Reconhecimento do Orador com Redes Neuronais

António Teixeira, Francisco Vaz

Resumo **- Neste trabalho apresenta-se uma experiência no uso de uma rede neuronal dinâmica, a Rede Gama Concentrada, na área do Reconhecimento do Orador, nas suas duas variantes de Verificação e Identificação.**

A abordagem utilizada é baseada no modelo clássico em Reconhecimento de Voz. Primeiro são extraidos atributos relevantes usando um banco de filtros. As propriedades são comparadas com referências previamente construidas usando a rede neuronal. A decisão é feita usando as saídas da rede neuronal. Foi utilizada uma rede dinâmica devido á natureza variável no tempo dos padrões.

Foi usada uma rede por cada orador cliente e vários oradores por rede. A primeira abordagem deu os melhores resultados, e por ser modular, torna possível a construção de sistemas com muitos clientes.

Os resultados obtidos são promissores e mostram que o método é robusto.

Abstract **- We present an experience in using a dynamic neural network , the focused gamma net, in the area of speaker recognition, in the two variants of verification and identification. The approach used is based on the classical model for speech recognition. First we extract relevant features with a filter bank. Features are compared with previously constructed references using a neural network. Decision is done using neural network outputs. A dynamic neural network, the gamma net, was used because the patterns are time varying.**

We tried the use of a network per speaker and several speakers per network. The first approach gave better results and , being modular, make possible the construction of large systems.

Results obtained are promising and show that the method is robust.

I. INTRODUÇÃO

A. Assunto

As previsões apontam para um uso maciço de processamento de voz nos sistemas de informação num futuro próximo.

Sendo a voz o meio de comunicação por excelência do ser humano todos estes novos serviços dependem da existência de capacidades de reconhecimento de palavras, reconhecimento do orador, tradução automática e identificação da linguagem utilizada.

O reconhecimento do orador será uma peça chave nestes sistemas devido à necessidade de segurança.

Como sabemos vozes de pessoas diferentes soam de uma forma diferente. Esta importante propriedade da voz, de ser dependente da pessoa, é o que torna possível, por exemplo, reconhecermos um amigo ao telefone. A faculdade de reconhecer uma pessoa apenas pela sua voz é conhecida como **Reconhecimento do Orador**.

O reconhecimento por ouvintes humanos é uma experiência comum conhecida desde sempre. Recentemente, com a disponibilidade de computadores digitais, cientistas interessados na área da voz resolveram explorar a possibilidade de reconhecer automaticamente pessoas utilizando apenas a sua voz o, que se designa por Reconhecimento Automático do Orador.

Apesar de em muitas áreas ser muito difícil igualar o desempenho humano, nesta área (Reconhecimento do Orador) os dados experimentais sugerem que o desempenho das máquinas em muitos casos execede ados seres humanos .

O objectivo final de todos os estudos em Reconhecimento do Orador é chegar a um sistema automático, independente do tempo, que replique a capacidade humana de rapidamente, com exactidão e independentemente do texto pronunciado efectuar o reconhecimento de uma pessoa apenas pela sua voz [1].

B. Aplicações

O Reconhecimento do Orador tem aplicação em **sistemas de segurança** como controlo de acesso (sistemas de fechadura actuados por voz para portas de casa e de automóveis); controlo de acesso a dados em computadores (*password* por voz); controlo de transacções através das linhas telefónicas (reserva de passagens aéreas ou movimentos bancários pelo telefone) [2]. Muitas instituições financeiras, assim como companhias fornecendo acesso limitado a bases de dados em computadores gostariam de poder oferecer serviços automáticos pelo telefone. Como números de código podem ser perdidos, roubados, ou esquecidos o reconhecimento através da voz (se suficientemente de confiança) pode ser uma alternativa [3]. Existe outro tipo de aplicações com muito interesse, num domínio diferente, como sejam a **investigação criminal** (comparação da voz desuspeitos de um crime) e **vigilância de canais de comunicação** [2]. No domínio da tecnologia **militar,** especialmente em sistemas de alta segurança, existe a possibiliade de aplicação destes sistemas. Seria útil para por exemplo aceitar ou rejeitar relatórios da linha da frente [4]. Uma aplicação completamente diferente é o uso destas técnicas para **adaptar os sistemas de Reconhecimento de Voz ás características do orador** para melhorar o desempenho destes sistemas [4].

C. Resumo Histórico da Investigação na Área

A investigação na área de Reconhecimento do Orador iniciou-se na década de 1960 com o estudo do reconhecimento por ouvintes humanos. Nesta altura também era comum o uso de espectogramas para reconhecimento visual [2].

