Rede Neural Artificial Aplicada em um Reconhecimento Automático de Voz Independentemente do Locutor

Conference Paper · November 2007

DOI: 10.13140/RG.2.1.4119.8329

6 authors, including:

Luiz Eduardo da Silva
Universidade Federal de Alfenas
13 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Germano Lambert-Torres
PS Solutions, Itajuba, Brazil
762 PUBLICATIONS **3,438** CITATIONS
SEE PROFILE

VI Congress of Logic Applied to Technology
Santos / SP / BRAZIL - November 21-23, 2007
The Logic as Base in the Development and Application of New Technologies
LAPTEC 2007

Rede Neural Artificial Aplicada em um Reconhecimento Automático de Voz Independentemente do Locutor

Luiz Eduardo da SILVA[1], Germano LAMBERT-TORRES[2], Wagner S. VIEIRA[2], Ciro R. SANTOS[2], Rômulo A. CARMINATI[2] e Helga G. MARTINS[2]
[1]*Universidade Federal de Alfenas*
Gabriel Monteiro da Silva, 714 - 37.130-000, Alfenas – MG - Brasil
[2]*Universidade Federal de Itajubá - UNIFEI*
Av. BPS 1303 – Itajubá – 37500-000 – MG – Brasil

Abstract. Este artigo apresenta uma implementação de Redes Neurais Artificiais (RNA) para o reconhecimento de quatro comandos de voz isolados independentemente do locutor. A partir de um banco de dados com os comandos pronunciados por diferentes narradores, técnicas de processamento do sinal de voz foram aplicadas para que se pudesse trabalhar com estes sinais na RNA. Grande parte do trabalho e sua finalização, foram feitas em MatLab. O projeto desenvolvido foi de caráter bastante empírico, demandando testes, análises gráfica e numérica, para que se alcançasse o objetivo final.

Keywords. Reconhecimento de Voz, Redes Neurais, Linear Predictive Conding - LPC.

Introdução

A constante busca para aperfeiçoar e estreitar o relacionamento entre homens e máquinas, tornando-o cada vez mais natural, não é nenhuma novidade. Em vista disto, um *Reconhecimento de Voz* é um sistema que possibilita a manipulação mais fácil e prática de equipamentos providos da capacidade de compreender a fala humana. Graças aos avanços tecnológicos na área de hardware esta tarefa está cada vez mais possível de ser realizada.

Um esquema de Reconhecimento Automático de Voz - RAV consiste em reconhecer automaticamente a identidade de um indivíduo comparando características provenientes do sinal de sua voz [1, 2]. Já um Reconhecimento Automático de Voz Independentemente do Locutor – RAVIL, não se preocupa em classificar quem está falando, mas sim o que é falado. O reconhecimento de voz vem se desenvolvendo a níveis avançados de desempenho e sendo utilizados em aplicações reais devido à evolução dos sistemas computacionais. São várias suas aplicações atualmente, como: Serviços de *telemarketing*; acesso a funcionalidades de aparelhos de telefonia móvel; sistemas de segurança (RAV); Brinquedos (RAV/RAVIL); Sistemas embarcados automotivos (RAV/RAVIL); etc. [3]. A base de um RAV consiste na comparação de

padrões. Para tal é necessário representar o sinal da voz e definir suas características mais adequadas para serem utilizadas nos padrões.

Este artigo propõe a construção de uma interface para computadores que reconheça um vocabulário de palavras de forma isolada e independente do locutor. A interface é composta por comandos de direcionamento aplicados a robôs, utilizando técnicas de processamento de sinais e técnicas de Inteligência Artificial, mais especificadamente, Redes Neurais.

