Anais do IX Congresso Brasileiro de Redes Neurais / Inteligência Computacional (IX CBRN)
Ouro Preto 25-28 de Outubro de 2009

RECONHECIMENTO DE COMANDO DE VOZ BASEADO EM FILTROS WAVELET UTILIZANDO REDES NEURAIS ARTIFICIAIS

IGOR S. PERETTA, GERSON F. M. LIMA, JOSIMEIRE A. TAVARES, KEIJI YAMANAKA

Núcleo de Engenharia de Computação, Faculdade de Engenharia Elétrica,
Universidade Federal de Uberlândia – Caixa Postal 593, 38400-902 Uberlândia, MG, BRASIL

E-mails: iperetta@gmail.com, gersonlima@ieee.org, josycbelo@gmail.com, keiji@ufu.br

Abstract— The use of voice commands as a new way of interaction between man and machine is the subject of several researches in recent years and has produced several commercial or freeware applications. However, considering the results achieved, there is still great development potential in this area. This work proposes the use of wavelet transform and wavelet packet filter bank as a main tool for feature extraction to feed a multi-layer artificial neural network to recognize a limited vocabulary of voice commands with low sensitivity to noise, independence for the pronunciation used and the identity of the speaker. A graphical interface was implemented to make the commands recognized by the artificial neural network able to control the movement and the color of an object inside a bi-dimensional virtual environment.

Keywords— Voice command recognition, DWT, wavelet packet filter bank, artificial neural network

Resumo— O uso de comandos de voz como nova forma de interação homem-máquina é alvo de diversas pesquisas nos últimos anos e já produziu diversos aplicativos comerciais ou de distribuição gratuita. No entanto, considerando os resultados conquistados, ainda existem grandes possibilidades de desenvolvimento nesta área. Este trabalho propõe o uso de transformadas *wavelet* e banco de filtros *wavelet packet* como ferramenta principal de extração de características para alimentar uma rede neural artificial multicamada para o reconhecimento de um vocabulário limitado de comandos de voz com baixa sensibilidade a ruídos e com independência da pronúncia utilizada e da identidade do locutor. Uma interface gráfica foi implementada para que os comandos reconhecidos pela rede neural artificial possam controlar o movimento e a cor de um objeto em um ambiente virtual bidimensional.

Palavras-chave— Reconhecimento de comando de voz, DWT, banco de filtros *wavelet packet*, rede neural artificial

1 Introdução

O reconhecimento da fala e de comandos de voz é uma extensa área de estudos com diversas aplicações possíveis no nosso cotidiano, como: tornar mais simples diversas tarefas diárias, possibilitar novas formas de interação homem-máquina, gerar controles inovadores para o desenvolvimento de realidade expandida, apoiar a inclusão de deficientes físicos com restrições severas de movimento. No entanto, mesmo após anos de pesquisa, não temos conhecimento de nenhum aplicativo que tenha obtido mais de 75% de precisão para locutores independentes [1]. Diversas estratégias já foram utilizadas, cada uma com suas vantagens e desvantagens [2] [3] [4] [5] [5] [6] [7].

Alguns fatores que impedem o sucesso absoluto nesta área são: possíveis indeterminações da qualidade dos equipamentos que serão utilizados na captura da voz; os diferentes níveis de ruído aos quais os aplicativos sempre estarão sujeitos [8]; as diferenças inerentes de cada locutor independente; a própria variação da fala de um mesmo locutor em situações distintas, por motivo de doença ou fadiga, ou mesmo pelo chamado efeito Lombard [9] (fenômeno em que o indivíduo altera sua produção vocal em ambientes ruidosos); a não compreensão da totalidade dos processos biológicos e cognitivos utilizados na audição humana.

Este trabalho parte de duas informações principais: a primeira é que as redes neurais artificiais têm obtido vários sucessos em reconhecimento de padrões de fala [2] [4] [10]. A segunda deve-se a uma característica da audição humana: a análise de sons feita pelos nossos ouvidos, ao menos em seu primeiro estágio determinado pela função resposta da cóclea humana, pode ser representada por transformadas *wavelet* [11]. O uso de funções *wavelet* foi indicado para aumentar a robustez a ruídos, emulando assim a resolução de freqüência da cóclea humana [2].

