Reconhecimento Robusto de Locutor Baseado nos Atributos ZCPAC

Dirceu G. da Silva, Carlos D. R. Cuadros e Abraham Alcaim

***Resumo*— Neste artigo é proposta a utilização das características baseadas no ZCPAC (*Zero Crossing with Peak Amplitude Cepstrum*) na tarefa de reconhecimento robusto de locutor dependente do texto. É mostrado que o ZCPAC supera o Mel-Cepstrum em todos os testes realizados com sinais degradados com ruído branco e colorido com diferentes valores de relação sinal-ruído (SNR).**

***Palavras-Chave*— Reconhecimento Robusto de Locutor, ZCPA, Ruído Aditivo, HMM.**

***Abstract*— In this paper we propose some improvements on ZCPA-based features extraction algorithm, in order to use them for noise-robust automatic speaker identification. We show that a properly chosen set of parameters as filterbank, longtime window lengths and channel-wise up-sampling factors can lead to increasingly better results than traditional MFCC in presence of AWGN, factory, F16 and babble noises at decreasing SNR's.**

***Keywords*— Robust Speaker Recognition, ZCPA, Additive Noise, HMM.**

I. Introdução

A tarefa de reconhecimento automático de locutor está bem desenvolvida nos dias atuais, há muitas técnicas de extração de características, de modelagem de locutor e de métodos de avaliação. Estas técnicas estão bem documentadas em vários tutoriais [1], [2], [3], [4]. Todavia o desempenho do reconhecimento de locutor sofre degradação em ambiente ruidoso inviabilizando o seu uso muitas das vezes. Ainda são necessários estudos para aumentar a robustez nos sistemas de reconhecimento, pois ela é fundamental em aplicações reais.

A robustez do reconhecimento pode ser obtida nos diferentes estágios de processamento de um sistema de reconhecimento de locutor. Ela pode pode ser obtida no pré–processamento, na extração de características, no sistema classificador ou na medida de similaridade. Na fase do pré–processamento, a referida robustez advém do uso de técnicas de realce da voz por meio de subtração espectral, canceladores adaptativos de ruído [5] e *beamforming* [6].

Na fase da extração de características, pode-se encontrar o uso de CMN (*cepstral mean normalization*) e transformadas utilizadas para compensar os efeitos do canal e do ruído aditivo [7]. No sistema classificador, a compensação pode ser realizada mediante o uso de modelos matemáticos que integram as características estatísticas da voz e do ruído [8]. Já para a medida de similaridade, existem compensações usando uma adequada combinação de várias medidas de similaridade [9]. O presente artigo foca o uso da segunda classe, ou seja, o estudo de uma característica robusta para o reconhecimento de locutor.

Dirceu G. da Silva e Abraham Alcaim: CETUC/PUC-Rio, Rua Marquês de São Vicente, 225, Rio de Janeiro, RJ, 22.453-900 (e-mail: alcaim,dirceu@cetuc.puc-rio.br).

Carlos D. R. Cuadros: Departamento de Engenharia de Telecomunicações, Universidade Federal Fluminense Escola de Engenharia, R. Passo da Pátria, 156 - São Domingos, Niterói, RJ, 24210-240, RJ - Brasil (Email: carlosdfresh@hotmail.com)

O modelo auditivo baseado no *Ensemble Interval Histogram* (EIH), proposto por Ghitza em [10] para reconhecimento automático de voz (RAV) em ruído, mostrou um desempenho melhor que o Mel-Cepstrum (*Mel- Frequency Cepstral Coefficients*-MFCC) para baixas SNR, desde que os valores dos níveis do EIH sejam escolhidos criteriosamente.

O EIH é composto de uma banco de filtros cocleares e um conjunto de detetores de cruzamento de nível na saída de cada um desses filtros. O banco de filtros modela a seletividade em freqüência ao longo da membrana basilar na cóclea. Seguido o banco de filtros há um arranjo de cinco detectores de níveis de cruzamento de zeros cuja função é simular a atividade das células pilosas internas. Os cruzamentos positivos de zeros detetados simulam os impulsos nervosos. Os níveis do arranjo de cruzamento de zeros podem ser interpretados como a resposta de um conjunto de fibras nervosas pertencentes a diferentes células pilosas e são distribuídos através de uma faixa de valores positivos, considerando a natureza retificadora das células pilosas.

