Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional (IX CBRN)
Ouro Preto 25-28 de Outubro de 2009

DETECÇÃO AUTOMÁTICA DE PATOLOGIAS DA LARINGE USANDO CODIFICAÇÃO POR PREDIÇÃO LINEAR E REDES NEURAIS MLP

João Vilian de Moraes Lima Marinus*, Herman Martins Gomes*, Joseana Macedo Fechine*, Silvana Cunha Costa†

*Universidade Federal de Campina Grande, Departamento de Sistemas e Computação, Rua Aprígio Veloso, 882, Bairro Universitário, Campina Grande, Paraíba, Brasil.

†Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Av. 1º de Maio, 720 - Jaguaribe, João Pessoa, Paraíba, Brasil.

Emails: vilian@dsc.ufcg.edu.br, hmg@dsc.ufcg.edu.br, joseana@dsc.ufcg.edu.br, silvana@ifpb.edu.br

Abstract— The laryngeal diseases affect many professionals who use their voices as the main working tool, such as teachers, for example. Advanced diagnosis techniques of these diseases are typically invasive, causing much discomfort to the patient. In recent years techniques of Digital Speech Processing has been investigated to obtain non-invasive systems to aid the diagnosis by a specialist. The work presented proposes a method of analysis that uses coefficients obtained from Linear Prediction Coding to represent the voice signals and Multilayer Perceptron Neural Networks for classification between normal voice and pathological voice. An experimental evaluation of the method has demonstrated that this is a promising approach for discriminating between pathological and normal voices, reaching a correct classification rate above 98%.

Keywords— Laryngeal Pathologies, Linear Prediction Coding, Multilayer Perceptron Neural Networks.

Resumo— As patologias da laringe afetam muitos profissionais que têm a voz como principal instrumento de trabalho, tais como professores, por exemplo. Técnicas avançadas de diagnóstico dessas patologias são tipicamente invasivas, causando muito desconforto ao paciente. Nos últimos anos, têm sido pesquisadas técnicas de Processamento Digital de Voz para obtenção de sistemas não invasivos para o auxílio ao diagnóstico por um especialista. O trabalho ora apresentado propõe um método de análise que utiliza coeficientes obtidos a partir da Codificação por Predição Linear para representação dos sinais de voz e Redes Neurais Multilayer Perceptron para classificação entre voz normal e voz patológica. Uma avaliação experimental demonstrou que se trata de um método promissor na discriminação entre voz normal e voz patológica, atingindo uma taxa de acerto superior a 98%.

Keywords— Patologias Laríngeas, Codificação por Predição Linear, Redes Neurais Multilayer Perceptron.

1 Introdução

As patologias da laringe se caracterizam, normalmente, por apresentar a disfonia como sintoma principal ou secundário (Fukuda, 2003). Disfonia é qualquer dificuldade de emissão vocal que modifique a produção normal da voz, causada por alterações orgânicas ou funcionais da laringe (Albernaz et al., 1997). As principais causas da disfonia são inadaptações fônicas, alterações psicoemocionais e o uso incorreto da voz (Fukuda, 2003). O grupo profissional em que existe uma maior ocorrência do problema é o dos professores (Costa et al., 2000; Gotaas and Starr, 1993; Koufman and Isaacson, 1991; Sapir et al., 1993).

Existem várias técnicas para o diagnóstico de patologias da laringe. Uma técnica bastante utilizada consiste na escuta da voz do paciente pelo médico que decide se há ou não patologia. Essa técnica é problemática por possuir um caráter subjetivo, estando sujeita a erros. Outras técnicas são aplicadas no sentido de evitar esse problema, tais como laringoscopia, glotografia, estroboscopia, electromiografia e videoquimografia (Kukharchik et al., 2007). É possível diagnosticar com precisão as mais diversas patologias da laringe a partir dessas técnicas. O problema é que elas são invasivas, causando muito desconforto ao paciente, o que gera resistência por parte deste no momento em que se efetua o exame, podendo causar distorções nos dados obtidos e, assim, produzir falsos diagnósticos (Adnene and Lamia, 2003; Alonso et al., 2001).

Nos últimos anos, estão sendo realizadas várias pesquisas na área de Processamento Digital de Sinais de Voz, visando o desenvolvimento de técnicas que permitam o diagnóstico preciso de patologias da laringe de maneira não invasiva, eliminando o desconforto no momento do exame.