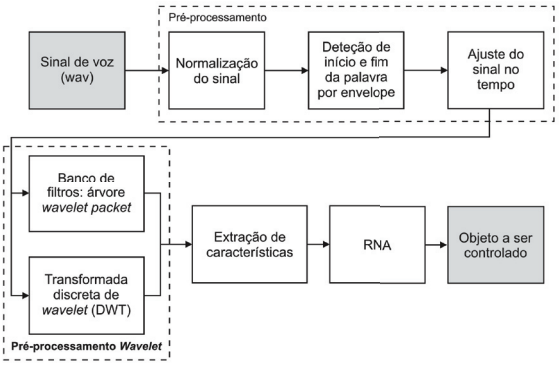

Figura 1. Diagrama funcional.

Assim, baseando-se em um banco de filtros do tipo *wavelet packet* com bandas de passagem inspiradas na audição humana [2] e na transformada discreta de *wavelet* (DWT) [11], foi desenvolvido um programa em MATLAB® para o reconhecimento de um vocabulário limitado a seis comandos de voz. Foram escolhidas as seguintes palavras: SOBE, DESCE, ESQUERDA, DIREITA, AZUL e VERMELHO.

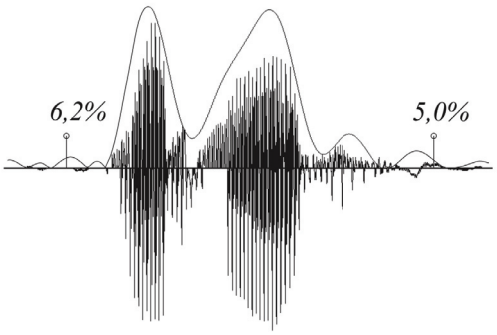

Figura 2. Sobreposição de um sinal de voz e seu envelope.

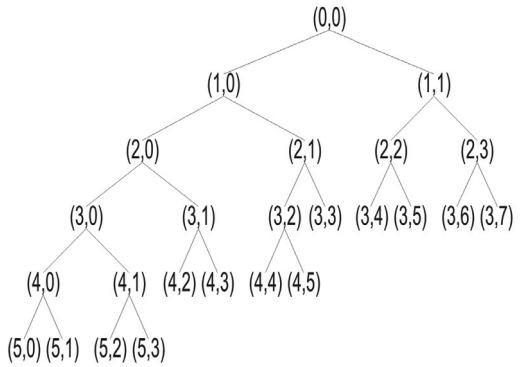

Figura 3. Nodos da árvore de decomposição da *wavelet packet*.

Com o uso do aplicativo Audacity® [12] de distribuição gratuita, foram capturadas três versões de cada uma dessas palavras faladas por 50 pessoas, sendo 30 homens e 20 mulheres na faixa etária de 17 a 40 anos, totalizando 900 amostras gravadas no formato de áudio *Waveform* (WAV). O programa foi configurado com freqüência de amostragem de 8kHz e comprimento de 16 bits por amostra de amplitude do sinal. As gravações foram realizadas com o uso de um microfone de computador comum, com impedância de 75Ω, em uma sala com fluxo regular de pessoas sem nenhum controle de nível de ruído. O objetivo foi desenvolver um aplicativo robusto a ruídos e com grande capacidade de generalização.

Estes dados foram processados pelas transformadas wavelet escolhidas e tiveram suas características extraídas para, agrupadas em vetores de entrada, possibilitarem o treinamento da rede neural artificial e o posterior reconhecimento dos comandos, como apresentado no diagrama funcional na Figura 1.

Para o reconhecimento dos comandos, foi utilizada uma RNA com arquitetura multicamada [13], treinada com 600 padrões de entrada. Os outros 300 padrões obtidos foram utilizados para verificar a capacidade de reconhecimento e generalização da rede.

Foi desenvolvida uma interface gráfica para que os comandos reconhecidos pudessem ser utilizados para o controle do movimento e alteração de características de um objeto em um ambiente virtual bidimensional (2-D).

Os resultados obtidos são apresentados no presente trabalho.

