Um Estudo da Aplicação de Redes Neurais Auto-Organizáveis para a Identificação Autônoma de Fonemas Portugueses

A Study on the Application of Self-Organizing Neural Networks for Autonomous Identification of Portuguese Phonemes

Raul Arthur Fernandes Rosa
Departamento de Engenharia de Computação e Automação Industrial
Universidade Estadual de Campinas - UNICAMP, Campinas, SP
rauleng2007@yahoo.com.br

Marcos Eduardo Valle
Departamento de Matemática
Universidade Estadual de Campinas - UNICAMP, Campinas, SP
valle@ime.unicamp.br

Resumo: O problema de reconhecimento automático de discurso tem se tornado alvo de estudos, em partes, devido à grande demanda por ferramentas que realizem esse processo de maneira rápida e eficaz. Este artigo investiga a aplicação da rede neural auto-organizável (SOM) de Kohonen para análise de fonemas da língua portuguesa em sinais de voz. Especificamente, a rede neural foi aplicada num difícil problema de reconhecimento de discurso contínuo, dependente do orador e com alfabeto aberto. Experimentos computacionais revelaram que a SOM é capaz de identificar corretamente muitos fonemas de um discurso contínuo. Porém, a rede também identificou alguns fonemas inoportunos que podem estar relacionados à transição de palavras, rotulação manual errada, ou que possuem sons semelhantes.

Palavras-chave: mapas auto-organizáveis; reconhecimento automático de discurso; redes neurais artificiais.

Recebido em 23/04/2012 - Aceito em 20/02/2013.

Abstract: The automatic speech recognition problem has become a target of studies, partially due to the high demand for tools that perform such process quickly and effectively. This paper investigates the application of the Kohonen's self-organizing neural network (SOM, Self-Organizing Map) for the analysis of Portuguese phonemes in speech signals. Specifically, the neural network has been applied to a difficult problem of continuous speech recognition, speaker-dependent, and open alphabet. Computational experiments revealed that the SOM can identify correctly many phonemes of a continuous speech. However, the network also identified some inopportune phonemes which are possibly related to the transition of words, to the wrong manual labeling, or to the fact that they have similar sounds.

Key words: self-organizing maps, automatic speech recognition, artificial neural network.

1 Introdução

Em termos gerais, o reconhecimento autônomo de discurso (ASR, acrônimo do termo inglês *automatic speech recognition*) consiste em extrair ou identificar informações relevantes em um discurso sem o auxílio contínuo de um ser humano. O ASR tem evoluído como uma das áreas líderes em ciência da computação [1]. As aplicações abrangem sistemas de atendimento ao cliente, telemarketing automatizados, máquinas (robôs, eletrodomésticos, computadores, etc.) controladas por voz, aplicativos para celulares, entre outras [2–6]. Os obstáculos encontrados numa aplicação envolvendo ASR incluem, por exemplo, presença de ruídos, diferenças entre sotaques dos oradores, grande número de fonemas e palavras, além de restrições na velocidade da realização da tarefa de reconhecimento.

Os estudos em ASR iniciaram em 1952 no Bell Labs com o reconhecimento de dígitos pronunciados via telefone [7]. Conforme os computadores evoluíram nos anos 1960s, novas técnicas baseadas em programação dinâmica foram desenvolvidas. Nos anos 1970s, sugiram grandes contribuições na área devido ao modelo linear

preditivo, que oferece uma forma eficiente de representar um sinal de voz [8]. O modelo linear preditivo continua sendo usado em muitas aplicações, embora tenha sido substituído pelo MFCC (acrônimo de *mel-frequency cepstral coefficients*) desenvolvido nos anos 1980s [8]. De fato, nos anos 1980s, surgiram diversos bancos de dados com sinais de voz e a técnica comum da época estava baseada no uso de *templates* para a identificação de fonemas. Posteriormente, técnicas baseadas nos modelos de Markov escondidos (HMM, do termo inglês *hidden Markov models*) foram empregados com a ideia de substituir os *templates* por modelos probabilísticos mais simples [9]. Finalmente, nos anos 1990s, foram desenvolvidos modelos baseados em *wavelets* [10], máquinas de vetores de suporte [11, 12] e redes neurais artificiais (ANNs, acrônimo do inglês *artificial neural networks*) [13, 14]. Uma revisão mais apurada das diversas técnicas pode ser encontrada em [15]. Atualmente, sistemas baseados em HMM bem como em técnicas baseadas em inteligência computacional e sistemas híbridos são utilizados em ASR [1, 16–19].

