Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 7º Congresso de Engenharia de Áudio
13ª Convenção Nacional da AES Brasil
26 a 28 de Maio de 2009, São Paulo, SP

Reconhecimento de Locutor baseado em Mascaramento Pleno em Freqüência por Oitavas

Sotero Filho, R. F. B. e de Oliveira, H. M.
Departamento de Eletrônica e Sistemas
Universidade Federal de Pernambuco - UFPE
Recife, Pernambuco, 50711-970, Brasil
rsotero@hotmail.com hmo@ufpe.br

RESUMO

Este artigo propõe um novo método de baixa complexidade computacional para reconhecimento de locutor, baseando-se em uma das propriedades-chave da percepção auditiva humana: o mascaramento acústico em freqüência. O vetor característico dos quadros do sinal de voz é representado pela média das amplitudes dos tons de mascaramento em cada oitava. Ambos os tipos de reconhecimento de locutor (de texto dependente e de texto independente) são estudados. Os resultados confirmam que o algoritmo proposto oferece um compromisso entre a complexidade e a taxa de identificações corretas, sendo atrativo para aplicações em sistemas embarcados.

ABSTRACT

This paper introduces a novel and low-complexity speaker identification technique. It is based on one of the key-properties of the human hearing perception: the auditory frequency masking. The feature vectors of voice frames are merely represented by the average amplitude of the greatest spectral samples within each octave. Both text-dependent and text-independent speaker recognition is investigated. Results corroborate a tradeoff between recognition efficiency and complexity of this kind of vocoder-based systems, which turns it attractive for embedded systems.

0 INTRODUÇÃO

Enquanto humanos, somos capazes de distinguir pessoas meramente ouvindo-as falar. Diferenças (ainda que sutis) de timbre, sotaque e/ou entonação, habilitam-nos a distinguir uma pessoa de outra apenas pela sua voz. Geralmente, curtos trechos de fala (2 a 3 segundos) são largamente suficientes para o reconhecimento de uma voz familiar.

A área de processamento de voz, que torna possível o reconhecimento de pessoas pela voz por meio de máquinas é chamada de "reconhecimento automático de locutor" (RAL). No RAL, determina-se a identidade de uma pessoa através da voz, com o propósito de controlar/restringir o acesso a redes, computadores, bases de dados, bem como restringir a disponibilização de informações confidenciais para pessoas não autorizadas, dentre várias outras aplicações [1].

Um sistema que trabalha com RAL calcula (por algum critério específico) a similaridade entre as características da voz do locutor que se deseja reconhecer, com as características de voz de um conjunto de locutores previamente armazenadas pelo sistema de reconhecimento.

O RAL divide-se em Verificação Automática de Locutor (VAL) e Identificação Automática de Locutor (IAL). Na VAL, faz-se uso de uma máquina para verificar

a identidade da voz de uma pessoa que a reivindicou [2]. Na literatura há outras denominações para a VAL, incluindo-se: verificação de voz, autenticação de locutor e autenticação de voz. Na VAL pode haver erros de dois tipos: a falsa aceitação (FA) de um locutor impostor, ou a falsa rejeição (FR) de um locutor verídico, [3], [4].

Na IAL não há a reivindicação de autenticidade. O sistema é que deverá decidir, dentre um determinado número *N* de locutores, qual o usuário correto ou se o mesmo é desconhecido dentre *N* possíveis locutores cadastrados [2]. A IAL pode ser implementada com rejeição ou sem rejeição. No primeiro caso, é estabelecido um limiar para cada usuário. Para o locutor ser considerado autêntico, a similaridade entre as características de sua elocução teste e as características extraídas de seu padrão deverá superar esse limiar. Em caso negativo, o locutor é considerado um impostor. Este trabalho é focado exclusivamente na Identificação Automática de Locutor sem rejeição.

O reconhecimento de locutor pode ser feito através do uso de um texto conhecido ou pode ser feito através de um texto arbitrário. No primeiro caso (reconhecimento dependente de texto), o texto ou frase é previamente conhecido pelo sistema que o utilizará para teste e para o treinamento. No segundo (reconhecimento independente de texto), não há especificação de texto. A tarefa de verificação é realizada com a comparação de um texto falado no momento do reconhecimento, com outro texto distinto, previamente gravado pelo sistema.

