Reconhecimento de Voz Contínua com Atributos PNCC e Métodos de Robustez WD e MAP

Christian Dayan Arcos, Marco Antonio Grivet e Abraham Alcaim

***Resumo*— A degradação do sinal de voz devido a condições adversas gera baixas taxas de acerto nos sistemas de reconhecimento de voz. Os autores propõem a mistura de dois métodos: *pré-extração de atributos* para realce de fala e *pós-extração de atributos* para compensação de características. Segundo seu foco principal, esses métodos estão orientados fundamentalmente a minimizar os desajustes causados pela inserção de ruído no sinal de voz. Estes métodos serão aplicados antes e depois da extração de atributos, respectivamente, conseguindo assim estimar o máximo possível o sinal limpo a partir da sua versão degradada.**

***Palavras-Chave*— Sinal, ruído, realce, compensação, atributos.**

***Abstract*— The degradation of the speech signal due to adverse conditions generates low accuracy rates in speech recognition systems. The authors propose mixing two methods: *pre-extraction of attributes* for speech enhancement and *post-extraction of attributes* for features compensation. According to their main focus, they are fundamentally oriented to minimize the misfit caused by noise insertion in the speech signal. These methods will be applied before and after the extraction of attributes, respectively, therefore allowing the best possible estimation of the clear signal from its degraded version.**

***Keywords*— Signal, degradation, enhancement, compensation, attributes.**

I. INTRODUÇÃO

O reconhecimento de fala, nos últimos tempos, tem evoluído de tal forma que a maior parte dos sistemas automáticos de reconhecimento de fala humana são utilizados para acesso telefônico, sistemas de informação, ajuda a descapacitados, controle de sistemas por voz, etc.

No entanto um dos maiores problemas destes sistemas de reconhecimento é a degradação do sinal quando a voz é obtida em ambientes ruidosos. Devido ao fato de que o ruído altera as estatísticas do sinal, baixos rendimentos são usualmente causados pelo descasamento entre as características de treinamento e as de reconhecimento.

Os métodos de reconhecimento de voz robusta podem ser divididos em três categorias [4]:

- Técnicas de realce de fala,
- Técnicas de compensação de atributos,
- Técnicas de adaptação de modelos.

Christian Dayan Arcos, Marco Antonio Grivet e Abraham Alcaim¸ Centro de Estudos em Telecomunicações da PUC-Rio (CETUC), Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: dayan3846@gmail.com, mgrivet@cetuc.puc-rio.br, alcaim@pucrio.br. Este trabalho foi parcialmente financiado pela CAPES.

O objetivo deste trabalho é analisar e testar técnicas de redução de ruído relacionadas com as duas primeiras categorias, a saber: i) a técnica Wavelet Denoising (WD) [5] para realce de fala e ii) a técnica Mapeamento de Histogramas (MAP) [6] para compensação de atributos. Além disso, o sinal de voz é representado como uma sequência de vetores de caraterísticas que contêm informação espectral de curtos períodos de tempo, usando como atributos os coeficientes cepstrais de potência normalizada (do Inglês, Power-Normalized Cepstral Coeficients - PNCC). As técnicas citadas são combinadas em etapas separadas, como mostrado na Fig. 1, visando gerar reconhecedores de voz que melhorem as taxas de reconhecimento e sejam capazes de operar em situações reais.

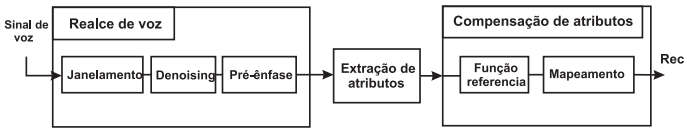

Fig. 1. Etapas do sistema proposto para reconhecimento de voz.

Este artigo estrutura-se da seguinte forma. A seção II inclui uma descrição dos atributos PNCC. As técnicas *pré-extração de atributos* e *pós-extração de atributos* são descritas nas seções III e IV, respectivamente. O procedimento experimental e a discussão dos resultados são apresentados na seção V. Finalmente, as principais conclusões são resumidas na seção VI.

