Universidade de São Paulo – USP
Escola de Engenharia Elétrica – EESC
Departamento de Engenharia Elétrica

Michel Alves Lacerda

RECONHECIMENTO DE LOCUTOR DEPENDENTE DE TEXTO

São Carlos
2010

MICHEL ALVES LACERDA

RECONHECIMENTO DE LOCUTOR DEPENDENTE DE TEXTO

Trabalho de Conclusão de Curso
apresentado à Escola de Engenharia
de São Carlos, da
Universidade de São Paulo

Curso de Engenharia Elétrica com
ênfase em Eletrônica

ORIENTADOR: Profº Dr. Rodrigo Capobianco Guido

São Carlos
2010

Ficha catalográfica preparada pela Seção de Tratamento da Informação do Serviço de Biblioteca – EESC/USP

Lacerda, Michel Alves

L131r Reconhecimento de locutor dependente de texto / Michel Alves Lacerda ; orientador Rodrigo Capobianco Guido –- São Carlos, 2010.

Monografia (Graduação em Engenharia Elétrica com ênfase em Eletrônica) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2010.

Dedicatória

Este trabalho é fruto de anos de estudo na universidade, que por sua vez é resultado dos anos de estudo desde as primeiras séries da escola. Todo esse tempo exigiu muito esforço, disciplina e por vezes, sacrifícios, e com toda certeza, dedico este conjunto a Deus, sem o qual nada teria começado, prosseguido e terminado.

Dedico este trabalho, também, aos meus pais João e Adalgisa, e a minha irmã Anna, que sempre me apoiaram e incentivaram em todos os momentos.

Agradecimentos

Agradeço a todas as pessoas que contribuíram direta ou indiretamente para a realização deste trabalho.

Em especial ao Professor Rodrigo Capobianco Guido pela orientação, disposição, dedicação, oportunidade e ajuda.

Aos meus amigos que apoiaram, incentivaram, estiveram comigo ou apenas "doaram" suas vozes para a realização do trabalho, à Jussara Ribeiro, João Paulo, Guilherme Nishizawa, Rodrigo Neves, Rodrigo Biondo, Samir Khaoule, Rafael Moya, Mateus Maia.

Aos professores que fundamentaram toda a base do conhecimento nesta jornada, do Departamento de Engenharia Elétrica - EESC, e ao próprio Departamento, assim como ao IFSC-USP , por oferecer toda estrutura necessária para realização do trabalho.

A frase mais empolgante de se ouvir na ciência, a que prenuncia novas descobertas, não é "Eureka!", mas sim "Isto é estranho..."

Isaac Asimov

Resumo

A Autenticação Biométrica por comandos de voz, especialidade pertencente à área de Processamento Digital de Sinais, das áreas de Engenharia Elétrica, Ciência da Computação e Física Aplicada, está em expansão e possui enorme aplicabilidade em diversos problemas do cotidiano. Com esse raciocínio em mente, o objetivo fundamental do presente trabalho é o de construir um algoritmo para a autenticação biométrica de locutor, com vocabulário limitado, do tipo *text-dependent*, baseados em redes neurais artificiais. Pretende-se criar uma abordagem eficiente em termos da taxa de reconhecimento, pesquisando qual é a melhor arquitetura de rede neural para a aplicação do algoritmo na área mencionada. O projeto pretende, focando as características da fala humana, ser uma ferramenta de autenticação biométrica que possui baixo custo computacional e desempenho satisfatório.

Palavras Chave: Autenticação Biométrica, Biometria, Processamento de Sinais, *Wavelets*, Comandos de Voz.

Abstract

The speech-based biometric authentication, which is an emerging field of digital signal processing, inside the areas of Electrical Engineering, Computer Science and Applied Physics, has a considerable applicability in real-life problems. Based on this idea, the fundamental objective of the proposed work is to build an algorithm for biometric authentication based on humans' voice analysis, considering a limited vocabulary, in *text-dependent* mode, and based on artificial neural networks. The intention is to create an efficient approach in terms of accuracy, proposing an efficient system architecture. By focusing on the characteristics of human voices, the project aims at serving as a tool for voice biometric authentication with a low computational cost and satisfactory performance.

Keywords: Biometric Authentication, Biometrics, Signal Processing, Wavelets, Voice Commands.

Conteúdo

Índice de Figuras

Índice de Equações

Índice de Tabelas

1. Introdução

A autenticação biométrica por comandos de voz [1], especialidade pertencente à área de processamento digital de sinais [2], [3] das áreas de Engenharia Elétrica, Ciência da Computação e Física Aplicada, está em expansão e possui enorme aplicabilidade em diversos campos [6][7].

Assim, o objetivo fundamental do presente trabalho foi o de criar um algoritmo para autenticação biométrica de locutores, com base em um vocabulário limitado, isto é, do tipo *text-dependent*, fundamentalmente utilizando uma Máquina de Vetor de Suporte (*Support Vector Machine* - SVM) [4], criando assim uma abordagem eficiente em termos da taxa de reconhecimento e estabelecendo uma correspondência, de forma a caracterizar a arquitetura e os parâmetros que proporcionem os melhores resultados.

O foco do reconhecimento de locutores está principalmente ligado a verificação mais detalhada das vozes dos mesmos, e tem sido um tema muito aplicado em diversos campos da nossa sociedade. Desde sistemas de segurança até em inteligência de complexos robôs, a identificação de sinais de voz e sua atribuição a uma pessoa é de suma importância. Portanto, aperfeiçoar o processo de reconhecimento de locutor, com novos algoritmos, com combinações de técnicas diferentes, e assegurando baixa complexidade computacional, significa melhorar inúmeros serviços e tecnologias.

Transformadas *Wavelet* Discretas (TWD) [11], Dimensões Fractais (DF) [12] e classificadores SVM [4] são técnicas importantes que têm sido utilizadas nas mais diversas aplicações. A capacidade da Transformada *Wavelet* de converter um sinal temporal para o domínio tempo-frequência, o grau de auto-similaridade de um sinal obtido pela Dimensão Fractal, assim como a potencialidade de um classificador em caracterizar certo conjunto de dados como pertencentes ou não a uma classe, são

qualidades que, em conjunto, podem compor um *speaker-verification algorithm* de baixa complexidade computacional.

Em vista das considerações anteriores, a abordagem eficiente proposta neste trabalho foi o desenvolvimento de um algoritmo para *speaker-verification* otimizado, no qual, a TWD foi utilizada para extrair as frequências formantes, e suas energias, dos sinais vozeados; a DF foi utilizada para obter o grau de auto-similaridade dos sinais; e, finalmente, um conjunto de SVMs é responsável pela classificação final.

A implementação, na sua totalidade, além de utilizar um computador pessoal comum com sistema operacional LINUX, foi proposta para permitir a utilização futura em tempo-real, com base em um *Digital Signal Processor* (DSP) e um DSPic [5].

Para a verificação da aplicabilidade do algoritmo desenvolvido, foi necessária a avaliação do seu funcionamento em condições similares às das diversas aplicações já citadas. Para tanto, foi realizada também a confecção do *hardware* necessário para um possível uso embarcado do DSP / DSPic. Isso para assegurar que as características necessárias dos sinais de entrada, transmissão e saída, sejam as ideais.

2. Revisão Bibliográfica

A revisão bibliográfica auxilia a compreender os conceitos e fundamentos principais para o entendimento dos objetivos, procedimentos, resultados e da conclusão. Os principais temas foram pesquisados e estudados para que seja possível uma melhor compreensão do projeto.

2.1 O Sistema de Produção de Fala

A fala humana é possível graças, primeiramente, ao mecanismo propulsor de ar formado basicamente pelos pulmões. Esse fluxo de ar proveniente dos pulmões passa, em seguida, pelas pregas vocais, que vibram, terminando no trato vocal e nasal, até ser expelido. A vibração das pregas vocais injeta pulsos de ar, periodicamente ou aperiodicamente, dependendo do tipo de som sendo produzido, na cavidade oral, que assume formato específico determinado pela língua, lábios, mandíbula e posição vertical da laringe. No caso de sons vozeados [9], as cordas vocais vibram em uma frequência periódica, o que não acontece para os sons não vozeados. O trato vocal pode ser modelado, conforme foi estudado, por uma função de transferência da forma tudo-pólo, onde cada pólo faz referência a uma frequência formante, ou seja, uma frequência de ressonância.

Sobre a fisiologia da fala, apesar de ser um mecanismo repleto de detalhes, é possível entender o aparato vocal realizando uma divisão do mesmo em subsistemas, destacando as características anatômicas mais relevantes na acústica das vogais:

respiratório: constituído pelos pulmões, músculos respiratórios, brônquios e a traquéia, sendo responsável pela geração da energia utilizada na produção de voz;
laríngeo: constituído por um conjunto de músculos, ligamentos e cartilagens sendo responsável pelo controle das pregas vocais (fonação) ;

supra-laríngeo: composto pelas regiões faringal, bucal e nasal, sendo responsável pela modulação do som gerado.

Basicamente, a literatura relata um modelo chamado de fonte-filtro, que separa os fenômenos acústicos em três grupos independentes: a fonte sonora, o filtro acústico e a irradiação. Consiste basicamente da propulsão de ar pelos pulmões, seguida de um processo de filtragem, ou equalização, realizado pelo trato vocal e elementos associados, conforme a figura 2.1.1.

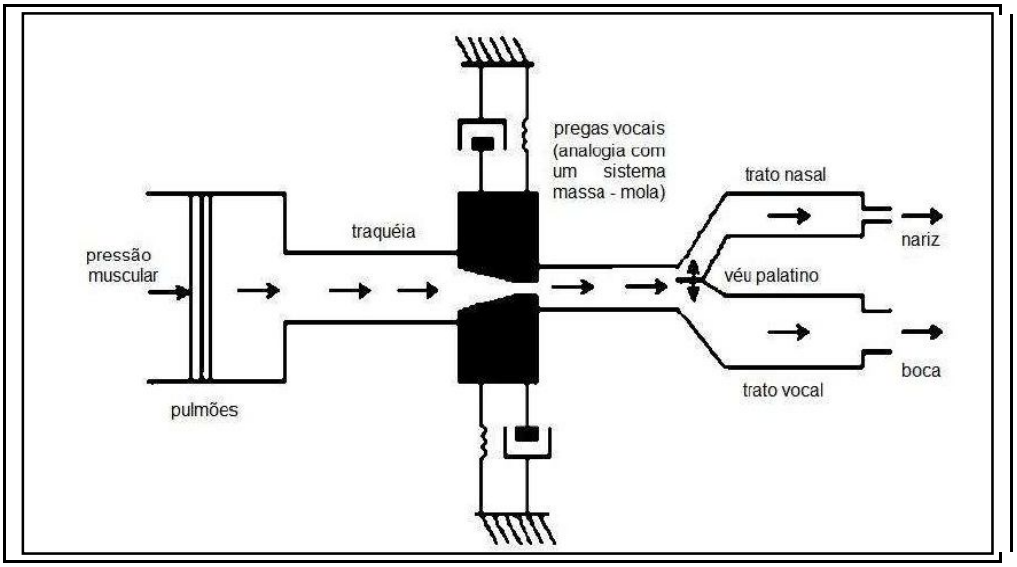

Figura 2.1. 1 Interpretação física simplificada do sistema gerador de voz

O pulmão pode ser considerado uma fonte propulsora de ar que torna possível a produção de voz. Esta massa de ar é conduzida até a laringe pela traquéia, onde são produzidos os pulsos glotais para a caracterização da voz. A faringe age como um ressonador e as pregas vocais controlam o fluxo de ar fornecido pelos pulmões, fazendo com que esse sinal de excitação seja periódico, vibrando em determinada frequência, ou aperiódico, similar a um sinal ruidoso. Se o sinal for periódico, este período é chamado de período de *pitch* e a voz produzida será classificada como vozeada (*voiced speech)*, caso contrário a voz será classificada como não vozeada (*unvoiced speech*).