1. Sistemas de Reconhecimento Automático de Voz

Os sistemas RAVIL também são classificados [1] como *dependentes do texto*, de modo que o locutor deve pronunciar um texto pré-definido para o reconhecimento. Ou *independentes do texto*, sistemas que não necessitam de texto pré-definido para o reconhecimento. É necessário definir a representação do sinal de voz e as características utilizadas para formar os padrões do reconhecimento. A abordagem de reconhecimento de padrões se baseia em métodos estatísticos [4] consistindo de quatro passos:

- Codificação do sinal da voz – Utiliza técnicas espectrais como a Codificação Preditiva Linear – *Linear Predictive Coding* (LPC) ou a Transformada Rápida de Fourier (FFT);
- Fase de treinamento – Padrões de referência são criados e são formados por algum método que preserva as características estatísticas da classe;
- Fase de classificação – Comparação entre uma locução desconhecida com os padrões de referência, calcula-se a distância espectral;
- Seleção da classe da locução desconhecida.

2. Codificação Preditiva Linear – *Linear Predictive Coding LPC*

A Codificação Preditiva Linear – Linear Predictive Coding LPC é uma das mais poderosas técnicas de análise de voz e um dos métodos mais utilizados para codificação de voz [5].

LPC é usada para transmitir informações de espectros tolerando erros de transmissão dos coeficientes dos filtros. Para um erro muito pequeno o espectro pode ser distorcido por inteiro ou pior, pode causar uma instabilidade no filtro de predição. LPC é um método de predição em que a amostra de sinal de voz baseada em várias amostras anteriores. De acordo com [6] existem os seguintes métodos para a obtenção dos coeficientes de predição linear, entre outros: Método de Covariância; Método da Autocorrelação; Método Lattice; Método de Estimação Espectral;

3. Redes Neurais

As Redes Neurais Artificiais (RNA) vem sendo implementada com sucesso para diversificados trabalhos, que têm em comum a essência de reconhecimento otimizado de padrões, a saber: Reconhecimento de imagens (caracteres, impressões digitais); Reconhecimento de voz (o comando pronunciado); Reconhecimento de tendências

financeiras. Nestas aplicações, são tratados problemas tipicamente não linearmente separáveis, assim para contornar esta dificuldade, utilizam-se as Redes Neurais Artificiais, inspirada na malha neural humana, com múltiplas camadas garantindo boas aproximações e estimativas no reconhecimento de padrões. Uma etapa extremamente importante na implementação de uma RNA é o *treinamento*. O treinamento das redes neurais consiste em estabelecer os pesos a partir das funções de treinamento específicas. A definição de qual algoritmo de treinamento utilizar depende de vários fatores, tais como a complexidade do problema, quantidade de dados de treinamento e a precisão e exatidão esperadas. Um modelo de trabalho de RNA Multicamadas que garante bons resultados é o *Perceptron,* MLP - Multilayer Perceptrons:

- Proposto por Rumelhart (1986);
- Redes com duas ou mais camadas de neurônios do tipo do Perceptron;
- Algoritmo de treinamento "Backpropagation error".
- Neste treinamento as entradas e as correspondentes saídas são utilizadas no treinamento da rede calibrando uma função de saída correlacionada com os vetores de entrada. Isso é possível devido a retro-propagação do erro na saída da rede, onde são feitos os ajustes de pesos da rede quando necessário [7, 8].

4. Metodologia

4.1. Preparação das Amostras de Voz

As amostras de voz foram coletadas utilizando-se um microfone o qual capta somente o som que incide perpendicularmente à sua superfície, com mínima captação de sons circundante. A gravação dos sinais foi feita através do software Audacity (versão 1.3) à taxa de 16 bits, formato Mono.

De posse do sinal de voz bruto, partiu-se para a etapa de edição, utilizando o mesmo software. Primeiramente, realizou-se a detecção dos ruídos presentes no ambiente ao longo de toda a extensão do sinal. Em seguida fez-se a remoção dos ruídos. A percepção do ruído é desempenhada durante o período pré ou pós-narração, períodos de "silêncio", quando o sinal relevante não se manifesta. Então, removeram-se os referidos períodos, nos quais não havia a voz do narrador e que, portanto, não representaram interesse para a identificação do comando, conforme Figuras 2.

