Métodos de Extração de Características e Classificação Automática de Desordens Vocais

Rafael Alberto dos Santos, * **Paulo Rogério Scalassara,** *
Wagner Endo *

* *Departamento Acadêmico da Elétrica, Universidade Tecnológica Federal do Paraná, PR, (e-mail: rafaelsantos.2015@alunos.utfpr.edu.br, [prscalassara, wendo]@utfpr.edu.br).*

Abstract: Vocal disorders may be present when a person's voice fails to fulfill its basic role of communication. These disorders can be detected by the variation of perceptual parameters of the voice, such as quality, pitch, and loudness. Changes in voice parameters can be measured and classified automatically through acoustic analysis. In this study, we compare feature extraction techniques for the automatic classification of voice disorders by means of support vector machines. These techniques are based on wavelet variance, mel spectrogram, and mel frequency cepstral coefficients. The voice signals are sustained vowel "a" utterances, with neutral pitch, belonging to healthy and pathological classes, specifically nodule on the vocal folds and Reinke's edema. These pathologies affect the vocal folds and alter acoustical parameters of voice signals. Using wavelet variance and mel spectrogram patterns the average resulting classification accuracy values were 84.4%, what was higher than the 82.2% obtained using the mel frequency cepstral coefficients.

Resumo: Distúrbios vocais podem existir quando não se consegue usar a voz para cumprir seu papel básico de comunicação. Esses distúrbios podem ser detectados pela variação de parâmetros perceptuais da voz, tais como qualidade, tom e volume. As alterações dos parâmetros da voz podem ser medidas e classificadas de forma automática através da análise acústica. Neste estudo, comparam-se técnicas de extração de características para a classificação automática de desordens vocais usando *support vector machines*. Essas técnicas são baseadas em variância *wavelet*, mel espectrograma e coeficientes cepstrais de frequência mel. Consideram-se sinais da vogal "a" sustentada, com tom neutro, pertencentes às classes saudável e patológicas, especificamente nódulo nas pregas vocais e edema de Reinke. Essas patologias afetam as pregas vocais e alteram parâmetros acústicos dos sinais de voz. Utilizando-se padrões de variância *wavelet* e mel espectrograma, foram obtidas acurácias médias de classificação de $84,4\%$, superior ao valor de $82,2\%$ obtido com o uso dos coeficientes cepstrais de frequência mel.

Keywords: wavelet; mel spectrogram; cepstral coefficients; support vector machine; voice; pathology.

Palavras-chaves: wavelet; mel espectrograma; coeficientes cepstrais; máquina de vetores de suporte; voz, patologia.

1. INTRODUÇÃO

A utilização de soluções à distância para avaliação e tratamento de saúde aumentou muito nos últimos anos, especialmente com a pandemia de COVID-19 (Verde et al., 2021). Em aplicações de distúrbios da laringe não foi diferente, principalmente com a facilidade de gravação da voz e interação com médicos por meio de smartphones, a chamada *m-Health* ou *mobile health* (saúde móvel, em tradução livre do inglês) (Verde et al., 2018, 2021).

Esses distúrbios da laringe, também chamados de disfonias, afetam a comunicação oral e ocorrem quando a pessoa não consegue usar a voz para sua função de transmissão verbal e emocional (Behlau, 2005). As disfonias podem ser percebidas pela diferença de qualidade, tom e volume da voz em relação às características típicas de falantes de mesma idade, gênero, formação cultural e localização geográfica (Stemple et al., 2020).

Uma das formas de se avaliar essas características da voz é extrair informações dos sinais coletados de voz pela utilização de medidas acústicas, sendo essa técnica conhecida como análise acústica. Essas medidas podem fornecer análises objetivas e não invasivas da função vocal e estruturas do aparelho fonador, contanto que sejam registradas e interpretadas de maneira correta (Stemple et al., 2020; Rabiner and Schafer, 2010).

Uma vantagem da análise acústica é ser um método não invasivo e que auxilia o diagnóstico clínico, evitando, algumas vezes, a necessidade de exames invasivos como a laringoscopia. Pode-se usar essa análise para acompanhamento pós-operatório de tireoide, como apresentado em Ortega et al. (2009), ou fornecer evidências dos resultados

do tratamento cirúrgico em pacientes com pólipo, como em Petrovic-Lazic et al. (2015).

Diversas técnicas podem ser utilizadas para a extração de características dos sinais de voz. Uma abordagem popular é a obtenção dos coeficientes cepstrais de frequência mel (MFCC - *mel frequency cepstral coefficients*) (Kadiri and Alku, 2020), com os quais se explora o princípio da audição e a propriedade de descorrelação do *cepstrum* (Gómez-García et al., 2019). Pode-se também utilizar medidas de previsibilidade como a entropia dos sinais de voz para extração de características (Scalassara et al., 2009; Al-Nasheri et al., 2018). Outra ferramenta comum em estudos de patologias da laringe é a transformada *wavelet* discreta (DWT - *discrete wavelet transform*) (Alves et al., 2021). Um exemplo é Fonseca et al. (2007), aonde usa-se decomposição com DWT em conjunto com coeficientes de predição linear para a classificação de distúrbios vocais, obtendo acurácia superior a 90%.

O objetivo deste estudo é apresentar algoritmos de extração de características dos sinais de voz e usá-los para classificar patologias da laringe. Para isso, realizam-se testes com três padrões dos sinais: variância *wavelet*, mel espectrograma e MFCC. A variância *wavelet* é calculada nos coeficientes gerados pela decomposição dos sinais de voz usando uma modificação da DWT, a *maximal overlap* DWT (MODWT). Já os MFCCs obtidos do mel espectrograma, sendo aplicada a transformada cosseno discreta (DCT - *discrete cosine transform*) nos coeficientes gerados pelo mel espectrograma.

