# A SOM-Based Analysis of Early Prosodic Acquisition of English by Brazilian Learners: Preliminary Results

Ana Cristina C. Silva[1], Ana Cristina P. Macedo[2], and Guilherme A. Barreto[3]

[1] Department of Linguistics, State University of Piauí, Brazil
cris0708@gmail.com
[2] Department of Linguistics, Federal University of Ceará, Brazil
pelosi@ufc.br
[3] Department of Teleinformatics Engineering, Federal University of Ceará, Brazil
guilherme@deti.ufc.br

**Abstract.** In this paper the SOM is used in an exploratory analysis of transfer phenomena from first language (L1) to the second language (L2) related to word/lexical stress. The basic hypothesis tested is whether the parameterization of the speech signal of the learner's utterances by standard signal processing techniques, such as Linear Predictive Coding (LPC), used to encode the input of the network results in efficient categorization of speakers by the SOM. Preliminary results indicates that the combination LPC+SOM is indeed able to produce well-defined clusters of speakers that possess similarities regarding the transfer of stress patterns among Brazilian students in learning English as a foreign language.

**Keywords:** Self-organizing map, word/lexical stress, linear predictive coding, U-matrix.

## 1 Introduction

Connectionist models have been playing an important role in language development in several areas, such as lexical and pronoun acquisition, syntactic sistematicity, language disorder modeling and prosodic analysis [16, 17, 20], just to mention a few. Most of these works are based on feedforward or recurrent supervised neural network architectures [4, 6, 8, 11], such as the MLP and Elman networks, but self-organizing neural network models have also been used as the primary linguistic model [9, 10, 13, 14, 18, 19].

For example, Li and co-workers [13, 14] simulated the lexical acquisition in infants using a self-organizing neural network model. The main objective of the research was to use the properties of topographic preservation of the Self-Organizing Map (SOM) [12] to study the emergence of linguistic categories and its organization throughout the stages of lexical learning. The model captured a series of important phenomena occurring in children's early lexical acquisition

J. Laaksonen and T. Honkela (Eds.): WSOM 2011, LNCS 6731, pp. 267–276, 2011.

and had significant implications for models of language acquisition based on self-organizing neural networks.

Also using the SOM, Gauthier et al. [10] studied whether and how children could learn prosodic focus directly from input continuous speech. The authors explored how the focus could be learned from acoustic continuous signals in Mandarin, which were produced with co-occurring lexical tones and by various speakers. The results of this study showed that neural networks can develop unsupervised groupings of specific focus from the continuous dynamic speech signal, produced by various speakers in various lexical tone conditions, which may eventually lead to the acquisition of the prosodic focus.

Of particular interest to the current paper is the lexical stress, one of the most important prosodic elements. The stress in English has multiple functions ranging from an emphatic role, through the contrastive power to indicate syntactic relationships between words and word parts, such as the oppositions of pairs of noun and verb words, e.g. (**OB**ject, ob**JECT**), (**DE**sert, de**SERT**), (**CON**flict, con**FLICT**), etc.

In Brazilian Portuguese (BP) there is a tendency the trisyllabic and polysyllabic nouns be paroxytone. Trisyllabic and polysyllabic verbs in BP suffer a tendency to be oxytone. There is another trend: that of trisyllabic and polysyllabic adjectives in BP are paroxytone. According to some studies in the area of phonology of interlanguage [1, 2, 3, 15, 21], the lexical stress is most responsible for cases of language transfer, i.e. the influence of the predominant accent of L1 (first language) in L2 (second language) learning.

Despite some previous works involving the connectionist modeling of prosodic features for language development and identification [4, 6, 10], to the best of our knowledge, there has been no systematic investigation nor an exploratory analysis of transfer phenomena from L1 to L2 related to word/lexical stress by means of connectionist model, such as the SOM. Furthermore, we are not aware of studies on the application of artificial neural networks to investigate how the knowledge of Brazilian learners of English is organized in relation to the acquisition of early L2 stress and transference of stress pattern from L1 to L2.

