UTILIZAÇÃO DE TÉCNICAS DE PROCESSAMENTO DIGITAL DE SINAIS PARA A IDENTIFICAÇÃO AUTOMÁTICA DE PESSOAS PELA VOZ

Adriano Petry, Adriano Zanuz, Dante Augusto Couto Barone
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre - RS
{adpetry, zanuz, barone}@inf.ufrgs.br

RESUMO

Este trabalho apresenta os resultados obtidos com a utilização de técnicas de processamento digital de sinais para a identificação automática de pessoas pela voz. São descritas as metodologias utilizadas nas etapas de aquisição do sinal vocal, detecção automática dos limites das palavras, pré-processamento, extração das características e classificação dos padrões vocais. Obteve-se com os testes realizados uma performance de 90% de aceitação das pessoas com direito de acesso e 96% de rejeição de impostores.

ABSTRACT

This work presents the results obtained using digital signal processing techniques to accomplish automatic people identity verification through speech. It is described the methodologies used in speech signal acquisition, automatic word limits detection, pre-processing, features extraction and vocal pattern recognition. The results have been provided 90% for registered people acceptation rate and 96% for impostors rejection rate.

1 INTRODUÇÃO

A área de processamento da fala possui uma grande variedade de aplicações. Por exemplo, sinais de voz podem ser analisados de uma forma particular para se obter uma melhor compressão de dados visando sua transmissão. Um uso mais recente para o processamento de voz é o reconhecimento automático de voz (RAV), onde uma pessoa pode interagir com máquinas através da fala. Outra aplicação são os sistemas de segurança, que podem analisar uma amostra de voz e identificar a pessoa que a produziu. Esses sistemas utilizam tecnologia para o reconhecimento automático de locutores (RAL).

Em aplicações de RAL, o sinal de voz não é apenas amostrado e classificado utilizando técnicas de reconhecimento de padrões. A identidade de um locutor está intrinsecamente associada às características fisiológicas e comportamentais da voz. As principais informações contidas no sinal devem ser filtradas do sinal amostrado. Para obter esta informação representativa, a voz é primeiramente pré-processada e então são utilizados algoritmos para extração dos parâmetros do sinal vocal. O conjunto de parâmetros de uma amostra de voz compõe um padrão, que pode ser classificado.

Este trabalho apresenta os resultados obtidos utilizando-se técnicas de processamento digital de sinais para a identificação automática de pessoas pela voz. A primeira seção mostra a metodologia empregada para realizar a aquisição automática do sinal vocal, e simultânea detecção dos limites das palavras. A seguir, são mostradas as técnicas usadas para a realização do pré-processamento do sinal e extração das características representativas do sinal pré-processado. Após, uma técnica para classificação dos padrões é mostrada. Concluindo, são analisados os resultados obtidos com a implementação prática das técnicas descritas anteriormente para o reconhecimento de pessoas pela voz.

2 AQUISIÇÃO DO SINAL DE VOZ E DETECÇÃO AUTOMÁTICA DOS LIMITES DAS PALAVRAS

A aquisição do sinal de voz é realizada utilizando-se um microfone conectado a um filtro passa baixas, cuja saída está ligada a uma placa de som instalada em um computador pessoal. A freqüência de corte do filtro é de 4,5KHz. A placa de som transforma o sinal analógico filtrado em amostras digitais, a uma taxa de 11025Hz, com resolução de 16 bits por amostra. Tais amostras são processadas por um algoritmo de detecção de limites de palavras (Luft, 1991), que identifica quando uma palavra iniciou e terminou, gravando os dados da palavra em um arquivo.

