Ensino de pronúncia mediado por computador: uma nova proposta de auto-avaliação para o aprendiz

Ana Cristina Cunha da Silva[1]
UESPI

Resumo: O ensino de pronúncia de alguns elementos prosódicos de língua inglesa tinha um papel pouco relevante até certo tempo atrás. Todavia, o cenário passou a mudar devido ao crescimento das tecnologias de reconhecimento automático de fala. Temas relacionados à inteligibilidade, ao sotaque e às variedades de língua inglesa faladas no mundo são considerados centrais para a resolução de muitas questões na lingüística aplicada. Pesquisas sobre o Ensino de Pronúncia Mediado por Computador indicam que tanto pesquisadores quanto professores de pronúncia vem fazendo maior uso dele para desenvolver novas teorias e práticas metodológicas que se alinhem com a realidade acústica da fala. O presente artigo apresenta uma proposta de auto-avaliação para o aprendiz de LE ao se basear nas vantagens de uma rede neural artificial, entre elas, a categorização do *input* de uma maneira não-supervisionada. A implementação dessa rede visa a fornecer um diagnóstico rápido sobre a *performance* do aprendiz ao alocá-lo em grupos de falantes com características lingüísticas semelhantes. A rede neural auto-organizável de Kohonen aliada às técnicas de parametrização do sinal de fala advindo da elocução de aprendizes pode fornecer um *feedback* avaliativo válido durante o processo de aprendizagem de pronúncia. A proposta de auto-avaliação consiste em um aplicativo para geração de mapas de classificação de desempenho acoplado a um ambiente computacional já existente que ofereça um *feedback* mais confiável ao aprendiz sobre a evolução do seu treinamento do conhecimento de pronúncia de língua inglesa.

Palavras-chave: ensino de pronúncia, rede neural auto-organizável, speech signal

Abstract: The teaching of pronunciation of prosodic elements of the English language had a role of little relevance some years ago. However, the picture began to change due to the growth of technologies for automatic speech recognition. Topics related to intelligibility, accent and varieties of English spoken in the world are considered central to the resolution of many issues in applied linguistics. Research on the Computer Assisted Pronunciation Teaching indicates that both researchers and teachers of pronunciation have been making greater use of it to develop new theories and methodological practices conjugated with the reality of speech acoustics. This paper presents a self-assessment proposal for the apprentice to the LE based on the advantages of an artificial neural network, among them the categorization of the input in a non-supervised fashion. The implementation of this network aims to provide a quick diagnostic on the learner's performance to allocate it in groups of speakers with similar linguistic features. The Kohonen network self-organizing techniques combined with the parameterization of the speech signal collected from the learners' speech signal can provide a valid evaluative feedback during the process of English learning pronunciation. The proposed self-assessment consists of an application for generating maps of learner's performance

[1] cris0708@gmail.com

classification coupled with an existing computing environment that provides a more reliable feedback to the learner on the evolution and training of their pronunciation knowledge.

Keywords: pronunciation teaching, self organizing neural net, speech signal.

1. Introdução

Nas últimas décadas, pesquisas sobre o *Ensino de Pronúncia Mediado por Computador (Computer-assisted Pronunciation Teaching)* (LEVIS, 2008; CHUN, 1998; MOLHOLT, 1988) indicam que tanto pesquisadores quanto professores de pronúncia vêm fazendo um maior uso da tecnologia computacional para desenvolver novas teorias e práticas metodológicas que se alinhem com a realidade acústica da fala. A teoria sobre processamento do sinal acústico da fala e as ferramentas de análise da fala são de fundamental importância para o desenvolvimento de recursos tecnológicos que auxiliem na aprendizagem de línguas estrangeiras.

Há uma infinidade de ferramentas e recursos tecnológicos para o treino de pronúncia de língua inglesa Atualmente o mercado disponibiliza uma ampla gama de opções para o aprendizado de pronúncia de língua inglesa: softwares que podem ser adquiridos por meio de licença (Pronunciation Power®. Ver figura 1) ou são oferecidos gratuitamente e que possibilitam a análise da síntese da fala, análise da forma da onda e reconhecimento da fala. Esses softwares podem ser instalados em pc's ou podem ser utilizados *on-line*. Há também os aplicativos que funcionam nos sistemas operacionais de celulares. Em todos eles, o usuário aprende e desenvolve o inglês de acordo com o seu próprio ritmo e disponibilidade.

