REVISTA DE LETRAS

USO DE REDES NEURAIS AUTO-ORGANIZÁVEIS NA DETERMINAÇÃO DO CONHECIMENTO PROSÓDICO DE APRENDIZES BRASILEIROS DE INGLÊS

Ana Cristina Cunha da Silva*

Resumo

Os mapas auto-organizáveis de Kohonen (Self-organizing maps - SOM) *têm sido bastante úteis para a visualização de dados de alta dimensão por serem capazes de comprimir informações da entrada e ainda preservar relações geométricas importantes nos dados (KOHONEN, 1998; KOHONEN, 2001). A rede de Kohonen apresenta enormes vantagens para a análise de fenômenos linguísticos, entre elas a categorização do* input *de uma maneira não-supervisionada, o que é muito útil em vários níveis de categorização linguística. Essa rede neural foi utilizada para facilitar a visualização da formação de aprendizes por grau de similaridade no que se referia à transferência de acento de L1 para L2. A aplicação da rede SOM neste contexto foi inovadora. Os resultados das simulações evidenciaram a organização dos falantes em grupos bem definidos e apontam o uso dessa rede como ferramenta bem-sucedida no auxílio da determinação de nível de proficiência em língua estrangeira no futuro.*

Palavras-chave*: Mapas auto-organizáveis. Acento. Língua inglesa. Nível de proficiência.*

1 INTRODUÇÃO

O presente artigo apresenta uma parte dos resultados de uma pesquisa de doutorado (Cf. SILVA, 2010), que teve como objetivo principal apresentar as redes neurais auto-organizáveis de Kohonen como uma ferramenta poderosa na análise do conhecimento acentual em aprendizes brasileiros de língua inglesa.

O acento lexical, um dos elementos prosódicos mais importantes, é o maior responsável pelos casos de transferência linguística, sotaque predominante de L1 (primeira língua) e fossilização em L2 (segunda língua) de acordo com os estudos da área de fonologia da interlíngua (ARCHIBALD, 1994; SILVA, 2005; MAIRS, 1989).

Modelos formais de aprendizado acentual explicam que há duas hipóteses sobre como os aprendizes constroem seus sistemas acentuais: por meio de transferência do padrão de L1 ou a construção de um novo sistema. Muitos estudos empíricos têm se dedicado a responder a essa questão, todavia não conseguiram de forma tão bem sucedida.

Os modelos conexionistas, por sua vez, mostram-se como uma ótima alternativa para a análise das regularidades estruturais do sistema acentual dos aprendizes em fase inicial de aquisição de língua inglesa devido à sua grande capacidade de generalização. De acordo com Vainio (2001, p. 3),

> *As redes neurais são conhecidas por sua capacidade de generalização de acordo com a similaridade de seus insumos, mas também por distinguir diferentes saídas de padrões de entrada que são similares apenas na superfície. Como conseqüência, as redes têm o poder de prever, depois de uma fase de aprendizagem adequada, mesmo os padrões nunca vistos antes.*[1]

Estudos que se utilizaram de redes de Kohonen para simular a aquisição linguística (LI ET AL., 2004; GAUTHIER ET AL., 2009) oferecem resultados promissores por advogarem a favor de uma comparação razoável entre as topologias e mecanismos de aprendizagem.

Gauthier et al. (2009) usaram modelos conexionistas para explorar se e como as crianças poderiam

* Doutora em Linguística pela Universidade Federal do Ceará. Professora Adjunta da Universidade Estadual do Piauí. Membro do Grupo de Estudos sobre Linguagem e Pensamento / Cognição e Linguística da Universidade Federal do Ceará.

[1] Trecho original: "Neural networks are known for their ability to generalize according to the similarity of their inputs but also to distinguish different outputs from input patterns that are similar only on the surface. As a consequence, networks have the power to predict, after an appropriate learning phase, even patterns they have never seen before.

aprender foco prosódico[2] diretamente de *input* de fala contínua. Em três simulações utilizando redes neurais auto-organizáveis, os autores exploraram como o foco poderia ser aprendido a partir de sinais acústicos contínuos em Mandarim, que foram produzidos com tons lexicais co-ocorrentes e por vários falantes. Os resultados deste estudo mostraram que redes neurais não-supervisionadas podem desenvolver agrupamentos específicos de foco a partir de sinais de fala dinâmicos contínuos, produzidos por vários falantes em várias condições de tom lexical, o que pode eventualmente conduzir à aquisição do foco.