Diferentes abordagens para extração de características foram propostas. Inicialmente, foram usados parâmetros como *pitch*, *jitter* , *shimmer*, quociente de perturbação de amplitude, quociente de perturbação do *pitch*, relação sinal-ruído, energia de ruído normalizada (Manfredi, 2000), dentre outros (Rosa et al., 2000; Wallen and Hansen, 1996). Entretanto, muitos desses parâmetros são baseados na extração da frequência fundamental, o que pode ser uma tarefa complexa devido à característica ruidosa do sinal afetado pela patologia (Boyanov et al., 1993; Manfredi et al., 1999).

Atualmente, são realizadas pesquisas envolvendo a Codificação por Predição Linear (Costa, 2008; Aguiar-Neto et al., 2008), Análise Cepstral e derivados (Costa, 2008; Aguiar-Neto et al., 2008; Martinez and Rufiner, 2000) e coeficientes melcepstrais (Costa, 2008; Aguiar-Neto et al., 2008;

Godino-Llorente and Gómez-Vilda, 2004; Godino-Llorente et al., 2006), que se baseiam em um modelo linear para o processo de produção da fala.

Entre as técnicas mais utilizados na construção de classificadores para a discriminação entre voz normal e voz patológica estão a Quantização Vetorial (Aguiar-Neto et al., 2008), a Mistura de Densidades Gaussianas (Godino-Llorente et al., 2006), as Máquinas de Suporte Vetorial (Kukharchik et al., 2007), os Modelos de Markov Escondidos (Costa, 2008) e Redes Neurais Artificiais, com destaque para as Redes Multilayer Perceptron (Martinez and Rufiner, 2000; Godino-Llorente and Gómez-Vilda, 2004).

O trabalho tratado neste artigo objetiva o desenvolvimento de um sistema de detecção de patologias da laringe para o auxílio ao diagnóstico por um especialista, que utilize coeficientes obtidos a partir da Codificação por Predição Linear (*LPC - Linear Prediction Coding*) como características representativas da fala, e Redes Neurais Multilayer Perceptron para classificação. Serão utilizados sinais de voz da vogal sustentada /a/ por ser esta bastante utilizada na área médica para diagnóstico de patologias na laringe, já que ao pronunciá-la, as dobras vocais vibram, permitindo observar as variações na voz e diagnosticar a presença ou não de patologias (Alonso et al., 2005).

O trabalho é direcionado para os casos de patologias laríngeas. Considerando o trato vocal saudável, as desordens presentes no sinal de voz serão atribuídas a alterações laríngeas. Dessa forma, é possível, a partir da Análise por Predição Linear, estimar o comportamento da voz na presença dessas patologias. A escolha das Redes Neurais MLP se deu devido ao seu desempenho e uso intenso na literatura de reconhecimento de padrões (Bishop, 1995), facilidade de implementação e de utilização de simuladores (WEKA, 2009).

Este artigo está organizado como segue. Na Seção 2, são apresentados os fundamentos de Processamento Digital de Sinais de Voz. A base de dados utilizada nos experimentos é apresentada na Seção 3. A descrição do método proposto encontra-se na Seção 4. Na Seção 5, são apresentados e discutidos os resultados obtidos. Por fim, a Seção 6 contém as conclusões e as sugestões para os trabalhos futuros.

2 Fundamentos de Processamento Digital de Sinais de Voz

Duas etapas são fundamentais para a maioria dos métodos de Processamento Digital de um sinal de voz: pré-processamento e extração de características. Nesta seção, é apresentada uma breve introdução sobre as técnicas comumente usadas para o pré-processamento do sinal de voz. Em seguida, será descrita a técnica utilizada para extração de características do sinal adotada neste trabalho, a Codificação por Predição Linear.

2.1 Pré-processamento do Sinal de Voz

Antes da extração de características do sinal de voz, é realizado um pré-processamento desse sinal. O pré-processamento é levado a efeito, comumente, a partir das etapas de pré-ênfase e janelamento do sinal (Papamichalis, 1997; Kil and Shin, 1996; Picone, 1993).

