Reconhecimento Automático de Locutor utilizando redes neurais artificiais

Autor
Rodrigo de Toledo Caropreso

Resumo— *Reconhecimento de padrões é um problema bastante estudado e abordado de diversas formas. Particularmente, o reconhecimento de fala mostra-se como uma área onde há ainda muito a ser explorado.* As redes neurais são uma técnica que auxiliam na tarefa de reconhecimento de padrões, dada suas características e vantagens frente à natureza não-estruturada dos padrões. Principalmente nos padrões de fala, onde uma estrutura exata é difícil de ser definida face a grande variabilidade existente, provocada por entonação, timbre, sotaque, etc. Visando diminuir esta variabilidade, algumas restrições são impostas às aplicações, procurando torná-las menos complexas.

Palavras-chave— Redes Neurais, locutor, reconhecimento, padrões, perceptron, cepstrum.

1 - Introdução

O reconhecimento da voz humana e de um interlocutor tem sido objeto de estudo cada vez mais frequente na atualidade [CORSI, 81],[JACOBS,88].

O Reconhecimento Automático de Locutores (RAL) tem despertado, durante as últimas décadas um profundo interesse por parte da comunidade científica e, por conta disso, tem sido fruto de amplas pesquisas.

Como resultado destas pesquisas foi possível elaborar procedimentos e algoritmos capazes de processar e efetuar o RAL. Entretanto, apesar dos avanços obtidos pelas pesquisas, ainda não foi possível obter uma máquina capaz de compreender toda a fala de um ser humano.

Ainda assim, o avanço da tecnologia, dos computadores e das telecomunicações tem servido como fator de ampliação das pesquisas na área do RAL e suas possibilidades: aplicações de automação bancária, controle de acesso à informações, sistemas de segurança e outras [HATON, 81].

De uma forma mais simples, podemos definir o RAL como sendo a capacidade de reconhecer uma determinada pessoa através da sua voz, sendo uma variação do Reconhecimento de Voz, outro assunto amplamente estudado por diversos pesquisadores [PICONE, 93], [FURUI, 89].

Entretanto, no estudo do Reconhecimento de Voz as variações referentes aos locutores não são consideradas, enquanto que no RAL. toda e qualquer variação intra-locutor é amplamente levada em consideração para as análises.

Em paralelo à estas pesquisas, os estudos na área de Inteligência Artificial e Redes Neurais Artificiais (RNA) têm se intensificado e novas técnicas vêm sendo introduzidas pela comunidade científica ao longo dos últimos anos. a exemplo de [SAMBUR, 75] e [ATAL, 76].

Diversos trabalhos sobre as Redes Neurais Artificiais [CASAGRANDE, 97],[ELMAN, 90], [JORDAN,86], [MAGNI, 98], [TMOSZCZUCK,98] mostram a viabilidade do uso de classificadores neurais. Os Classificadores do tipo Multi-Layer Perceptron (MLP) têm sido amplamente utilizados em problemas de identificação do locutor, devido à sua simplicidade e eficiência.

Por este motivo a rede MLP será usada como modelo de classificador padrão no estudo deste trabalho e seu desempenho servirá como balizador para comparação com as demais redes.

A proposta deste trabalho é avaliar a o desempenho de redes neurais artificiais no processo de reconhecimento automático de locutor.

2 – Descrição do problema

2.1 – O Processo de Formação da Voz

O problema de reconhecimento de fala é de difícil tratabilidade. A maior dificuldade é a sua natureza interdisciplinar. Além dessa, variabilidades acústicas, do transdutor, intra-locutor, entre locutores estão relacionadas com o problema. Mas. em contra partida, várias áreas podem ser beneficiadas com o uso desta técnica, tornando um campo desafiador para ser pesquisado.

Entretanto, no estudo do Reconhecimento de Voz as variações referentes aos locutores não são consideradas, enquanto que no RAL, toda e qualquer variação intra-locutor é amplamente levada em consideração para as análises. Alguns estudos abrangentes sobre o RAL podem ser vistos em [ATAL, 76], [CORSI, 81] e [GISH, 94].

Os sinais de voz são produzidos pelo aparelho fonador (ou trato vocal) humano, através de vibrações das cordas vocais em cavidades situadas entre a faringe e a boca[FLANAGAN, 72], [RABINER, 78].

O ar entra nos pulmões através da respiração e quando é expelido faz com que as cordas vocais vibrem. Dependendo da posição das articulações, as vibrações produzidas pelas cordas vocais sofrem modificações produzindo, assim, diferentes tipos de sons.

Estas modificações podem ser modeladas por uma função de transferência [RABINER, 78], [ATAL, 74], relacionando a entrada (excitação das cordas vocais) com a saída (produção da voz):

$$F(z) = \frac{G}{1 - \sum_{k=1}^{N} \alpha_k \cdot z^{-k}} \quad (1)$$

onde os α_k são parâmetros dependentes das características e dimensões do trato vocal e G representa o ganho do sistema.

