Reconhecimento do Orador

João Paulo Guedes, Pedro Miguel Figueirôa, António Teixeira, Francisco Vaz

***Resumo* – A necessidade de verificar a identidade duma pessoa tendo por base características intrínsecas à sua voz é um processo em evolução e de importância singular no acesso e processamento de informação, segurança e, em geral, nas telecomunicações.**

Neste trabalho procurou-se estudar várias possibilidades colocadas na implementação dum sistema automático de verificação do orador, independente do texto baseado no reconhecimento de fonemas. Pretendeu-se determinar qual o tipo de parâmetros a utilizar, a respectiva estabilidade no tempo e quais os fonemas a usar. Este trabalho utiliza redes neuronais artificiais como elemento de classificação.

***Abstract* - Verification of a person identity based in voice alone is an evolving technology with applications in secure access to information, access control to buildings, and telecommunications in general. This work addresses several of the possibilities available in the implementation of an automatic speaker verification system, independent of text, based in phoneme recognition. We studied which parameters to use, their stability over time, and also the phonemes to use for better performance. In this work we used feedforward neural networks in the verification process.**

I. INTRODUÇÃO[1]

Actualmente, qualquer projecto na área de processamento de voz, nomeadamente na área do reconhecimento do orador reveste-se da maior importância.

Define-se reconhecimento do orador como a tarefa que permite distinguir indivíduos baseando-se apenas nas características das suas vozes. Porém, é conveniente distinguir reconhecimento do orador do reconhecimento de voz. Este último procura reconhecer o que é dito e não quem o diz.

A verificação e identificação do orador são tarefas distintas no reconhecimento do orador: a identificação faz a classificação de uma voz pronunciada por um orador supostamente conhecida pelo sistema enquanto a verificação decide se essa voz pertence, de facto, a um utilizador conhecido pelo sistema. Este processo baseia-se numa decisão binária. Ou aceita ou rejeita a identificação dependendo de um limiar de decisão.

Nesta área existem dois tipos de sistemas: sistemas dependentes e independentes do texto. Os primeiros são treinados com poucas palavras e não apresentam a capacidade de aprendizagem a outras como os sistemas independentes.

Apesar de em muitas áreas ser muito difícil igualar o desempenho humano, nesta área os dados experimentais sugerem que o desempenho das máquinas em muitos casos excede a dos seres humanos [1].

O objectivo final de qualquer estudo 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].

II. INTERESSES NA ÁREA

Hoje em dia é da maior importância encontrar sistemas que permitam identificações pessoais com a maior segurança possível. Esta exigência é essencial em sistemas de segurança como sistemas de controlo de acesso pessoal, controlo de acesso à informação, controlo de transacções telefónicas automáticas.

O reconhecimento do orador constitui um exemplo de identificação pessoal biométrica, ou seja depende de características intrínsecas ao próprio indivíduo. Logo, estas técnicas são mais fiáveis e os seus atributos não podem ser esquecidos nem perdidos como nos casos de *passwords* ou cartões magnéticos. Esta técnica quando comparada com outras técnicas biométricas, como por exemplo, a utilização da retina do olho humano ou das impressões digitais, apresenta uma grande vantagem, uma vez que a recolha de dados de voz é fácil e o seu tratamento pouco dispendioso. Por outro lado, as tecnologias de identificação e verificação existentes têm taxas de erro inferiores a 2%.

III. MOTIVAÇÃO

Foi com esta motivação que ao longo do ano lectivo de 1997/98 se desenvolveu este projecto na área do reconhecimento do orador no seguimento de trabalhos anteriores [1].

As diferenças fundamentais ao nível do desenvolvimento prático deste projecto e do trabalho referido, consistem no modelo de análise utilizado, nos parâmetros, no treino da rede neuronal e da ferramenta

[1] Trabalho realizado no âmbito da disciplina de Projecto

utilizada. A tese de mestrado referenciada usava um modelo de banco de filtros e uma abordagem dependente do texto enquanto este projecto se baseia num modelo LPC e faz uma abordagem independente do texto

IV. MODELO DE RECONHECIMENTO

O modelo de reconhecimento apresenta três passos básicos: extracção de parâmetros, comparação de padrões e a decisão.

