ANÁLISE DA RESPOSTA DE REDES NEURAIS RECORRENTES DE JORDAN E ELMAN PARA O RECONHECIMENTO DE LOCUTOR COM DISTINTAS TÉCNICAS DE PRÉ-PROCESSAMENTO

RODRIGO T. CAROPRESO*, RICARDO A. S. FERNANDES†, IVAN N. SILVA‡

*Departamento de Desenvolvimento de Software - FIT - Flextronics Instituto de Tecnologia
Avenida Liberdade, 6315 - Prédio 4 - Bairro Iporanga - Sorocaba - CEP 18087-170

†Departamento de Engenharia Elétrica - CCET - Universidade Federal de São Carlos
Rod. Washington Luís, km 235 - SP-310, CEP 13565-905, São Carlos - SP

‡Departamento de Engenharia Elétrica - EESC - Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - CEP 13566-590, São Carlos - SP

Emails: rodrigo.caropreso@fit-tecnologia.org.br, ricardo.asf@ufscar.br, insilva@sc.usp.br

Abstract— The speaker recognition still stands as a field which a great potential on researching and analysis. Based upon this premise, this paper proposes a comparative analysis on the use of signal conditioning techniques altogether with an Elman and Jordan Recurrent Neural Networks for the speaker recognition task. The experiments have shown relevant results, with a faster learning curve when compared to a Multilayer Perceptron Neural Network, assuring the proposed method as a very good approach on this particular problem.

Keywords— Artificial Neural Networks, Speaker Recognition, Cepstrum, Elman, Recurrent Networks, Perceptron.

Resumo— O problema de classificação de padrões e particularmente, o reconhecimento de locutor mostra-se como uma área onde há ainda muito a ser explorado. Este artigo propõe uma análise comparativa do desempenho de Redes Neurais Artificiais Recorrentes de Elman e de Jordan para o reconhecimento de locutor, utilizando ferramentas destinadas ao pré-processamento e condicionamento de sinais. Com base nos experimentos, foi possível constatar o satisfatório desempenho alcançado pela metodologia proposta quando comparada à resposta de um Perceptron de Múltiplas Camadas, com maior rapidez de aprendizado para o paradigma neural recorrente.

Palavras-chave— Redes neurais artificiais, identificação de locutor, pré-processamento de sinais, Elman, Jordan, Redes Recorrentes.

1 Introdução

O processo de reconhecimento automático de locutores a partir de sua voz *(RAL)* é um problema de alta complexidade que tem recebido visibilidade e relevância em pesquisas devido ao crescimento das tecnologias de áudio e comunicação disponíveis.(Hansen and Hasan, 2015; Togneri and Pullella, 2011). Este problema consiste em identificar um indivíduo através da comparação de características extraídas de seu sinal de voz, onde se destacam dois domínios de problema distintos: identificação de locutor, onde determina-se a qual locutor dentre N indivíduos pertence uma dada elocução teste; e verificação de locutor, que é o processo de aceitar ou rejeitar a identidade pretensa de um locutor teste.

Assim, devido a relevância do tema, existem diversas abordagens na literatura que buscam oferecer os melhores resultados, podendo ser enumeradas em: técnicas probabilísticas, onde se mede a probabilidade condicional ao se comparar os dados com o modelo, tendo como destaque as os estudos dos Modelos Ocultos de Markov *(HMM)* e Modelos de Mistura de Gaussianas *(GMM)* (Bhaykar et al., 2013; Bao and Shen, 2016); ou técnicas determinísticas, baseadas em medidas de distância entre o sinal avaliado e o modelo *(codebook)*, onde cabe citar os estudos utilizando Quantização Vetorial (Hayet et al., 2014).

Ainda dentro da análise de extração de características para o RAL, as técnicas mais populares e ainda em estudo consideradas como estado da arte, utilizam os Coeficientes Mel-Cepstrais *(MFCCs)* (S and S, 2015; Singh et al., 2014; Jokic et al., 2015), Coeficientes de Predição Linear *(LPC)* (S and S, 2015) e Wavelets (Barbosa and Silva, 2015).

