Comparação dos Atributos MFCC, SSCH e PNCC para Reconhecimento Robusto de Voz Contínua

Jan Krueger Siqueira e Abraham Alcaim

***Resumo*—Este artigo compara o desempenho do reconhecimento de voz contínua – corrompida por ruído aditivo – utilizando três tipos de atributos: MFCC, SSCH e PNCC. Para a análise experimental, foram escolhidos o software HTK e as bases de dados TIMIT (de vozes) e NOISEX-92 (de ruídos). Os resultados mostram que o método PNCC apresenta maior robustez.**

Palavras-Chave—Reconhecimento Robusto de Voz Contínua, ruído, MFCC, SSCH, PNCC, HTK, TIMIT, NOISEX-92.

***Abstract*—This paper compares the performance of speech recognition – with continuous speech corrupted by additive noise – using three feature types: MFCC, SSCH e PNCC. For the experiment, the software HTK and the databases TIMIT (voice) and NOISEX-92 (noise) were chosen. Results show that the PNCC method has the best robustness.**

Keywords— Robust Continuous Speech Recognition, noise, MFCC, SSCH, PNCC, HTK, TIMIT, NOISEX-92.

I. INTRODUÇÃO

A área de reconhecimento de voz desenvolve sistemas que compreendem a fala humana – ou seja, que reconhecem aquilo que foi dito por uma pessoa. Deste modo, ações podem ser efetuadas de modo verbal, facilitando o trabalho do usuário (por exemplo, nos momentos em que suas mãos se encontram ocupadas). Desde 1952, esses sistemas se desenvolveram bastante e hoje estão presentes em diversas aplicações: ditado de textos, atendimento automático por telefone, auxílio a deficientes físicos, etc.

Um dos maiores problemas ainda presentes nesse campo de estudo é a robustez ao ruído. No momento da gravação, perturbações sonoras como falatório ou ronco de motores podem ser adicionadas ao sinal vocal. Assim, a onda acústica resultante passa a ter outras características além das que foram ditas pelo locutor, o que prejudica bastante a identificação das palavras. Por isso, o grande desafio é escolher um método que extraia bem as informações da fala, eliminando dados relativos aos sons externos.

Os atributos *Mel-Frequency Cepstral Coefficients* (MFCC) ainda são muito utilizados, porém sua eficácia cai rapidamente com a presença do ruído. Em [1], o método *Subband Spectral Centroid Histograms* (SSCH) foi aplicado e apresentou um bom resultado para uma base de voz pequena. Mais recentemente, [2] introduziu os atributos *Power-Normalized Cepstral Coefficients* (PNCC), também robustos ao ruído.

É importante ressaltar que não existe na literatura uma comparação entre o SSCH e o PNCC. Portanto, o objetivo deste artigo é avaliar seus desempenhos num teste unificado e com uma base de voz de vocabulário amplo.

Os três métodos – MFCC, SSCH e PNCC – serão descritos nas seções II, III e IV, respectivamente. As condições experimentais serão detalhadas na seção V. Finalmente, resultados e conclusões serão apresentados na seção VI.

II. ATRIBUTOS MFCC

Um sinal senoidal de 880 Hz não soa duas vezes mais agudo que um de 440 Hz e nem quatro vezes mais agudo que um de 220 Hz. Ou seja, a escala em Hertz não reflete bem a percepção auditiva humana. Para representá-la melhor, foi criada a escala mel [3]. Através de experimentos com diversos ouvintes, chegou-se na seguinte conversão de f (em Hertz) para m (em mel):

$$m = M(f) = 1125\ln\left(1+\frac{f}{700}\right) \tag{1}$$

$$f = M^{-1}(m) = 700\left(e^{\frac{m}{1125}}-1\right) \tag{2}$$

A nova medida se mostrou muito eficaz para extrair dados para o reconhecimento de voz. O procedimento consiste em dividir o espectro do sinal em B bandas, com frequências centrais igualmente espaçadas na escala mel. Para isso, são aplicados filtros digitais triangulares $H(k)$, como mostra a Figura 1:

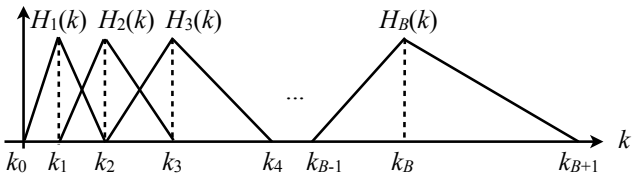

Fig. 1. Filtros triangulares igualmente espaçados na escala mel.