O trato vocal é composto por elementos que guardam íntima relação entre si, o que permite a modelagem do som e a projeção do mesmo no ambiente. Dependendo de como as estruturas seguintes às pregas vocais agem, pode-se ainda refinar esta classificação dos sinais de voz da seguinte forma [9]:

***Fricatives*: é um *unvoiced speech* que surge quando há fricção do ar em movimento contra a constrição, causando, em geral, uma turbulência de ar entre a língua e os dentes superiores. Exemplo: th na palavra *thin* de origem inglesa;**
Plosives*: é um *unvoiced speech* impulsivo, como t na palavra *top;
Whispers*: é um *unvoiced speech* onde uma barreira é criada nas pregas vocais de forma que elas permaneçam parcialmente fechadas e sem oscilação como ocorre quando se pronuncia o h na palavra *he;
Voiced fricatives*: são fonemas *voiced speech*, ou seja, de excitação periódica, porém misturados com ruído criado na constrição do trato vocal, atrás dos dentes e contra o palato. Exemplo: z na palavra *zebra;
***Unvoiced fricatives*: idem ao anterior, porém as pregas vocais não vibram simultaneamente com a fricação;**
***Voiced plosives*: são fonemas *voiced speech*, ou seja de excitação periódica, porém misturados com ruído impulsivo criado no trato vocal;**
Unvoiced plosives*: idem anterior, porém as pregas vocais não vibram simultaneamente com o impulso. Exemplo: b na palavra *boat;

Qualquer palavra ou frase pronunciada por um locutor pode ser dividida em fonemas, cada qual podendo ser classificado como explicado anteriormente. Outros conceitos necessários para compreensão da produção de voz, utilizando o modelo fonte-filtro são:

Formantes: correspondem as ressonâncias do trato vocal. Os formantes mais importantes para o reconhecimento de uma vogal são os três primeiros (F1, F2, F3). As freqüências dos formantes, particularmente as duas primeiras (F1 e F2), dependem do formato do trato vocal entre a glote e os lábios. O terceiro formante, F3, está relacionado com a ressonância do restante do trato vocal com a passagem da constrição de língua. F4 é altamente relacionado ao timbre vocal, ou seja, ao componente pessoal do som vocal;

Frequência fundamental (*pitch*): é a componente de freqüência da excitação pulmonar (F0), controlada pela vibração das pregas vocais no caso de *voiced speech*. Os homens apresentam frequência fundamental que vai, aproximadamente de 70 Hz até 130 Hz, enquanto que nas mulheres essa faixa costuma variar de 150 até 230 Hz, e, finalmente, podem ultrapassar 300Hz em crianças.

Devido à assimetria dos pulsos glotais, o espectro do sinal de excitação pulmonar que atravessa as pregas vocais possui, além da freqüência fundamental, uma série harmônicas.

2.2 *Speaker Verification*

Um sistema de *speaker verification* tem como objetivo a autenticação de uma pessoa, ou seja, a verificação, através da voz, se a pessoa pertence ou não ao conjunto de locutores "aceitos". Ao sistema cabe, portanto, tomar uma decisão binária em respeito à identidade do locutor [4]. Com aplicações em qualquer área que seja necessária a correlação vocal entre um locutor e um registro, a técnica de *speaker verification* tem como pontos fortes a simplicidade e naturalidade em relação a outros sistemas de verificação de identidade. O funcionamento do sistema procede da seguinte forma: a partir de uma prévia base de dados de características de usuários, o locutor fornece uma amostra de voz de onde são extraídas características para a classificação [5]. Caso haja o reconhecimento com a identidade em questão, o locutor será autenticado, caso contrário, o locutor será rejeitado. Neste trabalho um sistema equivalente é apresentado com um classificador SVM-*Wavelet*-Fractal.

2.3 Classificadores SVMs

Redes Neurais Artificiais [6] são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e

que adquirem conhecimento através da experiência. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios.

O sistema nervoso é formado por um conjunto extremamente complexo de células: os neurônios. Eles têm um papel essencial na determinação do funcionamento e comportamento do corpo humano e do raciocínio. Os neurônios são formados pelos dendritos, que são um conjunto de terminais de entrada, pelo corpo central, e pelos axônios que são longos terminais de saída.

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede. A propriedade mais importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas.

SVMs, que constituem uma das bases do presente projeto, são classificadores derivados das redes neurais artificiais [6], que possuem grande potencialidade para classificação e separação de características.

2.4 Transformada *Wavelet* Discreta

A TWD é uma função capaz de converter um sinal temporal para o domínio tempo-frequência. É possível visualizar o sinal em diferentes escalas. Para a execução da transformada, o sinal passa por uma filtragem, que pode ser representada pela Equação 2.4.1, ou mais especificamente pelas Equações 2.4.2 e 2.4.3, onde h[.] e g[.] são os filtros passa-baixas e passa-altas [7] , respectivamente. Todo o processo pode ser feito em vários níveis de decomposição, conforme

explicado na Figura 2.4.1, onde é exemplificado o funcionamento de uma Transformada *Wavelet* Discreta para um sinal s[.] de n amostras discretas e máxima freqüência PI, decomposto até o terceiro nível. Destaque para o espectro de freqüência e a quantidade de amostras presente em cada sub-banda.

$$y[n] = x[n] * t[n] = \sum_{k=0}^{n-1} t_k x_{2n-k}$$

Equação 2.4.1 : Filtragem da transformada

$$y_{passa-baixas}[\circ] = x[\circ] * h[\circ] = \sum_{k=0}^{n-1} h_k x_{2n-k}$$

Equação 2.4.2 : Filtragem passa-baixas

$$y_{passa-altas}[\circ] = x[\circ] * g[\circ] = \sum_{k=0}^{n-1} g_k x_{2n-k}$$

Equação 2.4.3 : Filtragem passa-altas

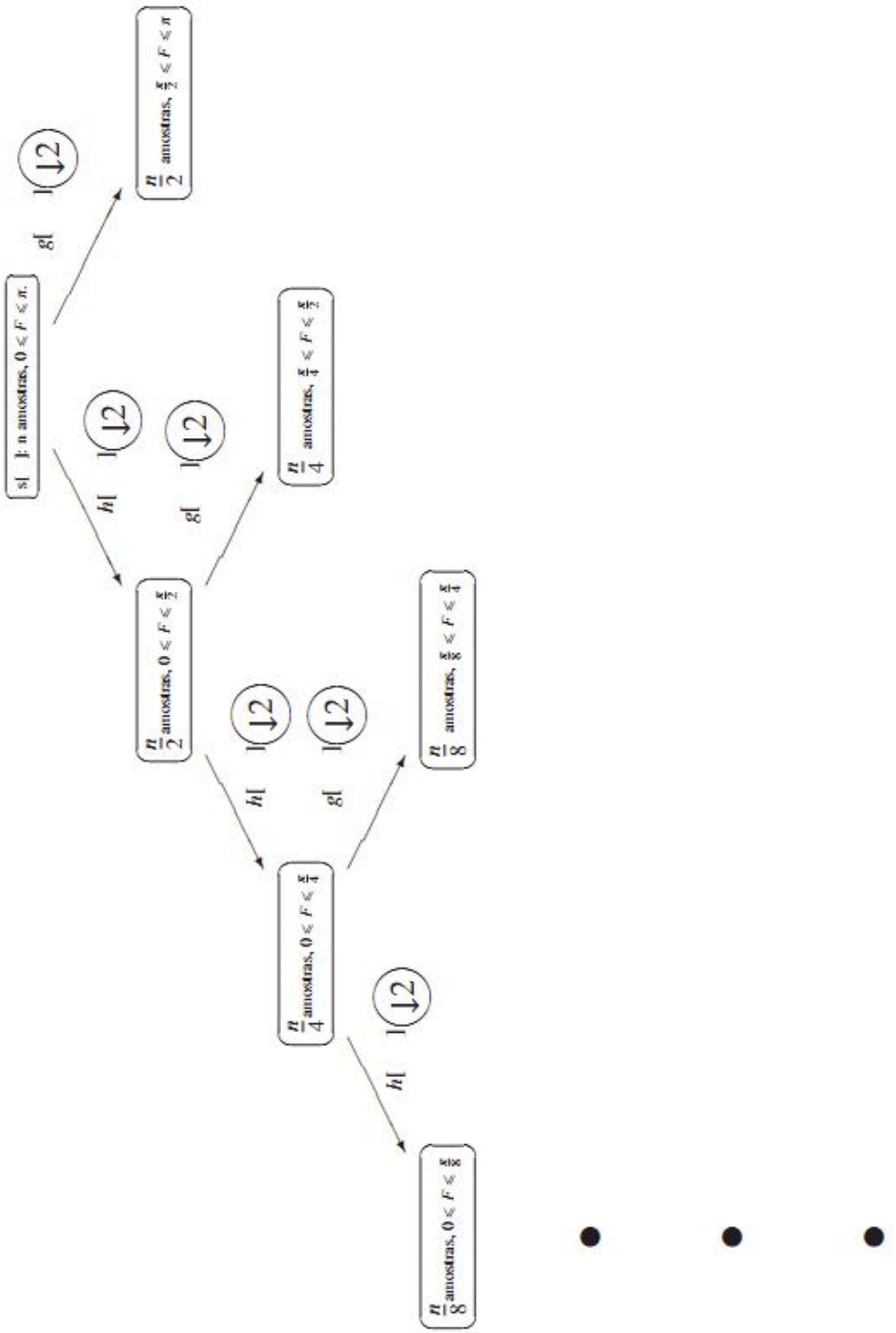

Figura 2.4. 1 Funcionamento da Transformada *Wavelet* Discreta.

A visualização em diferentes escalas é possível pela análise de multi-resolução proposta por Mallat. Para o cálculo da TWD de um sinal, aplica-se o algoritmo de Mallat, que envolve a multiplicação de duas matrizes para cada nível de transformação. Se A[.][.] é a matriz de coeficientes dos filtros e B[.] é o sinal original, então C[.] = A[.][.]B[.] corresponde ao sinal transformado, sendo que a disposição dos coeficientes nas matrizes está no Equação 2.4.4.

$$A[\quad][\quad] = \begin{pmatrix} h_0 & h_1 & h_2 & \dots & \dots & \dots & \dots & h_{n-1} & 0 & 0 & 0 & 0 & \dots & \dots & 0 & 0 \\ g_0 & g_1 & g_2 & \dots & \dots & \dots & \dots & g_{n-1} & 0 & 0 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & h_0 & h_1 & h_2 & \dots & \dots & \dots & h_{n-1} & h_n & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & g_0 & g_1 & g_2 & \dots & \dots & \dots & g_{n-1} & g_n & 0 & 0 & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot \\ \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot \\ \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot & & & \cdot \\ h_{n-1} & h_n & 0 & 0 & \dots & \dots & \dots & 0 & 0 & h_0 & h_1 & \dots & \dots & \dots & h_{n-3} & h_{n-2} \\ g_{n-1} & g_n & 0 & 0 & \dots & \dots & \dots & 0 & 0 & g_0 & g_1 & \dots & \dots & \dots & g_{n-3} & g_{n-2} \end{pmatrix},$$

$$B[\circ] = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \dots \\ \dots \\ \cdot \\ \cdot \\ \cdot \\ b_{n-2} \\ b_{n-1} \end{pmatrix}, \qquad C[\circ] = \begin{pmatrix} c_0 \\ c_{\frac{n}{2}} \\ c_1 \\ c_{\frac{n}{2}+1} \\ \dots \\ \dots \\ \cdot \\ \cdot \\ \cdot \\ c_{n-1} \\ c_{\frac{n}{2}-1} \end{pmatrix}.$$

Equação 2.4.4 : Matrizes envolvidas no cálculo da TWD.

Todo o processo da Transformada contribui para a extração das principais características (frequências formantes e suas energias) de um sinal vozeado, que contribuem para verificar um locutor.

2.5 Dimensão Fractal

A Dimensão Fractal mede o grau de auto-similaridade de um sinal. Relacionando tendências estatisticamente, obtêm se o expoente de Hurst que é uma estimativa numérica para previsão de séries temporais. Em uma amostra vozeada temos a tendência do comportamento do sinal refletindo-se no grau de auto-similaridade do mesmo; essa característica pode variar por diferentes fatores, e um deles é a proveniência do sinal, ou seja, o locutor.

2.6 Processadores Digitais de Sinais

O presente projeto foi implementado em um computador pessoal comum com ambiente Linux, entretanto, conforme já mencionado, uma das plataformas que podem ser utilizadas para funcionamento em tempo-real é o DSP, que consiste de um processador dedicado. Suas principais características são a alta velocidade, em instruções por segundo, e a execução rápida de processos computacionalmente custosos, como a Transformada de Fourier. O desempenho se dá com base em uma arquitetura que permite um ciclo de processamento controlado para responder em tempo-real, com atraso fixo, para cada entrada. Os DSPs disponíveis no *SpeechLab*-IFSC-USP, onde o projeto foi executado,são os *Analog Devices BlackFin*. A figura 2.6.1 exibe a foto de um desses kits.