O parâmetro de medida utilizado pelo algoritmo de detecção de limites de palavras é a energia média contida em um bloco. Um bloco é um conjunto composto por um número fixo de amostras de voz. O cálculo da energia média pode ser visto na equação 1. O início de uma possível palavra é considerado a partir do primeiro bloco onde a energia ultrapassa um limiar pré-determinado. Para que os blocos seguintes constituam realmente um início de palavra, a energia deles deve permanecer acima do limiar durante um período de tempo pré-estabelecido chamado tempo de início. Caso a energia de um bloco caia abaixo do limiar antes do fim deste período, as amostras até então armazenadas são descartadas e a procura pelo início de palavra recomeça. Caso a energia média dos blocos se mantenha superior ao limiar por todo o tempo de início, começamos a procura pelo fim da

palavra. O método de detecção do fim de palavra é idêntico ao método de detecção de início, descrito anteriormente. A diferença está na procura de blocos com energia inferior à estabelecida no limiar. Da mesma forma, os blocos devem permanecer com energia inferior ao limiar por um certo período de tempo para que se considere atingido o fim de palavra. Tal período de tempo é chamado tempo de fim. Os blocos que compõem o tempo de fim não são considerados como componentes da palavra. O algoritmo de detecção de limites de palavra é mostrado na figura 1.

$$En = \frac{1}{N}\sum_{k=0}^{N-1}\left|x(k)\right| \quad (1)$$

onde *En* é a energia média de um bloco composto por *N* amostras de voz, e *x(k)* é a amplitude da k-ésima amostra de voz de um bloco.

O sistema faz a detecção dos limites das palavras de forma automática. Assim, utilizou-se o algoritmo descrito para realizar tal tarefa, que permite ao usuário, a qualquer momento, iniciar uma locução sem a necessidade de "avisar" o sistema.

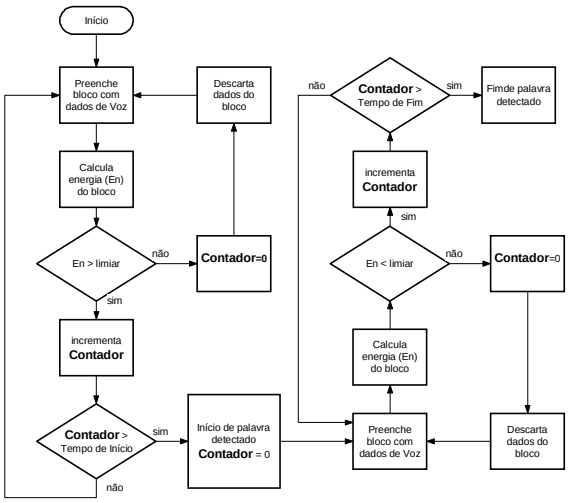


FIGURA 1: Algoritmo para detecção de limites de palavra.

Foram utilizados blocos de 32 bytes ou 16 amostras de voz, tempo de início de 70ms ou 48 blocos, tempo de fim igual a 150ms ou 103 blocos.

3 PRÉ-PROCESSAMENTO DO SINAL

Após a aquisição do sinal de voz, é realizado um pré-processamento nas amostras a fim de prepará-las para a extração de suas características (Papamichalis, 1987; Kil e Shin, 1996; Picone, 1993). Tais características são utilizadas no algoritmo de reconhecimento de padrões. O pré-processamento é composto pelas etapas de pré-ênfase, divisão do sinal em *frames* e janelamento.

3.1 Pré-ênfase

A pré-ênfase objetiva eliminar uma tendência espectral de aproximadamente -6dB/oitava na fala irradiada dos lábios. Essa distorção espectral não traz informação adicional e pode ser eliminada através da aplicação de um filtro, de resposta aproximadamente +6dB/oitava, que ocasionaria um nivelamento no espectro. Para um sistema digital, tais pré-ênfases podem ser implementadas como um circuito analógico, o qual precede o filtro e o amostrador, ou como uma operação digital no sinal amostrado, através de um filtro FIR de primeira ordem. O efeito de ascensão de +6dB/oitava pode ser obtido pela diferenciação da entrada. A equação 2 descreve o pré-enfatizamento realizado no sinal amostrado.

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

para $1 \le n < M$, onde *M* é o número de amostras do sinal amostrado *x(n)*, *y(n)* é o sinal pré-enfatizado e o parâmetro constante "*a*" é usualmente escolhido entre 0,9 e 1. Foi utilizado "*a*" igual a 0,95.