Além disso, o uso de sistemas inteligentes empregados na tarefa do RAL tem se mostrado promissores, conforme os estudos de (Barbosa and Silva, 2015; Chu and Chen, 2016), onde pode-se destacar o uso de Redes Neurais Artificiais em (Chu and Chen, 2016; Weng et al., 2014), mostrando que o paradigma de Redes Neurais Artificiais *(RNA)* ainda é relevante na tarefa de processamento do *RAL*.

Dentro do contexto supracitado, este trabalho propõe uma análise comparativa de desempenho de duas RNAs Recorrentes (Rede de Jordan e Rede de Elman) no processo do RAL, onde se combinou um pré-processamento de sinal baseado na extração dos Coeficientes Mel-Cepstrais *(MFCCs)* aos classificadores neurais. É importante destacar que o uso destas redes foi pouco explorado na li-

teratura e seu uso não foi descartado por estudos recentes, sendo este o ponto no qual o presente trabalho encontra sua contribuição.

Com o propósito de validar a arquitetura proposta, os resultados deste trabalho são comparados aos resultados obtidos em (Caropreso et al., 2013), onde a arquitetura de RAL foi empregada utilizando uma RNA do tipo Perceptron de Múltiplas Camadas *(PMC)*. Desta forma, é possível não somente analisar a influência do pré-processamento na taxa de acerto das redes recorrentes, mas também comparar o desempenho das redes recorrentes em relação ao PMC em iguais condições de simulação.

O presente trabalho divide-se da seguinte forma: na Seção 2, será descrito o pré-processamento do sinal de voz utilizado; a seção 3 descreve as Redes Recorrentes utilizadas no estudo; a Seção 4 apresenta a arquitetura do RAL baseado em RNA; na Seção 5, são apresentados os resultados obtidos pela metodologia proposta e na Seção 6, são feitas as considerações finais e conclusões do trabalho.

2 Pré-processamento de sinais

Destaca-se que tanto a base de dados de voz quanto a sequência de pré-processamento utilizada neste trabalho foram as mesmas utilizadas em (Caropreso et al., 2013) e, desta forma, o pré-processamento pode ser representado pelo diagrama da Figura 1:

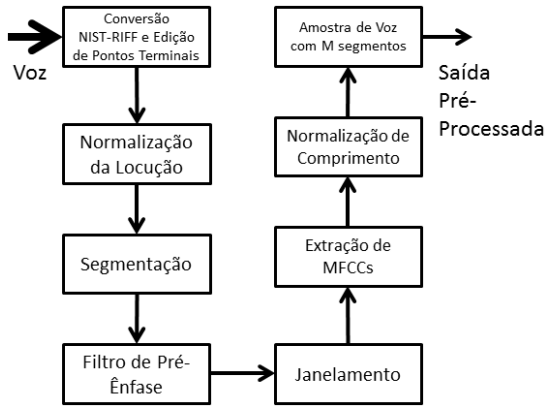

Figura 1: *Etapa de pré-processamento dos sinais*

Inicialmente, os sinais de voz foram convertidos do formato NIST para RIFF e, a seguir, foram normalizados a fim de minimizar interferências no sistema de reconhecimento.

Na sequência, as locuções foram segmentadas por meio de conjuntos de 512 pontos e submetidas a um filtro de pré-ênfase cuja função de transferência é dada pela Equação (1):

$$H(z) = 1 - 0,95z^{-1} \quad (1)$$

Na etapa seguinte, cada segmento passa por uma janela de Hamming, dada por (2):

$$W^{(n)}_{HAMMING} = 0,54 - 0,46\cos\left(2\pi\frac{n}{LJ}\right) \quad (2)$$

onde LJ é a largura da janela utilizada (512 pontos) e n é o valor da amostra ($0 \leq n \leq LJ$).