Todavia, um questionamento importante entra em cena: essas ferramentas para treinamento de pronúncia oferecem um *feedback* avaliativo aos aprendizes? Ferramentas tecnológicas há muito vêm sendo usadas para lançar luz sobre categorias fonológicas e, graças a pesquisas recentes em reconhecimento de fala, elas podem ser tranquilamente aplicadas no treinamento específico de algumas habilidades linguísticas.

Ao se conceber e planejar ferramentas tecnológicas de ensino de pronúncia, é natural pensar em algumas estratégias metodológicas a serem desenvolvidas a fim de fornecer aos aprendizes maior sucesso em seu processo de aprendizagem. Uma dessas estratégias seria a proporcionar ao aluno mecanismos de auto-avaliação.

Figura 1: interface do software Pronunciation Power.

2. A proposta de auto-avaliação para o aprendiz

É tarefa incansável dos pesquisadores da área de processamento de sinais da fala alinhar teorias de reconhecimento de fala com práticas metodológicas e realidade acústica da fala. Ainda há pouca aplicação de redes neurais artificiais na aquisição de L2 e, mais especificamente, na concepção de ferramentas de auto-avaliação de desempenho e habilidades lingüísticas.

Por isso, propomos o aproveitamento de uma interface pré-existente, como por exemplo um software ou qualquer outro ambiente de aprendizagem virtual de pronúncia de língua inglesa, para implementar o acoplamento de uma rede neural artificial (Rede de Kohonen) para o "julgamento" do desempenho linguísticos dos aprendizes. A essa nova ferramenta chamamos de JUMP (JUDGING, UNDERSTANDING AND MAPPING PRONUNCIATION PATTERNS).

A proposta aqui apresentada se baseia nos resultados preliminares de Silva (2010) e vem sendo desenvolvida em parceira com o Departamento de Engenharia de Teleinformática da Universidade Federal do Ceará e supervisão do Prof. Dr. Guilherme de Alencar Barreto (DETI-UFC). Os resultados da pesquisa citada logo acima objetivavam contribuir diretamente para a (re)formulação de teorias sobre a avaliação de desempenho e a determinação de nível de proficiência linguística em LE, dada a competência da rede neural de Kohonen em organizar

os grupos de indivíduos por características fônicas semelhantes. Essa rede neural provou ser uma ferramenta de visualização razoável para analisar a formação de agrupamentos a partir de simulação de representação de categorias lexicais. A presente proposta visa disponibilizar essa rede como meio para auxiliar, de forma direta e indireta, na determinação de níveis de proficiência linguística de aprendizes de L2.

Tal objetivo é totalmente viável porque a rede neural provou ser capaz de fazer generalizações sobre os processos fonológicos relacionados às vogais, consoantes e acento de palavra bem como discriminar os coeficientes LP e MFC dos aprendizes, assim segregando os grupos de aprendizes como base nas características fônicas semelhantes, um sinal claro de indicação de nível de proficiência linguística ou, no mínimo, de descrição de aquisição de L2. Na seção a seguir apresentaremos de forma sucinta como essa rede funciona.

3. A rede neural auto-organizável

O objetivo desta seção é apresentar os princípios básicos que regem a rede neural artificial auto-organizável, também chamada de mapa auto-organizável.

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).

O modelo de Kohonen torna possível o agrupamento de padrões similares em certas áreas da saída da rede, como também a identificação de padrões com características comuns. O agrupamento acontece devido ao fato de o mapa ser formado por neurônios (unidades mínimas de processamento).

O desenvolvimento de mapas auto-organizáveis como modelos neurais foi motivado a partir da característica distintiva de organização cerebral local em mapas topologicamente ordenados (mapas corticais do cérebro humano). O mapa auto-organizável foi criado para ser um sistema capaz de aprender por meio de auto-organização de forma neurobiologicamente inspirada.

Sob esta perspectiva, o aspecto mais relevante do sistema nervoso a ser levado em conta é a sua capacidade de formação topográfica de mapa, fato que o difere de outras redes neurais. Como dito por Kohonen (2001), “a localização espacial ou as coordenadas de uma

célula na rede correspondem a um domínio particular de padrões de sinais de entrada". O mapa computacional constitui-se, assim, como uma pedra fundamental básica na infraestrutura de processamento da informação do sistema nervoso.