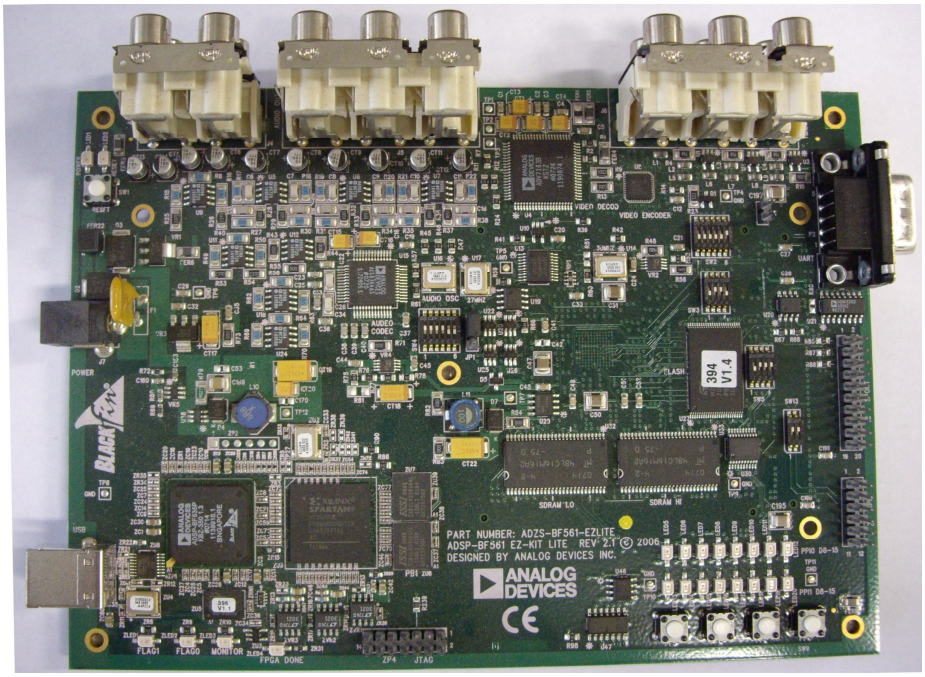

Figura 2.6. 1 *DSP Analog Devices - BlackFin* disponível no laboratório.

O DSP foi um dos dispositivos escolhidos para futura aplicação do algoritmo de reconhecimento de locutor dependente de vocabulário pela sua maleabilidade referente ao aporte de *softwares*, bem como pelas características estruturais que resultam em seu desempenho. Dentre as mais importantes está o alto paralelismo,

que permite processamento mais veloz; a alta taxa de *input/output*, permitindo que uma banda maior de dados seja tratada; rastreabilidade sobre eventos, permitindo que durante os testes os refinamentos sejam mais velozes; e alta precisão e velocidade, para que tenhamos no menor tempo a solução adequada.

Um segundo processador útil para efeitos de teste e acoplamento com os demais dispositivos é o DSPic 33F da *Microchip*. A figura 2.6.2 exibe o dispositivo, que é um microcontrolador com poder de processamento de um PIC, mas otimizado para processamento de sinais.

Figura 2.6. 2 *DSPic* disponível no laboratório.

O DSPic possui características suficientes para a nossa aplicação, desde frequência de amostragem de 48kHz, espaço para armazenamento de 4Mbits e as ferramentas de *software* para a gravação do algoritmo desenvolvido no *hardware*. O esquema de como é estruturado o *Kit* é exibido na figura 2.6.3.

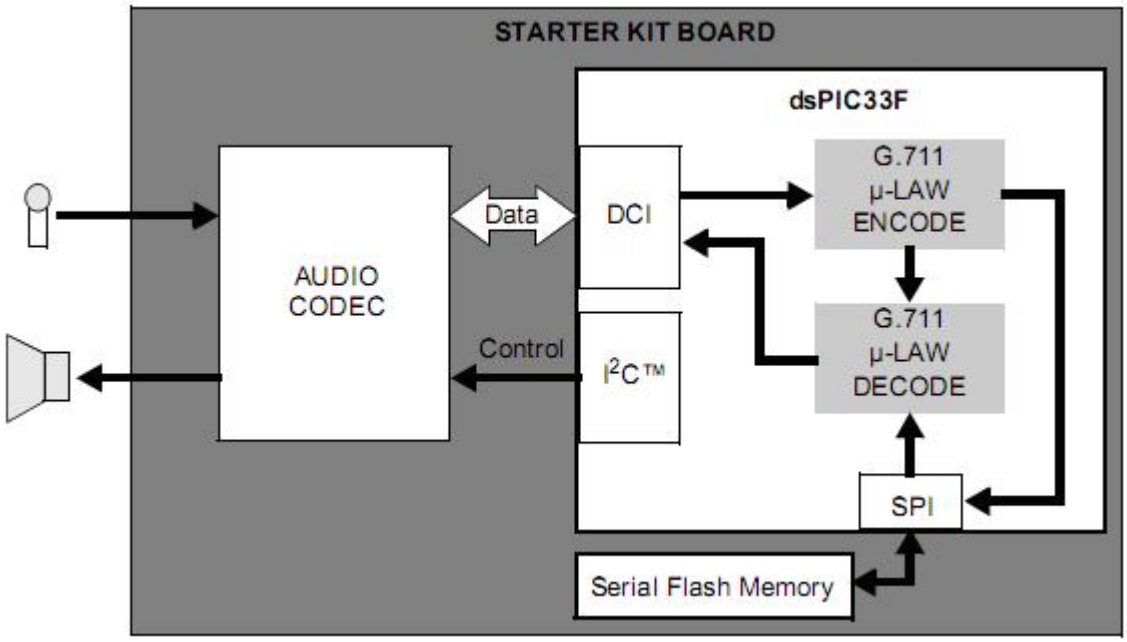

Figura 2.6. 3 Estrutura do *Kit DSPic*

3. Descrição do Sistema Proposto

Após a revisão bibliográfica, os algoritmos propostos para treinamento e para teste do sistema, que se encontram a seguir, foram implementados em linguagem C/C++ sob ambiente Linux.

3.1 Algoritmo Proposto para Treinamento do Sistema

INÍCIO
Passo T-1: Definir os sinais de voz a serem utilizados, que foram amostrados a 16000 amostras por segundo, quantizados com 16 bits ;
Passo T-2: Dividir os sinais em janelas de 256 amostras cada ;
Passo T-3: Para cada janela i, faça:
Passo T-3.1: subtrair a média de todas as 256 amostras do sinal, de tal forma que a freqüência 0 (nível "DC") seja removida ;
Passo T-3.2: se a janela corresponde a um trecho de *voiced speech*, então*:*
Passo T-3.2.1: obter a TWD nível máximo (8) da janela ;
Passo T-3.2.2: obter as energias das 5 primeiras bandas críticas do sinal, de acordo com a escala Bark ;
Passo T-3.2.3: obter a dimensão do fractal da janela corrente, com o método do espectro de potências ;
Passo T-3.2.4: obter as 5 primeiras freqüências formantes ;
Passo T-3.2.5: as 5 energias, as 5 formantes, além da dimensão do fractal, formam o vetor de características, com 11 parâmetros ;
Passo T-3.2.6: incorporar o vetor de características obtido no conjunto de treinamento do sistema, usando 1 como valor de saída, isto é, indicando que o locutor trata-se de um usuário autorizado ;
senão
descartar a janela;
FIM.

3.2 Descrição Detalhada

Para obterem-se os formantes é necessário obter o espectro das frequências do sinal de voz. Através do estudo e implementação de algoritmos, desenvolvidos no *Speechlab* para manipulação de arquivos *wave*, têm-se então ferramentas para visualizar os dados procurados nas amostras de um locutor. A partir dessas ferramentas foram estudados e desenvolvidos algoritmos capazes de, eficientemente, separar as frequências formantes necessárias bem como suas amplitudes. A técnica utilizada foi o processamento homomórfico para obtenção do cepstrum [7].

Fez parte do projeto, pesquisar diversos parâmetros, além dos formantes, para possibilitar a classificação e autenticação dos locutores, sendo que, dentre eles, foi estudado e implementado a dimensão fractal obtida por meio da técnica do espectro de potências associada ao expoente de Hurst [10] como mais uma característica a ser considerada.

Após a compreensão e aplicação desse conjunto de ferramentas, houve a integração e a implementação nos sinais amostrados de diversos locutores. Isso resultou num vetor de dados composto de formantes, energias das frequências respectivas e da dimensão fractal, caracterizando uma identificação própria para cada amostra, como na figura 3.2.1.

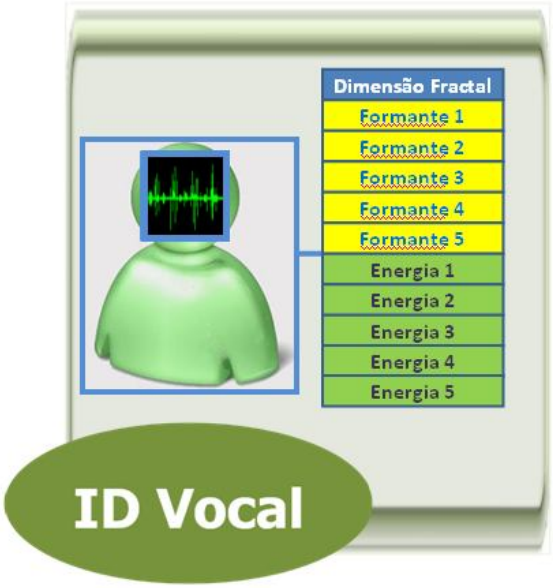

Figura 3.2. 1 Vetor de Características

A seguir, uma estrutura de SVMs foi utilizada com base em um treinamento supervisionado [4]. Particularmente, a SVM foi estruturada com *kernel* Gaussiano, contendo 11 pontos de entrada (mesma dimensão do vetor de características), 1 de saída (para a resposta binária: 1 ou -1) e, ainda, x unidades na camada intermediária, sendo x igual ao número de exemplos de treinamento utilizados para treinar o classificador.

4. Resultados

4.1 Detalhes do Experimento

Seguindo as atividades citadas, têm-se os resultados abaixo, seguindo a ordem de execução, e oos comentários respectivos. Detalhadamente, serão apresentados os experimentos realizados, para uma compreensão mais profunda do projeto.

Foram adquiridas as vozes de um grupo de 12 pessoas, com as características de idade, gênero, e histórico de complicações nas cordas vocais apresentado na Tabela 4.1.1.

Código do Voluntário	Gênero (Masculino – M / Feminino - F)	Idade	Complicações nas cordas vocais
1	F	47	Não
2	M	58	Não
3	M	21	Não
4	F	18	Não
5	M	21	Não
6	M	24	Não
7	M	22	Não
8	F	22	Não
9	M	19	Não
10	M	22	Não
11	M	23	Não
12	M	20	Não

Tabela 4.1.1 : Dados dos Locutores utilizados no experimento. Todos eles são considerados como "matriculados" ou aceitos pelo sistema.

Para o grupo dos 12 locutores, foi pedido que falasse, de forma normal, a frase “Eu amo processamento digital de sinais”. A aquisição foi realizada com um microfone de eletreto, com uma aquisição de 44100 Hz, comparável à qualidade de áudio de um CD comum. O tempo de duração da frase variou entre os participantes, gerando cerca de 720 janelas vozeadas, ao total.

Neste momento, é prudente lembrar que de acordo com a revisão bibliográfica, o algoritmo de verificação de usuário não necessita de amostras maiores que 0,050 segundos, pois nesse intervalo encontram-se de 7 a 10 períodos da forma de onda do fonema [8] [9].

Nas Figuras 4.1.1 e 4.1.2, é possível verificar um exemplo de aquisição de um sinal e o seu cepstrum correspondente. Já a tabela 4.1.2 e os gráficos das Figuras 4.1.3, 4.1.4, 4.1.5 e 4.1.6 ilustram características do sinal de cada locutor, dos locutores de sexo feminino, e dos locutores de sexo masculino, destacando que não há correlação aparente entre gênero neste grupo de participantes, reafirmando que as características vocais são determinadas pelo trato vocal, contribuindo assim para o reconhecimento do locutor [8] [9].