Com relação a (1), cabe comentar que o modelo do trato vocal representado por esta equação consiste de uma simplificação (modelo só de polos).

O modelo ideal do trato vocal inclui polos e zeros, onde estes últimos correspondem à modelagem dos fonemas nasais e fricativos. A simplificação adotada para obter a (1), faz com que apenas a informação de amplitude da densidade espectral de potência seja utilizada, o que é considerado aceitável para a grande maioria das aplicações uma vez que o ouvido humano não “percebe” a informação de fase e, no caso do modelo só de polos, apresenta uma resposta de fase mínima.

O modelo apresentado representa características individuais que dependem de um conjunto de fatores, entre os quais podemos destacar: variações nas formas físicas dos tratos vocal e nasal, ressonâncias nas cordas vocais, massa e elasticidade dos músculos das paredes dos tratos, posição da língua, espaçamento e formação dos dentes e viscosidade da saliva.

Por ser um sinal de variação temporal lenta, quando examinado em períodos de tempos suficiente pequenos (entre 5 e 100ms, embora nas aplicações práticas seja muito comum considerar intervalos inferiores a 10ms), suas características podem ser consideradas estacionárias.

2.2 – Sistemas Computacionais de Reconhecimento Automático de Locutor

Um sistema computacional que efetue o RAL, pode ser representado pelas seguintes etapas:

1. Fase de Preparação do Sistema
    - Codificação do Sinal de Voz;
    - Pré-Processamento;
    - Extração de Features;
    - Armazenamento dos Padrões;
2. Fase de Reconhecimento do Sistema (operação)
    - Codificação do Sinal de Voz;
    - Pré-Processamento;
    - Extração de Features;
    - Comparação e Análise;
    - Aplicação de Critérios de Decisão.

A representação do sinal de voz no meio digital é denominada digitalização e envolve as seguintes etapas: a *amostragem* que consiste em tomar amostras periodicamente no tempo, a *codificação* que determina o número de bits que irá representar cada uma das amostras obtidas e a *quantização* que fixa o valor representado pelos bits que codificam cada uma das amostras.

A etapa de pré-processamento se subdivide em:

- Edição dos sinais de voz;
- Normalização dos sinais de voz;
- Segmentação dos sinais de voz;
- Pré-ênfase;
- Janelamento;

A etapa de extração de atributos consiste em, a partir de um dado sinal de voz preprocessado, extrair os parâmetros representativos daquele sinal, obedecendo a critérios previamente estabelecidos, conforme reportado por [WOLF, 72].

Com a intensificação de pesquisas nesta área, ao longo dos anos, foram desenvolvidas diferentes técnicas para extração de atributos dos sinais de voz. Algumas destas técnicas foram inicialmente desenvolvidas para reconhecimento da fala, mas se

mostraram úteis para o RAL, sendo mais conhecidas e utilizadas, as seguintes:

- **Análise da energia**: Este tipo de análise tem a finalidade de obter a variação de energia ao longo de uma dada locução de voz, através do cálculo de parcelas de energia em intervalos ponderados por uma *Window Function* (janelas), como é mostrado em [SCHAFER,75], [GRAY,80]. O janelamento é efetuado tendo por objetivo realizar uma ponderação, destacando o trecho central da janela. Analisando a Teoria de Processamento Digital de Sinais (como descrito por [HARRIS, 78]), podem ser encontradas várias janelas que podem ser utilizadas, entre elas: retangular, *Hamming, Hanning, Kaiser, Blackman* e *Barlett*.

- **Análise espectral de Tempo Curto (STFT)**: A ideia deste tipo de análise é considerar um fato já exposto na modelagem do trato vocal**,** a saber: a propriedade de invariância no tempo para segmentos curtos de locuções de voz. A análise utiliza janelas de comprimento igual, como por exemplo janelas de *Hamming* e são calculadas as componentes de frequência do sinal em cada janela, através da *Short Time Fourier Transform (STFT).*

- **Bancos de Filtros**: A análise com Banco de Filtros é uma técnica utilizada para obter uma representação de uma locução de voz no domínio das frequências e consiste na aplicação de uma série de filtros passa-banda digitais, espaçados ao longo da faixa de frequências do sinal de voz (de zero a $f_{Nyquist}$).

- **Cepstro**: Com base no modelo utilizado para a geração do sinal de voz descrito na Seção 2.1 deste trabalho, o problema de análise do sinal de voz, em segmentos suficientemente curtos, pode ser encarado como um problema de separação das componentes de uma convolução da função excitação (pulsos quase periódicos ou ruído branco) com a função de resposta impulsiva do trato vocal, ou seja, um problema de deconvolução proveniente do modelo do trato vocal.

O cepstrum apresenta-se como uma excelente ferramenta para estimar a frequência fundamental de um trecho de fala, ou para determinar a presença do sinal de voz. Existem variantes (Delta_cepstrum) porém será utilizado no processamento de voz apresentado neste trabalho a versão mais consagrada na literatura envolvendo o RAL, conforme será descrito adiante.