2 Desenvolvimento

2.1 Pré-Processamento

Este é um dos mais importantes estágios em qualquer aplicação destinada a processamento de sinais. Seu objetivo é tratar o sinal de maneira a minimizar as variações apresentadas, padronizando as amostras obtidas.

No presente caso de reconhecimento de comandos de voz, foram realizadas algumas etapas de padronização. Na primeira etapa, são realizadas a normalização e a maximização dos sinais, de modo a enquadrá-los dentro do espectro de amplitude definido pelo intervalo entre -1 e 1.

A próxima etapa compreende a detecção de início e fim das palavras inseridas nos sinais capturados. Para isso, foi utilizado inicialmente um algoritmo de detecção de envelope que utiliza a transformada rápida de Fourier do sinal e um filtro passa-baixa com 4 Hz de freqüência de corte [2]. Sobrepondo a forma de onda desse envelope ao sinal normalizado, detectamos o início da palavra a 6,2% da amplitude máxima do envelope e o seu final a 5% da mesma amplitude máxima. Esses valores percentuais foram obtidos experimentalmente para as amostras capturadas e escolhidos por atingir os melhores resultados. Com o início e o fim da palavra, é isolado seu sinal equivalente.

Na última etapa deste estágio, é realizado o ajuste no tempo do sinal isolado. Para minimizar o impacto de diferentes velocidades de fala, foram enquadrados todos os sinais isolados dentro do período de 0,5 segundo (ou 4.000 amostras de amplitude do sinal),

Tabela 1. Banda de passagem do banco de filtros por nodo da árvore da *wavelet packet*.

Banda [Hz]	*Nodo da árvore wavelet packet*
0-125	(5,0)
125-250	(5,1)
250-375	(5,2)
375-500	(5,3)
500-750	(4,2)
750-1000	(4,3)
1000-1250	(4,4)
1250-1500	(4,5)
1500-2000	(3,3)
2000-2500	(3,4)
2500-3000	(3,5)
3000-3500	(3,6)
3500-4000	(3,7)

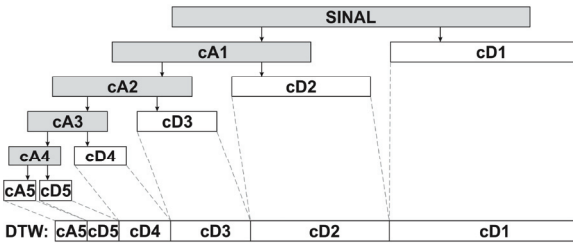

Figura 4. Decomposição do sinal pela DWT.

escolhido arbitrariamente. Com esse objetivo, foi desenvolvida uma função no MATLAB® para realizar o ajuste linearmente.

2.2 Pré-Processamento Wavelet

Este estágio é a principal contribuição deste trabalho. Com referência nos resultados de trabalhos que reforçam a utilização de wavelets em reconhecimento de padrões de voz [2] [3] [7] [10], foram utilizados o banco de filtros wavelet packet e a DWT para transformar os sinais obtidos até aqui, criando estratégias para aquisição de novas informações contidas nesses sinais.

São dois processos independentes utilizados para analisar o sinal. O primeiro consiste em submetê-lo a um banco de filtros *wavelet packet* para a geração da árvore de decomposição apresentada na Figura 3. Os coeficientes encontrados em cada nodo desta árvore correspondem às respostas no tempo do sinal filtrado com diferentes faixas de freqüência. As faixas de freqüências apresentadas na Tabela 1 foram concebidas para simular o sistema de audição humana [2].

O outro processo refere-se à aplicação da DWT ao sinal. Esta transformação decompõe o mesmo em coeficientes de aproximação (cA_i, com i = nível da *wavelet*) e coeficientes de detalhamento (cD_j, com j = 1, ..., nível escolhido da *wavelet*), conforme pode ser visto na Figura 4. A cada nível da *wavelet*, os números de coeficientes de aproximação e de detalhamento são reduzidos pela metade.

A *wavelet*-mãe escolhida para ambos os processos é a *Daubechies* 4 (`db4`) com nível 5 de decomposição.