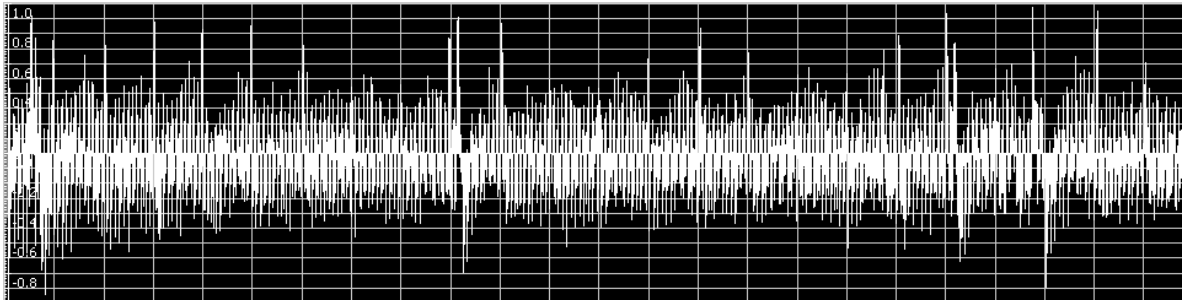

Figura 4.1. 1 Amostra de vogal "a" extraída de um locutor.

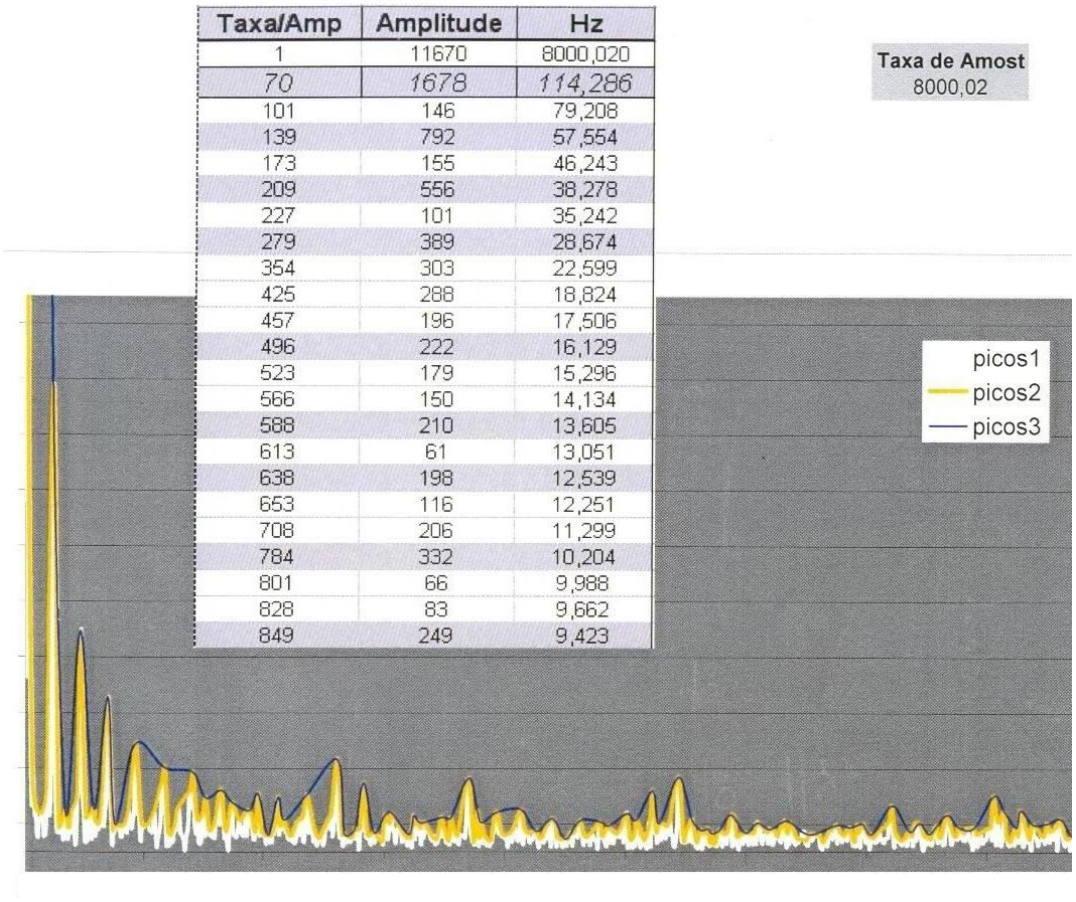

Figura 4.1. 2 Sinal exemplo no domínio *cepstral* com valores de freqüências formantes.

Locutor 1		Locutor 2		Locutor 3	
Índice	Amplitude	Índice	Amplitude	Índice	Amplitude
16	0,05194	12	0,010216	15	0,099042
22	0,04026	25	0,072755	34	0,035144
36	0,02467	42	0,017028	45	0,033546
46	0,03636	55	0,017028	56	0,011182
52	0,02077	68	0,027864	61	0,023962
Locutor 4		Locutor 5		Locutor 6	
Índice	Amplitude	Índice	Amplitude	Índice	Amplitude
21	0,051913	19	0,118761	11	0,124105
25	0,064208	29	0,07568	34	0,036993
42	0,021858	43	0,053356	43	0,02864
46	0,032787	55	0,020654	50	0,029833
57	0,032787	69	0,020654	61	0,019093
Locutor 7		Locutor 8		Locutor 9	
Índice	Amplitude	Índice	Amplitude	Índice	Amplitude
13	0,019908	13	0,049541	17	0,025788
20	0,013783	19	0,027523	29	0,018625
30	0,02144	33	0,022018	36	0,017192
51	0,012251	42	0,012844	41	0,02149
59	0,018377	54	0,014679	59	0,020057
Locutor 10		Locutor 11		Locutor 12	
Índice	Amplitude	Índice	Amplitude	Índice	Amplitude
10	0,083333	21	0,054007	17	0,151767
17	0,102713	34	0,057491	35	0,0665528
25	0,044574	41	0,04878	49	0,029106
34	0,065891	66	0,036585	63	0,051975
46	0,056202	75	0,019164	86	0,033264

Tabela 4.1.2 : Sinal transformado de cada locutor.

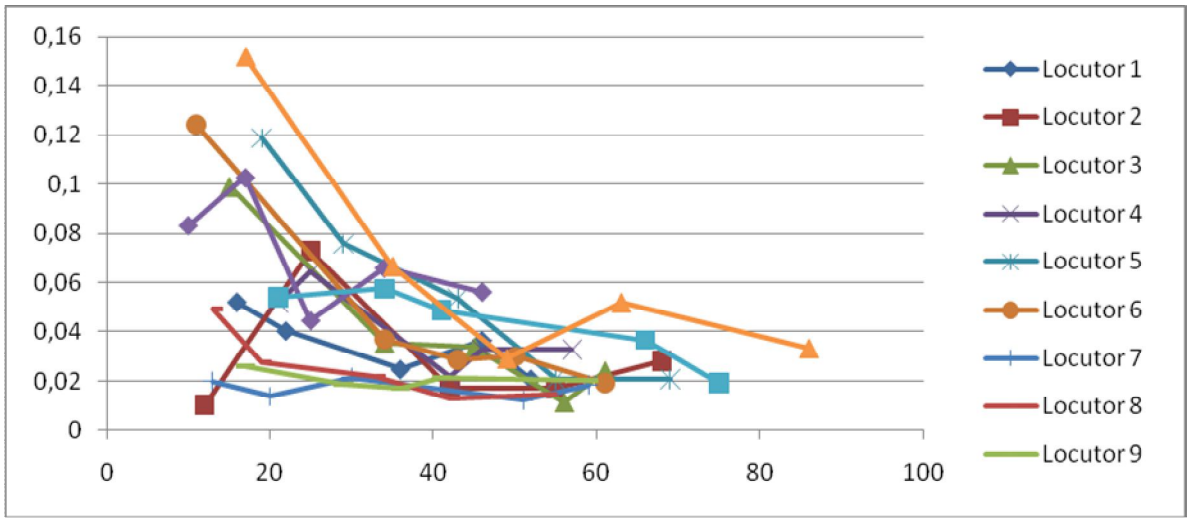

Figura 4.1. 3 Sinal de todos os locutores (Índice x Amplitude)

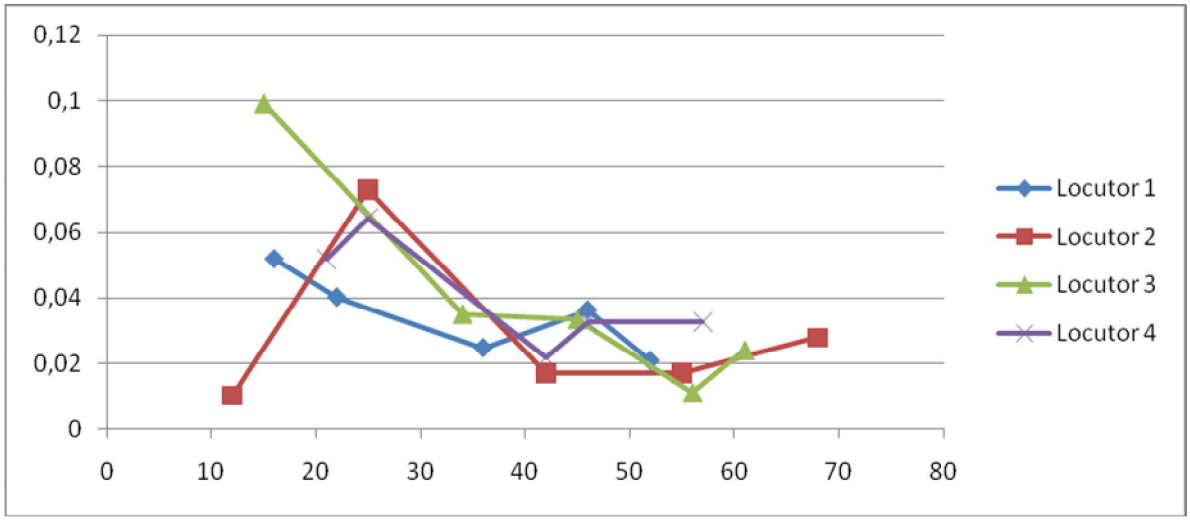

Figura 4.1. 4 Sinal de locutores do sexo feminino (Índice x Amplitude)

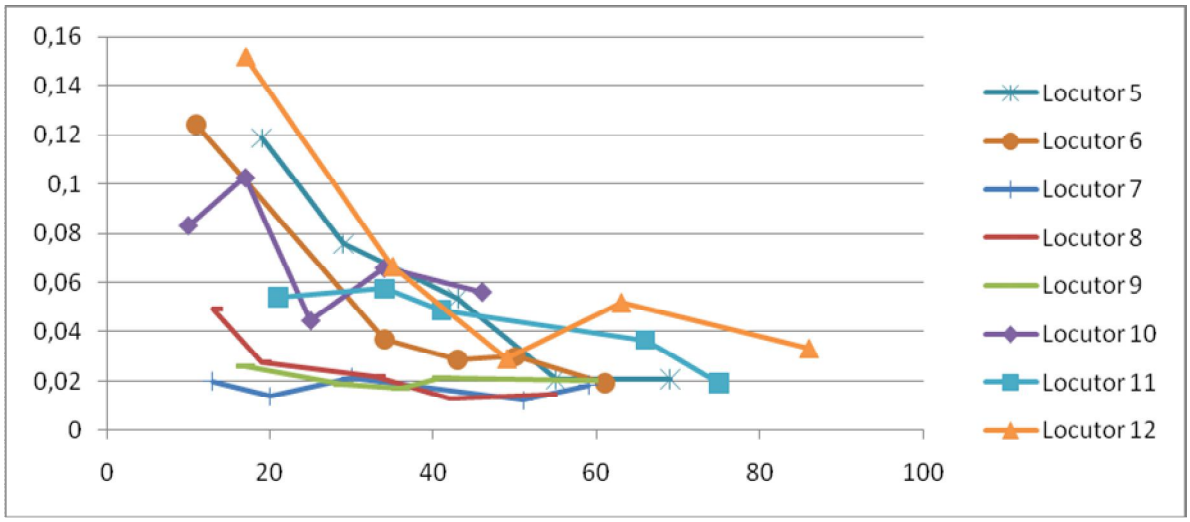

Figura 4.1. 5 Sinal dos locutores do sexo masculino (Índice x Amplitude)

identidade-locutor-1.txt - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
;EXP HURST(1)    FORMANTES(6)              ENERGIA NORMALIZADA(6)