A pré-enfase consiste na aplicação de um filtro FIR (*Finite Impulse Response* ou Resposta ao Impulso Finita) com o objetivo de atenuar os componentes de baixa freqüência do sinal de voz, minimizando o efeito da radiação do som pelos lábios e da variação da área da glote quando da elocução de um sinal (Sotomayor, 2003).

Após a pré-ênfase, é realizado o janelamento do sinal de voz, que consiste na divisão do sinal em pequenos segmentos, de forma a garantir a sua estacionariedade. Para tanto, o sinal é multiplicado por uma função janela (Haykin and Veen, 2002). Entre as funções janelas mais utilizadas, destaca-se a janela de Hamming, que consiste em uma função janela senoidal que, ao ser aplicada ao segmento do sinal de voz, mantém as características espectrais do centro do segmento, e elimina as transições abruptas das extremidades (Hamming, 1998).

2.2 Codificação por Predição Linear

A Codificação por Predição Linear (*Linear Prediction Coding* - LPC) utiliza um conjunto de técnicas que visa obter uma aproximação da voz amostrada a partir de uma combinação linear entre amostras passadas do sinal de voz e valores presentes e passados de uma entrada hipotética de um sistema, cuja saída é o sinal de voz (Costa, 1994).

O trato vocal é excitado durante a produção de voz por uma série de pulsos quase periódicos produzidos pelas cordas vocais no caso dos sons sonoros, e, no caso dos sons não-sonoros, por turbulência passando através das constrições do trato (Atal and Hanauer, 1971). Dessa forma, o conjunto de parâmetros obtidos pela predição linear podem ser usados para representar o trato vocal (Rabiner and Schafer, 1978), dado que o sistema linear é excitado por pulsos quase periódicos (em se tratando de uma excitação sonora) ou por ruído aleatório (em se tratando de uma excitação não sonora) (Costa, 1994).

Outra característica importante da codificação por predição linear reside no fato desta combinar os efeitos da excitação glotal, do trato vocal e da radiação (Atal and Hanauer, 1971).

Os coeficientes LPC, que irão compor o vetor de características, podem ser obtidos a partir da Equação 1.

$$\tilde{s}(n) = \sum_{k=1}^{K} c_k s(n-k) \quad (1)$$

em que K é a quantidade de coeficientes, $\tilde{s}(n)$ são as amostras de voz aproximadas (estimadas), c_k são os coeficientes LPC e $s(n-k)$ são as amostras do sinal de voz original.

Dentre os métodos utilizados para resolução dessa equação, tem-se o método da covariância (Atal and Hanauer, 1971); o método da autocorrelação (Makhoul, 1975); a formulação do filtro inverso (Rabiner and Schafer, 1978); a formulação da estimação espectral (Rabiner and Schafer, 1978); a formulação da máxima verossimilhança (Rabiner and Schafer, 1978) e a formulação do produto interno (Myers and Aleksander, 1989). No trabalho apresentado é utilizado o método da autocorrelação.

3 Base de Dados

A base de dados utilizada foi adquirida do Massachussetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab (Kay-Elemetrics, 1994). Esta base de dados contém 1400 arquivos de voz da vogal sustentada /a/, obtidas a partir de, aproximadamente, 700 pessoas. Os arquivos foram obtidos com baixo nível de ruído, distância constante do microfone, tamanho da amostra de 16 bits e frequência de amostragem de 25 kHz ou 50 kHz.

Dentre os arquivos da base de dados, foram utilizados 67 arquivos de voz de indivíduos com patologia, sendo 49 mulheres e 18 homens, 44 com Edema de Reinke e 23 com outras patologias, como nódulos, cistos e paralisia nas dobras vocais, e 53 arquivos de vozes normais, sendo 21 do sexo masculino e 32 do sexo feminino.

4 Método Proposto

O método proposto consiste de 3 módulos fundamentais: pré-processamento do sinal; extração de características (coeficientes LPC) para representação dos sinais de voz; e treinamento/classificação utilizando Redes Neurais Multilayer Perceptron (MLP).

Para a implementação do método, foi utilizada a linguagem de programação C para os módulos de pré-processamento e extração de características. As Redes Neurais foram simuladas utilizando o *software* WEKA (*Waikato Environment for Knowledge Analysis*) (WEKA, 2009), devido à facilidade de utilização, já que o *software* oferece uma interface de fácil uso, e consistência na implementação do algoritmo da rede MLP.