3.2 Divisão do Sinal em Frames e Janelamento

Em todas as aplicações práticas de processamento de sinais, é necessário trabalhar com "pequenas porções" ou *frames* do sinal, a não ser que o sinal seja de curtíssima duração. Isso é verdade especialmente se estivermos utilizando técnicas de análise convencionais de sistemas lineares invariantes no tempo (LTI). Nesse caso é necessário selecionar uma porção de sinal que possa ser razoavelmente assumida como estacionária. Formalmente, definimos um *frame* de voz como sendo o produto de uma janela discreta *w(n)* de tamanho *L* e terminando no tempo "*l*", com relação à seqüência de voz discreta (pré-enfatizada) *y(n)*, resultando na seleção de um pedaço do sinal pré-enfatizado, como mostra a equação 3.

$$f(n) = y(n).w(l-n) \quad (3)$$

onde *f(n)* é um *frame* do sinal pré-enfatizado *y(n)*, e *w(n)* é a janela aplicada.

A janela de hamming foi utilizada, por apresentar características espectrais interessantes e por atenuar a transição entre *frames* adjacentes. Sua descrição matemática pode ser vista na equação 4. As janelas usualmente são sobrepostas entre si, para que a variação dos parâmetros entre janelas sucessivas seja mais gradual. Foram utilizadas janelas de tamanho igual a 330 amostras ou 30ms, aplicadas a cada 10ms de sinal.

$$w(n) = \begin{cases} 0 & n < 0 \\ 0{,}54 - 0{,}46\cos\left(\dfrac{2\pi n}{330-1}\right) & 0 \le n < L \\ 0 & n \ge L \end{cases} \quad (4)$$

4 EXTRAÇÃO DOS PARÂMETROS

A partir das amostras que compõem um *frame* de voz, são utilizadas técnicas para obtenção dos coeficientes representantes do mesmo. Nesse trabalho, são utilizados dois tipos de coeficientes: os cepstrais e os mel-cepstrais. Esses coeficientes proporcionam uma redução no volume de dados, sem perda significativa de informação útil. Essa redução de dimensão nos fornece sistemas reconhecedores mais robustos e eficientes.

4.1 Coeficientes Mel-cepstrais

A análise homomórfica foi desenvolvida como uma forma de desconvoluir dois sinais. Análise homomórfica é considerada útil para o processamento da fala, pois oferece uma metodologia para a separação do sinal de excitação da resposta impulsiva do trato vocal. No modelamento matemático para a produção do sinal vocal (Deller, Proakis e Hansen, 1987; Rabiner e Juang, 1993), temos que um *frame f(n)* do sinal vocal (pré-enfatizado) *y(n)* pode ser escrito como o produto da convolução do sinal de excitação *u(n)* com a resposta impulsiva do trato vocal *h(n)*, como é visto na equação 5.

$$f(n) = u(n) \otimes h(n) \quad (5)$$

A representação no domínio freqüência desse processo através da aplicação da transformada de Fourier, transforma a operação de convolução em multiplicação. Aplicando-se a função logarítmica, transformamos a multiplicação na soma (ou sobreposição) de sinais, como mostra a equação 6.

$$\log\left(F\{f(n)\}\right) = \log\left(F\{u(n)\}\right) + \log\left(F\{h(n)\}\right) \quad (6)$$

onde $F\{\bullet\}$ representa a aplicação da transformada discreta de Fourier (DFT).

Aplicando-se a transformada inversa nesse sinal tem-se o *cepstrum* ou coeficientes cepstrais do sinal de voz. Sabe-se que a parcela do sinal de excitação varia mais rapidamente que a resposta impulsiva do trato vocal, então os dois sinais poderiam ser separados no domínio cepstral. Na prática, são utilizados apenas os primeiros coeficientes componentes do cepstrum. Tais coeficientes contém a informação relativa ao trato vocal, que está intimamente relacionada com o locutor.