1.486416 1 13 20 30 51 59 1.000000 0.019908 0.013783 0.021440 0.012251 0.018377
1.485276 1 12 36 50 73 91 1.000000 0.019769 0.011532 0.024712 0.037891 0.009885
1.484771 1 16 25 50 72 84 1.000000 0.016393 0.024590 0.014754 0.045902 0.018033
1.485040 1 16 23 46 52 62 1.000000 0.014754 0.011475 0.016393 0.011475 0.014754
1.487218 1 8 19 32 42 57 1.000000 0.026408 0.017606 0.014085 0.017606 0.021127
1.484127 1 20 36 44 55 65 1.000000 0.020638 0.022514 0.013133 0.015009 0.022514
1.482648 1 16 24 31 45 55 1.000000 0.028668 0.010118 0.013491 0.016863 0.026981
1.486161 1 9 19 33 53 59 1.000000 0.046552 0.022414 0.020690 0.010345 0.008621
1.486347 1 9 26 37 58 73 1.000000 0.038388 0.013436 0.019194 0.011516 0.065259
1.486815 1 18 34 39 45 54 1.000000 0.024621 0.005682 0.005682 0.009470 0.009470
1.487594 1 16 27 50 61 67 1.000000 0.023033 0.021113 0.019194 0.009597 0.024952
1.484641 1 6 23 37 42 59 1.000000 0.059846 0.021236 0.017375 0.011583 0.021236
1.484697 1 14 31 60 68 73 1.000000 0.032389 0.014170 0.022267 0.020243 0.046559
1.489416 1 17 23 34 41 45 1.000000 0.029668 0.029668 0.012216 0.012216 0.008726
1.485442 1 17 29 47 64 72 1.000000 0.017685 0.025723 0.011254 0.025723 0.038585
```

Figura 4.1. 6 Identidade da amostra

4.2 Resultados Primários do Experimento

Inicialmente, o processo descrito acima se mostrou satisfatório, com taxas de reconhecimento de 70% em média, em relação aos testes feitos com as identidades de algumas amostras de locutores. No detalhe das figuras 4.2.1 e 4.2.2 têm-se exemplos onde foram submetidas cinco e duas amostras diferentes, de um mesmo locutor, para a classificação. Quanto mais próximo do número 1 está o

resultado, maior é o reconhecimento da amostra. Pode-se perceber que no primeiro teste, apenas uma amostra não ficou próxima do valor esperado (um) utilizado para treinamento supervisionado, e no segundo nenhuma.

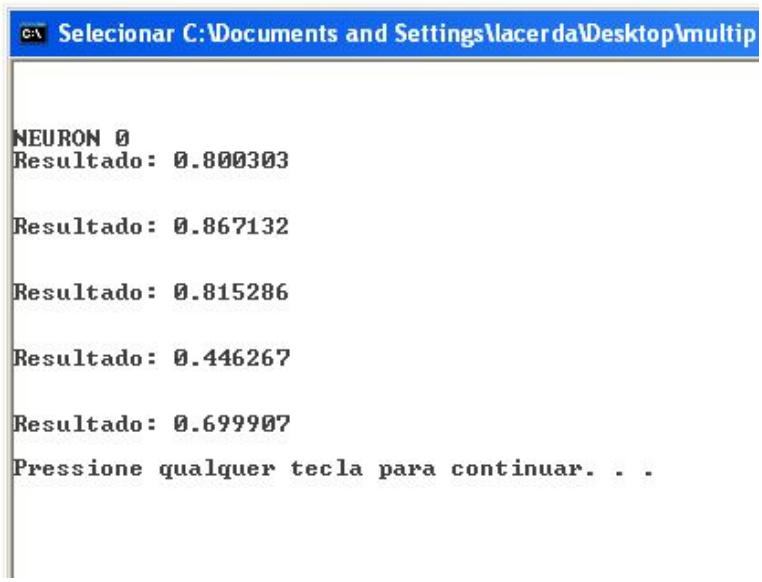

Figura 4.2. 1 Classificador analisando 5 amostras

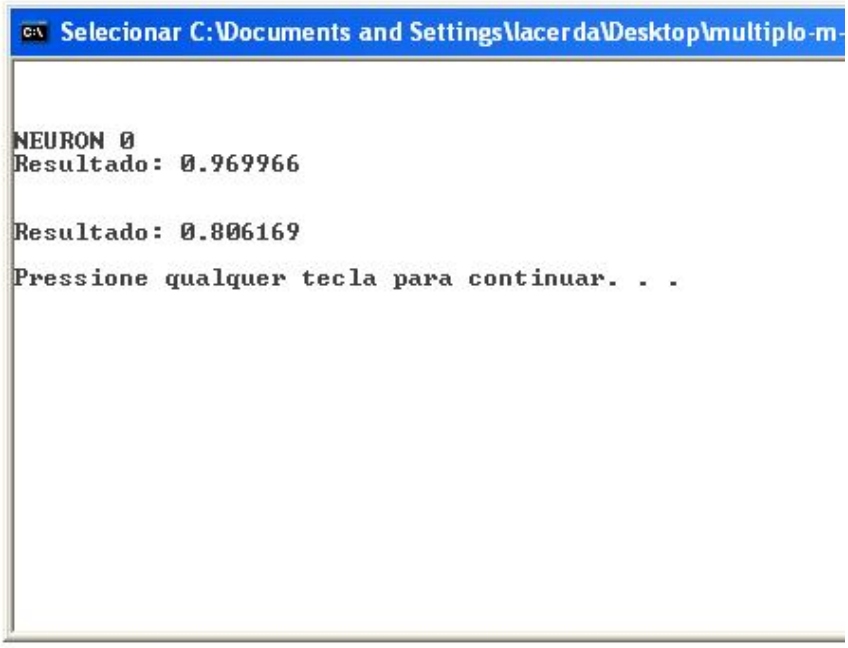

Figura 4.2. 2 Classificador analisando 2 amostras

4.3 Resultados Avançados

A partir de então, mais testes forma realizados utilizando um banco de dados maior de amostras, com o intuito de refinar a técnica de classificação. Nessa etapa, já tem-se disponíveis todas as amostras registradas, bem como seus vetores de identificação e bem relacionadas com os locutores da aquisição. Os experimentos a partir deste ponto têm caráter relacionado à SVM e de seu devido ajuste. As primeiras tentativas se resumiram à separação de amostras entre pessoas conhecidas pela rede, e pessoas desconhecidas. O gráfico de desempenho correspondente pode ser visualizado na Figura 4.3.1, onde cada número do eixo x representa uma pessoa, o eixo y representa a classificação dada pelo sistema quanto aos pontos dispersos, que constituem amostras da própria pessoa (em azul) e de pessoas desconhecidas pela rede (em vermelho).

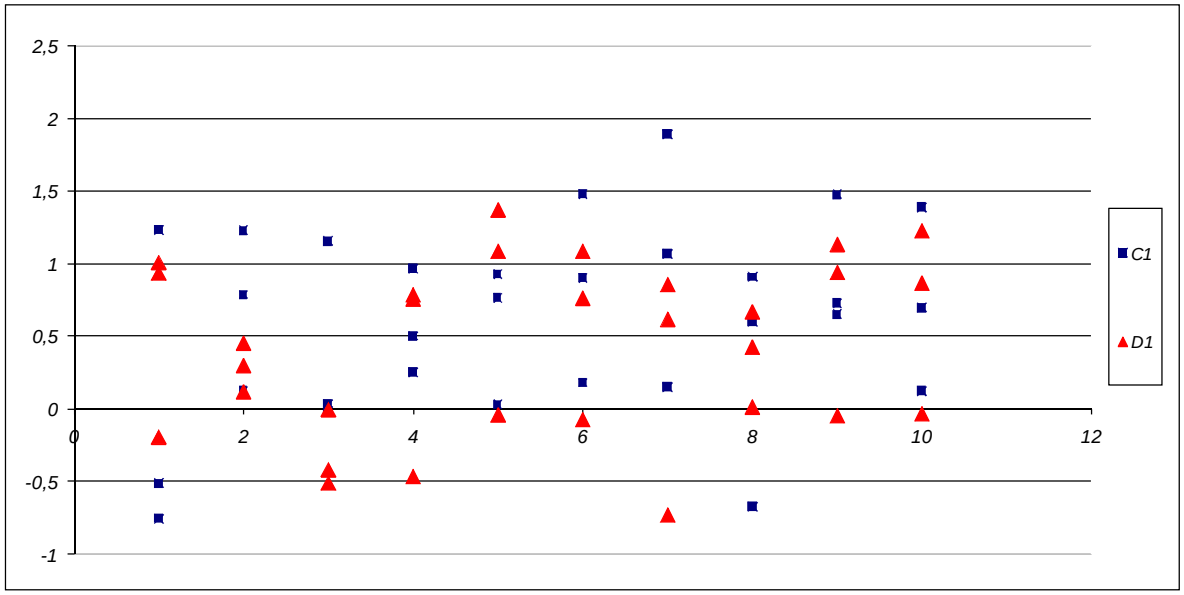

Figura 4.3. 1 Resultados

Inicialmente, duas classes foram definidas, sendo uma para as amostras conhecidas pela rede (com valor 1) e outro para as amostras desconhecidas pela rede (com valor -1). Com base na figura 4.3.1, é possível notar que alguns pontos estão próximos à classe "conhecida", outros a classificação de "desconhecido", e há ainda aqueles que não estão na região de fronteira ou distantes dos valores esperados. Assim sendo, foram observados alguns erros correspondentes à obtenção de valores 1 para locutores desconhecidos (pontos vermelhos) e valores -1 para locutores conhecidos (pontos azuis).

Inúmeros testes demonstraram que o esquema anterior de agrupamento foi pouco eficiente. Isso ocorreu devido ao fato de que se estava fornecendo um número fixo e limitado de exemplo para SVM representativos do universo de vozes fora do grupo conhecido. Assim, passou-se a trabalhar com apenas um agrupamento, onde somente o grupo de amostras de pessoas conhecidas é apresentado para a SVM. Neste último caso, obtiveram-se resultados muito mais eficazes, conforme demonstrado na Tabela 4.3.1, e no gráfico na Figura 4.3.2.

amostras	DESCON	CONHEC
1	0,972	1,002
2	0,962	1,000
3	0,965	1,000
4	0,982	0,995
5	0,996	1,004
6	0,998	1,001
7	0,981	1,000
8	0,989	0,980
9	0,934	1,000
10	0,997	1,001
11	0,997	1,003
12	0,977	1,001
13	0,995	0,999
14	0,985	1,002
15	0,996	0,994
16	0,990	0,996
17	0,997	1,003
18	0,999	0,999
19	0,982	1,004
20	1,003	1,002

Tabela 4.3.1 : Amostras classificadas pela SVM

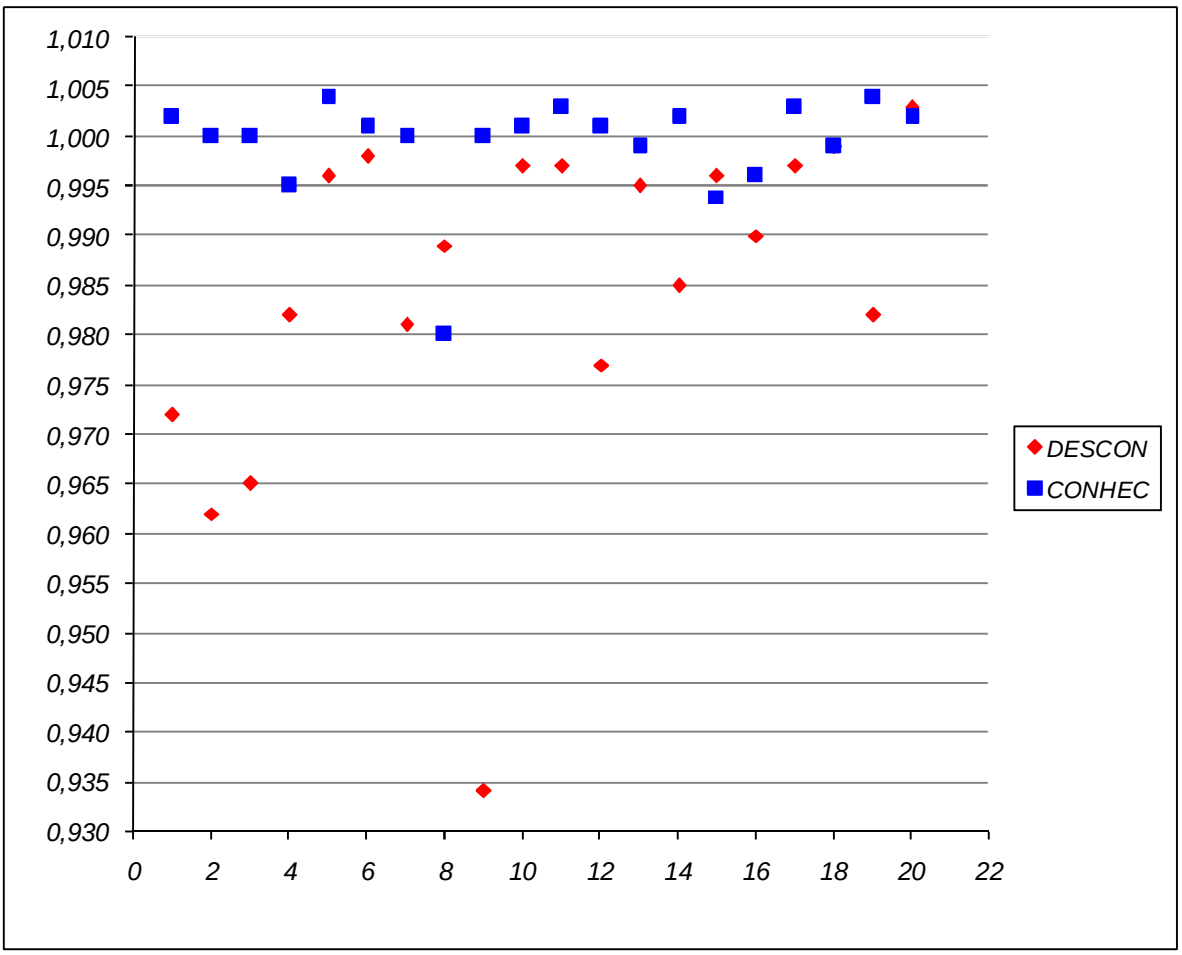

Figura 4.3. 2 Resposta da SVM treinada com base apenas nos locutores conhecidos após testes para refinamento.

Na figura 4.3.3 têm-se os mesmos dados da figura 4.3.2, porém com uma visualização diferenciada para contribuir em aspectos de resultados científicos. Nela é possível observar o comportamento e posicionamento das amostras de locutores conhecidos. As mesmas possuem um padrão de comportamento, que é situar-se na região de classificação 1,00 da SVM, e, portanto, caracterizando um reconhecimento de padrões, enquanto que as amostras de locutores desconhecidos estão fora desta faixa de 1,00 de classificação.

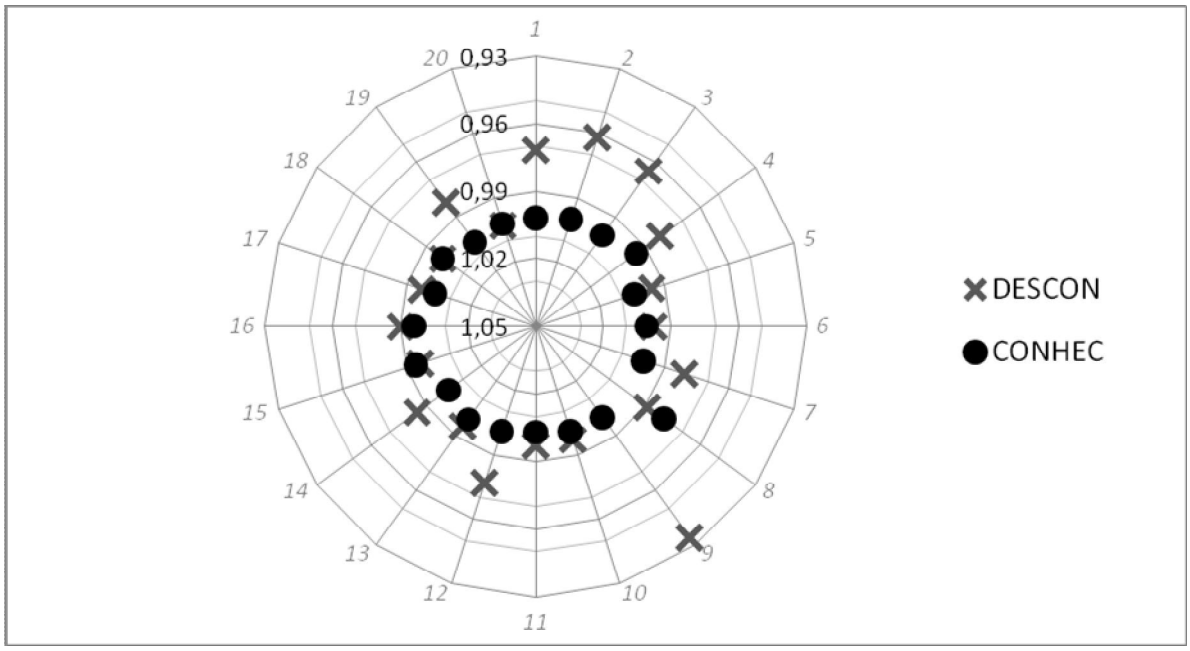

Figura 4.3. 3 Resposta da SVM treinada com base apenas nos locutores conhecidos após testes para refinamento, visualizada em teia.

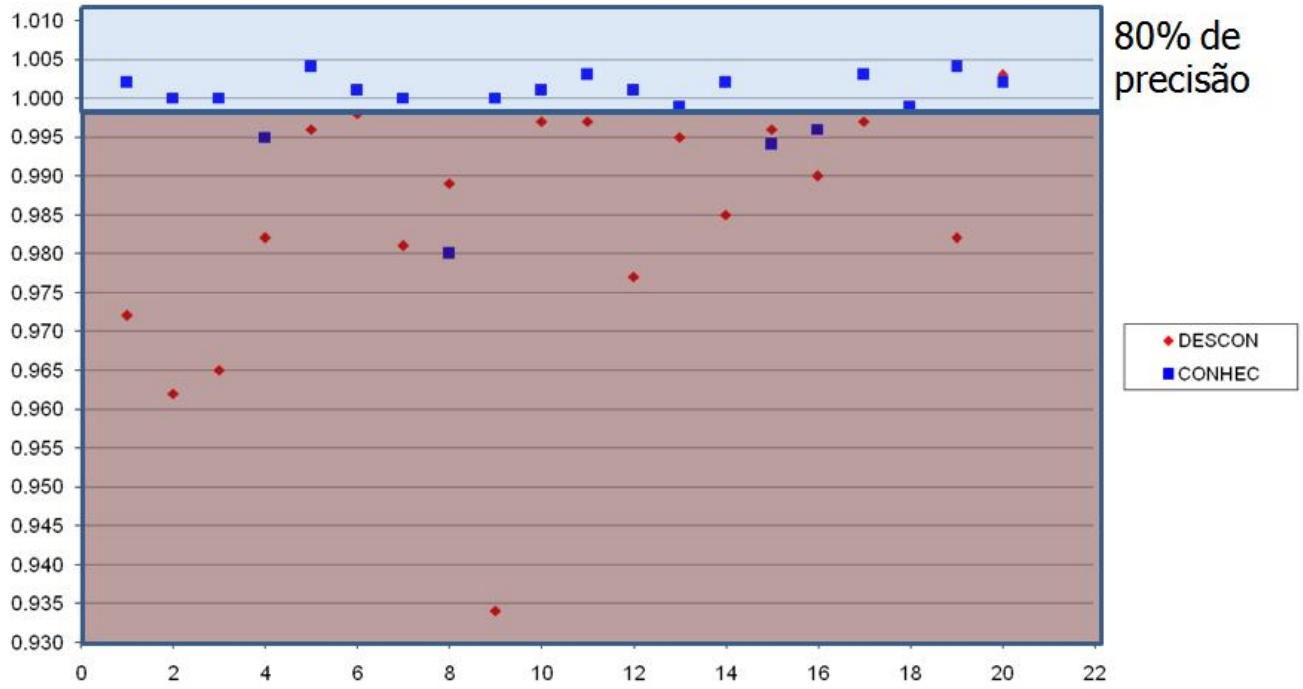

Figura 4.3. 4 Precisão de 80% na classificação dos locutores.

Visualiza-se que a classificação da SVM, agora com seus parâmetros ajustados atingiu sucesso, com uma verificação correta de 80% dos locutores testados, como reforçado na figura 4.3.4.

5. Trabalhos Futuros

Além do DSP ou do DSPic, o *hardware* necessário para a implementação do projeto ainda prevê para implementações futuras a possibilidade de um transmissor AM, onde podem captados os sinais de voz e transmitidos por radio freqüência, e um receptor AM, no qual podem ser recebidos os sinais de voz transmitidos por radio freqüência e passados ao processador, além das fontes de alimentação para, da forma apropriada, suprir as tensões utilizadas pelos dispositivos.

O transmissor AM é basicamente um dispositivo que transmite um sinal com modulação em amplitude. Nesse caso, trata-se dos sinais de vozes captados que, em seguida, serão recebidos por radio freqüência pelo receptor AM, que faz a demodulação e os passará ao DSP / DSPic para tratamento e classificação, conforme o algoritmo embarcado. Esses dispositivos devem ser alimentados com uma fonte CC (corrente contínua), que é uma configuração de circuito que transforma, retifica e acondiciona a tensão de corrente alternada da rede elétrica, em tensão de corrente contínua. A implementação do sistema está planejada conforme ilustra a figura 5.1.

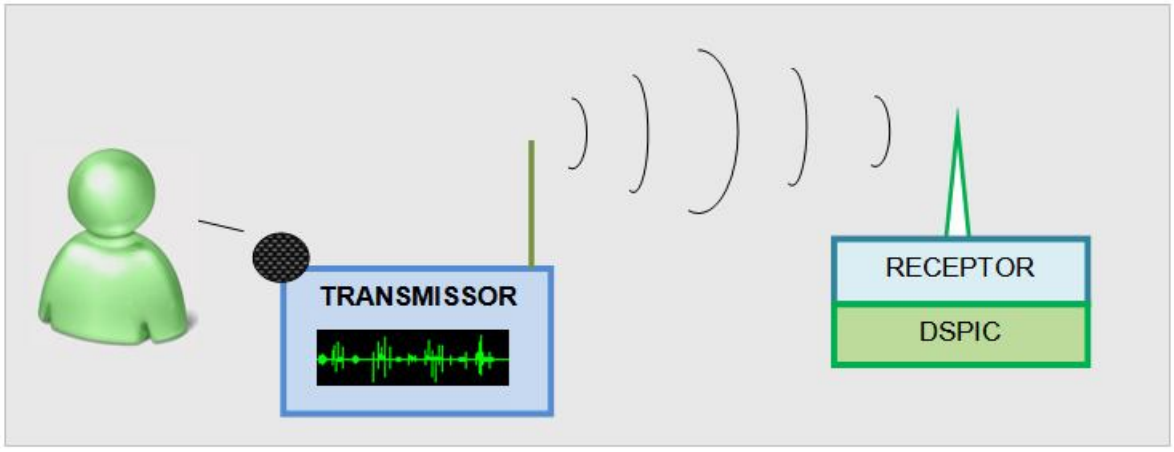

Figura 5. 1 Arquitetura do sistema proposto.

O uso de transmissor e receptor AM justifica-se pelo privilégio de ter-se uma estação fixa, na qual serão tratados os dados, e uma estação móvel, com a qual

poderá se captar os sinais de voz remotamente, e, com isso, utilizar o escopo do projeto dentro de um raio a partir da estação fixa, livremente, possibilitando maior número de aquisições, testes, e aplicações dentro das áreas de interesse citadas.

6. Artigo Elaborado

