Análise do Impacto da Subtração Espectral para Tratamento de Ruído no Reconhecimento de Palavras Isoladas com Coeficiente Mel-Cepstrais

Dunfrey Pires Aragão[1], Nathália Alves Rocha Batista[1], Daniella Dias C. da Silva[1], Caio Fernandes Gabi[2], Silvana Luciene do Nascimento Cunha Costa[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
Campina Grande, PB – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
João Pessoa, PB – Brasil

dunfrey.p.a@ieee.org,nathaliaarb@ieee.org,daniella.silva@ifpb.edu.br,
cfgabi@hotmail.com,silvanacunhacosta@gmail.com

Abstract.*Noisy environments can compromise speech recognition as the system cannot distinguish noise from the voice sample. In this study, we analyzed the efficiency of spectral subtraction technique with a half-wave rectification for the treatment of stationary noise. The effectiveness of this technique will be evaluated by the framework Sphinx, taking into account also the relationship between recognition rate, level of signal-noise ratio and number of Mel-cepstrais coefficients. As a result, it was observed that depending on the level of SNR, the spectral subtraction technique suppresses important frequencies, generating the musical noise, being preferable to allow the system to recognize the noisy signal.*

Resumo. *Ambientes ruidosos podem comprometer sistemas de reconhecimento de fala de modo que o sistema não consegue distinguir o ruído do sinal de voz. Neste trabalho foi analisada a eficiência da técnica de subtração espectral com retificação de meia onda para tratamento do ruído estacionário. A eficácia desta técnica foi avaliada no Framework Sphinx, levando em consideração a relação entre a taxa de reconhecimento, o nível de relação sinal-ruído e número de coeficientes Mel-cepstrais. Como resultado, observou-se que dependendo no nível de SNR, a técnica de subtração espectral suprime frequências importante do sinal, gerando o chamado ruído musical, sendo preferível permitir que o sistema utilize o sinal ruidoso.*

1. Introdução

Com o advento da tecnologia, as máquinas predominam em quase todos os cenários do cotidiano das pessoas, sejam elas computadores, eletrodomésticos, ou ainda dispositivos portáteis. A possibilidade de realizar a comunicação homem-máquina através da voz torna essa interação mais fácil e produtiva. Além disso, é a forma mais natural de comunicação humana, permitindo que as mãos e os olhos dos usuários fiquem disponíveis para outras tarefas. Sua estrutura é moldada pelas estruturas fonológicas,

sintáticas e prosódicas da língua, do ambiente acústico, do contexto em que a fala está sendo produzida (por exemplo, as pessoas falam de maneira diferente em ambientes ruidosos e silenciosos), e do canal através do qual é transmitida.

A complexidade dessas aplicações caracteriza-se ao longo de duas dimensões: o tamanho do vocabulário e a forma da elocução. Quanto maior o vocabulário, maior a dificuldade de reconhecimento, assim como uma conversação espontânea incluindo expressões como “hum...”, “ah!”, além de palavras parciais e/ou concatenadas, são bem mais difíceis de reconhecer do que palavras pronunciadas de uma maneira estritamente articulada e pausada [Ostendorf et al. 2008].

Enquanto que para o idioma inglês existem várias pesquisas e bons resultados, pouco se tem para o idioma português brasileiro [Alencar 2008] [Alcaim, 2008][Bresolin et al. 2008]. Para um sistema de reconhecimento de fala eficiente, torna-se necessária a adequação dos modelos acústicos e linguísticos para o português brasileiro, além de testes variando os principais parâmetros utilizados [Alencar 2008] [Alcaim 2008][Bresolin et al. 2008][Tevah 2006][Yared e Violaro 2005].

As tecnologias mais promissoras na área de reconhecimento de fala, como por exemplo, *HiddenMarkovModels*(HMM) utilizam métodos de modelagem estatística que aprendem por exemplos, e devem considerar no treinamento as possíveis variações da fala e o ambiente de utilização do processo do sistema de reconhecimento [Alencar 2008][Alcaim 2008][Bresolin et al. 2008].