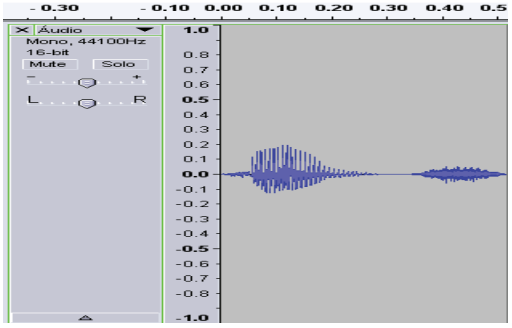

Figura 2 – Sinal de voz filtrado e editado.

O sinal editado foi exportado em formato “wave”, com taxa de 8 bits. Assim, conseguiu-se uma simplificação ainda maior da amostragem, devido à exigência do programa utilizado – MatLab. Este trabalha unicamente com arquivos em formato wave simplificado.

Assim, o formou-se um banco de dados constituído de 21 amostras de voz para treinamento supervisionado da rede neural, além de mais 7 amostras para posterior fase de testes. Dentre exemplares de vozes masculinas e femininas. Cada amostra deixa implícita a narração de quatro palavras, gravadas isoladamente: “back”, “front”, “left” e “right”.

4.2. Implementação e determinação do número ótimo de componentes LPC

As amostras de voz, em formato wave, foram importadas para o espaço de trabalho do MatLab. A partir disso, foi possível construir os gráficos discretizados das LPC dos sinais.

O objetivo, neste ponto, foi determinar o número de componentes LPC que forneceriam o melhor resultado, de forma a otimizar o trabalho da rede neural. Para tanto, imprimiu-se, em um único eixo cartesiano, os 28 gráficos discretizados referentes a cada comando de voz, totalizando quatro gráficos (“back”, “front”, “left” e “right”). Dessa forma, pode-se determinar o grau de coincidências entre pontos correspondentes da LPC, para um mesmo comando de voz, amostrado por diferentes locutores.

Para uma mesma abscissa, o maior nível de concentração dos referidos símbolos indica a ocorrência do maior grau de coincidência. Ou seja, pode-se analisar e estabelecer, empiricamente, o número ótimo de LPC que forneceria a maior relação de semelhança entre os mesmos comandos de voz, quando narrados por indivíduos distintos.

Paralelamente, deve-se, também, avaliar o grau de não-semelhança entre comandos diferentes. Esta análise foi feita empiricamente, a partir de gráficos que explicitam a curva de dispersão linear para cada comando de voz. Ficou consentido, então, que o número ótimo de componentes para cada LPC é 14.

4.3. Implementação da Rede Neural

Após a minuciosa análise efetuada a fim de se determinar os valores ótimos de LPC, trabalhou-se no MatLab com o modelo escolhido, através da ferramenta “nntool” provida pelo software. As amostras de LPC correspondentes a cada comando foram utilizadas como base de exemplo para o treinamento supervisionado da rede neural a ser implementada. Ao fim da fase de treinamento, a rede neural estabeleceu padrões de comparação para cada um dos quatro comandos.

A última etapa a ser cumprida é, finalmente, a bateria de testes para avaliar a capacidade da rede neural estruturada de reconhecer, apropriadamente, os comandos propostos de voz. Para tal, utilizou-se o banco de dados selecionado para experimentações, o qual foi composto de 7 amostras que, obviamente, não participaram do treinamento da rede neural. De posse do sinal correspondente a uma amostra de testes, inseriu-se na rede neural treinada, buscando um padrão de reconhecimento que possa identificá-lo como um dos quatro comandos em questão: “back”, “front”, “left” e “right”.