O artigo elaborado, intitulado "Wavelet based Speaker Verification" foi confeccionado paralelamente às atividades do projeto e publicado na edição de Novembro de 2010 no *International Journal on Wavelets, Multiresolution and Information Processing*. Nele, descreve-se o classificador aqui proposto.

O artigo encontra-se no Anexo A deste documento.

7. Conclusões

Após o estudo e implementação e desenvolvimento do projeto, é perceptível a realização satisfatória do mesmo, em vista dos resultados obtidos, tendo os experimentos e resultados sendo explicitados de forma detalhada, e com o alvo no produto final, uma abordagem eficiente para o reconhecimento de locutor. Além disso, o trabalho contribuiu significativamente para a formação do autor, para o andamento das pesquisas na área, bem como para a consolidação do grupo de pesquisas no qual o estudante está inserido.

Complementarmente o trabalho desenvolvido contribuiu para a pesquisa feita no *Speechlab*. E principalmente, foi possível desenvolver atividades integrativas, relacionando o tema trabalhado com os mais diversos itens estudados durante a graduação de Engenharia Elétrica.

7. Referências Bibliográficas

[1] Quartieri, T.F. Discrete-time Speech Signal Processing: Principles and Practice. New Jersey: Prentice-Hall, 2001.

[2] A.V. oppenhein e R.W. Schafer, "Discrete Time Signal Processing", 2. ed., Prentice-Hall, New York, 1999.

[3] S. Haykin e B.V. Veen, "Sinais e Sistemas", Bookman, Porto Alegre, 2001.

[4] Katagiri, S. handbook of Neural Networks for Speech Processing. Boston: Artech House, 2000.

[5] Lyons, R.D. Understandig Digital Signal Processing. 2 ed. New Jersey: Prentice Hall, 2004.

[6] Jain, A.; Hong, L.; Pankanti, S. Biometric Identifications. Communications of the ACM. N.43, v.2, p.90-98, 2000.

[7] Juang, B. H; Chou, W. Pattern Recognition in Speech and Language Processing. Boca Raton: CRC Press, 2003.

[8] Fant, G. Acoustic Theory of Speech Production. Mouton, The Haugue. 1970.

[9] Vieira, M.N. Uma Introdução à Acústica da Voz Cantada, I SMCT: AM, UFMG.

[10] Feder, J. Fractals. New York: Plenum Press, 1988.

[11] Chui, C.K. An Introduction to Wavelets: Wavelets Analysis and Its Applications. London: Academic Press, 1992.

[12] Feder,J. Fractals. New York: Plenum Press, 1988.

Anexo A

O artigo elaborado, intitulado "Wavelet based Speaker Verification" foi confeccionado paralelamente às atividades do projeto e publicado na edição de Novembro de 2010 no *International Journal on Wavelets, Multiresolution and Information Processing*. Neste anexo, o artigo aparece na formatação e diagramação idêntica a da publicação.

International Journal of Wavelets, Multiresolution
and Information Processing
Vol. 8, No. 6 (2010) 905–912

DOI: 10.1142/S0219691310003894

World Scientific
www.worldscientific.com

A WAVELET-BASED SPEAKER VERIFICATION ALGORITHM

MICHEL ALVES LACERDA, RODRIGO CAPOBIANCO GUIDO*,
LEONARDO MENDES DE SOUZA, PAULO RICARDO FRANCHI ZULATO
and JUSSARA RIBEIRO

SpeechLab/FFI/IFSC/USP, Department of Physics and Informatics
Institute of Physics at São Carlos, University of São Paulo
Avenida Trabalhador São Carlense 400
13566-590, São Carlos, SP, Brazil
**guido@ifsc.usp.br*

SHI-HUANG CHEN

Department of Computer Science and Information Engineering
Shu-Te University, N 59, Hengshan Rd., Yanchao
Kaohsiung County, 82445, Taiwan, R. O. C.

Received 20 May 2009
Revised 7 July 2010

This paper presents a study on wavelets and their characteristics for the specific purpose of serving as a feature extraction tool for speaker verification (SV), considering a Radial Basis Function (RBF) classifier, which is a particular type of Artificial Neural Network (ANN). Examining characteristics such as support-size, frequency and phase responses, amongst others, we show how Discrete Wavelet Transforms (DWTs), particularly the ones which derive from Finite Impulse Response (FIR) filters, can be used to extract important features from a speech signal which are useful for SV. Lastly, an SV algorithm based on the concepts presented is described.

Keywords: Speaker verification; discrete wavelet transform; FIR filters.

1. Introduction

Recently, we have witnessed much research in the areas of speech processing[1] and wavelets.[2–4] The former area consists of a wide field which spans many different applications: speech recognition, speaker recognition, speaker verification, spoken document summarization, voice conversion, and many others. Particularly, speaker verification (SV), which is the application treated here, consists of a procedure used to verify whether or not a speaker is authorized to gain access into some system. In this case, the algorithm does not identify who the speaker is, instead it presents

*Corresponding author.

only a binary output: *enrolled* or *not enrolled* in the database. The latter area of research, i.e. wavelet transform, has being intensively studied and used to solve a variety of problems in signal processing, mainly the ones involving time-frequency analysis.[5]

In the speech processing area, Mel Frequency Cepstral Coefficients (MFCCs), their deltas, and delta-deltas[1] are commonly used as input features which characterize a speaker, while Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs),[1] and Artificial Neural Networks (ANNs)[8] are intensively used as classifiers for speaker verification purposes. For any given classification problem, however, the accuracy depends on the particular classifier adopted and not only the parameters used as input to it, thus, it is not possible to say that a certain set of features are, in some sense, the best ones for some particular classification problem because the classifier itself highly influences the results.

This paper intends to show the role of the Discrete Wavelet Transform (DWT)[2] as being a feature extraction tool for SV purposes, discussing their important characteristics, and presenting a particular algorithm which uses DWTs in conjunction with a Radial Basis Function (RBF) ANN[8] for SV. This algorithm is characterized as being text-independent,[1] i.e. it verifies the speaker whatever the sentence or utterance that he or she is pronouncing.

The paper is organized as follows. Section 2 reviews the wavelet transform theory including particular characteristics which are important for speaker verification. The proposed system architecture and a discussion are presented in Sec. 3, tests and results are presented in Sec. 4, and, lastly, Sec. 5 describes the conclusions.

2. Review on the Literature: The Discrete Wavelet Transform

According to the DWT theory,[2] the jth-level decomposition of a given discrete (speech) signal, $f[n]$, can be written as[9]:

$$f[n] = \sum_{k=0}^{\frac{n}{2^j}-1} R_{j,k}[n]\phi_{j,k}[n] + \sum_{t=1}^{j}\sum_{k=0}^{\frac{n}{2^j}-1} S_{t,k}[n]\psi_{t,k}[n], \tag{2.1}$$

$\phi[n] = \sum_k h[k]\phi[2n-k]$ and $\psi[n] = \sum_k g[k]\phi[2n-k]$ being the scaling and wavelet functions, respectively, that form a Riesz basis[2] to write signal f, $R_{j,k}[n] = \langle f, \phi_{j,k}[n]\rangle$, $S_{t,k}[n] = \langle f, \psi_{t,k}[n]\rangle$, and $h[k]$ and $g[k] = (-1)^k h[N-k-1]$ being the quadrature mirror (QMF) low-pass and high-pass analysis filters,[2] respectively.