Entretanto, em praticamente todas as aplicações que tem como principal objeto o sinal de voz, a qualidade da comunicação é constantemente prejudicada pela presença de elementos que deterioram a informação. Tais elementos podem afetar o sinal de diversas formas, sobretudo reduzindo sua inteligibilidade e/ou afetando a eficiência de outros sistemas que se utilizarão desses sinais posteriormente, como reconhecedores ou codificadores de voz [Kanda 2010].

Ruídos presentes no meio ambiente interferem e dificultam o reconhecimento da fala, pois o sistema não tem como distingui-los do sinal de voz. Tais ruídos podem ser originados por outras pessoas conversando, barulhos de portas que abrem ou fecham, ar condicionado, entre outros fatores que produzam algum som além do sinal de fala que se deseja reconhecer [Mitra et al. 2012]. Além das características dos ruídos do ambiente, o tipo e posicionamento do microfone podem afetar o desempenho do sistema, uma vez que o mesmo pode gerar ruídos de interferência [Gilbert e Feng 2008] [Juang e Rabiner 1991].

Desta maneira, diante da necessidade de se reduzir, ou mesmo eliminar o ruído presente nos sinais de voz, torna-se necessário a utilização de uma técnica de tratamento de ruído. O tratamento utilizado, neste trabalho, será a Subtração Espectral com retificação de meia onda, e a eficácia desta técnica será avaliada pelo *Framework*Sphinx, levando em consideração, a relação entre a taxa de reconhecimento, nível da relação sinal-ruído e o número decoeficientes Mel-cepstrais.

Este documento está organizado da seguinte forma. Na Seção 2 são apresentadas as técnicas de subtração espectral avaliadas neste trabalho. Na Seção 3 são apresentados os principais conceitos relacionados aos coeficientes Mel-cepstrais. Na Seção 4 é realizada uma breve descrição do *framework* Sphinx bem como, da base de dados utilizada nos testes de reconhecimento. Na Seção 5 são descritos os experimentos

realizados e resultados obtidos. Por fim, na Seção 6 são apresentadas as conclusões deste trabalho.

2. Subtração Espectral com Retificação de Meia Onda

O método de Subtração Espectral é uma abordagem simples e eficaz para suprimir ruído de fundo estacionário. Este método é baseado no conceito de que o espectro na frequência do sinal é expresso como a soma do espectro de voz e espectro do ruído [Silva et al. 2007]. O processamento é feito inteiramente no domínio da frequência [Kanda 2010][Pagoraro 2000][Silva 2007].

Desta forma para que a técnica seja aplicada teremos que adotar que o sinal contaminado por ruído seja dado por:

$$y(n) = v(n) + r(n) \qquad (1)$$

, em que *y(n), v(n)* e *r(n)* é o sinal contaminado por ruído, sinal sem ruído e ruído aditivo, respectivamente.

Calculando a Transformada de Fourier da equação (1), teremos:

$$Y(\omega) = V(\omega) + R(\omega) \qquad (2)$$

Tomando-se o quadrado na equação (2), e considerando que o ruído é aditivo e descorrelacionado com o sinal de voz, temos que:

$$\hat{\mathrm{V}}^2(\omega) = Y^2(\omega) - \hat{R}^2(\omega) \qquad (3)$$

Como se pode observar nas equações (1) e (3), a Subtração Espectral dependem da estimação do espectro do ruído. A Subtração Espectral é aplicada apenas para o espectro de potência do sinal, ou do espetro de amplitude, preservando-se a fase do sinal ruidoso [Kanda 2010][Silva et al. 2007][Silva 2007].

