ANÁLISE MEL-CEPSTRAL NA DISCRIMINAÇÃO DE PATOLOGIAS LARÍNGEAS

G. L. Ribeiro1*, R.Gomes2*, S. C. Costa3*, W.C. A. Costa4*

* Unidade Acadêmica de Indústria – Engenharia Elétrica, IFPB, João Pessoa, Brasil
e-mail: girleneng@gmail.com

Resumo: Esta pesquisa dispõe-se a avaliar a eficiência da Análise Mel-cepstral na detecção de patologias laríngeas. Como características representativas dos sinais patológicos ou saudáveis são utilizados os coeficientes mel cepstrais, com abordagem não paramétrica no domínio da frequência (*MFCC – Mel Frequency Cepstral Coefficients*). As patologias consideradas são: edema e paralisia nas pregas vocais. Para a classificação dos sinais em patológico ou saudável, é empregada a arquitetura de Redes Neurais Artificiais (RNAs - Neural Networks). Os resultados obtidos, com acurácia acima de 94%, sugerem que o método utilizado pode ser indicado como uma ferramenta não invasiva de apoio ao diagnóstico de patologias laríngeas.
Palavras-chave: Patologias laríngeas, análise mel cepstral, analise acústica.

Abstract: *The research aims to evaluate the efficiency of the Mel-cepstral analysis in the detection of laryngeal pathologies. As the representative features of the healthy and pathological signals the mel frequency cepstral coefficients, with nonparametric approach in the frequency domain (MFCC - Mel Frequency Cepstral Coefficients). The considered pathologies are paralisys and vocal fold edema. For the signal classification on healthy or pathological Artificial Neural Networks architecture (ANN - Neural Networks) is used. The obtained results suggests that the employed method can be indicated as a noninvasive tool for diagnosis aid of laryngeal pathologies.*
Keywords: *Laryngeal pathology, mel ceptral, acoustic analysis.*

Introdução

A detecção de patologias laríngeas tem sido realizadas por análise acústica, como uma proposta objetiva e não invasiva, comparada às técnicas tradicionais como exames de videolaringoscopia [1]. Diversas abordagens são empregadas sugerindo técnicas baseadas em características obtidas no domínio do tempo (*pitch*, *shimmer*, *jitter*, coeficientes de predição linear, entre outras) [1] ou no domínio da frequência (relação harmônico-ruído, análise cepstral, transformada wavelets, entre outras) [2,3].

Sinais de vozes afetados por patologias mais severas apresentam dificuldades na obtenção da frequência de vibração das pregas vocais (correlato perceptual *pitch*) e, consequentemente, de outras medidas obtidas a partir de suas variações: em amplitude (*shimmer*), em frequência (*jitter*), entre outras. Sons ditos sonoros, a exemplo dos sons vocálicos, quase periódicos, são utilizados para obtenção do *pitch* ou da frequência fundamental [4]. Medidas que não dependam do *pitch* tornam-se mais interessantes para avaliar desordens vocais mais intensas, causadas por patologias mais severas, que atinge o sinal de voz afetando sua periodicidade, dificultando a análise do mesmo pelo *pitch* e medidas derivadas do mesmo [5].

Os coeficientes mel-cepstrais buscam adequar tons e sinais de voz à percepção humana, não linear [6]. Para cada tom com frequência *f*, medida em Hz, define-se um tom subjetivo medido na escala mel. O mel, então, é uma unidade de medida da frequência percebida de um tom. Nesta pesquisa, são apresentados os resultados obtidos na discriminação entre sinais saudáveis e sinais afetados por patologias na laringe (edema e paralisia nas pregas vocais) a partir da análise mel-cepstral. Os sinais de vozes são, então, analisados no domínio mel-cepstral e os coeficientes mel-cepstrais obtidos são submetidos a um classificador baseado em redes neurais para discriminar entre sinais patológicos e saudáveis e entre as patologias consideradas (edema e paralisia).

Materiais e métodos

Base de dados - Os sinais processados são oriundos da base de dados comercial *Disordered Voice Database*, *Model* 4337, da Kay Elemetrics amplamente empregada em trabalhos científicos internacionais, desenvolvida pelo Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab [7]. A base de dados é constituída de 710 gravações de sinais de vozes sendo 657 vozes patológicas e 53 vozes saudáveis, as gravações contém de 1 a 3 segundos da vogal sustentada /a/. Foram selecionados 149 sinais de vozes distribuídos em 53 sinais de vozes saudáveis, 44 sinais de vozes afetadas por edema de Reinke e 52 por paralisia nos nervos laríngeos.