Recentes pesquisas na área de reconhecimento de locutor visam reduzir a complexidade computacional de métodos já existentes, e que invariavelmente requerem grande carga computacional para o processamento. O trabalho publicado recentemente, [5], baseado em LS-SVM (*The Least Square Support Vector Machine*), transforma um problema de programação quadrática, do convencional *Support Vector Machine* (SVM), num problema de programação linear, reduzindo assim a complexidade computacional. Outras publicações recentes procuram aprimorar o desempenho dos métodos de reconhecimento em ambientes ruidosos, como em [6] e [7].

Visando trabalhar com uma técnica de baixa complexidade e com alta simplicidade de implementação, este trabalho apresenta os resultados obtidos utilizando-se técnicas de processamento digital de sinais para a identificação automática de pessoas pela voz, baseado em uma técnica nomeada de "mascaramento em freqüência por oitava".

Inicialmente são introduzidas as técnicas adotadas para a realização do pré-processamento do sinal e extração das características representativas do sinal pré-processado. Posteriormente, o processo de reconhecimento é descrito. Concluindo, são analisados os resultados obtidos, com a implementação prática das técnicas descritas neste artigo para o reconhecimento de falantes.

1 AQUISIÇÃO DE SINAIS DE VOZ

O processo de identificação do locutor tem início com a gravação das elocuções para o processamento. Isso é realizado utilizando um microfone, cuja saída está conectada a uma placa de som instalada em um computador. Essa tem a função de converter o sinal analógico de voz em amostras igualmente espaçadas no tempo, a uma taxa que pode ser previamente escolhida.

Do teorema da amostragem de Shannon [8], sabe-se que para não haver perda de informação, o sinal banda limitada em f_m Hz deve ser amostrado a uma taxa de pelo menos $2f_m$ amostras eqüiespaçadas por segundo. Tipicamente, a energia de um sinal de voz é concentrada numa faixa de freqüência de até 5 kHz, ainda que a realização (pronúncia) típica de fonemas fricativos (e.g. /s/) possua substancial parte da energia espectral acima desta freqüência. No entanto, como isso ocorre apenas para sons de natureza ruidosa, eles contêm pouca informação sobre o locutor (que se concentra mais nos sons vocálicos). Diante disso, em concordância com o Teorema da amostragem, um valor aceitável para amostragem de um sinal de voz típico na aplicação em vista deveria ser em torno de 10 kHz [9]. O valor escolhido nesse trabalho foi o de 8 kHz, utilizando 16 bits de resolução e 1 canal, *Mono*.

2 PRÉ-PROCESSAMENTO DO SINAL DE VOZ

Após adquirirem-se os dados e convertê-los em amostras digitais, passa-se à fase do pré-processamento dos mesmos. Essa etapa compreende a pré-ênfase, a detecção de pontos extremos (*endpoints),* segmentação dos dados em quadros (*frames*) e janelamento.

2.1 Pré-ênfase

Devido a características fisiológicas do sistema de produção da fala, o sinal de voz irradiado pelos lábios apresenta uma atenuação de aproximadamente 6 dB/ oitava nas altas freqüências. O filtro de pré-ênfase serve para compensar esta atenuação, antes da análise espectral, melhorando a eficiência da análise [10]; sendo a audição menos sensível a freqüências acima de 1 kHz do espectro, a pré-ênfase amplifica esta área do espectro, auxiliando os algoritmos de análise espectral na modelagem dos aspectos perceptualmente importantes do espectro da voz [11]. A resposta em freqüência do filtro pode ser representada por:

$$H(z) = 1 - az^{-1}. \quad (1)$$

Neste caso, a saída da pré-ênfase *y(n)* está relacionada à entrada *x(n)* pela equação diferença [12]:

$$y(n) = x(n) - a.x(n-1) \quad (2)$$

para $1 \leq n < M$, em que *M* é o número de amostras do sinal amostrado *x(n)*, *y(n)* é o sinal pré-enfatizado e a constante "*a*" é normalmente escolhido entre 0,9 e 1. No trabalho foi adotado um valor de "*a*" igual a 0,95 [11].