No modelo do EIH a quantidade de atividade nervosa gerada por um dado estímulo acústico é medida através da densidade de probabilidade de intervalo curto dos níveis de cruzamento de zeros do detetor. A estimação dessa densidade para um nível específico é obtida através do cálculo do histograma do número de cruzamentos de cada nível do detetor em relação aos intervalos de tempo entre eles. São considerados, apenas, os intervalos entre dois cruzamentos positivos de zero. Como a representação do sistema auditivo é realizada no domínio da freqüência, é calculado o histograma do inverso dos intervalos.

Infelizmente, este método é severamente influenciando pela escolha dos valores dos níveis. Além disso, não há nenhum método disponível para se escolher facilmente estes valores. Por outro lado, foi mostrado em [12] que a estimação de freqüência baseada nos níveis mais baixos são menos suscetíveis a ruído quando comparados aos níveis mais altos. Como consequência, Kim et. al. [12] propuseram a extração de características baseadas no ZCPA (Zero Crossings With Peak Amplitude) para a tarefa de reconhecimento de voz. Os experimentos realizados por Kim foram comparáveis com os obtidos pelo EIH e também mostraram-se melhores que o MFCC em ruído, com a vantagem da redução da complexidade computacional.

Neste artigo nós descrevemos brevemente a extração dos ZCPA Cepstrum (ZCPAC) e propomos a sua utilização na tarefa de reconhecimento automático de locutor dependente

do texto. Os passos para a extração do ZCPAC são dados na Seção II e uma descrição dos principais parâmetros do ZCPAC na Seção III. Resultados experimentais e comentários são apresentados na Seção IV. O ZCPAC é avaliado para diferentes relações de sinal-ruído, para um sistema de reconhecimento de locutor usando sinais de voz degradados pelos ruídos: gaussiano branco, fábrica, F16 e babble extraídos da base NOISEX. Esses resultados são comparados com os obtidos pelo MFCC nas mesmas condições de degradação. Por fim, na Seção V são apresentadas as conclusões do artigo.

II. Extração do ZCPAC

Em [10] foi mostrado através de testes para reconhecimento de voz, que as informações de intensidade fornecidas pelos níveis do EIH são importantes para o desempenho do sistema de reconhecimento, desde que os níveis sejam bem escolhidos. Motivado por este resultado, em [12] foi proposto uma modificação do EIH, mantendo um único nível para o detetor de cruzamento por zeros, enquanto a informação de intensidade foi preservada medindo-se a amplitude de pico entre sucessivos cruzamentos pelo zero. Por isso o nome *Zero-Crossing with Peak Amplitude*.

O procedimento para o cálculo do ZCPAC é mostrado na Figura 1. O sinal de voz de entrada, $s(n)$ é filtrado por um banco de K filtros perceptuais, gerando sinais $s_k(n)$. Cada um desses sinais é processado pelo detetor de cruzamentos pelo zero a fim de se determinar os instantes de cruzamentos ascendentes. Depois disso, cada par de sucessivos cruzamentos por zero, $z_k(i)$ e $z_k(i+1)$, o valor de pico $p_k(i)$ e o inverso do intervalo dos cruzamentos sucessivos $f_k(i)$ são calculados da seguinte forma:

$$p_k(i) = \max_{z_k(i) \leq n < z_k(i+1)} [s_k(n)] \quad (1)$$

$$f_k(i) = \frac{1}{z_k(i+1) - z_k(i)} \quad (2)$$

Em seguida, o eixo de freqüência é dividido pelo número de bins do histograma, R_j, onde j representa o índice de cada bin do histograma e R define a região de freqüência de cada bin. O histograma é construído com os valores de $f_k(i)$ levando-se em consideração todas as sub-bandas. Todavia ao invés de fazer a contagem dos bins por um, a contagem é feita tomando-se o logaritmo da amplitude de pico do sinal no intervalo dos cruzamentos. Então a contagem do j-ésimo bin do histograma é dado por:

$$bin(j) = \sum_{k=1}^{K} \sum_{i=1}^{N_{zc}} \psi[p_k(i)] \quad (3)$$

na qual $\psi[x] = \log(1+x)$.