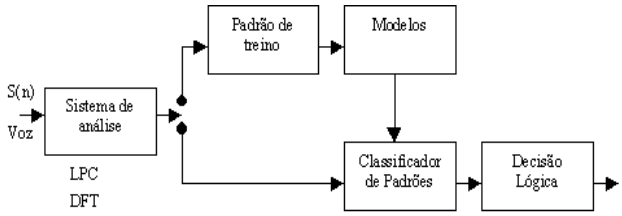

Figura 1 Modelo de Reconhecimento

Neste diagrama de blocos, que representa um sistema de reconhecimento tem-se na entrada um sinal de voz. Este sinal passa por um sistema de análise onde se extraem um conjunto de parâmetros representativos da informação de voz intrínsecas ao orador. Estes parâmetros pretendem-se o mais estáveis possíveis, ou seja, que dependam o menos possível dos factores que atribuem ao sinal variações ao longo do tempo. Inicialmente o sistema é "treinado" de forma a criar um padrão de referência para cada orador. Sempre que se pretender testar o sistema compara-se o padrão de referência de um dado orador com o seu padrão de teste seguindo-se um processo de decisão lógica. Sempre que se pretende adicionar um novo orador ao sistema tem que se treinar o sistema para esse mesmo utilizador.

V. BASE DE DADOS

De forma a tornar este projecto possível foi necessário criar de raiz uma base da dados em português uma vez que não existia nenhuma disponível. Os oradores escolhidos eram todos universitários do sexo masculino, com idades compreendidas entre os 26 e os 28 anos e originários de vários pontos do país.

Os dados recolhidos são compostos por 80 ficheiros (tipo .wav) por orador sendo cada um constituído por dez (10) palavras. Fez-se a opção de metade da recolha incidir em palavras que contêm vogais orais e as restantes vogais nasais.

A escolha de utilização de vogais em vez de consoantes ou ditongos não foi aleatória. Cada vogal pode ser caracterizada pela configuração do tracto vocal que é utilizado na sua produção. Como a configuração difere de orador para orador, as vogais podem fornecer uma pista para a identidade do orador, daí a opção do uso de vogais.

Em Neena Jain [3] é apresentada uma comparação entre várias classes de fonemas, sendo as vogais acentuadas as que obtêm um melhor desempenho.

Na tabela seguinte indicam-se as palavras utilizadas:

Vogais	
Orais	Nasais
PAPO	CANTANTE
BATO	PINTO
PICAR	CINCO
CATO	PONTO
TIPO	CANTO
FITO	FINTAR
ATINAR	TONTO
FITAR	CONSTANTE
KILO	TINTO
POLIR	CONTO

Tabela 1 - Palavras proferidas pelos Oradores

O *labelling* manual das vogais foi feito através duma observação do espectograma.

É de salientar que toda a recolha foi realizada em sala insonorizada recorrendo-se a *software* e *hardware* de nível profissional tendo a base de dados cerca de 101Mbytes[2].

VI. SISTEMA AUTOMÁTICO DE VERIFICAÇÃO

Um sistema automático de reconhecimento apresenta três blocos fundamentais, tal como se apresenta no seguinte diagrama de blocos:

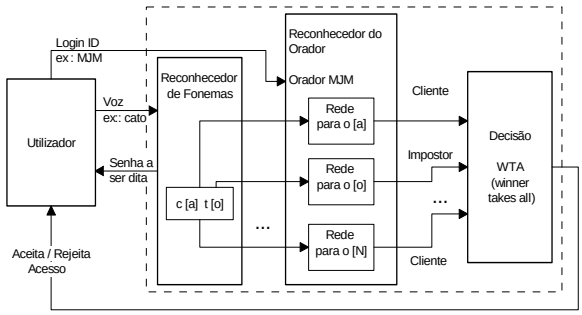

Figura 2 – Sistema automático de reconhecimento.

A interacção existente entre o utilizador e o sistema limita-se ao pedido por parte do sistema de uma senha a ser dita pelo utilizador e um *login* a ser introduzido.

O primeiro bloco é um reconhecedor de fonemas automático, que analisa o sinal de voz, isto é, reconhece certos fonemas e regista esta informação passando-a ao bloco seguinte. O bloco seguinte tem como entrada o sinal de voz com fonemas devidamente assinalados e um *login* fornecido pelo utilizador. Com estes dados procura na sua base de dados a rede neuronal correspondente a este orador e aos fonemas detectados. Seguidamente introduz na entrada da rede o fonema e estima a probabilidade do

[2] Agradecemos ao Director Técnico da Rádio Regional de Aveiro, José Ruivo, pela colaboração prestada.

fonema ter sido ou não proferido pelo suposto orador, ou seja, se é um cliente ou impostor. Este processo repete-se para cada um dos fonemas detectados. O bloco de decisão recolhe a informação do bloco anterior e decide se o utilizador é ou não um elemento do sistema, aceitando ou rejeitando o acesso. Existem vários processos de decisão sendo o mais comum a decisão por maioria. Assim se o maior número de entradas é de um cliente este aceita o seu acesso.