O objetivo de aprendizagem em um mapa auto-organizável é fazer com que diferentes partes da rede respondam de forma similar a certos padrões de entrada. Isto é parcialmente motivado pela forma como outras informações visuais, auditivas ou outras sensoriais são manipuladas em partes diferentes do córtex cerebral no cérebro humano (KOHONEN, 2001).

A maioria dos algoritmos de treinamento de redes neurais é inspirada, de forma direta ou indireta, pela lei de Hebb, que estipula que a intensidade de uma ligação sináptica entre dois neurônios aumenta se ambos são excitados simultaneamente. Numa rede neural baseada em aprendizagem Hebbiana, vários neurônios da camada de saída podem estar simultaneamente ativos. Já na aprendizagem competitiva, os neurônios da camada de saída "competem" entre si pelo direito de responder, ou seja, de permanecerem ativos para um dado estímulo de entrada. Ao final desta competição, apenas uma unidade permanecerá ativa, como resposta àquela informação da entrada. Com um único neurônio de saída sendo ativado a cada iteração, essa propriedade torna o algoritmo apropriado para descobrir características estatísticas salientes dentro dos dados que podem ser usados para classificar um conjunto de padrões de entrada (CASTRO, 2006).

É necessária uma breve descrição do algoritmo SOM original, introduzido por Kohonen (2001). Denota-se $\boldsymbol{m}_i(t) \in R^p$ como vetor de peso do iésimo neurônio no mapa. Depois de inicializar todos os pesos de vetores aleatoriamente ou de acordo com alguma heurística, cada iteração do algoritmo SOM envolve dois passos. Um primeiro, para um dado input de vetor $x(t) \in | R^p$, encontra-se o neurônio vencedor atualizado i*(t), como se segue:

$$i^*(t) = \arg\min_{\forall i}\{\|\mathbf{x}(t) - \mathbf{m}_i(t)\|\}.$$

eq. 1

Em que t denota as iterações do algoritmo. Então, é necessário ajustar os pesos de vetores do neurônio vencedor a aqueles que estão em sua vizinhança:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \eta(t)h(i^*, i; t)[\mathbf{x}(t) - \mathbf{m}_i(t)],$$

eq. 2

Em que 0 < η (t) < 1 é a taxa de aprendizagem e *h (i*, i; t)* é a ponderação de função gaussiana que limita a vizinhança do neurônio vencedor:

$$h(i^*, i; t) = \exp\left(-\frac{\|\mathbf{r}_i(t) - \mathbf{r}_{i^*}(t)\|^2}{2\sigma^2(t)}\right),$$

eq. 3

Em que $\mathbf{r}_i$(t) e $\mathbf{r}^*_i$(t) são respectivamente as posições dos neurônios *i* e *i** em uma matriz de saída pré-definida em que os neurônios estão organizados em nós, e σ (t) > 0 define o raio da função de vizinhança no tempo t. A fim de garantir a convergência do algoritmo, η (t) e σ (t) decaem exponencialmente no tempo de acordo com as seguintes expressões:

$$\eta(t) = \eta_0\left(\frac{\eta_T}{\eta_0}\right)^{(t/T)} \quad \text{e} \quad \sigma(t) = \sigma_0\left(\frac{\sigma_T}{\sigma_0}\right)^{(t/T)},$$

eq. 4

Em que η_0(σ_0) e η_T (σ_T) são os valores iniciais e finais de η (t) e σ (t).

A matriz-U é a ferramenta canônica para a demonstração da distância de estruturas dos dados de entrada em um mapa auto-organizável. Seus métodos têm sido amplamente usados para agrupamento de conjuntos de dados de alta dimensão.