2.3 Extração de Características

A partir dos estágios anteriores, são extraídos quatro grupos de características:

- Quantidade de picos no envelope detectado.
- Energia nos nodos da árvore da decomposição *wavelet packet*.
- Entropias de intervalos do sinal.
- Intensidades correlacionadas da DWT do sinal.

A quantidade de picos (máximos da função) encontrados no envelope é a primeira característica extraída. Como não houve controle algum da pronúncia dos locutores independentes amostrados, não foi verificado como realidade o fato de que a quantidade de picos do envelope equivaleria à quantidade de sílabas na palavra. O número de picos no envelope é o primeiro elemento do vetor de características.

Tabela 2. Intervalos definidos para cálculo de intensidades.

Decomposição do sinal	*Proporção com relação ao tamanho do sinal*	*Número de intervalos*
D1	2^{-1}	64
D2	2^{-2}	32
D3	2^{-3}	16
D4	2^{-4}	8
D5	2^{-5}	4
A5	2^{-5}	4

Os coeficientes encontrados na *wavelet packet* foram base para o cálculo da energia em cada nodo da árvore de decomposição. Esse grupo de características gera 13 novos elementos para o vetor de entrada.

O sinal ajustado da palavra é arbitrariamente dividido em 16 intervalos de 2^{-5} segundos para o cálculo das entropias de Shannon de cada intervalo. Os valores de entropia encontrados são 16 novos elementos para o vetor de entrada.

Os coeficientes da DWT do sinal foram divididos nos grupos cA_5, cD_5, cD_4, cD_3, cD_2 e cD_1. Cada grupo é dividido em intervalos definidos arbitrariamente, como pode ser visto na Tabela 2, e em cada intervalo é calculada a sua intensidade I, como na equação (1).

$$I = \frac{E}{A \cdot \Delta t} \qquad (1)$$

Onde: E é a energia no intervalo do sinal, A é a área da curva no intervalo do sinal e Δt é o intervalo de tempo considerado.

Como diferentes níveis de ruído podem alterar os cálculos de intensidade de determinado sinal, foi decidido correlacionar essas intensidades, referenciando-as sempre ao máximo de seu intervalo. Como teremos tantas intensidades correlacionadas quanto intervalos, são mais 128 elementos ao vetor de características.

Após a extração 158 características de todos os padrões de treinamento, a matriz de características é normalizada, enquadrando os seus elementos dentro do intervalo de -1 a 1. Devido à função de ativação dos neurônios da RNA escolhida, esse intervalo permite o melhor aprendizado da rede.

2.4 Rede Neural Artificial

Foi utilizada uma RNA com arquitetura multicamada (MLP), com 158 neurônios de entrada, 158 neurônios na camada oculta, e 6 neurônios de saída. A função de ativação dos neurônios escolhida foi a tangente

hiperbólica.

A RNA foi treinada com 600 padrões de entrada, cada qual com seu vetor de 158 características. Os outros 300 padrões gravados foram utilizados posteriormente para verificação de capacidade de reconhecimento da rede.

Os neurônios de saída equivalem cada qual a um comando a ser reconhecido. Os critérios utilizados para reconhecer um comando são:

- Se nenhum ou mais de um neurônio de saída é ativado, comando não reconhecido;
- Se apenas um neurônio é ativado, sua saída é comparada com um valor de confiabilidade; se for maior, o comando é reconhecido;
- Se a ativação é única, mas não atinge o valor mínimo de confiabilidade, o comando também não é reconhecido.

Para os resultados de testes apresentados neste trabalho, foram utilizadas três diferentes estratégias:

- Confiabilidade igual a 10%;
- Confiabilidade igual a 70%;
- Saída com maior valor positivo de ativação, aceitando mais de uma saída ativa.

2.5 A interface gráfica

Foi implementada uma interface gráfica com um ambiente bidimensional (2-D) para visualização dos resultados. Com base nos comandos reconhecidos pela RNA, um objeto é controlado ou alterado nesse ambiente. Os autores optaram por esse tipo de implementação por ser uma forma visual intuitiva que reflete os resultados de reconhecimento da RNA de maneira a minimizar o processo cognitivo do usuário.