A partir das janelas de Hamming, são obtidos os coeficientes Mel-Cepstrais, conforme a seguinte expressão:

$$C_p(j) = \frac{2}{K}\sum_{k=1}^{K}\log(X_k(j))\cos\left(p(k-0,5)\frac{\pi}{K}\right) \quad (3)$$

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-ésimo filtro calculada para a janela j. Neste trabalho são empregados 80 filtros com 24 Coeficientes Mel-Cepstrais, mantendo o padrão utilizado em (Caropreso et al., 2013).

Após a obtenção dos MFCCs, a última etapa do pré-processamento consiste na normalização do comprimento das locuções, a fim de evitar amostras excessivamente longas para um dado locutor.

As técnicas de alinhamento temporal utilizadas foram:

- Truncamento Simples *(TRUNC)* – consiste em eliminar o excesso de termos do vetor de locuções até que atinja o valor desejado para ser usado como entrada da RNA;
- *Trace Segmentation (TS)* – baseia-se 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 distintos;
- *Linear Time Warping (LTW)* – consiste em separar as linhas do vetor que compõe a locução em seções e calcular a média de cada seção de forma a produzir um único vetor de atributos por seção;
- *Minimal Temporal Information (MTIs)* – são conjuntos de segmentos consecutivos extraídos a partir da sequência de segmentos que representam uma locução (com sobreposição ou não), proposta por (Timoszczuk, 1998).

3 Redes Neurais Recorrentes

As Redes Neurais Artificiais são modelos computacionais inspirados no cérebro humano e capazes de adquirir e manter o conhecimento a partir de exemplos de um determinado processo ou problema (Haykin, 2001).

Dentro desta classe de modelos, as redes neurais recorrentes se destacam por possuírem uma ou mais conexões de realimentação. Assim, as redes recorrentes usam resultados de instantes anteriores (finais ou intermediários) nos cálculos de sua próxima saída, o que possibilita o reconhecimento de padrões espaço-temporais, bem como a geração de sequências como saída.

Existem diversas configurações de redes recorrentes, sendo que as duas topologias a serem usa-

das neste trabalho são as redes recorrentes de Jordan e Elman, também conhecidas como redes parcialmente recorrentes (Fausett, 1994), pois grande parte de suas conexões são *feedforward* (da entrada para a saída) e apenas um grupo específico de neurônios recebe sinais de realimentação dos resultados da rede no instante anterior.

As redes de Jordan (Jordan, 1997), são parcialmente recorrentes e possuem a realimentação da saída da rede para as unidades de contexto. Sua arquitetura pode ser visualizada por meio da Figura 2:

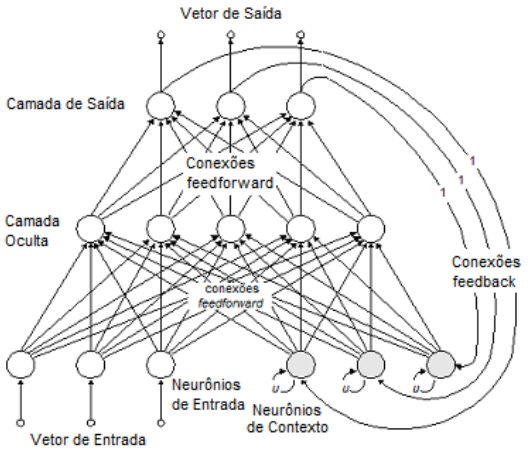

Figura 2: *Rede Recorrente de Jordan.*
Fonte: (Cabral Jr, 1999)

Alguns cuidados devem ser tomados no que diz respeito ao treinamento deste tipo de rede e sobre a condição inicial das unidades de contexto.

É importante ressaltar que a forma escolhida para inicializar a rede deve ser mantida durante a fase de utilização da mesma, garantindo que a sequência de saída gerada seja compatível com o treinamento.