Após as duas fases da rede, a **formação** do mapa, usando-se exemplos de entrada por meio de quantização vetorial e o **mapeamento**, é possível visualizar a disposição dos neurônios vencedores na grade de neurônios, como mostra a Figura abaixo:

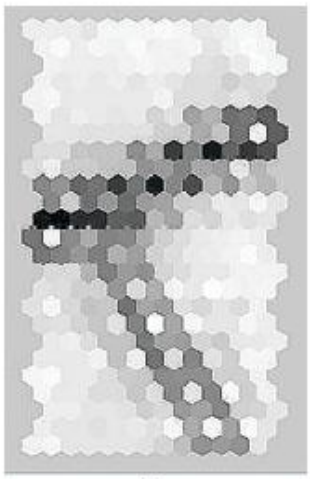

Figura 2: Exemplo de matriz-U matriz bidimensional com tons de cinza.

Fonte: BOSCARIOLI (2008, p. 63).

Em uma matriz-U bidimensional, pode-se visualizar região de agrupamento e região de separação. Os valores representados na matriz-U para um nó particular correspondem a distâncias entre o vetor de pesos deste nó e os vetores de pesos de seus vizinhos mais próximos. Numa grade quadrada, por exemplo, podem-se considerar os 4 ou 8 nós próximos ao BMU ou seis nós em uma grade hexagonal, conforme figura 4.9.

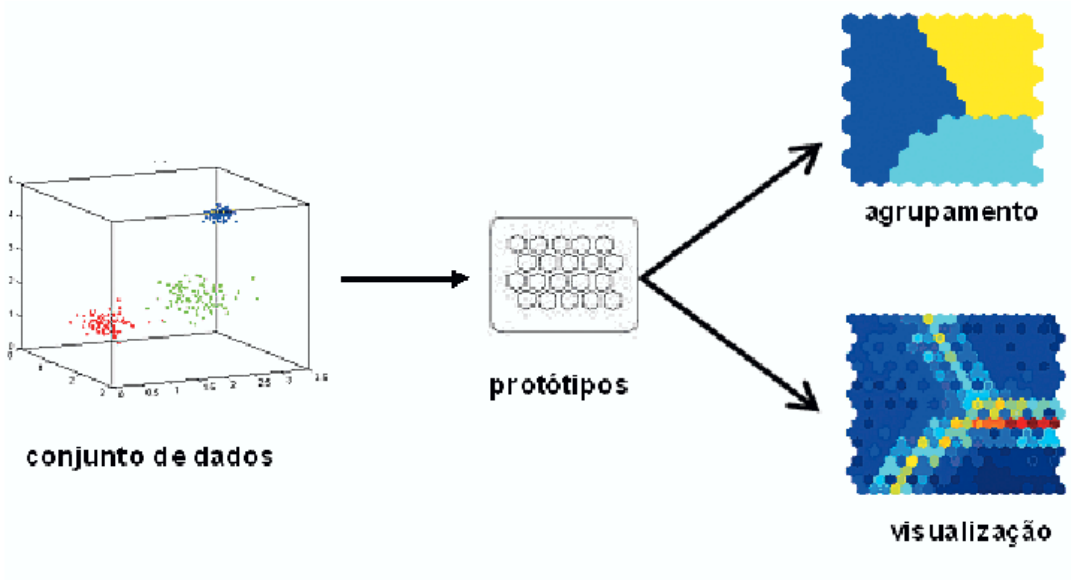

Figura 3: Análise de dados a partir do SOM.

Fonte: BOSCARIOLI (2008, p. 70).

4. O input da rede

A entrada da rede neural, também chamada de vetores de atributos ou de características, será processada pelos neurônios de modo a gerar uma representação compacta da informação presente nos dados originais. No caso da rede SOM, esta produzirá uma representação que preserva, entre neurônios vizinhos, relações de proximidade

(vizinhança) presentes nos dados originais. A representação gerada pela rede SOM é codificada nos vetores de pesos dos neurônios e pode ser visualizada através da matriz-U.

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). Além disso, o sinal de fala não é estacionário no tempo. Para a análise do sinal de fala, deve-se recorrer à segmentação do sinal e, nesse caso, considerá-lo *quasi* estacionário por partes.

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.

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.

5. A formação de agrupamentos

A fim de analisar a formação de grupos (clusters) nos mapas em função do conhecimento prosódico (acento) do aprendiz e do seu tempo de exposição ao idioma apresentamos brevemente os detalhes de uma simulação.

Com o objetivo de ajudar na interpretação do mapa, decidiu-se alocar 30 participantes em 5 níveis de desenvolvimento distintos, com base no critério de tempo de exposição ao idioma.