A seguir, serão descritas as fases de treinamento e classificação do sistema. A fase de treinamento é composta por 3 etapas (Figura 1). Na etapa de pré-processamento foi aplicada a pré-ênfase nos sinais de voz, com fator de pré-ênfase igual a 0.95. Em seguida, os sinais foram segmentados usando uma janela de Hamming de 20ms com sobreposição de 50%. Na etapa de extração de características, foram obtidos 12 coeficientes LPC para cada janela para obter uma boa representação de todos os formantes presentes no sinal (Rabiner and Schafer, 1978).

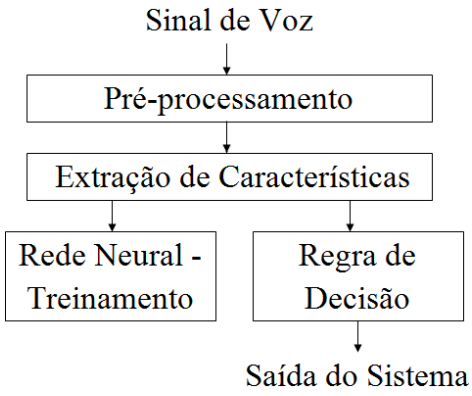

Figura 1: Etapas da fase de treinamento e classificação do sistema.

Na fase de treinamento da rede neural, foram investigadas 25 arquiteturas diferentes de Redes Neurais MLP, cada uma possuindo uma camada escondida. A camada de entrada de cada arquitetura possui 12 neurônios, um para cada coeficiente LPC, a camada escondida varia em cada arquitetura, tendo 2, 4, 6, 8, ..., 48, 50 neurônios, e a camada de saída de cada arquitetura possui 2 neurônios, para diferenciar entre voz normal e voz patológica.

Para cada arquitetura investigada, foram criadas 10 redes neurais com inicialização aleatória de pesos. Para cada rede, foram criados um conjunto de treinamento, com 50% dos coeficientes extraídos, um conjunto de validação para servir como uma das condições de parada, com 25% dos coeficientes, e um conjunto de testes, com os 25% restantes. Após o treinamento, foram realizados testes para definir qual a arquitetura de rede a ser utilizada na classificação entre voz normal e voz patológica. Para cada arquitetura da rede, foi obtida a média das taxas de acerto. A melhor arquitetura de rede encontrada foi utilizada na fase de classificação do sistema, que consiste das mesmas etapas de pré-processamento e extração de características da fase de treinamento, seguidas de uma regra de decisão (Figura 1).

5 Resultados e Discussão

Na Tabela 1 são apresentadas as médias das taxas de acerto para cada arquitetura da Rede Neural MLP investigada. A coluna n indica a quantidade de neurônios na camada escondida, as colunas *Normal(%)* e *Patológica(%)* indicam as médias de acer-

tos para voz normal e voz patológica e as colunas σ_N e σ_P os desvios padrão das duas colunas anteriores, respectivamente.

Tabela 1: Médias das taxas de acerto para cada arquitetura da Rede Neural MLP investigada.

n	Normal(%)	Patológica(%)	σ_N	σ_P
2	90,87	88,54	1,60	1,83
4	95,09	93,19	1,50	2,40
6	96,58	95,64	0,89	0,88
8	97,79	96,58	0,73	0,63
10	97,43	97,19	1,09	1,15
12	98,02	97,78	0,69	0,47
14	98,40	97,32	0,47	0,61
16	98,22	97,80	0,56	0,78
18	98,48	97,67	0,45	0,62
20	98,60	97,91	0,39	0,34
22	98,52	97,99	0,37	0,61
24	98,42	98,39	0,50	0,36
26	97,93	98,11	1,18	0,63
28	98,55	98,03	0,34	0,67
30	98,32	98,19	0,60	0,51
32	98,69	98,24	0,41	0,40
34	98,60	98,26	0,49	0,52
36	98,71	98,12	0,31	0,47
38	98,44	98,12	0,70	0,60
40	98,47	98,12	0,42	1,14
42	98,45	98,37	0,36	0,25
44	98,29	98,43	0,64	0,31
46	98,28	98,30	0,61	0,31
48	98,50	98,03	0,50	0,24
50	98,62	98,28	0,45	0,39