De modo geral, é possível expressar a B-ésima potência do espectro de potência do sinal de voz como [Silva et al 2007]:

$$\hat{V}^B(\omega) = Y^B(\omega) - n\hat{R}^B(\omega) \qquad (4)$$

, em que*B*é um inteiro (normalmente igual a Um ou a Dois) e *n* é um parâmetro para controlar a quantidade de ruído a ser subtraída do sinal de voz degradado [Silvaet al. 2007]. Consequentemente, a estimativa do espectro da voz pode ser recuperada através da equação:

$$\hat{V}(j\omega) = \left[Y^B(\omega) - n\hat{R}^B\right]^{\frac{1}{B}} e^{j(\omega)} \quad (5)$$

, sendo Ψa fase de $X(j\omega)$.

A subtração instantânea da potência de espectro resulta de $B = 2$ e $n = 1$, e a subtração da magnitude de espectro origina-se de $B = 1$ e n = 1.

A Subtração Espectral aplica-se apenas a sinais considerados estacionários. Para que este resultado seja obtido devem-se considerar amostras com curtos intervalos de tempo, entre 16 e 32ms. Desta forma, para estimar o ruído, é feita a segmentação do sinal com ruído *y(n)* em blocos de N amostras. Transformando, com DFT, cada bloco em um bloco de *N* amostras espectrais. Blocos sucessivos de amostras espectrais formam uma matriz bidimensional (com informações de tempo e frequência), denotada por $Y_T(j\omega)$, na qual *T* é o índice do número do bloco e designa a dimensão de tempo [Silva et al. 2007].

A estimação do ruído é expressa como:

$$\begin{cases} \hat{R}_T^B(\omega) = \alpha_A \hat{R}_{T-1}^B(\omega) + (1-\alpha_A) Y_{T-1}^B(\omega), se Y_T^B(\omega) \geq \hat{R}_{T-1}^B(\omega) \\ \hat{R}_T^B(\omega) = \alpha_D \hat{R}_{T-1}^B(\omega) + (1-\alpha_D) Y_{T-1}^B(\omega), se Y_T^B(\omega) < \hat{R}_{T-1}^B(\omega) \end{cases} \quad (6)$$

Sendoα_Ae α_D parâmetros que controlam as constantes de tempo das recursões, atualizando mais rapidamente nos casos em que o ruído é predominante no sinal, e mais lentamente nos casos em que ele não é tão relevante.

Devido à natureza aleatória do ruído podem ocorrer situações em que a magnitude do espectro do sinal contaminado por ruído seja menor que a magnitude média do ruído, resultando em valores negativos para a subtração espectral, gerando ruído musical [Silva 2007]. Neste caso é interessante a aplicação de um recurso conhecido como retificação de meia onda, que pode ser descrita como:

$$\hat{V}^B(\omega) = \begin{cases} Y^B(\omega) - n\hat{R}^B(\omega), se\ Y^B(\omega) \geq \hat{R}^B(\omega) \\ 0, \quad Caso\ Contrário \end{cases} \quad (7)$$

3. Coeficientes Mel-cepstrais

Os coeficientes Mel-cepstrais surgiram devido a estudos na área de psicoacústica que mostraram que a percepção humana das frequências de tons puros ou de sinais de voz não segue uma escala linear. Diante disso, foram definidas frequências subjetivas de tons puros, da seguinte forma: para cada tom com frequência *f*, medida em Hz, define-se um tom subjetivo medido em uma escala chamada mel. O mel, então, é uma medida da frequência percebida de um tom [Garau et al 2008].

Os coeficientes Mel-cepstrais são obtidos aplicando-se a transformada inversa de Fourier, com base na seguinte equação:

$$C_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k)e^{j\left(\frac{2\pi}{N}\right)kn} \qquad n = 1,2,\ldots,N \quad (8)$$

, sendo *E(k)* o número de pontos do sinal a ter sua banda crítica analisada, e sendo N_C o número de coeficientes desejado.