2.3 – Os Coeficientes Mel-Cepstrais

Os coeficientes Mel-Cepstrais têm sido os mais utilizados na tarefa do RAL. Para compreender melhor sua importância é fundamental levar em conta a avaliação de [PICONE, 93] sobre os sistemas de processamento de voz atuais, onde as representações utilizadas compartilham importantes características.

Em primeiro lugar são constituídas por informações de origem temporal e em segundo lugar tendem a ser perceptualmente compatíveis com o sistema auditivo humano, por exemplo o uso do logaritmo da energia (potência representada em decibéis) ou o uso de bancos de filtros distribuídos segundo escalas perceptuais (escala Bark ou Mel) na análise espectral, os quais procuram emular a resposta do ouvido humano.

A escala Mel foi apresentada pela primeira vez nos trabalhos de [STEVENS, 40].

As expressões a seguir definem o mapeamento das frequências acústicas *f* para as escalas perceptuais Bark ,definida como *critical band rate,* e Mel:

$$Bark = 13\, arctg\left(\frac{0.76\mathrm{f}}{1000}\right) + arctg\left(\frac{f^2}{7500^2}\right) \quad (2)$$

$$Mel = 2595 \log\left(\frac{f}{700}\right) \quad (3)$$

Os coeficientes Mel-Cepstrais são definidos como sendo a Transformada Discreta do cosseno (DCT) da saída logarítmica de um banco de filtros triangulares distribuídos de acordo com a escala Mel. A expressão abaixo define este conceito:

$$C_p(j) = \frac{2}{K}\sum_{k=1}^{K} \log\left(X_k(j)\right) \cdot \cos\left(p(k-0.5)\frac{\pi}{K}\right) \quad (4)$$

onde $C_p(j)$ é o *p-ésimo* coeficiente mel cepstral da janela *j*, *K* é o número de filtros e $X_k(j)$ é a energia do *k* filtro calculada para a janela *j*.

2.4 – Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs), pertencem a uma classe de sistemas que são constituídos de unidades processadoras simples.

Os estudos que deram origem às RNAs, foram motivados pelo objetivo de compreender o funcionamento do cérebro humano e a ideia de simular o comportamento do cérebro, e fizeram com que fossem buscados modelos matemáticos para os neurônios.

Uma rede neural resulta da interconexão de vários nós básicos em várias configurações. Uma das Redes mais conhecidas e utilizadas para diversos tipos de problema é o *Multi-Layer Perceptron* (ou Perceptron Multi-Camada - PMC)

O PMC faz parte de um conjunto de redes denominadas *feed-forward*, possuindo como estrutura básica camadas de neurônios na qual a saída de um neurônio numa camada alimenta todos os neurônios da camada seguinte. O fundamental desta estrutura é que não existem laços de realimentação.

Definido um vetor de entrada $x=[x_1, x_2, ...,x_n]$ e um vetor de saída $y=[y_1, y_2, ...,y_n]$, o PMC forma um mapeamento complexo $y=\Psi(w,x)$ da entrada da primeira camada para a saída da última camada, parametrizado pelos pesos sinápticos **w**.

Os conjuntos de Redes Neurais Artificiais compõem uma ferramenta muito abrangente e por isso, diferentes modelos foram criados, cada um com um determinado objetivo.

Cada um desses modelos possui uma arquitetura diferente além de uma maneira particular de treinamento. A arquitetura é determinada pela maneira que os neurônios de uma determinada rede estão organizados.

Dentre as arquiteturas mais conhecidas, pode-se citar:

- ***Redes FeedForward (camada única)***: Neste tipo de rede, tem-se uma camada de entrada e uma única camada de neurônio que é a própria camada de saída. São usadas em reconhecimento de padrões e memória associativa.
- ***Redes FeedForward (Multicamadas)***: Esse tipo de rede difere da anterior pela presença de uma ou mais camadas escondidas de neurônios. São comumente usadas em aproximador de funções, reconhecimento de padrões e identificação e controle.
- ***Redes Recorrentes***: São redes que contém retroalimentação entre neurônios de camadas diferentes. As aplicações são previsão/estimação, séries temporais, otimização e sistemas dinâmicos.
- ***Estrutura Lattice (ou reticulada)***: Consiste de um ARRAY de neurônios de uma ou mais dimensões, no qual os sinais de entrada são os mesmos para todos os neurônios. São usadas em grafos e aplicações que dependem de localização espacial dos neurônios visando retirada de características.

Além das arquiteturas, cada rede possui um algoritmo diferente de treinamento. O treinamento consiste em ajustar os pesos sinápticos e os limiares de um neurônio de forma que a aplicação de um conjunto de entradas produza um conjunto de saídas desejadas. Estes treinamentos podem ser classificados em:

- **Supervisionado**: A rede é treinada para fornecer a saída desejada a partir de um estímulo de entrada específica.
- **Não supervisionado**: Não há uma saída específica em relação aos estímulos de entrada. A rede se auto-organiza em relação às particularidades do conjunto de entrada.