Finalmente, a fim de reduzir a correlação dos dados é aplicada a DCT gerando o ZCPA Cepstrum (ZCPAC).

Do ponto de vista de processamento de sinais, o histograma do ZCPA pode ser visto como uma representação alternativa do espectro de voz. Isto está baseado no princípio da freqüência dominante [14] que estabelece que se há uma freqüência significativamente dominante no sinal, então o inverso do intervalo de cruzamento pelos zeros tende a tomar valores na vizinhança desta freqüência. Assim, o inverso do intervalo de cruzamento pelos zeros da k-ésima sub-banda pode ser visto como uma estimativa da freqüência dominante da sub-banda. Além disso, o pico do sinal entre cruzamentos sucessivos pode ser visto como uma medida da potência instantânea na sub-banda. Em suma, a construção do histograma do ZCPA consiste em atribuir a cada *bin* de freqüência, uma estimativa da potência da sub-banda correspondendo à freqüência dominante da sub-banda.

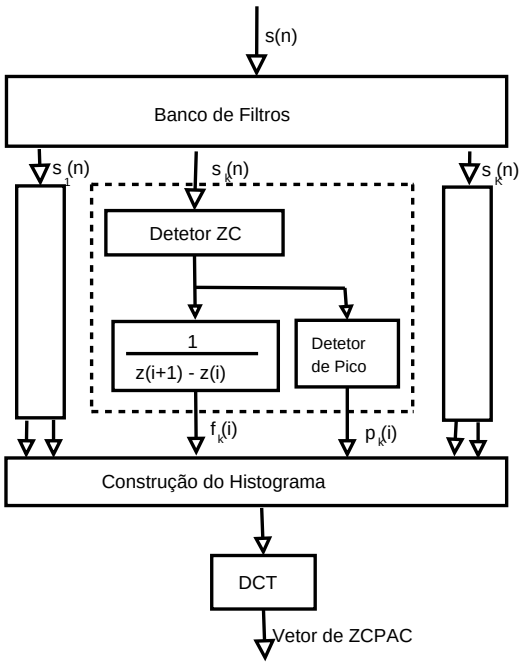

Fig. 1. Ilustração da Extração do ZCPA Cepstrum.

III. Principais Parâmetros para Extração do ZCPAC

Os principais parâmetros envolvidos na extração do ZCPAC são: banco de filtros, que envolve o tipo de filtro, o número de filtros e a largura de banda; e os parâmetros do histograma.

A. Banco de Filtros

Tipo de filtros: Inicialmente, foi utilizado o modelo de banco de filtros "cocleares" projetado por Lyon and Mead, o qual representa a propagação da onda ao longo da cóclea [10]. Todavia, resultados experimentais em [12] mostram que o uso de filtros FIR (*Finite Impulse Response*) implementados por janelas de Hamming podem ser mais eficientes e aumentam a taxa de reconhecimento, apesar do tipo do ruído e da SNR.

Número de bandas (canais): O número de canais (ou bandas) indicados para uso no ZCPA em [12] é de $K = 16$ até $K = 23$. Para sinais obtidos de sistema telefônico podem ser consideradas as 16 bandas de fc_2 ($\approx$ 150 Hz) a fc_{17} ($\approx$ 3400 Hz), na escala Bark. Esta escolha é conveniente por duas razões: ela atende a faixa de freqüência do sistema telefônico e é uma potência de 2, o que pode ser conveniente

para implementação. As bandas são dispostas segundo a escala Bark dada pela equação:

$$f_{Bark} = 13\ atan\left(\frac{0,76f}{1000}\right) + 3,5\ atan\left(\frac{f}{7500}\right)^2. \quad (4)$$

onde f é a freqüência em Hertz, e f_{Bark} e a freqüência perceptual em Bark correspondente.

