APLICAÇÃO DA REDE NEURAL DE KOHONEN NA IDENTIFICAÇÃO DE LOCUTOR

Gustavo Oliani David [1]; Thiago Antonio Grandi de Tolosa [2]

[1] Aluno de Iniciação Científica da Escola de Engenharia Mauá (EEM/CEUN-IMT);
[2] Professor da Escola de Engenharia Mauá (EEM/CEUN-IMT).

Resumo. *Neste trabalho é desenvolvida uma metodologia baseada na aplicação da rede neural de Kohonen para identificação de voz, dentre as várias aplicações possíveis em um sistema biométrico. Devido ao recente avanço tecnológico na área de reconhecimento de padrões, pretende-se que a solução do problema proposto seja realizada com a precisão necessária para atender sistemas de segurança de áudio ou apenas para identificar o locutor em um ambiente com várias fontes sonoras. Além da aplicação de um modelo baseado em redes neurais artificiais, a certificação sobre a identidade do locutor exige alguns desafios que serão investigados no trabalho, como por exemplo: como a voz de uma pessoa é associada a ela; como uma rede neural com treinamento não supervisionado realiza a associação da voz a uma pessoa e sua eficiência neste processo. Portanto, pretende-se com este projeto investigar a eficiência da utilização da rede neural de Kohonen para identificação do locutor, após ter sido treinada para tal. Na etapa de treinamento, a rede neural foi exposta à sinais de áudio de diferentes locutores em diferentes ambientes considerando intervalos de tempo iguais.*

Palavras-chave: Rede neural artificial de Kohonen, identificação de voz, Mel frequency cepstral coefficients.

Introdução

Uma rede neural artificial pode ser caracterizada como uma imitação de como o cérebro interpreta e processa informações. Assim como o cérebro usa a experiência adquirida para se desenvolver, a rede neural recebe informações do ambiente (entradas da rede) e determina pesos para criar a ligação do neurônio e a informação recebida. Portanto, cada neurônio ficará responsável por reconhecer aquela informação ou situações semelhantes no futuro.

A rede neural artificial, que funciona como um modelo de um sistema nervoso central aprende através das informações captadas do ambiente e se utiliza das forças das conexões entre seus neurônios, chamadas de "pesos sinápticos", para armazenar as informações por regularidade.

Alguns projetos envolvendo o tema deste trabalho foram realizados nos últimos anos utilizando metodologias diferentes para solução do problema. Destaca-se nas referências bibliográficas apresentadas o trabalho de Mafra (2002) que serviu para dar uma melhor direcionada neste projeto, por fazer uma pesquisa geral sobre algumas estruturas de redes neurais, algumas formas de extração de voz e por fim seu processo experimental, demonstrando resultados interessantes e tomando algumas conclusões bem esclarecedoras a respeito do assunto, por exemplo, sobre tempo de treinamento, alguns parâmetros do MFCC e quantidade de amostras.

Pode-se citar um trecho importante no desenvolvimento do trabalho de Mafra: "Com o detalhamento deste erro por frase, observou-se que as frases de teste mais curtas foram responsáveis pelos erros, com taxas de acerto de 100% para as frases de teste mais longas. Isto definiu um limite inferior para a duração das frases de teste de aproximadamente 2,6s.

De forma geral, pode-se dizer que o conjunto treinado com aproximadamente 17,5s de amostras de voz por locutor, é capaz de identificá-los com mais de 99% de taxa de acerto quando testado com locuções de duração superior a 2,8s, em modo independente de texto,

colocando-o muito próximo aos sistemas estado da arte na categoria. Os resultados dos testes indicam que este desempenho pode ser ainda melhorado pelo aumento do número de unidades das SOMs."

Mapas auto organizáveis de Kohonen (Redes Self Organizing Maps - SOM)

Em uma entrada com várias informações, como duas pessoas conversando, a rede detecta essas informações e procura por padrões nelas - por exemplo, quando a pessoa A está falando, ou quando a pessoa B está falando - enquanto designa um neurônio disponível para representar esse padrão no mapa. Em uma rede com dois neurônios, um deles ficaria responsável por identificar A e o outro por identificar B. Ao final do treinamento da rede, têm-se as informações separadas por padrões e distribuídas pelo mapa, ao mesmo tempo em que se separam essas informações em conjuntos semelhantes denominados "clusters". O que permite que, quando a rede "ouvir" a pessoa A falando novamente, seja capaz de identificar esta pessoa.

A rede SOM é assim chamada, pois, de acordo com o que for inserido na rede, ela irá se organizar sozinha para reconhecer os padrões e dar as saídas separadas. Isso ocorre devido ao processo de aprendizagem pelo qual ela passa, que pode ser resumido em três processos, competição, cooperação e adaptação sináptica, explicados abaixo, conforme Haykin (2001).