As características obtidas pelos algoritmos são parâmetros de entrada do classificador máquina de vetores de suporte, mais comumente conhecido como *support vector machine* (SVM). Como métrica de comparação, avalia-se a acurácia da classificação usando as características de cada abordagem individualmente. Escolheu-se o SVM por ser uma técnica popular em aplicações de detecção de patologias de laringe (Verde et al., 2018), além de ter processo de treinamento rápido com convergência garantida (Schölkopf and Smola, 2001).

O restante deste trabalho está dividido da seguinte forma: na Seção 2, apresentam-se as características dos sinais patológicos que são utilizados, além de uma breve revisão dos métodos de extração e classificação dessas características. Na Seção 3, descreve-se o banco de dados e a metologia utilizada; já na Seção 4, mostram-se os resultados obtidos e, por fim, as conclusões deste estudo na última seção.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, apresentam-se algumas informações sobre as patologias da laringe consideradas neste estudo. Também, descrevem-se brevemente as ferramentas utilizadas: MFCC, variância *wavelet* e classificador SVM.

2.1 Patologias da Laringe

Neste estudo, além dos sinais da classe saudável, tem-se duas classes patológicas: nódulo nas pregas vocais e edema de Reinke. Os nódulos são protuberâncias esbranquiçadas, geralmente bilaterais, na margem glótica de cada prega vocal, localizadas na junção terço médio anterior (Boone et al., 2014). Os efeitos resultantes na voz são variáveis e dependem das lesões. Segundo Stemple et al. (2020) os sintomas vocais includem disfonia leve a moderada, caracterizada por rugosidade, soprosidade e aumento da tensão da musculatura laríngea. A correspondência da rugosidade em parâmetros acústicos é a aperiodicidade de vibração das pregas vocais. A soprosidade corresponde a componentes de ruído devido a turbulência.

O edema de Reinke ocorre quando a camada superficial da lâmina própria (também chamada de espaço de Reinke) fica cheia de líquido viscoso devido a um trauma de longa data (Stemple et al., 2020). Geralmente, esse edema é bilateral, mas pode ser mais pronunciado de um lado (Boone et al., 2014). Os efeitos na qualidade da voz são disfonia leve a moderada, sendo caracterizada por tom mais baixo e rouquidão (Stemple et al., 2020). A rouquidão indica irregularidade de vibração das pregas vocais com presença de ruído.

Após esta descrição das patologias estudadas, apresentam-se, nas próximas seções, as ferramentas aplicadas no reconhecimento de padrões e classificação dos sinais de voz.

2.2 Mel Frequency Cepstral Coefficients

Uma das medidas acústicas utilizadas na detecção de distúrbios vocais é a frequência fundamental da voz, sendo essa medida relacionada a percepção sonora da frequência conhecida como *pitch*. O *pitch* é medido em mel (abreviação de *melody*), e pode ser relacionado com a frequência conforme a Equação (1), sendo F a frequência de um tom.

$$\text{pitch} = 2595 \log_{10}\left(1 + \frac{F}{700}\right), \tag{1}$$

A análise na escala mel pode ser realizada utilizando um banco de filtros, os quais são projetados com formas triangulares e igualmente espaçados na escala mel (Huang et al., 2001). Seja H_r o banco de filtros e X_m a transformada discreta de Fourier (DFT - *discrete Fourier transform*) da m-ésima janela do sinal x, pode-se escrever o mel espectrograma pela Equação (2), sendo R a quantidade de filtros e K a quantidade de amostras da DFT unilateral.

$$\text{MF}_m[r] = \sum_{k=0}^{K} |X_m[k]|^2 H_r[k], \quad r = 1, \ldots, R \tag{2}$$

Os MFCCs são obtidos pela Equação (3) com a aplicação da DCT, onde $n = 0, \ldots, N_{\text{mfcc}-1}$ e $\tilde{\beta}$ é $1/\sqrt{2}$ para $n = 0$ e 1 caso contrário.

$$\text{mfcc}_m[n] = \sqrt{\frac{2}{R}}\tilde{\beta} \sum_{r=0}^{R-1} \log(\text{MF}_m[k]) \cos\left[\frac{\pi n(2r+1)}{2R}\right] \tag{3}$$

A Figura 1 apresenta exemplos do mel espectrograma e MFCCs de um sinal da vogal "a" sustentada neutra de pessoa com laringe saudável, sendo λ a frequência na escala mel e *mfcc* os coeficientes cepstrais de frequência mel. O sinal temporal utilizado para criação dessas figuras será apresentado na Figura 2 (a).

2.3 Variância Wavelet

A variância *wavelet* é baseada na MODWT de uma série temporal, sendo os filtros dessa transformada versões escaladas dos filtros utilizados na DWT. O filtro *wavelet* é $\tilde{h}_l = h_l/\sqrt{2}$ e o filtro escala é $\tilde{g}_l = g_l/\sqrt{2}$ (Percival and Walden, 2000).