Rosemberg e Atal [5][6] fazem um ponto da situação do desenvolvimento até 1976. Era habitual usarem-se bancos de filtros para a extracção do *short-time spectrum* sendo usados o tom intensidade formantes e ceficientes de

predição linear (LPC) em alguns poucos sistemas. As técnicas de decisão eram muito rudimentares usando-se a distância Euclideana em muitos casos. Nesta altura já se utilizavam algoritmos de alinhamento dinâmicos para tratar as variações temporais da voz. Devido à sua maior simplicidade os sistemas eram dependentes do texto.

Tendo passado quase uma década [2][3] aumentou o interesse por sistemas independentes do texto o que levou ao uso de estatísticas recolhidas ao longo do tempo para representar o sinal de voz Os parâmetros utilizados evoluiram sendo usadas por exemplo análise cepstral e parâmetros LPC ortogonais. No capítulo da classificação é já vulgar a utilização do *Dynamic Time Warping* (DTW) e aparece a Quantificação Vectorial .

Em [7] são analisados alguns dos últimos progressos na área. No capítulo dos parâmetros é usual o uso de parâmetros LPC-cepstrais. Aparecem tambem os *Line Spectrum Pair* (LSP) e Δ–cepstrais. As técnicas actuais apresentadas incluem os Modelos de Markov não Observáveis (HMM) e várias variantes de *source coding* como Quantificação Vectorial, Quantificação Matricial, *Filler Template* e *Acoustic Segment Quantization*. Tem se tentado também o uso de redes neuronais nesta área.

Como é notório tem havido uma evolução quer no tipo de parâmetros usados quer nas técnicas usadas para efectuar a classificação.

D. Uso de Redes Neuronais

Algumas experiências foram relatadas na literatura sobre a utilização de modelos conexionistas para reconhecimento do orador. As primeiras referências são [8] usando Perceptrões multicamada com uma rede por orador, e [9][10] onde um orador é reconhecido por um *codebook* usando *Learning Vector Quantization* (LVQ). Ambas as abordagens são dependentes do texto.

Trabalhos mais recentes em sistemas independentes do texto usaram *Time Delay Neural Networks* (TDNN) [11] e *Predictive Neural Networks* [12]. A última usa parâmetros estáticos e dinâmicos, o que é necessário para atingir desempenhos superiores.

No nosso trabalho usamos uma forma diferente de utilizar a informação espectral dinâmica da voz, a rede gama [13].

II. CONCEITOS

A. Modelo

A abordagem habitual é extrair alguns atributos acústicos da voz e compará-los com um conjunto de referência previamente armazenado. Se existir uma semelhança grande, segundo um determinado critério, entre as propriedades de teste e da referência o orador é reconhecido [2].

De uma forma geral todos os sistemas de Reconhecimento do Orador são constituidos pelos seguintes blocos: Processamento acústico (extracção de parâmetros); Comparação de padrões; Referências e processo de Decisão (Figura 1).

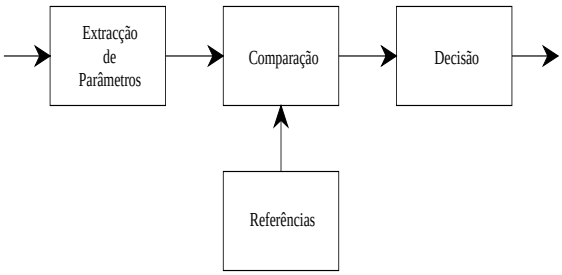

Fig 1 - Modelo do Sistema

Uma pronunciação de um orador desconhecido constitui a entrada do sistema. Esta entrada é analisada para dela se extrair as propriedades que caracterizam o orador. Na fase seguinte estas propriedades são comparadas com propriedades protótipo armazenadas na referência. Usando os resultados da comparação é finalmente efectuada uma decisão [7].

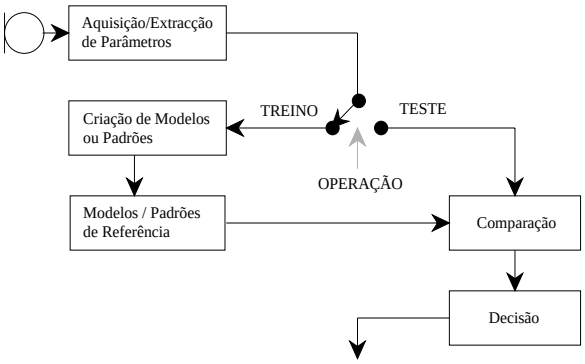

Fig 2 - Diagrama Operacional de um Sistema de Reconhecimento

B. Fases

Todas as tarefas de reconhecimento de padrões incluindo o Reconhecimento do Orador utilizam duas fases: treino e reconhecimento (Figura 2). Efectuado *off-line* e habitualmente combinando métodos manuais e automáticos a fase de treino establece as referências. A fase de reconhecimento é automática e usualmente em tempo real e tenta atribuir a uma entrada desconhecida uma identificação [3].