Metodologia- As principais etapas para a realização desse trabalho são: pré-processamento, análise e classificação dos sinais. Inicialmente, faz-se o pré-processamento do sinal de voz que consiste na pré-ênfase, segmentação, e janelamento dos sinais. O tamanho dos segmentos empregado é de 20 milissegundos, com sobreposição de 50% e, para o janelamento, foi utilizada a janela de Hamming. Na etapa seguinte, ocorreu a aquisição das características mel cepstrais, na qual foram extraídos 12 coeficientes

para cada segmento do sinal formando o vetor de características. No processo de formação do banco de características foram analisados 30 segmentos para cada sinal. Os vetores de características extraídos de cada sinal formaram uma matriz de características com dimensão 30x12 (12 coeficientes por segmento) correspondente a cada sinal analisado.

A classificação é realizada através da técnica de RNAs. A matriz de entrada para cada rede testada é a matriz de características 30x12 obtida nas etapas anteriores, formada pelos coeficientes mel cepstrais de cada sinal. Os padrões foram divididos em três subconjuntos: treino, validação e teste. Cada subconjunto possuiu um quantitativo de amostras de 60%, 20% e 20% respectivamente. Após a etapa de seleção das amostras em cada grupo específico, foram obtidas as medidas estatísticas de acurácia total, sensibilidade, especificidade e erro médio quadrático (EMQ). A quantidade de sinais, no classificador, foi equiparada com a classe com menor número de sinais.

A Rede neural *Perceptron* multicamadas (*multilayer Perceptron – MLP*), sua arquitetura e do tipo *feedforward*, e seu algoritmo de aprendizado, utilizado no sistema de treinamento, é baseado na regra delta generalizada ou *back-propagation*. A rede neural empregada consta de 12 entradas (referente aos 12 coeficientes mel-cepstrais por segmento), uma camada oculta com 10 neurônios e a camada de saída. A função sigmoide foi empregada para ajuste dos pesos. O treinamento de redes utilizando o algoritmo é comumente realizado mediante as duas fases, *forward* e *backward*, as quais induzem uma mudança automática nos pesos sinápticos e limiares para cada iteração, implicando-se em uma gradativa diminuição da soma dos erros produzidos pela resposta da rede frente àquelas desejadas [8].

Análise mel-cepstral – A diferença entre o cálculo dos coeficientes cepstrais e dos coeficientes mel-cepstrais está na aplicação de um banco de filtros digitais ao espectro real do sinal, antes da aplicação da função logarítmica. Tais filtros não estão linearmente espaçados no domínio da frequência. Esses filtros têm por objetivo aproximar a resposta humana a sinais sonoros.

O mapeamento entre a escala de frequência real, em Hz, e a escala de frequências percebida, em mel, é aproximadamente linear abaixo de 1000 Hz e, logarítmica, acima. Logo, o espaçamento dos filtros digitais deve respeitar a escala de frequências percebidas (escala Mel). A função de mapeamento da frequência acústica f (em Hz) para uma escala de frequências percebidas Mel (em mels) é dada por

$$F_{mel} = 2595 \cdot \log_{10}(1 + \frac{F_{linear}(Hz)}{700}), \qquad (1)$$

em que F_{linear} é a frequência linear (em Hz) e F_{mel} é a frequência percebida (em mel). Após o pré-processamento dos sinais, os coeficientes mel cepstrais são obtidos para cada segmento do sinal, de acordo com os seguintes passos [6]:

- É calculado do espectro de magnitude do sinal, $x(n)$, a partir do módulo da transformada de Fourier ($|\mathrm{FFT}(x(n)|^2$);
- Aplicação do banco de filtros triangulares em escala mel. São utilizados geralmente 20 filtros de formato triangular. No entanto, a quantidade de filtros é baseada na frequência de amostragem (F_a) ($3.\ln(F_a)$).
- Cálculo do logaritmo da energia de saída de cada filtro. A aplicação do logaritmo é necessária para a obtenção do cepstro.
- Finalmente, o processo de obtenção dos coeficientes MFCC pode ser matematicamente descrito por [8,9]:

$$c_{mel}(n) = \sum_{k=1}^{Nf} \log(Sf(k)).\cos[n(k-\frac{1}{2})].\frac{\pi}{Nf} \qquad n = 0, 1, ..., Nf \qquad (2)$$

em que Nf é o número de filtros digitais utilizados, $c_{mel}(n)$ é o n-ésimo coeficiente mel-cepstral e $Sf(k)$ é o sinal de saída do banco de filtros digitais, dado por

$$Sf(k) = \sum_{j=1}^{NFFT} W_k(j).X(j) \quad k = 1, ..., Nf, \qquad (3)$$

em que $W_k(j)$ são as janelas de ponderação triangulares associadas às escalas-mel e $X(j)$ é o espectro de magnitude da FFT de N pontos [3, 6, 10].

Utilizando o software Matlab® foi iniciada a fase de classificação dos grupos de sinais de voz disponíveis na base de dados supracitada. Os grupos foram dispostos em três combinações de grupos de sinais: paralisia e saudável; edema de Reinke e saudável; e edema e paralisia. Os resultados obtidos no processo de classificação são apresentados a seguir.

Resultados

Após processamento de cada sinal, foi realizada a etapa de classificação utilizando RNAs. Os resultados foram analisados por meio dos valores de acurácia total (relação entre o número de sinais classificados corretamente e a quantidade total de sinais), sensibilidade (corresponde ao número de vozes classificadas como patológicas corretamente, dividido pelo total de vozes patológicas usadas no teste), especificidade (corresponde ao número de vozes classificadas como saudáveis corretamente, dividido pelo número total de vozes saudáveis usadas no teste) e desempenho da classificação, que retorna o erro de classificação total da rede.

Inicialmente foi feita a classificação treinando a rede com todos os sinais, sem separação por gênero. Depois foi investigada a influência da separação por gênero na eficiência do classificador.

A Tabela 1 apresenta os valores de acurácia total, sensibilidade e especificidade obtidas para a classificação entre sinais de vozes afetados por paralisia e vozes saudáveis.

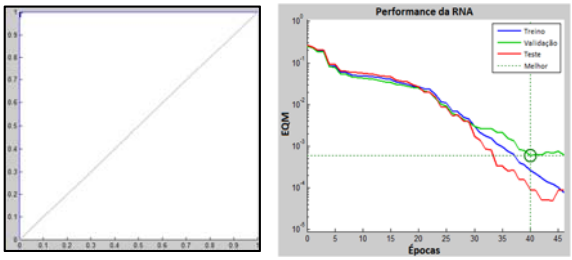
Na detecção da presença da patologia paralisia, o classificador já apresentou acurácia superior a 95% sem separação por gênero. Esta metodologia, no entanto,

permitiu uma melhor separação entre as patologias, Edema e Paralisia.

Tabela 1: Discriminação entre vozes afetadas por paralisia e vozes saudáveis, sem separação por gênero.

Medida de desempenho	Todos	Feminino	Masculino
Acurácia	100%	100%	100%
Sensibilidade	95.8%	100%	100%
Especificidade	97.8%	100%	100%

Na Figura 1, estão apresentados os gráficos correspondentes à curva ROC e ao desempenho da rede neural nas etapas de treino, validação e teste. O conjunto de validação é apresentado à rede a cada 46 épocas e calculado o EMQ, utilizado para determinar o melhor momento de parar o treinamento da rede, que obteve melhor resultado com EMQ entre 10^{-5} e 10^{0}. Esse erro é retropropagado e devidamente escalonado.



(a) (b)

Figura 1: Classificação entre voz com paralisia e voz saudável: (a) Curva ROC; (b) Desempenho da rede neural.

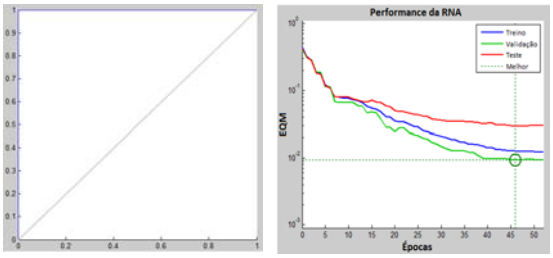
Na Tabela 2, estão apresentados os resultados obtidos para a classificação entre os grupos de vozes afetados por edema e voz saudável, com os sinais de teste. Para este caso, o classificador apresentou acurácia máxima, não sofrendo influência na separação por gênero.