Os mapas foram inicializados com 100 neurônios, um número escolhido heuristicamente por ser grande o suficiente para dar suporte à formação de grupos, mas pequeno o suficiente para evitar *overfitting* (superespecialização do aprendizado).

Logo abaixo está a matriz-U resultante do conjunto de dados formado pelos vetores de características dos 30 aprendizes e mais 4 falantes nativos pronunciando a palavra *object* (verbo), usando 10 coeficientes LP.

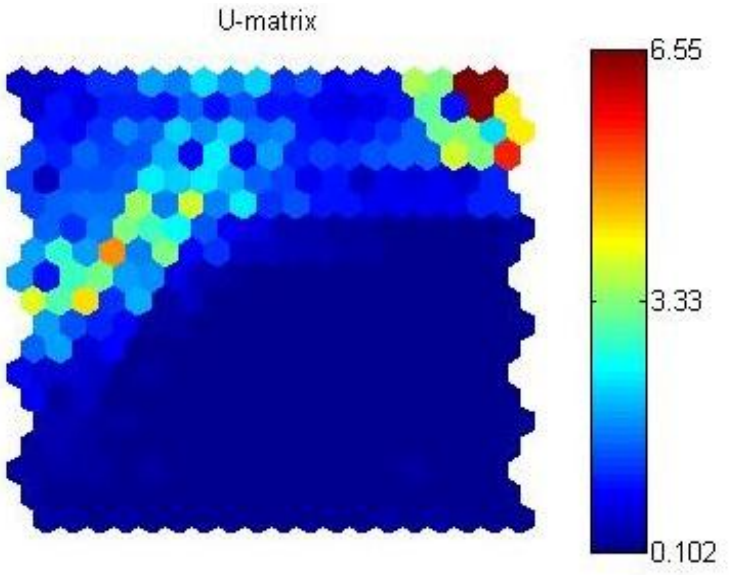

Figura 4: Matriz-U gerada para a rede 10 x 10 da palavra object (verbo) após treinamento de 250 épocas (ainda sem marcação dos grupos)

Fonte: Elaboração da autora.

A figura 4 mostra a matriz-U resultante do treinamento de um mapa auto-organizável bidimensional com 100 neurônios (10x10). As regiões escuras denotam a proximidade dos vetores de peso dos neurônios vizinhos no mapa. Observe-se que a parte inferior do centro à

direita possui neurônios próximos, formando o maior e principal agrupamento. Além disso, existe outro grupo no canto superior direito do mapa e outro sendo separado por uma zona de transição no canto superior esquerdo em que os dois grupos (*clusters*) estão separados por regiões claras (verde e amarelo). As zonas de fronteira são sempre representadas na matriz-U por neurônios de cores mais claras. Na teoria, as zonas de transição também podem ser consideradas como grupo.

Quanto à investigação do parâmetro que melhor facilitasse a visualização da segregação da rede, foram testados os coeficientes LP e os coeficientes cepstrais (MFC). A validação de agrupamentos pôde ser feita tanto via "Erro de Quantização" quanto por Índice DB. Na figura 5 (LPC 10), vê-se a sugestão de formação de três agrupamentos com erro de quantização final calculado a 0,427. Já na figura 3 tem-se o coeficiente Mel 10 sugerindo somente dois agrupamentos com erro de quantização superior (1,636), o que já descarta a futura utilização dessa matriz resultante de coeficientes MFC.