3 Resultados e Conclusão

O primeiro conjunto de testes é relacionado com os arquivos gravados do banco de dados do projeto. Para treinar a RNA, 2/3 dos sinais gravados foram utilizados (2 de 3 versões de cada comando de voz do mesmo locutor). Foram utilizados 600 padrões para testar a capacidade de aprendizagem da RNA e o restante, 300 padrões, para testar sua robustez.

Todos os testes foram realizados com as estratégias de 10% de confiabilidade, 70% de confiabilidade e considerando o maior valor de ativação.

Os dados obtidos nos testes de aprendizagem e de robustez foram unificados, considerando, portanto, o total de 900 padrões. Dentre os resultados encontrados, o que mais se destacou foi o total de acertos no reconhecimento de arquivos gravados, em que se obteve uma abrangência entre 69% a 84% do total de padrões reconhecidos (90% a 99% de reconhecimento, quando considerado apenas o resultado dos testes de aprendizagem, dependendo da palavra).

Entre as diferentes estratégias de reconhecimento de comandos, obteve-se melhor resultado de acerto para a estratégia de maior valor (75% a 84%). No entanto, apresentou maior percentual de erros (5% a 14%). Com a confiabilidade de 70%, temos uma redução do percentual de acertos (69% a 79%) e, conseqüentemente, o menor percentual de erros (3% a 7%). Esses comparativos podem ser vistos nos Gráficos 1 e 2.

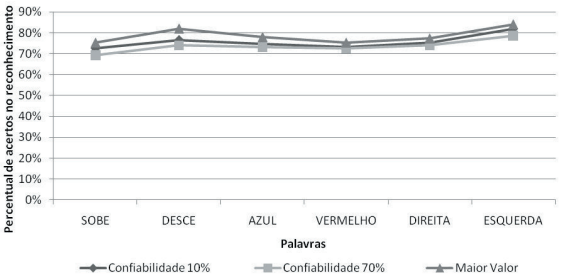

Gráfico 1. Comparativo entre acertos de reconhecimento em padrões gravados.

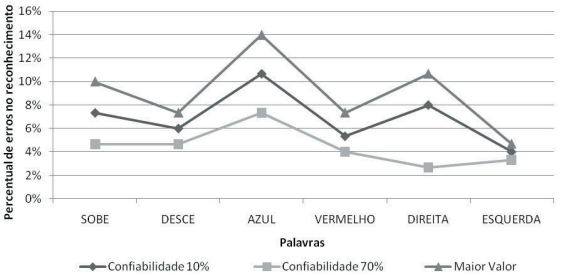

Gráfico 2. Comparativo entre erros de reconhecimento em padrões gravados.

O segundo conjunto de testes realizados tratou o reconhecimento em tempo real, com o auxílio de 5 locutores, em um ambiente sem controle de ruído. Dos locutores convidados, três já haviam participado da etapa de treinamento da RNA e dois locutores não, apresentando vozes inéditas. Cada locutor repetiu dez vezes cada palavra apresentada. O resultado, registrado no Gráfico 3, reflete a média de reconhecimento de todos esses locutores.

O conceito de "insucesso" utilizado significa que a RNA não pode identificar o comando no sinal analisado. Assim, o sistema declara que não pode reconhecer a palavra capturada como um comando válido. Quando a RNA não chegar a uma conclusão correta, é preferível decidir por uma resposta inconclusiva.

O desempenho da RNA gerou resultados promissores, comparáveis aos apresentados nos estudos referenciados nesse trabalho, embora tenha sido realizado em um ambiente não controlado para a coleta das amostras, sem controle de ruídos e nem regras de entonação para os locutores independentes.

Um dos resultados principais de ambos os conjuntos de testes (sinais gravados e tempo real) foi a baixa taxa de reconhecimento do comando "SOBE". Isso pode ser entendido ao verificar que o mesmo apresenta problemas no pré-processamento, no momento da detecção do início da palavra. Uma vez que a primeira

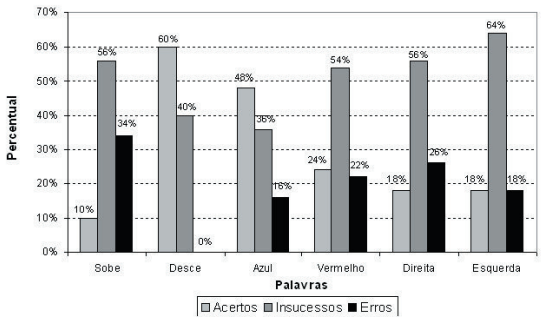

Gráfico 3. Comparativo dos resultados em tempo real.

sílaba possui som fricativo, o início da palavra tende a se confundir com o ruído ambiente, dificultando seu reconhecimento.

