X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

Detecção de Doenças da Laringe usando Transformada Wavelet e Redes Neurais Artificiais

Raphael Torres Santos Carvalho, Charles Casimiro Cavalcante, Paulo César Cortez
Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, Brasil
raphaeltscarvalho@hotmail.com, charles@gtel.ufc.br, cortez@lesc.ufc.br

Resumo – A quantidade de métodos não invasivos de diagnóstico tem aumentado devido à necessidade de exames simples, rápidos e indolores. Por conta do crescimento da tecnologia que fornece os meios necessários para a extração e processamento de sinais, novos métodos de análise têm sido desenvolvidos para compreender a complexidade dos sinais de voz. O presente artigo apresenta uma nova ideia para caracterizar os sinais de voz saudável e patológicos baseado em duas ferramentas matemáticas amplamente conhecidas na literatura, Transformada Wavelet (WT) e Redes Neurais Artificiais. Quatro classes de amostras foram utilizadas, uma de indivíduos saudáveis e as outras três de pessoas com nódulo na prega vocal, edema de Reinke e disfonia neurológica. Todas as amostras foram gravadas usando a vogal sustentada /a/ do Português Brasileiro. O trabalho mostra que a abordagem proposta usando WT é uma técnica adequada para discriminação entre vozes saudável e patológica.

Palavras-chave – Processamento de Voz, Transformada Wavelet, Redes Neurais Artificiais, Nódulo Vocal, Edema de Reinke, Disfonia Neurológica.

Abstract – The amount of non-invasive methods of diagnosis have increased due to the need for simple, quick and painless tests. Due to the growth of technology that provides the means for extraction and signal processing, new analytical methods have been developed to understand the complexity of the voice signals. This paper presents a new idea to characterize signals of healthy and pathological voice based on two mathematical tools widely known in the literature, Wavelet Transform (WT) and Artificial Neural Networks. Four classes of samples were used, one from healthy individuals and three from people with vocal fold nodules, Reinke's edema and neurological dysphonia. All the samples were recorded using the vowel /a/ in Brazilian Portuguese. The work shows that the proposed approach using WT is a suitable technique to discriminate between healthy and pathological voices.

Keywords – Voice Processing, Wavelet Transform, Artificial Neural Networks, Vocal Fold Nodules, Reinke's Edema, Neurological Dysphonia.

1. INTRODUÇÃO

A voz é um dos principais meios de comunicação humano e, como um sinal acústico, contém informações importantes sobre algumas características individuais. A estrutura biomecânica vocal, em associação com variáveis aerodinâmicas, desempenha um papel importante na produção da voz e está ligada às mudanças na qualidade vocal [1]. A qualidade da voz depende do modo de fechamento e abertura da glote e da vibração das pregas vocais. Certas alterações laríngeas impedem que as pregas vocais tenham uma vibração glotal harmônica. Os principais fatores que determinam a vibração vocal são [2]:

1. posição da prega vocal, ou a extensão em que as pregas vocais são aduzidas ou abduzidas;
2. mioelasticidade, ou o grau de elasticidade das pregas vocais (determinado pela posição e grau de tensão decorrente da contração do músculo vocal);
3. nível de pressão do ar através das pregas vocais.

As alterações das pregas vocais, devido à presença de patologias na laringe, causam mudanças significativas em seus padrões vibratórios, afetando a qualidade da voz. Certas alterações laríngeas podem causar dificuldade no diagnóstico, pois são muito semelhantes, apesar de apresentarem origens e alterações fisiopatológicas diferentes. Essas alterações podem ser classificadas como físicas, neuromusculares, traumáticas e psicogênicas, e todas afetam diretamente a qualidade da voz.

Algumas alterações laríngeas representam a grande maioria dos atendimentos de pacientes com alterações da voz e que não sejam devido a infecções das vias áreas superiores, onde se tem uma inflamação das pregas vocais e que geralmente cedem em alguns dias com a melhora do quadro geral. Estas alterações são: nódulo vocal, cisto vocal, pólipo vocal, edema de Reinke e sulco vocal. Na maioria dos casos, tais alterações laríngeas produzem uma rouquidão com características típicas de cada uma, principalmente quando analisadas por médicos otorrinolaringologistas ou profissionais da área da fonoaudiologia [2].

Diante do exposto, para conseguir avaliar a qualidade vocal do paciente, os médicos utilizam diversas técnicas. Algumas são técnicas subjetivas, baseadas na audição da voz pelo médico especialista, sendo essas totalmente dependentes da experiência do profissional. Outras técnicas permitem a inspeção direta da laringe e a visualização do comportamento vibratório das pregas vocais através do uso de técnicas laringoscópicas como a videolaringoscopia direta e a videoestroboscopia.

