Comparação entre as técnicas de MFCC e ZCPA para reconhecimento robusto de locutor em ambientes ruidosos

Carlos D. R. Cuadros, Edson Cataldo
UFF - Departamento de Matemática Aplicada
Programa de Pós-graduação em Engenharia de Telecomunicações
24.020-140, Niterói, RJ
E-mail: ecataldo@im.uff.br

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

José A. Apolinário Jr.
IME - Departamento de Engenharia Elétrica (SE/3)
Praça General Tibúrcio, 80
22.2990-270, Rio de Janeiro, RJ
E-mail: apolin@ime.eb.br

Resumo

Várias técnicas de extração de características têm sido propostas para sistemas de reconhecimento de locutor . Porém, poucos trabalhos têm sido dedicados ao desempenho desses sistemas em ambientes ruidosos, embora o interesse no assunto tenha crescido nos últimos anos . Este trabalho tem como objetivo comparar o desempenho de duas técnicas de extração de características, denominadas MFCC (*Mel frequency Cepstral Coefficients*) e ZCPAC (*Zero Crossing with Peak Amplitude Cepstrum*), em ambientes ruidosos.

1 Introdução

Embora a pesquisa em sistemas de reconhecimento de voz e de locutor [1], [2], [3] tenha crescido ao longo dos anos, apenas recentemente houve maior dedicação à aplicação de técnicas em ambientes ruidosos [7].

De modo geral, a robustez do sistema de reconhecimento de voz ou locutor pode ser obtida em diferentes estágios do processamento. Nesse trabalho, interessa-nos a robustez na fase de extração de características do sinal de voz. Em particular, discutiremos duas técnicas. A primeira técnica a ser tratada, MFCC (*Mel frequency Cepstral Coefficients) [9], é uma técnica mais conhecida e a segunda técnica, ZCPAC (*Zero Crossing with Peak Amplitude Cepstrum) [6], é mais recente. Devido às características de construção da técnica ZCPA, espera-se que seu desempenho seja melhor do que a técnica MFCC em ambientes ruidosos. Porém, para ambientes não-ruidosos, o MFCC tem proporcionado um bom desempenho nos sistemas de reconhecimento de locutor [7].

Na Sec. 2 apresentamos a técnica MFCC e uma aplicação em reconhecimento de voz, para o caso de dígitos concatenados. Na Sec. 3, a técnica ZCPAC é apresentada. Resultados provenientes da aplicação das duas técnicas a sinais de voz para reconhecimento de locutor dependente do texto são apresentados na Sec. 4 e, finalmente, na Sec. 5, conclusões são apresentadas.

2 Técnica MFCC

A técnica de extração de características denominada MFCC baseia-se no uso do espectro da

voz alterado segundo a escala *Mel*, uma escala que procura se aproximar de características de sensitividade do ouvido humano [9].

Os coeficientes MFCCs, assim chamados, são extraídos usando-se o seguinte protocolo: (1) aplica-se uma janela, geralmente de *Hamming* ao sinal de voz e em seguida obtém-se a transformada de Fourier do sinal resultante, (2) a amplitude da transformada de Fourier obtida é filtrada por janelas triangulares, seguindo a escala Mel, e a cada trecho resultante aplica-se o logaritmo, (3) Finalmente, aplica-se a transformada discreta de cosseno. Os coeficientes MFCCs são as amplitudes resultantes.

Os coeficientes *Mel-cepstrais* surgiram devido aos estudos na área de psicoacústica (ciência que estuda a percepção auditiva humana), pois verificou-se que a percepção humana de freqüências de tons puros ou de sinais de voz não seguem uma escala linear. Isto estimulou a criação de uma escala da seguinte forma: para cada tom com uma determinada freqüência, medida em Hz, associa-se um valor medido em uma escala chamada escala *Mel*. O *mel* é, então, uma unidade de medida da freqüência. Como referência, definiu-se a freqüência de 1 kHz, com potência 40 dB acima do limiar mínimo de audição do ouvido humano, equivalendo a 1000 $mels$. Os outros valores foram obtidos experimentalmente.