Arbitrando os limites k_0 e k_{B+1}, e fazendo a distância entre cada k_b ser a mesma na escala mel, temos:

$$f_b = M^{-1}\left(M(f_0)+b\frac{M(f_{B+1})-M(f_0)}{B+1}\right) \tag{3}$$

$$k_b = \frac{N}{F_S}f_b \tag{4}$$

onde N é o número de amostras da transformada discreta de Fourier (DFT) do sinal e F_S é a sua taxa de amostragem. Cada filtro é dado então por:

Jan Krueger Siqueira 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: janksiqueira@gmail.com, alcaim@puc-rio.br. Este trabalho foi parcialmente financiado pelo CNPq .

$$H_b(k) = \begin{cases} 0 & k < k_{b-1} \\ \dfrac{k - k_{b-1}}{k_b - k_{b-1}} & k_{b-1} \le k < k_b \\ \dfrac{k_{b+1} - k}{k_{b+1} - k_b} & k_b \le k < k_{b+1} \\ 0 & k \ge k_{b+1} \end{cases} \quad (5)$$

Definidos os filtros, os atributos MFCC são obtidos para cada quadro do sinal com as seguintes etapas:

1. A transformada discreta de Fourier é aplicada ao quadro, obtendo-se o espectro.
2. O espectro é dividido em bandas através dos filtros dados pela equação (5).
3. Calcula-se o logaritmo da energia de cada banda, já que esse tipo de não-linearidade é observado no sistema auditivo humano.
4. Por último, a transformada discreta do cosseno (DCT) é aplicada à sequência de logaritmos do item anterior, a fim de descorrelatá-los.

III. ATRIBUTOS SSCH

O método SSCH foi apresentado em [1] e [4], gerando resultados melhores do que outros atributos conhecidos na literatura. Seu argumento é que o formato do espectro do sinal de voz não se distorce muito com a adição de um ruído moderado. Por isso, a localização dos picos não se altera tanto. A Figura 2 ilustra isso com o espectro de um sinal de voz sem ruído e o mesmo espectro com ruído branco de razão sinal-ruído de 10 dB.

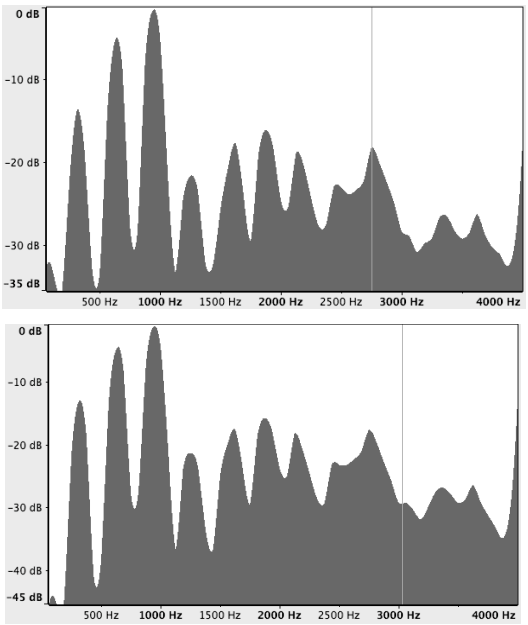

Fig. 2. Comparação entre os espectros da vogal "a" num sinal limpo e num sinal com ruído branco de razão sinal-ruído de 10 dB.