A videolaringoscopia direta é um exame realizado pelo médico com o objetivo de visualizar a laringe utilizando uma microcâmera. A videoestroboscopia permite a visualização do comportamento vibratório das pregas vocais [3]. Essas técnicas, embora sejam mais objetivas, são consideradas invasivas e causam desconforto aos pacientes, além de utilizarem instrumentos sofisticados e caros como fontes de luz especial, instrumentos endoscópicos e câmeras de vídeo especializadas [4].

A simplicidade e a natureza não invasiva têm feito das técnicas de processamento digital de sinais, por meio da análise acústica, uma eficiente ferramenta para o diagnóstico das desordens provocadas por patologias da laringe, classificação de doenças da voz e acompanhamento da evolução de tratamentos. Dessa forma, podem ser aplicadas como técnicas auxiliares aos métodos baseados na inspeção direta das cordas vocais, diminuindo a regularidade dos exames mais invasivos [4].

Em relação aos métodos tradicionais, as técnicas baseadas na análise acústica apresentam como vantagens: propiciar um exame mais confortável ao paciente, fornecer uma avaliação quantitativa e possibilitar o desenvolvimento de sistemas automáticos de auxílio ao diagnóstico por computador por um baixo custo. Pretende-se que a pesquisa descrita neste trabalho seja de valor para o auxílio ao diagnóstico médico na detecção de patologias da laringe através da utilização de um método não invasivo. Neste sentido, este artigo tem como objetivo caracterizar os sinais de voz saudável e patológicos com o auxílio de uma ferramenta matemática amplamente conhecida na literatura chamada Transformada Wavelet.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma breve descrição sobre as pregas vocais saudáveis e patológicas. Na Seção 3, as amostras de voz utilizadas no estudo são descritas em maiores detalhes e são apresentados os métodos utilizados no trabalho. Na Seção 4, a abordagem proposta de extração de atributos usando Transformada Wavelet Discreta é descrita. Na Seção 5, são apresentados os resultados da pesquisa usando essa nova abordagem e, por fim, a última seção apresenta os comentários finais sobre o trabalho.

2. FUNDAMENTOS DA PRODUÇÃO DA FALA

Nesta Seção, são descritos alguns detalhes sobre as pregas vocais saudáveis e patológicas, considerando os nódulos vocais e o edema de Reinke.

2.1 Pregas Vocais Saudáveis

A fala é uma das capacidades ou aptidões que os seres humanos possuem de comunicação, manifestando seus pensamentos, opiniões e sentimentos através dos vocábulos [5]. Existem duas principais fontes de características da fala, específicas aos locutores: as físicas e as adquiridas (ou aprendidas). As características físicas relacionam-se principalmente ao trato vocal, estrutura formada pelas cavidades que vão das pregas vocais até os lábios e o nariz [6].

A produção da voz humana depende de uma conjunção de vários mecanismos, como a ressonância do trato vocal, as pressões subglotal e supraglotal, as características biomecânicas dos tecidos do trato vocal e a oscilação do trato vocal [1]. As pregas vocais vibram com a passagem de ar vindo dos pulmões. Essa vibração das pregas vocais produz um som fraco e constituído de poucos harmônicos, que é amplificado quando passa pelas cavidades de ressonância (laringe, faringe, boca e nariz) e ganha "forma" final quando é articulado através de movimentos de língua, lábios, mandíbula, dentes e palato [6].

A passagem do ar pelas cavidades do trato vocal altera o espectro do som devido às ressonâncias, que formam picos de energia no espectro de frequência conhecidos como formantes. Através da análise espectral da fala produzida é possível estimar a forma do trato vocal [7].

De acordo com Hirano [8], a vibração do trato vocal determina a qualidade da voz. Em uma simples representação, consiste de cinco camadas de estrutura complexa: o epitélio, as camadas superficial, intermediária e profunda da lâmina própria (LP) da mucosa e o músculo vocal. A camada superficial da LP consiste principalmente de uma substância muito maleável amorfo conhecido como espaço de Reinke (RS). A camada intermediária da LP é composta principalmente de fibras elásticas, enquanto que a camada de fundo é composta basicamente por fibras colágenas. A estrutura consistindo de camada intermediária e profunda é chamada de ligamento vocal [1].

O movimento das pregas vocais pode ser observado como a sobreposição de duas componentes principais: a dinâmica do músculo vocal (corpo) e um dos epitélio e lâmina própria superficial (cobertura), conhecida como a onda mucosal. Este processo é descrito como uma onda que se propaga sobre o tecido de cobertura durante o ciclo de fonação, consistindo em um deslocamento dos tecidos em relação ao corpo [1].