Li et al. (2004) simularam a aquisição lexical em crianças utilizando um modelo de rede neural auto-organizável. O objetivo principal da pesquisa era usar as propriedades de preservação topográficas do SOM (*Self-Organizing Map*) para estudar a emergência e a organização de categorias linguísticas ao longo dos estágios de aprendizagem lexical. O modelo captou uma série de fenômenos importantes que ocorriam na aquisição lexical inicial das crianças, pois permitiu a representação de um ambiente em mudança linguística dinâmica na aprendizagem de línguas.

No entanto, há ainda uma escassez de estudos sobre a aplicação de redes neurais artificiais para visualizar a forma como o conhecimento de aprendizes de inglês em relação ao acento lexical e acento frasal estão organizados no que se refere à aquisição de acento de L2 e transferência do padrão acentual de L1 para L2.

O objetivo aqui apresentado neste artigo foi investigar se e como a rede de Kohonen seria capaz de evidenciar, por meio da formação de agrupamentos, grupos bem definidos de falantes que possuem similaridades quanto à transferência de padrões acentuais na aprendizagem de inglês como língua estrangeira.

A hipótese básica testada foi se a parametrização do sinal de fala por coeficientes LP (*Linear Predictive Coding)* ou MFC (*Mel-frequency cepstral coefficients)* como codificação do *input* da rede era eficiente na categorização dos falantes, justamente por esses coeficientes condensarem energia, frequência fundamental (F_0) e banda de frequência, informações genuinamente ligadas ao acento.

2 FUNDAMENTAÇÃO TEÓRICA

O processo de extração de características do sinal de fala é uma importante etapa na abordagem conexionista do processamento da fala e tomadas de decisão e classificação da rede neural. Essa etapa consiste na utilização de técnicas de transformação do sinal de fala original em uma representação matemática que permita a identificação de uma dada elocução, e é geralmente representado por um conjunto de vetores de características (SOUZA JR., 2009).

A codificação linear preditiva (*Linear Predictive Coding - LPC*) é uma técnica de parametrização e processamento do sinal da fala amplamente utilizada para a obtenção de coeficientes cepstrais nas áreas de reconhecimento automático de voz e sistemas de síntese texto-fala. Já os coeficientes mel-cepstrais (*Mel-frequency cepstral coefficients – MFCC)* consistem na representação de um espectro de potência de curta duração de um som, baseado na transformada de um cosseno linear de um espectro de força sobre uma escala mel de frequência não-linear.

Optou-se, então, pela análise de predição linear porque os coeficientes LP conseguem extrair a intensidade e a frequência do sinal de fala. Essas duas características são portadoras e indicadoras do elemento prosódico "acento". No inglês, o acento é a junção de três fatores perceptivos correlacionados: 1) quantidade/duração (medida em ms) relacionada com o tamanho da sílaba; 2) intensidade (medida em dB) relacionada à amplitude média alta e 3) altura (medida em Hz), ou seja, o valor de F_0 mais elevado na elocução.

Um mapa auto-organizável (SOM - *Self-Organizing Map*) é um tipo de rede neural artificial treinada por aprendizagem competitiva não-supervisionada baseada em princípios de auto-organização de sistemas que permite a representação de dados multidimensionais em espaços de dimensões menores (KOHONEN, 2001). Os mapas de Kohonen são utilizados usualmente como ferramenta de visualização de dados, permitindo que interrelações existentes em conjuntos complexos de dados possam ser percebidas e classificadas.

A matriz-U (matriz de distância unificada) é a ferramenta de visualização utilizada para analisar os resultados gerados pela rede SOM através da demonstração da distância de estruturas dos dados de entrada. Seus métodos têm sido amplamente usados para agrupamento de conjuntos de dados de alta dimensão. Metaforicamente falando, ela apresenta uma "paisagem" das relações de distância dos dados de entrada no espaço dos dados.

Na próxima seção estão descritos o *corpus*, o perfil dos participantes, os procedimentos e a metodologia do processamento dos dados de entrada da rede.