2.2 Detecção de pontos extremos (*endpoints*)

A fim de reduzir o tempo de processamento, e evitar que o ruído de fundo que ocorra antes e depois do sinal de voz prejudique o desempenho do reconhecimento [13], far-se-á o uso de um algoritmo (*voice activity detection* – VAD), que detecta os pontos extremos do sinal. Esse algoritmo baseia-se na metodologia criada por Rabiner e Sambur em 1975 e faz uso de duas medidas do sinal de voz: a energia e a taxa de cruzamento do zero obtidas em janelas de 10 ms

de duração do sinal. Um intervalo de 100 ms no início da elocução (10 janelas) é utilizado para efetuar uma estatística do ruído de fundo [14].

2.3 Segmentação dos dados em quadros e Janelamento

Após a detecção dos pontos extremos, o sinal de voz deve ser particionado em pequenos segmentos (*frames)* bem definidos, com o propósito de se obter trechos de voz razoavelmente assumidos como estacionários. Isso porque, sendo o sinal de voz um processo estocástico, e sabendo-se que o trato vocal muda de forma muito lentamente na voz contínua, muitas partes da onda acústica podem ser assumidas como estacionárias num intervalo de curtíssima duração (entre 10 e 40 ms). Este intervalo caracteriza o tamanho da janela a ser usada [15]. Neste trabalho, o tamanho da janela adotada será de 20 ms, um valor típico de muitas aplicações envolvendo voz.

O janelamento do sinal tem o objetivo de amortecer o efeito do "fenômeno Gibbs" [10], [16] que surge devido à descontinuidade das janelas [15].

Para o contexto da produção da voz, as características apresentadas, referentes ao janelamento de *Hamming*, mostram que este tipo de janela é mais eficiente quando comparada às janelas Retangular e de *Hanning*, com uma aproximação da janela ideal [16]. Assim sendo, essa foi a janela utilizada neste trabalho.

3 METODOLOGIA EMPREGADA

A idéia proposta baseou-se em umas das propriedades psico-acústicas da audição humana: o mascaramento auditivo ou "audibilidade diminuída de um som devido à presença de outro", podendo este ser em freqüência – foco do nosso trabalho – ou no tempo. O mascaramento auditivo em freqüência ocorre quando um som que normalmente poderia ser ouvido é mascarado por outro, de maior intensidade, que se encontra em uma freqüência próxima. Ou seja, o limiar de audição é modificado (aumentado) na região próxima à freqüência do som que causa a ocorrência do mascaramento, sendo que isto se deve à limitação da percepção de freqüências do sistema auditório humano.

Em função deste comportamento, o que método de reconhecimento proposto fará, *a priori*, é identificar casos de mascaramento em freqüência no espectro do sinal particionado em oitavas, e descartar sinais que "não seriam audíveis" devido a este fenômeno.

A tendência predominante dos padrões de reconhecimento existentes em utilizar coeficientes cepstrais e mel-cepstrais [16] para caracterizar um quadro de voz, não será aqui adotada. Em vez disso, utilizaremos a fração média das amplitudes das freqüências de mascaramento por oitava, como uma representação do padrão de voz. Essa nova abordagem reduz significativamente o volume de dados para processamento.

As algoritmos desenvolvidos para a extração das características do quadro de voz, geração e comparação dos padrões foram todos escritos na linguagem MATLAB® por ser uma linguagem muito difundida nos meios acadêmicos e de fácil implementação.

A seguir a metodologia abordada é descrita.

3.1 Extração das características do quadro de voz

O sinal gravado e amostrado (a uma taxa de 8 kHz) passará pelas etapas descritas no item 2, ou seja, da pré-ênfase, detecção dos pontos extremos, segmentação e janelamento. Posteriormente, para cada segmento do arquivo de voz janelado, será aplicada uma FFT de comprimento 160 (número de amostras contidas em um quadro de 20 ms de voz), obtendo-se assim a representação no domínio da freqüência do sinal, para cada quadro. Subseqüentemente, o espectro da magnitude do sinal é dividido em oitavas. A primeira oitava correspondendo à faixa de freqüências de 32 Hz – 64 Hz, a segunda indo de 64 Hz – 128 Hz, e assim por diante, até a sétima que corresponde à faixa de 2048 Hz a 4096 Hz.