Competição: Nessa etapa é determinado qual neurônio ficará responsável pelos padrões identificados na entrada. Para cada padrão de entrada, os neurônios atribuem um peso e o que possuir o maior ganhará a competição;

Cooperação: O neurônio vencedor determina a posição no mapa. Como em uma curva de vizinhança gaussiana, os neurônios mais próximos do que foi ativado são mais fortemente excitados do que os que estão mais distantes. Essa vizinhança ajuda o neurônio vencedor no processo de aprendizagem;

Adaptação Sináptica: de maneira simplificada, o neurônio vencedor para um determinado padrão de entrada melhora sua resposta para situações similares. O neurônio que identifica a pessoa B vai conseguir perceber ela melhor, por exemplo, quando ela estiver com a voz um pouco alterada ou com ruídos de fundo.

Uma característica bem importante a respeito do SOM é que ele não é dependente das palavras ditas, o que não o torna muito recomendado para um sistema de segurança, mas o deixa mais versátil em relação a identificar alguém. Uma aplicação para esse tipo de sistema poderia ser o de identificar pessoas que não deveriam estar em um determinado ambiente ou mesmo identificar alguém através de um sinal de voz, seja uma gravação ou uma chamada.

O MFCC (Mel-frequency Cepstral Coefficients)

O MEL é uma escala logarítmica, feita de tal forma para se adaptar a forma como o ouvido humano identifica as frequências. O MFCC é um coeficiente, que se utiliza dessa escala MEL. Para se chegar neste coeficiente é preciso tratar o sinal de áudio e transformá-lo em um espectro que também identifica as formantes do sinal (picos de energia de cada região amostral), pois nele é mais fácil de extrair as informações relevantes da voz, a partir disso o MFCC extrai os padrões de informações que servirão de entrada para a rede neural.

Material e Métodos

Com o intuito de fazer um trabalho diferente, que tivesse como fim, identificar locutores utilizando os mapas de Kohonen, foi feita uma pesquisa da revisão bibliográfica sobre o que poderia ser feito, a princípio. Com isso foi descoberto que o sistema não é dependente do que é dito, então o fato de ele poder identificar uma pessoa falando, não importando onde ela se encontra, se torna mais provável. Um código do MatLab, a plataforma na qual previamente se pretendia realizar o projeto, feito por Kamil Wojcicki (2011), foi encontrado no *site* da mathworks. A partir dele, alguns testes para determinar se o projeto iria

ou não funcionar foram conduzidos. O programa desenvolvido por Kamil extrai os MFCC da voz, a partir disso com a ferramenta de redes neurais do MatLab, nntool, foram obtidos alguns resultados.

As vozes foram adquiridas através de um aplicativo de celular, em formato .opus e transformado em .wav, através de um site que faz conversões de formatos de diversos arquivos, então editado no "Audacity" para que todos os arquivos tenham o mesmo tamanho (visto que, em MatLab os vetores de tamanhos diferentes são completados com zeros, o que interfere nas amostras), os áudios foram inseridos no programa, que devolve uma matriz de MFCC, essa matriz foi transformada em um vetor (posicionando as colunas uma em baixo da outra), visto que a ferramenta nntool considera cada vetor como um parâmetro de entrada e por fim, a rede fora criada e treinada com os parâmetros de entrada. A figura 1 mostra a janela obtida no MatLab para a estrutura da rede neural utilizada.