C. Verificação e Identificação

O Reconhecimento do Orador pode ser dividido em duas grandes classes: Identificação do Orador e Verificação do Orador. O problema mais geral Identificação do Orador pode ser definido da seguinte forma. Numa população de N oradores descobrir aquele a que corresponde o padrão de referência com mais semelhanças ao padrão do orador desconhecido, que constitui a entrada do sistema. A probabiliade total de uma decisão incorrecta é uma função monótona crescente do número de oradores.

O problema de Verificação do Orador pode definir-se da seguinte forma: Dado um padrão de um orador desconhecido conjuntamente com a identidade reclamada, que constituem a entrada do sistema, determinar se o padrão é suficientemente semelhante ao padrão de

referência associado à identidade reclamada para se aceitar a pretensão. Neste caso apenas é efectuada uma comparação independentemente do tamanho da população. Portanto a decisão é geralmente independente do tamanho da população. Uma premissa básica para muitos sistemas de verificação é a de que os utilizadores habituais são cooperantes, isto é, não tentam mudar o seu comportamento conscientemente de tentativa para tentativa.

D. Texto Utilizado

Uma dicotomia para sistemas de Reconhecimento do Orador é se existe um texto obrigatório para os oradores pronunciarem ou não. Em muitas aplicações o texto é fixo chamando-se estes sistemas **dependentes do texto**. Nos sistemas **independentes do texto** existem habitualmente restrições, tais como o tamanho, [5] e outros constrangimentos linguísticos, sendo muito raro o uso de texto completamente livre [14]. Os sistemas dependentes do texto precisam de um elevado grau de colaboração dos utilizadores tendo no entanto a vantagem de facilitarem bastante a comparação. Nos sistemas independentes do texto utilizam-se habitualmente estatísticas recolhidas ao longo de períodos alargados, como por exemplo a média e a variância. Esta abordagem tem o inconveniente de retirar a dimensão tempo. Mais recentemente nestes sistemas começaram a surgir técnicas baseadas em eventos de curta duração [7]. No que respeita às aplicações enquanto os sistemas com texto fixo são usados em sistemas de controlo de acesso (devido a haver cooperação por parte dos utilizadores) tem de usar-se texto livre em sistemas de vigilância (onde não há cooperação consciente).

E. Termos Habituais na Área

Apresentam-se de seguida alguns termos utilizados na área de Reconhecimento do Orador e que se irão aplicar ao longo deste texto.

Clientes - São os oradores conhecidos do sistema que reclamam a sua identidade verdadeira.

Impostores - São não clientes. Os impostores eventuais não imitam a voz de nenhum dos clientes apenas fazendo uso da sua voz natural

Imitadores - Classe de impostores que tentam entrar no sistema disfarçando a sua voz por forma a parecer-se de alguma forma com clientes do sistema.

Goats - Oradores que causam mais problemas ao sistema de reconhecimento.

Sheeps - Oradores que são facilmente reconhecidos.

F. Medição do Desempenho

As identidades usadas quer por clientes quer por impostores para tentar aceder ao sistema são geralmente apenas as pertencentes ao conjunto de clientes do sistema. É fácil rejeitar-se outras identidades usando por exemplo a consulta da tabela de clientes. No processo de verificação podem surgir 3 casos. O caso do uso da identidade de um não cliente como vimos é facilmente evitado distintos no tocante ao utilizador. Orador cliente do sistema que pretende verificar a sua identidade correcta Orador cliente que pretende fazer-se passar por outro Orador não cliente a tentar passar por cliente. Se defenirmos impostor como a pessoa que pretende assumir outra identidade os casos e são apenas variantes.

Podemos ter na entrada do sistema oradores Clientes ou Impostores e na saída pode obter-se cliente aceite ou não aceite.

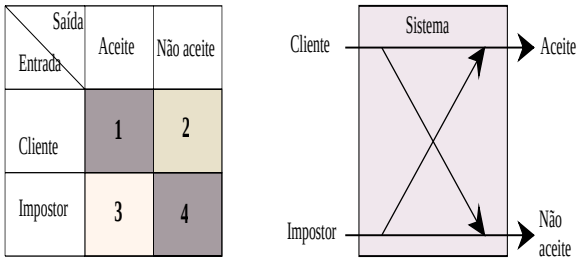

Fig 4 - Verificação - Casos

As situações 1 e 4 correspondem ao Acerto do sistema. A situação 2 corresponde a Falsos Negativos e a situação 3 a Falsos Positivos, tudo situações de erro. Os Falsos Positivos são designados geralmente na literatura por Falsas Aceitações (FA) e os Falsos Negativos por Falsas Rejeições (FR). A entrada Cliente pode ser decomposta em cada um dos clientes individuais para uma análise do comportamento de cada um deles.