É possível traçar uma comparação entre a frequência real (medida em Hz) e a frequência percebida (medida em mel). O mapeamento entre a escala de frequência real, em Hz, e a escala de frequências percebida, em mel, é aproximadamente linear abaixo de 1000 Hz e, logarítmica, acima [Picone et al 1993]. Dessa forma, algumas modificações foram realizadas na representação espectral de um sinal, de forma a favorecer sistemas de reconhecimento de fala e de locutor. Tais modificações consistiram na ponderação da escala de frequência para a escala mel e na incorporação do conceito de banda crítica. Ou seja, é utilizado o logaritmo da energia total das bandas críticas em torno das frequências mel. A aproximação mais utilizada para esse cálculo é a utilização de um banco de filtros triangulares (Figura 1), espaçados uniformemente em uma escala não linear (escala Mel) [Moller 1983].

Além dos Mel-cepstrais, é comum os sistemas utilizarem as derivadas temporais como componentes adicionais no vetor de atributos, colocando-as em um patamar de igual relevância ao dos coeficientes em si. O objetivo é refletir melhor as mudanças dinâmicas dos MFCC no tempo. As componentes de primeira derivada, conhecidas como Δ-Cepstrum ou Δ-MFCC, representam a velocidade com que o espectro Mel-cepstral varia e são facilmente computadas calculando a diferença entre coeficientes de *m* índices no passado e no futuro do tempo levado em consideração. É de grande valia também as componentes de aceleração do espectro Mel-Cepstral, por isso são utilizadas as derivadas de Δ-MFCC, chamadas de ΔΔ-Cepstrum ou ΔΔ-MFCC [Furui 1986][Bing-Hwang 2000].

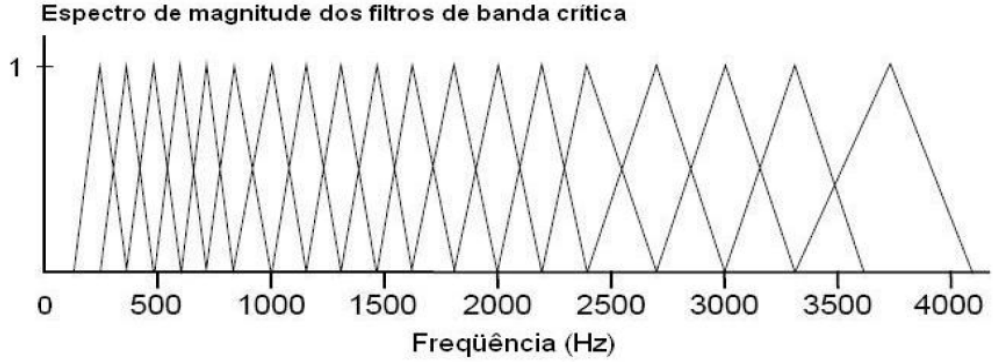

Figura 1 – Banco de filtros digitais na escala mel.

4. Sphinx e a Base de Dados

A fala quando gerada de modo espontâneo é mais relaxada, pois contém mais coarticulações, e, portanto é muito mais difícil de reconhecer do que quando gerada através de leitura ou apenas palavras isoladas. Os HMMs são estruturas poderosas que são capazes de modelar, ao mesmo tempo, as variabilidades acústicas e temporais do sinal de voz, e com isto, é extremamente útil para alcançar resultados de reconhecimento de fala satisfatórios. Com base nisto, neste trabalho foi utilizado o

*framework*Sphinx, o qual constitui o estado da arte na área de reconhecimento de fala, tendo como predominância os modelos estatísticos, notadamente baseados em Modelos Ocultos de Markov (*HiddenMarkovModels* – HMM).