When $f[n]$ is being analyzed, low-pass and high-pass filtering occur by discrete convolutions, followed by downsamplings by 2, and, therefore, the length of $h[k]$ and $g[k]$, N, is responsible for both *frequency selectivity*, Q, and *time resolution*, R. The value of N is defined here as being the *requirement* 1. Another important consideration, the *requirement* 2, is that linear phase filters $h[k]$ and $g[k]$ are desirable to avoid distortion in the filtered signal, i.e. symmetrical or anti-symmetrical impulse responses are certainly preferable. Requirements 1 and 2 will be discussed in the next section.

The calculation of the DWT can be performed with the use of Mallat's algorithm,[5] which requires only the filters $h[n]$ and $g[n]$ in addition to the input signal. Particularly, if A is the $N \times N$ matrix formed by the coefficients of $h[n]$ and $g[n]$, and B is the $N \times 1$ matrix containing the original input speech signal in the time-domain, then, $C = A \cdot B$ corresponds to the transformed signal in the wavelet domain. The even coefficients of C $(c_0, c_2, \ldots, c_{n-2})$ form the *approximation signal* and the odd ones $(c_1, c_3, \ldots, c_{n-1})$ form the *detail signal*. The matrices, that embed the processes of downsampling and wraparound,[5,6] are formed as follows:

$$A[\cdot][\cdot] = \begin{pmatrix}
h_0 & h_1 & h_2 & \cdots & \cdots & \cdots & \cdots & h_{n-1} & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
g_0 & g_1 & g_2 & \cdots & \cdots & \cdots & \cdots & g_{n-1} & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
0 & 0 & h_0 & h_1 & h_2 & \cdots & \cdots & \cdots & \cdots & h_{n-1} & 0 & 0 & \cdots & \cdots & 0 & 0 \\
0 & 0 & g_0 & g_1 & g_2 & \cdots & \cdots & \cdots & \cdots & g_{n-1} & 0 & 0 & \cdots & \cdots & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & & & \vdots & & \vdots & & & & \vdots & & & & \vdots & & \vdots \\
h_{n-2} & h_{n-1} & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & h_0 & h_1 & \cdots & \cdots & \cdots & h_{n-4} & h_{n-3} \\
g_{n-2} & g_{n-1} & 0 & 0 & \cdots & \cdots & \cdots & 0 & 0 & g_0 & g_1 & \cdots & \cdots & \cdots & g_{n-4} & g_{n-3}
\end{pmatrix},$$

$$B[\cdot] = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \cdots \\ \cdots \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{pmatrix}, \quad C[\cdot] = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \cdots \\ \cdots \\ \cdots \\ \vdots \\ c_{n-2} \\ c_{n-1} \end{pmatrix}.$$

Since no inverse transformation will be required, we will not concentrate our review on the functions ϕ and ψ. A closer look at Eq. (2.1) calls our attention to the fact that the inverse DWT, obtained by multiplying $A^{-1} = A^T$ by C, corresponds to writing $f[n]$ as a linear combination of such functions, so it would be important to study them with more details if the inversion were needed for our implementation. Instead of using the DWT itself, the proposed approach uses the full decomposition tree, i.e. the Discrete Wavelet-Packet Transform (DWPT),[5] and then the sub-bands are rearranged according to the natural frequency ordering.[7] All the concepts discussed above for the DWT are equally valid for the DWPT. Figure 1 shows the impulse responses shapes of the most common families of wavelets, that are characterized as being Finite Impulse Response (FIR) filters. Infinite Impulse Response (IIR) filters were not considered since they do not exhibit linear phase responses.

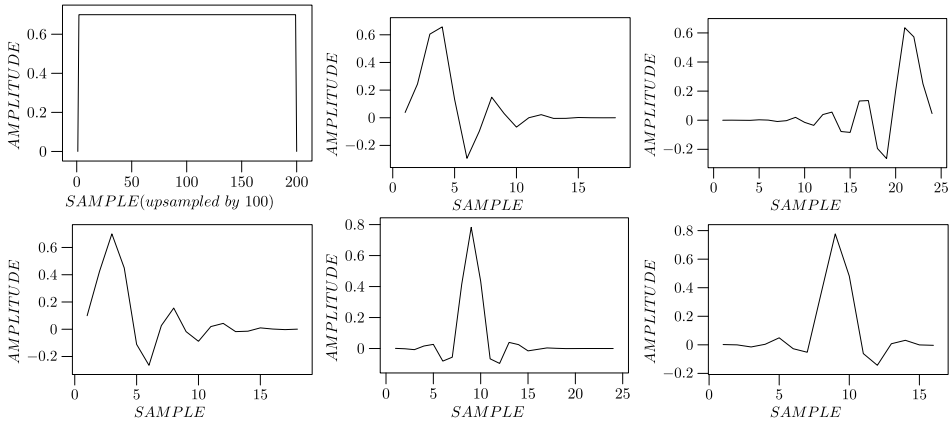

Fig. 1. Impulse responses' shapes of the wavelet filters Haar, Daubechies, Vaidyanathan, Beylkin, Coiflet, and Symmlet.

3. The Proposed Algorithm for SV

The proposed wavelet-based algorithm for SV is described in Tables 1 and 2. The former table describes the procedure required to train the classifier and the latter table describes the procedure to test and use it. As mentioned above, this algorithm

Table 1. The algorithm for training the proposed wavelet and RBF-based classifier.

- BEGINNING
- **STEP T-1**: Define the source speech signal to be analyzed, $s[\cdot]$, that should be sampled at 16,000 samples per second, 16-bit;
- **STEP T-2**: Divide $s[\cdot]$ into n frames with 256 samples each. If the last frame contains less than 256 samples, it can be zero-padded or discarded;
- **STEP T-3**: For each frame i of $s[\cdot]$, i.e. $s_i[\cdot]$, $(0 \leq i \leq n)$, do:
  - **STEP T-3.1**: Subtract the mean of $s_i[\cdot]$ from all its 256 samples so that the frequency 0 is removed;
  - **STEP T-3.2**: If $s_i[\cdot]$ is a voiced speech frame, then:
    - **STEP T-3.2.1**: Obtain the 8th level Discrete Wavelet-Packet Transform (DWPT) of $s_i[\cdot]$. The family of wavelets to be used, and the corresponding support-size of the filters, will be discussed later;
    - **STEP T-3.2.2**: Obtain the energies, i.e. the sum of the squared samples, of 21 sub-bands of DWTP($s_i[\cdot]$), according to the critical bands of the human auditory system[10] that are listed on Table 3;
    - **STEP T-3.2.3**: For each sub-signal j of DWTP($s_i[\cdot]$), $(0 \leq j \leq 20)$, do:
      - **STEP T-3.2.3.1**: Use the sub-signal j to train the RBF neural network number j, DWTP$_j$($s_i[\cdot]$) serving as input, and the value 1 serving as output to the classification scheme meaning that this is an authorized speaker, i.e. not an intruder;
  - else
    - do not consider this frame;
- END.

Table 2. The proposed procedure to test and use the classifier.

- BEGINNING
- **STEP U-1**: Create the array $R[\cdot]$ with 21 double-precision positions and initialize all of them with the value 0;
- **STEP U-2**: Define the test speech file, $t[\cdot]$, that should be sampled at 16,000 samples per second, 16-bit;
- **STEP U-3**: Divide $t[\cdot]$ into n frames with 256 samples each. If the last frame contains less than 256 samples, it can be zero-padded or discarded;
- **STEP U-4**: For each frame i of $t[\cdot]$, $(0 \leq i \leq n)$, do:
  - **STEP U-4.1**: Subtract the mean of $t_i[\cdot]$ from all its 256 samples so that the frequency 0 is removed;
  - **STEP U-4.2**: If i is a voiced speech frame then:
    - **STEP U-4.2.1**: Obtain the 8th level Discrete Wavelet-Packet Transform (DWPT) of $t_i[\cdot]$ by using the same family of wavelets used to train the classifier;
    - **STEP U-4.2.2**: Obtain the energies, i.e. the sum of the squared samples, of 21 sub-bands of DWTP($t_i[\cdot]$), according to the critical bands of the human auditory system[10] that are listed on Table 3;
    - **STEP U-4.2.3**: For each sub-signal j of DWTP($t_i[\cdot]$), $(0 \leq j \leq 20)$, do:
      - **STEP U-4.2.3.1**: Use the DWTP$_j(s_i[\cdot])$ as input to the RBF neural network number j, storing its output in the position j of the array $R[\cdot]$ in a cumulative way, i.e. $R_j \leftarrow R_j +$ *the corresponding RBF output*;
  - else
    - do not consider this frame;
- **STEP U-5**: For $(0 \leq i \leq 20)$ do:
  - **STEP U-5.1**: $R_i \leftarrow \frac{R_i}{n}$, i.e. obtain the 21 means of the RBF outputs;
- **STEP U-6**: Measure the Euclidean distance between the array $R[\cdot]$ and the array $A[\cdot] = \{1, 1, \ldots, 1\}$, both of them being 20-sample long. If the distance between $R[\cdot]$ and $A[\cdot]$ is less than the threshold δ, defined ahead, then the speaker is considered as being enrolled, otherwise he or she is not enrolled.