O procedimento foi repetido para todas as 7 amostras e, então, determinou-se a taxa de acertos e a taxa de precisão do sistema. Observou-se, também, que o treinamento da rede poderia, alternativamente, ser feito com vetores taxas de variação da LPC de comandos de voz. Esses vetores dizem como se comporta a variação ordenada entre cada componente de um vetor LPC, emitindo uma seqüência final com os seguintes valores: crescente ou decrescente. Sendo valor 1 para taxa de variação crescente e 0 para decrescente. Como os coeficientes nunca apresentavam o mesmo valor devido ao número elevado de casas decimais, a condição de taxa de variação nula não precisou ser considerada para este problema.

Suponha o seguinte vetor LPC: LPC = [**0.09** ; **1.2** ; **0.1** ; **0.02** ;**0.03** ; **2.5**]. Seu correspondente vetor de variação será: ΔLPC = [**1**; **0** ; **0**; **1** ; **1**].

A seqüência do vetor de variação representa as nuances da curva que pode ser esboçada ligando-se os componentes de um vetor LPC. Ainda, a partir de análises gráficas, verificou-se que as curvas de LPC seguiam uma mesma tendência de variação para mesmas palavras e se diferenciam pra com as de outras palavras. Desta forma o ΔLPC também constitui um parâmetro conveniente para se trabalhar na rede neural.

5- Configuração das Redes Neurais

O problema de reconhecimento de padrão de voz é caracterizado por um problema de análise descriminante não-linear, e para este são utilizadas as redes com configuração baseada em múltiplas. A configuração *backpropagation*, no MATLAB, possui duas vertentes, a *cascade-forward backpropagation* e *feed-forward backpropagation*.

Tanto a *cascade-forward* como *feed-forward* possuem N camadas que utilizam função de peso por produto escalar, função de entrada da rede, as específicas funções de transferência e a primeira camada possui pesos originados a partir da entrada. A diferença entre a *cascade* e a *feed* é a maneira de ajustar os pesos das camadas. Na *cascade-forward* cada camada subseqüente possui pesos ajustados tanto a partir da entrada como de todas as camadas anteriores. No entanto, a *feed-forward* ajusta os pesos apenas a partir das camadas anteriores. A análise dos resultados obtidos apresentará qual a influência desta diferença na convergência e precisão das redes neurais [9].

5.1 Funções de Transferência

Em múltiplas camadas as funções de transferência sigmoidais, *Logsig* e *Tansig*, são mais utilizadas, pelo fato de serem diferenciáveis.

A função *Logsig* gera saídas entre 0 e 1 a partir de entradas que variam de negativo a positivo infinito. Já a *Tansig* gera saídas entre -1 a 1. Como estabelecido, os valores das saídas variam entre 0 e 1, e a função de transferência proposta para o problema é a *Logsig*.

5.2 Funções de Treinamento

Existem diversas funções de treinamento disponíveis no MatLab, entretanto algumas delas apresentam melhores resultados para específicos problemas. De acordo com [9]

as funções trainrp (Resilient Backpropagation) e trainlm (Levenberg-Marquardt) geram redes neurais com uma melhor performance.

6. Treinamento das Redes Neurais

A partir dos coeficientes LPC e dos valores referentes às variações destes coeficientes, os quais são os parâmetros que compõe os vetores de entrada das redes neurais, a tabela de entrada das redes é criada.

É definida a tabela de saída utilizando como objetivo os quesitos de precisão e convergência das redes.

6.1 Tabela de Entrada

Composta por 84 vetores, visto que foram utilizados quatro amostras (palavras *back,front.left* e *right*) de 21 pessoas. Para a entrada de coeficientes LPC cada vetor é composto por 14 elementos. Valor este definido a partir de estudos dos números de coeficientes LPC. Para a entrada das variações dos coeficientes cada vetor é composto por 13 elementos, visto que a quantidade de variações é sempre uma unidade a menos do número total de coeficientes.