Na imagem do videolaringoscópio, mostrada na Figura 1(a), a fase de fechamento deve ser observado pelo contato total entre as bordas das duas pregas vocais resultantes de perturbações produzidas pelas ondas que viajam sobre as estruturas de cobertura respectiva, e produzindo o som da voz. Durante a respiração, as pregas vocais abertas permitem a passagem de ar através da glote, como mostrado na Figura 1(b).

2.2 Nódulos Vocais

O nódulo vocal é definido como uma lesão benigna, que ocorre em ambos os lados da prega vocal, estritamente simétrica na fronteira do terço anterior e médio da prega vocal e geralmente imóvel durante a fonação. A lesão está confinada à camada superficial da lâmina própria e é o resultado da colisão traumática e constante das pregas vocais, causados pela sobrecontração dos músculos da laringe intrínseca durante a fonação. De acordo com diversos estudos, a formação dos nódulos ocorre principalmente no ponto médio da prega vocal membranosa, onde as forças de impacto são os maiores e estes são geralmente bilateral [1].

(a) Fase de fechamento das pregas vocais, cujo contato pleno entre as duas fronteiras pregas vocais podem ser observados, resultando na produção de som da voz.

(b) Pregas vocais abertas durante a respiração, permitindo a passagem de ar através da glote.

Figura 1: imagens de videolaringoscopia coletadas no Departamento de Otorrinolaringologia e no Departamento de Cirurgia de Cabeça e Pescoço da Faculdade de Medicina de Ribeirão Preto, Estado de São Paulo, Brasil [1].

Durante a fase final da vibração das pregas, a presença de nódulos na camada exterior do tecido das pregas vocais inibe-as de serem completamente dobradas umas sobre as outras. Consequentemente, o fechamento da glote é inacabado, acrescentando ar turbulento ao sinal de voz. A fim de reduzir este efeito, os indivíduos aumentam a tensão muscular e a pressão subglótica, aumentando as forças de colisão vocal [9].

O uso excessivo e os hábitos vocais inadequados são fatores determinantes no desenvolvimento de nódulos nas pregas vocais, sendo frequentemente visto em pessoas que utilizam a voz profissionalmente, como professores, palestrantes e cantores, bem como em crianças.

Além do abuso vocal, outros fatores predisponentes têm sido destaque na gênese de nódulos de pregas vocais, como obstrução nasal, rinossinusite recorrente, insuficiência velofaríngea, hipoacusia, e refluxo gastroesofágico [10]. Os principais sintomas da presença de nódulos são rouquidão, falta de ar, fadiga vocal, desconforto na garganta e redução na extensão vocal durante o dia. Essa doença é a mais recorrente nos pacientes das clínicas de voz [1].

2.3 Edema de Reinke

O edema de Reinke é um inchaço generalizado das pregas vocais, devido ao acúmulo de líquido na camada superficial da lâmina própria [11]. Recebe este nome por se localizar no espaço anatômico com o nome do anatomista Reinke, que foi o primeiro a investigar a anatomia das pregas vocais. Ele descreveu a frouxa camada subepitelial das pregas vocais. Camada esta limitada acima pela linha arqueada superior e abaixo pela linha arqueada inferior na junção do epitélio cilíndrico com o escamoso [12]. É uma doença laríngea benigna que ocupa toda a prega vocal e faz saliência com a glote. Pode ocorrer em ambas as pregas vocais ou ser limitada a uma prega vocal, geralmente no início da doença. Nessa doença, o líquido acumulado no forro da submucosa do espaço de Reinke faz com que a cobertura das pregas vocais fique menos rígida e mais maciça, como uma estrutura de vibração, geralmente evoluindo para uma irritação crônica das pregas vocais, modificando a permeabilidade capilar dos tecidos [13].

Essa doença está frequentemente associada com fatores etiopatogênicos, como tabagismo, sinusite ou infecção do trato respiratório superior, e com o uso excessivo da voz, também conhecido com fonotrauma [1, 14]. O refluxo gastroesofágico ou irritação persistente também é considerado um fator que contribui para o edema de Reinke [12]. Entretanto, conforme citado por [1] e [15], ainda não há evidências se as condições alérgicas ou medicamentos também sejam fatores que contribuem para esta doença.

O edema de Reinke ocorre gradualmente ao longo do tempo, sendo percebido inicialmente pelo paciente com a voz em tom baixo. Com o tempo, a pronúncia da voz torna-se difícil e o paciente pode notar um aumento do esforço associado com a fala. Devido a esta característica gradual, os pacientes geralmente procuram um especialista somente após mudanças significativas na qualidade de voz [1].