Atualmente, muitos sistemas utilizam os coeficientes mel-cepstrais (*mel frequency cepstral coefficients* – MFCC) para o reconhecimento de voz (Picone, 1993; Deller, Proakis e Hansen, 1987). A análise mel-cepstral vem progressivamente substituindo a forma tradicional de utilização de parâmetros cepstais previamente descritos. A diferença entre o cálculo dos coeficientes cepstrais e dos coeficientes mel-cepstrais está na aplicação de um banco de filtros digitais ao espectro real do sinal, antes da aplicação da função logarítmica. Tais filtros não estão linearmente espaçados no domínio freqüência. O objetivo de tais filtros é uma tentativa de aproximar a resposta humana a sinais sonoros. Mel é a unidade de medida de freqüências ou picos percebidos de um tom. Essa unidade não corresponde linearmente à freqüência física bem como, aparentemente, o ouvido humano também não o faz. É possível traçar uma comparação entre a freqüência real (medida em Hz) e a freqüência percebida (medida em mels). Logo, o espaçamento dos filtros digitais deve respeitar a escala de freqüências percebidas (escala *Mel*). Podemos definir uma função para mapeamento da freqüência acústica *f* (em Hz) para uma escala de freqüências percebidas *Mel* (em mels) como mostra a equação 7.

$$Mel = 2595.\log_{10}\left(1 + \frac{f}{700}\right) \quad (7)$$

Uma forma de aplicar os filtros digitais de forma espaçada segundo a escala *Mel* seria primeiramente mapear as freqüências acústicas (em Hz) para a escala de freqüências percebidas (em mels), e após aplicar um banco de filtros espaçados linearmente nesse domínio (domínio mel). Isso corresponderia à aplicação de filtros digitais espaçados segundo a escala mel, no domínio das freqüências reais.

O processo de obtenção dos MFCC é matematicamente descrito na equação 8.

$$c(n) = \sum_{k=1}^{K} \log|S(k)|.\cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (8)$$

para $0 \le n < P$, onde *c(n)* é o n-ésimo coeficiente mel-cepstral, *P* é o número de coeficientes mel-cepstrais extraídos, *K* é o número de filtros digitais utilizados e *S(k)* é o sinal de saída do banco de filtros digitais. Usualmente é utilizada a transformada inversa do coseno (ICT) para obtenção dos MFCC ao invés da transformada inversa discreta de Fourier. A utilização dessa transformada inversa reforça a informação útil nos MFCC iniciais.

A aplicação de uma janela de ponderação aos coeficientes mel-cepstrais obtidos, chamada janela de janela de *liftro*, é amplamente empregada em reconhecimento de voz (Deller, Proakis e Hansen, 1987), apresentando bons resultados. Seu objetivo é enfatizar componentes com maior informação espectral útil. Tal ponderação *ln)* é dada pela equação 9.

$$l(n) = 1 + \frac{Q}{2}\sin\left(\frac{n\pi}{Q}\right) \quad (9)$$

para $0 \le n < P$, onde *Q* é uma constante chamada de coeficiente de liftro (usualmente igual a 22) e *P* é o número de coeficientes previamente extraídos.

5 CLASSIFICAÇÃO DOS PADRÕES

A partir dos coeficientes extraídos, procedemos ao reconhecimento de padrões. Neste trabalho, a técnica utilizada para o reconhecimento de padrões é a Quantização Vetorial Multisecção. Um padrão é o conjunto de vetores oriundos de uma locução. Denomina-se vetor o conjunto de coeficientes extraídos de um *frame* de voz. Existem três etapas distintas no processo de classificação de padrões através da quantização vetorial (Makhoul, Roucos e Gish,1985; Gray,1984; Burton,1987; Rosenberg e Soong,1987; Soong, Rosenberg e Juang,1987), que são a *geração de codebook*, a *quantização* de um padrão desconhecido, e a *comparação* ou medida de distorção. Adicionalmente, é mostrado como a técnica de quantização vetorial pode ser melhorada utilizando-se um artifício chamado *multisecção*.

Tendo-se um universo de vetores p-dimensionais, são estabelecidos vetores (também p-dimensionais) representantes desse universo. Esses vetores representantes são chamados centróides. O conjunto de todos os centróides é chamado de *codebook***.** Os centróides são em número muito menor que o número de vetores que compõem o universo. Assim, podemos discretizar qualquer vetor p-dimensional em um entre os centróides previamente treinados. Quantização é a conversão de um vetor de entrada em um código relacionado a vetores de mesma dimensão previamente treinados (centróides).