II. ATRIBUTOS PNCC

Estudos fisiológicos ao longo do tempo têm mostrado que os tons não são adequadamente representados em escalas lineares [1] [2]. Por esta razão tenta-se aproximar o comportamento auditivo humano por meio de escalas de frequências não lineares, como a conhecida escala MEL presente nos Mel-Frequency Cepstral coefficients (MFCC) [3]. Estes coeficientes são uma representação definida como o cepstrum de um sinal janelado no tempo, que tem sido derivado da aplicação da discrete Fourier transform (DFT), em escalas de frequências não lineares as quais aproximam-se ao comportamento do sistema auditivo humano. Porém, sua eficácia do reconhecimento diminui rapidamente em presença do ruído.

Recentemente, [7] introduziu um método mais eficiente chamado Power-Normalized Cepstral Coefficients. Sua eficiência se deve à adição de uma nova etapa de remoção de ruído, na qual através da média das energias de uma banda ao longo de alguns quadros consecutivos, consegue-se remover

parte considerável do ruído do sinal. A implementação detalhada pode se ver em [8]

Este procedimento é feito após a divisão do sinal em bandas de frequência superpostas, similar as utilizadas nos MFCC, cuja diferença é o uso de um novo tipo de banco de filtros Gammatome baseado na escala de Bandas Retangulares Equivalentes (ERB) [9].

Estes filtros possuem bandas de largura não uniforme e sobrepostas, como mostrado na Fig. 2, onde cada filtro representa a resposta em frequência relacionada com um ponto particular da membrana basilar.

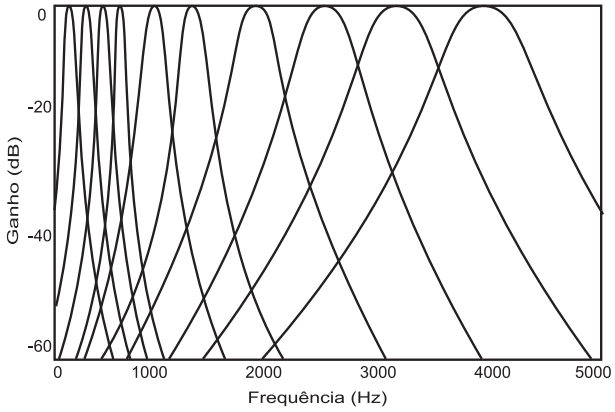

Fig. 2. Banco de filtros Gammatone.

A resposta ao impulso de cada filtro é dada por:

$$g_l t = t^{l-1} e^{-2\pi t 1{,}019\epsilon} \cos(2\pi f_c t) \qquad com \quad t \geq 0 \qquad (1)$$

onde l é a ordem do filtro e f_c a sua frequência central da banda. Assim, a largura de faixa de cada filtro é ajustada conforme as medidas da largura de ERB dos filtros auditivos humanos dados pela equação:

$$\epsilon(f_c) = 24,7 \left(4,37 \frac{f_c}{1000} + 1 \right) Hz \qquad (2)$$

Por último foi feita uma modificação na operação não-linear sobre a energia da banda. A função logarítmica dos MFCC apresenta uma grande inclinação para valores próximos de zero. Isso altera bastante os atributos MFCC quando se adiciona ruído a pequenos valores de energia. Por isso, foi escolhida a função de potenciação, que cresce mais suavemente. Esta função é descrita pela equação

$$y = x^{a0} \qquad (3)$$

onde a_0 é a constante de elevação da função de potenciação. A Fig. 3 ilustra o diagrama de blocos do sistema de parametrização PNCC e MFCC de cada quadro do sinal.

Desta maneira, os atributos PNCC são considerados uma evolução dos atributos MFCC.

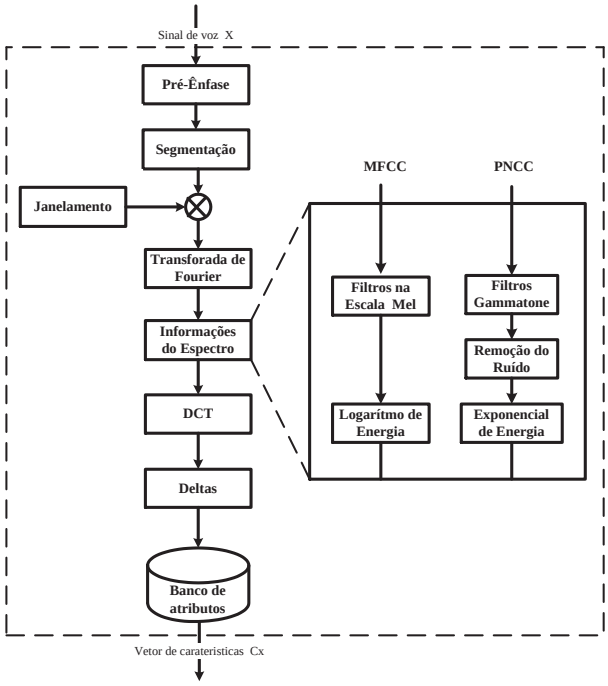

Fig. 3. Comparação dos métodos de parametrização.