Seja f uma freqüência dada em Hz. O valor associado a essa freqüência na escala *Mel* é denotado por $mel(f)$ e definido pela Eq. 1.

$$mel(f) = 1127\ \ell n\left(1 + \frac{f}{700}\right) \qquad (1)$$

Um outro importante critério relacionado ao conteúdo de freqüência de um sinal é o que chamamos de *banda crítica*. Alguns experimentos mostraram que algumas freqüências, em determinados sons, não podem ser individualmente identificadas pelo sistema auditivo humano, para certas faixas de freqüências. Quando uma componente de freqüência não pertence a essa faixa de freqüências, denominada de *banda crítica*, ela pode ser identificada. Uma explicação bem aceita é que a percepção, pelo sistema auditivo, de uma determinada freqüência é influenciada pela energia da *banda crítica* em torno da freqüência em questão.

Dessa forma, algumas modificações foram feitas na representação espectral de um sinal de forma a favorecer sistemas de reconhecimento de voz e de locutor. Tais modificações consistiram na ponderação da escala de freqüência para a escala mel e na incorporação do conceito de *banda crítica*. Ou seja, usa-se o logaritmo da energia total das bandas críticas em torno das freqüências *mel*. A aproximação mais utilizada para esse cálculo é a utilização de um banco de filtros triangulares, espaçados uniformemente em uma escala não linear (escala mel). O banco de filtros está representado na Fig. 1, no domínio da freqüência.

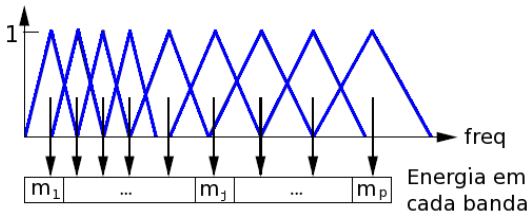

Figura 1: Banco de filtros usado na técnica MFCC.

Para a faixa de freqüências de interesse da voz humana, utilizam-se 20 (vinte) filtros centrados nas freqüências da escala *mel*. O espaçamento é de aproximadamente 150 $mels$ e a largura de banda de cada filtro triangular é de 300 $mels$.

A técnica de ponderação mel pode ser aplicada a vários tipos de representação espectral. No caso da técnica MFCC, utiliza-se a representação *cepstral* pois apresenta maior eficácia computacional, sendo chamada de *mel-cepstral*.

A extração dos chamados coeficientes *cepstrais* usando a escala mel (MFCC) é feita da forma descrita a seguir.

Inicialmente, divide-se o sinal de voz, digamos $s(n)$, em janelas. Para cada trecho m do sinal obtido, calcula-se a transformada discreta de Fourier, obtendo $S(\omega, m)$, no domínio da freqüência.

Faz-se então a convolução do sinal resultante com o banco de filtros, o que equivale a multiplicar no domínio da freqüência o sinal $S(\omega, m)$ pela transformada de Fourier do banco de filtros.

Obtém-se, assim, no domínio da freqüência, a função P, definida pela Eq. 2.

$$P(i) = \sum_{k=0}^{N/2} |S(k,m)|^2 H_i\left(k\frac{2\pi}{N}\right) \qquad (2)$$

com $i = 1, \ldots, N_f$. N é o número de pontos da transformada discreta de Fourier, N_f é o número de filtros triangulares e $\mid S(k,m) \mid$ é o módulo da amplitude na freqüência do $k-$ésimo ponto da $m-$ésima janela e $H_i(\omega)$ é a função resposta em freqüência do $i-$ésimo filtro triangular.