From the exposed, this article aims to investigate whether and how the SOM network is able to build well-defined groups (clusters) of speakers that possess similarities regarding the transfer of stress patterns among Brazilian students in learning English as a foreign language. The ultimate goal of this research is to use the SOM as a tool to evaluate the proficiency level of students. The basic hypothesis tested is whether the parameterization of the speech signal of the learner's utterances by standard signal processing techniques, such as Linear Predictive Coding (LPC), for encoding the input of the network is efficient in the categorization of speakers.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the SOM, the corpus and the speech parameterization technique used in this research. The results of computer simulations and comments about them are presented in Section 3. The paper is concluded in Section 4.

## 2 Methods

### 2.1 The Self-Organizing Map

In what follows, a brief description of the original SOM algorithm, introduced by Kohonen [12], is given. Let us denote $\mathbf{m}_i(t) \in \mathbb{R}^p$ as the weight vector of the $i$-th neuron in the map. After initializing all the weight vectors randomly or according to some heuristic, each iteration of the SOM algorithm involves two steps. First, for a given input vector $\mathbf{x}(t) \in \mathbb{R}^p$, we find the current winning neuron, $i^*(t)$, as follows

$$i^*(t) = \arg\min_{\forall i}\{\|\mathbf{x}(t) - \mathbf{m}_i(t)\|\}. \tag{1}$$

where $t$ denotes the iterations of the algorithm. Then, it is necessary to adjust the weight vectors of the winning neuron and of those neurons in its neighborhood:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \eta(t)h(i^*, i; t)[\mathbf{x}(t) - \mathbf{m}_i(t)], \tag{2}$$

where $0 < \eta(t) < 1$ is the learning rate and $h(i^*, i; t)$ is a Gaussian weighting function that limits the neighborhood of the winning neuron:

$$h(i^*, i; t) = \exp\left(-\frac{\|\mathbf{r}_i(t) - \mathbf{r}_{i^*}(t)\|^2}{2\sigma^2(t)}\right), \tag{3}$$

where $\mathbf{r}_i(t)$ and $\mathbf{r}_i^*(t)$, are respectively, the positions of neurons $i$ and $i^*$ in a predefined output array where the neurons are arranged in the nodes, and $\sigma(t) > 0$ defines the radius of the neighborhood function at time $t$. To guarantee convergence of the algorithm, $\eta(t)$ and $\sigma(t)$ decay exponentially in time according to the following expressions:

$$\eta(t) = \eta_0\left(\frac{\eta_T}{\eta_0}\right)^{(t/T)} \qquad \text{and} \qquad \sigma(t) = \sigma_0\left(\frac{\sigma_T}{\sigma_0}\right)^{(t/T)}, \tag{4}$$

where $\eta_0(\sigma_0)$ and $\eta_T(\sigma_T)$ are the initial and final values of $\eta(t)$ $(\sigma(t))$.

The incremental learning process defined by Eqs. (1 ) and (2) can often be replaced by the following batch computation version which is usually faster.

1. Initialize the weight vectors $\mathbf{m}_i$, $\forall i$.
2. For each neuron $i$, collect a list of all those input vectors $\mathbf{x}(t)$, whose most similar weight vector belongs to the neighborhood set $N_i$ of neuron $i$.
3. Take as the new weight vector $\mathbf{m}_i$ the mean over the respective list.
4. Repeat Step 2 a few times until convergence is reached.

Steps 2 and 3 of the batch SOM algorithm need less memory if at Step 2 one only make lists of the input vectors $\mathbf{x}(t)$ at those neurons that have been selected for winner, and at Step 3 we take the mean over the union of the lists that belong to the neighborhood set $N_i$ of neuron $i$.

In addition to usual vector quantization properties properties, the resulting ordered map also preserves the topology of the input samples in the sense that adjacent input patterns are mapped into adjacent neurons on the map. Due to this topology-preserving property, the SOM is able to cluster input information and spatial relationships of the data on the map.

### 2.2 Input Corpus and Data Representation

The corpus of this research is composed of interview recordings with 30 students (learners) of a higher education institution in the city of Fortaleza, federal state of Ceará, aged between 18 and 25, all Brazilians, of both genders, who have never traveled to an English-speaking country until the time of interview. It was decided to allocate the 30 participants in five different levels of development, using the criterion of length of exposure to language. Based on this fact, number of classroom hours accumulated in the discipline of English language obtained through interviews and questionnaires completed by participants was established as a circumstantial criterion of classification and organization of the individuals.