Como se está fazendo uso de uma taxa de amostragem de 8 kHz, cada amostra da magnitude do espectro corresponderá a uma amostra espectral múltipla de 50 Hz, sendo que a primeira amostra irá representar a componente DC de cada quadro de voz. Já que as raias espectrais caminham a passos de 50 Hz, a primeira oitava (de 32 Hz a 64 Hz), será representada pela amostra espectral de 50 Hz, a segunda oitava (64 Hz a 128 Hz) pela amostra de 100 Hz, a terceira (de 128 Hz a 256 Hz) pelas amostras de 150 Hz, 200 Hz e 250 Hz, e assim por diante.

Tabela 1 – Número de freqüências estimadas pela DFT de comprimento 160 em cada oitava do espectro vocal.

Oitava (Hz)	*# amostras espectrais/oitava*
32 - 64	1
64 - 128	1
128 - 256	3
256 - 512	5
512 - 1024	10
1024 - 2048	20
2048 - 4096	39

Terminado esse procedimento inicial, o algoritmo irá agora buscar em cada oitava, em todos os sete sub-bandas de voz do sinal, o ponto da FFT de maior magnitude, i.e., aquele que irá (potencialmente) mascarar os demais. Essa amostra espectral passará a ser o único representante dentro de cada oitava (por opção de complexidade reduzida). As demais serão descartadas, assumindo valor espectral nulo. O total de 80 freqüências oriundas da estimativa da DFT com N=160 é reduzido para 7 sobreviventes (retendo menos do que 5% das componentes espectrais). Portanto, cada quadro, agora, será representado, no domínio freqüencial, por 7 tons puros de mascaramento auditivo, um para cada oitava. Esta técnica é denominada aqui de mascaramento pleno de freqüência.

Definindo o vetor inicial de amostras espectrais, no i-ésimo quadro de voz, por $oct_j^{(i)}$ em que j representa o índice da oitava, tem-se:

$$oct_j^{(i)} = \left[a_{j,1}^{(i)}\, a_{j,2}^{(i)} a_{j,3}^{(i)} \ldots a_{j,N_j}^{(i)}\right], \quad \begin{cases} i = 1,2 \ldots, n \\ j = 1,2\ldots,7 \end{cases} \qquad (3)$$

sendo, $a_{j,k}^{(i)}$ a amplitude do k-ésimo ponto da FFT, na janela i e oitava j e N_j o número de amostras da j-ésima oitava.

Aplicando-se o procedimento de busca da amostra espectral de maior magnitude, vamos obter um novo vetor

$new\_oct_j^{(i)}$ sintetizado contendo N_j-1 zeros, e a única componente da amostra de mascaramento espectral correspondente ao $max\ (a_{j,k}^{(i)})$:

$$new\_oct_j^{(i)} = [0\ 0\ ...\ max\ (a_{j,k}^{(i)})...\ 0],\ k = 1,2...\ N_j. \quad (4)$$

A Figura 1 mostra o módulo do espectro de um quadro, de 20 ms, de uma locução usada para teste, antes e depois da simplificação por tons de mascaramento psico-acústico.

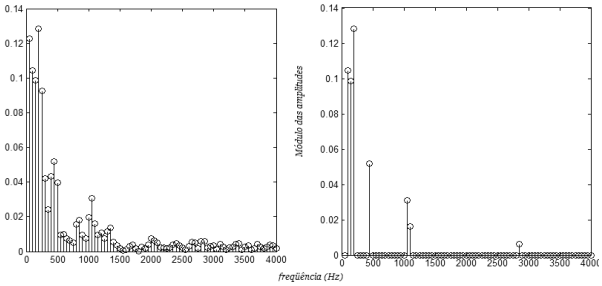

Figura 1 Representação do espectro de freqüência de um quadro de voz, para antes e depois do processo de mascaramento auditivo.