Como para aceder ao sistema um impostor tem de usar a identidade de um cliente, nos testes, é necessário simular a escolha do cliente. Em geral não é fácil usar dados reais, ou seja, impostores reais a escolherem um cliente para ludibriar o sistema. O uso de distribuições estatísticas é difícil e duvidoso. Pode evitar-se a questão de saber qual a identidade assumida pelo impostor, usando o pior caso, isto é, a identidade correspondente à saída maior e compará-la com o limiar. Temos assim o erro no caso de o impostor fazer-se passar por esse cliente. No caso de apenas existir um único cliente, não existe qualquer dúvida sobre a identidade que os impostores assumem, o que torna os testes fáceis de realizar.

No caso dos clientes aí sim já é relevante a identidade. Pode determinar-se quais os oradores mais fáceis e mais difíceis para o sistema (*sheeps* e *goats*).

Para indicar o desempenho de um sistema de verificação é comum apresentar-se a sua taxa de erro. O erro do sistema é representado habitualmente pelas taxas FA e FR ou por combinações destas. Além da média aritmética de FA e FR é muito usada a média geométrica. Na caracterização do desempenho neste trabalho usaremos esta última taxa.

Para a identificação, na entrada do bloco de decisão temos clientes e não clientes que são identificados como clientes e impostores, correcta e incorrectamente. As várias hipóteses encontram-se representadas na Figura 5.

Saída / Entrada	Impostor	Cliente 1	Cliente 2	...	Cliente N
Impostor	1	2	2		2
Cliente 1	3	1	4		4
Cliente 2	3	4	1		4
...					
Cliente N	3	4	4		1

Fig 5- Identificação - Casos Possiveis

A identificação correcta corresponde aos casos assinalados com 1. Os casos 2 e 3 são os Falsos Positivos e Falsos Negativos respectivamente. O caso 4 corresponde também a um erro, mas agora devido à troca de um cliente por outro. Em aplicações em que não é feita qualquer distinção em termos de privilégios entre os clientes este último caso não é considerado erro.

A taxa de acerto na identificação pode ser obtida dividindo os padrões classificados correctamente pelo total de padrões de teste. De forma análoga se obtêm as taxas de trocas, falsos negativos e falsos positivos. É no entanto usual utilizar percentagens separadas para cada orador, usando como medida global de desempenho a média aritmética destas.

III. SISTEMA IMPLEMENTADO

O sistema desenvolvido tem a arquitectura clássica na área, já apresentada. Devido à natureza da rede neuronal os blocos de comparação e referências são ambos implementados por esta. Os parâmetros adptáveis da rede constituem as referências, que são adaptadas na fase de treino. Na fase de teste a rede comporta-se como um comparador.

A. Extracção de Parâmetros

Num determinado sistema de processamento da voz apenas parte da informação é usada dependendo da finalidade do sistema. Tradicionalmente a informação é extrraída quer do sinal quer da sua transformada de Fourier. Mas devido à natureza única da voz e a sua ligação ao sistema auditivo podemos explorar formas alternativas que utilizem conhecimentos dos mecanismos auditivos para extracção de informação relevante para a nossa aplicação. O sucesso da utilização desta abordagem depende da quantidade de conhecimento que temos sobre o sistema auditivo, conhecimento este que é conseguido pela combinação de dados recolhidos por estudos psicológicos e fisiológicos.

As propriedades são extraídas usando um banco de filtros baseado num modelo que incorpora conhecimentos acerca do ouvido humano.

O banco cobre a banda de frequências entre os 200 e os 3300 Hz sendo composto por 18 Filtros passa-banda *GammaTone* [15] calculados segundo o trabalho de Moore e Glasberg [16].

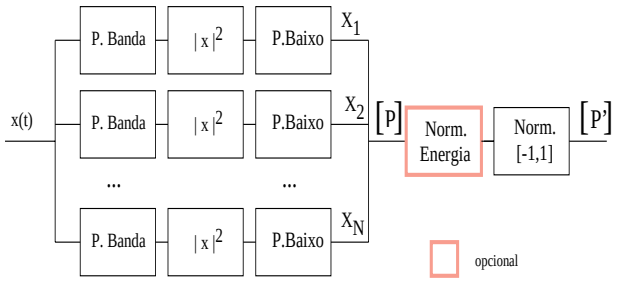

Fig 6- Extracção de Parâmetros - Banco de Filtros

Os sinais na saída do banco de filtros foram amostrados a 30 Hz.

A saída do Banco de Filtros para a frase escolhida consiste numa matriz, em que uma dimensão é a ordem do filtro e a outra pelo tempo. Esta matriz pode ser representada a 2 dimensões usando uma escala de cinzentos. Obtem-se assim uma aproximação ao Espectrograma (Figura 7).