A partir de uma análise da Tabela 1, pode-se observar que as redes com poucos neurônios na camada escondida apresentaram mais dificuldade em separar as classes, com taxas de acerto mais baixas. Conforme a quantidade de neurônios na camada escondida foi aumentando, a média das taxas de acerto também aumentou (Figura 2). A partir de 10 neurônios na camada escondida já foi obtida uma ótima separação entre classes, com médias de taxas de acerto acima dos 97%, e, a partir de 28 neurônios, as médias se estabilizaram entre 98% e 99%. Devido a essa estabilidade, os experimentos foram finalizados para a arquitetura com 50 neurônios na camada escondida. Outro aspecto a ser observado na Tabela 1 são os baixos valores de desvio padrão, indicando que os treinamentos para as diversas arquiteturas proporcionaram resultados similares. Destaca-se também, que esses valores diminuíram com o aumento do número de neurônios, estabilizando-se em valores abaixo de 1 a partir de 28 neurônios. Nota-se também, que as taxas de acerto para voz normal foram superiores às taxas de acerto para voz patológica na maioria dos casos, o que indica que em certos momentos a voz patológica não difere muito da voz normal, o que pode ocorrer nos casos em que a patologia é menos grave ou está em estágio inicial, dificultando assim a classificação. A arquitetura que apresentou os melhores resultados apresenta 36 neurônios na camada escondida. Na Tabela 2 está apresentada a matriz de confusão dos resultados para a melhor arquitetura encontrada.

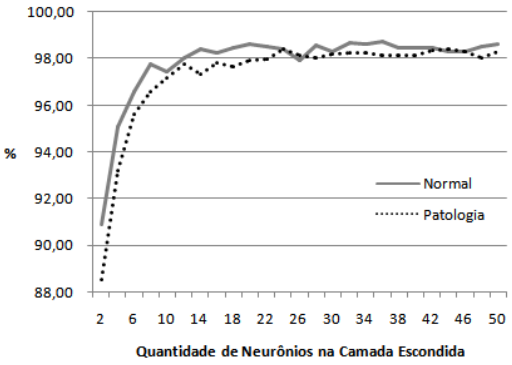

Figura 2: Média das taxas de acerto para cada Arquitetura da Rede Neural MLP investigada.

Tabela 2: Matriz de Confusão das taxas de acerto para a arquitetura com 36 neurônios na camada escondida.

	Normal(%)	Patologia(%)
Normal(%)	98,71	1,29
Patologia(%)	1,88	98,12

6 Conclusões e Sugestões para Trabalhos Futuros

A investigação de uso de coeficientes LPC para representar os sinais de voz associado a um processo de classificação utilizando Redes Neurais MLP demonstrou que esta abordagem consiste em um método promissor para a discriminação entre voz normal e voz patológica, atingindo uma taxa de acerto superior a 98%. Como sugestão para trabalhos futuros, espera-se aumentar o nível de especificidade na classificação, investigando a separação entre 3 classes: voz normal, voz com Edema de Reinke, que é a patologia de maior incidência na base de dados utilizada, e voz com outras patologias. Outro ponto a ser abordado será o uso de métodos de otimização (Yamazaki and Ludermir, 2003) para a obtenção da melhor arquitetura de rede.

Referências

Adnene, C. and Lamia, B. (2003). Analysis of pathological voices by speech processing, *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, pp. 365–367.

Aguiar-Neto, B. G., Costa, S. C. and Fechine, J. M. (2008). Lpc modeling and cepstral analysis applied to vocal fold pathology detection, *International Journal of Functional Informatics and Personalised Medicine Issue* **1**(2): 156–170.

Albernaz, P. L. M., Ganana, M. M., Fukuda, Y. and Munhoz, M. S. L. (1997). *Otorrinolaringologia para o clínico geral*, Byk.

Alonso, J. B., de Maria, F. D., Travieso, C. M. and Ferrer, M. A. (2005). Using nonlinear features for voice disorder detection, *International Conference on Non-linear Speech Processing*, pp. 94–106.

Alonso, J. B., Leon, J., Alonso, I. and Ferrer, M. A. (2001). Automatic detection of pathologies in the voice by hos based parameters, *EURASIP Journal on Applied Signal Processing* **4**: 275–284.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave, *The Journal of the Acourticol Society of America* pp. 637–655.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford.