III. Wavelet Denoising como técnica de Realce de Fala

Esta técnica tem como objetivo eliminar o ruído antes da parametrização do sinal, através de um processo que tenta estimar o sinal de voz original a partir de sua versão contaminada ou degradada.

Segundo [10] o sinal de voz contaminado $y(n)$ é modelado pela soma do sinal de voz $x(n)$ e o sinal de ruído $r(n)$, onde este último é assumido como um processo aleatório estacionário em intervalos curtos de tempo e estatisticamente decorrelatado do sinal de fala.

$$y(n) = x(n) + r(n) \qquad (4)$$

A técnica Wavelet Denoising tem por objetivo transformar $y(n)$ em um sinal $x'(n)$, o mais similar possível do sinal $x(n)$. Este processo de redução de ruído é conhecido como *denoising* e é baseado em transformadas Wavelet. É composto de três etapas básicas que são ilustradas no diagrama de blocos da Fig 4.

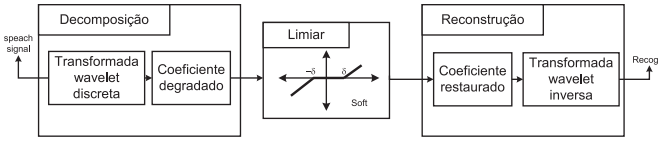

Fig. 4. Diagrama de blocos da técnica wavelet denoising.

A transformada Wavelet discreta é usada no bloco de decomposição que recebe os dados do sinal e a transformada wavelet discreta inversa é calculada sobre os coeficientes

Wavelet alterados. Estas duas transformadas são apresentadas nas equações (5) e (6), respectivamente.

A transformada Wavelet pode ser considerada como uma evolução da transformada de Fourier, que é usada para decompor e filtrar o sinal de voz de entrada, a fim de remover a redundância de coeficientes entre escalas.

$$W(a,b,n) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} g(n)\psi_{a,b}(n) \tag{5}$$

$$W^{-1}(a,b,n) = \frac{1}{\sqrt{N}} \sum_{a} \sum_{b} W(a,b,n)\psi_{a,b}(n) \tag{6}$$

onde $W(a,b,n)$ são os coeficientes da transformada Wavelet, $\psi_{a,b}(n)$ é a família de funções Wavelet com escala a e posição b e $W^{-1}(a,b,n)$ é a transformada inversa Wavelet.

O objetivo principal do bloco central representado na Fig.4 é a redução de ruído através da técnica soft-thresholding equação (7) apresentada em [11] onde através do limiar δ equação (8) determina-se os valores W da transformada que são ou não significativos perante os demais, tornando assim uma transição suave [8] e resultando em melhores taxas de acerto no reconhecimento.

$$W_s = \begin{cases} \text{sgn}(W)(\mid W \mid -\delta) & \mid W \mid \geq \delta \\ 0, & \text{caso contrário.} \end{cases} \tag{7}$$

$$\delta = s\sqrt{2\ln(L)} \tag{8}$$

onde $sgn(W)$ é o sinal de $W(+1,\ 0 \text{ ou } -1)$, L é o tamanho do sinal de entrada e s é a estimativa da intensidade do ruído dada por

$$s = \frac{median(|W_{a,b}|)}{0.6745} \tag{9}$$

Assim s determinara se um valor da transformada é significativo o não perante os demais.

IV. Técnica de Compensação de Atributos Usando Mapeamento de Histogramas

Esta técnica age sobre as características parametrizadas e tem por objetivo reduzir as variações entre o sinal de áudio utilizado para treinar os modelos e as condições reais do sinal de áudio de avaliação. Ela é chamada de técnica de pós-extração de atributos e é projetada para compensar os efeitos não lineares causados pelo ruído.