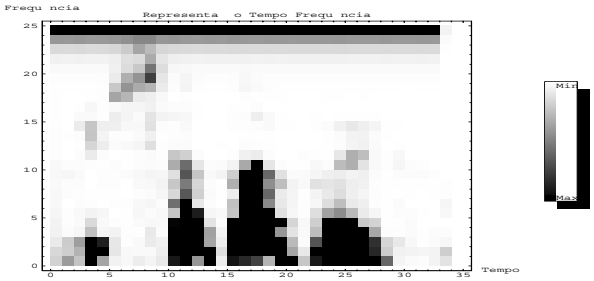

Fig 7 - Saída do Banco de Filtros para a frase "Universidade de Aveiro"

B. Comparação e Referências

Derivado do facto dos padrões extraidos, pelo banco de filtros, serem variante no tempo a rede neuronal utilizada, a Rede Gama Concentrada, possui uma camada de entrada dinâmica. Esta rede neuronal é uma simplificação do modelo de convolução que utiliza atrasos para introdução de dinâmica no modelo [13].

A rede gama é também uma simplificação do modelo *AutoRegressive Moving Average* (ARMA) [13].

Das simplificações obteve-se um modelo neuronal aditivo com estabilidade trivial. Por ser aditivo poderia usar-se o *back-propagation-thru-time* (BPTT) ou o *real-time-recurrent-learning* (RTRL). Mas, o primeiro tem o problema do desdobramento no tempo, criando redes muito grandes no caso de sequências no tempo. O segundo apenas se pode aplicar a redes pequenas, devido às necessidades de processamento.

O uso de arquitecturas *prewired* [17], onde apenas se usa estruturas dinâmicas na camada de entrada, torna possível o uso de uma regra de aprendizagem simples, derivada utilizando a técnica de retropropagação do erro (*back propagation*) [18].

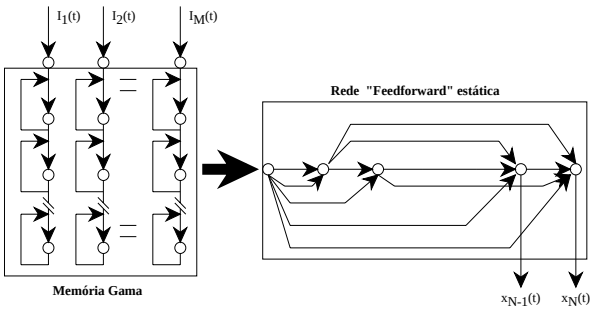

Fig 8 - A Rede Gama Concentrada

Para o treino precisamos de definir o sinal desejado. Como temos uma sequência no tempo, a tarefa é mais complicada que no caso estático habitual.

Habitualmente é fácil escolher o sinal desejado (d(t)) para o instante final. Na aprendizagem baseada em aproximação a escolha natural é ,

$$d_i(t_f) = \begin{cases} 1, i = k \\ 0, i \neq k \end{cases}$$

onde t_f é a duração da sequência e k a unidade correspondente à classe apresentada na entrada. No caso de treino discriminativo queremos maximizar a distância entre as saídas correctas e as incorrectas, isto é algo como,

$$\varepsilon(-y_k(t_f)) + \sum_{i \neq k} y_i(t_f)$$

onde $\varepsilon(.)$ é uma função de saturação monótona crescente. É difícil dizer qual destes métodos dá melhores resultados na prática. Optamos pelo uso do primeiro pela sua simplicidade.

A escolha das trajectórias do sinal desejado para para as unidades de saída é muito menos clara. Foi proposto usar *don't cares*, erro nulo qualquer que seja saída, assim como funções crescentes, como rampas e exponenciais par $t<t_f$, algumas vezes combinadas com funções decrescentes, para os outros nodos de saída [19] [20]. O sinal desejado correcto é função das probabilidades condicionais *a posteriori*, assim como da arquitectura da rede. Geralmente não será aproximado correctamente usando funções escolhidas arbitráriamente como rampas e exponenciais. Por outro lado o uso de *don't cares* conduz a um treino lento, pois só é possível fazer a realimentação do erro no instante final da sequência. Foi proposto o uso de *constraints* nas trajectórias num esforço para tornar mais rápida a aprendizagem e possivelmente aumentar o desempenho [21]. No entanto esta abordagem ainda não foi comprovada em aplicações práticas. Embora pagando bastante em termos de velocidade de aprendizagem usamos *don't cares* nos testes especificar contruir e simular redes neuronais requer um esforço considerável. O uso de ferramentas pode reduzir significativamente o esforço necessário. E fornecendo uma estrutura conceptual torna possível a simulação de redes complexas.

A implementação foi efectuada usando o *Stuttgart Neural Networks Simulator* (SNNS) [22]. O simulador de Redes Gama e o treino foram introduzidos no simulador, extendendo a aplicação deste simulador a novas áreas [23].