3. Extração e Classificação de Atributos de Sinais da Voz

3.1 Amostras de Voz

Neste trabalho, são utilizados sinais de voz coletados de 60 voluntários, com idade entre 18 e 90 anos, divididos em 4 grupos de aproximadamente mesmo tamanho. O primeiro grupo é composto de amostras de voz de pessoas saudáveis, sem patologias na laringe. O segundo grupo é composto por pessoas com nódulos vocais em diferentes estágios de evolução da doença de acordo com Scalassara [16]. O terceiro grupo é composto por amostras de pessoas com edema de Reinke. E o último grupo é composto por pacientes que apresentam diferentes distúrbios neurológicos, como AVC (Acidente Vascular Cerebral), Doença de Huntington, Doenças de Parkinson, ELA (Esclerose Lateral Amiotrófica), Mononeurite múltipla, Mitocondropatia, Distrofia de Duchenne, Distrofia Miotônica e Distonia Cervical, e que são agrupados como pacientes com disfonia neurológica.

Essas amostras de voz são parte de um banco de dados de voz do Grupo de Bioengenharia da Escola de Engenharia de São Carlos da Universidade de São Paulo, Brasil. Esses sinais foram coletados ao longo dos últimos 10 anos e foram utilizados em diversos estudos [1, 16, 17]. Os pacientes submetidos à análise foram diagnosticados por médicos do Departamento de Otorrinolaringologia e do Departamento de Cabeça e Pescoço do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Estado de São Paulo, Brasil, por meio de vídeo-laringoscópio e luz estroboscópica. Os indivíduos do grupo de pessoas saudáveis também foram diagnosticados para provar a ausência de qualquer patologia.

De acordo com [1], as amostras foram gravadas usando uma metodologia simular à apresentada em [18], na qual os indivíduos foram convidados a produzir a vogal sustentada /a/ mantendo o tom e o nível de intensidade confortáveis por cerca de 5 segundos. No momento da aquisição da voz, verificaram se o indivíduo poderia lidar com o intervalo de fonação, pois a manutenção do enunciado provoca um aumento da frequência fundamental e uma estabilidade artificial sobre a sua produção [17]. Das diversas coletas consecutivas, foi selecionado o sinal com menor variabilidade da voz. Todas as amostras de vozes foram quantizadas em amplitude com 16 bits e gravada em formato WAV mono-canal para preservar a fidelidade do sinal. A frequência de amostragem foi de 22.050 Hz [1].

3.2 Transformada *Wavelet*

A análise por Transformada de Fourier assume que o sinal é estacionário no tempo, porém, no caso de sinais de fala, isto não é verdade. No processamento dos sinais de fala, uma versão janelada do sinal é usada com a suposição de estacionaridade durante esta janela. Este método é conhecido como Transformada de Fourier de Tempo Curto (STFT), onde se uma janela de curta duração for escolhida, a frequência de resolução será pequena. Assim, fixando o tamanho da janela, a resolução tempo-frequência conseguida pela STFT é também fixada [19].

Para superar o problema da resolução fixa, a Transformada *Wavelet* utiliza uma janela de tamanho adaptativo, que utiliza mais tempo para baixas frequências e menos tempo para altas frequências [20]. Essa análise pode ser usada para um sinal que tem componentes de alta-freqüência de curta duração e componentes de baixa-frequência de longa duração, como no caso da fala [19].

A função base usada na Transformada *Wavelet* é localizada tanto no tempo como na freqüência. Todas as funções *wavelet* são versões escalonadas de uma função protótipo $\psi(t)$, também conhecida como *wavelet* 'mãe', de média zero e centrada na vizinhança de $t = 0$, dada por [7]

$$\psi_{\tau,a}(t) = a^{-\frac{1}{2}} \cdot \psi\left(\frac{t-\tau}{a}\right), \tag{1}$$

em que os parâmetros τ e a são chamados parâmetros de translação e escalamento, respectivamente. O termo $a^{-1/2}$ é usado para normalização da energia. A Transformada *Wavelet* Discreta (DWT) de um sinal $x(t)$, em que $x \in \mathbf{L^2}$, é dada pela equação [7]

$$D(j,k) = 2^{-\frac{j}{2}} \sum_i x(i) \cdot \psi^*\left(2^{-j}i - k\right), \tag{2}$$

em que i, j e k são inteiros. Escolhendo o fator de escalamento como diádico (2^j), a transformada resultante é conhecida como DWT diádica.

Se a decomposição *wavelet* for computada para uma escala 2^j, a representação do sinal resultante não é completa. As componentes de baixa frequência correspondentes as escalas maiores que 2^j são avaliadas usando [7].

$$A(j,k) = 2^{-\frac{j}{2}} \sum_i x(i) \cdot \phi^*\left(2^{-j}i - k\right), \tag{3}$$

em que $\phi^*(i)$ é o complexo conjugado da função de escalamento $\phi(i)$.