6.2 Tabela de Saída

O número de vetores da tabela de saída é o mesmo das tabelas de entrada, porém cada qual com 9 elementos "digitais", ou seja, podem assumir valores "0" ou "1". São definidos, os vetores, com 9 elementos cada para facilitar a análise de convergência e precisão das redes. Este valor não poderia ser muito baixo nem elevado, pois dificultaria as análises já mencionadas, porém deveria ser suficiente para distinguir os vetores de saída.

7. Testes e Resultados

Os testes foram realizados a partir de amostras de voz de pessoas que não participaram do treinamento das redes.

Foram analisados dois tipos de respostas, uma referente às redes neurais individuais e outra referente ao acoplamento de duas redes distintas. Cada rede neural pode apresentar como saída as palavras *back, front, left. right* e indefinido.

O acoplamento de duas redes neurais consiste em analisar as respostas das redes e determinar como saída uma resposta que seja mais conveniente. Esta determinação utiliza como parâmetro apenas a precisão das redes. Isto é, se a reposta de uma das redes foi a palavra *back* e da outra rede foi a palavra *right,* o acoplamento das redes determina que a resposta será *back* se a primeira rede tem uma porcentagem de acertos de palavra *back* maior do que a porcentagem da segunda rede em relação a palavra *right*, caso contrario a resposta é a palavra *right*.

Se uma das redes apresenta como resposta o indefinido e a outra qualquer uma das palavras, o acoplamento determina que a resposta será qualquer uma das palavras e

nunca o indefinido. A resposta será o indefinido somente quando as duas redes apresentar como resposta o indefinido. E quando a porcentagem de acerto de uma determinada palavra de certa rede for igual ao de outra palavra da outra rede, o acoplamento determina que a resposta será a palavra da rede que apresenta uma maior porcentagem no total (média aritmética das porcentagens das quatro palavras).

O treinamento das redes neurais, através das funções de treinamento *trainrp* e *trainlm* com configuração *backpropagation*, é realizado a partir de dois tipos de entradas, um dos tipos são os coeficientes LPC e o outro as taxas de variações destes coeficientes.

Vários testes foram realizados, desde através do acoplamento de redes de mesma função de treinamento e mesmo tipo de entrada até a partir de redes neurais simples.

Os resultados de maior desempenho são do acoplamento de uma rede com função de treinamento trainrp e outra com função trainlm, ambas com dados de entrada sendo os coeficientes LPC. Os resultados obtidos estão nas tabelas 3, 4 e 5.

Tabela 3 – Resultado parcial Rede Neural 1.

Rede Neural 1 – Função de Treinamento *Trainrp*			
Configuração		Resultados	
Função de Transferência	LOGSIG	Acertos *Back*	85,71%
Número de Camadas	5	Acertos *Front*	71,43%
Número de Neurônios - Camada 1	14	Acertos *Left*	42,85%
Número de Neurônios - Camada 2	18	Acertos *Right*	71,43%
Número de Neurônios - Camada 3, 4, 5	9	Resposta Indefinido	10,71%
Quantidade de Amostras no Teste	28	Total de Acertos	67,85%

Tabela 4 – Resultado parcial Rede Neural 2.

Rede Neural 2 – Função de Treinamento *Trainlm*			
Configuração		Resultados	
Função de Transferência	LOGSIG	Acertos *Back*	71,43%
Número de Camadas	4	Acertos *Front*	42,86%
Número de Neurônios - Camada 1	13	Acertos *Left*	71,43%
Número de Neurônios - Camada 2	18	Acertos *Right*	57,14%
Número de Neurônios - Camada 3 e 4	9	Resposta Indefinido	3,57%
Quantidade de Amostras no Teste	28	Total de Acertos	60,71%

Tabela 5 – Resultado Final Rede Neural 1 acoplada a Rede Neural 2.