Observando isso, surgiu a ideia de representar o formato do espectro através do centroide e da potência de suas sub-bandas. O sinal é dividido em quadros e as etapas a seguir são aplicadas a cada um:

1. A densidade espectral de potência (DEP) é encontrada elevando ao quadrado o módulo da transformada discreta de Fourier do quadro.
2. A DEP é dividida em B bandas de frequências superpostas, num procedimento similar ao descrito na seção anterior. A diferença é que a escala Bark é usada ao invés da mel, e sua relação com a frequência em Hertz é dada por:

$$a = A(f) = \frac{26.81 f}{f + 1960} - 0.53 \quad (6)$$

3. Para cada banda, o centroide é obtido através da equação:

$$C_b = \frac{\sum_{k=1}^{N} k H_b(k) P(k)}{\sum_{k=1}^{N} H_b(k) P(k)} \quad (7)$$

com $P(k)$ sendo a DEP, $H_b(k)$ o filtro da banda b no domínio da frequência digital e N o número de amostras da transformada discreta de Fourier.

4. A energia da banda é calculada.
5. Após aplicar 3 e 4 em cada banda, o eixo das frequências é dividido em I intervalos de mesmo comprimento e sem sobreposição. Para cada um deles, são somadas todas as energias dos centroides que estiverem no intervalo. A coletânea das somas forma o histograma.
6. Por fim, a transformada discreta de cosseno é aplicada nos valores do histograma, a fim de descorrelatá-los.

IV. ATRIBUTOS PNCC

Os atributos PNCC foram apresentados em [2] e [5] como uma evolução dos MFCC, alterando algumas de suas etapas para torná-lo mais robusto.

A primeira modificação se dá na divisão das bandas (etapa 2 do MFCC). Ao invés de B filtros triangulares baseados na escala mel, são aplicados B filtros gammatone [6]. Eles representam bem a resposta impulsional da membrana basilar do ouvido humano, cuja expressão no domínio do tempo é dada por:

$$g(t) = a t^{n-1} e^{-2\pi c_b t} \cos(2\pi f_b t + \phi) \quad (8)$$

onde a é a amplitude, n é a ordem do filtro, c_b é o comprimento de banda, f_b é a frequência central da banda e ϕ é a fase. Baseado nesse comportamento, a escala Equivalent Rectangular Bandwidth (ERB) foi criada, e seus valores em função de f (em Hertz) são iguais a:

$$e = ERB(f_b) = 24.7(1 + 0.00437 f_b) \quad (9)$$

Com a mesma lógica utilizada no MFCC, as frequências centrais f_b são espaçadas em intervalos de mesmo comprimento na escala ERB, como mostra a Figura 3. Os detalhes da implementação se encontram em [7].

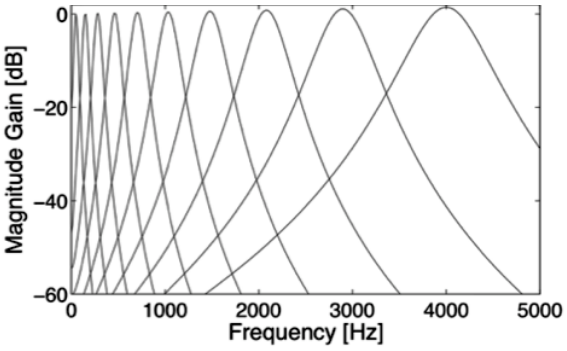

Fig. 3. Exemplo de um banco de filtros gammatone.

A segunda modificação adiciona uma nova etapa. A Figura 2 mostrou o efeito do ruído no espectro de voz. Nota-se que houve uma elevação geral na curva. Portanto, é interessante remover esse acréscimo após a divisão do sinal em bandas, aprofundando seus vales. Isso é feito com a média das energias de uma banda ao longo de alguns quadros consecutivos, pois o ruído costuma ser mais estacionário que a onda de voz. A implementação detalhada é um pouco extensa para ser explicada aqui, mas pode ser encontrada nos artigos originais.

A terceira e última modificação foi feita na operação não-linear sobre a energia da banda (etapa 3 do MFCC). A função logarítmica 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. A energia então é elevada a uma constante a_0 determinada experimentalmente.

Em resumo, os atributos PNCC são extraídos de cada quadro do sinal com as seguintes etapas:

1. A transformada discreta de Fourier é aplicada ao quadro, obtendo-se o espectro.
2. O espectro é dividido em bandas através dos filtros gammatone, dados pela equação (8).
3. O ruído de cada banda é estimado e removido.
4. A energia de cada banda é calculada e elevada a uma constante a_0.
5. Por último, a transformada discreta do cosseno (DCT) é aplicada à sequência de valores do item anterior, a fim de descorrelatá-los.