A DWT pode ser vista como um processo de filtragem do sinal, usando um filtro passa-baixas (escalamento) e um filtro passa-altas (*wavelet*). Então, o primeiro nível de decomposição DWT de um sinal divide em duas faixas, uma versão passa-baixas e uma versão passa- altas do sinal. A versão passa-baixas fornece a representação aproximada do sinal, enquanto a passa-altas indica os detalhes ou variações de altas frequências. O segundo nível de decomposição é executado sobre a versão passa-baixas do primeiro nível de decomposição, como mostrado na Figura 2. Então, a decomposição *wavelet* resulta em uma árvore cuja estrutura é dita recursiva.

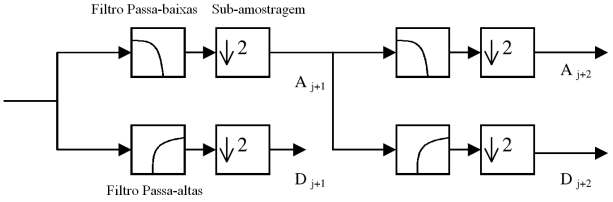

Figura 2: decomposição de sinal usando DWT diádica [19].

3.3 Reconhecimento de Padrões e Redes Neurais Artificiais

O Reconhecimento de Padrões (RP) trata da classificação de uma estrutura de dados através de um conjunto de propriedades ou características, envolvendo técnicas para a atribuição dos padrões a suas respectivas classes, de forma automática ou com a menor intervenção humana possível. Um padrão é uma descrição de um objeto e a classe de padrões é uma família de objetos que compartilham uma mesma propriedade. Reconhecimento de voz, impressão digital, estrutura de íris e diagnósticos médicos são alguns exemplos de aplicações de RP [7].

A classificação de padrões pode ser definida como um problema relacionado com a determinação de uma fronteira de decisão que consegue distinguir diferentes padrões em um conjunto de classes dentro de um espaço de características d-dimensional, em que d é o número de características. A classificação, então, pode ser realizada, e pode ser entendida, de maneira geral, pela partição do espaço de atributos em um número finito de regiões de tal forma que objetos de uma mesma classe recaiam, pelo menos em tese, sempre dentro de uma mesma região. Existem diversos classificadores conhecidos na literatura, sendo as redes neurais artificiais (RNA) um dos mais conhecidos e, em função disso, foi utilizado neste trabalho para classificar as doenças da laringe.

Modelos de RNAs são especificados pela topologia da rede, características dos neurônios e regras de aprendizagem ou treinamento. O termo topologia refere-se à estrutura da rede como um todo, especificando como as entradas, as saídas e as camadas escondidas são interconectadas [21]. Neste trabalho, é utilizada a topologia do Perceptron Multi-Camadas (MLP), por ser uma das RNAs mais utilizadas para separar dados não linearmente separáveis.

O número de neurônios das camadas de entrada e saída é determinado conforme o problema em questão. Já o número de camadas ocultas e o de neurônios nelas são definidos de forma intuitiva, não havendo, portanto, uma regra que defina o seu número. Se a quantidade de neurônios escolhidos for muito pequena, isto pode fazer com que apenas alguns neurônios especializem-se em características não úteis, tais como ruído. Se o número de neurônios for insuficiente, pode acontecer da rede não conseguir aprender os padrões desejados, tornando a especificação do número de neurônios uma etapa importante para construção de uma boa RNA [21].

4. ABORDAGEM PROPOSTA

A fim de ser analisado, selecionou-se um trecho de cada amostra de voz com características acústicas adequadas, a fim de eliminar fenômenos transitórios e variações da voz presentes na gravação. Para todas as amostras, selecionou-se cerca de 1 s de amostra de voz e dividiu-se em 40 partes de 512 pontos, de aproximadamente 24 ms cada. Antes de aplicar a técnica de extração de características, os sinais foram pré-processados através de uma normalização para resultar em sinais com amplitudes entre 0 e 1.

Para identificar uma dada patologia da laringe, algumas características do sinal de voz no tempo/frequência ou em algum outro domínio devem ser conhecidos [19]. Através da extração de características, o espaço de dados é transformado num espaço de características que possui a mesma dimensão do espaço de dados original, porém é representado por um número reduzido de características efetivas.