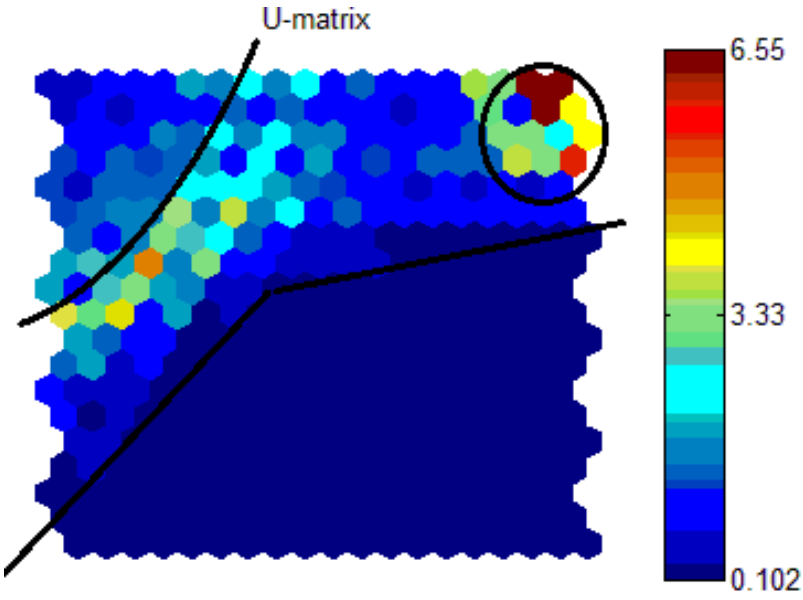

Figura 5: Matriz-U com entrada formada por 10 coeficientes LP. Índice DB indica 3 clusters e o erro de quantização final é igual a 0,427

Fonte: Elaboração da autora.

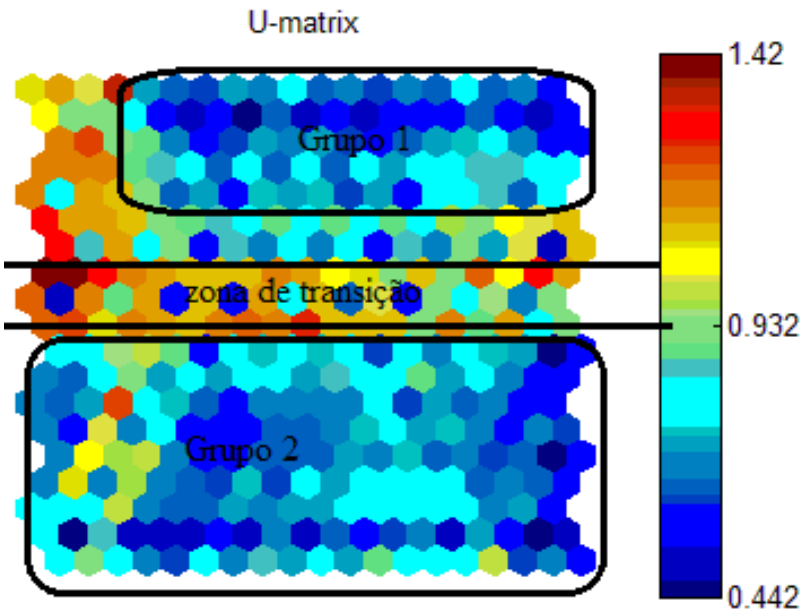

Figura 6: Matriz-U com entrada formada por 10 coeficientes MFC. Índice DB indica 2 clusters e o erro de quantização final é igual a 1,636.

Fonte: Elaboração da autora.

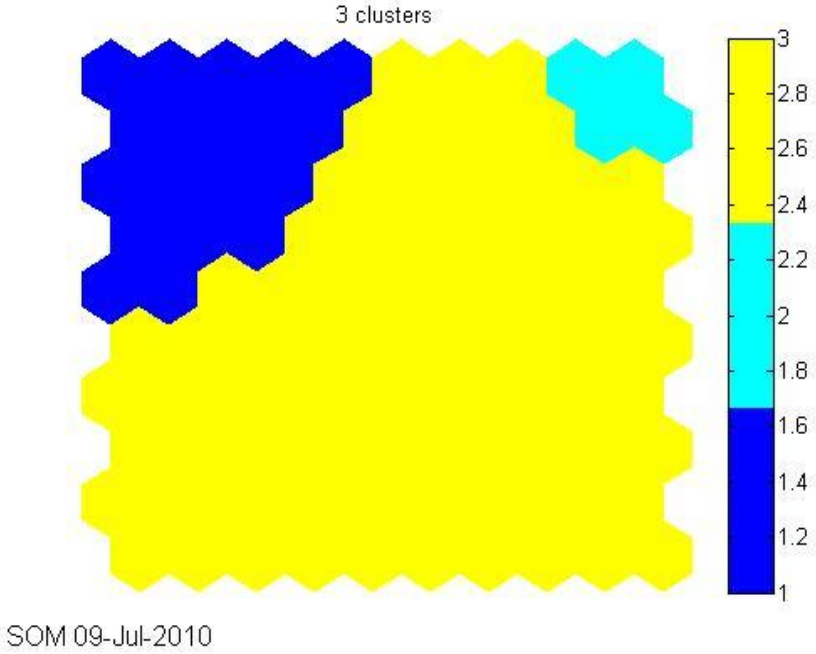

Figura 7: Representação dos grupos no mapa treinado para k=3.