Neste trabalho será utilizada uma rede Perceptron Multi-Camada juntamente com o sistema de pré-processamento de sinais de voz.

3 – Solução do problema

Esta seção irá detalhar a estrutura do sistema de RAL, composto pelos seguintes componentes:

- A Base de Dados;
- O Pré-Processamento;
- A Extração de atributos*;*
- *Classificação dos Padrões (RNA);*
- As Regras de decisão.

3.1 – A Base de Dados

A base de dados de voz utilizada neste trabalho será a Speaker Recognition v1.0, desenvolvida pelo *Center of Spoken Language Understanding (CSLU) do Oregon Graduate Institute (OGI)*, e consiste na

gravação de amostras de voz de 90 participantes ao telefone.

Cada participante gravou sua voz em doze sessões ao longo de um período de vários meses. Por se tratar de uma base americana, as frases foram construídas no idioma inglês.

O desenvolvimento do problema irá consistir na seleção das amostras, treinamento e operação da rede neural escolhida, com a respectiva análise dos resultados em termos de desempenho no reconhecimento.

3.2 – O Pré-Processamento

Cada amostra de voz foi normalizada pelo seu valor máximo, com o objetivo de minimizar a possibilidade de interferências no sistema de reconhecimento causadas por grandes diferenças de energia entre as amostras.

A seguir as locuções foram segmentadas através de janelas retangulares de 512 pontos (duração de 64ms) com taxa de sobreposição de 50%.

Após a segmentação, as locuções foram submetidas a um filtro de pré-ênfase cuja função de transferência é dada por:

$$H(z) = 1 - 0.95z^{-1} \quad (5)$$

Seguiu-se à pré-ênfase, uma etapa de janelamento, onde cada segmento passa por uma janela de *Hamming*, muito utilizada na literatura, dada pela seguinte expressão:

$$W_{HAMMING}(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{LJ}\right) \quad (6)$$

onde LJ é a largura da janela utilizada (512 neste caso).

Neste ponto, inicia-se a extração dos atributos de cada segmento e conforme descrito anteriormente, serão utilizados os coeficientes mel-cepstrais, pelos bons resultados observados na literatura.

Para uma primeira análise, a escolha do número de filtros a ser utilizado baseou-se no trabalho de [MAGNI, 98], utilizando 80 filtros com 24 MFCCs.

O coeficiente MFCC(0) foi desprezado, uma vez que a sua informação, a energia do segmento de amostras, influencia de maneira indesejável o sistema de reconhecimento.

Um dos maiores problemas na tarefa de RAL, ocorre devido a flutuações no padrão de fala com relação ao eixo do tempo, bem como variações espectrais.

Desta forma os vetores de padrão da fala, gerados pelas frases pronunciadas pelos locutores devem ser modificados para se adequar à entrada da rede, ao mesmo tempo que devem manter suas características inerentes ao locutor.

Vários algoritmos foram desenvolvidos e estão disponíveis na literatura, a fim de realizar o "alinhamento temporal" do padrão de entrada para a RNA, sendo que a performance da Rede está diretamente ligada ao tipo de alinhamento utilizado.

3.3 – Trace Segmentation

O algoritmo conhecido como *Trace Segmentation* (TS) consiste em um método não-linear para a normalização de sequências temporais de *frames* de representação de voz.

Introduzido por [KUHN, 91], foi bastante utilizado no passado em conjunto com métodos de Programação Dinâmica (*Dynamic Programming – DP*) para reconhecimento isolado de palavras onde, utilizando 2 bancos de dados para treinamento e teste de reconhecimento de palavras, o desempenho gerado pela utilização do algoritmo TS se manteve abaixo da performance do DTW (Dynamic Time Warping) em torno de 10%.

A ideia do TS é baseada na hipótese de que, apesar de diferenças temporais, para sinais de voz de uma mesma categoria, as flutuações no espectro de frequências ao longo do tempo irão ocorrer na mesma sequência, porém em intervalos de tempo de comprimentos diferentes.

Assim, a sequência de *frames* (ou linhas) que representam uma dada locução pode ser interpretada como uma trajetória (*trace*) em um espaço de dimensionalidade igual ao tamanho de cada *frame* (vetor de atributos). Como a trajetória é característica de cada categoria de locuções, a proposta do TS é codificar a trajetória como um vetor de dimensionalidade fixa, para que este possa servir de entrada para uma RNA *Feed-Forward* convencional, que apresenta um número bem definido e fixo de entradas na camada de entrada.

3.4 – Minimal Temporal Information (MTI)

Toda locução de voz, sendo mais longa ou mais curta, apresentará número de segmentos diferente.

Quando se busca o reconhecimento de palavras, costuma-se normalizar o tamanho das locuções.

Entretanto, em problemas do tipo RAL, isto não é necessário como demonstrou [TMOSZCZUK, 98] em seu trabalho, desde que em seu lugar seja utilizada uma outra técnica de caracterização das informações presentes nas amostras de voz.

A técnica proposta por [TMOSZCZUK, 98] baseia-se no agrupamento dos segmentos das locuções em *Minimal Temporal Information* (MTI).

