios dos formantes verifica-se que nos sujeitos não saudáveis não diagnosticados através do primeiro pico na envolvente espectral, a diferença entre o primeiro formante e o segundo formante é maior que num orador saudável.

Verifica-se através da análise da Figura 4-8, que num número considerável de sujeitos não saudáveis com valores da frequência do primeiro pico da envolvente espectral acima de 550 Hz existe uma diferença considerável entre o primeiro e o segundo formante (F2-F1) que é superior à dos oradores saudáveis nas mesmas frequências. De facto existe uma quantidade considerável de oradores que conseguem ser correctamente classificados com esta característica.

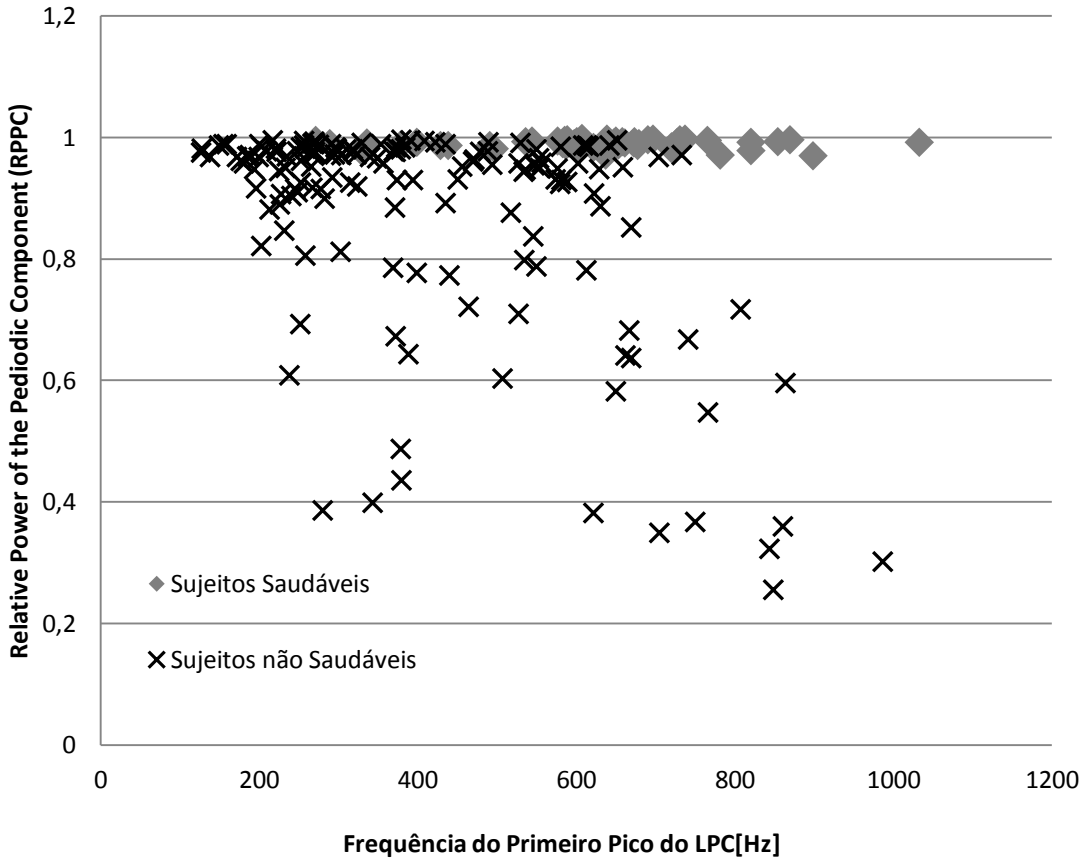

Figura 4-9 - Valor médio da frequência do primeiro pico da envolvente espectral estimada com LPC de ordem 30 e o RPPC, sujeitos saudáveis e não saudaveis na base de dados da MEEI.

Por fim, analisaram-se os resultados do RPPC em função do primeiro pico da autocorrelação. O cálculo do RPPC foi efectuado com base no período fundamental tendo sido calculada uma primeira autocorrelação, numa janela de 30 ms, para determinar qual a frequência da frequência fundamental. Posteriormente, foi calculado o valor do primeiro pico da autocorrelação numa janela com uma dimensão igual ao de 6 períodos fundamentais. A autocorrelação obtida foi também normalizada para retirar o efeito da janela rectangular

usada para a segmentação das tramas. A normalização consiste em multiplicar a autocorrelação do sinal pela autocorrelação da janela de análise. Neste caso, sendo a janela de análise rectangular, a sua autocorrelação vai ser uma janela triangular. O valor do RRPC foi encontrado na vizinhança dos valores da frequência fundamental que foi estimado na primeira iteração do algoritmo.

A Figura 4-9 mostra os valores médios do RPPC estimados para todos os sujeitos da base de dados. Como esperado, verifica-se que os sujeitos diagnosticados com patologias têm um valor de RPPC mais baixo que os sujeitos saudáveis. Este facto deve-se a características típicas de vozes patológicas como sejam uma menor periodicidade ou um menor vozeamento. Verifica-se, de facto, que nos sujeitos diagnosticados com patologias da voz em que não é detectado o primeiro pico do LPC abaixo de 550 Hz existem um numero significativo de sujeitos que têm os valores de RPPC mais baixos, com valores abaixo de 0,6 existindo mesmo um conjunto de 7 sujeitos cujo valor de RPPC está abaixo de 0,4 e nos quais não foi detectado o primeiro pico do espectro LPC com uma frequência inferior a 550 Hz.

4.3.2 Árvore de decisão

Após uma análise preliminar das características extraídas do sinal de fala, o objectivo foi desenvolver uma árvore de decisão de fácil interpretação para a classificação de vozes patológicas. A primeira abordagem consistiu em usar o treino automático de árvores de decisão para obter alguns resultados preliminares, mas acima de tudo o objectivo foi de entender quais os valores das características, analisadas no ponto anterior, que iriam ser usados no projecto da árvore final. Para a implementação das árvores de decisão foi usada a *Matlab Statistics Toolbox* sendo o treino das árvores realizado com o método de *Gini's Diversity Index* que mede a dispersão dos dados permitindo conhecer qual a relevância de cada característica.

Para avaliar o impacto na classificação da frequência do primeiro pico da envolvente espectral, a sua largura de banda e a diferença entre F2 e F1 e o RPPC, foi criado um conjunto de treino e um conjunto de teste com metade dos dados para cada. Nos sujeitos diagnosticados com patologias a selecção dos elementos de treino e teste foi realizada ao nível da patologia. Uma vez que na base de dados usada estão presentes 5 patologias, os sinais de cada patologia

foram divididos em treino e teste. O resultado foi 27 sujeitos saudáveis e 79 sujeitos não saudáveis para treino sendo os restantes 26 saudáveis e 77 não saudáveis usados para o teste.

A primeira abordagem no estudo do impacto das quatro características apresentadas no ponto anterior foi realizada em pares de patologias, agrupadas em função da frequência do primeiro pico do LPC. O resultado foi a criação de três árvores de classificação que permitiram obter os valores de decisão aproximados dos 4 parâmetros, assim como a sua relevância. Por fim, realizou-se um treino com os quatro parâmetros por forma a obter um resultado de referência. Os resultados obtidos estão apresentados na tabela seguinte.

Tabela 4-8 – Resultados obtidos obtidos com árvores de decisão geradas pelo MatLab no reconhecimento de vozes patologicas.

Parâmetros	**Conjunto de Treino**		**Conjunto de Teste**		**Taxa de acerto no conjunto de teste**	**Taxa de Reconhecimento Total**
	Saudável	**Não Saudável**	**Saudável**	**Não Saudável**		
F1 e *BW1*	9/27	74/79	8/26	72/74	80%	79,1%
F1 e *RPPC*	24/27	73/79	19/26	68/74	87%	89,3%
F1 e *RPPC (com poda)*	19/27	75/79	17/26	71/74	88%	88,3%
F1 e *F2-F1*	10/27	73/79	11/26	72/74	83%	80,6%
F1 e *F2-F1 (com poda)*	19/27	66/79	18/26	71/74	89%	84,5%
Todas as características anteriores	18/27	74/79	18/26	68/74	86%	86,4%

Verifica-se que nos segmentos de teste e de treino os resultados obtidos são idênticos na maior parte dos casos, pelo que se pode depreender que não existe sobre-adaptação das árvores no conjunto de treino. Com a frequência do primeiro pico e a largura de banda BW1, obtém-se cerca de 80% de acertos na base de dados, sendo este o valor mais baixo obtido com árvores de decisão. Por outro lado, a árvore obtida com o F1 e o RPPC obtém o melhor resultado para toda a base de dados, incluindo o conjunto de treino e o de teste, com 89,3%. A árvore de decisão obtida com F1 e a diferença entre formantes obtém os melhores resultados no conjunto de teste com 89%, sendo que este resultado foi obtido usando um processo de

poda. Os resultados obtidos com as todas as características (F1, BW1, RPPC e F2-F1) foram de 86% de acertos no conjunto de teste e 86,4% no conjunto de treino, ficando aquém dos melhores resultados nesta fase.

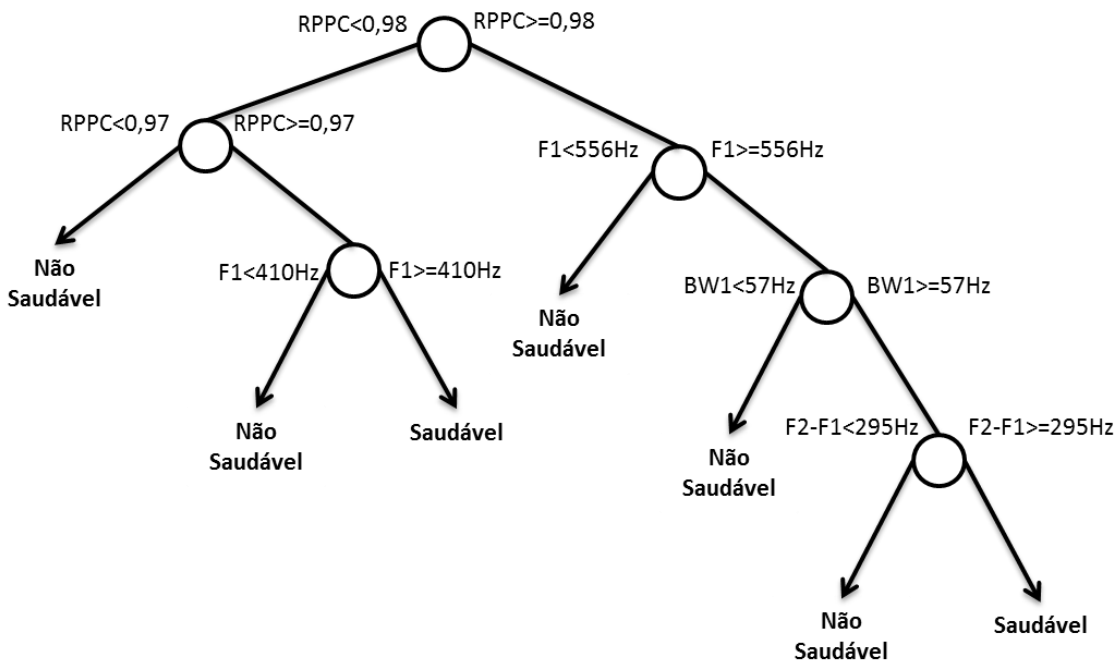

Figura 4-10 – Exemplo de uma árvore gerada pelo MATLAB.

Verifica-se que sempre que é usado o RPPC este nó surge em primeiro lugar sendo o valor de decisão deste nó cerca de 0,98 (valor do RPPC). Verifica-se também que em vários nós das árvores geradas automaticamente surgem valores entre 550 Hz e 580Hz para o valor da frequência de F1 e F2-F1 surge com valores que rondam os 600 Hz. Contudo, nos valores da largura de banda de F1 parece não existir nenhuma convergência. A Figura 4-10 mostra a árvore gerada pelo MATLAB usando as 4 características descritas anteriormente. O resultado obtido é mostrado na última linha da Tabela 4-8.

A Figura 4-11 apresenta a árvore de decisão. No primeiro nó da árvore é avaliado o valor do RPPC. Com se pode verificar na Figura 4-9, para grande parte dos sujeitos não saudáveis o RPPC está abaixo de 0,96 e não existe nenhum sujeito saudável com um valor abaixo deste. De facto, considerando este valor como limiar, 76% dos sujeitos não saudáveis são classificados correctamente. Esta é portanto a característica usada no primeiro nó da árvore de decisão melhorada.

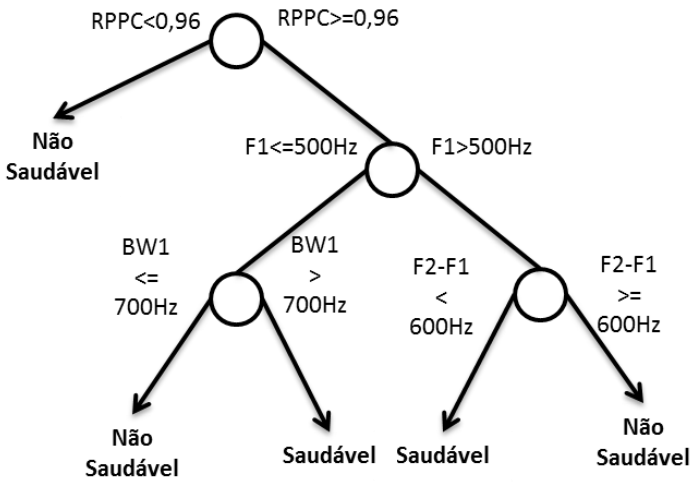

Figura 4-11 – Árvore de decisão optimizada.