O mapeamento de histogramas é uma técnica similar à utilizada para processamento de imagens, usualmente conhecida como Equalização de Histogramas [13]. Ela tem sido aplicada para compensar os efeitos não lineares causados pelo ruído sobre a representação da voz. Após se obter o vetor de atributos, cada um dos seus componentes é equalizado de forma independente, através de uma transformação $z(k)$ aplicada a cada um destes componentes. Busca-se transformar a distribuição de probabilidade da voz contaminada $f(.)$ em uma distribuição de probabilidade de referência $h(.)$, como mostrado na Fig. 5, atenuando assim os efeitos do ruído presente na função de distribuição original. Obviamente o mapeamento $z(k)$ do valor cepstral n ao valor mapeado m deve conservar a principal propriedade de distribuição estatística, ou seja,

$$\int_{k=-\infty}^{n} f(k)\,dk = \int_{z=-\infty}^{m} h(z)dz \tag{10}$$

A função distribuição de probabilidade de referência escolhida é a distribuição de Gauss com média zero e variância unitária, já que segundo [14] esta distribuição é frequentemente utilizada em Modelos Ocultos de Markov (do Inglês, Hidden Markov Models - HMMs) na forma de mistura de Gaussianas.

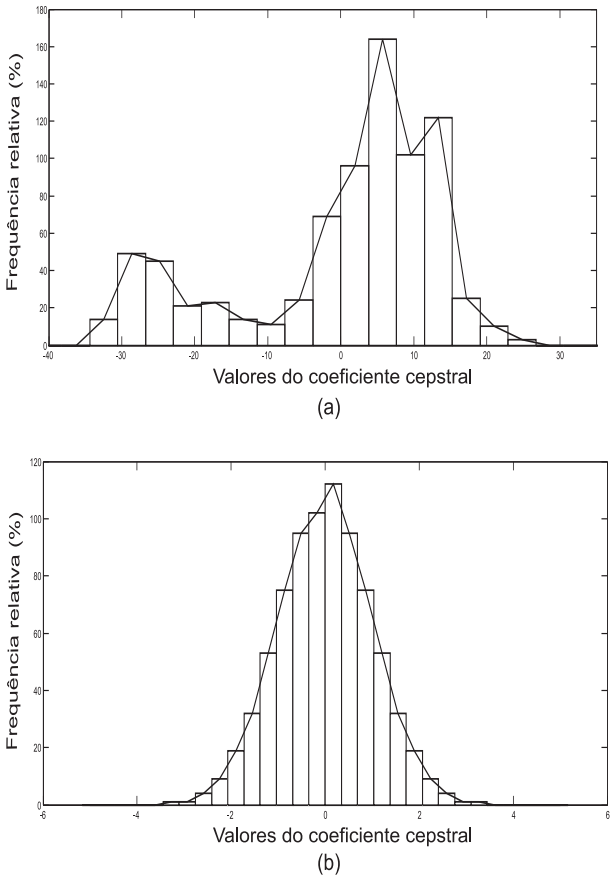

Fig. 5. Característica do mapeamento de histogramas. (a) Histograma dos atributos cepstrais originais, (b) Histograma dos atributos cepstrais mapeados.

V. Resultados Experimentais

A robustez dos métodos propostos foi avaliada através de experimentos de reconhecimento sob condições de ruído usando dois bancos de dados: o TIMIT, o qual possui um total de 6300 sentenças pronunciadas por 630 pessoas das quais 70% são homens e 30% mulheres, onde cada um pronuncia 10 frases abrangendo os diversos sotaques do inglês americano para ambos os sexos. Para este trabalho foram utilizadas em total 4620 sentenças para treinar e para criar o modelo de linguagem e 1000 sentenças para teste.

Para obter as amostras corrompidas de teste, basta tomar as amostras limpas de voz e adicionar um sinal de ruído sobre ela.

$$r(t) = \left(\frac{1}{\sqrt{E_{ra(t)}}} \sqrt{\frac{E_s}{10^{\frac{SNR}{10}}}} \right) ra(t) \qquad (11)$$

onde $E_{ra(t)}$ e a energia média do sinal de ruído original, E_s Energia média do sinal limpo e $ra(t)$ é o sinal de ruído original

Dependendo da intensidade ruído SNR, a voz será corrompida com uma intensidade maior ou menor. Este trabalho selecionou sinais de ruído a partir de uma segunda base de dados: A NOISEX-92, que contém arquivos de som de diversas naturezas, um deles é o ruído branco. Trechos aleatórios desse sinais foram adicionados às amostras de teste, com razões sinal-ruído de 15 e 10dB.