Em seguida, define-se o conjunto de pontos $E(k)$ dados por:

$$E(k) = \begin{cases} \ell og[P(i)] \ , \ k = k_i \\ 0 \ , \ k \in [0, N-1] \end{cases} \qquad (3)$$

Os coeficientes *mel-cepstrais* $C_{mel}(n)$ são então obtidos com o uso da Transformada Inversa (discreta) de Fourier. Após algumas manipulações algébricas, chega-se a

$$C_{mel}(n) = \sum_{i=1}^{N_f} E(k_i) cos\left(\frac{2\pi}{N} k_i n\right)$$

sendo $n = 1, 2, \ldots, N_c$, onde N_c é o número de coeficientes *mel-cepstrais* desejado, N_f é o número de filtros e k_i é o centro do $i-$ésimo filtro.

Com a finalidade de apresentar um exemplo usando a técnica MFCC para a tarefa de reconhecimento de voz, no caso de dígitos concatenados, foi utilizada uma base de voz com 50 locutores femininos e 50 locutores masculinos dos quais cada um deles repete três vezes em Português os números: "zero", "um", "dois", "três", "quatro", "cinco", "seis", "meia", "sete", "oito", "nove".

Cada gravação tem uma taxa de amostragem de 11025 Hz e 16 bits de resolução com um só canal e gravados em ambiente de escritório.

O sistema de classificação utilizado foi um HMM (*Hidden Markov Models*) de 10 estados com 3 misturas de gaussianas por estado. Foi usada a técnica *MFCC*, considerando-se os 15 primeiros coeficientes (menos C_0) mais sua primeira e segunda derivadas.

Os resultados são apresentados em termos da *matriz de confusão*, mostrada na Tab. 1. A taxa de acerto foi de 98%.

Tabela 1: Matriz de confusão

	0	1	2	3	4	5	6	7	8	9
0	22	0	0	0	0	0	0	0	0	0
1	0	32	0	1	0	0	0	0	0	1
2	0	0	27	0	0	0	0	0	0	0
3	0	0	0	34	0	0	3	0	0	0
4	0	0	0	0	32	0	0	0	0	0
5	0	0	0	0	0	38	0	0	0	0
6	0	0	0	0	0	0	29	0	0	0
7	0	0	0	0	0	0	0	28	0	0
8	0	0	0	0	0	0	0	0	33	0
9	0	0	0	0	0	0	0	0	0	19

3 Técnica ZCPA

A técnica ZCPA originou-se do modelo auditivo baseado na técnica *EIH* (Ensemble Interval Histogram), proposto por Ghitza [5] para reconhecimento automático de voz em ambientes ruidosos. Esta técnica mostrou-se melhor do que a técnica MFCC para baixos valores de relação Sinal-Ruído (SNR), desde que os valores dos níveis do EIH sejam bem escolhidos. O EIH é composto de um 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 ao banco de filtros há um arranjo de cinco detectores de cruzamento de níveis cuja função é simular a atividade das células pilosas internas.

No modelo do EIH, a quantidade de atividade nervosa gerada por um dado estímulo acústico é medida através da densidade de probabilidade da diferença entre dos cruzamentos de nível consecutivos. 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 (cruzamento por zero quando a função é crescente). 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 [6] que a estimação de freqüência baseada nos níveis mais baixos são menos suscetíveis a ruído quando compara-

dos aos níveis mais altos. Como consequência, Kim et. al. [6] 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 em relação ao EIH.

Neste artigo nós descrevemos brevemente a extração dos ZCPA Cepstrum (ZCPAC), utilizamo-lo na tarefa de reconhecimento automático de locutor dependente do texto e comparamos seus resultados com a técnica MFCC em ambientes ruidosos e não-ruidosos.