VII. ETIQUETAGEM

Neste trabalho não foi implementado o primeiro bloco, devido à indisponibilidade temporal para a sua execução. Utilizou-se uma etiquetagem (ou *labelling*) manual, especificamente para as vogais orais e nasais anteriormente referidas. Na figura seguinte ilustra-se o processo de etiquetagem:

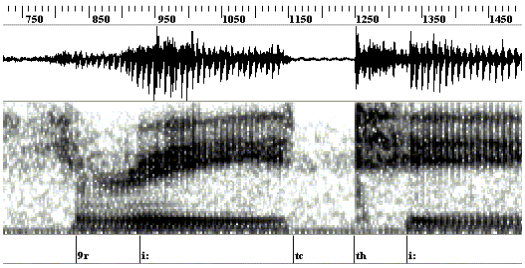

Figura 3- Etiquetagem ou *labelling*

VIII. FERRAMENTAS

As ferramentas utilizadas quer para a etiquetagem quer no restante procedimento, foram o *Tcl/Tk* e o *CSLU Toolkit*. *Tcl* significa "*Tool command Language*" e devido à sua flexibilidade e extensão de comandos è ideal para a implementação de aplicações que usem um grande agregado de ferramentas. O *Tcl* apresenta muitas extensões, sendo a mais comum e importante o *Tk*. Esta permite ao utilizador criar e manipular interfaces. A outra ferramenta, o *CSLU Toolkit* – "*Center for Spoken Language Understanding*" – fornece um ambiente baseado no *Tcl* para investigação e desenvolvimento de sistemas de voz. O *Toolkit* permitiu a manipulação de ficheiros do tipo .wav, a etiquetagem manual baseada no espectograma, a extracção de parâmetros e o uso de redes neuronais artificiais.

Foi, também, apartir deste *Toolkit* que se desenvolveu um processo de treino de redes neuronais. Originalmente foi desenvolvido para reconhecer palavras, no entanto, após alteração de algumas *scripts* tornou-se numa solução para este trabalho.

IX. TREINO, DESENVOLVIMENTO E TESTE DAS REDES NEURONAIS

A. Ficheiros

Neste processo destacam-se três tipos de ficheiros: *.wav*, *.txt* e *label files*. O primeiro contém a forma de onda correspondente à palavra (ou grupo de palavras) a ser reconhecida. Os ficheiros .txt contêm uma transcrição, em texto, das palavras presentes na *wave file*. Apesar de neste caso este tipo de ficheiro não ser utilizado, tem de se indicar ao *Toolkit* a sua possível localização, sendo do ponto de vista deste considerados NULL. A *label file*, que usualmente têm a extensão .phn, .cat ou .wrd, contém os *labels* da onda, ou seja, onde ocorre certo fonema no registo de voz. Como em geral se usa palavras com vários tipos de fonemas (neste caos, por exemplo, o [a] , [i] e [u]) com que se pretende executar a tarefa de reconhecimento, na *label* file pode haver vários tipos de fonemas indicados, este tipo de file é o resultado da etiquetagem. Estes ficheiros têm o seguinte formato:

```
MillisecondsPerFrame: <value>
END OF HEADER
   <begin_time_1> <end_time_1> <label_1>
…
   <begin_time_n> <end_time_n> <label_n>
```

onde:

< value > é o número de milisegundos num *frame* de voz (usualmente este valor é 1).

<begin_time> é a altura em que inicia a < *label* >.

<end_time> é a altura em que se termina a < *label* >.

< *label* > é a palavra, fonema ou categoria para o segmento de voz.