O treinamento de uma Rede de Jordan faz uso do algoritmo de *Backpropagation*, o qual é adaptado para contemplar as realimentações e neurônios de contexto. Assim, é necessário definir as notações utilizadas no algoritmo de treinamento da rede, onde: T é o numero de entradas da rede, C é o número de neurônios de contexto; F corresponde ao numero de neurônios da primeira camada da rede, H da camada oculta e L da camada de saída; x é o vetor de entrada da rede e t é o vetor de saídas; w_ij são os pesos sinápticos do i-ésimo neurônio de entrada para o j-ésimo neurônio da camada oculta; w_jk são os peso da conexão do j-ésimo neurônio da camada oculta para o k-ésimo neurônio de saída; G(.) e θ correspondem respectivamente ao limiar e à função de ativação de cada neurônio; μ é um peso fixo de autoexcitação dos neurônios de contexto e η representa a taxa de aprendizado da rede no *Backpropagation*. O *Forward Step* do algoritmo de aprendizado da Rede de Jordan pode ser descrito pelas equações abaixo:

$$H = T + L \tag{4}$$
$$C = L \tag{5}$$
$$F = T + C \tag{6}$$

$$i_j^{oculta} = \sum_{i=1}^{F} x_i \cdot w_{ij} + \theta_j, 1 \leq j \leq H \tag{7}$$
$$O_j^{oculta} = G(i_j^{oculta}) \tag{8}$$
$$i_j^{saida} = \sum_{i=1}^{H} O_j^{oculta} \cdot w_{jk} + \theta_k, 1 \leq k \leq L \tag{9}$$
$$O_k^{saida} = G(i_k^{saida}) \tag{10}$$

E o *Backward Step* da Rede de Jordan é dado pelas seguintes equações:

Erro de Saída:
$$\epsilon_k = t_k - O_k^{saida}, 1 \leq k \leq L \tag{11}$$

Atualização da Camada de Saída:
$$\Delta w_k^{saida} = \epsilon_k \cdot G'(i_k^{saida}), 1 \leq k \leq L \tag{12}$$
$$w_{jk}^{novo} = w_{jk}^{velho} + \eta \cdot \Delta w_k^{saida} \cdot O_j^{oculta}, \quad 1 \leq k \leq L \; e \; 1 \leq j \leq H \tag{13}$$
$$\theta_k^{novo} = \theta_k^{velho} \cdot \eta \cdot \Delta w_k^{saida} \tag{14}$$

Atualização da Camada Oculta
$$w_j^{oculta} = \left(\sum_{k=1}^{L} \Delta w_k^{saida} \cdot w_{jk} \right) \cdot G'(i_j^{oculta}) \tag{15}$$
$$w_{ij}^{novo} = w_{ij}^{velho} + \eta \cdot \Delta w_j^{oculta} \cdot x_i, \quad 1 \leq i \leq Le1 \leq j \leq H \tag{16}$$
$$\theta_j^{novo} = \theta_j^{velho} \cdot \eta \cdot \Delta w_k^{oculta} \tag{17}$$

Atualização da Camada de Contexto
$$x_{k+T} = \mu \cdot x_{k+T} \cdot O_k^{saida}, 1 \leq k \leq L \tag{18}$$

As Redes de Elman (Elman, 1990) também são parcialmente recorrentes, onde as conexões de realimentação são feitas das saídas anteriores dos neurônios da camada oculta para as unidades de contexto sendo que, neste caso, cada neurônio da camada oculta transfere sua saída para apenas um neurônio de contexto. Sua arquitetura pode ser visualizada por meio da Figura 3:

Deve ser observado que esta arquitetura obriga que o número de neurônios de contexto seja igual ao número de neurônios da camada oculta, uma vez que é feita uma cópia um para um. O treinamento de uma Rede de Elman apresenta muitas semelhanças com o algoritmo de aprendizado da Rede de Jordan, com diferenças necessárias para contemplar as realimentações e neurônios de contexto. Assim, as equações 5 a 17 podem ser mantidas e a equação 18, referente à atualização da camada de contexto, é alterada para:

$$x_{k+T} = O_j^{oculta}, 1 \leq k \leq H \tag{19}$$

É importante ressaltar que, enquanto na Rede de Jordan a camada de contexto possui a mesma