Para o segundo nível do projecto da árvore foi usado o valor da frequência do primeiro pico do LPC, F1. O objectivo seguinte é separar os 24% dos sujeitos diagnosticados com patologias dos sujeitos saudáveis. Apesar das árvores de decisão projectadas automaticamente apontarem para valores superiores a 550 Hz, na árvore automática e tendo em conta as figuras referidas anteriormente, usou-se o valor de 500 Hz, sendo que, se o valor de F1 for inferior a este, irá ser avaliada a largura de banda do formante, BW1, e se for superior irá ser avaliada F2-F1.

Deste modo, o terceiro nível de decisão é constituído por dois nós. Para BW1 superior a 700 Hz, o sujeito é considerado saudável. Para valores de F2-F1 inferiores a 600 Hz, valor neste caso indicado pelas árvores automáticas, o sujeito é considerado saudável. Tendo em conta estas considerações projectou-se a árvore de decisão apresentada na Figura 4-11 e os resultados obtidos são apresentados na Tabela 4-9.

Tabela 4-9 – Resultados obtidos na árvore de decisão optimizada.

Valores dos Parâmetros	**Conjunto de Treino**		**Conjunto de Teste**		**Taxa de acerto no conjunto de teste**	**Taxa de Reconhecimento Total**
	Saudável	**Não Saudável**	**Saudável**	**Não Saudável**		
Figura 4-11	23/27	73/79	24/26	72/74	96%	93,2%
F1=600 Hz	20/27	73/79	18/26	74/74	92%	89,8%
RPPC=0,97 F1=350 Hz	24/27	68/79	24/26	69/74	93%	89,8%

A árvore optimizada permite obter 96% de acertos no conjunto de teste e 93,2% de acertos considerando toda a base de dados, na mesma base de dados usada no projecto de árvores automáticas.

No diagnóstico automático de vozes patológicas é preferível maximizar a correcta classificação de sujeitos com patologias, mesmo à custa de considerar alguns sujeitos saudáveis como sofrendo de uma patologia. Na Tabela 4-9 pode-se verificar que, se o valor de F1 passar para 600 Hz mantendo o resto da árvore, vão ser classificados correctamente todos os sujeitos não saudáveis do conjunto de teste, sendo que no conjunto de treino o resultado permanece inalterado. Contudo, nesta abordagem, o erro entre os sujeitos saudáveis aumenta e como consequência a taxa de reconhecimento total desce para 92% no conjunto de teste e para 89,8% quando usada toda a base de dados.

Uma medida de qualidade é o EER (*Equal Error Rate*) onde a taxa de erro nos sujeitos saudáveis é idêntica à taxa de erro nos sujeitos com patologias. Neste caso alterou-se o valor do RPPC para 0,97 e F1 para 350Hz, obtendo-se assim um valor de ERR de 10,2%, tendo em conta o conjunto de treino e o de teste.

Verifica-se que a árvore optimizada melhora os resultados obtidos pelas árvores de decisão geradas no MatLab. Os resultados obtidos de forma automática não eram beneficiados quando eram usadas as quatro características aqui apresentadas. Na árvore optimizada conseguiu-se uma taxa de reconhecimento de 93,2% em toda a base de dados, contra 89,3% nas árvores automáticas. A diferença é ainda maior se for considerado apenas o conjunto de teste, conseguindo-se neste caso 96% de acertos contra 89%.

Se o nó do RRPC for retirado da árvore optimizada, ficando a decisão dependente apenas de características da envolvente espectral, consegue-se 87,4% de acertos em toda a base de dados enquanto ao usar apenas o valor de F1 se atinge uma taxa de reconhecimento de 77,2%. Por outro lado, o uso de largura de banda de F1 tem um impacto positivo na identificação de mais 5 sujeitos saudáveis diagnosticados correctamente e o valor de F2-F1 classifica correctamente mais 8 sujeitos com patologias. Os resultados obtidos com F1 demostram também que, nesta base de dados, em mais sujeitos e em diferentes patologias, o primeiro pico do LPC para uma ordem elevada contêm informação sobre vozes patológicas.

4.4 Validação e discussão

Este capítulo descreveu a presença de um pico na envolvente espectral que se verifica na maioria das vozes patológicas antes da frequência do primeiro formante, e que não estando presente em oradores saudáveis, contem efectivamente informação relevante na detecção de vozes patológicas.

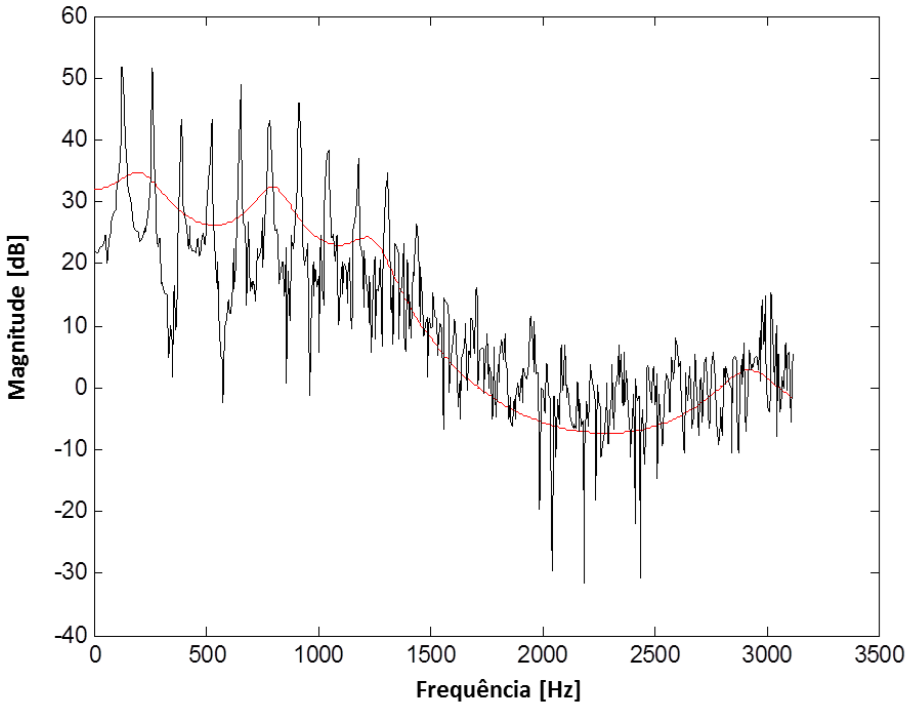

Figura 4-12 – Espectro e envolvente espectral do orador não saudável nº199, foi detectado primeiro pico em 209 Hz, $RPPC_{med}$= 0,977 e r_1=0,970.

Analisando alguns espectros, como por exemplo o espectro da Figura 4-12, verifica-se que este pico está a modelar a primeira e a segunda harmónica do sinal. Estas duas harmónicas, cujas amplitudes são designadas na literatura por H1 e H2, são alvo de vários estudos [80]. Esses estudos revelam que a diferença entre estas duas amplitudes está estritamente relacionada com as vozes soprosas, sendo que nestes casos o valor de H1-H2 deverá ser inferior numa voz normal, ou seja, o valor de H2 é superior a H1 numa voz não soprosa.

Vozes soprosas decorrem da incapacidade da pregas vocais fecharem completamente, dando origem à passagem de ar de forma indesejada, sem que esse ar permita a sua correcta vibração, fenómeno que está presente em muitas das patologias. Alguns factores podem, no entanto, fazer que este primeiro pico não seja detectado. O ruído e a consequentemente

diminuição do declive espectral são apenas um exemplo. Quanto maior for a quantidade de ar que passa indevidamente pelas pregas vocais maior vai ser o ruído no sinal, ruído este que está presente tipicamente nas médias e altas frequências, provocando a diminuição do declive espectral, como mostrado na Figura 4-13.

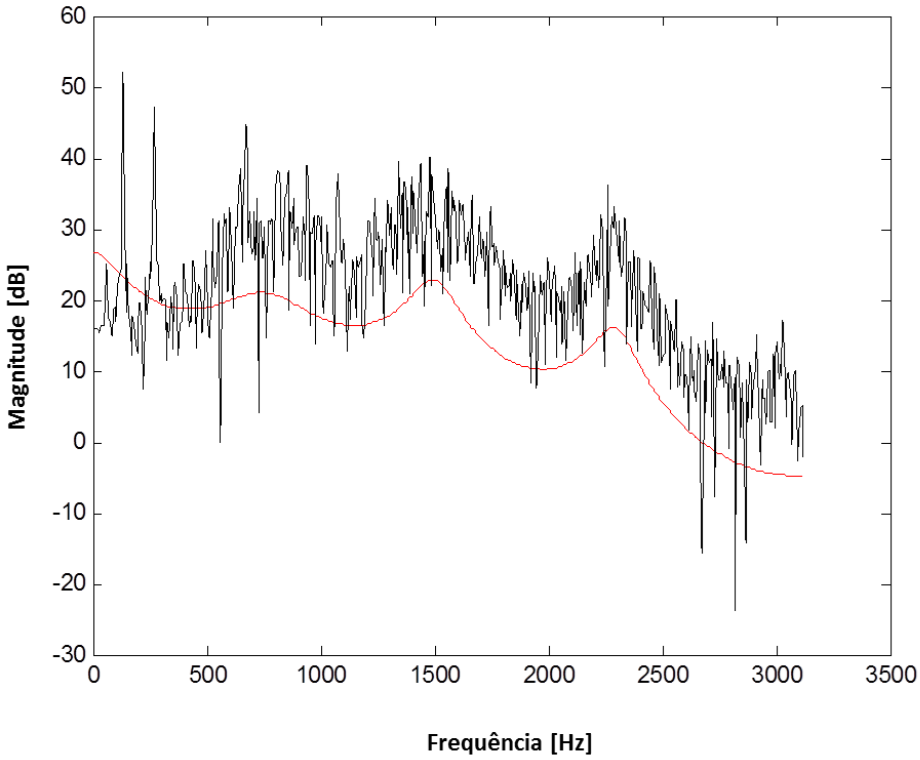

Figura 4-13 - Espectro e envolvente espectral do orador não saudável nº161, $RPPC_{med}$= 0,54 (não foi detectado primeiro pico), r_1=0,93, foi detectado por F2-F1 maior 600 Hz.

Na figura anterior verifica-se que a presença de ruído diminui o declive espectral. Deste modo deixa de ser possível caracterizar a primeira e a segunda harmónica pelo primeiro pico da envolvente espectral. Neste caso, a voz com patologia seria detectada através do valor do RPPC que tem um valor muito baixo devido à presença de ruído.

Observe-se agora o caso de um sujeito não saudável que não foi detectado pelo sistema, Figura 4-14, e o espectro de um orador saudável, Figura 4-15. Nestes dois casos verifica-se que os espectros, valores dos declives espectrais e RPPC são em tudo idênticos. De facto, o valor do RPPC é mesmo superior no orador não saudável, sendo visível uma menor presença de ruído

no espectro. Neste caso é impossível diagnosticar correctamente sujeitos com patologia utilizando os métodos apresentados o que leva obviamente a erros.

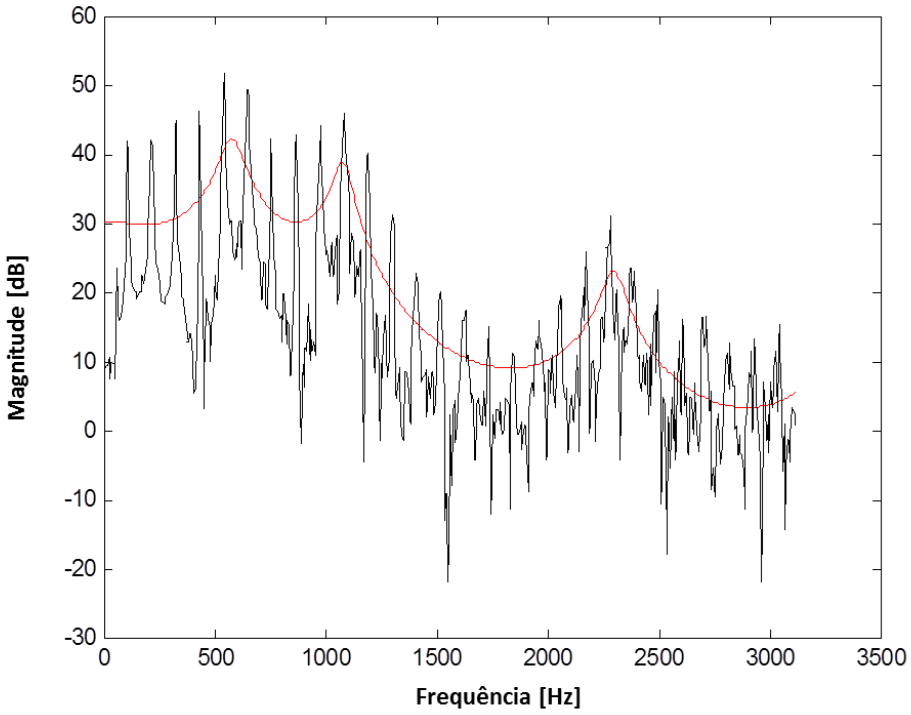

Figura 4-14 - Espectro e envolvente espectral do orador não saudável nº180, $RPPC_{med}$=0,984 (não foi detectado primeiro pico) , r_1=0,979, não foi detectado pelo sistema.

O primeiro pico da envolvente espectral também pode modelar apenas a primeira harmónica da frequência fundamental. Neste caso, a energia da primeira harmónica prevalece sobre as restantes podendo as restantes harmónicas serem ocultadas por ruído. Nestes casos, a largura de banda do primeiro pico do LPC é normalmente menor do que quando é modelada a primeira e a segunda harmónica. Ocorre também num número inferior de casos que seja modelada apenas a segunda harmónica e tipicamente a largura de banda tende a aumentar ainda mais. Para que seja detectado o primeiro pico do LPC é necessário que a energia das primeiras harmónicas, nomeadamente a primeira e a segunda, tenham uma amplitude superior às restantes, sendo que a amplitude do primeiro pico de LPC é tipicamente superior à amplitude do primeiro formate. De notar que, para a detecção deste pico, não foi utilizada pré-enfase que tem por objectivo aumentar a amplitude do espectro nas altas frequências. No entanto, há que ter em conta que em casos de intenso ruído nas harmónicas das médias e

altas frequências com a diminuição do declive espectral pode não ser possível a detecção do primeiro pico do LPC.