A MTI tem por objetivo pesquisar a possibilidade de caracterização do locutor por meio de estruturas temporais obtidas a partir do vetor de MFCCs utilizados como parâmetros para o reconhecimento do locutor.

As MTIs são conjuntos de segmentos consecutivos extraídos a partir da sequência de segmentos que representam uma locução.

Ao longo de uma locução as MTI são consecutivas e podem apresentar repetição ou não de segmentos utilizados na MTI imediatamente anterior. A figura abaixo apresenta de forma gráfica o método de construção das MTIs.

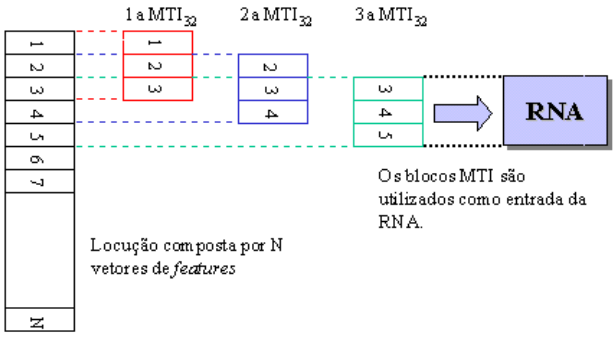

Figura 1 – Construção de MTIs

Será utilizada neste trabalho a MTI_{31}, sem perda de generalidade, uma vez que o objetivo proposto é avaliar o desempenho da RNA na tarefa do RAL e o ônus computacional é menor (devido a um menor nível de interpolação entre os segmentos).

3.5 – Arquitetura do sistema de RAL

Para N locutores do grupo de treino, o sistema compõe-se de uma única Rede PMC, com N saídas, onde cada saída irá representar um locutor.

A Rede será então treinada de forma que, para um sinal de teste, a associação será feita a partir do maior valor de saída (*winner takes all),* de forma imediata.

A utilização das MTIs propostas por [TMOSZCZUCK, 98] como forma de apresentar a locução à entrada da rede elimina a necessidade de se normalizar as locuções. Porém, como consequência deste procedimento, surge a necessidade de se criar um critério de decisão adicional, pois, para cada locução apresentada, será gerado um vetor com M saídas (uma saída para cada MTI).

O 2º critério de decisão utilizado neste trabalho consiste em avaliar o vetor de saídas, e o locutor será dado pela saída apontada pelo maior número de MTIs da locução apresentada.

A fim de tornar o sistema mais flexível e permitir a identificação de impostores, algumas informações devem ser acrescentadas ao 2º critério de decisão: a utilização de um *threshold* τ_1 , de forma que, se a porcentagem correspondente a maioria das MTIs apontadas for maior do que τ_1 $_1$, então diz-se que o sistema **decidiu** favoravelmente ao locutor, caso contrário o status será **indefinido**

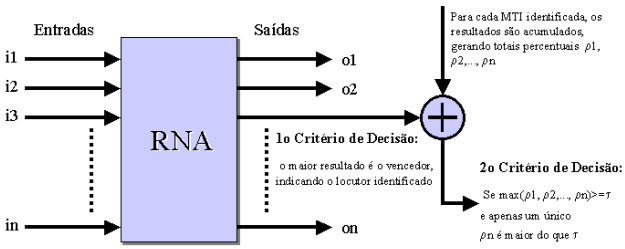

Figura 2 – Arquitetura do Sistema de RAL

4 – Resultados obtidos

O sistema foi construído de forma a analisar 10 locutores escolhidos da base de dados descrita na seção anterior, e o conjunto de "locuções" de análise foi composta da seguinte forma:

- 10 locuções (por locutor) para "treinamento" do sistema de reconhecimento;
- 6 locuções (por locutor) para a verificação de desempenho do sistema.

Para o Pré-Processamento foram aplicadas as técnicas descritas anteriormente (Janelamento, MFCCs com 80 filtros e 24 coeficientes, Trace Segmentation sem Interpolação). Para o alinhamento temporal, foi aplicada a MTI_{31} nos segmentos.

Em uma primeira análise, a RNA foi treinada para com diferentes valores para os neurônios da camada oculta, a fim de determinar um valor mais adequado para a camada oculta utilizada no sistema. Os valores escolhidos foram: NH={35, 45, 55, 65, 75}.

A taxa de aprendizado η foi mantida constante

em 0.015 para todos os treinamentos.

Foram levantados gráficos da taxa de acerto de cada topologia acima em função da variação do threshold τ_1 na faixa de 0,1 a 1,0.

A variável τ_1 indica qual a porcentagem de segmentos de uma dada locução foi associada a um dado locutor pelo sistema. Assim é possível sintonizar o sistema em função desta variável.

Cabe ressaltar porém que valores baixos de τ_1 podem conduzir a decisões erradas por parte do sistema, uma vez que apenas uma pequena parte do sinal de voz acaba sendo associado a um locutor, conferindo baixa confiabilidade à rede.