5.1 Geração de Codebook

O objetivo da etapa de geração de *codebook* é estabelecer referências para posterior classificação. Nessa etapa, é criado um *codebook* por locutor cadastrado no sistema. Nesse caso, é estabelecido também um limiar associado a cada *codebook*.

A etapa de geração de *codebook* consiste na geração dos níveis discretos que cada vetor poderá assumir – os centróides. Esses níveis são armazenados em um *codebook*. Para a geração de tal *codebook*, é utilizado um grupo de vetores de treinamento. Os centróides encontrados nessa etapa devem ser os melhores representantes dos vetores de treinamento. Em outras palavras, os centróides devem ser tais que minimizem o somatório da distorção de cada vetor de treinamento relacionado com seu respectivo centróide. Um determinado *codebook* é chamado ótimo (ou globalmente ótimo) se, para um número *T* de centróides, a distorção média de todos os vetores de treinamento (distorção média global), quando comparados com o centróide associado, é menor que a distorção média produzida por qualquer outro *codebook* de *T* centróides. A associação entre um vetor qualquer e o centróide que melhor o representa é feita utilizando a regra de seleção pela mínima distorção (*nearest neighbor*). O cálculo da distorção média global (DMG) é realizado de acordo com a equação 10.

$$D = \frac{1}{Tr}\sum_{n=1}^{Tr} d\left[x_n, y_n\right] \quad (10)$$

onde *D* é a distorção média global, *Tr* é o número de vetores de treinamento, x_n é o n-ésimo vetor de treinamento, y_n é o centróide associado a x_n, e $d[x, y]$ é a medida de distorção entre os vetores *x* e *y*.

O algoritmo para geração de centróides utilizado é o Linde, Buzo e Gray (LBG) (Makhoul, Roucos e Gish,1985; Soong, Rosenberg e Juang,1987).

5.2 Quantização

A quantização de um padrão é a escolha do centróide que melhor representa cada vetor. Isso é feito levando-se em conta a distância entre o vetor em questão e todos os centróides existentes no *codebook* que está sendo utilizado.

O método mais simples para realizar uma quantização de uma seqüência de vetores, chamado *full search*, seria o de comparar cada vetor da seqüência com todos os centróides armazenados no *codebook*. O centróide mais "semelhante" é assumido como representante do vetor em questão. O processamento necessário para executarmos uma quantização através desse método, porém, é enorme. Um outro método, chamado *tree search*, armazena todos os centróides que foram sendo duplicados e considerados "estáveis" ou prontos para próxima duplicação, e a quantização é realizada através da comparação entre o vetor em questão e os dois centróides que as comparações anteriores indicaram. A figura 2 ilustra o método *tree search*, onde o vetor a ser quantizado é primeiramente comparado com os dois centróides iniciais. Decidido qual dos dois centróides é mais "semelhante", o vetor é comparado então com os próximos dois centróides que foram derivados do primeiro centróide "vencedor", e assim por diante. O centróide associado ao vetor em questão será escolhido na última comparação.

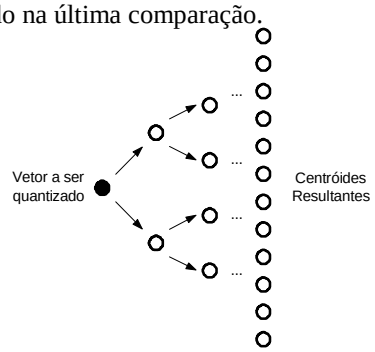


FIGURA 2: Método *tree search* para quantização de um vetor.

5.3 Comparação

Como última etapa do processo de classificação de padrões temos a comparação entre dois vetores.

A comparação é realizada através do cálculo da distorção entre eles. Há vários tipos de medidas de distorção entre vetores que podem ser utilizadas em reconhecimento de voz. A medida de distorção mínima ou euclidiana é a medida mais conhecida. Outro tipo de medida de distorção que aproxima com maior fidelidade a medida de distorção de máxima verossimilhança (que é a medida que espelha a distorção entre variáveis aleatórias com distribuição gaussiana), é chamada de distância de Mahalanobis, descrita pela equação 11.

$$d[x,y] = (x-y)^T \sum^{-1} (x-y) \quad (11)$$

onde $d[x,y]$ é a medida de distorção entre o vetor *x* e o vetor *y*, e $\sum$ é a matriz diagonal de covariância do vetor *y*.