Fonte: Elaboração da autora.

Conforme o número de agrupamentos identificados, a figura 6 exibe a organização dos grupos no mapa. Ao se interpretar o mapa, chega-se à conclusão de que três grupos representam melhor a distribuição dos dados. Assim, pode-se dizer que a rede sugere que os aprendizes estão separados em três grupos, cada um contendo propriedades linguísticas relevantes que os diferenciam.

Contrastando a figura 7 (mapa colorido sugerindo a formação de grupos) com a figura 8, que contém a representação dos grupos numerados na matriz-U, verifica-se a correspondência entre as áreas separadas no mapa colorido com as áreas circuladas e numeradas na matriz-U. A área azul corresponde ao grupo 4, a área de cor azul-piscina corresponde ao grupo 1 e a zona de cor amarela ao grupo 2.

Como se pode observar, o que garante um julgamento completo e confiável da saída da rede é a combinação da análise da matriz-U colorida, o mapa rotulado com as informações acerca do nível de proficiência lingüística do aprendiz e o mapa colorido. A análise individual de um só exemplo de representação da saída da rede não garante o sucesso na interpretação dos dados.

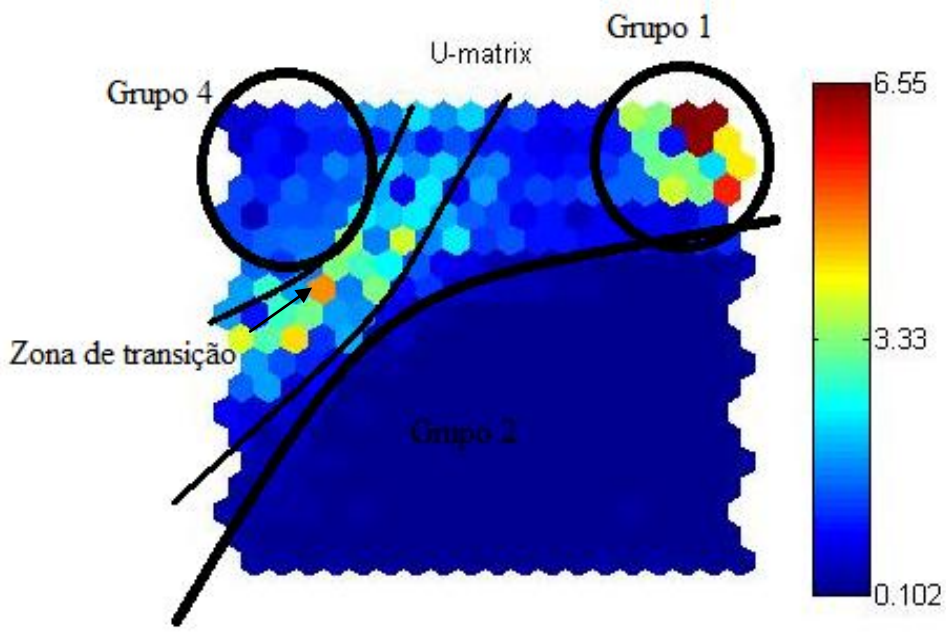

Figura 8: Matriz-U marcada com a representação dos grupos sugeridos pelo índice DB.

Fonte: Elaboração da autora.

Na matriz-U da figura acima, os vetores atípicos são identificados por regiões pequenas e separadas das demais por áreas de transição (SILVA, 2004). Isso significa dizer que a região correspondente ao grupo 4 possui rótulos vencedores para neurônios representantes de dados atípicos.