Acoplamento - Rede Neural 1 com Rede Neural 2	
Acertos *Back*	100%
Acertos *Front*	85,71%
Acertos *Left*	57,14%
Acertos *Right*	71,43%
Resposta Indefinido	0%
Acertos Total	78,57%

8. Conclusão

Este artigo apresenta a potencialidade da utilização das Redes Neurais Artificiais para a resolução do problema de Reconhecimento Automático de Voz Independentemente do Locutor - RAVIL, uma vez que a RNA consegue estabelecer padrões comparativos para estes comandos. A obtenção de bons resultados não compete apenas às RNAs, mas também à confiabilidade dos vetores de entrada. Logo, o trabalho meticuloso do tratamento dos sinais, o qual teve por finalidade diminuir a taxa de amostragem de 16 bits para 8 bits, eliminar os ruídos e aplicar o LPC, foi de significativa importância para a análise do problema. No entanto, como intuito da proposta de RAVIL é reconhecer o comando de voz e não o locutor, as informações perdidas neste tratamento não influenciaram nos resultados obtidos. Com a observação de que a taxa de variação dos coeficientes LPC poderia apresentar certo padrão, além dos coeficientes LPC, os esforços também foram direcionados na análise deste parâmetro. E foi observado que a taxa de variação realmente trata-se de um padrão, porém as RNAs com vetores de entrada sendo os coeficientes LPC apresentaram um melhor desempenho. Conclui-se que os coeficientes LPC se apresentam mais como um padrão do que suas taxas de variações, mesmo assim estas são convenientes para se trabalhar com RNA. É de suma importância expressar que as configurações *cascade-forward backpropagation* e *feed-forward backpropagation* não tiveram diferenças significativas para este problema, pois apresentaram resultados bastante próximos, mesmo utilizando de maneiras diferentes para ajustar os pesos das camadas de neurônios. O acoplamento das redes neurais individuais foi um artifício que melhorou de forma expressiva o desempenho. O melhor acoplamento apresentou 0% de respostas indefinidas e obteve porcentagens de acertos maiores do que as redes individuais. É importante citar que os melhores acoplamentos não surgiram de RNAs com bons resultados para todos os comandos, mas sim de RNAs que eram especialistas nos comandos os quais a outra rede não era, e vice-versa. Estas redes foram treinadas por funções de treinamento diferentes, uma *trainrp* e outra *trainlm*, porém sendo os mesmos vetores de entrada, os coeficientes LPC.

Referências

[1] Timoszczuk, A. P., *"Reconhecimento Automático do Locutor com Redes Neurais Pulsadas"*, Tese de Doutorado, Escola Politécnica da Universidade de São Paulo, São Paulo, 2004.

[2] Furui, S., *"Digital Speech Processing, Synthesis, and Recognition."*, New York, Marcel Dekker, 1989.

[3] Rabiner, L. R., *"Applications of Voice Processing to Telecommunications"*, Proceedings of the IEEE, v. 82, n. 2, p 197-228, Feb. 1999.

[4] Rabiner, L. and Juan, B. H., *"Fundamentals of Speech Recognition"*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[5] Schafer, R. W. and Rabiner, L. R., "*Digital Representations of Speech Signals*", Proceedings of the IEEE, v.63, n. 4, p662-677, April, 1975.

[6] Bezerra, M. R., "Reconhecimento Automático de Locutor para fins Forenses, Utilizando Técnicas de Redes Neurais", Dissertação Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, 1994.

[7] Demuth, H., Beale, M., and Hagan, M., "*Neural Network Toolbox 5 – User's Guide. Matlab"*, http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf, 05/08/2007.

[8] Perez,C. Ricardo. Ferramenta Para Reconhecimento de Padrões de Defeitos em Linhas de Transmissão. Projeto de Iniciação Científica. CQE. Abril, 2007.

[9] Demuth, Haward. Beale, Mark. Hagan, Martin. Neural Network Toolbox 5 – User's Guide. Matlab. Encontrado em http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf. Acessado em 05 de agosto de 2007 ás 15:00hs.