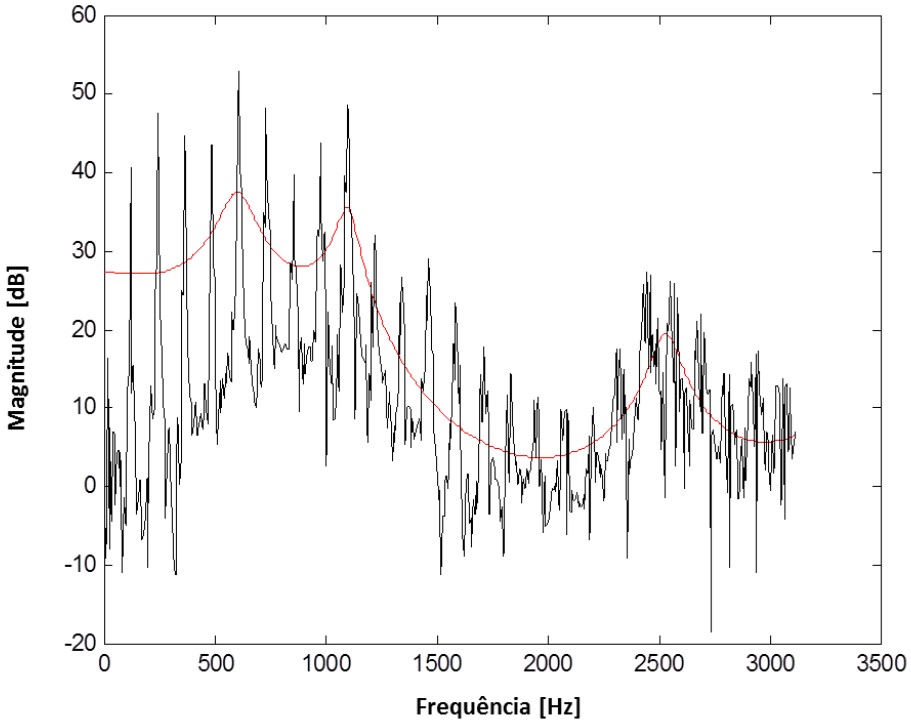

Figura 4-15- Espectro e envolvente espectral do orador saudável nº 15 com espectro muito idêntico ao anterior, $RPPC_{med}$=0,979, r_1=0,972.

O facto de não ser detectado o primeiro pico do LPC em oradores com patologia das pregas vocais, sendo que nestes casos também foi diagnosticado um valor baixo do RPPC, indica que o valor do primeiro pico do LPC pode ser de facto um diagnóstico preliminar no diagnóstico de patologias das pregas vocais. Uma vez que este primeiro pico está relacionado com as harmónicas da frequência fundamental, em fases avançadas da doença pode não existir vozeamento, sendo que este facto impossibilita a detecção do primeiro pico da envolvente espectral.

4.5 Conclusões

Este capítulo apresentou a caracterização de vozes patológicas através da envolvente espectral para uma ordem de LPC de 30. Com esta ordem de LPC foram apresentadas novas características para avaliação e reconhecimento de vozes com patologias. Com os vários testes efectuados em duas bases de dados diferentes é permitido concluir que:

a) Existe um primeiro pico na envolvente espectral que modela a frequência das duas primeiras harmónicas e que contêm informação das vozes patológicas.

b) A caracterização de vozes patológicas usando o valor da frequência do primeiro pico da envolvente espectral e a largura de banda deste formante permite classificar 100% dos sujeitos (47 no total, dos quais 31 com patologia) da base de dados da Universidade de São Paulo.

c) Estas características permitem classificar correctamente 79,1% dos sujeitos (206 no total, dos quais 153 com patologias) da base de dados da MEEI.

d) Verifica-se que as mesmas características não são dependentes do fonema.

e) A introdução do RPPC como medida de ruido e de aperiodicidade do sinal, e a diferença entre valores de frequência dos formantes aumentou a taxa de reconhecimento na base de dados da MEEI de 79.1% para 93,2%.

f) A análise dos espectros das vozes patológicas indica que as características baseadas na envolvente espectral podem ser um método de diagnóstico preliminar antes da degradação efectiva da qualidade vocal.

5 Sistemas de identificação de patologias da voz baseados em fala contínua

Analisando os trabalhos descritos no estado da arte verifica-se que o reconhecimentos de vozes patológicas através de sinais de fala utiliza com maior frequência a vogal /a/, com menor frequência outras vogais e esporadicamente fala contínua. Num dos casos utilizando fala contínua [32], apenas foram usados os segmentos vozeados porque é conhecida a dificuldade na detecção dos segmentos vozeados precisamente porque o vozeamento deficiente é uma das características das vozes patológicas. Verifica-se também em [30], onde foram usados segmentos vozeados e não vozeados, que a taxa de reconhecimento de vozes patológicas com fala contínua e a vogal /a/ foram idênticos.

Por outro lado, visto que a identificação de patologias laríngeas utilizando processamento de sinais de fala se encontra numa fase insipiente ainda com poucos trabalhos publicados, a avaliação da fala contínua ainda não foi devidamente explorada. No entanto, é opinião de vários otorrinolaringologistas e terapeutas da fala que é possível, e até mais fácil, através da audição de fala contínua diagnosticar qual a patologia que afecta o paciente. Este facto sugere que características perceptivas do sinal de fala contínua poderão ter informação relevante no reconhecimento de patologias da voz.

A abordagem mais comum no diagnóstico de patologias laríngeas consiste na detecção de patologias. Esta por sua vez reside num processo de verificação onde se prossupõe à partida que o sujeito já tem uma determinada patologia. O resultado produzido pelo sistema de reconhecimento é posteriormente comparado com um limiar, sendo o diagnóstico realizado com base nessa comparação. De facto, assumir *a priori* que um sujeito tem uma determinada patologia pode não ser a abordagem mais correcta. Poderá ser preferível optar por um sistema de identificação, sendo neste caso o sinal de fala comparado com sinais de fala de diversas patologias e tendo como resultado um diagnóstico efectuado entre esse conjunto de patologias. Numa situação de rastreio este método será certamente uma mais-valia, uma vez que poderá ser realizado por um individuo que não seja especialista.

Neste capítulo são apresentados os resultados de vários sistemas que comparam os resultados do diagnóstico de patologias usando fala contínua e vogais sustentadas. Estes sistemas avaliam

sujeitos pertencentes a três classes: saudáveis; diagnosticados com patologias laríngeas fisiológicas (edema e nódulos); e diagnosticados com patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). Os vários sistemas implementados são avaliados com base em várias características do sinal de fala e em 4 tipos de reconhecedores. Para optimizar a taxa de reconhecimento, os melhores sistemas individuais são combinados em árvore.

Finalmente, de modo a ser avaliado o impacto da diminuição da largura de banda no sinal na identificação de vozes patológicas e na identificação de patologias da voz, todos os sinais de fala são reamostrados com uma frequência de 8 kHz. O capítulo termina com um resumo do trabalho efectuado e com as conclusões obtidas.

5.1 Materiais e métodos

O estudo da identificação de patologias desenvolvido com base em sinais de fala contínua consiste, numa primeira abordagem, na implementação de quatro reconhecedores usando SVM, GMM e duas implementações de Discriminador Linear. Os parâmetros MFCC foram extraídos do sinal de fala para a vogal /a/ e para fala contínua.

De seguida, são apresentadas as bases de dados utilizadas assim como a descrição da implementação dos sistemas e as medidas usadas para a avaliação dos sistemas.

5.1.1 Base de dados

Os vários sistemas implementados para identificação de patologias laríngeas avaliaram sujeitos repartidos por três classes: saudáveis, diagnosticados com patologias fisiológicas e diagnosticados com patologias neuromusculares. Para o efeito foi criado um subconjunto da base de dados da MEEI composto por sujeitos pertencentes às três classes conforme a Tabela 5-1.

Para todos os sujeitos seleccionados foram adquiridos sinais de fala contínua (*rainbow passage*) e da vogal /a/ sendo a frequência de amostragem de todos os sinais de 25 kHz. Para manter a uniformidade deste parâmetro nos 53 sujeitos saudáveis, 17 foram descartados pois nesses casos os sinais em fala contínua foram adquiridos com uma frequência de amostragem de 10 kHz. Foram seleccionados 59 sujeitos com patologias fisiológicas, diagnosticados com edemas e nódulos. Associou-se estas duas patologias pois são, a seguir à paralisia das pregas

vocais, aquelas que têm maior ocorrência na base de dados e também porque o edema pode ser uma patologia preliminar dos nódulos. Note-se que existem mesmo dois sujeitos na base de dados a quem foram diagnosticadas as duas patologias simultaneamente. Por fim, para balancear os dados ao nível das patologias, foram seleccionados 59 sujeitos diagnosticados com paralisia unilateral.

Tabela 5-1 - Base de dados usados na identificação de patologias laríngeas.

Patologia		**Género Masculino**	**Género Feminino**	**Total**
Patologia Laríngea Fisiológica (PLF)	Nódulos	1	18	19
	Edema	9	31	40
Patologia Laríngea Neuromuscular (PLN)	Paralisia unilateral	29	30	59
Saudáveis		14	22	36

5.1.2 Implementação dos sistemas

Dada a quantidade limitada de dados disponíveis e assumindo que o ideal seria usar cerca de 3/4 dos dados para treino e 1/4 dos dados para teste, foi implementado o método de validação cruzada [81], o que permite que todos os dados sejam avaliados no conjunto de teste. Assim, foram criados 4 conjuntos de treino com os respectivos conjuntos de teste. Deste modo os resultados apresentados neste capítulo referem-se não a um mas a quatros sistemas, sendo que no total todos os sujeitos apresentados na Tabela 5-1 são avaliados.

Numa primeira fase, para avaliar o desempenho do sistema utilizando fala contínua na identificação de patologias, foram implementados 4 sistemas baseados em SVM, GMM, e dois tipos de Discriminadores Lineares. Foram usadas como características dos sinais de fala os MFCC, tipicamente usados em várias aplicações de reconhecimento, inclusive no reconhecimento de vozes patológicas.

Os resultados apresentados nas tabelas seguintes têm, na maior parte dos casos, os resultados dos quatro classificadores implementados. A implementação de 4 classificadores diferentes tem como objectivo verificar se os resultados obtidos por um classificador são consistentes

para as vogais e para fala contínua para as várias características extraídas do sinal de fala ou se poderão existir diferenças significativas entre os vários classificadores.

Para implementar um sistema de identificação de 3 classes baseado em SVM e dado que este é um classificador binário, a implementação foi realizada com recurso à metodologia *one-against-one*, (OAO) [82] obtendo-se assim três reconhecedores. Dois dos SVM foram treinados com os sujeitos saudáveis contra a classe de patologias laríngeas fisiológicas e contra a classe de patologias laríngeas neuromusculares, respectivamente. O terceiro SVM foi treinado entre as duas classes de patologias. Em testes preliminares verificou-se que seria mais vantajoso usar um *kernel* gaussiano. O teste de um sujeito no reconhecedor SVM consiste na classificação trama a trama. Um sujeito é considerado saudável se a média das tramas classificadas como saudáveis nos dois primeiros classificadores for superior a 50%. Caso contrário o sujeito é classificado como não saudável e o terceiro SVM diagnosticará a patologia.

A implementação do reconhecedor baseado em GMM é mais directa. Para cada classe é criado um modelo de misturas gaussianas. Neste caso serão criados 3 modelos, um para cada classe, de *N* misturas gaussianas. Todos os modelos foram treinados com valores de *N* compreendidos entre 4 e 64. As várias tabelas mostradas ao longo deste capítulo apresentam os melhores resultados obtidos, com a respectiva indicação de *N*. O teste consiste no cálculo da verosimilhança de todas as tramas das características do sujeito de teste para cada um dos modelos. O sujeito é classificado no modelo que obtiver maior verosimilhança.

Para o discriminador linear foram adoptadas duas abordagens. A primeira cria 3 classificadores lineares e usa a metodologia *one-against-one*, permitindo comparar resultados directamente com o sistema implementado com SVM. Neste caso o teste também seguiu a mesma sequência de utilização dos classificadores. Na segunda abordagem é implementado um discriminador linear com 3 classes, sendo que o sujeito é classificado na classe onde foram classificadas mais tramas.

Os parâmetros extraídos dos sinais de fala foram os MFCC+ΔMFCC, LSF+DLSF e MLSF+DMLSF. Adicionalmente foi adicionada energia e Δenergia no caso de serem usados MFCC. Todas as características foram extraídas de tramas com 20 ms de duração com um andamento de 10 ms. Para os sinais de fala contínua, as zonas de silêncio foram removidas usando o algoritmo

proposto por Lamel [83]. Para cada sujeito as características extraídas do sinal de fala foram normalizadas com média nula e variância unitária. Para todas as características foram testadas as ordens 8, 12, 16 e 20.

5.1.3 Medidas

A avaliação dos sistemas consiste na classificação dos sujeitos numa das 3 classes. Os resultados apresentados consistem na taxa de exactidão total do sistema (ACC - *overall accuracy*) na sensibilidade ou *true positive rate* (TPR) e na precisão ou *positive predictive rate* (PPV). Estas medidas são estimadas tipicamente para sistemas binários, ou seja envolvendo apenas 2 classes, sendo que são definidos: *true positive* (*TP*), *false positive* (FP), *false negative* (FN) e *true negative* (TN). Como os sistemas desenvolvidos avaliam três classes assume-se que quando uma classe é considerada positiva as outras duas são consideradas negativas. As expressões para o cálculo das diversas medidas são apresentadas de seguida.