O uso da norma euclidiana para calcular a distância que um falante, cujo conjunto de dados seja inserido posteriormente na rede, esteja de outro indivíduo com características já mapeadas pode ser facilmente implementado. É justamente onde nossa proposta se firma. Por exemplo, pode-se treinar a rede com vetores de atributos de um número igual de falantes por nível de proficiência (e.g.: 5 falantes do nível A1[2], 5 do nível B2, 5 do nível C1) juntamente

[2] Referência ao Marco Comum Europeu (Common European Framework of Reference - CERF) de Proficiência Lingüística, disponível em http://www.cambridgeesol.org/exams/exams-info/cefr.html

com os vetores de características de um aprendiz que ainda não se submeteu a uma classificação rígida antes. É possível determinar o nível em que esse novo aprendiz se encontra sem o uso obrigatório de um teste padrão. Basta visualizar a topografia dos dados na matriz-U para depreender em que agrupamento ele se insere após o treinamento da rede, se nos agrupamentos A1, B2 ou C3 ou em outros.

6. Considerações finais

A vantagem de se usar o mapa auto-organizável (SOM) para estudar processos de aprendizagem reside não somente em seus princípios auto-organizacionais (competição entre neurônios por recursos limitados e cooperação implementada pela função vizinhança), mas também em sua visualização e propriedades de abstração (KOHONEN, 1998).

A rede SOM ainda é capaz de representar estruturas hierárquicas implícitas e modelar aspectos específicos de redes neurais biológicas. Constitui-se como um representante de um novo paradigma na inteligência artificial e na modelagem cognitiva; é uma ferramenta para análise estatística; é uma ferramenta para o desenvolvimento de aplicações complexas (HONKELA, 1997).

Além da rede de Kohonen se mostrar como uma ferramenta de auxílio na determinação de nível de proficiência lingüística, uma outra contribuição da rede SOM é uso da norma euclidiana para calcular a distância que um falante. Pode-se treinar a rede com vetores de atributos de um número igual de falantes por nível de proficiência (e.g.: 5 falantes do nível A14 , 5 do nível B2, 5 do nível C1) juntamente com o vetores de características de um aprendiz que ainda não se submeteu a uma classificação rígida antes. É possível determinar o nível em que esse novo aprendiz se encontra sem o uso obrigatório de um teste padrão. Basta visualizar a topografia dos dados na matriz-U para depreender em que agrupamento ele se insere após o treinamento da rede, se nos agrupamentos A1, B2 ou C3 ou em outros.

Em suma, a rede foi capaz de reconhecer padrões recorrentes nas formas de agrupamento dos neurônios vencedores (protótipos), refletindo o grau de aproximação e baseando-se em medidas de similaridade nos dados de entrada da rede. Tal tarefa desempenhada pela rede é um indicativo (ou por que não dizer sugestão) de nível de

proficiência que deve ser retomado em trabalhos futuros que se dediquem à análise dos critérios de determinação, de classificação e avaliação linguística em língua estrangeira.

Referências

BOSCARIOLI, C. Análise de agrupamentos baseada na topologia dos dados e em mapas auto-organizáveis. Tese de doutorado. Escola Politécnica da Universidade de São Paulo, 2008.

CASTRO, L. N. de. Fundamentals on Neurocomputing. Basic concepts, algorithms, and applications. Taylor e Francis Group, 2006.

CHUN, Dorothy M. Signal analysis software for teaching discourse intonation. Language Learning & Technology. volume 2, number 1, pp. 74-93, july 1998.

Disponível em <http://llt.msu.edu/vol2num1/article4/> Acesso em: 23 mar. 2010. HONKELA, T. Self-organizing maps in natural language processing. Doctorate thesis. University of Helsink.1997.

KOHONEN, T. Self-Organizing Maps. Springer, 3a. ed. 2001.LEVIS, J. Computer Technology In Teaching And Researching Pronunciation. Annual Review of Applied Linguistics, CUP, 2007.

MOLHOLT, G. Computer-assisted instruction in pronunciation for Chinese speakers of American English. *TESOL Quarterly, 22*(1), 91-111, 1988.

SILVA, A. C. C. da. 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.

SILVA, M. A. S. da. Mapas auto-organizáveis na análise exploratória de dados geoespaciais multivariados. Dissertação de mestrado em computação aplicada. INPE, 2004.

SOUZA JR. A. H. de. Avaliação de rede neurais auto-organizáveis para reconhecimento de voz em sistemas embarcados. Dissertação de mestrado.