Largura de Banda: Resultados experimentais em RAV [13] têm mostrado a conveniência de fixar a largura de banda bw de cada um dos K canais, em cerca de 2 ou 3 vezes a banda crítica perceptual $BW_{critica}(fc_k)$ dada pela equação [15]:

$$BW_{critica}(f) = 25 + 75\left[1 + 1,4\left(\frac{f}{1000}\right)^2\right]^{0,69}. \quad (5)$$

onde f é dado em Hz.

B. Parâmetros do Histograma

Há dois fatores que afetam as propriedades do histograma: a alocação de raias(*bins*) na faixa de freqüência e a escolha do tamanho da janela em que serão realizados a deteção dos cruzamentos pelo zero e os cálculos para construção do espectro.

1) Alocação de raias: A alocação das raias de freqüência é feita de acordo com a escala Bark, conforme a Eq. (4). A largura, R, de cada raia é dada pela Eq. (5), ou seja, à medida que a freqüência aumenta a largura de R também aumenta, levando o histograma a uma polarização nas altas freqüências.

2) Definição da janela de observação: Através de várias medições realizadas em [10], foi mostrado que o ouvido humano responde com uma alta resolução em freqüência e pobre resolução no tempo para baixas freqüências e vice-versa para as altas freqüências (filtros de Q–constante). Isso pode implementado com o uso de janelas temporais de tamanhos distintos.

Além disso, a fim de que o sinal filtrado em cada canal $s_k(n)$ tenha o mesmo número de períodos, o comprimento da janela para o k-ésimo canal deve ser idealmente $L_k = Np/fc_k$, onde Np é o número de períodos desejado [12]. Isto leva em consideração que o sinal é senoidal com freqüência igual à freqüência central de cada canal. Desse modo, o comprimento da janela torna-se longo para as baixas freqüências e curto para as altas. O que resulta numa alta resolução em freqüência e baixa resolução no tempo para as baixas freqüências e vice-versa para as altas freqüências, conforme já comentado. Considerando $Np = 20$, note que $L_k = 20/fc_k$ leva a comprimentos de janelas muito diferentes das bandas mais altas para as mais baixas. Para ilustrar esse problema, consideremos o caso de $K = 16$, $fc_2 = 150$ Hz e $fc_{17} = 3400$ Hz, de acordo com a freqüência central bark obtida por (4). Neste caso, teremos $L_2 = 133$ ms e $L_{17} = 6$ ms. Obviamente, $L_{17} = 6$ ms é muito curto, uma vez que este valor é menor que um período de pitch e muito menor que $L_2 = 133$ ms. A fim de reduzir esse problema, [13] utilizou uma nova expressão para o cálculo do comprimento das janelas para o RAV:

$$L_k = \frac{Np}{\sqrt{fc_k/1000}}\ \ milisegundos. \quad (6)$$

Para o mesmo exemplo, (6) resulta em $L_{17} \approx 11$ ms e $L_2 \approx 52$ ms, produzindo uma excursão muito mais aceitável. A utilização de janelas de comprimento elevado, parece ser uma característica intrínsica da estimação de freqüência no domínio do tempo, e isto está relacionado à necessidade de múltiplos pares de ZC positivos (ascendentes) para uma boa precisão na estimação (boa resolução em freqüência).

Outra observação importante diz respeito à relação entre o período central $t_k = 1/fc_k$ e a freqüência de amostragem f_s. A fim se ter uma boa precisão na estimação dos cruzamentos pelos zeros e em conseqüência uma boa estimação da freqüência, utilizam-se, normalmente, interpoladores na saídas dos bancos de filtros.