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (5.1)$$

$$TPR = TP / (TP + FN) \quad (5.2)$$

$$PPV = TP / (TP + FP) \quad (5.3)$$

O uso destas medidas permite avaliar o comportamento dos sistemas para cada uma das classes. A medida de ACC, que mostra a taxa de reconhecimento do sistema, pode ocultar desequilíbrios entre a taxa de reconhecimento das classes. Assim, é importante verificar, dentro de cada classe através da medida TPR, o impacto do tipo de sinal de fala (vogal /a/ ou fala continua) e das características extraídas do sinal de fala. Por outro lado, é importante perceber quais as classes que têm maior correlação através da análise das medidas de TPR e PPV. Outra abordagem seria a apresentação de resultados com base nas matrizes de confusão dos sistemas. Contudo este método foi evitado, sendo usado apenas no ponto seguinte, pois dada a quantidade de sistemas analisados não se obteria uma análise eficiente dos resultados.

5.2 Identificação de patologias da voz analisando vogais e fala contínua

O primeiro estudo realizado usando fala contínua é apresentado em [8] e teve como principal objectivo verificar o desempenho dos sistemas utilizando fala contínua e usando MFCC na identificação de patologias laríngeas. Nesta primeira abordagem foram implementados apenas dois reconhecedores baseados em SVM e GMM tendo sido realizados vários testes, nomeadamente com o valor da ordem das características e com o número de misturas Gaussianas. Os melhores resultados são apresentados na Tabela 5-2 e na Tabela 5-3.

Neste estudo inicial pode constatar-se que a taxa de reconhecimento total dos sistemas implementados com fala contínua foi superior à obtida com a vogal /a/ verificando-se que existe também uma maior correlação entre as classes das patologias. Por outro lado, existe uma descorrelação evidente entre as classes das patologias e a classe de sujeitos saudáveis. Este resultado era previsível pois esses sujeitos já tinham sido usados em estudos anteriores onde se obtiveram taxas de acertos entre saudáveis e não saudáveis superiores a 90%.

Tabela 5-2 - Resultados para o classificador SVM com MFCC de ordem 20, para fala contínua e vogal /a/. Entre parêntesis resultados para a vogal /a/. ACC: 72%(69%).

Classificação / Original	**Saudáveis**	**Patologia Laríngea Fisiológica (PLF)**	**Patologia Laríngea Neuromuscular (PLN)**	**Sensibilidade (TPR)**
Saudáveis	34 (33)	0 (1)	2 (2)	**94,4%** (91,6%)
Patologia Laríngea Fisiológica (PLF)	1 (9)	44(32)	14 (18)	74,5% (54,2%)
Patologia Laríngea Neuromuscular (PLN)	2 (3)	24 (15)	33 (41)	56% **(69,4%)**
Precisão (PPV)	**91,8%** (73%)	64,7% **(66,6%)**	67,3% (67,2%)	

Analisando apenas os resultados entre patologias verifica-se que a fala contínua obtém melhores resultados na identificação de patologias fisiólogas e que a vogal /a/ obtém sempre os melhores resultados na identificação de patologias neuromusculares. Observa-se ainda que os sistemas implementados com fala contínua obtêm uma melhoria significativa na

sensibilidade e na precisão dos oradores saudáveis. No caso da precisão, os valores são no mínimo de 18% melhores em valor absoluto quando analisada a fala contínua.

Tabela 5-3 Resultados para o classificador GMM com 16 misturas com MFCC de ordem 20 para fala contínua e ordem 8 vogal /a/. Entre parêntesis resultados para a vogal /a/. ACC: 72%(69%).

Classificação / Original	**Saudáveis**	**Patologia Laríngea Fisiológica (PLF)**	**Patologia Laríngea Neuromuscular (PLN)**	**Sensibilidade (TPR)**
Saudáveis	33 (30)	3 (2)	0 (4)	91,6% (83,3%)
Patologia Laríngea Fisiológica (PLF)	1 (10)	45 (32)	13 (17)	**76,2%** (54,2%)
Patologia Laríngea Neuromuscular (PLN)	3 (4)	20 (15)	36 (40)	61% (67,7%)
Precisão (PPV)	89,1% (68%)	66,2% (65,3%)	**73,4% (**65,5%)	

Analisando apenas os resultados entre patologias verifica-se que a fala contínua obtém melhores resultados na identificação de patologias fisiólogas e que a vogal /a/ obtém sempre os melhores resultados na identificação de patologias neuromusculares. Observa-se ainda que os sistemas implementados com fala contínua obtêm uma melhoria significativa na sensibilidade e na precisão dos oradores saudáveis. No caso da precisão, os valores são no mínimo de 18% melhores em valor absoluto quando analisada a fala contínua.

Este estudo permite concluir que a fala contínua tem resultados promissores na identificação de vozes patológicas e na identificação de patologias laríngeas, mas que a utilização da vogal /a/ poderá também ter informação relevante na identificação de patologias. Outra conclusão é que é possível, através de características que modelam o tracto vocal (filtro), realizar a identificação de patologias que afectam as pregas vocais (fonte). Esta conclusão vai de encontro às teorias que assumem como hipótese que os pacientes, na presença destas patologias laríngeas, usam o tracto vocal para colmatar a incapacidade de realizar o correcto vozeamento.

5.3 *Identificação de patologias da voz analisando a informação dos formantes e características perceptivas*

Na sequência dos resultados obtidos no ponto anterior o estudo da aplicação da fala contínua no reconhecimento de patologias laríngeas avançou em duas vertentes. A primeira para verificar o seu desempenho em dois Discriminadores Lineares. A segunda para extrair novas características do sinal de fala nomeadamente os parâmetros LSF e MLSF. Os LSF permitem a caracterização directa do tracto vocal através da estimação dos formantes e da largura de banda. Os MLSF introduzem nestes últimos a informação perceptiva através de um banco de filtros de escala Mel. Os resultados deste trabalho foram apresentados em [2].

A principal motivação para os testes seguintes é a confirmação das conclusões obtidas no ponto anterior. Pretende-se aqui verificar que para os MFCC os desempenhos da fala contínua para os diversos classificadores em relação à vogal /a/ se mantêm consistentes, qual o desempenho das novas características no reconhecimento de patologias e ainda se este desempenho é consistente com os resultados obtidos com MFCC e com fala contínua.

Tabela 5-4 Resultados MFCC+Delta MFCC. Classificadores baseados em Discriminadores Lineares assinalados na tabela.

Classificador	**3 Classes GMM (16 Mixtures)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (8)	FC (12)	/a/ (8)	FC (12)	/a/ (20)	FC (12)	/a/ (20)	FC (20)
Saudável TPR	83,3	91,6	88,9	**97,2**	88,9	**97,2**	91,6	94,4
Saudável PPV	68	89,1	78,0	71,4	69,6	71,4	73	**91,8**
PLF TPR	54,2	**76,2**	69,5	62,7	57,6	64,4	54,2	74,5
PLF PPV	65,3	66,2	62,1	63,8	61,8	60,3	**66,6**	64,7
PLN TPR	67,7	61	59,3	54,2	64,4	52,5	**69,4**	56
PLN PPV	65,5	73,4	**74,5**	68,1	71,7	73,8	67,2	67,3
ACC	66	**74**	70,1	67,5	67,5	67,5	69	72

Resultados na tabela em [%]

Analisando os resultados obtidos na Tabela 5-4, verifica-se que os dois novos classificadores implementados obtêm resultados em linha com os apresentados anteriormente. De facto, a taxa de reconhecimento nos oradores saudáveis foi melhor para fala contínua quando usado o discriminador de 3 classes e a sensibilidade nos sujeitos com patologias neuromusculares foi superior na vogal /a/. Contudo, os resultados do ACC foram, em qualquer dos casos, inferiores aos obtidos anteriormente, não se conseguindo tirar conclusões relevantes.

Tabela 5-5 - Resultados LSF+DLSF

Classificador	**3 Classes GMM (8 Mix) (32 Mix)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (12)	FC (12)	/a/ (12)	FC (12)	/a/ (20)	FC (12)	/a/ (12)	FC (12)
Saudável TPR	69,4	97,2	55,6	**100**	77,8	**100**	69,4	**100**
Saudável PPV	75,8	92,1	74,1	83,7	70,0	78,3	67,6	**92,3**
PLF TPR	66,1	76,3	61,0	74,6	54,2	72,9	67,8	**81,4**
PLF PPV	65,0	**69,2**	63,2	65,7	61,5	66,2	63,5	67,6
PLN TPR	67,8	62,7	**78,0**	61,0	66,1	54,2	66,1	59,3
PLN PPV	65,6	72,5	65,7	76,6	55,7	74,4	72,2	**79,5**
ACC	67,5	76,0	59,7	73,4	64,3	72,1	67,5	**77,3**

Resultados na tabela em [%]

Por outro lado, na Tabela 5-5 e Tabela 5-6 os resultados mostram-se bem mais consistentes. Para todos os reconhecedores implementados, num total de 8, independentemente das características, os resultados do ACC são sempre superiores para a fala contínua do que para a vogal /a/. O melhor resultado obtido com fala contínua é de 77,9% usando MLSF em GMM, sendo este valor também superior que os resultados alcançados com MFCC. O sistema com fala contínua também demostra melhor desempenho na sensibilidade e precisão dos sujeitos saudáveis e na sensibilidade dos sujeitos com patologias fisiológicas. Por outro lado, também se verifica que a vogal /a/ obtém os melhores resultados na sensibilidade sobre os sujeitos com patologias neuromusculares.

Tabela 5-6 - Resultados MLSF+DMLSF

Classificador	**3 Classes GMM (8 Mix) (32 Mix)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (20)	FC (12)	/a/ (16)	FC (20)	/a/ (12)	FC (20)	/a/ (16)	FC (20)
Saudável TPR	80,6	**97,2**	58,3	94,4	72,2	97,2	80,6	97,2
Saudável PPV	70,7	**85,4**	80,8	85,0	63,4	79,5	72,5	87,5
PLF TPR	61,0	78,0	69,5	83,1	69,5	**81,4**	62,7	83,1
PLF PPV	69,2	**73,0**	64,1	69,0	66,1	68,6	67,3	70,0
PLN TPR	69,5	66,1	**72,9**	61,0	66,1	55,9	66,1	55,9
PLN PPV	67,2	78,0	67,2	**83,7**	76,5	82,5	66,1	75,0
ACC	68,8	**77,9**	68,2	77,3	68,8	75,3	68,2	76,0

Resultados na tabela em [%]

Analisando em particular os resultados das duas características LSF e MLSF verifica-se que, com a primeira se obtêm melhores resultados na discriminação de sujeitos saudáveis enquanto com a segunda se obtêm melhores resultados na discriminação de sujeitos com patologias fisiológicas. Os LSF obtêm 100% de acertos nos sujeitos saudáveis para três dos quatros discriminadores implementados. Os MLSF obtêm mais de 80% de acertos, em três dos quatro classificadores implementados, nos sujeitos diagnosticados com patologias fisiológicas.

Na comparação dos resultados do ACC para todos os reconhecedores, Figura 5-1, verifica-se que a fala contínua obtém os melhores resultados. Apenas em um dos reconhecedores implementados a taxa de reconhecimento é melhor usando a vogal /a/. Verifica-se também que os MFCC não conseguem obter melhores resultados que os LSF ou MLSF. Estes últimos conseguem três dos quatro melhores resultados nos valores do ACC.

Os resultados obtidos pelos vários sistemas permitem consolidar as conclusões obtidas no ponto anterior existindo, ao nível dos classificadores, pequenas diferenças que convém salientar. O GMM, treinado com as características perceptuais MFCC e MLSF, extraídas da fala contínua, obtêm os melhores resultados de ACC. Contudo, estes resultados estão dependentes dos valores de inicialização do treino dos modelos, pelo que os resultados podem ter algumas flutuações. Por outro lado, analisando a discriminação entre sujeitos saudáveis e não saudáveis, verifica-se que, para qualquer das características extraídas do sinal de fala, os

melhores resultados são obtidos com SVM e fala contínua. Por fim, verifica-se que para a vogal /a/, é o reconhecedor implementado com Discriminador Linear que obtém os melhores resultados usando LSF e MLSF. Estes dois últimos reconhecedores, SVM e Discriminador Linear, não sofrem das questões de inicialização que afectam os GMM. Nestes casos, os resultados permanecem sempre iguais desde que os dados de treino sejam iguais. Contudo, o GMM apresenta algumas vantagens uma vez que os modelos de treino são independentes entre si, ou seja, uma classe pode ser retreinada sem que isto afecte os modelos das outras classes. O sistema pode também aumentar o número de classes a reconhecer sem ter de ser integralmente retreinado, bastando para isso criar modelos para as novas classes e estas serem usadas no procedimento de teste.