Podemos reescrever a equação 11, como mostra a equação 12.

$$d[x,y] = \frac{(x_1-y_1)^2}{\sigma_1} + \frac{(x_2-y_2)^2}{\sigma_2} + \ldots + \frac{(x_p-y_p)^2}{\sigma_p} \quad (12)$$

onde $d[x,y]$ é a medida de distorção entre o vetor *x* e o vetor *y*, *p* é a dimensão do vetor *y* e do vetor *x*, x_i é a i-ésima componente dimensional do vetor *x*, y_i é a i-ésima componente dimensional do vetor *y*, σ_i é a variância da i-ésima componente dimensional do vetor *y*.

5.4 Multisecção

A técnica de quantização vetorial descrita utiliza um *codebook* para cada locutor. Essa técnica não preserva a característica temporal de tais amostras. Essa falta de caracterização explícita de aspectos seqüenciais das amostras pode ser remediada. Utiliza-se para isso a técnica chamada Quantização Vetorial Multiseção (*Multi-Section Vector Quantization* – MSVQ).

O processo de MSVQ é idêntico ao processo usual de quantização vetorial, com a diferença que as amostras de treinamento são divididas em pedaços de mesmo tamanho. Uma palavra é falada com determinada duração, e se for repetida pelo mesmo locutor terá sua duração diferente da anterior. Mas se repartíssemos a palavra em partes de tamanhos idênticos, teríamos cada pedaço com, praticamente, a mesma informação. Geramos então um *codebook* para cada pedaço da palavra. Quando desejássemos avaliar uma palavra a ser classificada, repartiríamos também tal palavra e cada pedaço seria avaliado com o *codebook* correspondente. Isso evita que vetores localizados, por exemplo, no final da amostra a ser classificada sejam associados a centróides gerados com vetores do início das amostras de treinamento. Dessa forma, a MSVQ faz com que cada vetor da amostra a ser classificada só possa ser associado a um centróide com alguma correspondência temporal com tal vetor. A figura 3 ilustra o processo de treinamento ou geração dos *codebooks* utilizando, por exemplo, *N* amostras de treinamento, subdivididas em três partes iguais. Quando uma amostra fosse classificada, ela seria também repartida em três partes iguais e cada parte só poderia ter seus vetores associados aos centróides pertencentes ao *codebook* correspondente àquela parte.

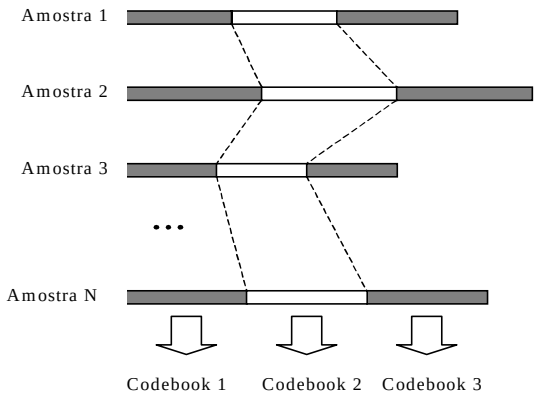


FIGURA 3: Treinamento da quantização multiseção

5.5 Metodologia Utilizada para Classificação

Toda a análise teórica mostrada anteriormente não evidencia claramente como é utilizada a técnica de quantização vetorial para a classificação de um padrão vocal desconhecido. Após o sistema ter sido treinado, é efetuada a *verificação de locutor*, que definirá se o usuário foi previamente cadastrado.