The participants' utterances were recorded in the software *SoundForge* (version 5.0) in WAV audio files at a sampling rate of 44.1 KHz with 16-bit resolution, single channel (mono). After this phase, each word representing the lexical item to be investigated (e.g. object, separate, desert, etc.) was manually segmented with the help of a phonetician using the same software.

As the speech signal cannot be directly used to feed the network because it contains thousands of samples, which would make their processing very slow, and also for being very noisy, which makes it extremely difficult to extract knowledge, the solution is to represent it numerically with a set of coefficients obtained from the application of mathematical techniques such as linear prediction coefficients and mel-cepstral coefficients, with the speech signal divided into multiple frames. Thus, the speech signal of the learners is numerically represented by coefficient vector sets computed using the PRAAT software [5].

### 2.3 Speech Signal Parametrization (Feature Extraction)

The process of feature extraction of the speech signal is a crucial step in the connectionist approach to pattern classification and clustering. This step consists in applying standard signal processing techniques to the original speech signal in order to convert it to more suitable compact mathematical representation that permits the identification of a given utterance by a connectionist model.

Linear predictive coding (LPC) is a signal processing technique widely used for the parametrization of the speech signal in several applications, such as speech compression, speech synthesis and speech recognition [7]. Roughly speaking, the LPC[1] technique represents small segments (or frames) of the speech signal by the coefficients of autoregressive (AR) linear predictors. For example, if the speech signal has 500 frames, it will be parameterized by a set of 500 coefficient vectors. To assure stationarity, each frame usually has a short duration ($\sim$ 10-30ms).

The set of LPC coefficient (LPCC) vectors associated with the utterance of a given word are then organized along the rows of a matrix of coefficients. For

[1] LPC coefficients can extract the intensity and frequency of the speech signal. These two characteristics are closely associated with the prosodic element "accent". In English, the stress is the junction of three perceptual factors interrelated: 1) quantity / length (measured in ms) related to the size of the syllable, 2) intensity (measured in dB) related to amplitude and 3) height (measured in Hz), i.e., the value of higher F0 in an utterance.

example, if 500 coefficient vectors generated, one vector for each frame, the corresponding matrix of coefficients has 500 rows. The number of columns of this matrix is equal to order of the AR predictor used in the LPC analysis. The matrices of coefficients are then used to train the SOM.

### 2.4 SOM Training and Data Visualization

Four simulations were ran with parameters that varied according to the need to adjust to the phenomenon in question. The experiments were design in order to verify whether the network could organize (discriminate) learners depending on the transference of stress pattern from L1 to L2 are detailed below. When applied to the problem of interest, the simulation process of the SOM and the analysis of results of the training involves the following steps:

1. Startup and training (learning) of the network;
2. Evaluation of the quality of the map using the quantization error (QE) and topological error (TE);
3. Generation of the $U$-matrix and labeled map after each training run;
4. Validation of clusters through the Davies-Bouldin index (DB);
5. Tabulation of the data for all outcome measures of network performance (the quantization error and topological error).

All simulations were conducted using a two-dimensional hexagonal SOM, with hexagonal neighborhood structure, Gaussian neighborhood function, random initiation of weights and batch learning. For all the experiments we simulated a $5 \times 5$ SOM, for 250 epochs (50 for rough training, 200 for fine tuning) with initial and final neighborhood of 4 and 1, respectively. The maximum numbers of clusters used by the DB index was set to 10. These specifications proved adequate to treat the phenomena in question. The SOM toolbox [23] was used to run all the experiments to be described.

As mentioned in the Subsection 2.3, every word uttered by a speaker generates a coefficient matrix. In order to identify this speaker in a posterior analysis of the results, it is necessary to label the data (row) vectors in that matrix as belonging to that particular speaker. For this purposes, an alphanumeric label is appended to each row vector in an additional column. Finally, the text files containing labeled data related to the utterance of a specific word for all the speakers are concatenated into a single file.

It is noteworthy that in addition to the label that identifies the speaker, other labels can be associated with a given coefficient matrix of that speaker. For instance, a second label can identify the linguistic category in which the word pronounced is inserted. This Multi-Label (ML) Analysis is introduced in this paper with the goal of determining which labeling is more appropriate to the type of parameterization used. In other words, ML analysis can help inferring which linguistic properties of the speech signal are encoded in the LPC coefficients.