V. Condições de Teste

Nos experimentos, o banco de dados de voz utilizado foi o TIMIT, que abrange os diversos sotaques do inglês americano para ambos os sexos. Seu vocabulário consiste em 6234 palavras. 630 locutores pronunciaram 10 frases cada, gerando 4620 sentenças para treino e 1680 para teste.

Já os ruídos foram selecionados da base NOISEX-92. Ela contém arquivos de som de diversas naturezas, tais como falatório e ruído branco. Trechos aleatórios desses sinais foram adicionados às amostras de teste, em diversas razões sinal-ruído. O treinamento foi realizado apenas com sinais limpos.

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

Nos três métodos, os sinais de áudio tiveram a taxa de amostragem reduzida para 8 kHz, buscando simular a faixa do sistema telefônico. O filtro de pré-ênfase $H(z) = 1 - 0.97z^{-1}$ foi então aplicado, seguido da divisão em quadros de 25 ms com um passo de 10 ms entre um e outro (superposição de quadros). Cada quadro foi multiplicado por uma janela de Hamming e depois submetido à extração dos atributos com os seguintes parâmetros:

- MFCC: B = 26 bandas e apenas os 12 primeiros valores da DCT foram considerados.
- SSCH: B = 60 bandas, I = 15 intervalos e todos os 15 valores da DCT foram considerados.
- PNCC: B = 40 bandas com filtros de ordem n = 1, expoente a_0 = 1/15 e apenas os 20 primeiros valores da DCT foram considerados.

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

VI. Resultados e Conclusões

A taxa de acerto de um reconhecedor de voz é dada pela expressão:

$$T = 100\frac{N - S - D - I}{N} \tag{10}$$

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.

Os resultados se encontram nas Tabelas 1 e 2.

TABELA I. Taxas de acerto para amostras corrompidas por ruído branco, para diversas razões sinal-ruído

	limpo	20 dB	15 dB	10 dB
MFCC	84.19 %	69.51 %	45.96 %	18.20 %
SSCH	75.88 %	68.49 %	51.99 %	29.92 %
PNCC	88.40 %	79.98 %	72.81 %	64.96 %

TABELA II. Taxas de acerto para amostras corrompidas por falatório, para diversas razões sinal-ruído

	limpo	15 dB	10 dB	5 dB
MFCC	84.19 %	72.13 %	43.59 %	11.60 %
SSCH	75.88 %	65.76 %	48.24 %	16.15 %
PNCC	88.40 %	79.18 %	69.17 %	39.59 %

O método SSCH só se mostrou mais vantajoso que o MFCC para os casos de menor razão sinal-ruído. Já o PNCC

forneceu os melhores resultados em todos os cenários, superando o MFCC até mesmo quando não há ruído.

Conclui-se então que, dentro do cenário considerado neste trabalho, os atributos PNCC são os mais adequados para reconhecimento de voz em presença de ruído aditivo.

AGRADECIMENTOS

Os autores agradecem o CNPq pelo financiamento da pesquisa.

REFERÊNCIAS

[1] B. Gajic e K. K. Paliwal, "Robust Parameters for Speech Recognition Based on Subband Spectral Centroid Histograms", *Eurospeech 2001*, v. 1, pp. 591–594, Setembro 2001.

[2] C. Kim e R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction", *INTERSPEECH-2009*, pp. 28–31, Setembro 2009.

[3] X. Huang, A. Acero e H-W Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

[4] B. Gajic, "Auditory Based Methods for Robust Speech Feature Extraction", *Telektronikk*, v. 2, pp. 45-58, 2003.

[5] C. Kim e R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring", *Proc. IEEE ICASSP*, 2010.

[6] R. D. Patterson e J. Holdsworth, "A functional model of neural activity patterns and auditory images", *Advances in Speech, Hearing and Language Processing*, pp. 547-563, 1996.

[7] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank", Tech. Rep. 35, Apple Computer Inc., 1993

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