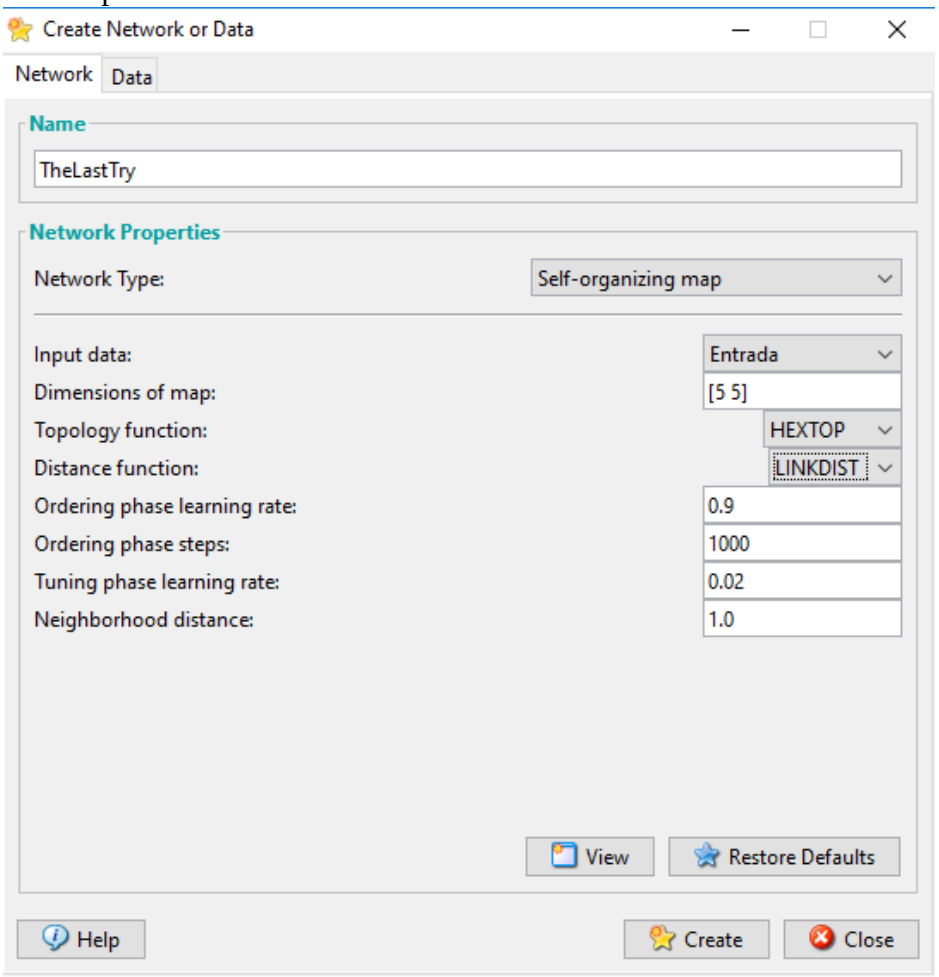

Figura 1 – Estrutura da rede de Kohonen utilizada no trabalho.

O tipo de rede foi alterado para o desejado, *SOM*, os dados de entrada foram os vetores de amostras tratados e a dimensão do mapa foi escolhida arbitrariamente, levando em conta apenas que a rede não poderia ser pequena demais, para que os vizinhos não se coincidam durante o aprendizado.

Os seguintes parâmetros permaneceram inalterados,

- *Topology function: HEXTOP;*
- *Distance function: LINKDIST;*
- *Ordering phase learning rate:* 0.9*;*
- *Ordering phase steps:* 1000*;*
- *Turning phase learning rate:* 0.02*;*
- *Neighborhood distance:* 1.0.

Em um estudo prévio, foram coletadas as vozes gravadas de duas pessoas, um homem e uma mulher, considerando alguns segundos na aquisição e os testes para reconhecimento utilizando apenas dois neurônios. As amostras criadas, a partir do MFCC, foram compostas por vetores de tamanho 3094 no MatLab. Os resultados apresentados pela rede não foram muito satisfatórios.

Considerando os resultados prévios com pouca eficiência no reconhecimento do locutor, foram realizadas alterações tanto na estrutura da rede neural quanto no tamanho dos arquivos de áudio utilizados para o treinamento. A rede neural foi modificada com uma maior quantidade de neurônios (foi escolhido arbitrariamente 25) criando uma matriz de 5x5 neurônios e o tempo dos áudios que serviram como informações de entrada com 77974 de tamanho correspondendo a 1 minuto de áudio.

Para o processo de convergência da rede foi utilizado um número de iterações proporcional a 500 vezes a quantidade de neurônios presentes nela, considerando os critérios estipulados por Haykin(2001) e Mafra(2002). Após os testes realizados com a rede já treinada, esperava-se que a rede fosse capaz de identificar os respectivos locutores, assim como "soubesse dizer" se não fosse nenhum deles. Os resultados foram mais satisfatórios.

Resultados e Discussão

Inicialmente houve uma desconfiança de que a baixa eficiência nos estudos iniciais poderia estar relacionada com a quantidade de informações do conjunto de treinamento da rede neural e que o programa não estava se comportando apropriadamente na obtenção dos coeficientes. Isso provocou a coleta de mais amostras, mas depois de vários estudos foi comprovado que o problema não era esse. Posteriormente descobriu-se, em conjunto com as conclusões feitas por Mafra(2002), que deveriam ser utilizados mais do que dois neurônios na estrutura da rede SOM, o que acabou fazendo sentido devido a interação que cada um dos neurônios têm com os seus vizinhos, no mapa gerado. As mudanças resultaram na figura 2, onde os hexágonos em azul demonstram os neurônios que foram utilizados e os números exibem a quantidade de informações em cada um mostrando que há uma separação em dois aglomerados separados na matriz 5x5 de neurônios.