B. *Descrição Geral*

O processo de treino, desenvolvimento e teste está ilustrado no seguinte fluxograma:

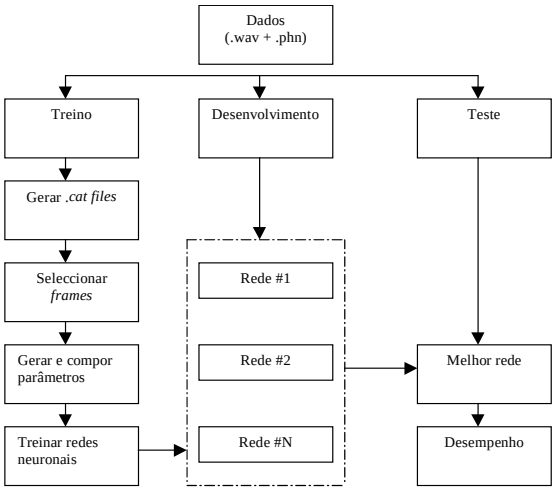

Figura 4 – Processamento de treino, desenvolvimento e teste

Como se observa, divide-se os dados em três conjuntos: treino, desenvolvimento e teste. Com o conjunto de treino geram-se as *.cat files*. Teve de se alterar o código de forma a criar as *.cat files* no formato pretendido. Houve necessidade de associar a um cliente ou impostor um determinado padrão, ou seja, se pretendermos que o cliente seja AMG, por exemplo, todos os outros têm de ser impostores. Para qualquer pessoa isto é evidente, no entanto, temos de arranjar um processo de indicar à máquina que tem de associar certos *frames* a um cliente ou a um impostor. Para este processo de associação usamos as *files .phn* que contêm os *labels* da onda. Esta associação é feita da seguinte forma:

1. Determinar qual é o parâmetro de entrada (cliente ou impostor).
2. Procurar em todos os ficheiros *.phn* a vogal do cliente.
3. Escrever as *.cat files* correspondentes, substituindo o local onde aparece a vogal por cliente e as restantes vogais são marcadas como <x>.
4. Para gerar as *.cat files* dos impostores sucede um processo análogo. A vogal dos impostores passa a estar marcada como <impostor> e as outras vogais como <x>.

O passo seguinte é seleccionar *frames* para treinar a rede, ou seja, criar uma listagem (neste caso binária) das *files*, dos *frames* de cada *file* e a categoria correspondente a estes *frames* (cliente ou impostor).

- Parâmetros

O bloco seguinte gera e compõe os parâmetros. Aqui, define-se objectivamente o tipo de parâmetros a usar: LPC, Cepstra, PLP, RASTA, etc. Após conclusão deste processo inicia-se um outro que torna aleatória a ordem dos vectores de treino de forma a ajudar o programa de treino a aprender melhor a informação.

- Treino das redes neuronais

Finalmente começa o treino das redes neuronais. Geram-se vários ficheiros de pesos da rede (um por iteração) até atingir 30 iterações. Note-se que as trinta iterações foi o valor considerado suficiente para que a aprendizagem ocorre-se, sem que houvesse excesso *de* treino. Assim da fase de treino obtemos trinta redes neuronais, uma por cada iteração.

- Desenvolvimento e teste da rede neuronal

Coloca-se agora a questão: qual delas "aprendeu" melhor, ou seja, qual a que tem melhor desempenho ? Para responder a esta questão, utilizamos o conjunto de desenvolvimento, podendo observar qual o desempenho de cada rede. Obtemos desta forma a melhor rede. Para finalizar o processo testamos a rede com o conjunto de teste, obtendo o desempenho da rede para uma vogal (neste caso) de um dado orador. Este processo ocorre para todas as vogais de todos os oradores. Para cada rede pretende-se que haja uma decisão, ou seja, se o fonema pertence a um cliente ou a um impostor. Assim o processo de decisão é bastante importante como tal, passa-se a descreve-lo:

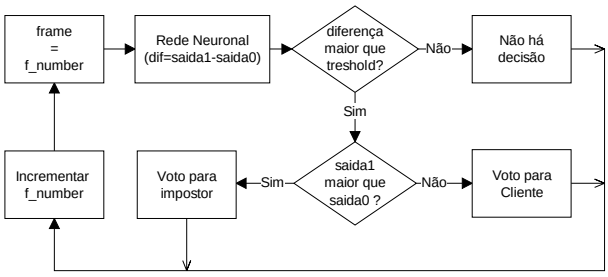

Figura 5 Fluxograma do processo de decisão

Coloca-se um *frame* de cada vez na entrada da rede. A rede tem duas saídas correspondendo a cliente e impostor. Analisa-se a diferença entre as duas saídas (dif), caso seja menor que um dado limiar não se toma qualquer decisão. Caso seja superior decide-se pela maior saída: cliente ou impostor. Este processo repete-se até que todos os *frames* sejam analisados. Após serem consideradas todos os *frames*, a categoria com mais votos será escolhida.