Este artigo contém um estudo da aplicação de ANNs para o ASR, precisamente para a identificação de fonemas. Uma ANN é um modelo matemático inspirado no cérebro humano, onde as unidades básicas de processamento são os neurônios [13, 14, 20]. Os estudos das redes neurais artificiais iniciaram em 1943, por McCulloch e Pitts [21]. No início dos anos 1980s, Kohonen apresentou um modelo de ANN auto-organizável conhecida como mapa auto-organizável e referida como SOM (acrônimo do termo inglês *self-organizing maps*) [22, 23]. A SOM, e suas variações, representam a classe mais popular de ANNs com aprendizado não-supervisionado, isto é, sem professor [14]. A SOM de Kohonen foi aplicada com sucesso em diversas áreas, incluindo estatística, processamento de sinais, teoria de controle, análise financeira, física experimental, química e medicina [23]. Com efeito, a SOM pode ser empregada em problemas de dimensões grandes e não-lineares, incluindo a extração de características em imagens e padrões acústicos como discursos. Além disso, a SOM pode ser usada para estabelecer uma correspondência entre as entradas e uma tabela de unidades – geralmente com uma ou duas dimensões – que preserva as relações topológicas e a distribuição de probabilidade dos dados [13].

Em 1988, Kohonen apresentou uma aplicação da SOM para o reconhecimento autônomo de fonemas da língua finlandesa [24]. Recentemente, Venkateswarlu e Kumani apresentaram diversas técnicas para extrair características relevantes de sinais de áudio apropriadas para serem usados como entrada para a SOM [18]. Behi et al. apresentaram uma nova variação da SOM baseada em neurônios impulsivos [25]. Modelos de SOM baseados numa abordagem probabilística podem ser encontrados em [17, 19]. Uma aplicação da SOM para o reconhecimento de discurso na língua portuguesa pode ser encontrada em [26]. Nessa referência, Sousa et al. consideram um problema de ASR com palavras isoladas de um alfabeto restrito composto apenas por números e operações aritméticas elementares como, por exemplo, as palavras "um", "dois", "vezes", "igual", "mais", etc.

Este artigo também contém uma aplicação da SOM para a identificação de fonemas da língua portuguesa em sinais de áudio. Contudo, diferente da abordagem apresentada por Sousa et al., um problema de reconhecimento de discurso contínuo, dependente do orador e com alfabeto aberto é estudado. Esse problema é mais complexo pois, além da ausência de pausas e sinalizações de início e fim das palavras, há também uma quantidade muito maior de informação. Portanto, este artigo apresenta uma avaliação do desempenho da SOM num problema mais complexo relacionado à identificação de discurso.

O artigo está organizado da seguinte forma. A próxima seção contém a fundamentação teórica, com a descrição do mapa auto-organizável e dos *mel-frequency cepstral coefficients*. A seção 3 apresenta os experimentos computacionais. O artigo termina com a conclusão na seção 4.

2 Fundamentação teórica

Esta seção apresenta os principais conceitos utilizados no artigo. Especificamente, primeiro é exposta uma descrição do mapa auto-organizável de Kohonen. Posteriormente, descreve-se a ferramenta utilizada para extrair as características de um discurso, o *mel-frequency cepstral coefficients*. Ambos conceitos são apresentados como módulos que foram implementados em rotinas diferentes. Em vista disso, cada uma das subseções seguintes segue notação própria. Por exemplo, a letra w é

usada para denotar os pesos sinápticos do mapa auto-organizável na subseção 2.1. A mesma letra é usada na subseção 2.2 para representar o conceito de janela.

2.1 Mapa auto-organizável

O córtex cerebral humano é organizado de modo que sensações diferentes excitam regiões distintas. Em outras palavras, diferentes áreas do córtex são ativadas por diferentes estímulos. O mapa auto-organizável SOM, de Kohonen, também referido como rede de Kohonen, é desenvolvido a partir dessa característica do córtex cerebral.

Resumidamente, a rede de Kohonen é determinada por uma regra de aprendizagem não-supervisionada composta de três processos [14, 23]. O primeiro é um processo de competição entre os neurônios da rede para um dado estímulo. Essa competição é determinada a partir dos valores apresentados por uma função discriminante que relaciona o padrão de entrada (estímulo) a cada um dos neurônios da rede. O neurônio que apresenta o maior valor da função discriminante é o vencedor. O segundo processo é de cooperação, no qual o neurônio vencedor excita os seus neurônios vizinhos. A noção de vizinhança espacial é definida *a priori*. O último processo, chamado de adaptação sináptica, consiste em atualizar os valores dos pesos sinápticos dos neurônios ativados.

Como consequência dos três processos da regra de aprendizagem, a localização espacial de um neurônio na rede de Kohonen é indicativo das características estatísticas intrínsecas contidas nos padrões de entrada [22, 27]. Em vista disso, os neurônios são geralmente arranjados em uma rede uni ou bidimensional, dado que cada neurônio tenha um conjunto de neurônios vizinhos. Consequentemente, a rede de Kohonen pode ser vista como uma transformação não-linear que projeta padrões de entrada de dimensões arbitrárias em um mapa discreto uni ou bidimensional. O mapa se forma seguindo um processo adaptativo, organizando-se de uma maneira topologicamente ordenada e simulando as características do córtex cerebral. Um exemplo interessante e bem visual da SOM, no qual é formado um mapa através das características de mamíferos e aves, foi apresentado por Ritter e Kohonen em [28] e estudado posteriormente por Haykin [14] e Bezdek [29].