3 METODOLOGIA

O *corpus* desta pesquisa é composto das gravações das entrevistas com 30 alunos de uma instituição de ensino superior da cidade de Fortaleza-CE, com idade entre 18 e 25 anos, todos brasileiros, de ambos os sexos e que não tinham feito nenhuma viagem para um país de língua inglesa até o momento da entrevista.

Antes do processo de simulação na rede neural e com o objetivo de ajudar na interpretação do mapa, decidiu-

[2] O foco é uma função comunicativa, que serve para colocar ênfase em uma determinada parte de um enunciado.

-se alocar os 30 participantes em 5 níveis de desenvolvimento distintos, valendo o critério de tempo de exposição ao idioma[3].

Com base neste fato, estabeleceu-se como critério circunstancial de classificação e organização dos indivíduos a quantidade de horas-aula acumulada na disciplina de língua inglesa obtida por meio de entrevistas e questionários respondidos pelos participantes.

As elocuções dos participantes foram gravadas e digitalizadas no programa *Sound Forge* versão 5.0 em arquivos de áudio tipo WAV (extensão .wav), a uma taxa de amostragem de 44.100 mil amostras por segundo (44.1 KHz) e resolução de 16 bits, mono.

Após essa fase, procedeu-se à segmentação de cada frase e consequentemente cada palavra que representava o item lexical a ser investigado (*e.g. object, separate, desert, etc*).

O sinal de fala não pode ser usado diretamente para alimentar a rede por conter milhares de amostras, o que tornaria seu processamento muito lento e também por ser muito ruidoso, o que dificulta sobremaneira a extração de conhecimento (SOUZA JR., 2009, p. 10).

A solução é representá-lo numericamente em um conjunto de coeficientes obtidos a partir da aplicação de técnicas matemáticas, tais como coeficientes de predição linear e/ou coeficientes mel-cepstrais, ao sinal de fala subdividido em vários frames. Assim, conjuntos de vetores de coeficientes passam a representar numericamente o sinal de fala dos aprendizes. Diz-se, então, que o sinal de fala é parametrizado pelos coeficientes LP/MFC.

Os vetores de coeficientes LP ou MFC são então organizados ao longo das linhas de uma matriz de coeficientes. Para o exemplo dado, são gerados 100 vetores de coeficientes, um vetor para cada frame, que correspondem a 100 linhas da matriz de coeficientes. O número de colunas dessa matriz é igual ao número de coeficientes extraídos da análise LPC ou MFCC (no caso, 10, 15 ou 20). As matrizes de coeficientes são extraídas e salvas em uma planilha do tipo Excel, sendo posteriormente convertidas em arquivos de texto (ASCII) para serem usadas no treinamento da rede neural.

É importante enfatizar que cada palavra pronunciada por um falante gera uma matriz de coeficientes. Logo, para identificar qual falante pronunciou qual palavra, "faz-se necessária uma etapa de rotulação dos dados. Partindo do arquivo-texto em que uma certa matriz de coeficientes está armazenada, adicionou-se uma última coluna a esta matriz contendo um rótulo para identificar aquele conjunto de dados". Por fim, os arquivos de texto relativos à elocução de uma palavra específica por todos os falantes são concatenados em um só arquivo com o auxílio do Matlab.

Os rótulos podem identificar tanto o falante, quanto a categoria linguística na qual se insere a palavra pronunciada. A rotulação é feita a fim de relacionar os conjuntos de vetores a seus locutores.

A rede SOM, assim como a maioria dos modelos conexionistas, necessita que alguns parâmetros sejam pré-especificados de modo a garantir um correto treinamento da mesma. Kohonen (2001) ressalta que a determinação do tamanho do SOM e dos parâmetros de aprendizagem é um processo empírico, baseado na experiência do usuário e em métodos de tentativa e erro. Após alguns testes preliminares, foi escolhido um mapa bidimensional, com arranjo hexagonal (geralmente utilizado por ter uma boa capacidade de projeção de dados), função de vizinhança gaussiana, iniciação linear dos pesos e aprendizagem em lote. Conforme será mostrado, estas especificações para a arquitetura escolhida mostraram-se aparentemente adequadas para tratar os fenômenos em questão.

Vale ressaltar que são utilizados os mesmos vetores de características, porém com rótulos diferentes. A análise denominada Multi-Rótulo (*Multi-Label Analysis*) é considerada uma abordagem nova na análise de agrupamentos para este fim e tem como objetivo último verificar qual rotulação é mais adequada ao tipo de parametrização usada (LPC/MFC). A estratégia de análise Multi-Rótulo permite, em princípio, verificar que propriedades linguísticas do sinal de fala são melhor codificadas nos coeficientes LP/MFC.