Tabela 2: Grupos de vozes com edema de Reinke e voz saudável.

Medida de desempenho	Todos	Feminino	Masculino
Acurácia	100%	100%	100%
Sensibilidade	100%	100%	100%
Especificidade	100%	100%	100%

Na Figura 2, estão apresentados os gráficos correspondentes à curva ROC e ao desempenho da rede neural nas etapas de treino, validação e teste, para a classificação entre vozes com edema de Reinke e voz saudável.

Na Figura 3 é apresentada a matriz de confusão para a classificação entre os sinais afetados por paralisia e os sinais afetados por edema nas pregas vocais. Neste caso, a separação sem gênero, apresentou acurácia total de 78,3%. Neste caso, 76,2% dos sinais com edema (classe 1) e 80,7% dos sinais com paralisia (classe 2) foram classificados corretamente. Os números na matriz de confusão se referem à classificação por segmento (fase de teste), sendo 20% do total de sinais.



(a) (b)

Figura 2: Classificação entre Edema de Reinke e voz saudável: (a) Curva ROC; (b) Desempenho da rede neural.

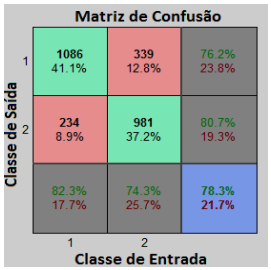
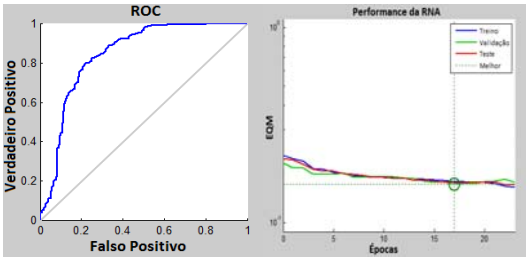


Figura 3: Matriz de confusão entre os grupos de vozes com edema e com paralisia, sem separação por gênero.

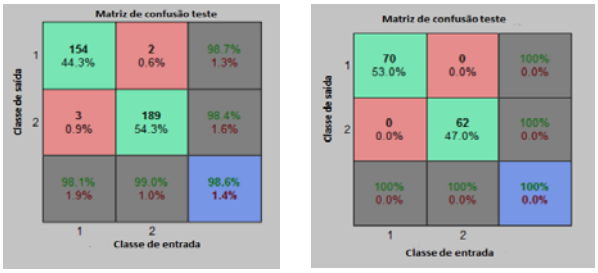


(a) (b)

Figura 4: Classificação entre vozes com edema e com paralisia, sem separação por gênero: (a) Curva ROC; (b) Desempenho da rede neural.

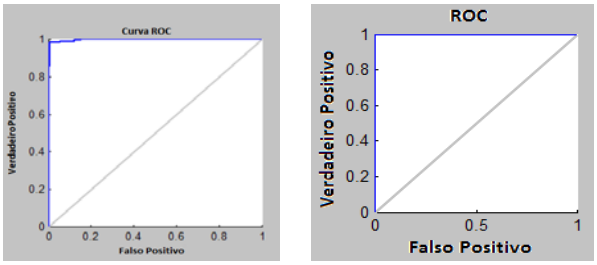
No entanto, ao realizar a separação, o ganho no desempenho do classificador foi bastante significativo. Nas Figuras 5 e 6 são mostrados os resultados para a classificação entre edema (1) e paralisia (2), com os sinais do gênero feminino (acurácia de 98,6%) e masculino (acurácia de 100%), respectivamente.

Na Figura 6 são apresentadas as curvas ROC e na Figura 7 o desempenho da rede neural na classificação entre vozes com edema e vozes com paralisia. Os sistemas que possuem respostas entre a linha diagonal e a curva perfeita são considerados adequados à análise realizada. O grupo masculino obteve maior acurácia.