O algoritmo que realiza os processos de competição, cooperação e adaptação é chamado algoritmo SOM e está apresentado de forma simplificada no algoritmo 1. Para tanto, um padrão (vetor) de entrada, apresentado a rede no instante t, será representado por

$$\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_m(t)]^T, \quad \forall t = 0, 1, 2, \ldots, \tag{2.1}$$

em que m é a dimensão do espaço de entrada. O vetor peso sináptico do k-ésimo neurônio da rede no instante t será denotado por

$$\mathbf{w}_k(t) = [w_{k1}(t), w_{k2}(t), \ldots, w_{km}(t)]^T. \tag{2.2}$$

Observe que o vetor peso sináptico tem a mesma dimensão do espaço de entrada. O número de neurônios da rede será denotado por l.

O parâmetro da taxa de aprendizagem, ϵ, e a função de vizinhança, $h_{k,i(\mathbf{x})}(t)$, possuem as seguintes características [14, 23]:

- A vizinhança espacial, ou função de vizinhança, $h_{k,i(\mathbf{x})}(t)$ assume o valor máximo no neurônio vencedor $i(\mathbf{x})$ e decresce com o aumento da distância espacial $d_{k,i(\mathbf{x})}$ entre o neurônio vencedor $i(\mathbf{x})$ e o neurônio excitado k.
- O tamanho da vizinhança espacial deve decrescer com o aumento de t.
- O parâmetro da taxa de aprendizagem permanece com valor fixo para todo instante t.

A função de vizinhança $h_{k,i(\mathbf{x})}(t)$ adotada neste artigo foi a função Gaussiana

$$h_{k,i(\mathbf{x})}(t) = \exp\left(-\frac{d_{k,i}}{\sigma(t)^2}\right). \tag{2.3}$$

A distância espacial $d_{k,i(\mathbf{x})}$, no caso bidimensional, foi definida através da equação

$$d_{k,i(\mathbf{x})} = \left\| r_k - r_{i(\mathbf{x})} \right\|, \tag{2.4}$$

em que r_k e $r_{i(\mathbf{x})}$ definem, respectivamente, a posição dos neurônios k e $i(\mathbf{x})$ na rede. Ambos r_k e $r_{i(\mathbf{x})}$ foram definidos como sendo pontos do conjunto discreto $\{1, 2, \ldots, N_L\} \times \{1, 2, \ldots, N_C\}$, que representa uma malha $N_L \times N_C$. O parâmetro

$\sigma(t)$ mede o grau com o qual neurônios vizinhos ao neurônio vencedor participam do processo de aprendizagem. Seguindo a sugestão de Ritter e Kohonen, a função σ foi definida como

$$\sigma(t) = \sigma_i \left(\frac{\sigma_f}{\sigma_i} \right)^{t/t_f}, \quad \text{para } t = 0, 1, 2, \ldots, \tag{2.5}$$

em que σ_i é o valor de $\sigma(t)$ na inicialização do algoritmo SOM, σ_f é o valor final e t_f é o número máximo de iterações.

Algoritmo 1: Algoritmo SOM

Inicialização: Os pesos sinápticos dos neurônios da rede, $\mathbf{w}_k(0)$ para $k = 1, 2, \ldots, l$, são inicializados ou com valores pequenos e arbitrários ou com valores dos dados de entrada selecionados de maneira aleatória. Em ambos os casos, os pesos sinápticos iniciais devem ser dois a dois distintos. Forneça o maior número permitido de iterações t_f e defina $t = 0$.
enquanto *o mapa auto-organizável apresentar alterações significativas em sua forma ou* $t \leq t_f$ **faça**

1. **Amostragem:** Escolha aleatoriamente um padrão de entrada $\mathbf{x}(t)$ que será apresentado à rede.
2. **Casamento por Similaridade:** Encontre o neurônio vencedor através da equação
$$i(\mathbf{x}) = \underset{k=1:l}{\operatorname{argmin}} ||\mathbf{x}(t) - \mathbf{w}_k||.$$
Neste artigo, utilizamos a distância euclidiana na função discreta $i(\mathbf{x})$.
3. **Atualização:** Após a escolha do neurônio vencedor, os vetores dos pesos sináptico de todos os neurônios são ajustados utilizando a equação
$$\mathbf{w}_k(t+1) = \mathbf{w}_k(t) + \epsilon h_{k,i(\mathbf{x})}(t)(\mathbf{x}(t) - \mathbf{w}_k(t)),$$
em que ϵ é o parâmetro da taxa de aprendizagem e $h_{k,i(\mathbf{x})}(t)$ é a função de vizinhança centrada em torno do neurônio vencedor $i(\mathbf{x})$.
4. Faça $t \leftarrow t + 1$.

2.2 Cálculo dos *mel-frequency cepstral coefficients*

Esta subseção descreve os passos necessários para calcular o *mel-frequency cepstral coefficients* (MFCCs) de um sinal de voz. Os MFCCs capturam as informações que serão usadas para a identificação dos fonemas e para a síntese do mapa fonético. A abordagem adotada neste artigo está baseada na referência [30].