Finally, the $U$-matrix [22] is used as a tool to visualize the clusters formed during the learning process.

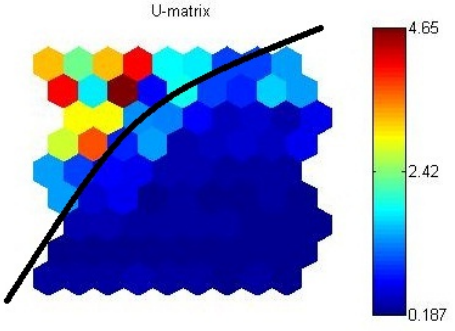


**Fig. 1.** U-matrix revealing the formation of two major groups, one probably related to speakers who transfer the stress pattern and the other to speakers who do not transfer

## 3 Results and Discussion

This simulation aims at investigating whether the SOM would be able to organize the speakers in clusters, according to the process of transferring the stress pattern of Brazilian Portuguese into English. All the 30 speakers were asked to utter 30 different English sentences containing situations where certain words of interest act sometimes as a verb or as a noun. In this paper, we report only the results obtained for the sentence 'I object to going to a bar', where the word of interest is the verb 'obJECT'. The full corpus is available to the interested reader upon request.

Three types of graphics were generated after SOM training: U-Matrix, labeled map (majority rule) and clustered map. The U-matrix and the clustered map requires no labeled data to be constructed. The labeled map is more useful for our purposes if labeled data are available since labels may provide a better understanding of speakers' organization as a function of their linguistic abilities. It is worth pointing out that all the SOM computations are performed using unlabeled data, i.e. it runs totally in an unsupervised way. The labels are used only in the analysis of the results.

Two criteria were followed for labeling purposes. At first, the speakers' labels carry no information about errors in L2 stress, i.e., the transfer pattern of L1 to L2. In this case, the data from a given speaker is labeled by a number indicating his/her formal education level in L2 studies (i.e. period in an English course) and his/her order in the interview process. For example, the label '608' denotes a speaker in the 6th semester ranked 8th in the list of individuals interviewed for this research. The second labeling criterion added the characters "er" to the label when a speaker misses the pronunciation, i.e. when he/she transfers the pattern from L1 (Brazilian Portuguese) to L2 (English). For example, the label '203er' denotes a speaker in the 2nd semester, ranked 3rd in the interview sequence and who missed the pronunciation.

**Table 1.** DB index values for different values of $K$

| $K=2$ | $K=3$ | $K=2$ | $K=3$ | $K=2$ | $K=3$ | $K=2$ | $K=3$ | $K=10$ |
|---|---|---|---|---|---|---|---|---|
| 0.4306 | 0.8494 | 0.6018 | 0.6100 | 0.7914 | 0.6484 | 0.7520 | 3.1829 | 11.3072 |

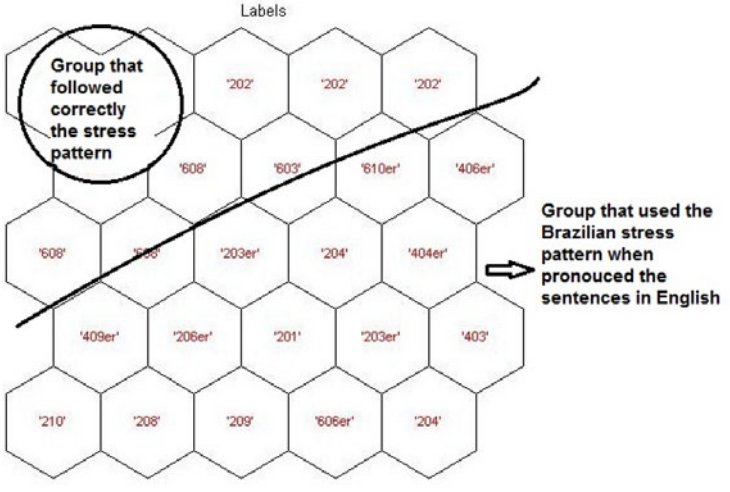


**Fig. 2.** Labeled map associated to the U-matrix shown in Figure 1, confirming the expectation of two major groups of students, one containing mainly individuals who transfer the BP stress pattern and one that does not transfer