C. Decisão

Os valores de saída da rede neuronal são usados pelo bloco de decisão para aceitar ou rejeitar um pedido de verificação, ou para determinar a identidade do orador, no caso de indentificação.

O processo utilizado consiste em determinar a unidade de saída com o maior valor e usar um limiar, no caso de identificação (Figura 9).

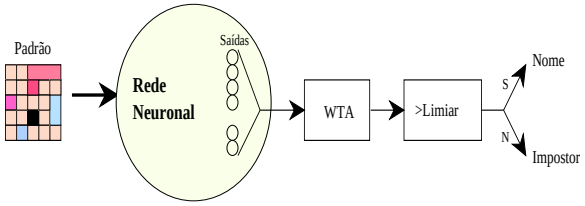

Fig 9 - Decisão -Identificação

No caso da Verificação o procedimento normal é apenas utilizar a informação da saída que corresponde à identidade pretendida. Se esta saída fôr superior ao limiar o pedido é aceite, caso contrário é rejeitado.

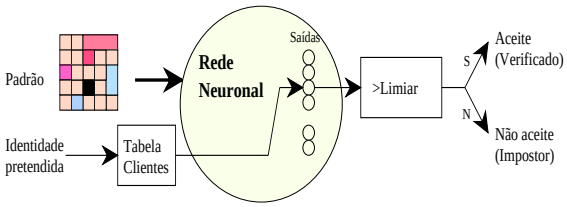

Fig 10 - Decisão -Verificação

Quando o bloco de decisão está inserido num sistema de reconhecimento apenas lhe é exigido que decida. No entanto em sistemas experimentais é necessário obter informações sobre o desempenho do sistema. Englobamos por isso na implementação deste bloco a elaboração das tabelas de confusão e das estatísticas do desempenho.

A decisão no nosso sistema consiste num programa desenvolvido em C que utiliza como entrada o ficheiro de resultados da rede neuronal e fornece á saída as estatísticas (acertos, falsos negativos, falsos positivos e trocas) e ainda as tabelas de confusão. O único parâmetro do programa é o valor do limiar.

IV. DADOS

Para a recolha dos dados para este trabalho foi desenvolvida uma aplicação, para facilitar a gestão dos dados e o trabalho do operador responsável pela recolha ,tornando a tarefa o menos entediante possível ao orador. O desenvolvimento da aplicação foi feito num computador NeXT.

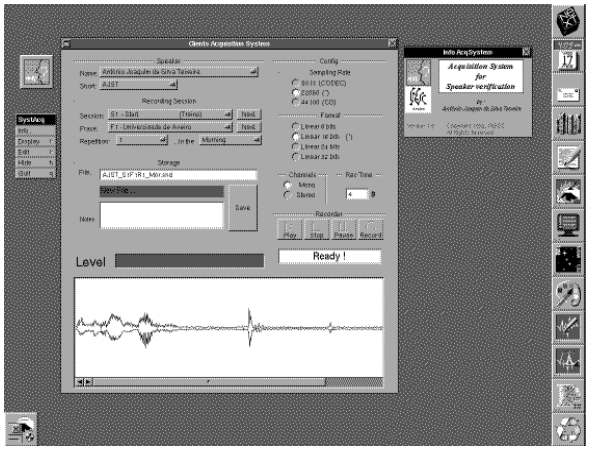

Fig 11 - Ecran da aplicação de recolha

A frase usada ("Universidade de Aveiro") foi escolhida tendo como objectivo a representação a variedade fonética da lingua Portuguesa [24], começando e acabando em vogais para facilitar a segmentação [25].

Os dados foram recolhidos num Laboratório com ruído moderado (e por vezes intenso !), e a frase foi segmentada de uma forma primitiva, utilizando um editor do sinal.

Recolhemos dados de 15 indíviduos do sexo masculino (5 clientes e 10 impostores) em várias sessões de 10 pronunciações cada. Os impostores gravaram 2 sessões, 2 dos clientes 10 sessões, e os 3 outros 7 sessões. A duração média das frases foi de aproximadamente 1 segundo.

V. RESULTADOS

A. Definição dos Parâmetros do Sistema

Efectuamos diversos testes preliminares para definição da topologia da rede a utilizar.

Apenas foi utilizada uma camada escondida com um número variável de unidades. Os melhores resultados foram obtidos com 14 e 16 unidades. Para a camada de saída foi usado um número de unidades igual ao número de clientes, existindo uma correspondência directa entre as unidades e os clientes, sem qualquer codificação. A activação de uma unidade, dependente do processo de decisão, indica um cliente, e não activação de todas as unidades representa um impostor.

Os testes mostraram que é necessário parar o treino com um erro razoavelmente alto, para evitar o excesso de treino e melhorar a generalização. Escolhemos parar o treino para um taxa de falsos positivos igual à taxa de falsos negativos, e usamos esse valor do erro nos testes subsequentes.