X RESULTADOS

Apresenta-se em seguida os resultados das várias comparações efectuadas desde o início do procedimento experimental. Estes são apresentados sob a forma de tabela estando em destaque o tipo de comparação em causa.

A. 5 frames versus 1 frame

Pretendeu-se comparar o desempenho do sistema usando ou não contexto.

	Perceptual Linear Prediction					
	Vogais Orais – Sessão 1 – 5 *frames*					
	A		I		u/o	
	Dev	Teste	Dev	Teste	Dev	Teste
Média	100	97	100	96	98.3	93
	Vogais Orais – Sessão 1 – 1 *frame*					
Média	76.7	76	80	80	80	78

Tabela 2 – 5 frames vs. 1frame

Observa-se que, para o conjunto de teste, quando usamos apenas um *frame* o desempenho do sistema cai extraordinariamente. Em termos médios chega a atingir uma queda de 20% e pontualmente cerca de 35%. Este resultado já era esperado pois estamos a usar menos informação dos *frames* quando apenas se usa um.

B. Orais versus Nasais

Pretendeu-se com esta comparação avaliar qual o tipo de vogal mais indicada para reconhecimento do orador: as vogais orais ou nasais ?

	Perceptual Linear Prediction – 5 *frames*					
	A		i		u/o	
	Dev	Teste	Dev	Teste	Dev	Teste
Média Orais	100	97	100	96	98.3	93
Média Nasais	84	89	89	88	91	92

Tabela 3 – Orais vs. Nasais (PLP)

	RASTA – 5 *frames*					
	a		I		u/o	
	Dev	Teste	Dev	Teste	Dev	Teste
Média Orais	98.7	95	98.3	93	96.7	95
Média Nasais	86	83	86	84	91	90

Tabela 4 – Orais vs. Nasais (RASTA)

Por análise das tabelas observa-se que as vogais nasais fazem com que o desempenho da rede diminua. Quando se obteve os resultados referentes aos parâmetros PLP, poder-se-ia supor que a descida dos resultados se devesse essencialmente ao tipo de parâmetros usados. Pondo esta hipótese recorreu-se a outro tipo de parâmetro para verificar se este comportamento se mantinha ou não. Após obtenção dos resultados com parâmetros RASTA confirmou-se a tendência de descida do desempenho, quando passamos de vogais orais para nasais.

C. Comparação de Parâmetros

A questão seguinte colocou-se ao nível dos parâmetros. Qual dos parâmetros conduz a um melhor desempenho? Para tentar responder a esta questão efectuámos uma comparação entre vários parâmetros, tendo sempre em consideração os resultados anteriores. Assim mantivemos como constantes o tipo de vogal (oral) e o número de *frames* (cinco):

	Vogais Orais – Sessão 1 – 5 *frames*					
	a		I		u/o	
	Dev	Teste	Dev	Teste	Dev	Teste
Cepstra	100	99	100	100	99	99
Rasta	98.7	95	98.3	93	96.7	95
PLP	100	97	100	96	98.3	93

Tabela 5 – Desempenho dos vários parâmetros

Desta comparação, torna-se óbvio que o melhor desempenho é conseguido pelo parâmetro CEPSTRA. Este obteve um desempenho para todas as vogais orais de cerca de 100%. No caso dos parâmetros PLP estes já só têm um desempenho que oscila entre os 93 e 97%.

D. Sessão1 versus Sessão2

Quando a rede é treinada esta apresenta um dado tipo de desempenho. Será que este se mantém constante no tempo? Será que alterações de voz ao longo do tempo não influenciarão o desempenho do sistema? Com a comparação seguinte tentaremos esclarecer esta questão.

A base de dados subjacente a este projecto contém duas sessões de gravações espaçadas no tempo (cerca de 1.5 meses). Com a primeira sessão pretendia-se treinar e testar uma rede. A segunda sessão tinha como objectivo único determinar se o desempenho se mantinha estável após um espaço temporal. Nas tabelas seguintes apresentam-se os resultados de ambas as sessões de forma a ser possível comparar os desempenhos. Outra vertente desta comparação também se prende com os parâmetros utilizados, ou seja, pretende-se também avaliar se um dado parâmetro, apesar de anteriormente não ter obtido um bom desempenho, mantém o seu comportamento.