Por outro lado, valores excessivamente altos (próximos a 1) não necessariamente conduzem a uma taxa de acerto maior, porque significa que a RNA deve classificar quase a totalidade da locução a uma única pessoa, o que é muito difícil quando se trata de sinais de voz.

Forçar o sistema a responder com um valor muito alto de τ_1 pode resultar em overfitting, tornando o sistema não-confiável (assim como ocorre para τ_1 baixo).

Valores adequados para τ_1 parecem estar situados entre 0,7 e 0,85. Nesta faixa, uma parcela suficientemente significativa da frase (mais de 70%) foi classificada pelo sistema como sendo de um único locutor, dando um grau de confiança bastante aceitável para a identificação correta do locutor em questão.

Analisando-se então os gráficos dos treinamentos, e levando-se em conta os valores de τ_1 acima, foi possível observar que o treinamento para NH=55 obteve uma taxa de acerto cerca de 5% acima das demais (cujos gráficos encontram-se no anexo deste documento):

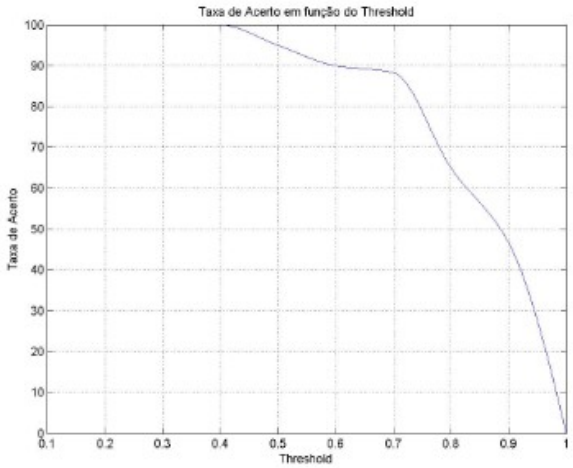

Figura 3 – Taxa de acerto para NH=55

Na faixa próxima a $\tau_1 = 0,7$, a taxa de acerto do sistema ficou em torno de de 86%, um resultado bastante robusto para o sistema de RAL.

Levando-se em conta o valor de NH=55 como sendo o ideal para a topologia apresentada, foram feitos treinamentos adicionais para pré-processamento de sinal utilizando-se Trace Segmentation, Truncamento simples e Linear Time Warping.

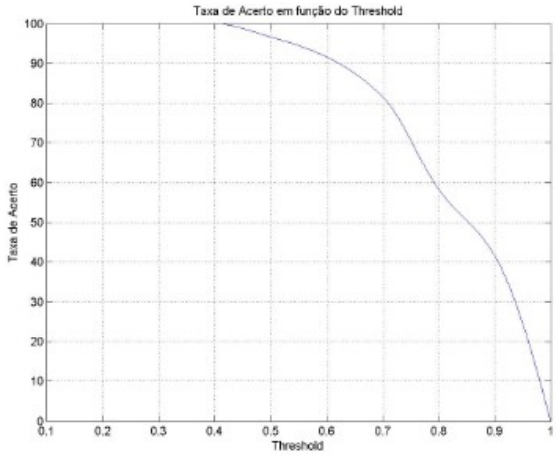

Figura 4 – Taxa de acerto para truncamento

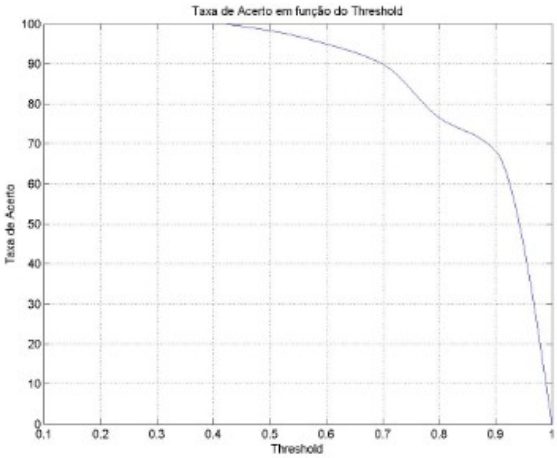

Figura 5 – taxa de acerto para Trace Segmentation

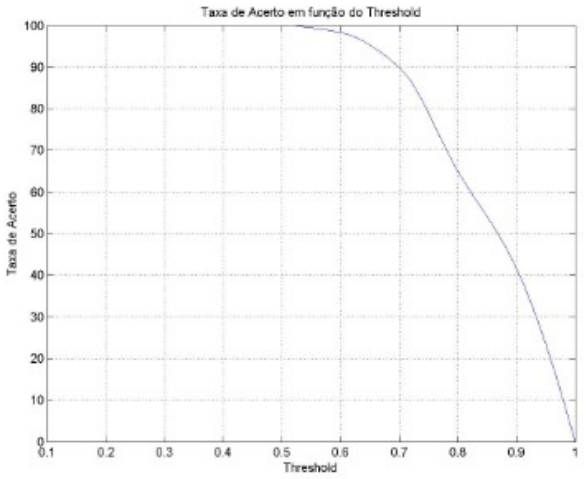

Figura 6 – Taxa de acerto para Linear Time Warping

A partir dos gráficos gerados, pode-se observar nitidamente que as técnicas de TS e LTW melhoram sensivelmente a eficiência do sistema, cuja taxa de acerto para $\tau_1 = 0,7$ fica acima de 90% com uma resposta mais plana para uma faixa maior de τ_1, principalmente para a técnica de Trace Segmentation.