Os sinais corrompidos e os limpos foram usados para obter os resultados.

O sistema de reconhecimento foi implementado com a ferramenta HMM Toolkit (HTK)[15]. Modelos Ocultos de Markov (HMMs) com mistura de 8 gaussianas foram gerados a fim de representar os trifones do idioma inglês. A partir de todas as frases listadas no banco TIMIT, estimou-se um modelo de linguagem de trigramas.

A representação da voz é baseada em parametrização PNCC onde o sinal de voz foi amostrado a 8 kHz e segmentado em quadros que são representados por um vector de atributos com os seguintes parâmetros:

- PNCC: B = 40 bandas com filtros de ordem 1, expoente $a_0 = 1/15$ e apenas os 20 primeiros valores da Discrete Cosine Transform (DCT) foram considerados.

Finalmente, os coeficientes delta dos atributos foram incluídos, dobrando a quantidade de valores por quadro.

O desempenho do reconhecedor foi avaliado pela proporção de palavras corretas nas frases de teste, calculado como:

$$R(\%) = 100\frac{N-(S+D+I)}{N} \qquad (12)$$

onde *N* é o número de palavras usadas no teste, *S* é o número de palavras substituídas, *D* é o número de palavras deletadas e *I* é o número de palavras inseridas.

Quatro tipos de testes foram realizados a fim de avaliar o sistema de reconhecimento proposto. No primeiro caso, um sistema de referência (*Ref.*) baseado unicamente em atributos PNCC é avaliado em condições limpas e posteriormente corrompidas com ruído branco a 15 dB e 10 dB. Um segundo e terceiro teste foi realizado visando obter um sistema mais robusto ao efeito de adição de ruído. Para isso foram usadas as mesmas condições do sistema (*Ref*) mas aplicando o método *wavelet denoising (WD)* e *mapeamento de histogramas (MAP)* antes e depois de parametrização respectivamente. Os testes foram feitos da seguinte maneira

- *Wavelet Denoising* (WD) + *Ref*
- *Ref* + *Mapeamento de Histogramas* (MAP)

os resultados são apresentados nas Tabelas II. e I respectivamente.

TABELA I
TAXAS DE ACERTOS UTILIZANDO WD + PNCC

SNR [dB]	Ref	Ref + WD
limpo	86.58%	70.19%
15	80.89%	81.80%
10	74.52%	75.65%

TABELA II
TAXA DE ACERTO UTILIZANDO PNCC + MAP

SNR [dB]	Ref	Ref + MAP
limpo	86.58%	89.19%
15	80.89%	83.50%
10	74.52%	78.73%

comparando as Tabelas II e I, pode-se ver o efeito positivo dos métodos propostos apresentando taxas de acerto ligeiramente maiores em todos os cenários inclusive no cenário limpo, para o caso do MAP.

Finalmente, foi realizada uma mistura dos métodos de robustez aqui propostos, com o objetivo de constatar se há uma melhora no resultado de cada cenário, a mistura foi feita da seguinte maneira

- *Wavelet Denoising* (WD) + *Ref* + *Mapeamento de Histogramas* (MAP)

E seu desempenho é avaliado na Fig. 6

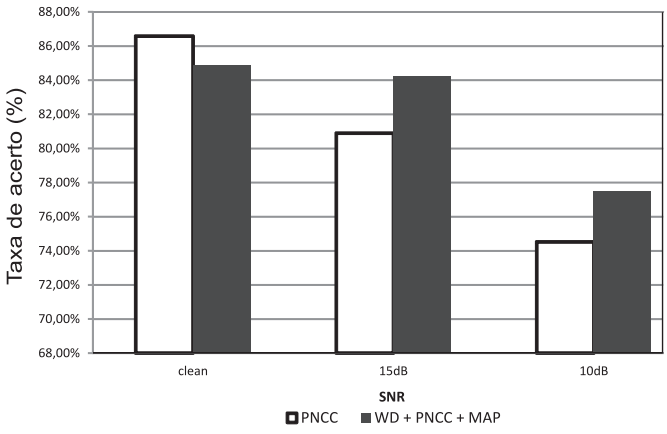

Fig. 6. Taxa de acerto do reconhecedor.