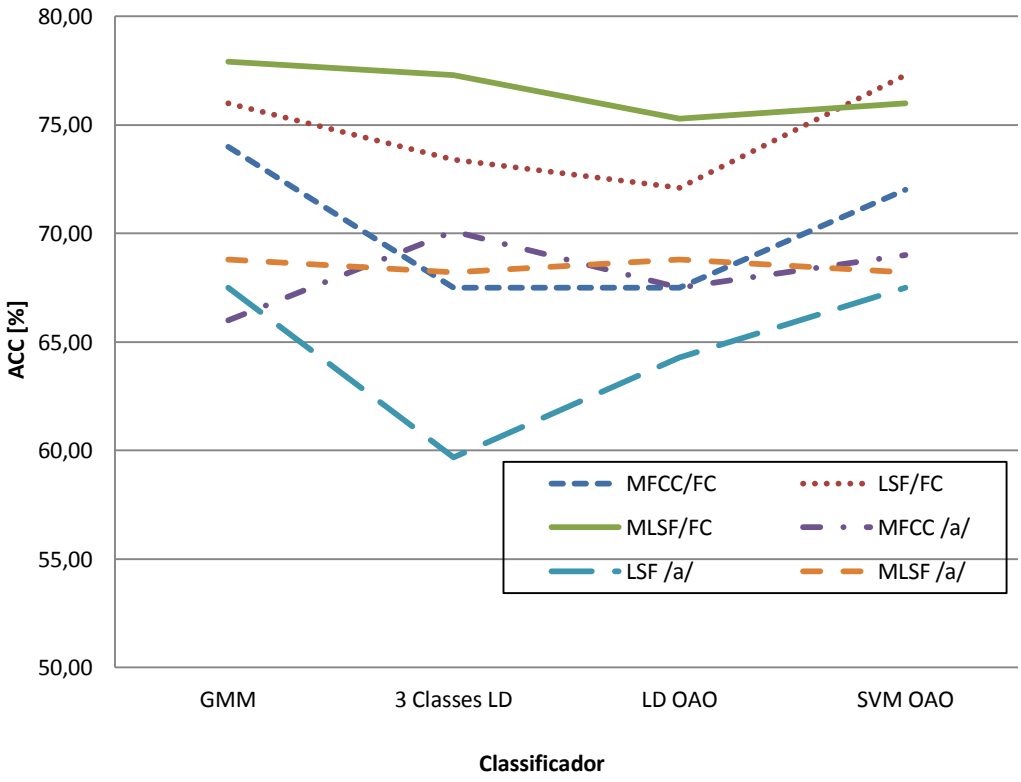

Figura 5-1 – Comparação dos ACC para todos os reconhecedores.

5.4 Classificador hierárquico e combinação de sistemas

Nos estudos apresentados anteriormente ficou patente que, apesar da análise do sinal de fala contínua apresentar, no geral, melhores resultados do que os obtidos com a análise da vogal /a/, esta última contém informação relevante que permite o diagnóstico de patologias neuromusculares. Tendo em conta esta conclusão e outras relacionadas com a questão do uso de características extraídas do sinal de fala com e sem informação perceptual, pode ser vantajosa a combinação de vários sistemas de modo a optimizar a taxa de reconhecimento. De modo a testar esta hipótese foram seleccionados e combinados os sistemas que obtiveram os melhores resultados para cada classe. Este trabalho foi apresentado em [3].

O desenho da topologia final é apresentado na Figura 5-2, onde é criado um classificador hierárquico com a combinação de vários classificadores. Nesta topologia, o primeiro nó classifica o sujeito como sendo saudável ou não saudável. Se o sujeito for classificado com sendo não saudável será avaliado por um segundo nó, no qual será diagnosticada a patologia.

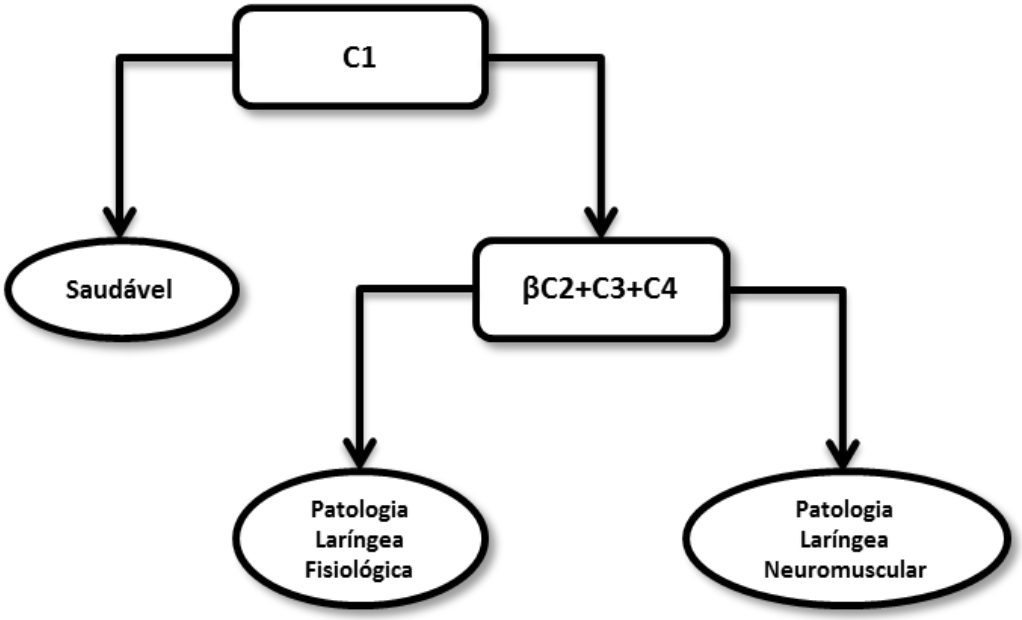

Figura 5-2 - Classificador hierárquico

O classificador C1 tem como objectivo diagnosticar oradores não saudáveis, sendo portanto melhor maximizar a precisão da classe saudáveis. Analisando os resultados das tabelas apresentadas no ponto anterior verifica-se na Tabela 5-5 que o sistema implementado com SVM, LSF e fala contínua obtém os melhores resultados nesta métrica. Contudo, os resultados

mostrados nas tabelas têm como objectivo a maximização do valor do ACC. Nos testes efectuados verifica-se que, com o mesmo sistema, se a ordem dos LSF for 16 em vez de 12 existe uma ligeira melhoria no valor da precisão dos oradores saudáveis subindo neste caso de 92,3% para 94,7%. Este valor traduz o diagnóstico errado de dois sujeitos com patologias neuromusculares.

O classificador C2 maximiza o diagnóstico de sujeitos com patologias fisiológicas. Neste caso o sistema implementado usa MLSF e fala contínua num discriminador de 3 classes (ver Tabela 5-6). Para maximizar o diagnóstico de patologias neuromusculares foram usados dois classificadores (C3 e C4). Ambos os classificadores usam Discriminadores Lineares de 3 classes e a análise da vogal /a/. No primeiro são usados MLSF e no segundo LSF. O resultado obtido pela combinação dos vários classificadores é apresentado na Tabela 5-7, onde são mostrados os valores obtidos por cada classificador assim como os melhores resultados obtidos por um sistema individual.

Tabela 5-7 - Resultados obtido com combinação de classificadores

Classificador	**SVM**	**3 Classes LD**			**GMM 32mix Melhor Sistema Individual**	**Combinação de Sistemas**
	C1	**C2**	**C3**	**C4**		**C1;C2;C3;C4**
Sinal de Fala (Ordem do parâmetro)	CS (LSF-16)	CS (MLSF-20)	/a/ (MSLF-16)	/a/ (LSF-12)	CS MLSF(12)	Figura **5-2**
Saudável TPR	**100**	94,4	58,3	55,6	97,2	**100**
Saudável PPV	**94,7**	85,0	80,8	74,1	85,4	**94,7**
PLF TPR	77,9	**83,1**	69,5	61,0	78,0	**83,1**
PLF PPV	63,8	69,0	64,1	63,2	73,0	**79,6**
PLN TPR	52,5	61,0	72,9	**78,0**	66,1	76,3
PLN PPV	70,4	**83,7**	67,2	65,7	78,0	81,8
ACC	73,3	**77,3**	68,2	59,7	77,9	**84,4**

Resultados na tabela em [%]

Analisando os resultados da tabela anterior verifica-se que existe um ganho expressivo na taxa de reconhecimento do sistema. O melhor resultado obtido com um sistema individual, GMM

com MSLF e fala contínua, apresentava 77,9% de acertos. Com a combinação de classificadores, o sistema apresenta 84,4% de acertos. A sensibilidade e precisão da classe dos sujeitos saudáveis obtém os mesmos valores que o classificador C1. A sensibilidade da classe dos sujeitos diagnosticados com patologias fisiológicas obtém também valores iguais a C2, sendo que o valor da precisão melhora. Apenas os resultados da sensibilidade e precisão da classe dos sujeitos diagnosticados com patologias neuromusculares não apresentam os melhores resultados que os sistemas C4 e C2, respectivamente.

Tabela 5-8 – Resultados na identificação de vozes patológicas, primeiro nó do classificador hierárquico.

Classificador	**C1**
Saudável TPR	100
Saudável PPV	97,7
Não Saudável TPR [%]	98,3
Não Saudável PPV[%]	100
ACC [%]	98,7

Resultados na tabela em [%]

Tabela 5-9 – Resultados na identificação de patologias da voz, segundo nó do classificador hierárquico.

Classificador	**Combinação de C2; C3; C4**
PLF TPR [%]	83,1
PLF PPV [%]	80,3
PLN TPR [%]	79,7
PLN PPV [%]	82,4
ACC [%]	81,3

Resultados na tabela em [%]

Considerando apenas o primeiro nó do sistema hierárquico, onde se classificam os sujeitos como saudáveis e não saudáveis, verifica-se que a taxa de reconhecimento nestas duas classes é de 98,7%. Neste caso apenas dois sujeitos não saudáveis são classificados como saudáveis, como se pode ver na Tabela 5-8.

Se por outro lado for descartado o primeiro nó que realiza a classificação entre oradores saudáveis e não saudáveis, e considerando apenas o segundo nó onde se realiza o diagnóstico de patologias, verifica-se que os dois sujeitos que são classificados erradamente como saudáveis são classificados correctamente no segundo nó. Isto significa que entre patologias,

com a associação dos três sistemas, C2, C3 e C4, se consegue aumentar a sensibilidade nas patologias neuromusculares de 76,3% para 79,7%, o que representa uma taxa de reconhecimento entre patologias de 81,3%, Tabela 5-9.

Conclui-se portanto, como seria de esperar, que a associação dos melhores classificadores individuais, como os melhores resultados em cada situação particular é uma opção vantajosa. Comparativamente com o melhor sistema individual, todas as métricas avaliadas obtiveram resultados superiores. Este tipo de opção, já implementada noutro tipo de sistemas de reconhecimento, nomeadamente no reconhecimento de orador, é certamente uma mais-valia na implementação de sistemas de diagnóstico de patologias laríngeas.

5.5 Impacto da diminuição da largura de banda nos sinais de fala nos sistemas de reconhecimento de patologias da voz

Neste ponto pretende-se estudar qual o impacto na taxa de reconhecimento nos reconhecedores implementados no ponto 5.3 usando os mesmos sinais mas amostrados a uma frequência de 8 kHz. Deste modo, pretende-se analisar qual a necessidade das frequências mais altas do espectro e o impacto destas no reconhecimento de patologias, nomeadamente nos sistemas que usam fala continua.

Os sinais de fala com largura de banda de 12,5 kHz foram filtrados e reamostrados ficando com uma largura de banda de 4 kHz. Estes sinais foram analisados com os mesmos reconhecedores (GMM,SVM e DL) e com os mesmos parâmetros dos sinais de fala (MFCC, MLSF e LSF). Foram testadas várias ordens e foram testados os sinais de fala continua e a vogal sustentada /a/, sendo os resultados apresentados nas tabelas seguintes aqueles maximizam a taxa de reconhecimento.

Tabela 5-10 – Resultados dos sistemas com MFCC para sinais amostrados a 8 kHz.

Classificador	**3 Classes GMM (8 Mix) (32 Mix)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (16)	FC (20)	/a/ (12)	FC (16)	/a/ (12)	FC (20)	/a/ (12)	FC (20)
Saudável TPR	88,9	88,9	91,7	50	88,9	69,4	91,7	83,3
Saudável PPV	77,5	91,4	82,5	81,8	84,2	75,8	86,8	100
PLF TPR	74,6	79,7	72,9	83,1	74,6	76,3	66,1	57,6
PLF PPV	67,7	62,7	67,2	53,3	65,7	58,4	65	55,7
PLN TPR	66,1	54,3	61,1	49,2	61	49,2	61	42,4
PLN PPV	76,5	72,7	70,6	72,5	73,5	65,9	64,3	40,3
ACC	74,7	72,1	74,7	62,3	72,7	64,3	70,1	57,8

Resultados na tabela em [%]

Tabela 5-11 – Resultados dos sistemas com MLSF para sinais amostrados a 8 kHz.

Classificador	**3 Classes GMM (8 Mix) (32 Mix)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (20)	FC (12)	/a/ (24)	FC (12)	/a/ (20)	FC (20)	/a/ (16)	FC (20)
Saudável TPR	86,1	94,4	88, 9	88,9	88,9	91,7	88,9	75
Saudável PPV	77,5	87,2	86,5	78	78	70,2	80	75
PLF TPR	64,4	79,6	59,3	88,1	59,3	88,1	59,3	35,6
PLF PPV	69,1	67,2	63,6	56,5	71,4	61,1	66	42
PLN TPR	69,5	57,6	66,1	35,6	69,5	30,5	67,8	52,5
PLN PPV	69,5	75,5	62,9	65,6	64	81,8	65,6	45,6
ACC	71,4	74,7	68,8	68,2	70,1	66,9	69,5	51,3

Resultados na tabela em [%]

Tabela 5-12 – Resultados dos sistemas com LSF para sinais amostrados a 8 kHz.