Em comparação com o Treinamento feito apenas usando Truncamento de Segmento (janela retangular), o desempenho do sistema cai sensivelmente em comparação com os anteriores (em torno de 80% para $\tau_1 = 0,7$), sugerindo mais uma vez que as técnicas de alinhamento temporal são relevantes na seleção de parâmetros e no desempenho da Rede classificadora.

5 – Conclusões

Este trabalho teve por objetivo mostrar a viabilidade do uso de Redes Neurais Artificiais no processo de classificação dos padrões associados ao problema do Reconhecimento Automático do Locutor (RAL).

A partir dos testes realizados, fica clara a influência do pré-processamento na taxa de desempenho da RNA do sistema. As técnicas de pré-processamento consagradas na literatura realmente mostram resultados superiores se comparados a técnicas mais simples como a utilização de Janelas Retangulares.

O sistema apresentado mostrou melhor desempenho para a Técnica de Trace Segmentation, cuja taxa de acerto ficou superior à técnica de Linear Time Warping, bastante referenciada na literatura.

A utilização do *threshold* de decisão τ_1 facilitou as análises de resultados, uma vez que permitiu a visualização do comportamento de cada RNA ao longo de uma faixa de resultados, além de permitir um ajuste adicional para a taxa de acerto dos sistema, sem que haja a necessidade de se recorrer a um novo treinamento.

Os testes confirmaram o paradigma das MTI como possíveis unidades que carregam informações importantes sobre o locutor, facilitando o treinamento através da redução da dimensionalidade da camada de entrada das RNA. Os resultados obtidos estão de acordo com os estudos de [TMOSZCZUCK, 98], com os melhores desempenhos ocorrendo para a família MTI_{3X}, com especial destaque para MTI_{32};

Como trabalhos futuros, outros paradigmas de Redes Neurais podem ser analisados, como por exemplo, as Redes Recorrentes, além de outras arquiteturas como por exemplo, um CNN (Concurrent Neural Netbook), proposta por [CATARON, 01], onde são utilizadas várias Redes Interconectadas, cada uma treinada especificamente para identificar um único locutor, rejeitando os demais (portanto, com uma só saída).

Referências

[ATAL, 76] - ATAL, B.S. Automatic recognition of speakers from their voices. Proceedings of the IEEE, v.64, n.4, p.460-75, Apr. 1976.

[CASAGRANDE, 97] - CASAGRANDE, R. Redes neurais artificiais com retardos temporais aplicadas ao reconhecimento automático do locutor. São Paulo, 1997, 80p. Dissertação de Mestrado, EPUSP.

[CATARON, 01] - CATARON A.; NEAGOE, V.E. **Concurrent Neural Networks for Speaker Recognition**. In: IEEE International Conference on Telecommunications*,* Romania, Bucharest, 2001.

[CORSI, 81] - CORSI. P. Speaker recognition: a survey. In: Proceedings of the NATO Advanced Study Institute, Bonas, 1981. Automatic Speech Analisys and Recognition. Dordrecht, D. Reidel Publishing company, 1982. pp.277-308.

[ELMAN, 90] - ELMAN, J. Finding Structure in Time. Cognitive Science 14, pp. 179-211, 1990.

[FLANAGAN, 72] – FLANAGAN, J. L. **Speech Analysis, Synthesis and Perception.** 2nd Edition, Springer-Verlag*,* NY, 1972.

[FURUI, 89] - S. Furui. Digital Speech Processing, Synthesis and Recognition. Mareei Decker Inc., NY. 1989.

[GISH, 94] - GISH, H.; SCHMIDT. M. Text-independent speaker identification. IEEE Signal Processing Magazine, v.l 1, n.4, p. 18-32, Oct. 1994.

[HATON, 81] - HATON, J.P. Automatic Speech Analysis and Recognition. Proceedings of the NATO, France, 1981.

[JACOBS, 88] - JACOBS, RA. Increased rates of convergence through learning rate adaptation. Neural Networks, vol.l, pp.295-307, 1988.

[JORDAN.86] - JORDAN, M Serial Order: A Parallel Distributed Approach. Institute for Cognitive Science Report 8604, University of Califórnia. San Diego, 1986.

[MAGNI, 98] - MAGNI, A.B. Reconhecimento automático do locutor com coeficientes Mel-Cepstrais e redes neurais artificiais. São Paulo, I38p., Dissertação de Mestrado, EPUSP, 1998.