Quando aplicado ao problema de interesse, tem-se que o processo de simulação da rede SOM e a análise do resultado do treinamento envolvem os seguintes passos:

1. Inicialização e treinamento (aprendizagem) da rede;
2. Avaliação da qualidade do mapa gerado usando os índices - erro de quantização (E_Q) e erro topológico (E_T);
3. Geração e armazenamento da matriz-U e mapas rotulados para cada resultado de treinamento;
4. Validação dos agrupamentos através do índice Davies-Bouldin (DB);
5. Tabulação no programa Excel de todos os resultados das medidas de desempenho da rede (erro de quantização e erro topológico).

Um conjunto de quatro simulações foi rodado com parâmetros que variavam conforme a necessidade de ajuste com o fenômeno em questão. Todas as simulações foram conduzidas numa rede neural auto-organizável bidimensional, hexagonal, com função de vizinhança gaussiana e aprendizagem em lote. Apresentaremos a seguir os resultados da simulação que verificou se a rede conseguia orga-

[3] Para maiores informações e critérios mais detalhados sobre a determinação de nível de proficiência, consultar o ACTFL (*American Council on the Teaching of Foreign Languages)* Proficiency Guidelines em http://www.actfl.org/i4a/pages/index.cfm?pageid=1 bem como The Common European Framework of Reference em http://www.coe.int/t/dg4/linguistic/cadre_en.asp.

nizar os aprendizes em função da transferência do padrão acentual de L1 para L2. Essa simulação contribui também para resolver questões de determinação de nível de proficiência linguística.

4 SIMULAÇÃO E DISCUSSÃO DOS RESULTADOS

A presente simulação objetivou investigar se a rede seria capaz de evidenciar, através da formação de agrupamentos, os processos de transferência do padrão acentual do português brasileiro[4] para o inglês[5], mais especificamente o acento primário nas categorias lexicais substantivo, verbo e adjetivo.

Especificou-se a rede em uma topologia bidimensional com 25 neurônios (5x5) em vizinhança hexagonal, já que se teria que vislumbrar o aparecimento de dois grupos somente – o grupo que transferiria o padrão do português para o inglês e o grupo que não transferiria. Os parâmetros de treinamento da rede são mostrados na tabela 1 abaixo:

As duas figuras abaixo (matriz-U - figura 1 - e mapa rotulado – figura 2) são relativas à formação do mapa que teve como entrada uma matriz de dados de falantes rotulados por um número (semestre e ranqueamento) acrescido do rótulo "er" quando o aprendiz erra a pronúncia ao transferir o padrão de L1 para L2. Vê-se claramente a sinalização de formação de dois grupos: um que transfere o padrão acentual (grupo maior e mais coeso) e um que não transfere (grupo mais disperso).

No mapa rotulado com a identificação dos aprendizes quanto ao semestre adicionado do padrão de transferência para a palavra *object* (verbo) usando coeficientes LP 10, há uma correspondência entre as topologias dos respectivos mapas. O mapa rotulado corrobora a formação do mesmo número de agrupamentos nas mesmas áreas de formação na matriz-U. Cabe ressaltar que a interpretação do mapa de Kohonen se vale das duas formas de visualização para garantir uma análise fidedigna do conjunto de dados.

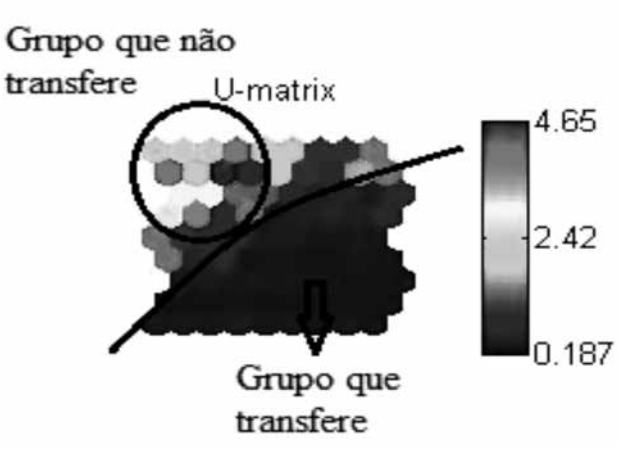

Figura 1 - Matriz-U evidenciando a formação de 2 grupos; um que contém indivíduos que transferem o acento primário e outro que não transfere (Fonte: Silva, 2010).