Classificador	**3 Classes GMM (8 Mix) (32 Mix)**		**3 Classes LD**		**LD (OAO)**		**SVM (OAO)**	
Sinal de Fala (Ordem do parâmetro)	/a/ (20)	FC (20)	/a/ (20)	FC (20)	/a/ (20)	FC (20)	/a/ (20)	FC (12)
Saudável TPR	77,8	88,9	80,6	83,3	80,6	88,9	86,1	77,8
Saudável PPV	75,7	88,9	80,6	81,1	78,4	72,7	79,5	56
PLF TPR	69,5	79,7	59,3	84,7	57,6	81,4	67,8	52,5
PLF PPV	65,1	64,4	72,9	60,9	70,8	61,5	72,7	63,3
PLN TPR	64,4	57,6	76,3	47,4	76,3	40,7	72,8	52,5
PLN PPV	70,4	75,6	64,9	80	65,2	75	71,6	56,4
ACC	69,5	73,4	70,8	70,1	70,1	67,5	74	58,4

Resultados na tabela em [%]

Ao diminuir a largura de banda do sinal retira-se riqueza espectral eliminando-se frequências entre 4 e 12,5 kHz. Comparando os dois tipos de sinais testados, vogal /a/ e fala continua, verifica-se que a taxa de reconhecimento final dos reconhecedores que avalia fala continua desce ligeiramente. Por outro lado, os reconhecedores que realizaram o reconhecimento baseados na vogal /a/ tiveram um ligeiro aumento da taxa de reconhecimento, sobretudo devido a um aumento na identificação de sujeitos saudáveis, sem obter melhores resultados que os melhores sistemas que usam sinais com 12,5 kHz de largura de banda, como veremos de seguida. Na generalidade, sem particularizar o tipo de reconhecedor ou a característica do sinal de fala ou classe (patológica ou não) verifica-se que uma largura de banda menor tem impacto negativo na fala continua pois neste caso existem mais fonemas onde as diferenças espectrais entre estes variam consoante a patologia. No caso da vogal /a/ este impacto será menor parecendo que numa primeira analise esta abordagem poderá ter vantagens. Os melhores resultados obtidos representam uma taxa de reconhecimento de 74,7%, em três reconhecedores, dois dos quais usam MFCC e a vogal /a/ como sinal de fala e o outro usa MLSF com fala contínua. Comparativamente, nos reconhecedores implementados com sinais amostrados a 25 kHz o melhor resultado obtido foi 77,9%, Tabela 5-13, e existem vários sistemas implementados que têm taxas de acertos superiores a 75% mas usando os sinais de fala continua com maior largura de banda.

Tabela 5-13 – Melhores sistemas para sinais a 8 kHz comparativamente com os melhores resultados a 25kHz

	Melhores sistemas a 8 kHz			**Melhor sistema 25 kHz**
Classificador	**GMM.**	**3 Classes LD**	**GMM**	**GMM**
Sinal de Fala (Parâmetro - Ordem)	/a/ MFCC-16	/a/ MFCC-12	FC MLSF-12	FC MSLF-12
Saudável TPR	88,9	91,7	94,4	**97,2**
Saudável PPV	77,5	82,5	**87,2**	85,4
PLF TPR	74,6	72,9	**79,7**	78,0
PLF PPV	67,7	67,2	67,1	**73,0**
PLN TPR	**66,1**	61	57,6	66,1
PLN PPV	76,5	70,6	75,6	**78,0**
ACC	74,7	74,7	74,7	**77,9**

Resultados na tabela em [%]

Observando a tabela anterior verifica-se que os sistemas implementados com MFCC e a vogal /a/ tiveram um desempenho inferior da identificação de vozes saudáveis e patológicas e na identificação de patologias fisiológicas. Os resultados na identificação de patologias neuromusculares neste sistema foram iguais ao melhor sistema que usa sinais amostrados a 25 kHz. O outro sistema que usa MLSF com fala contínua, amostrada a 8 kHz, até conseguiu dois dos melhores resultados, nomeadamente obteve a melhor taxa de identificação de patologias fisiológicas. Contudo obteve uma taxa de reconhecimento de patologias neuromuscular mais baixa comparativamente com o melhor sistema com fala amostrada a 25 kHz.

De facto pode verificar-se que nos resultados apresentados nas três primeiras tabelas deste ponto, que os reconhecedores implementados com fala contínua amostrada a 8 kHz ficaram desbalanceados ao nível da identificação de patologias, isto é a taxa de identificação de patologias fisiológicas aumentou mas a taxa de reconhecimento de patologias de neuromusculares diminuiu. Existem vários reconhecedores que independentemente das características do sinal de fala usadas têm taxas de identificação acima de 80% na identificação de patologias fisiológicas e resultados inferiores a 50% na identificação de patologias

neuromusculares, sendo o pior exemplo o Discriminador Linear implementado com MLSF que obteve 88% de acertos nos sujeitos com patologias fisiológicas e 30% de acertos nos sujeitos com patologias neuromusculares. Ou seja, está-se na presença de um sistema de classificação que não contribui para um diagnóstico fidedigno.

Verifica-se também que os sistemas implementados com a vogal /a/ amostrada a 8 kHz obtiveram melhoria do desempenho geral comparativamente aos sistemas implementados com o mesmo sinal amostrado a 25 kHz. Esta melhoria deve-se a um aumento do reconhecimento de sujeitos saudáveis, mas analisado a identificação patologias neuromuscular verifica-se que efectivamente esta taxa de reconhecimento diminui nestes reconhecedores quando analisam o mesmo sinal com largura de banda de 8 kHz. Recorde-se que este sinal foi usado no ponto anterior na implementação de um sistema hierárquico, para melhorar o diagnóstico de sujeitos com patologias neuromusculares. Com sinais de 8 kHz verifica-se que, independentemente do reconhecedor ou das características do sinal de fala, a taxa de identificação de patologias neuromusculares diminui, uma vez que a eliminação das altas frequências acaba por eliminar questões espectrais relativas à soprosidade no sinal de fala, prejudicando a identificação desta patologia.

A diminuição da largura de banda dos sinais tem como principal consequência a eliminação do ruido das altas frequências, representativo da soprosidade nas vozes patologias, que se encontra com maior predominância nas patologias neuromusculares estudas neste trabalho, a paralisia unilateral das pregas vocais. Como resultado a taxa de identificação nesta patologia, até mesmo nas vogais, diminui, assim como a identificação de vozes patológicas e sujeitos saudáveis quando usada fala continua. No entanto, os sistemas implementados com sinais amostrados 8 kHz tem um melhor desempenho quando comparados com os sistemas que usaram sinais amostrados a 25 kHz. Nos sistemas com fala contínua, verifica-se efectivamente, que o espectro das altas frequências contem informação que permite melhores taxas de reconhecimento e um melhor balanceamento dos reconhecedores, pois apesar dos sinais amostrados a 8 kHz obterem as melhores taxas de identificação nas patologias fisiologias, a identificação das patologias neuromusculares não obtém padrões aceitáveis.

5.6 Validação e discussão

Nos pontos anteriores foram descritos e avaliados vários sistemas de identificação de patologias da voz, identificando três classes: sujeitos saudáveis, sujeitos com patologias fisiológicas e sujeitos com patologias neuromusculares. Foram implementados quatro classificadores, cada um avaliando o sinal de fala contínua e a vogal /a/, através de três parâmetros MFCC, LSF e MLSF.

O primeiro propósito é avaliar o desempenho dos sistemas utilizando fala contínua no reconhecimento das patologias da fala. O facto de serem implementados vários reconhecedores e com várias características tem como intenção verificar se existe convergência neste sentido, ou seja, se de um modo geral os resultados obtidos pela análise da fala contínua são superiores aos resultados obtidos pela análise da tradicional vogal /a/. Esta tese é suportada pelo facto de profissionais ligados ao diagnóstico de patologias da voz entenderem ser possível prever o diagnóstico através da audição da fala do paciente.

Também se pretende entender se os parâmetros do sinal de fala que modelam o trato vocal permitem efectuar a identificação de patologias laríngeas. Apesar da patologia laríngea afetar o desempenho das pregas vocais, ou seja, a fonte, alguns autores indicam que os parâmetros do sinal de fala que contêm informação do trato vocal, ou seja do filtro, contém informação que permite caracterizar a patologia. A análise de vários parâmetros do sinal de fala visa entender se parâmetros nunca testados no reconhecimento de patologias como os LSF e os MLSF, que contêm a informação directa dos formantes e da sua largura de banda, contêm informação que permita a identificação de patologias da voz. Esta tese indica que, apesar das patologias analisadas estarem presentes nas pregas vocais, um sujeito afectado por uma patologia poderá usar o tracto vocal para compensar a desvantagem provocada pelo mau funcionamento das pregas vocais. Também é sabido, até pela análise do capítulo anterior, que o vozeamento deficiente ou a sua ausência poderá afectar a relação sinal ruído, o declive espectral e a energia dos formantes.

Os resultados apresentados mostram que a fala contínua tem potencial na identificação de patologias laríngeas, pois nos parâmetros de fala usados demostrou obter resultados superiores aos obtido com a vogal /a/. Apenas num dos classificadores os resultados na taxa de reconhecimento foi inferior na fala contínua em relação à vogal /a/. Nos restantes

classificadores os resultados da fala contínua foram iguais ou superiores. Analisando em detalhe os resultados obtidos verifica-se que a fala contínua obtém sistematicamente melhores resultados na classificação de sujeitos saudáveis e de sujeitos com patologias fisiológicas. Por outro lado, os sistemas que analisam a vogal /a/ obtêm melhores resultados na identificação de sujeitos diagnosticados com patologias neuromusculares.

Analisandos os resultados obtidos pelos vários parâmetros do sinal de fala verifica-se que, no geral, os MLSF obtêm o melhor desempenho, seguidos pelos LSF com uma diferença inferior a 1%. De facto os MLSF obtêm os melhores resultados na taxa de reconhecimento total em três dos quatro classificadores implementados. Apenas no classificador implementado com SVM os MLSF apresentaram resultados inferiores aos LSF. Os MFCC, parâmetros tipicamente usados em aplicações de reconhecimento de fala, obtiveram sempre resultados inferiores aos outros dois parâmetros. Os LSF obtêm melhores resultados na identificação de sujeitos saudáveis uma vez que em três dos quatros sistemas implantados com fala contínua a taxa de identificação é 100%. Os MLSF obtêm a melhor taxa de identificação no diagnóstico de sujeitos com patologias fisiológicas, com dois sistemas a obterem taxas de identificação acima de 83%. Nos resultados deste parâmetro na identificação de sujeitos saudáveis apenas errou em dois ou três oradores dependendo dos casos. Este facto, aliado à melhor taxa de identificação de sujeitos com patologias fisiologias, permitiu aos MLSF obterem os melhores resultados na taxa de reconhecimento total.

Dado que os LSF e MLSF contêm informação directa dos formante e consequentemente do tracto vocal, este facto pode indiciar que efectivamente os sujeitos com patologias podem recorrer ao tracto vocal para colmatar o handicap provocado pelas pregas vocais. Analisando os resultados da taxa de reconhecimento total este facto é mais evidente quando é usada fala contínua. A vogal /a/ é mais propícia para a identificação de sujeitos diagnosticados com patologias neuromusculares.

Ao nível dos classificadores deve-se destacar o desempenho do classificador implementado com discriminador linear de 3 classes, que usando LSF e MLSF e analisando a vogal /a/ obteve os melhores resultados na identificação de patologias neuromusculares, com 78% e 72,9% respectivamente.

Os parâmetros LSF, que não contêm informação perceptiva, obtêm excelentes resultados na identificação de vozes patológicas e obtêm o melhor resultado na identificação de sujeitos diagnosticados com patologias neuromusculares. Por outro lado, os parâmetros MLSF, que contêm informação perceptiva, obtêm melhores resultados da identificação nos sujeitos diagnosticados com patologias fisiológicas e obtêm resultados ligeiramente inferiores (cerca de 5% em valor absoluto) na identificação de sujeitos diagnosticados com patologias neuromusculares comparativamente com LSF. Aparentemente, se apenas se levar em consideração as duas classes de patologias, os parâmetros MLSF obtêm uma ligeira vantagem no valor médio na detenção de patologias, nomeadamente quando extraídos do sinal de fala contínua. Este facto pode indiciar que a informação perceptiva pode ser determinante na identificação de patologias.

De modo a melhorar a identificação de patologias, foi implementado um classificador hierárquico, que tira partido das vantagens individuais descritas anteriormente. Na implementação deste classificador foram utilizados dois classificadores com LSF e dois classificadores com MLSF. Um dos classificadores, implementado com LSF e fala contínua, é responsável pela identificação de vozes patológicas num primeiro nó. No segundo nó, são combinados três discriminadores lineares que usam como características LSF com a vogal /a/, MLSF com a vogal /a/ e fala contínua. Nos resultados obtidos verifica-se que cinco das sete métricas analisadas obtêm resultados melhores ou iguais aos resultados dos classificadores individuais. As duas métricas onde tal não acontece são precisamente na sensibilidade e na precisão dos sujeitos diagnosticados com patologia neuromuscular mas com diferenças inferiores a 2% no valor absoluto. De notar que o classificador hierárquico identifica correctamente todos os sujeitos saudáveis, mas falha na identificação de dois sujeitos não saudáveis. Estes dois sujeitos diagnosticados com patologias neuromusculares, caso fosse eliminado o primeiro nó, seriam correctamente identificados no nó que identifica as patologias. Verifica-se assim que efetivamente a implementação do classificador hierárquico, assim como a combinação de vários classificadores, introduz uma mais-valia no diagnóstico de patologias da voz.

No decorrer desta investigação surge uma publicação [51] no reconhecimento de patologias da voz que propõem um sistema baseado em GMM e parâmetros ceptras retirados de fala continua no reconhecimento de patologias da fala. Nessa publicação os autores reconhecem

que a primeira contribuição apresentada neste capítulo [8] foi a primeira a propor sinais de fala continua no reconhecimento de patologias da voz. Apesar de os trabalhos não usarem a mesma metodologia ao nível das patologias analisadas, pois em [51] são analisadas cinco patologias (pólipos, nódulos, paralisia, disfonia espasmódica e queratoses) no total de 173 sujeitos, e não é usada por exemplo edemas que é a segunda patologia mais comum na base de dados da MEEI. Não está claro também se são retirados sujeitos com múltiplas patologias. Esta abordagem corresponde em algumas patologias a três sujeitos no conjunto de teste. Nos trabalhos apresentados neste capítulo analisam-se duas classes de patologias com 59 ficheiros cada, o que corresponde a quinze ou catorze sujeitos no conjunto de teste. Neste contexto a combinação de classificadores apresentada em [3] apresenta 79,7% de taxa de reconhecimento na paralisia unilateral contra 78% na mesma patologia em [51].