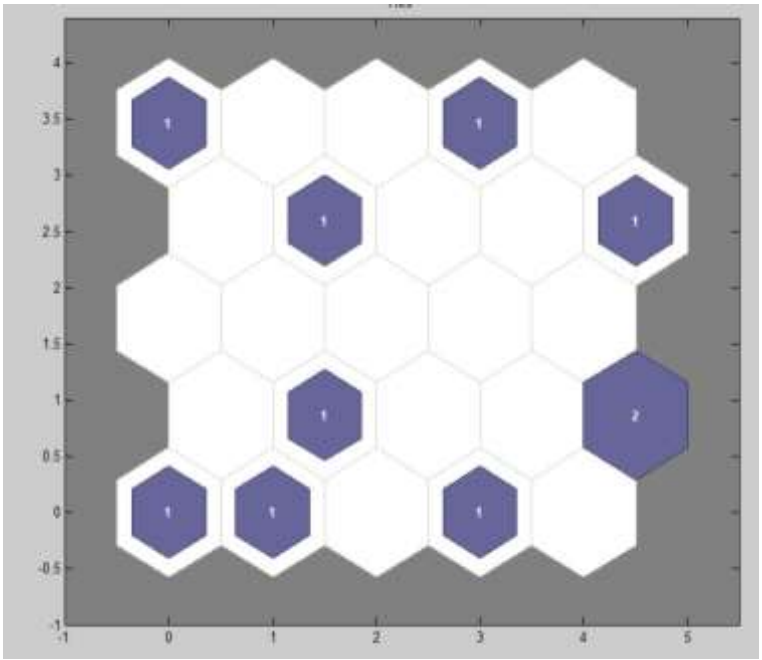

Figura 2 – Disposição dos dados de entrada em dois aglomerados na matriz 5x5.

Foi escolhida uma quantidade maior que o mínimo de neurônios para a rede, pensando também em uma possível expansão na quantidade de locutores. Definidas as variáveis do

sistema, resultaram na figura 2, que mostra o mapa, com a distribuição dos neurônios no mesmo e a versão final do programa já treinado.

Conclusões

A proposta inicial do projeto não previa o desenvolvimento de um programa computacional para realizar a extração dos coeficientes necessários para o treinamento e posterior reconhecimento na rede neural, portanto um código já pronto foi utilizado, que pode ser alterado num trabalho futuro para melhorar os resultados obtidos e que seja dedicado para o objetivo de se identificar locutores.

Durante a pesquisa foi notado que existem duas principais formas de se estabelecer um padrão para identificação de locutor. Se o sistema é ou não dependente de texto: quando ele é, o sistema se torna mais seguro podendo ser utilizado para uma rede de segurança como acesso a um cofre, por exemplo; quando não é dependente de texto, o sistema se torna mais versátil, o que é ruim para segurança, mas por outro lado, melhor para quando se busca identificar alguém em diversas situações.

A teoria, tanto de redes neurais, que engloba Learning Vector Quantization (LVQ), Hidden Markov Models (HMM), Self-Organizing Maps (SOM), quanto das formas de extração das características das vozes e dos filtros necessários para construção de um projeto físico são desafiadoras em entendimento de seus funcionamentos e interessantes para realizar em um projeto futuro também (montar uma rede neural física ou mais portátil). A quantidade de informações sobre o assunto abordado neste trabalho é extensa tanto em material disponível na internet quanto em livros que explicam sobre as bases teóricas que foram usadas neste projeto e isso faz com que seja mais fácil sua compreensão.

Referências Bibliográficas

BRAGA, P. (Maio/2006) Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais. Universidade de Pernambuco.

BRANDÃO, A. (Junho/2005) Redes Neurais Artificiais Aplicadas ao Reconhecimento de Comandos de voz. UNIVERSIDADE FEDERAL DE VIÇOSA.

HASAN, R. e Jamil, M. e Rabbani, G. e Rahman, S. (Dezembro/2004) Speaker identification using MEL frequency cepstral coefficients.

HAYKIN, (2001) Simon. Redes Neurais: princípios e prática 2° edição.

MAFRA, A. (2002) Reconhecimento Automático de Locutor em Modo Independente de Texto por Self-Organizig Maps. Escola Politécnica da Universidade de São Paulo.

MUDA, L. e Begam, M. e Elamvazuthi, I. (Março/2010) Voice Recognition Algorithms using MEL Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3*.

PRAHALLAD, K. Speech Technology: A Practical Introduction. Carnegie Mellon University & International Institute of Information Technology Hyderabad.

ZUBEN, F. Attux, R. Rede Neural de Kohonen e Aprendizado Não-Supervisionado. DCA/FEEC/Unicamp & DECOM/FEEC/Unicamp.