1. **Pré-Ênfase:** Como em um sinal de voz as frequências baixas apresentam maior energia que as frequências altas, o primeiro passo consiste em aumentar a energia nas frequências altas. Para tanto, utilizou-se um filtro de primeira ordem passa-altas dado pela seguinte equação para todo n:

$$y(n) = s(n) - \alpha s(n-1). \tag{2.6}$$

   Aqui, s denota o sinal de entrada de tamanho N, y o sinal filtrado e α é uma constante com valores no intervalo $[0,9;1]$. O valor $\alpha = 0,95$ foi adotado neste artigo.

2. **Escolha da Janela:** Um discurso é um sinal não estacionário, uma vez que suas características dependem dos fonemas nele contidos. A transformada de Fourier e, consequentemente, o cálculo do MFCC, é desenvolvida para sinais estacionários. Contudo, é possível converter um sinal não estacionário numa sequência de sinais estacionários menores, compostos de vários trechos do sinal original. Nos experimentos computacionais, o sinal original foi particionado em trechos de tamanho N que corresponde à 20ms. Os trechos foram selecionados a cada 10ms do sinal de voz.

   Além disso, antes do cálculo da transformada de Fourier discreta, cada trecho do sinal foi multiplicado pela janela de Hamming. A janela de Hamming suaviza os extremos do trecho do sinal, evitando descontinuidades que podem gerar problemas na análise de Fourier. A janela de Hamming é determinada

pela equação

$$w(n) = \begin{cases} 0,54 - 0,46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1, \\ 0, & \text{caso contrário,} \end{cases} \tag{2.7}$$

em que n é a variável de tempo e N é o tamanho da janela (iniciando em $n = 0$).

3. **Transformada de Fourier Discreta:** A transformada de Fourier discreta (DFT, do termo inglês *discrete Fourier transform*) é usada para extrair informações sobre o espectro de um sinal. Com efeito, a DFT revela a quantidade de energia que o sinal contém nas diferentes faixas de frequência. A saída da DFT, denotada por $X(k)$, é um número complexo com a magnitude e a fase da k-ésima componente de frequência no sinal. A DFT de x é determinada pela equação

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \quad k = 0, \ldots, N-1. \tag{2.8}$$

Nessa equação, N denota o tamanho de x e j representa a unidade imaginária, i.e., $j = \sqrt{-1}$. Na prática, a DFT é determinada utilizando a transformada de Fourier rápida (FFT, do termo inglês *fast Fourier transform*), que reduz significativamente o esforço computacional.

4. **Filtro Mel e Cálculo do Log:** A audição humana não é igualmente sensível para todas as faixas de frequência. De fato, ela é menos sensível a frequências maiores que 1000 Hz. O filtro mel foi utilizado para simular essa característica [31, 32].

   O filtro mel é composto por 10 faixas de frequência com espaçamento linear até 1000 Hz e escala logarítmica até 10 kHz.

   O filtro é aplicado em todas os trechos do sinal do discurso no intuito de simular a percepção humana. Além disso, calcula-se o logaritmo de todos os

valores obtidos pois, nos humanos, a resposta a um nível de sinal é logarítmica. Sobretudo, o logaritmo deixa as estimativas menos sensíveis a variações de potência causadas pela proximidade ou não da boca do narrador ao microfone durante na gravação do discurso.

5. ***Cepstrum* – O Inverso da Transformada de Fourier Discreta:** O termo "*cepstrum*" foi introduzido em 1963 por Borget, Healy, e Tukey em um artigo chamado "*The Quefrency Alanysis of Time Series for Echoes*" [33]. O *cepstrum* é o espectro de potência do log do espectro de potência de um sinal. Para um sinal de tempo-discreto, a melhor definição do *cepstrum* é o inverso da transformada de Fourier discreta (IDFT, do termo inglês *inverse discrete Fourier transform*) do logaritmo da magnitude da DFT. Matematicamente, o *cepstrum*, denotado por $c(n)$, é dado pela seguinte equação em que X denota a DFT de um sinal x:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)|e^{+j\frac{2\pi}{N}kn}, \quad n = 0,\ldots,N-1. \tag{2.9}$$

A forma de onda criada pela fala humana é fruto do movimento glotal, que gera uma forma de onda em uma frequência fundamental. Essa onda passa pelo trato vocal, que devido ao seu formato, é basicamente um filtro. Porém, as informações produzidas pelo movimento glotal não são importantes para distinguir os diferentes sons vocais. A informação mais relevante para a detecção vocal é o filtro. O *cepstrum* consegue, de certa forma, separar a fonte do filtro de um som.

Após aplicar o *cepstrum*, o trecho do sinal que estava no domínio da frequência volta ao domínio do tempo. Logo nas primeiras amostras está a informação necessária para a identificação dos fonemas. Desse modo, apenas os primeiros 12 valores *cepstrais* foram utilizados neste artigo. Esses 12 coeficientes representam exatamente as informações sobre o filtro formado pelo trato vocal, separando as informações desnecessárias contidas na fonte como o movimento glotal.