O Sphinx [LEE et al 1990] foi desenvolvido e lançado no ano 2000 pela Universidade Carnegie Mellon (CMU - *Carnegie MellonUniversity*) para estimular a criação de ferramentas e aplicativos de processamento de voz e fazer avançar o estado da arte em reconhecimento de fala, bem como, nas áreas correlatas, como sistemas de diálogo e síntese de voz. Os termos de licença para as ferramentas e bibliotecas disponíveis no Sphinx são derivados do modelo de distribuição de *software* da Berkeley (BSD - *Berkeley Software Distribution*). Não há restrição alguma quanto ao uso comercial ou redistribuição.

O Sphinx ainda possui uma ferramenta chamada SphinxTrain que é utilizada para o treinamento do modelo acústico. Este módulo permite o uso tanto de HMMs contínuos como semi-contínuos no treinamento [Oliveira et al. 2010]. Para os testes de reconhecimento, foi utilizada a base de dados desenvolvida em [Yared e Violaro 2005]. Optou-se pelo uso dessa base, uma vez que, sua estrutura é adequada para o teste de aplicações para reconhecimento de fala contínua para o idioma português. A base é formada por diferentes locutores: foram gravadas amostras de voz com 46 locutores, sendo 24 do gênero masculino e 22 femininos, de diferentes idades (18 a 60 anos), graus de instrução e cidades do Brasil. Além disso, foram utilizadas 200 frases foneticamente balanceadas [Alcaim et al. 1992].

5. Procedimentos e Resultados

Os algoritmos de subtração espectral, simples e com retificação de meia onda, foram implementados no *software*Matlab. O mesmo também foi utilizado para inserir, de forma artificial, o Ruído Banco (*Additive White GaussianNoise - AWGN*) aos sinais de voz originais da base de dados. Em seguida, estes sinais foram tratados de duas formas diferentes. Primeiro, através da Subtração Espectral simples, e em seguida com retificação de meia onda.

Após a geração dos sinais com ruído, realizaram-se os testes de reconhecimento no Sphinx. Vale salientar que se considerou o reconhecimento de palavras isoladas e independente de locutor, formando 400 sentenças e totalizando 2628 palavras. A taxa de erro para os sinais sem ruído foi de aproximadamente 8%. Na Tabela 1 são apresentadas as taxas erro para o reconhecimento dos sinais com ruído assim como, para os sinais processados pelos dois algoritmos avaliados neste trabalho.

Tabela 1 – Taxa de erro para o reconhecimento dos sinais com ruído, processados com retificação de meia onda e sem retificação de meia onda.

	Sinal Ruidoso (%)	Sinal processado com retificação de meia onda (%)	Sinal processado sem retificação de meia onda (%)
10	91,1	94,4	94,6
12	84,3	94,	94
14	73	93,6	93,8
16	57	93,5	92
18	45	92,6	94,1
20	38,6	91,9	93,5

Analisando a Tabela 1, é possível observar que ao tratar o sinal ruidoso com as técnicas em questão, no pior caso, a taxa de erro chega a 94%. A Subtração Espectral com retificação de meia-onda alcançou uma taxa de reconhecimento apenas um pouco melhor que a técnica tradicional. Acredita-se que essa queda no reconhecimento seja devido ao ruído musical, que é inserido no sinal, ao se processar o mesmo com as técnicas analisadas. Na Tabela 1 também foram apresentados os resultados obtidos, ao variar a relação sinal ruído (*SignalNoise Rate* – SNR) entre 10 e 20. Estes resultados são ilustrados graficamente na Figura 2.

Com a variação da SNR, foi possível observar que à medida que a mesma aumenta, a taxa de erro diminui. No entanto, é importante destacar que a taxa de erro para o sinal ruidoso ainda é menor do que para o sinal processado, independente da técnica utilizada. Por exemplo, para uma SNR igual a 20, enquanto a taxa de erro é de 38,6% para o sinal ruidoso original, para as técnicas subtração espectral com e sem retificação de meia onda, este valor aumenta para 91,9% e 93,5%, respectivamente.