A Fig 6 apresenta os resultados para a mistura dos métodos usando as mesmas condições dos testes anteriores. Pode-se ver que os resultados correspondentes às condições ideais foram prejudicados devido ao fato de que quando um método de eliminação de ruído é aplicado ao sinal limpo, ele remove parte da informação do sinal, provocando assim perda de inteligibilidade do discurso. Em contraste, nos casos de ambiente altamente ruidoso, existe uma melhora significativa porque o método de mapeamento de histogramas promove uma

estimação mais eficiente das transformações não lineares que compensam grande parte dos efeitos do ruído. Cabe ressaltar, que não existe nenhuma referência na literatura técnica, onde os dois métodos aqui discutidos sejam conjuntamente empregados e seu desempenho avaliado.

VI. Conclusões

Este artigo discute duas formas diferentes de melhorar o desempenho do reconhecimento de voz contínua na presença de ruído aditivo através de métodos de compensação e realce de voz. O Wavelet Denoising no domínio do espectro de magnitude é implementado e as suas vantagens são ilustradas. Uma vez que o método Wavelet combina as informações dos domínios do tempo e frequência, ele fornece uma informação mais completa e detalhada do sinal. Esta abordagem em conjunto com o método de Mapeamento de Histogramas no campo dos coeficientes cepstrais promove a melhora das taxas de reconhecimento. Isto se deve à capacidade do Mapeamento de Histogramas para compensar as distorções não lineares restantes após da utilização de Wavelet Denoising. Isto pode ser observado nos resultados apresentados no artigo que revelam ser uma mistura eficaz e de um baixo custo computacional. Quando a SNR(razão sinal-ruído) é de 15 e 10 dB, a taxa de reconhecimento cai (relativamente ao uso de sinal limpo), respectivamente, de 80.89% e 74.52% quando se usa apenas os PNCC e de 84.19% e 77.47% quando se usa a combinação *WD + PNCC + MAP*.

Agradecimentos

Os autores agradecem à CAPES.

Referências

[1] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing melfrequency cepstral coefficients on the power spectrum," *IEEE International Acoustics, Speech, and Signal Processing*, pp. 73?-76, 2001.

[2] F. Miyara, "Introducción a la Psicoacústica," pp. 20–22, 1999.

[3] Ch. Ittichaichareon, S. Suksri, T. Yingthawornsuk, " Speech Recognition using MFCC," *International Conference on Computer Graphics, Simulation and Modeling (ICGSM?2012)*, pp. 135–138, 2012.

[4] J. Bellegarda, "Statistical techniques for robust ASR: review and perspectives," *Proc. of EuroSpeech*, pp. KN 33–36, 1997.

[5] O. Farooq, and S. Datta, "Wavelet-based denoising for robust feature extraction for speech recognition," *Electronics Letters*, pp. 163–165, 2003.

[6] A. De La Torre, J.C. Segura, C. Benitez, A.M. Peinado, and A.J Rubio, "Non-linear transformations of the feature space for robust speech recognition," *Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 401–404, 2002.

[7] C. Kim, and R. Steam, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," *Acoustics, Speech, and Signal Processing (ICASSP)*, pp 28–31, 2010.

[8] K. Finnian, and N. Harte. "A comparison of auditory features for robust speech recognition." *EUSIPCO*, pp 1968-1972, 2010.

[9] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory physiology and perception, Oxford: Pergamon Press*, pp. 429–446, 1992.

[10] J.C. Segura, A. De La Torre, M.C. Benitez, and A.M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and task," *Proc. of Eurospeech*, pp. 221-224, 2001.

[11] C.A. Medina, A. Alcaim, and J.A. Apolinario Jr, "Wavelet denoising of speech using neural networks for threshold selection," *Electronics Letters*, pp. 1869–1871, 2003.

[12] D.L. Donoho,"Denoising by soft-thresholding," *Information Theory, IEEE Transactions on*, pp. 613–627, 1995.

[13] A. Alcaim, and C.A. Oliveira, "Fundamentos do Processamento de Sinais de Voz e Imagem," *Editora Interciência*, pp. 226–228, 2011.

[14] G. Saon, S. Dharanipragada, and D. Povey, "Feature space gaussianization," *Acoustics, Speech, and Signal Processing (ICASSP)*,pp. 329–332, 2004.

[15] S. Young, et al. "The HTK Book: HTK Tools and Reference Manuals," *Entropic* , 1999