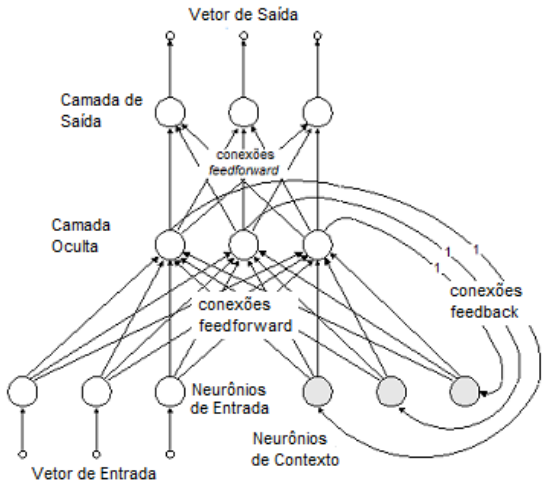

Figura 3: *Rede Recorrente de Elman.*
Fonte: (Cabral Jr, 1999)

dimensão da camada de saída, na Rede de Elman, a camada de contexto possui a mesma dimensão da camada oculta.

4 Arquitetura do RAL baseado em Redes Neurais Artificiais

Após a etapa de pré-processamento, cada sinal de voz de um dado locutor produz um conjunto de MTIs e faz-se necessário o uso de um mecanismo capaz de realizar a tarefa de classificação destes conjuntos, onde é aplicado o classificador neural. Assim, neste trabalho, o tal classificador é representado pelas redes de Elman e Jordan, treinadas por meio do *Backpropagation*. A entrada de cada classificador conterá um conjunto de MTIs (cada MTI possui uma quantidade fixa **P** de parâmetros, assim a RNA terá **P** entradas). O classificador também terá **m** saídas, onde cada uma representa um locutor, conforme Figura 4:

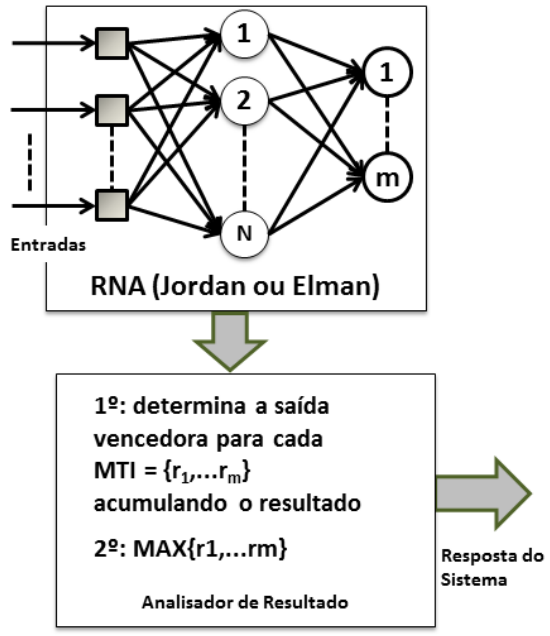

Figura 4: *Identificador de Locutor baseado em RNA*

O critério de decisão da rede é realizado em duas etapas, a saber:

1. Verifica-se qual das saídas apresentou o maior valor e, portanto, o locutor reconhecimento será o que representa esta saída *(winner takes all)*;

2. A partir de um *threshold* τ, quando o valor relativo de MTIs for superior a τ e este conjunto for único (representando um único locutor L), a saída do sistema irá considerar o reconhecimento bem sucedido e o locutor identificado será o indivíduo L. Caso contrário, a saída do sistema será "Locutor Indefinido".

5 Resultados Obtidos

O sistema foi construído de forma a analisar 10 locutores escolhidos da base de dados, onde 10 locuções (por locutor) foram separadas para o "treinamento"da RNA e 6 locuções (por locutor) foram separadas para validar o desempenho da RNA. O treinamento de referência do Perceptron de Múltiplas Camadas (Caropreso et al., 2013), que corresponde ao melhor resultado atingido pelo conjunto de treinamentos do sistema, apresentou a seguinte topologia: 55 neurônios na camada oculta e taxa de aprendizado $\eta = 0{,}015$.