O procedimento para o cálculo do ZCPAC é o seguinte: 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 detector 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:

$$p_k(i) = \max_{z_k(i) \leq n < z_k(i+1)} [s_k(n)] \quad (4)$$

$$f_k(i) = \frac{1}{z_k(i+1) - z_k(i)} \quad (5)$$

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 o incremento dos *bins* por um, o incremento é feito tomando-se o logaritmo da amplitude de pico do sinal no intervalo dos cruzamentos. A contagem do j-ésimo *bin* do histograma é dada, então, por:

$$bin(j) = \sum_{k=1}^{K} \sum_{i=1}^{N_{zc}} \psi[p_k(i)] \quad (6)$$

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

Finalmente, a fim de reduzir a correlação dos dados é aplicada a DCT (transformada discreta de cosseno) gerando os coeficientes *cepstrais* do ZCPA. A Fig.2 ilustra a extração do ZCPA.

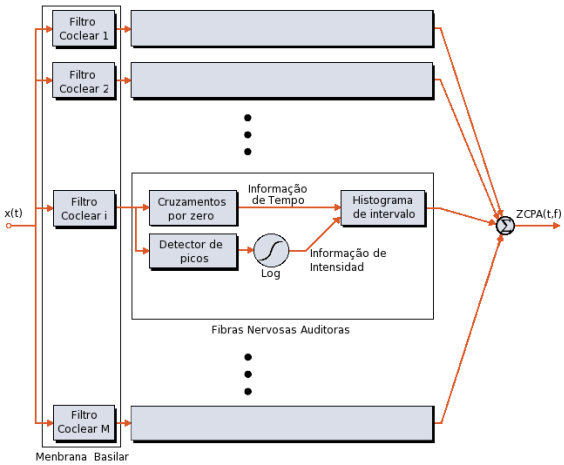

Figura 2: Extração do ZCPA [6].

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 [8] 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.

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.

3.1 Banco de Filtros

Neste estágio o sinal $s(n)$ é decomposto em K componentes $s_i(n)$ por K filtros passa-bandas com freqüência central fc_i e largura de banda bw_i, $i = 1, 2, 3...K$, dispostas linearmente ao longo da escala Bark (ou Mel), carregando in-

formações perceptuais da voz [9].

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 [5]. Todavia, resultados experimentais em [6] 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. A Fig.3 mostra os tipos de filtros usados.

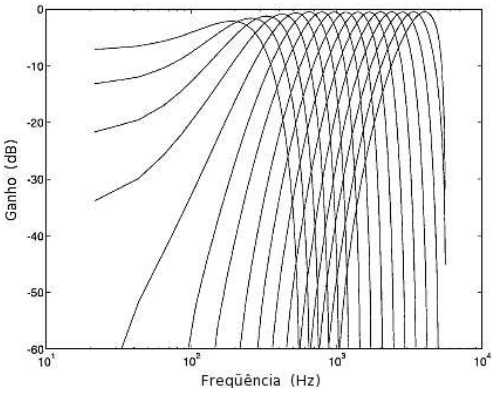

Figura 3: Banco de filtros usado na técnica ZCPA.

Número de bandas (canais): O número de canais (ou bandas) indicados para uso no ZCPA em [6] 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:

$$f_{Bark} = 13\, atan\left(\frac{0,76f}{1000}\right) + 3,5\, atan\left(\frac{f}{7500}\right)^2, \quad (7)$$

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

Largura de Banda: Resultados experimentais em Reconhecimento Automático de Voz (RAV) [7] 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 [9]:

$$BW_{critica}(f) = 25 + 75\left[1 + 1,4\left(\frac{f}{1000}\right)^2\right]^{0,69}, \quad (8)$$

onde f é dado em Hz.

3.2 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 detecção dos cruzamentos pelo zero e os cálculos para construção do espectro.

3.2.1 Alocação de raias

A alocação das raias de freqüência é feita de acordo com a escala Bark, conforme a Eq. (7). A largura, R, de cada raia é dada pela Eq. (8), ou seja, à medida que a freqüência aumenta a largura de R também aumenta, levando o histograma a uma polarização nas altas freqüências.

3.2.2 Definição da janela de observação

Através de várias medições realizadas em [5], 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 ser implementado com o uso de janelas temporais de tamanhos distintos.