[PICONE, 93] - PICONE, J.W. Signal modeling techniques in specch recognition. Proceedings of the IEEE, v.8l, n.9, p.1215-47, Sept. 1993.

[RABINER, 78] – RABINER, L.R.; SCHAFER, R.W. **Digital processing of speech signals**. Englewood Cliffs, Prentice Hall, 1978.

[SAMBUR. 75] - SAMBUR, M.R. Selection of acoustic features for speaker identification. IEEE Transactíons on Acoustics, Speech and Signal Processing, v.23, n.2, p. 176-82, Apr. 1975.

[STEVENS, 40] – STENVENS, S.S.; VOLKMAN, J. **The relation of pitch to frequency.** American Journal of Psychology, vol.53, p.329, 1940. citado em DELL93.

[TMOSZCZUK. 98] - TMOSZCZUK, A.P. Reconhecimento automático do locutor com redes neurais artificiais do tipo radial basis function (RBF) e Minimal Temporal Information (MTI). São Paulo, Dissertação de Mestrado, 135p. EPUSP, 1998.

[WOLF, 72] - WOLF, J.J. **Efficient acoustic parameters for speaker recognition**. The Journal of the Acoustical Society of America, v.51, n.6, p.2044-56, 1972.

ANEXOS

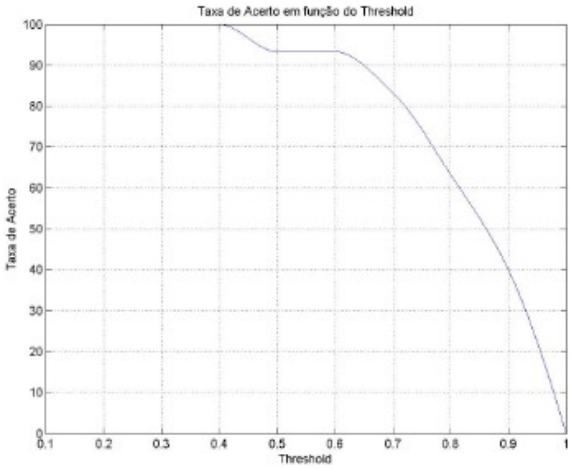

Figura 7 – Taxa de acerto para NH=35

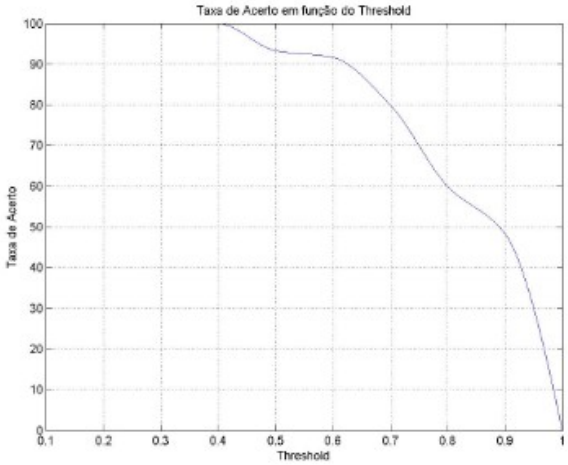

Figura 7 – Taxa de acerto para NH=75

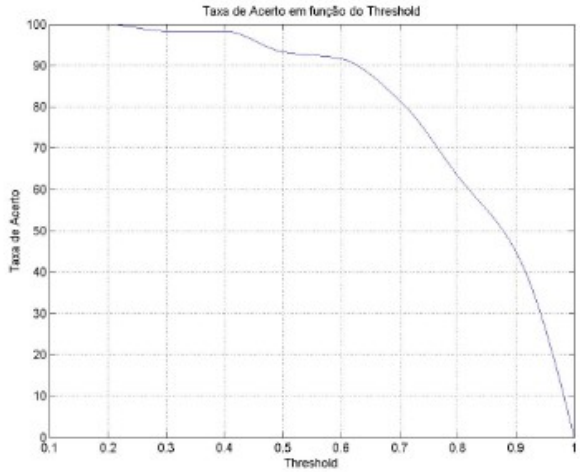

Figura 7 – Taxa de acerto para NH=45

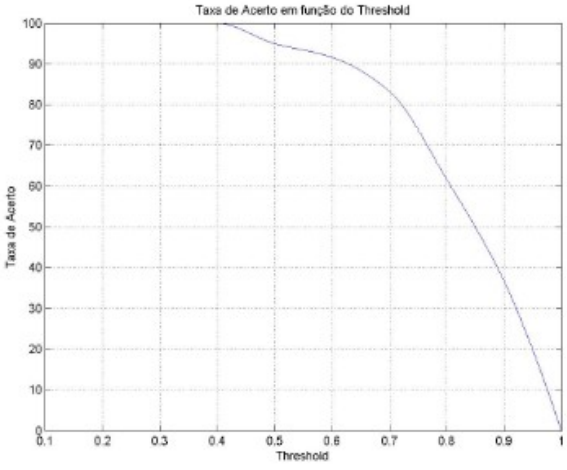

Figura 6 – Taxa de acerto para NH=65