Foram realizados treinamentos utilizando *TRUNC*, *TS* e *LTW* juntamente com a MTI31, onde o *TS* apresentou os melhores resultados em função da variação do *threshold* τ, na faixa [0,1; 1,0], conforme pode ser visto na Figura 5 .

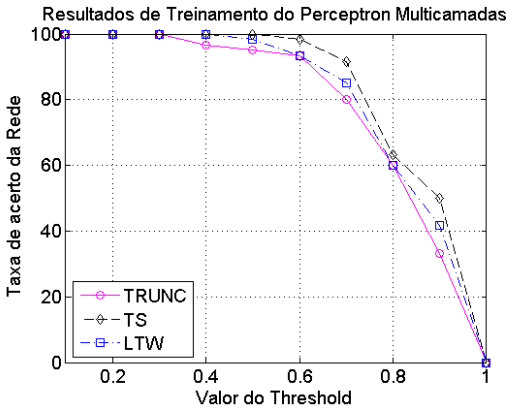

Figura 5: *Taxa de acerto do (PMC)*

Devido à característica dinâmica das redes recorrentes, as MTIs foram apresentadas por partes na camada de entrada da rede, ou seja, cada vetor componente é introduzido na rede, associando-se o respectivo valor de saída somente após o processamento do último vetor-componente, configurando assim a exibição de uma MTI ao sistema.

Os resultados obtidos podem ser visualizados por meio da Figura 6, onde é possível notar valores adequados para τ estão entre 0,5 e 0,6, onde mais de 80% do conjunto de MTIs foi classificada pela RNA como sendo pertencente a apenas um

locutor. Neste cenário, destaca-se a técnica de *TS* é que apresenta a melhor taxa de acerto da rede, superior a 90% para um *threshold* de até 0,55.

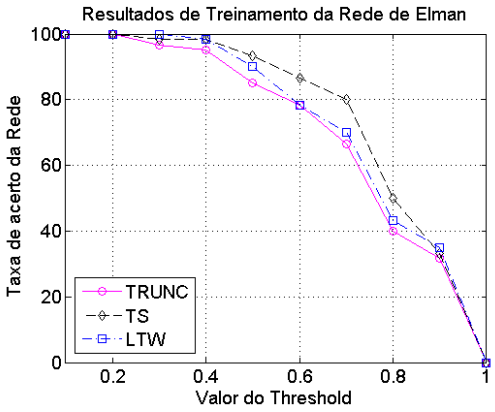

Figura 6: *Taxas de acerto da Rede de Elman*

Para a Rede de Jordan, os resultados podem ser analisados por meio da Figura 7, onde verificou-se que valores adequados para τ estão entre 0,5 e 0,6, sendo que mais de 80% do conjunto de MTIs foi classificado pela rede como pertencente a apenas um locutor. Neste cenário, a técnica de TS apresenta a melhor taxa de acerto da rede, a qual é superior a 90% para um *threshold* de até 0,5.

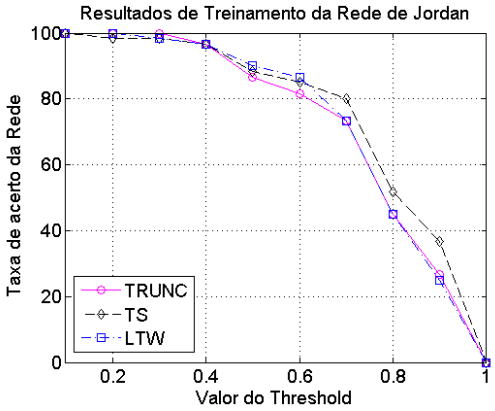

Figura 7: *Taxa de acerto da Rede de Elman (TS)*

Desta forma, tendo como referência de melhor resultado o desempenho para *TS*, é possível consolidar o resultado das 3 topologias estudadas na Figura 8 abaixo:

A partir dos resultados obtidos, nota-se que o desempenho das redes recorrentes para a tarefa de RAL, ficou equivalente ao Perceptron de Múltiplas Camadas nos mesmos cenários de estudo, para τ = até 0,5, com uma vantagem de 5% a 10% para o PMC.