O método proposto consiste em aplicar a Transformada Wavelet Discreta sobre um frame de áudio, com duração de 10 a 30 milissegundos (ms), extraindo algumas características deste frame. A *wavelet*-base escolhida é a *Daubechies*, que são um família de *wavelets* ortogonais definidas para a DWT e caracterizadas por um número máximo de momentos nulos para um dado suporte. Para cada tipo de *wavelet* desta classe, existe uma função de escalamento (também chamada de *wavelet* 'pai') que gera uma análise multi-resolução ortogonal. Em geral, as *wavelets Daubechies* são escolhidas por ter um maior número N de momentos nulos para $\psi(t)$ e por dar largura de suporte $2N-1$, em que N é a ordem da *wavelet*. Entre as $2N-1$ possíveis soluções, é escolhida a solução cujo filtro de escalamento tenha fase extrema [22].

Nesse método, é escolhido um frame de 512 amostras, cuja duração é de aproximadamente 24 ms. Este *frame* é dividido em 4 sub-*frames* de duração de 6 ms para poder acomodar rápidas oscilações e permitir acompanhar a evolução temporal da energia média por amostra em cada faixa. Como o sinal de fala é estacionário para duração de até 10 ms (devido à limitação física do movimento das articulações da produção da fala), qualquer redução adicional na duração do frame não é útil [19].

A decomposição *wavelet* é aplicada em cada sub-*frame* e a energia dos coeficientes *wavelet* em cada faixa de freqüência é calculada. Esta energia é normalizada pelo número de amostras na faixa correspondente, resultando desse modo uma energia média por amostra em cada faixa. A normalização é essencial porque cada faixa terá um número diferente de amostras. Estas energias médias por amostra, para diferentes faixas, são usadas como características para classificação. As características extraídas em um sub-*frame* base não fornecem somente a energia em cada faixa, mas fornecem também uma ideia da variação temporal da energia em cada faixa.

A quantidade de características extraídas depende do nível de decomposição de cada sub-*frame*. Uma decomposição nível N de um sub-*frame* extrai $N+1$ características. Desta forma, sobre um frame de 512 amostras, extrai-se um total de $4N+4$ características. Uma vez extraídas as características, o problema consiste em obter uma função discriminante para separar as diferentes classes presentes no espaço de características. A Tabela 1 apresenta um resumo do algoritmo proposto.

O classificador utilizado para classificar os atributos extraídos é uma rede neural MLP com 2 camadas, sendo uma camada oculta cujo número de neurônios q foi determinado usando a Regra de Komolgorov, $q = 2 \cdot n + 1$, em que n é o número de atributos usados na classificação. Os parâmetros de treinamento da rede neural são escolhidos para se obter um estudo mais preciso e são: 200 épocas de treinamento, MSE desejado de 10^{-4}, passo de aprendizagem de $0,01$ e fator de momento $0,2$.

Tabela 1: resumo do algoritmo usado para a análise do sinal de voz.

1. Avaliação das amostras de voz
   - 1.1 Eliminação dos fonemas indesejados
   - 1.2 Seleção das partes estacionárias WSS
2. Pré-processamento do sinal
   - 2.1 Normalização da amplitude
3. Transformada Wavelet Discreta
   Para cada frame de 512 pontos
   - 3.1 Dividir em 4 sub-*frames*
   - 3.2 Aplicar a decomposição wavelet em cada sub-*frame*
   - 3.3 Calcular a energia média em cada faixa de frequência de cada sub-*frame*
   - 3.4 Agrupar as energias médias em um vetor de atributos

Para avaliar o desempenho de classificação, o conjunto de dados é dividido em dois conjuntos: um de treinamento e outro de teste. As amostras são embaralhadas aleatoriamente e 80% delas são atribuídas ao conjunto de treinamento, enquanto os 20% restantes são utilizados no teste. A avaliação da técnica é feita com base nas taxas de acerto média, máxima, mínima e desvio-padrão. Os resultados são extraídos de 20 simulações independentes.

5. RESULTADOS

Nesta seção são descritos os resultados e as avaliações deste trabalho baseados nas métricas descritas na seção anterior.

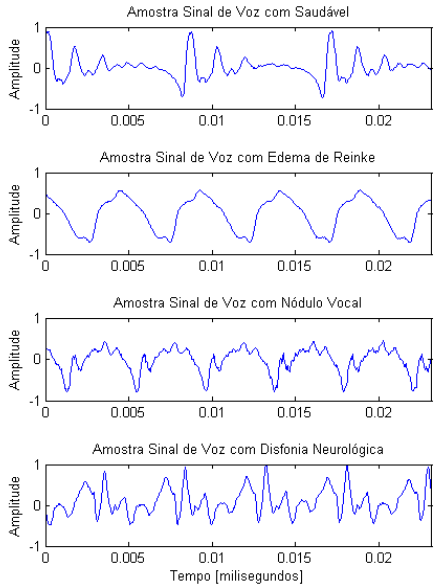

Figura 3: exemplos de sinal de voz para vogais sustentadas /a/.