Embora não fosse esperada uma redução significativa na taxa de erro, devido ao ruído musical que é inserido pelas técnicas em questão, o fato do reconhecimento ser melhor com o sinal ruidoso original do que com o sinal processado, gerou uma grande surpresa.

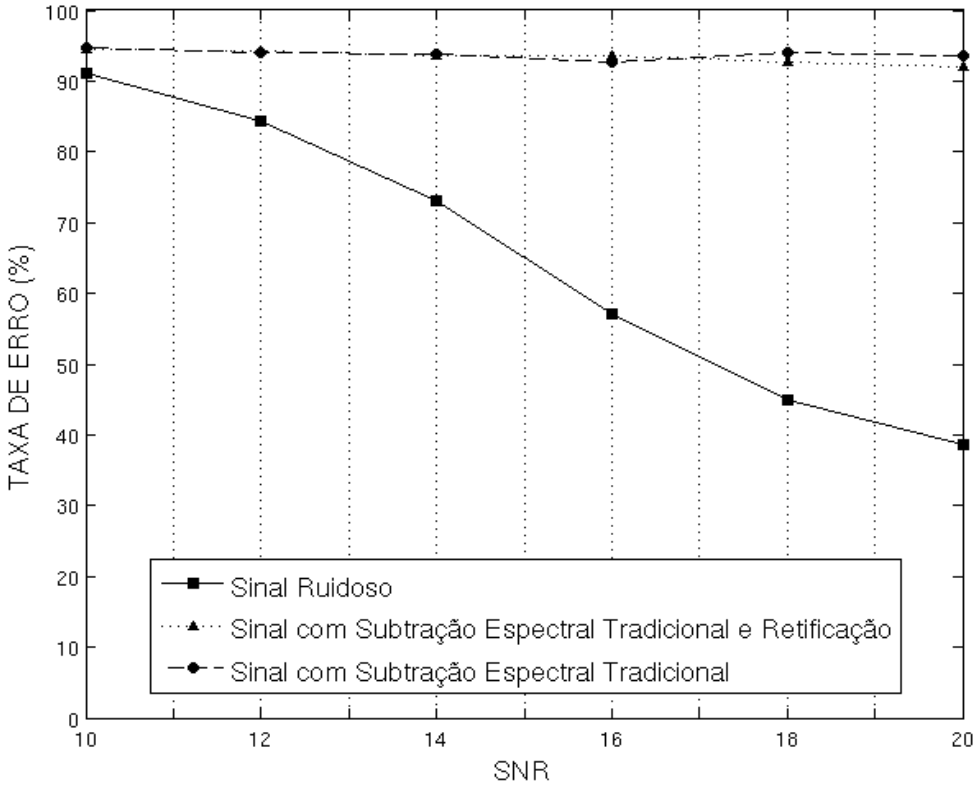

Figura 2 - Taxa de erro para o reconhecimento de palavras com SNR variando entre 10 e 20.

Assim, diante dos resultados obtidos, realizaram-se novos experimentos para verificar se ao variar o número de coeficientes MFCC, estes valores sofreriam alguma mudança. Os valores apresentados na Tabela 1 e Figura 2 são para 13 coeficientes Mel-

cepstrais, que é o valor padrão utilizado pelo Sphinx. Na segunda bateria de experimentos, foram realizados testes variando de 9 a 16 coeficientes Mel-cepstrais e SNR igual a 20. Os resultados são apresentados na Tabela 2 e Figura 3.

Tabela 2 – Taxa de erro do reconhecimento de palavras variando o número de MFCC para sinais com ruído, processados com retificação de meia onda e sem retificação de meia onda.