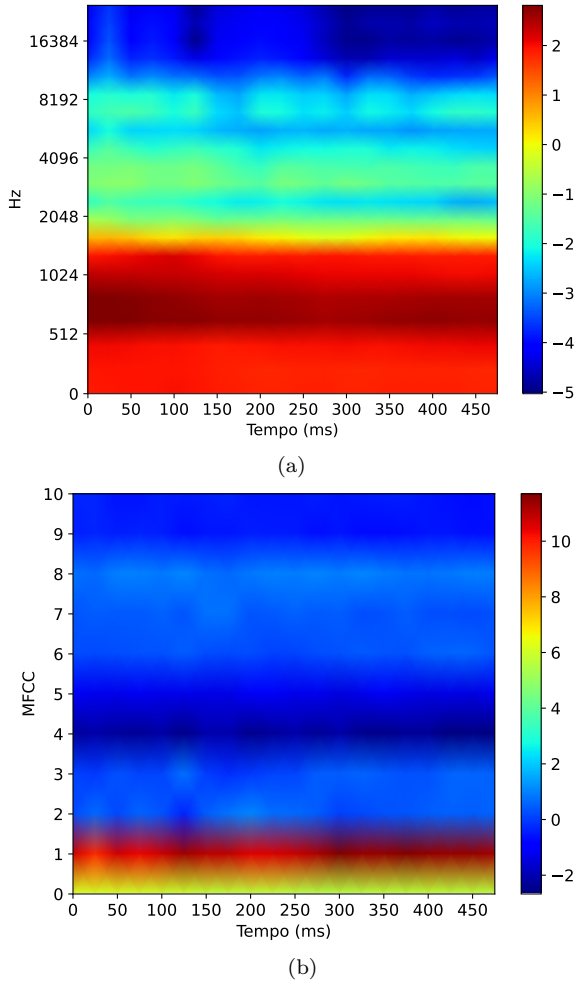


Figura 1. Extração de características em um sinal da vogal "a" neutra. (a) Mel espectrograma. (b) MFCC.

A decomposição pela MODWT é expressa pelas Eqs. (4) e (5), onde j é o nível da decomposição e L_j é o comprimento dos filtros da Equação (6).

$$\widetilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N}, \quad t = 0, \ldots, N-1, \tag{4}$$

$$\widetilde{V}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{g}_{j,l} X_{t-l \bmod N}, \quad t = 0, \ldots, N-1, \tag{5}$$

$$L_j = (2^j - 1)(L-1) + 1 \tag{6}$$

A decomposição de uma série temporal utilizando a MODWT permite uma análise estatística do sinal como função da escala (Cornish et al., 2006). Pode-se estimar a variância *wavelet* não enviesada, conforme a Equação (7), onde $M_j = N - L_j + 1$ é a quantidade de amostras do sinal (N) retirando os coeficientes de fronteira no nível j.

$$\hat{v}_X = (\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \widetilde{W}_{j,t}^2 \tag{7}$$

2.4 Support Vector Machine

SVM é um algoritmo de aprendizagem supervisionada que utiliza hiperplanos de decisão com o objetivo de otimizar a separação de classes. Para a classificação binária de classes linearmente separáveis, os dados de treinamento $\mathbf{x}_i \in \mathcal{R}^d$ são rotulados $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, l$ sendo l a quantidade de dados de treinamento com rótulos $y_i \in \{-1,\ 1\}$. O hiperplano de decisão é definido em (8), onde $\mathbf{w}$ é o vetor de pesos e b o parâmetro de tendência.

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b, \tag{8}$$

Seja d_+ e d_- as menores distâncias entre o hiperplano de decisão $\mathbf{w}^T\mathbf{x}_i + b = 0$ e os pontos com rótulos positivos e negativos. Define-se a margem como a soma $(d_+) + (d_-)$. Se os dados de treino são linearmente separáveis, por definição, tem-se pelo menos um $\mathbf{w}$ e b que satisfaz o conjunto de inequações (9).

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0 \quad \forall i. \tag{9}$$

Pode-se resolver esse problema minimizando $\|\mathbf{w}\|^2$, sendo equivalente a maximizar a margem, sujeito às restrições de (9). Porém, quando os dados não forem linearmente separáveis, o algoritmo não encontrará uma solução viável para o problema (Burges, 1998).

Para dados não separáveis, pode-se relaxar as restrições em (9) pela introdução de variáveis de folga $\xi_i \geq 0$, conforme a Equação (10), onde $\phi(\cdot)$ é a transformação fixa do vetor de características $\mathbf{x}_i$, sendo utilizada quando o limite de decisão é não linear.

$$y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, l \tag{10}$$

Pode-se maximizar a margem ao minimizar a Equação (11), sendo penalizados os dados que forem classificados incorretamente (Bishop, 2006), onde $C > 0$ controla a troca entre a variável de folga e a margem.

$$C\sum_{i=1}^{l} \xi_i + \frac{1}{2}\|\mathbf{w}\|^2, \tag{11}$$

Para encontrar a solução de (11), utilizam-se multiplicadores de Lagrange $a_i \geq 0$, Equação (12), onde $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ é definido como uma função núcleo. A Equação (12) está sujeita às restrições em (13).

$$\mathcal{L}(\mathbf{a}) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{12}$$

$$\sum_{i=1}^{l} a_l y_i = 0; \quad 0 \leq a_i \leq C, \quad i = 1, \ldots, l. \tag{13}$$

A classificação é realizada ao modificar (8), conforme (14).

$$f(\mathbf{x}) = \sum_{i=1}^{l} a_i y_i k(\mathbf{x}, \mathbf{x}_i) + b \tag{14}$$