IV. Estudo Experimental

Com a finalidade de avaliar o desempenho do ZCPAC para a tarefa de reconhecimento de locutor, foi utilizada uma base de voz dependente do texto contendo 25 locutores (17 homens e 8 mulheres). Cada locutor falou 2 sentenças: E1 - *O prazo está terminando*, a qual é predominantemente composta por fonemas orais e E2 - *Amanhã ligo de novo*, onde a predominância é por fonemas nasais. Cada locutor repetiu 60 vezes cada uma das 2 frases. Trinta deles foram usadas para treinamento e o restante para teste de reconhecimento. O sistema de classificação utilizado foi um HMM (*Hidden Marlov Models*) com modelo esquerda-direita, com 10 estados e 1 gaussiana por estado. O HTK [18] foi utilizado para treinamento e teste.

As características foram extraídas a cada 10 ms. Foram extraídos os coeficientes Cepstrum do ZCPA (sem o coeficiente C_0) e suas derivadas Δ e Δ-Δ. Para obter esses coeficientes foram empregados 17 filtros FIR de ordem 61 obtidos da janela de Hamming, uniformemente espaçados na escala Bark, com largura de banda igual a 2 barks. O comprimento da janela da k-ésima sub-banda (dado em ms) foi calculado usando a equação (6) onde foi utilizado $Np = 30$, resultando em janelas de comprimento entre 16 e 77 ms. O histograma foi composto de 100 bins. Como as larguras das bandas críticas nas altas freqüências são superiores às larguras nas freqüências mais baixas, ocorre uma polarização à medida que a freqüência aumenta. Para compensar este efeito, foi feita uma normalização com respeito a freqüência no histograma. Para comparação, foi considerado o uso do MFCC (sem o coeficiente C_0) e suas derivadas Δ e Δ-Δ. Para a extração do Mel-Cepstrum, foram utilizados 22 filtros triangulares. Como o desempenho dos sistemas de reconhecimento dependem da janela de tempo das derivadas [4], foram testadas 4 janelas de tempo diferentes: 2, 5, 8 e 11 quadros. Foi escolhida a janela de tempo de 8 quadros por apresentar o melhor desempenho.

Com a finalidade de avaliar a robustez do ZCPAC em ambientes ruidosos, foram somados aos sinais de teste 4 tipos de ruído: gaussiano branco, fábrica, F16 e babble, todos extraídos da base NOISEX. Foram considerados também diferentes valores de SNR. A relação sinal-ruído foi calculada pela razão entre o quadro de maior potência e a potência média do ruído para um dado arquivo de voz.

As Tabelas I e II apresentam os resultados obtidos com as características estáticas e dinâmicas, respectivamente, para a

TABELA I

TAXA DE RECONHECIMENTO EM % USANDO AS CARACTERÍSTICAS MFCC E ZCPAC ESTÁTICAS PARA A FRASE E1 EM RUÍDO BRANCO

	MFCC				
Nr Coef \ SNR (dB)	Limpo	20	15	10	5
12 coef	100,00	38,40	20,00	7,87	0,93
15 coef	99,87	53,07	33,60	13,87	4,27
18 coef	99,47	59,60	39,33	29,47	18,27
20 coef	99,47	67,20	40,40	17,07	11,07
	ZCPAC				
12 coef	97,73	96,80	87,07	51,47	12,27
15 coef	98,53	97,47	92,40	60,40	14,27
18 coef	98,53	97,73	92,27	64,53	18,13
20 coef	98,13	96,53	89,47	62,93	16,13

TABELA II

TAXA DE RECONHECIMENTO EM % USANDO CARACTERÍSTICAS MFCC E ZCPAC ESTÁTICAS E DINÂMICAS (Δ E Δ-Δ, DELTA-WINDOW $= 8$) PARA A FASE E1 EM RUÍDO BRANCO

	MFCC				
Nr Coef \ SNR (dB)	Limpo	20	15	10	5
12 coef	99,60	65,73	37,33	17,33	9,07
15 coef	99,87	79,07	51,33	26,93	11,73
18 coef	99,07	81,33	52,67	28,00	11,20
20 coef	98,80	83,87	55,20	27,07	10,80
	ZCPAC				
12 coef	98,00	97,87	96,93	89,07	41,47
15 coef	98,40	98,27	97,87	90,13	41,33
18 coef	98,13	98,13	97,33	90,67	40,93
20 coef	98,00	97,60	96,80	89,20	40,00