Finalmente foi realizado um estudo para entender de que modo as frequências mais altas dos sinais tem impacto no reconhecimento das patologias da voz. Assim os sinais de fala foram reamostrados para 8 kHz. A diminuição da largura de banda dos sinais tem como principal consequência a eliminação do ruido das altas frequências, representativo da soprosidade nas vozes patologias, que se encontra com maior predominância nas patologias neuromusculares estudas neste trabalho, a paralisia unilateral das pregas vocais. Como resultado a taxa de identificação nesta patologia diminui, assim como a identificação de vozes patológicas e sujeitos saudáveis quando usada fala continua. No entanto estes sistemas implementados com sinais amostrados 8 kHz tem um melhor desempenho quando comparados com os sistemas que usam sinais amostrados a 25 kHz. Nos sistemas com fala contínua, verifica-se que efectivamente, o espectro das altas frequências contem informação que permitem melhores taxas de reconhecimento e um melhor balanceamento dos reconhecedores, pois apesar dos sinais amostrados a 8 kHz obterem as melhores taxas de identificação nas patologias fisiologias, a identificação das patologias neuromusculares não obtém padrões aceitáveis.

5.7 Conclusões

Este capítulo abordou o tema de reconhecimento de patologias da voz. Foram construídos vários sistemas individuais que avaliaram três: sujeitos saudáveis; sujeitos diagnosticados com patologias laríngeas fisiológicas (edemas e nódulos) e sujeitos diagnosticados com patologias laríngeas neuromusculares (paralisia unilateral das pregas vocais). A investigação pretendeu avaliar dois tipos de sinais de fala, a vogal /a/ tipicamente usada em reconhecimento de vozes

patologias e o sinal de fala continua, proposto nesta tese. Paralelamente foram avaliadas as características do sinal da fala, MFCC, MLSF e LSF. Com estes parâmetros consegue-se avaliar a utilização de informação perceptual ou não perceptual e a utilização directa de informação sobre os formantes.

Os sistemas individuais foram combinados de modo a optimizar os resultados obtidos originando um sistema hierárquico que permite num primeiro nó realizar o reconhecimento de voz patológicas e num segundo nó realizar o reconhecimento de patologias. Este sistema hierárquico permite ainda a introdução de novas patologias de teste sem ter que se retreinar todo o sistema.

As principais conclusões do trabalho apresentado neste capítulo são:

a) Os LSF, que contêm informação directa dos formantes, quando extraídos de sinais de fala contínua apresentam a melhor taxa de reconhecimento na identificação de vozes patológicas.

b) Os MLSF, que contêm informação perceptiva, extraídos da fala contínua apresentam a melhor taxa de reconhecimento nos sistemas individuais com 77,9%. Paralelamente apresentam o melhor resultado na identificação de patologias laríngeas fisiológicas com 83,1%.

c) Os MFFC, tipicamente usados em aplicações de reconhecimento de fala, não apresentam nenhum dos melhores valores de reconhecimento nas métricas avaliadas.

d) A vogal /a/ apresenta os melhores resultados na identificação de sujeitos com patologias laríngeas neuromusculares, com 78%. Este facto indicia que o esforço na produção de vozeamentos sustentados pode ajudar na detecção desta patologia.

e) No geral pode-se concluir através dos parâmetros do sinal de fala estudados que apesar das patologias estarem presentes na laringe, existem perturbações no espectro devido ao esforço vocal e pela possível compensação do tracto vocal devido à ineficiência da laringe, que permite o reconhecimento de patologias da voz.

f) O sistema hierárquico proposto apresenta uma melhoria na taxa de reconhecimento de 77,9% para 84,4%, assim como uma melhoria em 5 das 7 métricas avaliadas. Este

sistema tem ainda a vantagem de ser modular e facilitar a ampliação para outras patologias.

g) A diminuição da largura de banda dos sinais de fala tem um impacto negativo na taxa de reconhecimento, principalmente nos sistemas que analisam fala contínua, significando que altas frequências contêm informação revelante para o reconhecimento das patologias da voz.

6 Conclusões

Esta tese investigou novas abordagens no reconhecimento de vozes com patologias e também na identificação de patologias da voz usando sinais de fala. Neste último ponto, onde se pretende identificar a patologia presente nas pregas vocais, continuam a existir muito poucas contribuições comparativamente ao reconhecimento de vozes patológicas, onde apenas se pretender diagnosticar se o sujeito tem uma qualquer patologia na voz.

6.1 Visão geral da tese

O principal objectivo desta tese foi encontrar novas características no sinal de fala que permitam identificar patologias da voz, quer sejam parâmetros do sinal de fala ou diferentes tipos de sinais de fala como seja a análise de vogais sustentadas ou de fala contínua. Existe, no entanto, uma limitação nos tipos de sinais de fala devido à inexistência de bases de dados com sinais provenientes de sujeitos diagnosticados com patologias da voz. Apesar disso, conseguiu-se utilizar duas bases de dados de referência no reconhecimento de patologias da voz: a base de dados da Universidade de São Paulo e a base de dados da MEEI.

As contribuições produzidas nesta tese podem ser divididas em duas partes: caracterização de vozes patológicas através da envolvente espectral, descrita no Capítulo 4 e sistemas de identificação de patologias laríngeas baseados em fala contínua descrita no Capítulo 5. O penúltimo ponto de cada um destes capítulos faz uma análise detalhada da validação e discussão dos resultados obtidos. O último ponto de cada capítulo apresenta um resumo das conclusões consideradas mais relevantes.

A primeira parte, descrita no capítulo 4, apresenta contribuições na identificação de voz ao nível da análise espectral, na detecção de um pico no espectro relacionado com a energia da primeira e segunda harmónica e na relação desta energia com o ruído presente no espectro. O estudo envolveu duas bases de dados de sinais de fala contendo sujeitos saudáveis e sujeitos diagnosticados com várias patologias laríngeas como edemas, pólipos, nódulos, paralisia nas pregas vocais e queratoses. O método apresentado indica a presença de um pico na envolvente espectral antes do primeiro formante. Este pico é facilmente detectado mesmo em vozes patológicas com valores de relação harmónica-ruído semelhantes às vozes de sujeitos

saudáveis. Esta característica está também presente em vozes soprosas onde a presença de ruído não atinge níveis considerados relevantes. Nos casos em que a soprosidade é evidente, verifica-se a degradação da componente harmónica do sinal e a diminuição do declive espectral, pelo que em alguns casos não é possível a detecção do primeiro pico que modela a primeira e a segunda harmónica do sinal. Verifica-se, de facto, que em vozes com soprosidade elevada a relação HNR baixa não sendo possível detectar este pico. Outro facto importante que também se verificou, é que este pico está presente não só na vogal /a/ utilizada normalmente neste tipo de estudos, como também nas vogais /e/ e /i/, sendo portanto uma característica independente do fonema.

Este método de identificação de vozes com patologias permite um processo simples e eficaz na identificação de sujeitos com patologias da voz, que pode ser particularmente útil em situações de rastreio onde não esteja presente um médico especialista. O método apresentado nesta tese, baseado em árvores de decisão, permite taxas de acertos superiores a 95%, o que está em linha com todos os outros trabalhos apresentados para as mesmas bases de dados.

A segunda parte desta tese, descrita no capítulo 5, consiste em demonstrar que na fala contínua, as características que modelam o tracto vocal e as características perceptivas do sinal de fala têm contribuições significativas na identificação de vozes com patologias e na identificação das patologias. Neste estudo foi usada apenas a base de dados da MEEI pois é a única comercialmente disponível que contém sinais de fala contínua. Desta base de dados foram criadas três classes de sujeitos repartidos por sujeitos saudáveis, sujeitos com patologias fisiologias (edema e nódulo) e sujeitos com patologias neuromusculares (paralisia unilateral das pregas vocais). De facto, verifica-se que os sujeitos com patologias podem usar o tracto vocal para compensar a incapacidade glotal. Por outro lado, a mesma patologia, ou patologias idênticas do ponto de vista funcional, tendem a produzir perturbações semelhantes no sinal de fala. Neste caso, o uso de características perceptivas pode ser uma mais-valia quando se pretende fazer o diagnóstico de patologias. Para tal, foram desenvolvidos vários classificadores (SVM, GMM e Discriminadores Lineares), usando vários parâmetros do sinal de fala (MFCC, MLSF e LSF) e sinais de fala contínua e a vogal /a/. Verificou-se que o uso da fala contínua, juntamente com o parâmetro MSLF que contém informação perceptiva, obtém a melhor taxa de reconhecimento com 77,9%. Por outro lado, através da análise mais cuidada dos resultados, verifica-se que não se deve descartar a vogal /a/ no reconhecimento de patologias. De facto,

verifica-se que este sinal apresenta a melhor taxa de reconhecimento na identificação de oradores diagnosticados com paralisia unilateral das pregas vocais, quer usando MLSF, quer usando LSF. Este resultado pode ser explicado pelo facto da produção de uma vogal sustentada requerer um maior esforço ao nível das pregas vogais, originado um maior impacto nesse sinal de fala por parte desta patologia. Tendo em conta estas considerações, desenvolveu-se um sistema hierárquico, onde vários sistemas individuais são combinados de modo a optimizar os resultados. Este sistema tem a vantagem de ser modular e facilitar a ampliação para outras patologias. O sistema hierárquico obteve uma taxa de reconhecimento de 84,4%, obtendo melhores resultados em 5 das 7 métricas avaliadas quando comparados com os sistemas individuais.

6.2 Contribuições

No início desta investigação foram colocadas duas questões de investigação nas quais resultaram as seguintes contribuições.

Questão 1: O uso de parâmetros de modelação do trato vocal, usados em reconhecimento de fala/orador, serão fiáveis no reconhecimento de patologias da voz? Quais serão os parâmetros que permitem obter melhores resultados na discriminação de patologias?

Pode-se concluir que, efectivamente, existem parâmetros que modelam o tracto vocal que permitem a classificação de vozes patológicas e também a identificação de patologias e que os sujeitos que padecem de patologias da voz usam o tracto vocal para compensar o vozeamento deficiente das pregas vocais. No caso do primeiro pico do LPC, este modela a soprosidade das vozes, não sendo um típico formante do tracto vocal mas uma alteração atípica no espectro que permite diagnosticar vozes patológicas. Parâmetros como os MLSF e os LSF permitem diagnosticar sujeitos com patológicas laríngeas fisiológicas e neuromusculares com melhor desempenho que os MFCC. Os parâmetros MFCC são tipicamente usados em várias aplicações de reconhecimento de fala embora não transportem informação directa dos formantes. Os parâmetros MLSF e LSF têm essa informação implícita, pelo que se conclui que estas características têm impacto na identificação das patologias.

Questão 2: Será que o uso de vários fonemas, ou mesmo fala contínua, no reconhecimento de patologias da voz, pode aumentar a taxa de reconhecimento?

Quanto ao tipo de sinais de fala, verifica-se que a fala contínua obtém melhores resultados na identificação de vozes patológicas e na identificação de patologias fisiológicas. Por outro lado, a vogal /a/ obtém melhores resultados na identificação de patologias neuromusculares. Pode-se portanto concluir que, dependendo das patologias e do esforço vocal, ambos os sinais têm informação relevante, sendo que não se deve *a priori* descartar nenhum dos sinais. Assim esta tese propõe um reconhecedor onde é usada a combinação de sistemas usando fala contínua e vogais sustentadas.

6.3 Disseminação de resultados

A investigação e consequentemente as contribuições deram origem a uma publicação em revista indexada na *Science Citation Index*, e seis artigos publicados em conferências científicas internacionais, sendo duas delas indexadas na Scopus e uma na Scopus e na PubMed. Estas publicações encontram-se apresentadas na Tabela 6-1.

Tabela 6-1 – Resumo das publicações de artigos científicos.

Ano	Título	Tipo
2016	H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Hierarchical classification and system combination for automatically identifying physiological and neuromuscular laryngeal pathologies ," *J. Voice*, Elsevier	Artigo de revista
2015	H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Voice Pathologies Identification Speech: signals , features and classifiers evaluation," 19th Conf. SPA 2015 Signal Process. Algorithms, Archit. Arrange. Appl., IEEE	Artigo de conferência
2015	Hugo Cordeiro, Isabel Guimarães, José Fonseca, Carlos Meneses, "Diagnosis of larynx pathologies using speech signal processing," First World Student Meeting on SLT, Revista Portuguesa de Terapia da Fala, Vol III	Artigo de conferência (resumo)
2015	H. Cordeiro, J. Fonseca, and C. Meneses, "Continuous Speech Classification Systems for Voice Pathologies Identification," in Technological Innovation for Cloud-Based Engineering Systems, Springer International Publishing, pp. 217–224.	Artigo de conferência
2014	H. Cordeiro, J. Fonseca, and C. Meneses, "Spectral Envelope and Periodic Component in Classification Trees for Pathological Voice Diagnostic," Eng. Med. Biol. Soc. (EMBC), 2014 36th Annu. Int. Conf. IEEE, pp. 4607–4610	Artigo de conferência

2014	H. Cordeiro, J. M. Fonseca, and C. M. Ribeiro, “Reinke’s Edema and Nodules Identification in Vowels Using Spectral Features and Pitch Jitter,” Procedia Technol., vol. 17, pp. 202–208, Elsevier	Artigo de conferência
2013	H. Cordeiro, J. M. Fonseca, and C. M. Ribeiro, “LPC Spectrum First Peak Analysis for Voice Pathology Detection,” Procedia Technol., vol. 9, pp. 1104–1111, Elsevier	Artigo de conferência

6.4 Desenvolvimento futuro

No desenvolvimento da investigação deparou-se com algumas limitações que não permitiram avançar em determinados sentidos. Ultrapassadas essas limitações, que se prendem essencialmente com a inexistência de uma base de dados de sinais de fala mais completa, os pontos seguintes propõem um conjunto de propostas como trabalho futuro.