É importante que a rede generalize bem o que é um cliente, mas é também essencial que a rede generalize o que é um impostor. Os testes mostraram que a rede necessita de um número superior de padrões de treino de impostores em relação ao número de padrões de treino de clientes.

B. Verificação

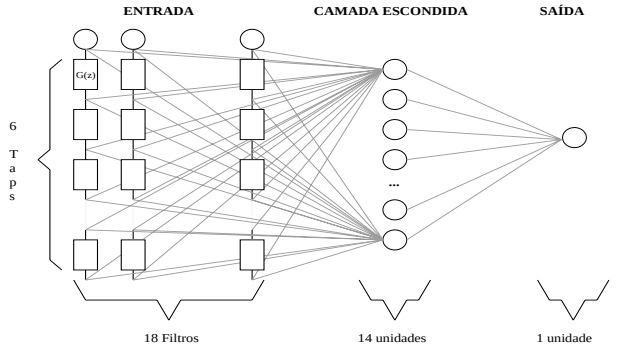

Fig 12 - Rede com apenas 1 saída

Para a verificação foi testada a utilização de uma rede para cada um dos clientes. A rede possuía apenas uma unidade de saída (Figura 12). Através de treino, a rede é especializada no reconhecimento de um orador em particular. Para adicionar outro cliente ao sistema não existe a necessidade de retreinar as redes já existentes, o sistema é modular. Este tipo de abordagem conduz, também, a redes menores que podem ser treinadas mais rápidamente e que generalizam melhor.

Os melhores resultados obtidos foram 8 % de falsa rejeição de clientes e 6.15 % de falsas aceitações de impostores. O pior resultado (usando a voz do primeiro autor) foi de 26 % para a falsa rejeição e de 22.31 % para a falsa aceitação.

Alguns dos resultados são apresentados na Tabela I (em %).

TABELA I

ALGUNS RESULTADOS DA VERIFICAÇÃO

Orador	FR	FA	$(FAxFR)^{1/2}$
PJG	8	6.15	7
AJST	26	22.31	24
JLVR	0	10	---

C. Identificação

Foi tambem efectuado um teste de identificação, usando uma rede com 5 unidades de saída, correspondentes ao mesmo número de clientes.

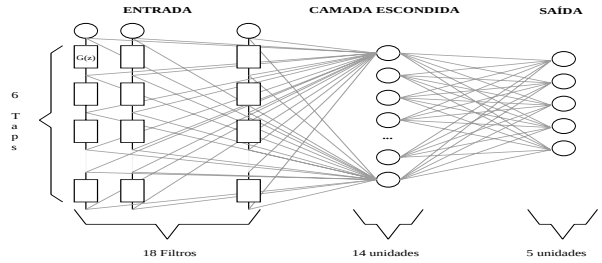

Fig 13 - Rede com 5 saídas

A média de decisões correctas para os 5 clientes foi de 72 %. O pior orador, o mesmo qu causou problemas na identificação, apenas atingiu 40 % e os outros 4 atingiram os 60, 85, 90 e 95 %. Os impostores foram rejeitados correctamente 62 % dos casos. Os impostores com

padrões de treino foram rejeitados, correctamente, 82 % dos testes, mas os outros impostores, sem treino, apenas foram rejeitados correctamente 52 % dos testes.

CONCLUSÕES

Apesar das más condições na recolha do sinal e da pequena base de dados utilizada, dos testes efectuados podem tirar-se as seguintes conclusões:

- Os resultados na verificação podem ser considerados bons, atendendo às limitações (ruído, pequena base de dados, segmentação pouco precisa, utilização de apenas uma frase). No entanto existem problemas para a utilização prática, tais como a melhor topologia ser dependente do orador; o desempenho também variar com o orador; e o treino ser bastante demorado.
- Os resultados da identificação são inferiores aos da verificação. A deterioração dos resultados pode estar relacionada com o reduzido número de padrões de treino (em especial de impostores).
- O método é robusto, pois sem grandes cuidados na recolha e segmentação obtiveram-se resultados aceitáveis.
- Os parâmetros extraídos pelo banco de filtros possuem informação sobre a identidade do orador, além de informação sobre as palavras proferidas.
- A rede gama concentrada tem capacidade para extraír dos parâmetros diferênças entre os oradores.

AGRADECIMENTOS

Este trabalho foi desenvolvido graças à Bolsa de Mestrado Nº BM-1763/91- IA da JNICT.

Agradecemos ao grupo da Universidade de Estugarda que desenvolveu o simulador de redes neuronais SNNS e ao Computational Neuroengineering Laboratory da Universidade da Florida.