Treinamento

Para a realização do processo de treinamento do sistema, tem-se um *codebook* para cada um dos locutores cadastrados no sistema. O *codebook* de um determinado locutor é obtido utilizando amostras de voz daquele locutor específico. É também estabelecido um limiar associado a cada *codebook*. O cálculo do limiar associado a cada *codebook* assume distribuição gaussiana (Burton,1987). Ele é estabelecido levando-se em conta as distorções geradas por dois tipos de amostras de voz: as que geraram o *codebook* as que sabidamente não pertencem àquele locutor. O limiar é calculado de acordo com a equação 13.

$$L_i = \frac{\mu_i^{in}\sigma_i^{out} + \mu_i^{out}\sigma_i^{in}}{\sigma_i^{out} + \sigma_i^{in}} \quad (13)$$

onde *Li* é o limiar associado ao i-ésimo *codebook*, μ_i^{in} é a média entre os valores de distorção obtidos com as amostras de voz que pertencem ao i-ésimo locutor, μ_i^{out} é a média entre os valores de distorção obtidos com as amostras de voz que não

pertencem ao i-ésimo locutor, σ_i^{in} é o desvio padrão correspondente às distorções obtidas com as amostras de voz que pertencem ao i-ésimo locutor, e σ_i^{out} é o desvio padrão correspondente às distorções obtidas com as amostras de voz que não pertencem ao i-ésimo locutor.

Verificação de Locutor

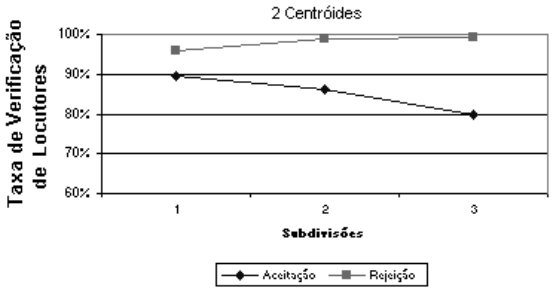
Para verificar a amostra de voz como pertencente a um determinado locutor primeiramente calcula-se a distorção acumulada obtida quantizando o padrão desconhecido com o *codebook* daquele locutor e comparando os vetores resultantes com os vetores do padrão inicial, vetor a vetor. Após, é feita uma comparação entre tal distorção e o limiar associado. Se a distorção acumulada for superior ao limiar, a amostra de voz que está sendo avaliada não é reconhecida como pertencente ao locutor em questão. Caso contrário, aceita-se a amostra de voz como pertencente àquele locutor. Faz-se realmente apenas uma verificação de locutor, ou seja, a amostra avaliada pertence ou não ao locutor.

6 TESTES REALIZADOS

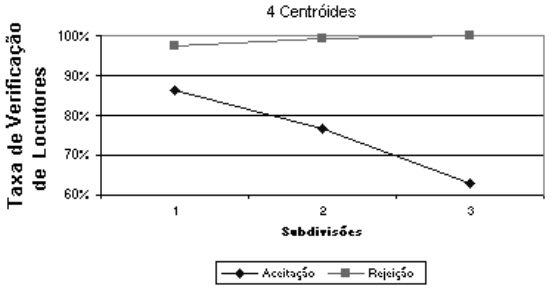
O treinamento de locutores foi realizado utilizando 50 locutores. Foram usadas 10 repetições da palavra "andar" para cada locutor, totalizando 500 arquivos. O cálculo do limiar necessita amostras de voz faladas por locutores não cadastrados. Assim, para cada um dos 50 locutores também são utilizados no treinamento 1 repetição da palavra "andar" falada por 37 locutores diferentes do 50 cadastrados.

No reconhecimento dos locutores cada um dos 50 locutores tentou acessar 10 vezes o sistema no qual foi cadastrado. Essas verificações compõem a taxa de aceitação de locutores com direito de acesso, ou apenas *aceitação*. Cada um dos 50 locutores também tentou 4500 vezes o acesso sem permissão. Essas tentativas de acesso compõem a taxa de rejeição de impostores, ou apenas *rejeição*.

Na figura 4 nota-se que a rejeição a impostores aumenta, quando aumenta-se o número de centróides em cada *codebook* de 2 (figura 4-a) para 4 (figura 4-b). Entretanto, a taxa de aceitação de locutores cadastrados cai de forma mais acentuada. Como nosso sistema deve otimizar ambas taxas, *codebooks* com 2 centróides são preferidos, apresentando inclusive menor necessidade de processamento.