Finalmente, conforme descrito em [8], o cálculo da IDFT foi efetuado utilizando o inverso da transformada do cosseno discreta (DCT, *discrete cosine transform*). A DCT é dada pela equação

$$c(n) = \beta(k) \sum_{k=0}^{N-1} X(k) \cos\left(\frac{\pi k(2n+1)}{2N}\right), \quad \text{para } n = 0, \ldots, N-1, \quad (2.10)$$

com

$$\beta(k) = \begin{cases} \frac{1}{\sqrt{N}}, & n = 0, \\ \sqrt{\frac{2}{N}}, & 1 \leq n \leq N-1, \end{cases} \quad (2.11)$$

em que N é o tamanho de X. Dessa forma, o vetor dos coeficientes MFCC de um trecho de 20ms do sinal de discurso foi definido neste artigo como $\mathbf{c} = [c(1), c(2), \ldots, c(12)]$. Esclarecimentos com respeito a relação entre a DCT e a IDFT podem ser obtidos em [34].

3 Experimentos computacionais

Os experimentos computacionais foram realizados considerando a frase:

> "Olá amigo ouvinte, hoje temos por título o seguinte tema: Envelhecer, sim, é inevitável, mas crescer é opcional;" (3.12)

e as sentenças "cinco", "assim", "pense nisso" e "afinal". Tanto a frase como as quatro sentenças foram extraídas de uma gravação do programa "Maturidade em Foco" de Marcelo Caires na Rádio UEL. Os fonemas da frase e das sentenças foram identificados manualmente, utilizando um editor de áudio digital gratuito, o *software* Audacity®. Essa informação foi utilizada tanto para a elaboração do mapa fonético como para teste e validação da rede de Kohonen como uma aplicação ASR.

Primeiramente, a frase (3.12) foi particionada em intervalos de 20ms e com deslocamento de 10ms. Foram extraídos 12 coeficientes MFCC de cada partição conforme descrito na subseção 2.2. Obteve-se assim um conjunto de dados de treinamento com 797 vetores com 12 elementos, cada vetor representando um fragmento

de 20ms da frase (3.12). Por fim, com os dados de treinamento, o algoritmo 1 foi utilizado para sintetizar o mapa fonético de dimensão $N_L \times N_C$.

Após a execução do algoritmo 1, os neurônios da rede foram rotulados com base na identificação manual dos fonemas da frase (3.12). Especificamente, para cada neurônio da rede, identificou-se qual o fonema fornece a maior similaridade, ou seja, cujo vetor de coeficientes MFCC é o mais próximo do vetor dos pesos sinápticos. O neurônio foi então rotulado com o fonema identificado manualmente. É importante ressaltar que a duração dos fonemas portugueses variam, em geral, entre 47ms e 229ms [35]. Portanto, tanto os dados utilizados para o treinamento como os neurônios da rede não representam exatamente um fonema, mas parte dele. Somente os trechos identificados manualmente representam completamente o fonema. Ainda assim, refere-se à rede de Kohonen obtida ao final do processo como um "mapa fonético".

O mapa fonético pode ser utilizado para o ASR. Precisamente, tal como no processo de treinamento da rede, uma sentença ou frase desconhecida S é inicialmente particionada em trechos de 20ms e com deslocamento de 10ms. Para cada trecho, o vetor contendo os 12 coeficientes MFCC é apresentado à rede e o rótulo do neurônio com maior similaridade é atribuído ao fragmento do discurso apresentado. Em termos matemáticos, o ASR baseado na rede de Kohonen resulta numa sequência finita de rótulos $\mathbf{r} = \{r_1, r_2, \ldots, r_L\}$, em que L denota o número de partições do discurso apresentado S. Por exemplo, a sentença $S_1 =$ "Olá amigo" tem duração total de 700ms e, portanto, resulta em $L = 27$ partições de 20ms. A seguinte sequência de rótulos foi obtida após apresentar essa sentença a um mapa fonético de dimensão 38×14:

$$\mathbf{r}_1 = \begin{Bmatrix} \text{“o”, “o”, “o”, “l”, “l”, “l”, “l”, “l”, “a”, “a”, “a”, “a”, “a”, “a”,} \\ \text{“n”, “m”, “m”, “m”, “n”, “m”, “m”, “i”,“g”,“g”, “o”, “g”} \end{Bmatrix}. \quad (3.13)$$

Note que a transcrição fonética correta da sentença "Olá amigo" é "o, l, a, m, i, g, o". Portanto a rede de Kohonen foi capaz de identificar todos os fonemas dessa sentença. Contudo, ela também identificou o fonema "n" que, embora similar ao "m", não pertence ao discurso apresentado.

Considerando que um fonema pode ter duração maior que 20ms, adotou-se como medida do desempenho da rede como ASR a razão entre o número de rótulos que correspondem a fonemas do discurso pelo número total de fragmentos. Formalmente, a razão de reconhecimento $\mathscr{R}$ de um discurso S é dado por

$$\mathscr{R}(S) = \frac{1}{L}\sum_{k=1}^{L}\chi_S(r_k), \tag{3.14}$$

em que χ_S denota a função indicadora do conjunto dos fonemas do discurso S, ou seja,

$$\chi_S(r) = \begin{cases} 1, & \text{se } r \text{ é um fonema do discurso } S, \\ 0, & \text{caso contrário.} \end{cases} \tag{3.15}$$

Por exemplo, a razão de reconhecimento da sentença $S_1 =$ "Olá amigo" é:

$$\mathscr{R}(S_1) = \frac{25}{27} = 0,9259. \tag{3.16}$$

Observe que a razão de reconhecimento de um discurso S satisfaz a equação $\mathscr{R}(S) = 1$ se todos os rótulos da sequência **r** correspondem a fonemas de S. A recíproca, porém, não é verdadeira. Por exemplo, se todos os rótulos de **r** fossem associados ao fonema "o", então a razão de reconhecimento resultaria 1. Todavia, situações como a desse exemplo não foram observadas na prática.