REFERÊNCIAS

[1] Ronald S. Cheung e Bruce A. Eisenstein, "Feature Selection via Dynamic Programming for Text-Independent Speaker Identification", IEEE Trans. on Acoustics, *Speech and Signal Processing*, ASSP-26, n.5, pp. 397-403, Outubro 1978

[2] George R. Doddington, "Speaker Recognition - Identifying People by their Voices", *Proc. IEEE*, 73 (11), pp. 1651-1664, Novembro 1985

[3] Douglas O' Shaughnessy, "Speaker Recognition", *IEEE ASSP Magazine*, pp. 4-17, Outubro 1986

[4] Clifford J. Weinstein, "Opportunities for Advanced Speech Processing in Military Computer-Based Systems", *Proc. IEEE*, 79 (11), pp. 1626-1641, Novembro 1991

[5] Aaron E. Rosemberg, "Automatic Speaker Verification: A Review", *Proc. IEEE*, 64(4), pp. 475-487, Abril 1976

[6] Bishnu Atal, "Automatic Recognition of Speakers from their Voices", *Proc. IEEE*, 64(4), pp. 460-474, Abril 1976

[7] Aaron E. Rosemberg and Frank K. Soong, "Recent Research in Automatic Speaker Recognition", In *Advances in Speech Signal Processing*, pp. 701-738, 1992

[8] J. Oglesby and J. S. Mason, "Optimization of Neural Models for Speaker Identification",In *Proceedings ICASSP*, pp. S5.1, Abril 1990

[9] Y. Bennani and F. Fogelman and P. Gallinari, "A Connectionist Approach to Speaker Identifiaction", In *Proceedings ICASSP*, pp. S5.1 Abril 1990

[10] Y. Bennani and F. Fogelman and P. Gallinari, "Text-Dependent Speaker Identification Using Learning Vector Quantization", In *Proceedings INNC*, Paris, France, Julho 1990

[11] Younés Bennani and Patrick Gallinari, "A Modular Connectionist Architecture For Text-Independent Talker Identification", In *Proceedings IJCNN*, Julho 1991

[12] Hiroaki Hattori, "Text-Independent Speaker Recognition Using Neural Networks",In *Proceedings ICASSP*, pp. II-153-157,1992

[13] Bert de Vries e José C. Principe , "The Gamma Model - A New Neural Model for Temporal Processing", *Neural Networks*, 5, pp. 565-576, 1992

[14] John D . Markel e Steven B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base", *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-27, n.1, pp. 74-82, Fev. 1979

[15] John Holdsworth, Ian Nimmo-Smith, Roy Patterson e Peter Rice, "Implementing a GammaTone Filter Bank", *Annex C of the SVOS Final Report (Part A: The Auditory Filter Bank)*, Fev. 1988

[16] B. C. J. Moore e B. R. Glasberg, "Cochlear Modeling", *J. Acoust. Soc. Am.*, 74, pp. 750-753, 1983

[17] M. C. Mozer, "A focused back propagation algorithm for temporal pattern recognition", *Complex Systems*, 3, pp. 349-381, 1989

[18] Tomás Oliveira e Silva, "Generalized Feed-forward Filters: Some Theorectical Results", In *Proceedings ICASSP*, pp. 109-112, vol. III,1993

[19] K. P. Unnikrishnan, J. J. Hopfield e D. W. Tank, "Connected-digit speaker-dependent speech recognition using a neural network with time-delayed connections", *IEEE Trans. on Signal Processing*, vol. 39, n. 3, pp. 698-713, 1991

[20] R. L. Watrous, B. Ladendorf e G. Kuhn, "Complete gradient optimization of recurrent network applied to /b/,/d/,/g/ discrimination", *J. Acoust. Soc. Am.*, 87(3), pp. 1301-1309, Mar. 1990

[21] Bert de Vries, Leslie Dias e John Pearson, "Learning with target trajectory constraints for Sequence Classification tasks", In *Proceedings ICASSP*, pp. 525-528, vol. II,1993

[22] Andreas Zell, Niels Mache, Ralf Hübner, Michael Schmalzl, Tilman Sommer, Günter Mamier e Michael Vogt, "SNNS - Stuttgart Neural Network Simulator", Universität Stuttgart, Institute for Parallel and Distributed High Performance Systems (IPVR), User Manual Number 8/92, Fed. Rep. of Germany, 1992

[23] António J. S. Teixeira e Francisco A. C. Vaz, "Using SNNS in Speech Recognition", In *Proceedings of the Workshop on Simulation of Neural Networks with SNNS (SNNS'93)*, Universitaet Stuttgart, IPVR, Federal Republic of Germany, Setembro 1993

[24] M. R. D. Martins, *Ouvir Falar - Introdução à Fonética do Português.* CAMINHO, 1988.

[25] L. R. Rabiner e R. W. Schafer, *Digital Processing os Speech Signals.* Prentice-Hall, Englewood Cliffs, NJ, 1978