No entanto, vale destacar que foi observada uma diferença relevante no processo de convergência do PMC na comparação com as redes recorrentes, onde o algoritmo do PMC convergiu em torno de 3000 épocas de treinamento enquanto as redes recorrentes convergiram com menos de 300 épocas, tornando-se evidente a influência da camada

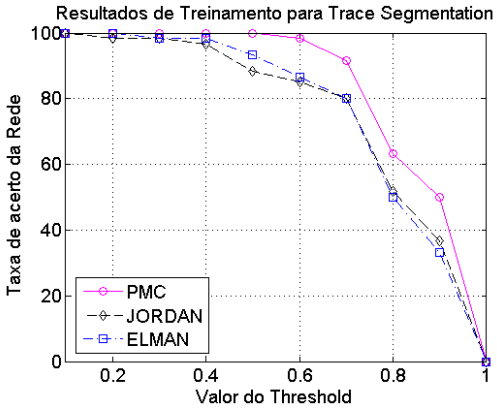

Figura 8: *Taxa de acerto das topologias de RNA (TS)*

de contexto na velocidade de aprendizado dessas redes ao se tratar de problemas de RAL.

Desta forma, ao se comparar a evolução do Erro Quadrático Médio (EQM) do PMC com os respectivos valores de EQM das redes recorrentes usando *TS*, foi possível observar os resultados apresentados nas Figuras 9 e 10, onde nesta última, o eixo das abscissas foi mantido em 250 épocas para as três topologias para permitir melhor visualização da progressão do erro ao longo dos treinamentos.