TABELA III

TAXA DE RECONHECIMENTO EM % USANDO CARACTERÍSTICAS ESTÁTICAS E DINÂMICAS (Δ E Δ-Δ) MFCC E ZCPAC PARA A FRASE E1, COM 15 COEFICIENTES, DELTA-WINDOW $= 8$

	MFCC				
Noise\ SNR (dB)	Limpo	20	15	10	5
Branco	99,87	79,07	51,33	26,93	11,73
Fábrica	99,87	98,67	89,07	70,67	28,53
F16	99,87	89,47	55,20	20,40	6,40
Babble	99,87	98,27	87,60	44,67	12,53
	ZCPAC				
Branco	98,40	98,27	97,87	90,13	41,33
Fábrica	98,40	97,60	98,00	94,40	65,73
F16	98,40	97,73	96,40	73,60	23,33
Babble	98,40	98,00	96,00	72,40	29,73

TABELA IV

TAXA DE RECONHECIMENTO EM % USANDO CARACTERÍSTICAS ESTÁTICAS E DINÂMICAS (Δ E Δ-Δ) MFCC E ZCPAC PARA A FRASE E2, COM 15 COEFICIENTES, DELTA-WINDOW $= 8$

	MFCC				
Noise \ SNR (dB)	Limpo	20	15	10	5
Branco	100,00	82,93	54,93	21,33	7,47
Fábrica	100,00	99,47	94,67	71,07	30,80
F16	100,00	95,20	69,87	21,87	5,87
Babble	100,00	99,33	93,33	64,67	25,73
	ZCPAC				
Branco	98,93	98,80	98,00	92,80	72,27
Fábrica	98,93	98,93	98,53	93,33	74,13
F16	99,33	99,07	96,80	83,07	37,07
Babble	98,93	98,67	96,93	85,47	38,00

sentença E1 e ruído gaussiano branco. Nestas tabelas foram avaliados os MFCC e o ZCPAC para observações com diferentes números de coeficientes e para valores de SNR de 5, 10, 15 e 20 dB, bem como para o sinal limpo. Pode ser observado que em baixas SNR o desempenho do ZCPAC é muito melhor que o do MFCC em todas as situações. No entanto, há uma queda significativa de 10 dB para 5 dB. Isto pode ser devido ao fato de que o ruído em 5 dB, torna-se o sinal dominante em boa parte das sub-bandas, principalmente nas altas freqüências, onde os formantes possuem potência mais baixa. Todavia, apesar disso, o ZCPAC ainda supera consideravelmente o desempenho do MFCC, pois nas freqüências mais baixas a potência dos formantes está acima do nível do ruído, sendo detetado pelo ZCPAC. Já o MFCC faz uso de todo o espectro nas sub-bandas - picos e vales - e por esta razão ele é afetado mais intensamente pelo ruído que o ZCPAC, por outro lado essa característica favorece nos casos dos sinais limpos e em alguns casos de ruído com 20 dB de SNR. Das Tabelas I e II observa-se também que as características dinâmicas desempenham um papel fundamental para o desempenho do reconhecimento robusto de locutor.

As Tabelas III e IV mostram os resultados obtidos para as frases E1 e E2, respectivamente, quando as características dinâmicas são incluídas nos vetores de observação com 15 coeficientes. Estes resultados foram obtidos para os ruídos gaussiano branco, fábrica, F16 e babble. Pode-se ver que em todos os casos o desempenho do ZCPAC superou significativamente o MFCC, exceto para 20 dB nos ruídos fábrica e babble. Pode ser observado também que os resultados para a frase E2, a qual é predominantemente composta de sons nasais, superam os resultados da frase E1. Este fato está de acordo com os resultados relatados em [17], onde foi verificado que os sons nasais favorecem o desempenho dos reconhecedores de locutor por conterem informações principalmente do trato nasal.