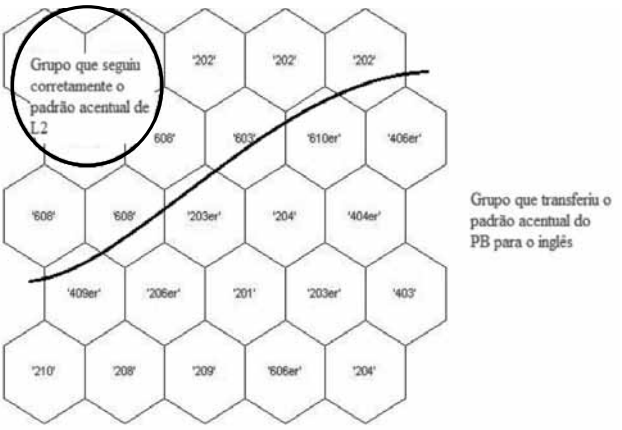

Figura 2 - Mapa rotulado com "erro" para os conjuntos de dados referentes à palavra object (verbo), usando LPC 10, treinados com dados dos 30 aprendizes. (Fonte: Silva, 2010).

Tabela 1 – Parâmetros de treinamento da rede SOM para a presente simulação (Formação de agrupamentos em função da transferência do padrão acentual de L1 para L2).

Parâmetro	**Opção Escolhida**
Tipo de treinamento	*batch (em lote)*
Topologia do mapa	*Bidimensional (planar)*
Dimensão do mapa	*5 x 5*
Estrutura da vizinhança	*Hexagonal*
Número máximo de clusters	*k = 10*
Número de repetições	*1*
Épocas	*250 (50 – rough training, 200 – fine tuning)*
Raio vizinhança inicial	*2*
Raio vizinhança final	*1*
Taxa de aprendizagem inicial	*Não é utilizado no modo de treinamento batch*
Taxa de aprendizagem final	*Não é utilizado no modo de treinamento batch*

Fonte: Silva (2010).

[4] Para mais informações sobre o padrão acentual do português consultar Bisol (1994) e Lee (1994).
[5] Para mais informações sobre o padrão acentual do inglês consultar Prator e Robinett (1985) e Kreidler (1989).

A fim de obter uma compreensão maior da organização dos dados, foram seguidos dois critérios de rotulação da rede. No primeiro, mantiveram-se os números de identificação dos falantes sem nenhuma informação sobre seus erros na computação do acento de L2, ou seja, a transferência do padrão de L1 para L2. No segundo critério, adicionou-se o rótulo "er" para indicar o erro na computação do acento primário de palavra daquele falante.

Os vetores de atributos carregam informação que pode ter uma interpretação ampla. Todavia, as categorias linguísticas que emergiram na rede puderam dar margem de interpretação sobre o fenômeno de transferência acentual. Tal fato não parece ser fruto do acaso, pois a rede organiza os dados por similaridade e os vetores referentes aos indivíduos que cometeram erros de substituição vocálica e consonantal, bem como os que realizaram o processo de inserção e apagamento de segmentos vocálicos, também emergiram durante o processo de treinamento e atualização dos pesos sinápticos da rede.

Os neurônios que aparecem sem a notação do rótulo "er" representam os vetores de peso dos neurônios vencedores dos grupos de aprendizes que não transferiram o padrão acentual do português para o inglês. A análise dos rótulos de neurônios vencedores que ficaram em segundo lugar pôde garantir a segregação correta da rede, como mostra a figura abaixo.

Durante o preparo do *input* da rede, rotulou-se o conjunto de dados com diferentes codificações - rótulo alfanumérico (e.g.: 201er), rótulo "transfer" e "no transfer", e rótulo 'er'. Em todos os resultados para a palavra *object* (LPC 10), a matriz-U seguiu o mesmo padrão de segregação, não importando o rótulo dado ao conjunto de dados. Tal fato faz chegar a uma conclusão importante: as linhas de exemplos de vetores de atributos são única e exclusivamente processadas sem levar em consideração a variação do título do rótulo durante a convergência da rede.