Obtidos todos os vetores $new\_oct_j^{(i)}$, o algoritmo obtém, para cada oitava, uma matriz M_j cujas linhas são formadas por todos os n vetores $new\_oct_j^{(i)}$ do arquivo. Esse procedimento será útil para calcular as médias dos "tons" de mascaramento.

$$M_j = (m_{j,k}) = \begin{bmatrix} new\_oct_j^{(1)} \\ new\_oct_j^{(2)} \\ new\_oct_j^{(3)} \\ \vdots \\ new\_oct_j^{(n)} \end{bmatrix} \quad (5)$$

$$= \begin{bmatrix} 0 & 0 & \max(a_{j,k}^{(1)}) & \dots & 0 \\ 0 & 0 & 0 & \vdots & \max(a_{j,k}^{(2)}) \\ \max(a_{j,k}^{(3)}) & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \max(a_{j,k}^{(N_j)}) & 0 & \dots & 0 \end{bmatrix}$$

Calculando-se a média de cada coluna da matriz M_j, obtém-se a participação média de cada amostra espectral de mascaramento (múltiplos de 50 Hz), no sinal de voz, resultando no vetor:

$$\overline{m_j} = [m_{j,1}\ m_{j,2}\ ...\ m_{j,Nj}], \quad (6)$$

em que, $m_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \max(a_{j,k}^{(i)})$, e k representa o índice no qual existam amostras espectrais de mascaramento.

Em seguida, todas as componentes do vetor, $\overline{m_j}$ são somadas. Essa soma representará a participação média dos "tons" de mascaramento dentro de sua respectiva oitava.

$$s_j = \sum_{k=1}^{N_j} m_{j,k}\ . \quad (7)$$

Esses s_j assim definidos formarão o vetor s_{total}:

$$s_{total} = [s_1\ s_2 \dots s_7]. \quad (8)$$

Os parâmetros obtidos pela etapa anterior são diretamente proporcionais aos níveis de energia dos sinais coletados, fator que pode deturpar a classificação incorretamente. Para realizar a normalização dessas amplitudes, faz-se a divisão do vetor s_{total} pela soma de todas as suas componentes.

Normalizando, o vetor s_{total} encontra-se, enfim, o vetor característica do sinal de voz, com apenas 7 componentes, representantes do número de oitavas, o qual será usado para a comparação com as locuções testes:

$$s_{norm} = \frac{1}{\sum_{j=1}^{7} s_j}[s_1\ s_2\ ...\ s_7]. \quad (9)$$

3.2 Geração dos padrões de locutores

A geração do padrão de cada locutor é feita obtendo a média de todos os vetores representantes das características do sinal de voz, das elocuções reservadas para o treinamento.

3.3 Comparação dos padrões de voz

Como última etapa do processo de identificação, tem-se a comparação entre dois vetores. A comparação é realizada através do cálculo da distorção entre eles. Há várias medidas de distorção entre vetores que podem ser utilizadas em reconhecimento de locutor. A medida de distorção mínima ou euclidiana, a medida mais conhecida, foi aquela utilizada. Simulações de desempenho pela alteração das métricas de comparação dos locutores precisam ser conduzidas, a fim de selecionar a mais adequada, i.e., aquela de melhor compromisso complexidade versus taxa de reconhecimento. A seleção do locutor é realizada com base na técnica simples de *template matching* via distância euclidiana entre o vetor de característica e os vetores armazenados para os locutores cadastrados. A Figura 2 no anexo ilustra o algoritmo de reconhecimento de locutor proposto neste trabalho.

4 RESULTADOS

Foram realizados dois tipos de testes. No primeiro deles, a identificação dos locutores é feita fazendo uso de uma mesma frase padrão para todos os locutores (reconhecimento dependente de texto). No segundo caso, a identificação é feita com textos escolhidos aleatoriamente no momento da gravação (reconhecimento independente de texto). Todas as gravações foram realizadas com o mesmo microfone, numa sala que não teve nenhuma preparação especial destinada à redução de ecos ou mesmo a eliminação total de ruído de fundo. Nos experimentos realizados a eficiência do algoritmo foi também testada na ausência da pré-ênfase. Os resultados são comentados a seguir.

4.1 IAL Dependente de Texto

Para a realização desse teste faz-se necessário o pré-conhecimento de textos ou frases. Duas frases são consideradas adequadas para reconhecimento de locutor, por apresentarem grande quantidade de fonemas nasalados e vocalizados [15]. São elas: "O prazo tá terminando" e "Amanhã ligo de novo". A segunda opção foi à selecionada para realização dos testes audiométricos.