- END.

is text-independent, i.e. it verifies a speaker independent of the sentence or utterance that he or she is pronouncing. To do so, it analyzes only the voiced speech tags of the spoken sentence, i.e. the parts which contain a quasi-periodic excitation signal produced by the lungs in conjunction with the vocal folds.[1] The technique used to separate between voiced and unvoiced tags is the zero-crossing rate, described in Ref. 1.

To sum up, the algorithm uses 21 RBF ANNs which receive the energies of 21 corresponding critical sub-bands of frequencies from the source speech frames. Therefore, each RBF is trained to interpret the vocal information which is related to specific parts of the human vocal system, such as excitation and its control by the lungs and vocal folds, acoustic resonances of the vocal tract, i.e. formant frequencies,[1] and so on.

The intention of the proposed algorithm is that, during the filtering procedures, each sub-band contains a particular set of frequencies with a superposition of the

adjacent sub-bands, such as the triangular filtering scheme used to obtain the Mel Frequency Cepstral Coefficients (MFCCs),[1] that are traditionally used as feature extractors in speech processing. To do so and to define the value of N mentioned in the last section, which is directly related to requirements 1 and 2, we have to choose Haar wavelets. Although the frequency response of such family of filters is far away from being ideal, it is the only one which provides linear phase response and the desired overlap between each adjacent sub-band, just as the filters used with MFCCs do. Furthermore, this family of wavelets contains filters with support-size equals 2 which provide a very fast convolution scheme.

4. Tests and Results

We have tested the proposed approach with a set of voices collected from 40 people (20 men and 20 women) of all ages. Each person's voice were recorded for five seconds with a professional digital recorder using a high impedance crystal capsule, while the individual was pronouncing a sentence that was chosen randomly among the ones we extracted from a digital signal processing book written in Brazilian Portuguese language. Fifty percent of each speaker's data was used to train the system and the other 50% was used to test it, ensuring that a large set of voiced phonemes were applied during the training stage. The results, that are shown in Fig. 2, demonstrate the efficacy of the proposed method and the proposed wavelet family used, i.e. Haar. Adopting the threshold $\delta = 0.997$, obtained empirically, we got only four enrolled speakers and two non-enrolled speakers who were wrongly classified, which represent, respectively, 10 and 5% of the total. When we enroll

Table 3. The 25 critical bands of the human auditory system. As the speech signals are sampled at 16,000 Hz, i.e. contain frequencies under 8,000 Hz, only the first 21 bands (0 to 20) are used. The last used band, i.e. band 20, ranges from 6,400 to 7,700 Hz, however, for the current application we approximate this to the range 6,400 to 8,000 Hz. The columns IS and FS represent, respectively, the initial and final samples of the 8th-level DWPT of a speech frame, which is 256-sample long, that approximately correspond to the critical band frequency content.

Band	Frequency (Hz)	IS	FS	Band	Frequency (Hz)	IS	FS
0	0–100	0	2	13	2,000–2,320	64	73
1	100–200	3	5	14	2,320–2,700	74	85
2	200–300	6	9	15	2,700–3,150	86	100
3	300–400	10	12	16	3,150–3,700	101	118
4	400–510	13	15	17	3,700–4,400	119	141
5	510–630	16	19	18	4,400–5,300	142	170
6	630–770	20	24	19	5,300–6,400	171	205
7	770–920	25	28	20	6,400–7,700	206	255
8	920–1,080	29	34	21	7,700–9,500	—	—
9	1,080–1,270	35	40	22	9,500–12,000	—	—
10	1,270–1,480	41	46	23	12,000–15,500	—	—
11	1,480–1,720	47	54	24	15,500–22,050	—	—
12	1,720–2,000	55	63				

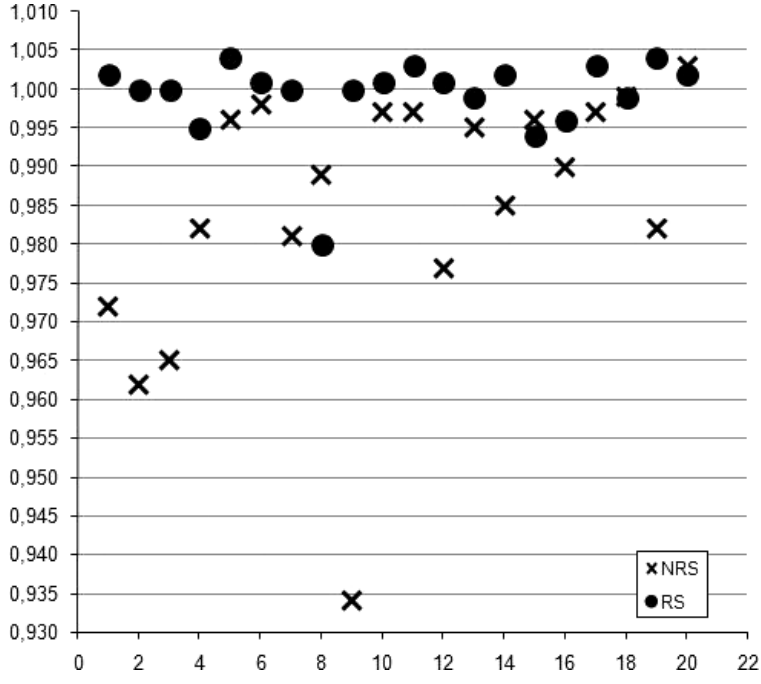

Fig. 2. Results of the tests on the trained system using the procedure described in the text. (NRS): intruder speaker; (RS): enrolled speaker. The threshold δ adopted is 0.997 in this case. All the sentences used required over 5 seconds to be uttered, therefore only the 5 initial seconds were used, always discarding the initial 2,048 samples to avoid the introduction of artifacts.

more speakers in the database, the previous reported error rates tend to decrease in a quasi-linear proportion, as we could observe. These values can be considered as being a good result due to the simplicity of the proposed algorithm, showing that the RBF classifier can certainly benefit from the wavelet-based tool for parameter extraction. Tests were also performed to the other families of wavelets discussed in the text, however, no effective improvement was noted, furthermore, a delay was introduced proportionally to the support-size of the filter used. We observed that the algorithm has a low order of computational complexity, thus, its real-time implementation based on a Field Programmable Gate Array (FPGA) is being planned. For this implementation, the fact of using Haar wavelets is certainly a tremendous advantage which facilitates the implementation using languages such as VHDL.

5. Conclusions

We proposed an interesting method for speaker verification that is based on wavelets that has a low computational cost. The proposed approach runs relatively fast due to the reduced support-size of the Haar wavelets and the optimized implementation

of the RBF ANNs. The results show the effectiveness of the algorithm, which can also be extended for use in real time.

Acknowledgments

We wish to thank the State of São Paulo Research Foundation for the grants given to this work under process nr. 2005/00015-1.

References

1. L. Deng and O. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach* (Marcel Dekker Inc., 2003).
2. P. S. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance* (Institute of Physics Publishing, Edinburg, 2002).
3. C.-C. Chiu, C.-M. Chuang and C.-Y. Hsu, Discrete wavelet transform applied on personal identity verification with ECG signal, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(3) (2009) 341–355.
4. A. B. Mabrouk, H. Kortas and Z. Dhifaoui, A wavelet support vector machine coupled method for time series prediction, *Int. J. Wavelets Multiresolut. Inf. Process.* **6**(6) (2008) 851–868.
5. G. Strang and T. Nguyen, *Wavelets and Filter Banks* (Wellesley-Cambridge Press, 1997).
6. R. S. Pathak and A. Pathak, On convolution for wavelet transform, *Int. J. Wavelets Multiresolut. Inf. Process.* **6**(5) (2008) 739–747.
7. A. Jensen and A. Cour-Harbo, *Ripples in Mathematics: The Discrete Wavelet Transform* (Springer-Verlag, 2000).
8. S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd edn. (Prentice-Hall, Upper Saddle River, New Jersey, 1998).
9. R. C. Guido, L. S. Vieira, S. Barbon Jr., F. L. Sanchez, C. D. Maciel, E. S. Fonseca and J. C. Pereira, A neural-wavelet architecture for voice conversion, *Neurocomputing* **71** (2007) 174–180.
10. M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*, 2nd edn. (Kluwer Academic Publishers, Massachusetts, 2003).