A dimensão da rede de Kohonen, bem como outros parâmetros empregados no algoritmo 1, foram determinados utilizando o discurso que contém a palavra "cinco", cujo conjunto de fonemas é {"s", "i˜", "k", "o"}. Especificamente, foram fixados os parâmetros $\sigma_i = 3$, $\sigma_f = 0,02$, $n_{max} = 10000$ e $\epsilon = 0,3$, enquanto as dimensões N_C e N_L da rede variaram ambas no conjunto $M = \{10, 14, \ldots, 54, 58\}$. Esse processo foi repetido cinco vezes a fim de obter uma média do desempenho de cada um dos pares N_C e N_L. A maior razão de reconhecimento, com o valor 0,4565, foi obtida para o par $(N_L, N_C) = (38, 14)$.

Posteriormente, ainda utilizando a palavra "cinco", foram determinados sequencialmente os melhores valores para os parâmetros n_{max}, σ_i, σ_f e ϵ. A maior razão de

reconhecimento foi obtida sintetizando uma rede de dimensões $N_L = 38$ e $N_C = 14$ com os parâmetros $\sigma_i = 8$, $\sigma_f = 0{,}73$, $n_{max} = 16000$ e $\epsilon = 0{,}11$. A razão de reconhecimento da palavra "cinco" foi $\mathscr{R}(S_2) = 0{,}6522$, que foi determinada sobre a seguinte sequência de rótulos:

$$\mathbf{r}_2 = \left\{ \begin{array}{l} \text{“s”, “s”, “v”, “i˜”, “i˜”, “i˜”, “i˜”, “m”, “n”, “n”, “m”, “m”, “n”,} \\ \text{“n”, “o”, “o”, “o”, “o”, “o”, “o”, “o”, “o”, “o”} \end{array} \right\}. \quad (3.17)$$

Observe que o fonema "k" não foi identificado pela rede de Kohonen. De fato, esse fonema aparece somente uma vez na frase utilizada para o treinamento da rede. Especificamente, na palavra "crescer". Desse modo, devido à diferentes sonoridades do fonema, não foi possível seu reconhecimento.

Tabela 1. Resumo do desempenho da rede de Kohonen.

Treinamento	
Palavra	Razão de reconhecimento
"Olá amigo"	0,9259
Ajuste dos parâmetros	
Palavra	Razão de reconhecimento
"cinco"	0,6522
Validação	
Palavra	Razão de reconhecimento
"assim"	0,5625
"pense nisso"	0,7895
"afinal"	0,7200

A tabela 1 sintetiza o desempenho final da aplicação da rede de Kohonen no problema de ASR considerado neste artigo. Precisamente, essa tabela apresenta a razão de reconhecimento das sentenças: $S_1 =$ "Olá amigo", $S_2 =$ "cinco", $S_3 =$ "assim", $S_4 =$ "pense nisso" e $S_5 =$ "afinal". Note que a primeira sentença pertence ao conjunto de treinamento. A palavra "cinco", embora não pertença ao conjunto de treinamento, foi utilizada no ajuste da dimensão da rede e dos parâmetros do algoritmo 1. Portanto, os valores mais significativos referem-se à razão de reconhecimento das sentenças S_3, S_4 e S_5. Com efeito, o mapa fonético forneceu a seguinte sequência de

rótulos para a sentença “assim”, cuja transcrição fonética é “a, s, i˜”:

$$\mathbf{r}_3 = \left\{ \begin{matrix} \text{“v”, “s”, “s”, “s”, “s”, “s”, “s”, “v”, “n”, “i˜”, “i˜”, “i˜”, “n”, “n”,} \\ \text{“n”, “n”} \end{matrix} \right\}. \quad (3.18)$$

Note que o mapa fonético não foi capaz de identificar o fonema “a”. Além disso, a sequência de rótulos (3.18) contém os fonemas “n” e “v” que não deveriam ser reconhecidos. Esses dois fonemas, representando consoantes, são difíceis de identificar mesmo manualmente e devem ser interpretados como ruídos na aplicação ASR baseada na rede de Kohonen.