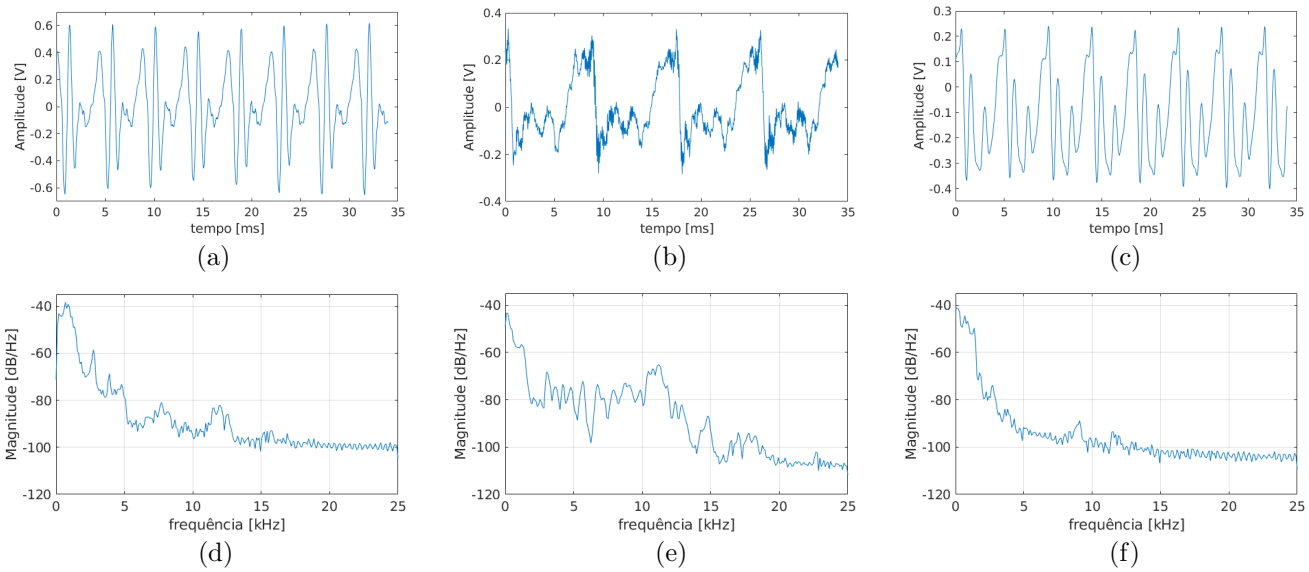


Figura 2. Exemplos de sinais da vogal "a" neutra sustentada para as classes: (a) saudável, (b) edema de Reinke e (c) nódulo nas pregas vocais. Espectros de potência dos sinais anteriores, respectivamente (d), (e) e (f).

Na próxima seção, apresentam-se o banco de dados de sinais de voz e a metodologia proposta neste estudo.

3. MATERIAIS E MÉTODOS

Inicia-se esta seção descrevendo o banco de sinais de voz, seguido pela explicação da metodologia proposta para extração e classificação de características dos sinais.

3.1 Banco de Dados

Os sinais de voz utilizados neste estudo foram obtidos do *Saarbruecken Voice Database* (SVD) (Barry and Pützer, 2021), sendo gravações da vogal "a" neutra sustentada, com frequência de amostragem de 50 kHz no formato *wave*. Foram obtidos 45 sinais, 38 de mulheres e 7 de homens, sendo 15 para cada classe. A reduzida quantidade de amostras dificulta a generalização do método SVM, porém, com a utilização de validação cruzada, pode-se obter um classificador adequado (Domingos, 2012).

A primeira classe é formada por vozes de pessoas com laringes saudáveis e idades entre 23 e 62 anos. Já a segunda classe é formada por vozes de pessoas com edema de Reinke e idades entre 42 e 64 anos. Por fim, a terceira classe é formada por vozes de pessoas com nódulo nas pregas vocais e idades entre 18 e 64 anos. A Tabela 1 sumariza os grupos e o intervalo de idade de cada um.

Tabela 1. Características das classes de sinais de voz.

Classe	Sinais	Idades
Saudável	15	45-65
Edema	15	43-65
Nódulo	15	19-64

A Figura 2 mostra exemplos de sinais das três classes sendo (a) saudável, (b) edema de Reinke e (c) nódulo nas pregas vocais. Observa-se a quasi-periodicidade dos sinais, característica dos sons vocálicos sustentados (Rabiner and Schafer, 2010), porém os casos patológicos possuem diferenças visíveis de quantidade de ruído para edema e alterações de amplitude para nódulo. Outras características não são claramente observáveis (Stemple et al., 2020). As figuras (d), (e) e (f) apresentam as densidades espectrais de potência respectivamente dos sinais em (a), (b) e (c).

A seguir, descreve-se a metodologia para extração e classificação das características obtidas dos sinais de voz.

3.2 Metodologia

Os passos para a implementação do método proposto neste estudo para classificação de desordens vocais é apresentado de forma simplificada na Figura 3. Todas as etapas serão descritas com ênfase na extração de características dos sinais e sua classificação.

Os sinais obtidos do SVD foram padronizados para conterem 26.518 amostras, equivalente a $530,36$ ms conforme a frequência de amostragem. Essa quantidade foi escolhida por ser o maior valor comum a todos os sinais analisados. Em seguida, realizou-se a normalização da amplitude dos sinais pelo maior valor absoluto, ou seja, as amplitudes foram ajustadas para o intervalo $[-1,\ 1]$.

Extração de Características. Nesta etapa, são utilizadas três abordagens: MODWT, mel espectrograma e MFCC. Os atributos obtidos com cada uma delas são separados com SVM de forma a se avaliar qual obtém os melhores resultados de classificação. Assim, faz-se necessário treinar o classificador para cada método de extração de característica, como indicado na Figura 3.