Quatro exemplos de sinais de voz, cada um referente a uma das classes, são mostrados na Figura 3. O primeiro é um típico sinal de voz saudável para a vogal sustentada /a/, o segundo é um sinal de Edema de Reinke, o terceiro é uma sinal de Nódulo Vocal e o último é de um sinal caracterizado como Disfonia Neurológica.

As estatísticas da taxa de acerto e o desvio padrão da taxa de acerto obtidos pelo método proposto,usando as amostras de voz saudável, nódulo vocal e edema de Reinke, são mostradas na Tabela 2, em que observa-se que a taxa média de acerto varia entre 82% e 94%. O melhor resultado obtido foi para uma decomposição *wavelet* nível 6 em cada sub-*frame*.

As taxas de acerto médio para as três classes analisadas é mostrado na Tabela 3. Nesta Tabela verifica-se que os melhores resultados para as classes Edema de Reinke e Nódulo Vocal ocorrem para uma decomposição *wavelet* nível 6, com taxas acima de 90%. Enquanto para a classe de vozes saudáveis, o melhor resultado ocorreu para nível 7, com quase 100% de acerto.

As taxas de acerto médio por classe para a configuração com 6 níveis de decomposição *wavelet* em cada sub-*frame*, destacadas

Tabela 2: desempenho de Classificação para as classes: voz saudável, nódulo vocal e edema de Reinke.

Nível DWT	Taxas de Reconhecimento(%)			
	mínima	*média*	*máxima*	*desvio padrão*
3	79,35	82,23	84,51	1,98
4	88,38	90,01	91,95	1,30
5	89,53	90,27	90,93	0,55
6	**91,57**	**93,03**	**94,00**	**0,70**
7	84,51	86,63	88,32	1,67

Tabela 3: desempenho de Classificação por Classe (Voz Saudável, Nódulo Vocal e Edema de Reinke).

Nível DWT	Taxa (%)		
	Ed. Reinke	Nódulo	Saudável
3	81,27	74,71	94,61
4	89,50	88,62	92,05
5	87,41	89,25	94,39
6	**91,19**	**90,37**	**97,46**
7	82,47	81,91	99,51

na Tabela 3, em que que todas as classes obtiveram uma taxa de acerto médio acima de 90%, demonstram que a abordagem proposta permite discriminar essas duas doenças da laringe com boa probabilidade de acerto.

Tabela 4: desempenho de Classificação para todas as classes.

Nível DWT	Taxas de Reconhecimento(%)			
	mínima	*média*	*máxima*	*desvio padrão*
3	76,37	77,15	78,22	0,68
4	77,34	79,41	81,05	1,37
5	78,22	79,45	80,66	1,16
6	85,35	86,86	89,16	1,41
7	87,70	89,14	90,53	1,30

Tabela 5: desempenho de Classificação por Classe (incluindo Disfonia Neurológica).

Nível DWT	Taxa (%)			
	Edema de Reinke	Nódulo	Neurológica	Saudável
3	79,33	78,18	74,07	76,89
4	83,76	79,09	72,43	82,31
5	82,55	77,26	74,50	83,38
6	88,15	84,14	81,85	93,14
7	87,61	86,56	89,40	93,33

Incluindo a classe de pessoas com disfonia neurológica, ocorre uma diminuição de 7% na taxa de acerto médio para a configuração com 6 níveis DWT em cada sub-*frame*, conforme mostrado na Tabela 4. Para obter resultados equivalentes aos obtidos com apenas 3 classes, foi necessário aumentar o número de níveis de decomposição *wavelet*, ou seja, do número de características extraídas.

Esta queda no desempenho de classificação com a adição da classe Disfonia Neurológica ocorre devido a essa classe ter algumas características espectrais parecidas com a classe de voz saudável e por ter amostras de voz de pacientes com diferentes doenças de origem neurológica.

6. CONCLUSÃO

O presente trabalho explorou o poder de uma ferramenta matemática amplamente conhecida, Transformada Discreta Wavelet, para extrair características de sinais de voz, que permitissem classificá-las, usando RNA do tipo MLP, em quatro possíveis classes: saudáveis, nódulos, edema de Reinke e disfonia neurológica.

Os resultados apresentados indicam que o método proposto é bastante eficaz, por obter resultados equivalentes aos apresentados em outros trabalhos com mesmo banco de amostras [1, 16]. Esses resultados demonstram que a abordagem proposta permite discriminar as vozes saudáveis e patológicas com taxas de acerto superiores a 90%. Algumas melhorias no método proposto estão sendo realizadas para aumentar a separação entre as todas as classes, principalmente a de Disfonia Neurológica.

Como perspectivas futuras, pretende-se ter um conjunto de atributos que permita classificar as quatro classes de amostras com taxas próximas às obtidas apenas com três classes.