Foram gravadas 40 repetições para 10 locutores diferentes (7 do sexo masculino e 3 do sexo feminino), das quais 20 serão utilizadas para a geração do padrão de cada locutor e outros 20 serão utilizados para a comparação dos padrões, totalizando 400 elocuções. Os resultados dos testes seguem na Tabela 2.

Tabela 2 – Resultado dos testes para o reconhecimento de locutor dependente de texto.

Pré-ênfase	*Identificações corretas*	*Identificações incorretas*	*Eficiência*
Sim	174	26	87,0 %
Não	183	17	91,5%

Como se pode observar pela Tabela 2, na ausência da pré-ênfase o algoritmo tornou-se mais eficiente.

4.2 IAL Independente de Texto

Nesse teste, utilizaram-se oito textos, escolhidos aleatoriamente, de aproximadamente 10 segundos de duração, para 12 locutores diferentes. Quatro desses textos foram usados para a geração do padrão de cada locutor. Os outros quatro textos foram utilizados para as comparações dos padrões. Os resultados são sumarizados na Tabela 3.

Tabela 3 – Resultado dos testes para o reconhecimento de locutor independente de texto.

Pré-ênfase	*Identificações corretas*	*Identificações incorretas*	*Eficiência*
Sim	39	9	81,25 %
Não	44	4	91,66 %

5 DISCUSSÃO E CONCLUSÕES

Ficou constatado nesse artigo que o mascaramento em freqüência fazendo uso de um único ponto da FFT sobrevivente por oitava pode ser útil no reconhecimento de locutor. A síntese do sinal de áudio proveniente de um vocoder contendo apenas o espectro "ultra-simplificado" (com único sobrevivente por oitava, e.g. Fig.1) fornece um sinal perfeitamente inteligível, a partir do qual se reconhece facilmente o falante. Assim, a despeito da qualidade "metálica e artificial" da voz sintética (vide arquivo anexo *sotero-reconhecimento-2.wav*), típica de vocoders, as informações suficientes para o reconhecimento não são destruídas. O processo descrito tem como atrativo a simplicidade, pois cada "padrão de voz" é resumido em um único vetor de sete componentes associadas às oitavas distintas. Adicionalmente, o classificador padrão usando cadeias de Markov escondidas (HMM) é substituído pela técnica simples de *template matching* via distância euclidiana entre os vetores. Foi observada uma maior taxa de acertos do algoritmo para o reconhecimento dependente de texto. De modo surpreendente para as expectativas iniciais, constatou-se que o filtro de pré-ênfase comprometeu um pouco a eficiência das identificações. De fato, ao enfatizar componentes espectrais mais sensíveis a distorções e ruído, obtém-se melhor qualidade e um sinal de voz mais natural. Porém, os resultados indicam que tais componentes não são cruciais no reconhecimento. Os resultados preliminares apresentados são promissores. Mesmo que a taxa de reconhecimentos corretos nesta versão inicial seja inferior a 95% − restringindo seu uso imediato em algumas aplicações comerciais– aprimoramentos simples podem ser introduzidos (e.g. considerar mais de um sobrevivente em bandas de maior freqüência) visando reduzir a taxa de falhas. Este tópico encontra-se atualmente sob investigação, além de uma análise do comportamento do vetor de características para diferentes falantes, ou seja, quão bem ele consegue "espalhar" timbres diferentes no espaço de características (algo como a característica de decorrelação dos coeficientes MFCC).

A técnica de mascaramento espectral pleno "lembra" a abordagem de estatística mínima suficiente [17]. É como se fossem descartadas as informações espectrais irrelevantes no processo de estimação. Detalhes práticos suplementares merecem investigação. A transformada de comprimento N=160 usa bases mistas e visando simplicidade de implementação de *hardware* ou DSP, pode-se alterar a duração da janela. Com janelas de 32 mseg (ou 16 mseg) é possível usar o algoritmo *butterfly* (radix-2) [16], restando investigar o impacto na eficiência.