Para a primeira técnica, MODWT, utilizam-se os coeficientes de aproximação e detalhes, conforme a Equação (7). A cada nível de decomposição, tem-se o vetor de características de comprimento igual ao número desse nível mais um. O máximo nível para o cálculo da variância depende da quantidade de amostras do sinal (N) e do comprimento

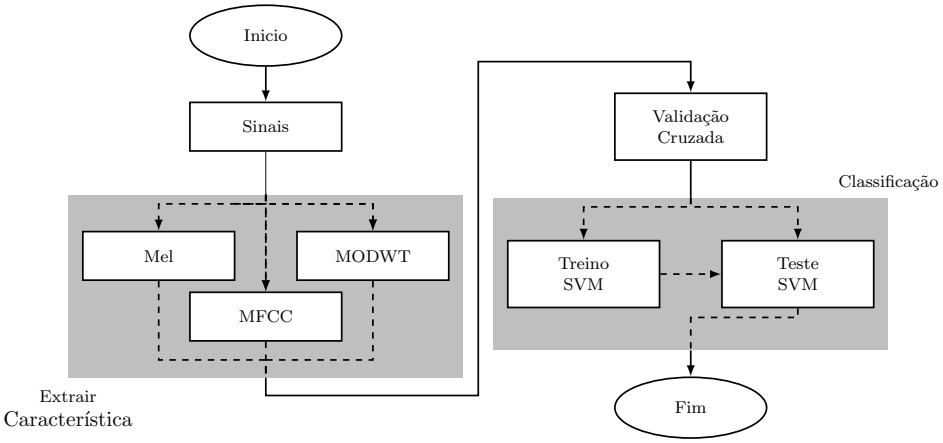


Figura 3. Fluxograma simplificado do método proposto neste estudo para classificação de desordens vocais.

do filtro (L), conforme a Equação (15), sendo $\lfloor \cdot \rfloor$ a função de arredondamento para baixo.

$$J_0 = \lfloor \log_2 (N/(L-1)+1) \rfloor, \tag{15}$$

A acurácia do método depende de alguns parâmetros, tais como a escolha do filtro e o nível da decomposição. Guido et al. (2005) utilizaram diferentes famílias de *wavelets* para a classificação de desordens vocais. Dentre os filtros mais comuns, concluíram que o de Daubechies apresentou as melhores acurácias, em torno de 85%. Então, foram realizados testes com essa família de filtros, sendo observado que o filtro de Daubechies com 2 coeficientes apresenta bons resultados ao decompor os sinais até o nível 11. Portanto, esse filtro será utilizado neste estudo para a análise dos sinais de voz. Assim, usando esse nível de decomposição, tem-se um coeficiente de aproximação e 11 componentes de detalhe. A variância desses 12 componentes é usada como vetor de características para o classificador SVM.

Para realizar a extração de característica através das técnicas MFCC e mel espectrograma, utiliza-se a transformada de Fourier de tempo curto (STFT - *short-time Fourier transform*) com janela de Hamming de comprimento de 50 ms com 50% de sobreposição entre as janelas. Devido à padronização da quantidade de amostras dos sinais, ao se realizar a STFT, tem-se um total de 20 janelas. Após testes prévios com os sinais do banco de dados, escolheu-se a quantidade de filtros de 21, sendo esses filtros espalhados no espectro de frequências no intervalo de 20 a 20.000 Hz, o qual engloba todas as frequências audíveis pelo ouvido humano. A Figura 4 apresenta a distribuição desses filtros na frequência (normalizada em π radianos), sendo a amplitude máxima unitária.

Para transformar o espectrograma em mel espectrograma, realiza-se a multiplicação matricial entre o resultado da STFT e o banco de filtros na escala mel e aplica-se a função logaritmo na base 10 na matriz resultante. A extração de características com o método MFCC é a continuação do mel espectrograma, sendo aplicada a DCT no mel espectrograma e mantidos apenas os primeiros 11 coeficientes. Conforme Deller Jr. et al. (2000), o primeiro coeficiente do MFCC é geralmente uma medida ponderada do logaritmo da energia, assim, pode-se substituir esse

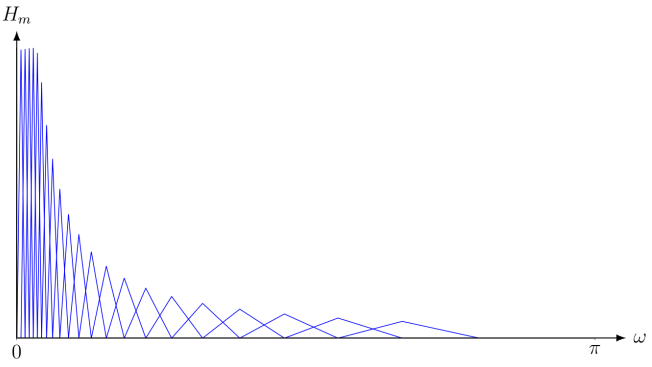


Figura 4. Banco de filtros utilizados para o cálculo do mel espectrograma e MFCC.

coeficiente pelo logaritmo da energia, como indicado por Rabiner and Schafer (2010). A Figura 5 mostra as etapas para a obtenção do mel espectrograma (Mel no diagrama) e do MFCC.

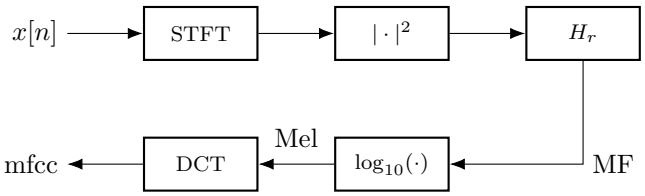


Figura 5. Diagrama em blocos para o cálculo do mel espectrograma e MFCC.