De um modo similar, o mapa fonético forneceu respectivamente as seguintes sequências de rótulos para as sentenças “pense nisso” e “afinal”:

$$\mathbf{r}_4 = \left\{ \begin{matrix} \text{“m”, “e”, “e”, “i˜”, “i˜”, “i˜”, “n”, “v”, “s”, “s”, “v”, “n”, “n”,} \\ \text{“i˜”, “n”, “n”, “n”, “n”, “m”, “i˜”, “i˜”, “i”, “m”, “i”, “n”, “v”,} \\ \text{“s”, “s”, “s”, “s”, “s”, “v”, “r”, “i˜”, “i˜”, “i”, “i˜”, “i˜”} \end{matrix} \right\}, \quad (3.19)$$

e

$$\mathbf{r}_5 = \left\{ \begin{matrix} \text{“a”, “a”, “l”, “m”, “i”, “i”, “i”, “i”, “n”, “i”, “m”, “n”, “n”, “n”,} \\ \text{“n”, “a”, “a”, “a”, “a”, “a”, “l”, “l”, “l”, “l”, “l”} \end{matrix} \right\}. \quad (3.20)$$

As transcrições fonéticas dessas duas sentenças são “p, e˜, s, e, n, i, s, o” e “a, f, i, n, a, l”, respectivamente. Por um lado, a rede de Kohonen não conseguiu identificar todos os fonemas presentes nas duas sentenças. Em particular, o fonema “f” da sentença S_5 não foi identificado porque que não está presente no conjunto de treinamento. Novamente, observa-se a presença de ruído manifestado através dos fonemas “n” e “v”. Por outro lado, o mapa fonético cumpriu sua tarefa ao reconhecer a maioria dos fonemas presentes nas sentenças. Em outras palavras, apesar da falha no reconhecimento de fonemas que não estão presentes, a rede teve êxito em reconhecer muitos dos fonemas que compõem as sentenças.

Concluindo, existe na literatura diferentes abordagens para o ASR usando redes neurais auto-organizáveis [18, 25, 26]. Em alguns casos, observa-se que a rede neural apresenta desempenho entre 0,80 e 0,98. Esses números, porém, não seguem

o mesmo critério de avaliação adotado neste artigo. Em Venkateswarlu e Kumari [18], por exemplo, o método de avaliação segue o princípio de correto ou não. Especificamente, apresenta-se um fonema ou palavra à rede, dito por vários locutores, e é contabilizado o número de acertos e erros na identificação para cada locutor. Diferentemente, neste trabalho, a avaliação é feita através da recuperação de uma sentença completa e contínua no qual os fonemas não foram isolados *a priori*.

4 Considerações finais

Diferente de muitos trabalhos disponíveis na literatura, este trabalho apresenta um estudo da aplicação do mapa auto-organizável (SOM) de Kohonen, em conjunto com o *mel-frequency cepstral coeficientes* (MFCC), para a identificação de fonemas num discurso contínuo, dependente de orador e com alfabeto aberto. Em acordo com os resultados obtidos por Kohonen, para a identificação de fonemas da língua finlandesa, os experimentos computacionais revelaram que a SOM é capaz de identificar corretamente muitos fonemas de um discurso contínuo. Com efeito, a razão de reconhecimento do discurso usado para treinamento foi 0,9259. Porém, a rede de Kohonen identificou também fonemas que não fazem parte da sentença apresentada à rede. Observou-se que muitos dos fonemas inoportunos possuem um som muito semelhante a fonemas corretos, como é o caso do “m” e do “n”. Ressalta-se, porém, que alguns fonemas inoportunos também podem estar relacionados à transição de uma palavra ou sílaba ou decorrer de uma possível falha no processo manual de identificação dos fonemas.

Uma vez que um fonema da língua portuguesa pode ser influenciado pelos fonemas antecessores e sucessores, trabalhos futuros podem utilizar bi- ou tri-fonemas no lugar de fonemas. Além disso, o desempenho do processo de aprendizagem e reconhecimento podem ser melhorados, utilizando algoritmos que sinalizam o início e o fim dos fonemas ou das palavras. Por fim, estudos futuros também podem ser conduzidos na aplicação da SOM para a identificação de fonemas independente do orador, ou seja, discursos proferidos por pessoas diferentes seriam utilizados tanto para o treinamento como para a avaliação da SOM como ASR.

5 Agradecimentos

Este trabalho contou com o apoio da Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Paraná e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo no. 304240/2011-7.

6 Referências

[1] GANGAPUTRA, K. Error minimization in phoneme based automated speech recognition for similar sounding phonemes. In *International Conference on Communication Technology and System Design*, n.30, Procedia Engineering, p.402–409, 2011.

[2] CHONG, J.; GONINA, E.; KOLOSSA, D.; ZEILER, S.; KEUTZER, K. An automatic speech recognition application framework for highly parallel implementations on the gpu. Tech. Rep. UCB/EECS-2012-47, EECS Department, University of California, Berkeley, Apr 2012.

[3] KOMER, J. L.; GEPNER, J. E.; SHERWOOD, C. G. Automatic speech recognition system and method for aircraft, February 2010. U.S. Patent 7,912,592 B2.

[4] LAVANIA, K. K.; SHARMA, S.; SHARMA, K. K. Reviewing human-machine interaction through speech recognition approaches and analyzing an approach for designing an efficient system. *Int Comp Appl*, v.38, n.3, p.26–32, 2012.

[5] MCCARLEY, J. S.; QIAN, L. R. Visualizing automatic speech recognition and machine, January 2012. U.S. Patent 2012/0010869 A1.