Uma comparação rigorosa entre a complexidade e o compromisso com o desempenho do algoritmo de reconhecimento do locutor entre diferentes técnicas IAL não foi realizada. Porém o principal mérito desta nova abordagem é oferecer uma taxa de reconhecimento razoável, porém demandando uma complexidade computacional substancialmente inferior àquela requerida por outras técnicas consagradas (e.g., HMM, redes neurais, quantização vetorial etc.). Vale lembrar que as complexidades (por janela de 20 ms) exigidas pela FFT (N=160) e algoritmo de seleção do maior elemento de uma lista (Tabela 1) são desprezíveis para os comprimentos requeridos. A adaptação do método para uso de wavelets discretas [8], tornando-o mais atrativo, também se encontra em investigação. Outro aproveitamento possível deste algoritmo é nos casos em que a base de locutores é demasiadamente extensa. Este método rápido pode ser aplicado, selecionando um locutor provável, incluído em uma subclasse de locutores potenciais. Este é então eliminado da base original, repetindo o processo de forma a escolher um segundo locutor potencial. O procedimento é iterado até gerar um número pré-estabelecido de locutores potenciais (base reduzida). Esta aplicação prévia não requer taxas de acerto excessivamente altas, 90% é bastante razoável. Um método sofisticado (alto custo computacional e alta eficiência) é aplicado para identificar o locutor dentro desta base reduzida. Outra situação de potencial interesse para este método é no monitoramento em tempo real de telefonemas em prédios (empresas, repartições, etc.) que possuem centrais telefônicas. Com centenas de ligações simultâneas e diferentes ramais, como selecionar gravações (autorizadas) de conversações envolvendo indivíduo sob suspeição? Supõe-se disponível um trecho previamente gravado (e.g., primeiro contato de um seqüestrador, chantagista, corrupto, terrorista etc.) para constituir a informação de treinamento do locutor alvo. Neste caso, taxas de FA e FR aceitáveis podem ser maiores do que em aplicações comerciais típicas. Assim, situações em tempo real – nas quais há parca disponibilidade de recursos (como em sistemas embarcados) – esta técnica pode se tornar bastante atrativa.

AGRADECIMENTOS- Os autores agradecem a revisores anônimos por sugestões valiosas para aperfeiçoar a apresentação deste trabalho.

6 REFERÊNCIAS

[1] Oliveira, M.P.B., "*Verificação Automática de locutor, Dependente do Texto, Utilizando Sistemas Híbridos MLP/HMM*" Dissertação de Mestrado – Instituto Militar de Engenharia / IME - 2001.

[2] Campbell Jr, J.P., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, September, vol.85, n 9. (1997).

[3] Atal, B.S. "Automatic Recognition of Speakers from Theirs Voices", *Proceedings of the IEEE*, April, vol 64, n 64, pp 460-475 (1976).

[4] Rosemberg, A.E. "Automatic Speaker Verification: A Review", *Proceedings of the IEEE*, April vol. 64, n 4, pp. 475-487 (1976).

[5] Dan, Z. Zheng, S. Sun S. and Dong, R. "Speaker Recognition based on LV-SVM" – *The 3rd International Conference on Innovative Computing Information and Control* (ICICIC'08), 2008.

[6] Wang, N. Ching, P.C. Zheng N.H. and Tan Lee – "Robust Speaker Recognition Using Both Vocal Source and Vocal Tract Features Estimated from Noisy Input Utterances", *IEEE International Symposium on Signal Processing and Information Technology*, 2007.

[7] Shao Y. and Wang D., "Robust Speaker Recognition Using Binary Time-Frequency Masks"- *IEEE International Conference on Acoustic, Speech and Signal Processing* 2006 (ICASSP 2006).

[8] De Oliveira, H.M., *Análise de sinais para Engenheiros – Uma abordagem via Wavelets*, Brasport, 2007.

[9] Diniz, S.S. "*Uso de Técnicas Neurais para o Reconhecimento de Comandos à Voz*". Dissertação de Mestrado, IME, Rio de Janeiro, 1997.

[10] Rabiner, L.R.; Schafer, R.W. *Digital processing of speech signals*. New Jersey: Prentice Hall, 1978.

[11] Silva, D.D.C, "*Desenvolvimento de um IP Core de Pré-Processamento Digital de Sinais de Voz para Aplicações em Sistemas Embutidos*", Dissertação de Mestrado, UFCG, Campina Grande, 2006.