4 Resultados

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. Para classificação foram utilizados HMMs (hidden markov models) com modelos esquerda-direita, com 10

estados e 1 gaussiana por estado. O pacote HTK [10] foi utilizado para treinamento e teste. Cada gravação tem uma taxa de amostragem de 8000 Hz e 16 bits de resolução com um só canal e gravados em ambiente de escritório. Acrescentou-se aos sinais ruído branco gaussiano e o teste foi feito com diferentes relações sinal-ruído (SNR): 0 db, 5 dB, 10 dB, 15 dB e 20 dB.

As características foram extraídas a cada 10 ms. Foram extraídos os coeficientes *cepstrais* do ZCPA (sem o coeficiente C_0) e suas primeira e segunda derivadas. Para obter esses coeficientes foram empregados 17 filtros FIR de ordem 61 obtidos a partir de janelas 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 $Np = 30$, 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 à freqüência no histograma. Para comparação, foi considerado o uso do MFCC (sem o coeficiente C_0) e suas primeira e segunda derivadas . 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.

Através dos resultados mostrados na Tab. 2, podemos concluir que a técnica ZCPA é mais robusta que a técnica MFCC em ambientes ruidosos. À medida que a relação sinal-ruído diminui, a superioridade da técnica ZCPA mostra-se mais ainda evidente. Porém, para ambientes não ruidosos (sinais limpos), a técnica MFCC mostrou-se um pouco mais eficiente que a ZCPA.

Tabela 2: Resultados das simulações.

	MFCC				
	Limpo	20dB	15dB	10dB	5dB
12 coef.	**100**	74	40,27	13,87	6,67
15 coef.	99,87	78,13	45,73	14,67	**8,4**
18 coef.	99,87	84,53	52,53	20,13	7,73
20 coef.	99,73	**87,2**	**56,27**	**24,27**	7,07
25 coef.	99,6	80,27	56	21,6	4,13
	ZCPA				
12 coef.	**99,47**	**98,93**	96,8	88,27	46,4
15 coef.	99,07	98,4	**96,93**	**90,67**	55,73
18 coef.	97,73	97,6	95,33	90	58,53
20 coef.	98	97,6	96	89,87	58,4
25 coef.	96,4	96,4	95,33	88,67	**60,13**

5 Conclusões

Neste trabalho, apresentamos as técnicas de MFCC e ZCPAC e fizemos comparações entre os resultados obtidos para a tarefa de reconhecimento robusto de locutor dependente do texto. A superioridade do desempenho do ZCPAC sobre o MFCC foi confirmada no caso de ambientes ruidosos. Foi mostrado que o desempenho do reconhecimento do MFCC é melhor que o ZCPAC unicamente no sinal claro, livre de ruído. 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.

Referências

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

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

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

[4] S. Furui, Recent advances in speaker recognition, *Pattern Recognition Letter*, 18, (1977), 859-872.

[5] O. Ghitza, Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, 2, nr. 1, (1994), 115-131.

[6] D. Kim, S. Lee and R.M. Kil, Auditory Processing of Speech Signals for Robust Speech

Recognition in Real-World Noisy Environments, *IEEE Transactions on Speech and Audio Processing*, 7, nr. 1, (1999), p. 55-68.

[7] B. Gajic, e K. Paliwal,Robust Speech Recognition Using Features Based On Zero Crossing With Peak Amplitudes., *ICASSP 2003*, (2003), 64-67.

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

[9] J. Picone, Signal Modeling Techniques in Speech Recognition, *Proceedings of IEEE*, 81, nr. 9, (1993), 1215-47.

[10] Young, S.J., Evermann G., Gales M.,Hain T., Kershaw D., Moore G. , Odell J., Ollason D., Povey D., Valtchev V., Woodland P. The HTK Book, verbatimhttp://htk.eng.cam.ac.uk/.