(a)



(b)

FIGURA 4: Taxa de verificação de locutor para 14 MFCC utilizando (a) 2 centróides por *codebook* e (b) 4 centróides por *codebook*

Podemos notar na figura 5 uma queda gradual na taxa de rejeição de impostores, com o aumento do número de coeficientes mel-cepstrais utilizados. A taxa de aceitação de locutores com acesso não varia muito no intervalo para o número de coeficientes proposto. Fica claro, então, que a utilização de coeficientes excessivos prejudica a verificação de locutores, da mesma forma que o faz com as taxas de reconhecimento de comandos.

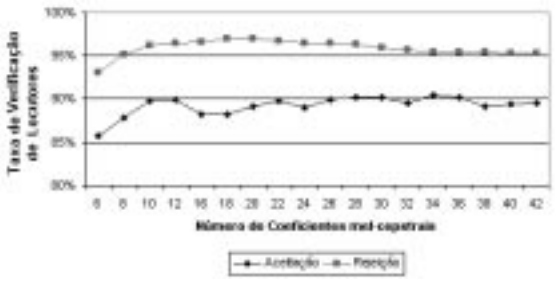


FIGURA 5: Taxa de verificação de locutor variando o número de coeficientes mel-cepstrais

Para os testes realizados, o melhor resultado que maximiza as taxas de aceitação e rejeição apresentou 90% de aceitação de locutores com direito de acesso e 96% de rejeição de impostores.

7 CONCLUSÕES

Neste trabalho foram apresentadas técnicas utilizadas para o reconhecimento automático de pessoas pela voz.

O reconhecimento de locutor é uma área específica que ainda tem muito a desenvolver. A extração de novos coeficientes, que representem de forma mais fiel as características singulares do locutor poderá nos levar à criação de sistemas mais confiáveis e robustos. Seria ainda importante que tais coeficientes fossem cada vez mais independentes do estado emocional do locutor, de anomalias momentâneas no aparelho respiratório (como resfriados ou inflamações) ou do ambiente onde é realizada a aquisição da voz. Isso facilitaria a implementação de sistemas onde a segurança e confiabilidade é imperativo, como o sistema descrito anteriormente.

AGRADECIMENTOS

Os autores agradecem a Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por seu apoio financeiro.

REFERÊNCIAS BIBLIOGRÁFICAS

LUFT, Joel Augusto, “Reconhecimento Automático de Voz para Palavras Isoladas e Independente de Locutor”. Dissertação de mestrado, Universidade Federal do Rio Grande do Sul - Escola de Engenharia, 1991.

PAPAMICHALIS, Panos E., “Pratical Approaches to Speech Coding”. Prentice Hall, 1987.

KIL, David H.; SHIN, Frances B., “Pattern Recognition and Prediction with Applications to Signal Characterization”. AIP Press, American Institute of Physics, 1996.

PICONE, Joseph W, “Signal Modeling Techniques in Speech Recognition”. Proceedings of The IEEE, vol. 81, no. 9, september 1993, 1215-1247.

DELLER, John R.; PROAKIS, John G.; HANSEN, John H. L., “Discrete-time Processing of Speech Signals”. Prentice Hall, 1987.

RABINER, Lawrence; JUANG, Biing-Hwang, “Fundamentals of Speech Recognition”. Prentice Hall, 1993.

MAKHOUL, John; ROUCOS, Salim; GISH, Herbert, “Vector Quantization in Speech Coding”. Proceedings of the IEEE, vol. 73, no. 11, novembre 1985, 1551-1587, 1985.

GRAY, Robert M., “Vector Quantization”. Readings in Speech Recognition, 75-100, 1984.

BURTON, David K., “Text-Dependent Speaker Verification Using Vector Quantization Source Coding”. IEEE Transactions on Acoustics, Speech, and Signal Processing, , vol. ASSP-35, 133-143, 1987.

ROSENBERG, A. E.; SOONG, F. K., “Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes”. Computer Speech and Language, 22, 143-157, 1987.

SOONG, Frank K., ROSENBERG, Aaron E., JUANG, Biing-Hwang, RABINER, Lawrence R., “A Vector Quantization Approach to Speaker Recognition”. AT&T Technical Journal, vol. 66, Issue 2, march/april 1987, 14-26, 1987.