Nota-se que o objetivo do método é a classificação de sinais de voz, então transformam-se as matrizes de características em vetores. A matriz de características do mel espectrograma é composta por 21 linhas e 20 colunas e a do MFCC por 11 linhas e 20 colunas. Portanto, ao vetorizar as matrizes, o vetor de características para os métodos mel espectrograma e MFCC, ficam com comprimentos de 420 e 220, respectivamente.

Validação Cruzada e Classificação. Esta primeira é geralmente utilizada quando os dados de treino e teste são limitados, ou seja, quando se tem um pequeno conjunto de dados disponíveis. Nesse método, o conjunto de dados disponível de comprimento N é dividido em K_v subcon-

juntos, sendo $K_v > 1$. Então, o treino ocorre em todos os subconjuntos, exceto por um, e o erro de classificação é medido pelo subconjunto deixado de fora do treino. Este processo se repete para um total de K_v testes, cada vez utilizando um subconjunto diferente para o cálculo do erro. Assim, a cada iteração k, gera-se um erro de classificação e_k, sendo realizada a média dos erros para avaliação do classificador, Equação (16).

$$\overline{e} = \frac{1}{K_v} \sum_{k=1}^{K_v} e_k \tag{16}$$

Quando o conjunto de dados disponível é severamente limitado, utiliza-se o método conhecido como *leave-one-out* (Silva et al., 2016). Nesse caso, $N-1$ dados são utilizados para o treino e o teste é realizado com a amostra restante. Neste estudo, utiliza-se a técnica *leave-one-out* devido à limitação do banco de dados disponível.

Após a separação do conjunto de dados em treino e teste, o conjunto de treino é normalizado, sendo aplicada a mesma normalização no conjunto de teste. Para isso, utiliza-se o método *z-score*, no qual se subtrai a média e a divisão pelo desvio padrão de cada atributo (Silva et al., 2016).

Os vetores normalizados são as entradas do classificador SVM, sendo escolhido o *kernel* de funções de base radial (RBF). Para se obter melhor acurácia na classificação dos dados, os hiperparâmetros do método SVM são otimizados para que, na média dos experimentos, o erro seja minimizado (Mantovani et al., 2015).

Além do erro (ou acurácia do método), pode-se calcular outras métricas, tais como a sensibilidade e a especificidade. A Equação (17) define as expressões utilizadas para calcular essas três métricas.

$$\begin{aligned} \text{Acc. } (\%) &= 100 \times \frac{vp + vn}{vp + vn + fp + fn}, \\ \text{Sens. } (\%) &= 100 \times \frac{vp}{vp + fn}, \\ \text{Espec. } (\%) &= 100 \times \frac{vn}{vn + fp}, \end{aligned} \tag{17}$$

sendo vp verdadeiro positivo, fn falso negativo, vn verdadeiro negativo e fp falso positivo.

A seguir, apresentam-se os resultados das classificações usando técnica SVM com entradas dadas pelas características discutidas nesta seção.

4. RESULTADOS

Nesta seção, serão discutidos os resultados do classificador SVM quando se usam as características obtidas com MODWT, mel espectrograma e MFCC.

As Tabelas 2 a 4 apresentam as acurácias médias, sensibilidades e especificidades de classificação entre as classes de sinais de voz para os três métodos de extração de características. Esses valores são médias dos resultados obtidos na validação cruzada *leave-one-out* empregada no classificador. Para separação entre sinais saudáveis e patológicos, nota-se que a classificação usando características extraídas com MODWT e mel espectrograma obtiveram acurácia média ligeiramente acima da obtida usando os atributos de MFCC. Na sensibilidade, que mede o quão sensível é o método em detectar distúrbios, a MODWT foi superior ao mel espectrograma, sendo inferior na especificidade.

Tabela 2. Métricas do SVM para características obtidas com MODWT.

Grupos	**MODWT**		
	Acc.	Sens.	Espec.
Patológico/Saudável	84,4	90,0	73,3
Edema/Saudável	86,7	93,3	80,0
Nódulo/Saudável	73,3	73,3	73,3
Edema/Nódulo	76,7	80,0	73,3

Tabela 3. Métricas do SVM para características obtidas com Mel espectrograma.

Grupos	**Mel espectrograma**		
	Acc.	Sens.	Espec.
Patológico/Saudável	84,4	90,0	73,3
Edema/Saudável	73,3	66,7	80,0
Nódulo/Saudável	56,7	46,7	66,7
Edema/Nódulo	70,0	66,7	73,3

Tabela 4. Métricas do SVM para características obtidas com MFCC.

Grupos	**MFCC**		
	Acc.	Sens.	Espec.
Patológico/Saudável	82,2	90,0	66,7
Edema/Saudável	80,0	86,7	73,3
Nódulo/Saudável	60,0	40,0	80,0
Edema/Nódulo	83,3	86,7	80,0

Com relação à separação entre sinais saudáveis e com edema de Reinke, novamente, as melhores métricas foram obtidas usando MODWT. O mesmo ocorreu para o grupo saudável versus nódulo, porém, a melhor especificidade foi resultado dos atributos de MFCC. Para sensibilidade, mel espectrograma e MFCC resultaram em valores bem inferiores ao da MODWT. Por fim, para discriminação entre os sinais patológicos, as melhores métricas foram do MFCC, entretanto próximas dos outros dois métodos. A maior dificuldade na classificação foi separar sinais saudáveis de sinais de nódulo com acurácia máxima de 73,3% com MODWT. Para os outros grupos, foram obtidas acurácias acima de 80% com MODWT ou MFCC.