V. CONCLUSÃO

Neste trabalho foi proposto o uso do ZCPAC para a tarefa de reconhecimento robusto de locutor dependente do texto. A superioridade do desempenho do ZCPAC sobre o MFCC foi confirmada através do uso de sinais de voz distorcidos por ruído gaussiano branco, de fábrica, F16 e o babble. Foi mostrado que o desempenho do reconhecimento do MFCC é melhor que o ZCPAC unicamente no sinal limpo ou para a SNR de 20 dB nos casos de ruído de fábrica e babble. Porém nos demais casos o ZCPAC superou o MFCC. Foi mostrado também que o desempenho do reconhecimento é melhor para a frase E2, cujo conteúdo fonético é basicamente nasal. A continuação deste trabalho será o estudo mais aprofundado dos parâmetros do ZCPAC para que seja feita uma adaptação para a tarefa de reconhecimento robusto de locutor independente do texto, além de uma avaliação do ZCPAC utilizando-se outras bases de voz para a tarefa de reconhecimento de locutor.

REFERÊNCIAS

[1] B. S. Atal, *Automatic Recognition of Speakers From Their Voices*. Proceedings of IEEE, v. 64, nr. 4, p. 460-475, April 1976.

[2] A. E. Rosemberg, *Automatic Speaker Verification: A Review*. Proceedings of IEEE, v. 64, nr. 4, p. 475-487, April 1976.

[3] G. R. Doddington, *Speaker Recognition - Identifying people by their voices*. Proceedings of IEEE, v. 73, nr. 11, p. 1651-1664, November 1985.

[4] S.Furui, *Recent advances in speaker recognition.* Pattern Recognition Letter, v. 18, p. 859-872, 1997.

[5] M. Gabrea and C. Tadj *Speaker enhancement for speaker identification*. International Workshop on Acoustic Echo and Noise Control, Darmstadt, Germany, September 2001.

[6] I. Mccowan, J. Pelecanos and S. Sridharan, *Robust speaker reognition using microphone array*. Proceedings of 2001: A speaker odyssey, June 2001.

[7] R. J. Mammone, Zhang Xiaoyu, R. P. Ramachandran, *Robust speaker recognition: a feature-based approach*. IEEE Signal Processing Magazine, v. 13, Issue 5, p. 58-71, September 1996.

[8] R.C. Rose,E.M.Hofstetter and D.A.Reynolds *Integrated models of signal and background with applicatin to speaker identification in*. IEEE Transaction on Speech and Audio Processing, 2(2):245-257, April 1994.

[9] Y. A. Solewicz, *Noise robustness in forensic speaker verification*. Proceedings of 2001: A speaker odyssey, June 2001.

[10] Ghitza, Oded. *Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition.* IEEE Transactions on Speech and Audio Processing, v. 2, nr. 1, p. 115-131, January 1994.

[11] Doh-Suk Kim, Soo-Young Lee and R.M. Kil, X. Zhu *Auditory model for robust speech recognition in real worls noisy enviroments*. Eletronics Letters, v. 33, nr. 1, p. 12-13, January 1997.

[12] Kim, Doh-Suk, Lee, Soo-Young and Kil R.M., *Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments*. IEEE Transactions on Speech and Audio Processing, v. 7, nr. 1, p. 55-68, January 1999.

[13] Gajic, Bojana e Paliwal, Kuldip K., *Robust Speech Recognition Using Features Based On Zero Crossing With Peak Amplitudes.* ICASSP 2003, p. 64-67, 2003.

[14] B. Kedem, *Spactral Analysis and Discrimination by Zero-Crossings*. Proceedings of the IEEE, v. 74, nr. 11, November 1986.

[15] Picone, Joseph W. *Signal Modeling Techniques in Speech Recognition*. Proceedings of IEEE, v. 81, nr. 9, p. 1215-47, September 1993.

[16] Sachin S. Kajarekar, *Analysis of Variability in Speech with Applications to Speaker and Speaker Recognition*. Phd dissertation, Oregon Healt and Science University, July 2002.

[17] James W. Glenn and Norbert Kleiner, *Speaker Identification Based on Nasal Phonation*. The Journal of the Acoustic Society of America, v. 43, nr. 02, February 1968.

[18] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., Woodland P., *The HTK Book - Version 3.0*. Microsoft Corporation, July 2000.