MFCC	Sinal Ruidoso (%)	Sinal processado com retificação de meia onda (%)	Sinal processado sem retificação de meia onda (%)
9	44,3	90,5	91,5
10	43,3	92,1	93,2
11	40,4	90,4	92,9
12	40,6	92,8	93
13	39,5	93,5	93,4
14	38,5	92,2	93,2
15	42,1	93,6	94,2
16	43,7	95,5	93,5

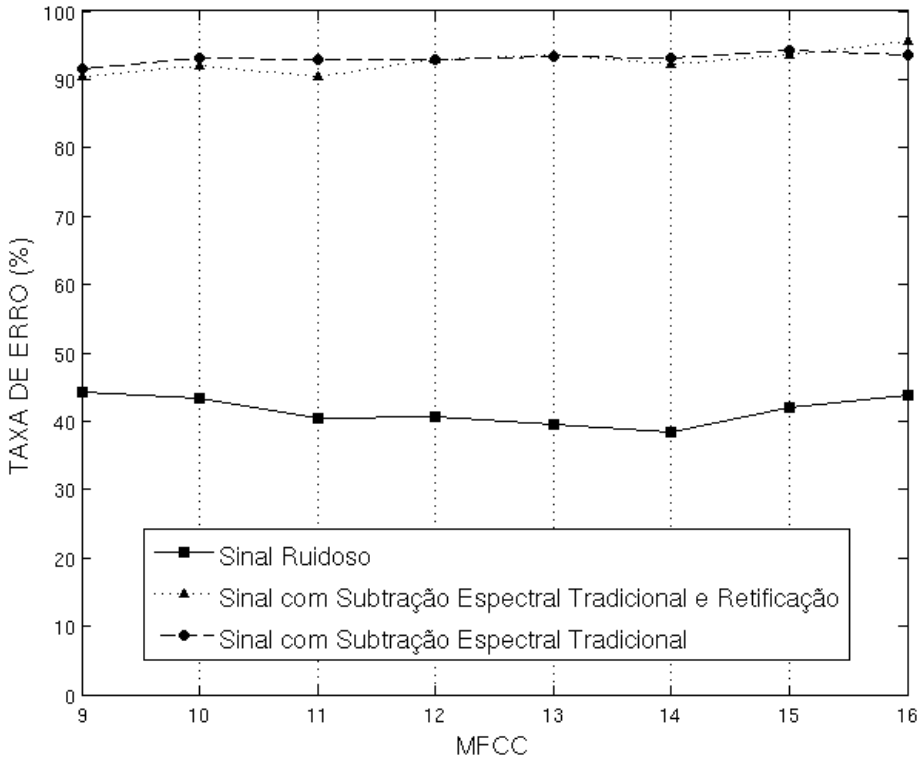

Figura 3 - Taxa de erro variando o número de MFCC de 9 a 16 e SNR igual a 20.

6. Conclusões

Neste trabalho foram realizados testes de reconhecimento automático da fala utilizando o*framework* Sphinx, de modo a validar a eficiência da técnica de subtração espectral para tratamento de ruído ambiental estacionário. Diferente de outros trabalhos relacionados que fazem essa validação via análise espectral, neste foi possível verificar a eficiência da técnica diretamente em um sistema de reconhecimento, sendo aplicados os testes em arquivos em sua forma original, como também em arquivos com o acrescento de ruídos AWGN, variando o nível SNR de 10 a 20, e em arquivos processados pela técnica de estudo, objetivando a análise de supressão de ruídos.

Antes de iniciar os experimentos acreditava-se que, apesar das limitações citadas da técnica de subtração espectral, seria obtido algum aumento no reconhecimento de palavras isoladas. Contudo, após a análise dos resultados, foi possível observar que dependendo do nível SNR, é preferível utilizar o sinal ruidoso a processá-lo com a técnica de subtração espectral. Diante dos resultados obtidos na primeira etapa dos testes, foi levantada a hipótese de que, ao variar o número de coeficientes Mel-Ceptrais, poderia haver alguma mudança neste comportamento. No entanto, mesmo com essa variação, o aumento no reconhecimento não foi significativo mesmo levando em consideração os testes realizados em diferentes níveis Mel-cepstrais.