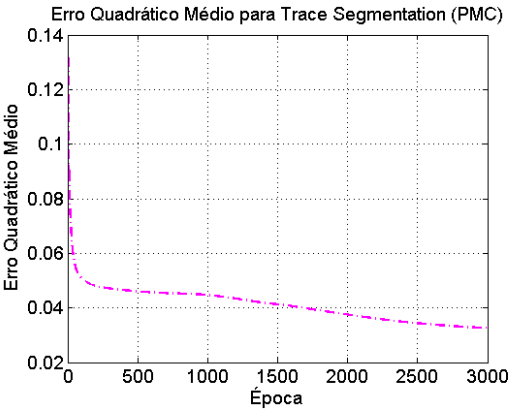

Figura 9: *EQM para o PMC no treinamento (TS)*

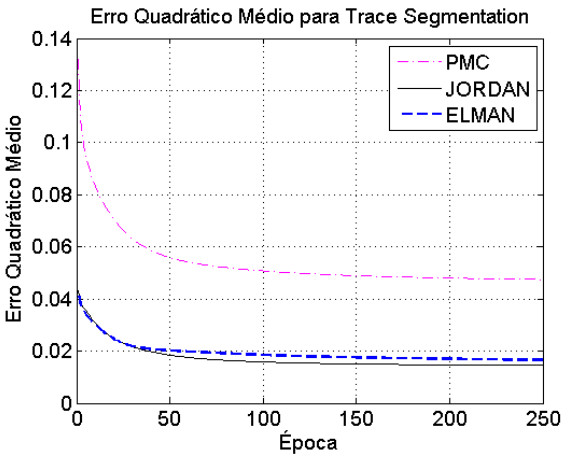

Figura 10: *EQM das redes em 250 épocas (TS)*

6 Conclusões

A partir dos testes realizados, fica evidenciada a viabilidade do uso de redes recorrentes no reconhecimento de locutor, onde se destaca a influência da camada de contexto destas redes no desempenho do RAL, cuja velocidade de aprendizado é muito superior à do PMC. Analisando-se os cenários de pré-processamento utilizados, foi possível confirmar os resultados obtidos para o estudo de referência, onde os melhores resultados foram obtidos para TS ao invés de *LTW*. Cabe ressaltar que a técnica *LTW* possui um desempenho quase idêntico ao *TS*, exceto para valores maiores de *threshold*, onde o *TS* se mantém a rede com uma taxa de acerto mais elevada (para τ=0,7 a diferença na taxa de acerto entre *TS* e *LTW* é cerca de 10% a mais para a primeira técnica). A utilização do *threshold* de decisão τ se mantém como boa opção para as análises de resultados, uma vez que permite a visualização do comportamento de cada RNA ao longo de uma faixa de valores, bem como o ajuste adicional na taxa de acerto do sistema, sem a necessidade de um novo treinamento. Como trabalhos futuros pode-se sugerir a análise de outras arquiteturas de RNA e treinamentos com o propósito de se verificar a influencia do valor da ponderação da camada de contexto e a variação da camada oculta na taxa de desempenho das redes recorrentes.

Agradecimentos

Os autores agradecem ao FIT - Flextronics Instituto de Tecnologia por seu apoio financeiro.

Referências

Bao, L. and Shen, X. (2016). Improved gaussian mixture model and application in speaker recognition, *2016 2nd International Conference on Control, Automation and Robotics (ICCAR)*, pp. 387–390.

Barbosa, F. G. and Silva, W. L. S. (2015). Support vector machines, mel-frequency cepstral coefficients and the discrete cosine transform applied on voice based biometric authentication, *SAI Intelligent Systems Conference (IntelliSys), 2015*, pp. 1032–1039.

Bhaykar, M., Yadav, J. and Rao, K. S. (2013). Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm, *Communications (NCC), 2013 National Conference on*, pp. 1–5.

Cabral Jr, E. (1999). Redes neurais artificiais: um curso teórico e prático para engenheiros e cientistas, *Editora da USP: São Paulo* .

Caropreso, R., Negri, L. H. B. L., Fernandes, R. A. S. and Silva, I. N. (2013). Análise da resposta de redes neurais do tipo mlp para reconhecimento de locutor considerando distintas ferramentas de pré-processamento., pp. 1–6.

Chu, W. and Chen, R. (2016). Speaker cluster-based speaker adaptive training for deep neural network acoustic modeling, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5295–5299.

Elman, J. L. (1990). Finding structure in time, *Cognitive Science* **14**(2): 179–211.

Fausett, L. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*, Prentice-Hall, Inc.

Hansen, J. H. L. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review, *IEEE Signal Processing Magazine* **32**(6): 74–99.

Hayet, D., Radia, A., Akila, D. and Tayeb, L. M. (2014). Session variability in automatic speaker verification, *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*, pp. 185–190.

Haykin, S. (2001). Redes neurais: princípios e prática, *Bookman* p. 900.

Jokic, I. D., Jokic, S. D., Delic, V. D. and Perie, Z. H. (2015). Mel-frequency cepstral coefficients as features for automatic speaker recognition, *Telecommunications Forum Telfor (TELFOR), 2015 23rd*, pp. 419–424.

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach, *Advances in psychology* **121**: 471–495.

S, Y. and S, V. (2015). Real time voice identification based gear control system in lmv using mfcc, *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*, pp. 1–7.

Singh, A. K., Singh, R. and Dwivedi, A. (2014). Mel frequency cepstral coefficients based text independent automatic speaker recognition using matlab, *Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on*, pp. 524–527.

Timoszczuk, A. P. (1998). *Reconhecimento automático do locutor com redes neurais artificiais do tipo Radial Basis Function (RBF) e Minimal Temporal Information (MTI).*, PhD thesis, Escola Politécnica da Universidade de São Paulo.

Togneri, R. and Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues, *IEEE Circuits and Systems Magazine* **11**(2): 23–61.

Weng, C., Yu, D., Watanabe, S. and Juang, B. H. F. (2014). Recurrent deep neural networks for robust speech recognition, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5532–5536.