REFERÊNCIAS

[1] 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". *Applied Mathematics and Computation*, vol. 270, no. 1, pp. 95 – 108, 2009.

[2] I. C. Zwetsch, R. D. R. Fagundes, T. Russomano and D. Scolari. "Processamento digital de sinais no diagnóstico diferencial de doenças laríngeas benignas". *Scientia Medica*, vol. 16, no. 3, Set 2006.

[3] A. Parraga. "Aplicação de Transformada Wavelet Packet na análise e classificação de sinais e vozes patológicas". Master's thesis, Universidade Federal do Rio Grande do Sul, Escola de Engenharia, 2002.

[4] H. H. Falcão, S. Correia, S. C. Costa, J. M. Fechine and B. Aguiar Neto. "O Uso da Entropia na Discriminação de vozes patológicas". *Salvador, Brasil*, 2008.

[5] M. B. Paula. "Reconhecimento de palavras faladas utilizando redes neurais artificiais". *Monografia de Final de Curso, Universidade Federal de Pelotas*, 2000.

[6] A. H. D. Souza Júnior. "Avaliação de Redes Neurais Auto-organizáveis para reconhecimento de voz em sistemas embarcados". *Dissertação de Mestrado, Universidade Federal do Ceará*, 2009.

[7] R. T. S. Carvalho. "Estudo Comparativo de Técnicas de Extração de Características para Reconhecimento de Fonemas". *Monografia de Final de Curso, Depart. de Eng. de Teleinformática, Universidade Federal do Ceará*, Dezembro 2009.

[8] M. Hirano. "Vocal mechanisms in singing: Laryngological and phoniatric aspects,". *Journal of Voice*, vol. 2, no. 1, pp. 51 – 69, 1988.

[9] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh and C. Vaughan. "Phonatory function associated with hyperfunctionally related vocal fold lesions". *Journal of Voice*, vol. 4, no. 1, pp. 52 – 63, 1990.

[10] R. H. G. Martins, J. Defaveri, M. A. C. Domingues, R. de Albuquerque e Silva and A. Fabro. "Vocal Fold Nodules: Morphological and Immunohistochemical Investigations". *Journal of Voice*, vol. 24, no. 5, pp. 531 – 539, 2010.

[11] B. M. J. Neves, J. A. G. Neto and P. Pontes. "Diferenciação histopatológica e imunoistoquímica das alterações epiteliais no nódulo vocal em relação aos pólipos e ao edema de laringe". *Revista Brasileira de Otorrinolaringologia*, vol. 70, pp. 439 – 448, 08 2004.

[12] O. Kleinsasser. *Microlaringoscopia e Microcirurgia da Laringe*, volume 2. Manole, São Paulo, Brasil, 1997.

[13] M. Hirano. "Structure of the vocal folds in normal and disease states: anatomical and physical studies". *Proc. of the Conf. on the Assessment of Vocal Pathology*, pp. 11–30, 1981.

[14] M. H. L. de Abreu. "Edema de Reinke: aspectos gerais e tratamento". *Monografia de Final de Curso*.

[15] M. Greene. *Distúrbios da Voz*, volume 4. Manole, São Paulo, Brasil, 1989.

[16] P. R. Scalassara, C. D. Maciel, R. C. Guido, J. C. Pereira, E. S. Fonseca, A. N. Montagnoli, S. B. Júnior, L. S. Vieira and F. L. Sanchez. "Autoregressive decomposition and pole tracking applied to vocal fold nodule signals". *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1360 – 1367, 2007.

[17] M. de Oliveira Rosa, J. Pereira and M. Grellet. "Adaptive estimation of residue signal for voice pathology diagnosis". *IEEE Trans. on Biom. Eng.*, vol. 47, no. 1, pp. 96 –104, jan 2000.

[18] V. Uloza, V. Saferis and I. Uloziene. "Perceptual and Acoustic Assessment of Voice Pathology and the Efficacy of Endolaryngeal Phonomicrosurgery". *Journal of Voice*, vol. 19, no. 1, pp. 138 – 145, 2005.

[19] O. Farooq and S. Datta. "Phoneme recognition using wavelet based features". *Information Sciences*, vol. 150, no. 1-2, pp. 5–15, Mar 2003.

[20] O. Rioul and M. Vetterli. "Wavelets and signal processing". *Signal Processing Magazine, IEEE*, 1991.

[21] S. S. Haykin. *Redes Neurais: Princípios e Prática*. Bookman, Porto Alegre, second edition, 2001.

[22] I. Daubechies. *Ten lectures on wavelets*. Society for Industrial Mathematics, Philadelphia, Pennsylvania, 1992.