As diferenças apresentadas nas outras taxas de reconhecimento de ambos os conjuntos de testes podem ser entendidas por problemas no momento da captura e do pré-processamento dos sinais.

Para o prosseguimento deste trabalho, pretende-se:

- Aumentar o banco de vozes de locutores independentes, gerando melhorias na capacidade de generalização da RNA.
- Melhorar os algoritmos utilizados no pré-processamento dos sinais.
- Alterar configurações da RNA para ampliar sua capacidade de generalização (número de ciclos de treinamento e taxa de aprendizagem).
- Pesquisar ou desenvolver uma nova wavelet mãe que possa se aproximar dos processos utilizados pela audição humana.
- Estudar outras arquiteturas de RNA, híbridas ou não, buscando agilizar o treinamento sem perder a capacidade de generalização.

Agradecimentos

Aos Professores Doutores Edna Lúcia Flôres, Gilberto Arantes Carrijo, José Wilson Resende e Luciano Vieira Lima que, de forma direta ou indireta, colaboraram para a realização deste trabalho. Agradecemos também aos alunos da Universidade Federal de Uberlândia (UFU) que contribuíram solicitamente na coleta de amostras e também aos Professores Doutores Edgard Lamounier e Alexandre Cardoso pela cessão do espaço de trabalho e ambiente de testes.

Referências Bibliográficas

[1] Uma Maheswari, N., Kabilan, A.P., Venkatesh, R. (2008). Speaker independent speech system based on phoneme identification, *Computing, Communication and Networking*. pp. 1-6

[2] Gandhiraj, R., Sathidevi, P.S. (2007). Auditory-Based Wavelet Packet Filterbank for Speech Recognition Using Neural Network, *Advanced Computing and Communications*, pp. 666-673.

[3] Chen, S. H., Wu, H. T., Chen C. H., Ruan, J. C., Truong, T.K. (2005). Robust voice activity detection algorithm based on the perceptual wavelet packet transform, *Intelligent Signal Processing and Communication Systems*, pp. 45-48.

[4] Omar, M., Abdalgader, K. (2004). Designing A Neural Network Based Audio Classification System, *Masters Thesis*, Universiti Utara Malaysia.

[5] Sun, H., Ma, B., Li, H. (2008). An Efficient Feature Selection Method for Speaker Recognition, *Chinese Spoken Language Processing, ISCSLP '08*, pp. 1-4.

[6] Dakkak, O., Harba, Y. (2006). Vocal Commands to a Robot by an Isolated Words Recognition System using HMM, *Information and Communication Technologies ICTTA '06*, 2nd Volume 1, pp. 1219-1224.

[7] Siafarikas, M., Mporas, I., Ganchev, T., Fakotakis, N. (2008). Speech Recognition using Wavelet Packet Features, *Journal of Wavelet Theory and Applications*, Volume 2, Number 1, pp. 41-59.

[8] Ming, J. (2006). Noise compensation for speech recognition with arbitrary additive noise, *Audio, Speech, and Language Processing, IEEE Transactions*, Volume 14, Issue 3, pp. 833-844.

[9] Lau, P. (2008). The Lombard Effect as a Communicative Phenomenon, *UC Berkeley Phonology Lab Annual Report '08*.

[10] Crovato, C. D. P., Schuck Jr, A. (2007). The Use of Wavelet Packet Transform and Artificials Neural Networks in Analysis and Classification of Dysphonic Voices, *IEEE Transactions on Biomedical Engineering*, v. 54, pp. 1898-1900.

[11] Daubechies, I. (1992), *Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics*, SIAM Ed.

[12] http://audacity.sourceforge.net/?lang=pt

[13] Fausett, L. (1994) *Fundamentals of neural networks*, Prentice-Hall.