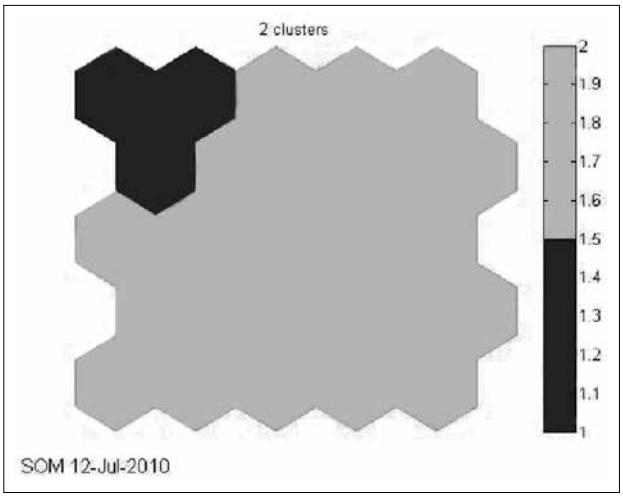

Figura 4 - Mapa colorido sugerindo a formação de dois grupos, conforme indicação do índice Davies-Bouldin (Fonte: Silva, 2010).

O mapa colorido logo acima na figura 4 corrobora o mapa rotulado e a matriz-U. É importante ressaltar que dentro desses dois grupos grandes (o grupo que transfere o padrão acentual e o que não transfere) pode haver subgrupos (subclusters) que ao serem minuciosamente analisados e

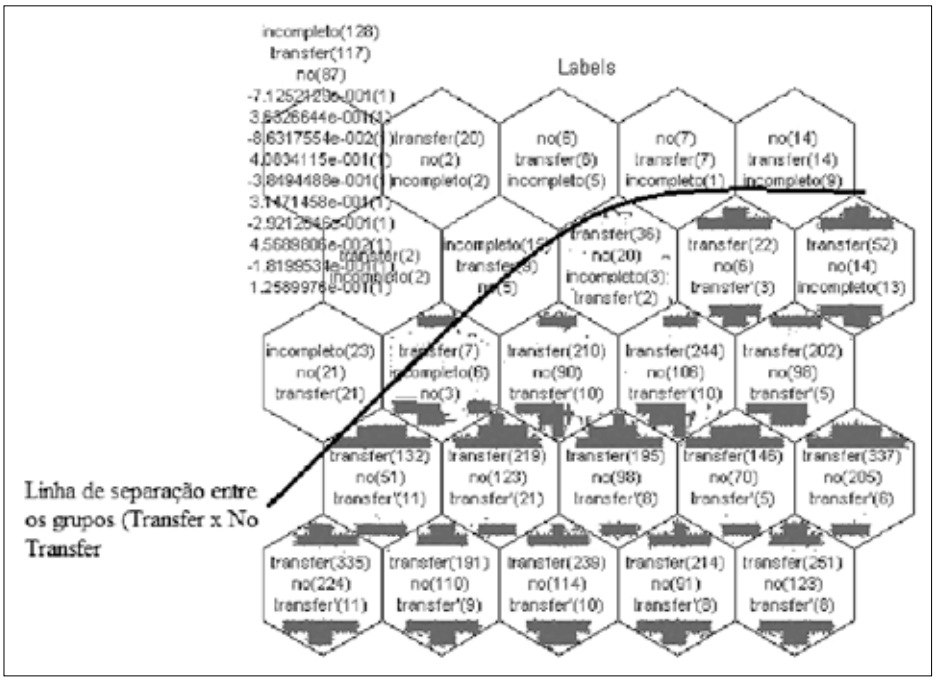

Figura 3 - Mapa rotulado com informações sobre a freqüência das classes (rótulos) para cada neurônio vencedor. (Fonte: Silva, 2010).

de forma isolada podem revelar informações ricas para a análise linguística das elocuções dos aprendizes bem como contribuir para a compreensão da organização do conjunto de dados.

Após o término da presente simulação com a palavra *object* (verbo), ainda foram feitas simulações para mais 19 itens lexicais restantes com conjunto de dados extraídos usando coeficientes MFC de diferentes tamanhos. Em todos os outros resultados das simulações seguintes, observou-se a emergência do mesmo padrão de atividade neural.