Boyanov, B., Ivanov, T., Hadjitodorov, S. and Chollet, G. (1993). Robust hybrid pitch detector, *Electron. Lett.* **29**(22): 1924–1926.

Costa, H. O., Duprat, A., Eckley, C. and Silva, M. A. A. (2000). Caracterização do profissional da voz para o laringologista, *Revista Brasileira de Otorrinolaringologia* **66**(2): 129–134.

Costa, S. C. (2008). *Análise Acústica, baseadan no Modelo Linear de produção da fala, para discriminação de vozes patológicas*, PhD thesis, Universidade Federal de Campina Grande. Doutorado em Engenharia Eltrica.

Costa, W. C. A. (1994). *Reconhecimento de fala utilizando modelos de markov escondidos (hmm's) de densidades contínuas*, Master's thesis, Universidade Federal da Paraba.

Fukuda, Y. (2003). *Otorrinolaringologia*, Manole.

Godino-Llorente, J. I. and Gómez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, *IEEE Transactions on Biomedical Engineering* **51**(2): 380–384.

Godino-Llorente, J. I., Gómez-Vilda, P. and Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters, *IEEE Transactions on Biomedical Engineering* **53**(10): 1943–1953.

Gotaas, C. and Starr, C. D. (1993). Vocal fatigue among teachers, *Folia Fhoniatr* **45**: 120–129.

Hamming, R. W. (1998). *Digital Filters*, 3 edn, Courier Dover Publications.

Haykin, S. and Veen, B. V. (2002). *Signals and systems*, Wiley.

Kay-Elemetrics (1994). *Kay Elemetrics Corp. Disordered Voice Database, Model 4337*, 3 edn, Massachussetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab.

Kil, D. H. and Shin, F. B. (1996). *Pattern Recognition and Prediction with Applications to Signal Characterization*, AIP Press, American Institute of Physics.

Koufman, J. A. and Isaacson, E. G. (1991). *Voice Disorders*, Otol Clin NA.

Kukharchik, P., Martynov, D., Kheidorov, I. and Kotov, O. (2007). Vocal fold pathology detection using modified wavelet-like features and support vector machines, *Proceedings of 15th European Signal Processing Conference*.

Makhoul, J. (1975). Linear prediction: A tutorial review, *Proceedings of the IEEE*, pp. 561–580.

Manfredi, C. (2000). Adaptive noise energy estimation in pathological speech signals, *IEEE Trans. Biomedical Engineering* **47**(11): 1538–1543.

Manfredi, C., Pierazzi, L. and Bruscaglioni, P. (1999). Pitch estimation for noise retrieval in time and frequency domain, *Med. Biol. Eng. Comput.* **37**(2): 532–533.

Martinez, C. E. and Rufiner, H. L. (2000). Acoustic analysis of speech for detection of laryngeal pathologies, *Proceedings of the 22nd Annual EMBS International Conference*, pp. 23–28.

Myers, C. and Aleksander, I. (1989). Output functions for probabilistic logic nodes, *Proc. IEE International Conference on Artificial Neural Networks*, pp. 310–314.

Papamichalis, P. E. (1997). *Pratical Approaches to Speech Coding*, Prentice Hall.

Picone, J. W. (1993). Signal modeling techniques in speech recognition, *Proceedings of The IEEE*, pp. 1215–1247.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice Hall.

Rosa, M. O., Pereira, J. C. and Grellet, M. (2000). Adaptive estimation of residue signal for voice pathology diagnosis, *IEEE Trans. Biomedical Engineering* **47**(1): 96–104.

Sapir, S., Keidar, A. and Mathers-Schmidt, B. (1993). Vocal attrition inteachers: Survey findings, *Eur J Disord Commun* **28**: 177–185.

Sotomayor, C. A. M. (2003). *Realce de voz aplicado verificação automática de locutor*, Master's thesis, Instituto Militar de Engenharia.

Wallen, E. J. and Hansen, J. H. (1996). A screening test for speech pathology assessment using bjective quality measures, *ICSLP 96. Proc.* **2**: 776–779.

WEKA (2009). http://www.cs.waikato.ac.nz/ml/weka/.

Yamazaki, A. and Ludermir, T. B. (2003). Neural network training with global optimization techniques, *International Journal of Neural Systems* **13**(2): 77–86.