[12] Petry, A., Zanuz, A. e Barone, D.A.C., "Reconhecimento Automático de Pessoas pela Voz usando técnicas de Processamento Digital de Sinais. *SEMAC*, *Semana de Computação da UNESP*, 2000.

[13] Rabiner, L.; Juang, B.H., *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall, 1993. 507p.

[14] Paranaguá, E.D.S., "*Reconhecimento de Locutor Utilizando Modelos de Markov Escondidos Contínuos*", Dissertação de Mestrado, IME, Rio de Janeiro-RJ, 1997.

[15] Bezerra, M.R. "*Reconhecimento Automático de Locutor para Fins Forenses, Utilizando Técnicas de Redes Neurais*", Dissertação de Mestrado, IME, Rio de Janeiro, 2001.

[16] Oppenheim, A.V. & Schafer, R.W. *Digital-Time Signal Processing,* Prentice-Hall, Inc, Englewood Cliffs, New Jersey, 1989.

[17] Ferguson, T., *Mathematical Statistics: a Decision Theoretic Approach*, New York, Academic Press, 1967.

ANEXO

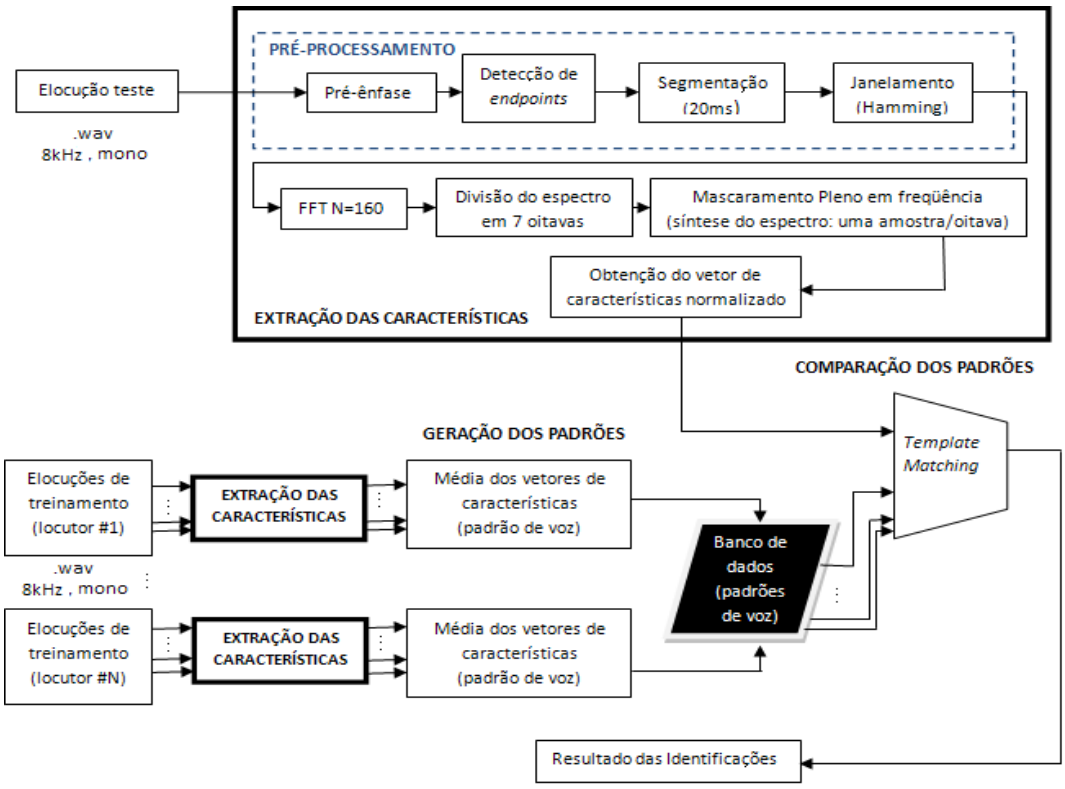

Figura 2 Diagrama de blocos de um sistema de reconhecimento de locutor com base no mascaramento de freqüências por oitava.