5 CONCLUSÕES

A pesquisa contou com a modelagem da aprendizagem de padrões linguísticos de L2, utilizando o sinal de fala numericamente codificado como *input* de uma rede neural não supervisionada. Partiu-se da hipótese de que a parametrização do sinal de fala por meio da abordagem de coeficientes LP ou MFC seria eficiente na categorização dos falantes por características prosódicas.

A nova proposta teórico-metodológica apresentada aqui por meio da Análise Multi-rótulo (*Multi-label Analysis*) é inédita e a ela estão atreladas as vantagens do uso da rede de Kohonen (rede neural não-supervisionada), que também tem utilização inédita para esta área investigada, assim como o uso da técnica de LPC e MFCC na codificação do *input* da rede.

A segregação do mapa em regiões de agrupamentos densos sugeriu que os sujeitos foram agrupados por características fonético-acústicas semelhantes. De acordo com as rodadas de experimentos, confirmou-se que a rede discriminou os falantes por características prosódicas e os organizou de acordo com as similaridades relativas a essas características.

A simulação apresentada tinha como objetivo investigar se a rede seria capaz de captar as diferenças e semelhanças nos aprendizes quanto à transferência de padrões acentuais de L1 para L2 e assim agrupá-los por similaridade. Pôde-se perceber claramente a emergência de padrões similares em todos os testes. Em resumo, a rede foi capaz de agrupar os falantes que transferiram os padrões acentuais do português para o inglês e, de forma análoga, também foi capaz de segregar os grupos no mapa.

A rede de Kohonen também pode vir a ser usada para avaliar a distância que um grupo de aprendizes está de um grupo de falantes nativos. De posse de medidas (Distância Euclidiana) pode-se comparar os segmentos realizados em cada neurônio vencedor ou nos agrupamentos formados com os dados de falantes nativos, um claro sinal de que essa rede pode vir a ser aplicada como ferramenta no contexto de determinação de nível de proficiência em língua estrangeira.

6 REFERÊNCIAS

ARCHIBALD, J. A formal model of prosodic learning. *Second Language Research* 10, 3, 215-240, 1994.

BISOL, L. O acento e o pé binário. *Letras de Hoje*. Porto Alegre, v. 29, n. 4, p. 25-36, dez. 1994.

GAUTHIER, B; SHI, R; XU, YI. Learning Prosodic Focus from Continuous Speech Input: A Neural Network Exploration. *Language Learning and Development*, 5: 94–114, 2009.

KOHONEN, T. The self-organizing map. *Neurocomputing*, 21, 1–6, 1998.

_____. *Self-organizing Maps*. 3rd ed. Berlin: Springer, 2001.

KREIDLER, C. W. *The pronunciation of English:* A course book in phonology. Oxford: Blackwell, 1989.

LEE, S. H. A regra do acento do português. *Letras de Hoje*, Porto Alegre. V. 29, nº 4, p. 37-42, dezembro 1994.

LI, P.; FARKAS, I.; MacWHINNEY, B. Early lexical development in a self-organizing neural network. *Neural networks* 17, 1345 – 1362, 2004.

MAIRS, J. L. Stress assignment in interlanguage phonology: an analysis of the stress system of Spanish speakers learning English. In: Gass, M & Schatcther, J. (orgs.) *Linguistic Perspectives on Second Language Acquisition*. Cambridge, USA: Cambridge University Press, 1989.

PRATOR, Jr., C. H.; ROBINETT, B. W. *Manual of American English pronunciation.* 4. ed, Orlando: Harcourt Brace & Company, 1985.

SOUZA JR. A. H. de. *Avaliação de redes neurais auto-organizáveis para reconhecimento de voz em sistemas embarcados*. Dissertação (Mestrado em Engelharia de Teleinformática) - UFC. Fortaleza, 2009.

SILVA, A. C. C. da. *A produção e a percepção do acento em pares mínimos de língua inglesa por aprendizes brasileiros*. Dissertação (Mestrado em Linguística) – UFC, Fortaleza, 2005.

_____. *O uso de redes neurais auto-organizáveis para a análise do conhecimento acentual em aprendizes brasileiros de língua inglesa*. (Doutorado em Linguística) – UFC, Fortaleza, 2010.

VAINIO. M. *Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis.* Doctorate dissertation, 2001.