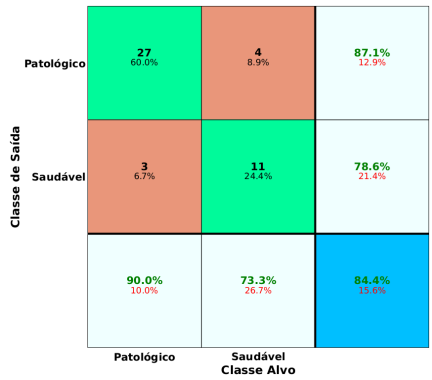
Outra forma de avaliar o desempenho é usando matrizes de confusão, que mostram quais as classes das amostras classificadas correta e incorretamente. As linhas da matriz correspondem às classes preditas pelo classificador e as colunas mostram as classes alvo, ou seja, as verdadeiras. A diagonal principal da matriz mostra a classificação correta e fora da diagonal principal estão as classificações incorretas (Theodoridis and Koutroumbas, 2003). Devido a utilização do método de validação cruzada *leave-one-out*, as matrizes de confusão para os três tipos de atributos foram obtidas pela soma das matrizes de cada teste.

Pode-se calcular outras métricas como a precisão e o valor preditivo negativo (VPN), sendo essas métricas definidas

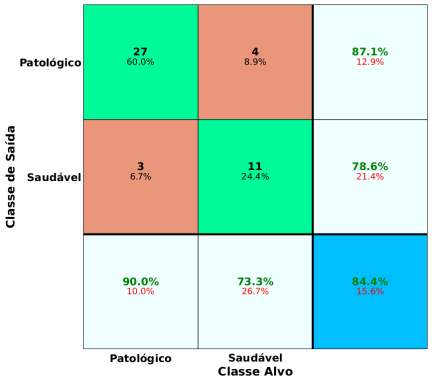
na Equação (18).

$$\text{Prec. } (\%) = 100 \times \frac{vp}{vp + fp},$$
$$\text{VPN. } (\%) = 100 \times \frac{vn}{vn + fn}. \tag{18}$$

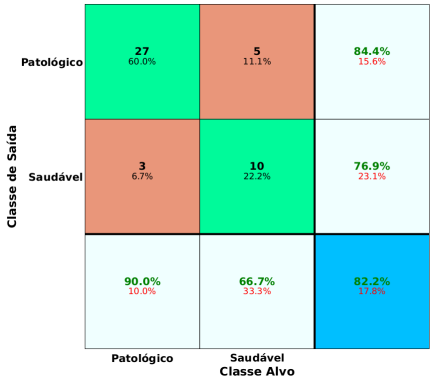
Para o grupo patológico versus saudável, calcularam-se essas novas métricas e, junto com os valores de acurácia, sensibilidade e especificada das Tabelas 2 a 4, são apresentadas nas matrizes de confusão da Figura 6. A precisão e o valor preditivo negativo são adicionados na última coluna, a sensibilidade e especificidade na última linha e a acurácia adicionada na última posição da diagonal principal.



(a)



(b)



(c)

Figura 6. Matriz de confusão do SVM com (a) MODWT, (b) mel espectrograma e (c) MFCC para o grupo patológico/saudável.

A Figura 6(a) apresenta a matriz de confusão dos resultados com uso de características obtidas com a MODWT. Além da acurácia, sensibilidade e especificidade de $84,4\%$, $90,0\%$ e $73,3\%$ respectivamente, tem-se precisão e VPN de $87,1\%$ e $78,6\%$ respectivamente. As figuras em (b) e (c) apresentam as matrizes dos resultados de mel espectrograma e MFCC. Com base nessas matrizes, observa-se que as métricas para as três abordagens na separação de sinais saudáveis e patológicos foram similares, com pequena desvantagem para os atributos obtidos com MFCC.

Para a classificação usando mel espectrograma, tem-se quantidade de características extraídas dos sinais de voz superior do que usando MODWT, sendo 420 e 12, respectivamente. Assim, é preferível usar os padrões MODWT, pois consegue-se a mesma acurácia que o mel espectrograma com quantidade menor de características. O MFCC utiliza uma quantidade menor de características do que o mel espectrograma, entretanto essa técnica apresentou as piores métricas nos testes, a não ser para o grupo edema/nódulo.

Para comparação, Ali et al. (2017) usou MFCC para a extração de características de sinais de vozes em três bancos de dados, incluindo o SVD. A melhor acurácia obtida nesse banco de dados foi 80.2%, entretanto, são classificadas diferentes desordens vocais. Já em Fonseca et al. (2007), tem-se acurácia superior a 90% ao utilizar a DWT em conjunto com coeficientes de predição linear. Porém, o banco de dados utilizado não é público. Assim, a comparação direta com outros métodos se torna difícil, visto que, alguns estudos usam bancos de dados privados ou analisam diferentes patologias.

5. CONCLUSÃO

Este estudo comparou diferentes métodos de extração de características para a classificação de desordens vocais. Utilizaram-se a variância *wavelet*, obtida pela decomposição dos sinais de voz pela MODWT, o mel espectrograma, que realiza uma análise dos sinais na escala de *pitch*, e o MFCC, que também utiliza a análise na escala de *pitch* e a propriedade de descorrelação do *cepstrum*.

Usando esses padrões em um classificador SVM, foram obtidas acurácias acima de 80% na discriminação de sinais saudáveis de sinais patológicos (edema de Reinke e nódulo nas pregas vocais), com destaque para variância *wavelet* e mel espectrograma com $84,4\%$ de acerto. Apesar de se obter a mesma acurácia do que usando mel espectrograma, com a variância *wavelet*, tem-se maior sensibilidade, ou seja, menor erro dentro da classe patológica. Entretanto, para discriminação entre os sinais patológicos, ou seja, diferenciar os sinais de edema dos de nódulo, a melhor abordagem foi com atributos de MFCC, com acurácia de $83,3\%$ contra $76,7\%$ e $70,0\%$ dos outros métodos.