[6] WALKER, N. R.; CEDERGREN, H.; TROFIMOVICH, P.; GATBONTON, E. Automatic speech recognition for call: A task-specific application for training nurses. *Can Mod Lang Rev/ La Revue canadienne des langues vivantes*, v.4, n.67, p.459–479, 2011.

[7] O'SHAUGHNESSY, D. Automatic speech recognition: History, methods and challenges. *Pattern Recogn*, v.41, p.2965–2979, 2008.

[8] RABINER, L.; SCHAFER, R. Theory and aplications of digital speech processing. Prentice Hall, Upper Saddle River, NJ, 2010.

[9] LEVINSON, S. E. Mathematical models for speech technology. John Wiley and Sons, Hoboken, NJ, 2005.

[10] O'SHAUGHNESSY, D. Speech communications: Human and machines. IEEE Press, Piscataway, NJ, 2000.

[11] VAPNIK, V. N. Statistical learning theory. John Wiley and Sons, Hoboken, NJ, 1998.

[12] VAPNIK, V. N. The nature of statistical learning theory, 2 ed. Springer, Berlin, Heidelberg, 1999.

[13] HASSOUN, M. H. Fundamentals of artificial neural networks. MIT Press, Cambridge, MA, 1995.

[14] HAYKIN, S. Neural networks: A comprehensive foundation. Prentice Hall, Upper Saddle River, NJ, 1999.

[15] DAS, S. Speech recognition technique: A review. *Int Eng Res Appl v.2*, n.3, p.2071–2087, 2012.

[16] BAGHDASARYAN, A. G.; BEEX, A. A. L. Automatic phoneme recognition with segmental hidden Markov models. In *Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, p. 569–574, 2011.

[17] CRUZ, R.; ORTIZ, A.; BARBANCHO, A. M.; BARBANCHO, I. Unsupervised classification of audio signals by self-organizing maps and bayesian labeling. In *Proceedings of the 7th international conference on Hybrid Artificial Intelligent Systems - Volume Part I* (Berlin, Heidelberg,), HAIS'12, Springer-Verlag, p. 61–70, 2012.

[18] VENKATESWARLU, R.; KUMARI, R. Novel approach for speech recognition by using self-organized maps. In *International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, IEEE, p. 215–222, 2011.

[19] WANG, Y.; VAN HAMME, H. Gaussian selection using self-organizing map for automatic speech recognition. In *Proceedings of the 8th International Conference on Advances in Self-Organizing Maps* (Berlin, Heidelberg,), WSOM'11, Springer-Verlag, p. 218–227, 2011.

[20] ANDERSON, J. An Introduction to Neural Networks. MIT Press, Cambridge, MA, 1995.

[21] MCCULLOCH, W.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *B Math Biophys* , v.5, p.115–133, 1943.

[22] KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biol Cybern*, v.43, p.59–69, 1982.

[23] KOHONEN, T. Self-Organizing Maps, third extended edition ed., Springer Series in Information Sciences, v.30, Springer, Berlin, Heidelberg, 2001.

[24] KOHONEN, T. The "neural"phonetic typewriter. *IEEE Computer*, v.21, n.3, p.11–22, 1988.

[25] BEHI, T.; AROUS, N.; ELLOUZE, N. New variant of the self organizing map in pulsed neural networks to improve phoneme recognition in continuous speech. *Int Comp Appl*, v.46, n.15, p.34–40, 2012.

[26] SOUZA JÚNIOR, A.; BARRETO, G.; VARELA, A. A speech recognition system for embedded applications using the SOM and TS-SOM networks. In *Self Organizing Maps: Applications and Novel Algorithm Design*, J. Mwasiagi, Ed. IN-TECH publishing, Viena, Áustria, p. 97–108, 2010.

[27] KOHONEN, T. The self-organizing maps. *Pr Inst Electr Elect* , v.78, p. 1464–1480, 1990.

[28] RITTER, H.; KOHONEN, T. Self-semantic maps. In *Biol Cybern*, v.61, p.241–254, 1989.

[29] BEZDEK, J. H.; PAL, N. R. A note on self-organizing semantic maps. *IEEE T on Neural Networ*, v.6, n.5, p.1029–1036, 1995.

[30] JURAFSKY, D.; MARTIN, J. H. Speech and language processing: An introduction to natural language processing, computational linguistic, and speech recognition. Prentice Hall, Upper Saddle River, NJ, 2008.

[31] STEVENS, S. S.; VOLKMANN, J. The relation of pitch frequency: A revised scale. *Am J Psychol*, v.53, n.3, p. 329–353, 1940.

[32] STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *Am J Psychol*, v.8, p.185–190, 1937.

[33] BORGET, B.; HEALY, M.; TUKEY, J. The quefrency alanysis of time series for echoes. *Proc Sym Time Series Analysis* p.209–243, 1963.

[34] OPPENHEIM, A.; SCHAFER, R. Discrete-Time Signal Processing. Prentice-Hall, Upper Saddle River, NJ, 1989.

[35] BARBOSA, P. A. At least two macrorhythmic units are necessary for modeling brazilian portuguese duration: emphasis on automatic segmental duration generation. *Cadernos de Estudos Linguísticos* , v.31, p.33–53, 1996.