1) Desenvolver um sistema de reconhecimento de patologias da voz dependente do género. Os parâmetros do sinal de fala analisados nesta investigação são altamente dependentes do género. A separação de género *a priori* certamente que permitirá um aumento na taxa de reconhecimento.

2) Nos sinais de fala contínua usados nesta investigação, todos os oradores pronunciam a mesma frase, o que torna o sistema de fala contínua dependente do texto. Será também importante realizar testes no âmbito de sinais de fala independentes do texto. Neste caso os sistemas serão independentes do texto e sinais de fala em termos de material serão completamente descorrelacionados. No entanto, não são conhecidas base de dados para esse efeito.

3) A extensão do estudo da identificação de patologias da voz a outras patologias. No capítulo 5 desta tese foram estudadas duas classes de patologias. Estas patologias são representadas por 59 sujeitos por cada classe. Na base de dados disponível uma terceira classe teria menos de 20 sujeitos. Para se validarem os resultados de mais patologias, é necessário ter um maior número de vozes com patologias de modo a criar classes mais consistentes para a validação dos resultados obtidos.

4) A implementação de uma aplicação que permita o rastreio de patologias da voz baseado em sinais de fala. Tendo em conta que no reconhecimento de vozes patológicas são

apresentados vários sistemas com taxas de acerto acima de 90%, o desenvolvimento de uma aplicação de rastreio para ser usada por pessoal não especialista é uma mais-valia no diagnóstico precoce de patologias da voz.

Bibliografia

[1] I. Guimarães, *A Ciência e a Arte da Voz Humana*. ESSA – Escola Superior de Saúde de Alcoitão, 2007.

[2] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Voice Pathologies Identification Speech: signals , features and classifiers evaluation," *19th Conf. SPA 2015 Signal Process. Algorithms, Archit. Arrange. Appl.*, 2015.

[3] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Hierarchical classification and system combination for automatically identifying physiological and neuromuscular laryngeal pathologies," *J. Voice*, 2016.

[4] H. T. Cordeiro, J. M. Fonseca, and C. M. Ribeiro, "LPC Spectrum First Peak Analysis for Voice Pathology Detection," *Procedia Technol.*, vol. 9, pp. 1104–1111, 2013.

[5] H. T. Cordeiro, J. M. Fonseca, and C. M. Ribeiro, "Reinke's Edema and Nodules Identification in Vowels Using Spectral Features and Pitch Jitter," *Procedia Technol.*, vol. 17, pp. 202–208, 2014.

[6] H. Cordeiro, J. Fonseca, and C. Meneses, "Spectral Envelope and Periodic Component in Classification Trees for Pathological Voice Diagnostic," *Eng. Med. Biol. Soc. (EMBC), 2014 36th Annu. Int. Conf. IEEE*, pp. 4607–4610, 2014.

[7] H. Cordeiro, J. Fonseca, and C. Meneses, "Voice Pathology Detection and Identification using speech processing techniques," *DOCEIS*, 2012.

[8] H. Cordeiro, J. Fonseca, and C. Meneses, "Continuous Speech Classification Systems for Voice Pathologies Identification," in *Technological Innovation for Cloud-Based Engineering Systems*, Springer International Publishing, 2015, pp. 217–224.

[9] J. D. Gibson, "Speech Coding Methods, Standards, and Applications," *IEEE Circuits and Systems Magazine*, vol. 5, no. 4. pp. 30–49, 2005.

[10] B. S. Atal, V. Cuperman, and A. Gersho, *Speech and Audio Coding for Wireless and Network Applications*. Springer Science & Business Media, 2012.

[11] Y. Agiomyrgiannakis, “Vocaine the vocoder and applications in speech synthesis,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015, vol. 2015-Augus, pp. 4230–4234.

[12] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on hidden Markov models,” *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[13] W. Mattheyses and W. Verhelst, “Audiovisual speech synthesis: An overview of the state-of-the-art,” *Speech Commun.*, vol. 66, pp. 182–217, 2015.

[14] J. R. Bellegarda and C. Monz, “State of the art in statistical methods for language and speech processing,” *Comput. Speech Lang.*, vol. 35, pp. 163–184, 2015.

[15] T. Kinnunen and H. Li, “An Overview of Text-Independent Speak er Recognition : from Features to supervectors,” *Speech Commun.*, vol. 1, no. 1, pp. 12–40, 2009.

[16] G. Fant, *Acoustic Theory of Speech Production.* Mouton De Gruyter., 1970.

[17] C. Meneses, “Processamento de Fala,” *Inst. Super. Eng. Lisboa*, 2011.

[18] M. Gordon and P. Ladefoged, “Phonation types : a cross-linguistic overview,” 2001.

[19] K. Omori, H. Kojima, R. Kakani, D. H. Slavit, and S. M. Blaugrund, “Acoustic characteristics of rough voice: Subharmonics,” *J. Voice*, vol. 11, no. 1, pp. 40–47, 1997.

[20] M. D. O. Rosa, J. C. Pereira, and M. Grellet, “Adaptive estimation of residue signal for voice pathology diagnosis,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, 2000.

[21] M. E. and E. Infirmary, *Voice disorders database, (Version 1.03 cd-rom)*. Kay Elemetrics Corp., Lincoln Park, NJ, 1994.

[22] P. Lieberman, “Some acoustic measures of the fundamental periodicity of normal and pathologic larynges,” *J. Acoust. Soc. Am.*, vol. 35, pp. 344–353, 1963.

[23] S. Iwata, “Periodicities of pitch perturbations in normal and pathological larynges,” *Laryngoscope*, vol. 82, pp. 87–96, 1972.

[24] N. Pinto and I. Titze, “No TitleUnification of perturbation measures in speech signals,” *J. Acoust. Soc. Am.*, vol. 89, no. 3, pp. 1278–1289, 1990.

[25] M. Vasilakis and Y. Stylianou, “A mathematical model for accurate measure of jitter,” in *5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2007.

[26] L. C. Oliveira, D. G. Silva, and M. Andrea, “Jitter estimation algorithms for detection of pathological voices,” *EURASIP J. Adv. Signal Process.*, 2009.

[27] E. Yumoto, W. J. Gould, and T. Baer, “Harmonics-to-noise ratio as an index of the degree of hoarseness.,” *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1544–1550, 1982.

[28] P. J. Murphy, “Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis.,” *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2866–2881, 1999.

[29] C. Ferrer, E. González, M. E. Hernández-Díaz, D. Torres, and A. Del Toro, “Removing the influence of shimmer in the calculation of harmonics-to-noise ratios using ensemble-averages in voice signals,” *EURASIP J. Adv. Signal Process.*, 2009.

[30] A. a Dibazar, S. Narayanad, and T. W. Berger, “Feature Analysis for Automatic Detection of Pathological Speech,” *36th Asilomar Conf, Signal, Syst. Comput. 2002*, 2002.

[31] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, “Identification of voice disorders using long-time features and support vector machine with different feature reduction methods,” *J. Voice*, vol. 25, no. 6, 2011.

[32] V. Parsa and D. G. Jamieson, “Acoustic discrimination of pathological voice: sustained vowels versus continuous speech.,” *J. Speech. Lang. Hear. Res.*, vol. 44, no. 2, pp. 327–339, 2001.

[33] J. W. Lee, H. G. Kang, J. Y. Choi, and Y. I. Son, “An investigation of vocal tract characteristics for acoustic discrimination of pathological voices,” *Biomed Res. Int.*, 2013.

[34] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, “Methodological issues in the development of automatic systems for voice pathology detection,” *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, 2006.

[35] E. S. Fonseca, R. C. Guido, P. R. Scalassara, C. D. Maciel, and J. C. Pereira, “Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders,” *Comput. Biol. Med.*, vol. 37, no. 4, pp. 571–578, 2007.

[36] K. Shama, A. Krishna, and N. U. Cholayya, “Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology,” *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007.

[37] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, “Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, 2011.

[38] X. Wang, J. Zhang, and Y. Yan, “Discrimination between pathological and normal voices using GMM-SVM approach,” *J. Voice*, vol. 25, no. 1, pp. 38–43, 2011.

[39] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, “On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices.,” *Logoped. Phoniatr. Vocol.*, vol. 36, no. 2, pp. 60–69, 2011.

[40] A. a. Dibazar, T. W. Berger, and S. S. Narayanan, “Pathological voice assessment,” *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 1669–1673, 2006.

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

[42] P. T. Hosseini, F. Almasganj, T. Emami, R. Behroozmand, S. Gharibzade, and F. Torabinezhad, “Local discriminant wavelet packet basis for voice pathology classification,” *2nd Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2008*, pp. 2052–2055, 2008.

[43] E. Fonseca and J. Pereira, “Normal versus pathological voice signals,” *IEEE Eng. Med. Biol. Mag.*, vol. 28, no. 5, pp. 44–48, 2009.

[44] R. T. S. Carvalho, C. C. Cavalcante, and P. C. Cortez, “Wavelet transform and artificial neural networks applied to voice disorders identification,” *2011 Third World Congr. Nat. Biol. Inspired Comput.*, pp. 371–376, 2011.

[45] M. Markaki and Y. Stylianou, “Using modulation spectra for voice pathology detection and classification,” *Proc. 31st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. Eng. Futur. Biomed. EMBC 2009*, pp. 2514–2517, 2009.

[46] L. Atlas and S. A. Shamma, “Joint acoustic and modulation frequency,” *EURASIP J. Appl. Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[47] H. C. Peng, F. H. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[48] M. Markaki and Y. Stylianou, “Voice pathology detection and discrimination based on modulation spectral features,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 1938–1948, 2011.

[49] G. Muhammad, M. Alsulaiman, A. Mahmood, and Z. Ali, “Automatic voice disorder classification using vowel formants,” *Proc. - IEEE Int. Conf. Multimed. Expo*, pp. 0–5, 2011.

[50] G. Muhammad and M. Melhem, “Pathological voice detection and binary classification using MPEG-7 audio features,” *Biomed. Signal Process. Control*, vol. 11, no. 1, pp. 1–9, 2014.

[51] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, “Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model.,” *J. Voice*, 2015.

[52] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond*. 2005.

[53] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[54] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, 1995.

[55] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[56] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4. pp. 430–451, 2004.

[57] V. Wan, P. Street, S. Sheffield, and W. M. Campbell, “Support Vector Machines for Speaker Verification and Identification,” *Neural Networks Signal Process PROC IEEE*, vol. 2, pp. 775–784, 2000.

[58] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *Signal Process. Lett. IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[59] J. Mariéthoz and S. Bengio, “A kernel trick for sequences applied to text-independent speaker verification systems,” *Pattern Recognit.*, vol. 40, no. 8, pp. 2315–2324, 2007.

[60] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Computer Speech & Language*, vol. 20, no. 2–3. pp. 210–229, 2006.

[61] M. Ben, M. Betser, F. Bimbot, and G. Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” in *Proceedings of the International Conference on Spoken Language Processing, ICSLP*, 2004, vol. 2004, p. Issue: 2.

[62] J. Bonastre and N. Scheffer, “ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition,” *ISCA-IEEE Speaker Odyssey 2008*, pp. 2–9, 2008.

[63] M. N. Vieira, F. R. McInnes, and M. a Jack, “On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures.,” *J. Acoust. Soc. Am.*, vol. 111, no. 2, pp. 1045–1055, 2002.

[64] P. a. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[65] B. Yegnanarayana, C. D'Alessandro, and V. Darsinos, “An iterative algorithm for decomposition of speech signals into periodic and aperiodic components,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 1–11, 1998.

[66] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *IFA Proc.*, vol. 17, pp. 97–110, 1993.

[67] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust.*, vol. 28, no. 4, 1980.

[68] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals.,” *J. Acoust. Soc. Am.*, vol. 57, 1975.

[69] F. Soong and B. Juang, “Line spectrum pair (LSP) and speech data compression,” *ICASSP ’84. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 9, 1984.

[70] J. Crosmer and T. . I. Barnwell, “A low bit rate segment vocoder based on line spectrum pairs,” *ICASSP ’85. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 10, 1985.

[71] C. M. Ribeiro and I. M. Trancoso, “Speaker adaptation in a phonetic vocoding environment,” *1999 IEEE Work. Speech Coding Proceedings. Model. Coders,* 1999.

[72] T. F. Quatieri, R. B. Dunn, D. A. Reynolds, J. P. Campbell, and E. Singer, “Speaker recognition using G.729 speech codec parameters,” *2000 IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 2, 2000.

[73] H. Cordeiro and C. M. Ribeiro, “Speaker characterization with MLSFs,” in *IEEE Odyssey 2006: Workshop on Speaker and Language Recognition*, 2006.

[74] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” *Proc. Fifth Annu. ACM Work. Comput. Learn. Theory*, pp. 144–152, 1992.

[75] A. Ng, “CS229 Lecture notes -- Support Vector Machines,” *Intell. Syst. their Appl. IEEE*, vol. pt.1, no. x, pp. 1–25, 2000.

[76] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” *Elements*, vol. 1, pp. 337–387, 2009.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[78] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, 1995.

[79] T. Pang-Ning, M. Steinbach, and V. Kumar, “Introduction to data mining,” *Libr. Congr.*, p. 796, 2006.

[80] R. Wayland and A. Jongman, “Acoustic correlates of breathy and clear vowels: The case of Khmer,” *J. Phon.*, vol. 31, no. 2, pp. 181–201, 2003.

[81] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *International Joint Conference on Artificial Intelligence*, 1995, vol. 14, no. 12, pp. 1137–1143.

[82] C. Hsu and C. Lin, “A comparison of methods for multiclass support vector machines,” *Neural Networks, IEEE Trans.*, vol. 13, no. 2, pp. 415–425, 2002.

[83] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, “An Improved Endpoint Detector for Isolated,” *IEEE Trans. Acoust.*, vol. ASSP-29, no. 4, pp. 777–785, 1981.