Em uma abordagem sequencial, pode-se utilizar um classificador SVM para separar casos patológicos de saudáveis e, depois, discriminar as duas patologias com outro classificador. Nesse caso, com base nos resultados anteriores, seria interessante usar MODWT para o primeiro classificador e MFCC para o segundo.

Mostrou-se que a variância *wavelet* se compara às técnicas usando mel espectrograma e MFCC, porém com

a vantagem de usar um vetor de características menor. Portanto, o presente trabalho contribui para a análise automática de desordens vocais ao comparar técnicas de extração de características em um banco de dados público e gratuito. Como o SVD possui poucos sinais, pretende-se, em trabalhos futuros, buscar novos bancos de dados, além de se adicionar sinais de novas patologias. Com maior quantidade de amostras, seria possível trocar a validação cruzada *leave-one-out* por *k-folds*, permitindo uma avaliação mais significativa dos resultados dos classificadores, em especial, empregando métodos estatísticos para verificar a equivalência das propostas de extratores de características.

AGRADECIMENTOS

Os autores agradecem o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Universidade Tecnológica Federal do Paraná (UTFPR). Por fim, também agradecem os pesquisadores do Instituto de Fonética da Universidade de Saarland por disponibilizarem o *Saarbruecken Voice Database*.

REFERÊNCIAS

Al-Nasheri, A. et al. (2018). Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, 6, 6961–6974.

Ali, Z. et al. (2017). Intra-and inter-database study for arabic, english, and german databases: do conventional speech features detect voice pathology? *Journal of Voice*, 31(3), 386.e1–386.e8.

Alves, M. et al. (2021). Voice disorders detection through multiband cepstral features of sustained vowel. *Journal of Voice*, (no prelo), 1–10.

Barry, W. and Pützer, M. (2021). Saarbrucken voice database. Institute of Phonetics, Saarland University. URL `http://www.stimmdatenbank.coli.uni-saarland.de`.

Behlau, M. (2005). *Voz: o livro do especialista*. Revinter, Rio de Janeiro.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin, Heidelberg.

Boone, D.R., McFarlane, S.C., Von Berg, S.L., and Zraick, R.I. (2014). *The voice and voice therapy*. Pearson, Boston.

Burges, C.J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167.

Cornish, C.R., Bretherton, C.S., and Percival, D.B. (2006). Maximal overlap wavelet statistical analysis with application to atmospheric turbulence. *Boundary-Layer Meteorology*, 119(2), 339–374.

Deller Jr., J.R., Hansen, J.H.L., and Proakis, J.G. (2000). *Discrete-Time Processing of Speech Signals*. Wiley, New York, NY.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.

Fonseca, E.S. et al. (2007). Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders. *Computers in Biology and Medicine*, 37(4), 571–578.

Gómez-García, J.A., Moro-Velázquez, L., and Godino-Llorente, J.I. (2019). On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51, 181–199.

Guido, R.C. et al. (2005). Trying different wavelets on the search for voice disorders sorting. In *Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory (SSST'05)*, 495–499. IEEE, Tuskegee, USA.

Huang, X., Acero, A., Hon, H.W., and Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, Upper Saddle River, NJ.

Kadiri, S.R. and Alku, P. (2020). Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing*, 14(2), 367–379.

Mantovani, R.G. et al. (2015). To tune or not to tune: Recommending when to adjust svm hyper-parameters via meta-learning. In *Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Killarney, Ireland.

Ortega, J., Cassinello, N., Dorcaratto, D., and Leopaldi, E. (2009). Computerized acoustic voice analysis and subjective scaled evaluation of the voice can avoid the need for laryngoscopy after thyroid surgery. *Surgery*, 145(3), 265–271.

Percival, D.B. and Walden, A.T. (2000). *Wavelet methods for time series analysis*, volume 4. Cambridge university press, Cambridge.

Petrovic-Lazic, M. et al. (2015). Acoustic and perceptual characteristics of the voice in patients with vocal polyps after surgery and voice therapy. *Journal of Voice*, 29(2), 241–246.

Rabiner, L. and Schafer, R. (2010). *Theory and Applications of Digital Speech Processing*. Pearson, New York, NY.

Scalassara, P.R., Maciel, C.D., and Pereira, J.C. (2009). Predictability analysis of voice signals. *IEEE Engineering in Medicine and Biology Magazine*, 28(5), 30–34.

Schölkopf, B. and Smola, A.J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Silva, I.N., Spatti, D.H., and Flauzino, R.A. (2016). *Redes Neurais Artificiais - Para Engenharia e Ciências Aplicadas*. Artliber, São Paulo, 2a. edition.

Stemple, J.C., Roy, N., and Klaben, B.K. (2020). *Clinical voice pathology: Theory and management*. Plural Publishing, San Diego, CA.

Theodoridis, S. and Koutroumbas, K. (2003). *Pattern Recognition*. Elsevier, San Diego, CA, 2nd edition.

Verde, L., Pietro, G.D., and Sannino, G. (2018). Voice disorder identification by using machine learning techniques. *IEEE Access*, 6, 16246–16255.

Verde, L. et al. (2021). Exploring the use of artificial intelligence techniques to detect the presence of coronavirus covid-19 through speech and voice analysis. *IEEE Access*, 9, 65750–65757.