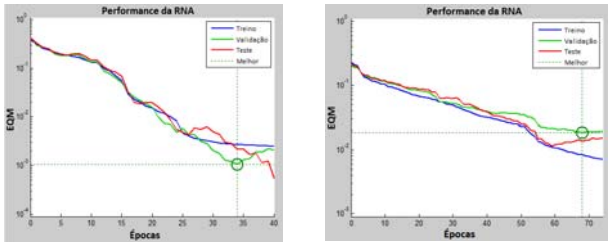
(a) (b)

Figura 5: Matriz de confusão dos grupos Edema e Paralisia: (a) feminino; (b) masculino.



(a) (b)

Figura 6: Curvas ROC para os grupos edema e paralisia para os gêneros: (a) feminino; (b) masculino.



(a) (b)

Figura 7: Desempenho do classificador para discriminação entre edema e paralisia, para os gêneros: (a) feminino; (b) masculino.

Discussão

O sistema proposto se mostrou capaz de classificar os sinais de voz da base de dados estudada com valores de acurácia com alto grau de sucesso, 100% no melhor caso e 76,2% no pior caso, sem separação por gênero. A separação por gênero apresentou resultado relevante na discriminação entre as patologias consideradas, com uma acurácia entre 98,6% e 100%. Esta separação pode ser realizada numa etapa anterior à classificação pelo profissional da área médica e indicado ao sistema de classificação, tornando a classificação mais rápida e eficiente. Os resultados indicam que os coeficientes mel-cepstrais conseguem representar bem a desordem vocal presente nos sinais afetados pelas patologias edema e paralisia nas pregas vocais, por se aproximar da percepção auditiva humana. Este mesmo fator confunde as patologias, cuja percepção melhora na separação por gênero. Para que um sistema de apoio ao diagnóstico de patologias diversas possa ser implementado, torna-se necessário treinar o sistema com outras patologias e avaliar o desempenho do mesmo.

Foram obtidas taxas de classificação entre 98,6% e 100%, na discriminação entre patologias (edema e paralisia) e de 100% entre sinais saudáveis e patológicos, com separação por gênero. Portanto, os parâmetros mel-cepstrais podem ser considerados para serem empregados numa ferramenta computacional para auxílio ao pré-diagnóstico de patologias laríngeas e na avaliação de tratamentos terapêuticos e pós-cirúrgicos.

Agradecimentos

Ao CNPq/Projeto PIBITI e ao IFPB pelo suporte à esquisa. À UFCG pelo fornecimento da base de dados.

eferências

[1] Costa SLNC. Análise Acústica, Baseada no Modelo Linear de Produção da Fala, para Discriminação de Vozes Patológicas. Tese de doutorado, Universidade Federal de Campina Grande (UFCG), 2008, 141p.

[2] Umapathy K, Krishnan S, Parsa V, Jamieson DG. Discrimination of Pathological Voices Using a Time-Frequency Approach. IEEE Transactions On Biomedical Engineering, Vol. 52., No. 3, March, 2005.

[3] Zwetsch, IC, Ribeiro, RD, Fagundes, TR, Scolari, D. Processamento Digital de Sinais no Diagnóstico Diferencial de Doenças Laríngeas Benignas. Scientia Medica, Porto Alegre: PUCRS, Vol. 16, n. 3, jul./set. 2006.

[4] Rabiner, LR, Schafer, RW. Digital processing of speech signals. New Jersey: Prentice-Hall, 1978.

[5] Godino-Llorente JI, Gómez-Vilda P, Blanco VM. Dimensionality Reduction of a Patholgical Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Transactions on Biomedical Engineering, Vol. 53, No. 10, October, 2006.

[6] Deller Jr. R, Proakis JG, Hansen JHL. Discrete-time Processing of Speech Signals. Macmillan Publishing Co., 1993.

[7] Kay Elemetrics, Kay Elemetrics Corp. Disordered Voice Database1.03 ed. 1994.

[8] Silva IN, Spatti DH, Fllauzino AR. Redes Neurais Artificiais: para engenharia e ciências aplicadas. São Paulo: Artliber, 2010.

[9] O'Shaughnessy D. Speech Communications: Human and Machine, 2nd Edition, NY, IEEE Press, 2000.

[10] Andreão RV. Implementação em Tempo Real de Um Sistema de Reconhecimento de Dígitos Conectados. Universidade Estadual de Campinas - Faculdade de Engenharia Elétrica e de Computação. Dissertação de Mestrado. Janeiro 2001.