Diante do exposto, foi concluído que, para o reconhecimento automático da fala, o tratamento de ruído com a técnica de subtração espectral, com ou sem retificação de meia-onda, não é uma boa alternativa, poiso efeito da supressão de frequências importantes do sinal gera outro tipo de ruído, o ruído musical que é um tipo de flutuação aleatória do ruído.Este fenômeno pode ser explicado pela SNR que conduzem a picos degenerados no espectro processado. Quando o sinal é reconstruído no domínio do tempo, estes picos resultam em senoidais curtas cujas frequências variam de quadro para quadro. Em particular, o ruído musical é muito irritante durante as pausas de fala e em condições de baixa SNR quando não é mascarado pelo sinal de voz.[Esch2009]. O estado da arte indica que outras técnicas podem ser utilizadas para esta tarefa, como por exemplo, Filtragem de Wiener, Filtragem Adaptativa de Wiener e Wavelets. As próximas etapas deste trabalho consistirão na realização de experimentos semelhantes aos apresentados neste artigo, para as técnicas acima citadas.

Referências

ALCAIM, A. et al.. Frequência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no rio de janeiro. Revista da Sociedade Brasileira de Telecomunicações (SBrT), vol. 7, no. 1, pp. 23–41, 1992.

ALENCAR, Vladimir FabregasSurigué de; ALCAIM, Abraham. Lsf and lpc - derived features for large vocabulary distributed continuous speech recognition in brazilianportuguese. 2008.

MOLLER, A. Auditory Physiology. Academic Press, New York, 1983

BRESOLIN, A. A. et al. Digit recognition using wavelet and svm in brazilianportuguese. 2008.

BIING-HWANG, J.; FURUI, S., "Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication," *Proceedings of the IEEE* , vol.88, no.8, pp.1142,1165, Aug. 2000

ESCH, T.; VARY, P., "Efficient musical noise suppression for speech enhancement system," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* , vol., no., pp.4409,4412, 19-24 April 2009

FURUI, S., "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on* , vol.29, no.2, pp.254,272, Apr 1981

GARAU, G.; RENALS, S., "Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.16, no.3, pp.508,518, March 2008

GILBERT, M.; FENG, J. Speech and language processing over the web. Signal Processing Magazine, IEEE, vol. 25, no. 3, pp. 18–28, 2008.

JUANG, B. H.; RABINER, L. R..Hidden Markov Models for speech recognition.Technometrics, vol. 33, no. 3, pp. 251–272, 1991.

KANDA, A. Z. Estudo e implementação de uma técnica de redução de ruído em sinais de voz baseada na subtração espectral e em critérios psicoacústicos. Dissertação, UNESP, IlhaSolteira, Brazil, 2010.

LEE, K. et al..An overview of the Sphinx speech recognition system. 1990.

MITRA, V. et al.. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4117–4120, 2012.

OSTENDORF, M. et al..Speech segmentation and spoken document processing.IEEE_M_SP, 25(3):59–69, 2008.

PAGORARO, T. F..Algoritmos robustos de reconhecimento de voz aplicados a verificação de locutor. Dissertação, UNICAMP, Campinas, SP, Brazil, 2000.

PICONE, J.W., "Signal modeling techniques in speech recognition," *Proceedings of the IEEE* , vol.81, no.9, pp.1215,1247, Sep 1993

SILVA, L. A. d.. Filtros de kalman no tempo e frequência discretos combinados com subtração espectral. Dissertação, USP, São Carlos, SP, Brazil, 2007.

SILVA, V. R. da. et al.. Algoritmos para redução de ruído em sinais de áudio. 2007.

TEVAH, R. T.. Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro. Tese, 2006.

YARED, G. F. G.; VIOLARO, Fábio. Algoritmo para redução do número de parâmetros de modelos hmm utilizados em sistemas de reconhecimento de fala contínua. 2008.

YNOGUTI, C. A.; VIOLARO, F.. A brazilianportuguese speech database.2008.