	Vogais Orais – 5 *frames*					
	Sessão1			Sessão2		
	a	i	U	A	i	u
Cepstra	99	100	99	89.4	89.9	90
Rasta	95	93	95	85.4	86.3	84
PLP	97	96	93	87.2	89.2	84.8

Tabela 6 - Vogais Orais: Sessão 1 vs. Sessão2

	Vogais Nasais – 5 *frames*					
	Sessão1			Sessão2		
	a	i	U	A	i	u
Rasta	83	84	90	85.6	88.6	87.8
PLP	89	88	92	87	89.2	90.8

Tabela 7 – Vogais Nasais : Sessão1 vs. Sessão 2

Como se pode observar em geral o desempenho diminuiu consideravelmente de uma sessão para a outra. No caso das vogais orais, este facto torna-se claro para todos os casos. No entanto, tem de se registar que no caso das vogais nasais o desempenho aumentou nalgumas situações, o que pode indicar que, apesar destas numa fase de treino terem pior desempenho, mantêm uma certa invariância ao longo do tempo.

XI. DISCUSSÃO E CONCLUSÕES

Depois de fazer a análise dos resultados obtidos e salvaguardando alguns pontos que possam não ter sido

exaustivamente testados chegámos às seguintes conclusões:

1. O uso de vários *frames* melhora o desempenho do sistema.
2. Actualmente considera-se que as vogais são os fonemas mais úteis na tarefa de reconhecimento. Dentro das vogais , as vogais orais indiciam, no âmbito do estudo realizado, um melhor desempenho do que as nasais. Este facto poderá ser devido às vogais nasais apresentarem uma maior coarticulação, ou seja, são mais dinâmicas e, consequentemente, mais complexas, ou pelo facto da ordem dos modelos LPC não ser adequada ás vogais nasais.
3. Apesar de poucos tipos de parâmetros terem sido testados, aqueles que apresentaram melhores resultados foram os cepstrais confirmando a bibliografia consultada.
4. Quando se tenta determinar qual dos parâmetros apresenta maior estabilidade temporal não se consegue chegar a nenhuma conclusão satisfatória. Se por um lado houve um decréscimo significativo no desempenho por parte do sistema, para as vogais orais após o segundo teste, por outro, verificou-se que as vogais nasais conseguiram manter ou melhorar o seu desempenho, que pode ser indicativo de uma maior estabilidade. Torna-se assim inconclusiva a nossa comparação.

XII. TRABALHO FUTURO

Existe um número de problemas interessantes que deveriam ser explorados em futuros estudos. Alguns deles incluem-se nos seguintes itens:

1. Implementar o sistema de reconhecedor de Fonemas (1º bloco da figura 2).
2. Criar um protótipo funcional em *Tcl/Tk* visando uma implementação automática do reconhecedor.
3. Aumentar a base de dados trabalhando com um conjunto mais alargado de pessoas.
4. Melhorar a relação desempenho/limiar do elemento de decisão, visando estabelecer uma relação óptima para este mesmo elemento.
5. Assumindo que haverá sempre variação é necessário determinar com exactidão o tempo necessário para retreinar a rede, ou seja, actualizar a rede neuronal.
6. Atendendo à disparidade do número de amostras entre os clientes e os impostores utilizadas no treino da rede neuronal no futuro dever-se-á tentar recolher mais amostras de forma a tentar diminuir esta diferença. Apesar de não serem apresentados os resultados obtidos em termos de acerto (exemplo: número de vezes que um cliente ou impostor foi reconhecido como cliente/impostor) notou-se que a rede teve uma maior facilidade de identificar os impostores como impostores, relativamente a qualquer outra combinação possível. O número de amostras para impostores foi sempre muito superior à dos clientes, ou seja, a probabilidade *a priori* da classe impostor é superior à probabilidade da classe cliente.
7. Estender o conceito implementado para as vogais a outro tipo de fonemas, ou mesmo associá-los. Assim tornar-se-ia o reconhecedor dependente de vários tipos de parâmetros.
8. Trabalhar com imitadores.

BIBLIOGRAFIA

[1] António J S Teixeira, Reconhecimento do Orador com Redes Neuronais, Tese de Mestrado, Universidade de Aveiro, 1993.

[2] António Teixeira, Francisco Vaz, “Reconhecimento do Orador com Redes Neuronais” , Revista do Dep. Electronica e Telecomunicações da Universidade de Aveiro, vol. 1, Nº 1, Janeiro 1994.

[3] Neena Jain, “A New Aproach to voice dialing”, Thesis for the degree Master of Science in Electrical Engineering, OGI, 1992

[4] Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of speech recognition, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[5] Aaron E. Rosenberg, “ Recent Research in Automatic Speaker Recognition” , AT&T Laboratories