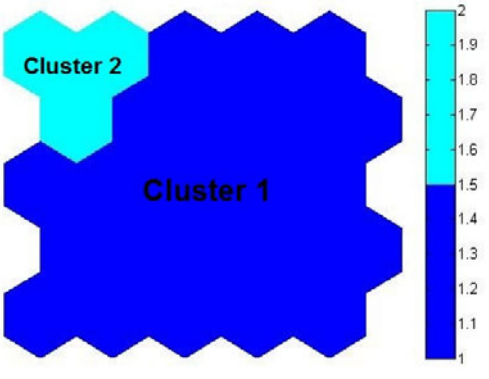


**Fig. 3.** Clustered map suggesting the existence of two well-defined clusters, according to Davies-Bouldin index

Figures 1 to 3 illustrate the obtained results. Errors in the pronunciation of this word occur when it is pronounced as a noun (**OB**ject), instead of a verb (ob**JECT**). For each speaker, the speech signal segment corresponding to the word "Object" is manually selected and parameterized by the LPC method. The order of the AR predictor for this experiment was set to 10 and the duration of each frame was set to 25ms.

The resulting U-matrix is shown in Figure 1. By analyzing this figure one can clearly see the signs of formation of two clusters: a larger group (the one with prototypes closer to each other - in blue in the figure) and a smaller group (the one with prototypes more separated from each other - in different colors in the figure). Based on the a priori analysis of the frequency of occurrence of the labels provided by the phonetician, the larger group probably is the one containing the individuals who transfer the primary stress. This hypothesis can only be confirmed by analyzing the labeled map shown in Figure 2.

Neuron labeling is carried out by the majority rule, i.e. the neuron inherits the label that occurs most frequently among the data vectors (LPCC vectors) mapped to that neuron. The labeled map confirms the hypothesis of two groups raised by the U-matrix. The neurons whose labels include the characters "er" are located below the solid line separating the map into two parts.

It is worth noting that this clear separation of students was carried out in an unsupervised way by the SOM using solely the information provided by the LPC coefficients, i.e. the network organizes the students by similarity between their feature (LPCC) vectors only. No a priori linguistic knowledge was used during the feature extraction process nor the SOM training. Label information was indeed provided by an expert but it is used only for the purpose of interpretation of the trained map.

The clustered map in Figure 3 adds corroborating evidence to the results provided by the labeled map and the $U$-matrix concerning the emergence of two well-defined groups. Clustering of the SOM was carried out using the $K$-means algorithm, varying $K$ from 1 to 10, following the approach proposed in [23]. The optimal value for $K$, according to the DB index, was $K_{opt} = 2$ (see Table 1).

## 4 Conclusion

The preliminary results presented in this paper can serve as a starting point to demonstrate that an unsupervised neural network can be useful to visualize the cluster formation of prosody-related linguistic phenomenon, in this case, the transference of lexical stress. We started from the assumption that the parameterization of the speech signal through the LPC coefficients would be effective in the categorization of speakers for prosodic features.

The segregation of the map in regions of well-defined clusters suggested that the learners were grouped by similar phonetic-acoustic features. According to the rounds of experiments, it was confirmed that the network discriminated speakers according to prosodic features and organized them according to similarities on these characteristics. Importantly, within these two large groups (the group that

transfers the BP stress pattern and what does not transfer) there can be subgroups (subclusters) which, when closely examined in isolation, might reveal rich information for the linguistic analysis of learner's utterances as well as to contribute to understanding the organization of the data set. We are currently, developing experiments to analyze these subgroups.

Further tests are to be made and with more results, we hope to perfect the proposed SOM-based methodology and use it in the future as a tool for determining the language proficiency level classification in foreign languages.

**Acknowledgements.** The authors thank FUNCAP and CAPES (Brazilian agencies for promoting science) for the financial support to this research.

## References

1. Albini, A.B.: The influence of the portuguese language in accentuation of english words by brazilian students (in portuguese). Revista Prolíngua 2(1), 44–56 (2009)
2. Archibald, J.: A formal model of learning L2 prosodic phonology. Second Language Research 10(3), 215–240 (1994)
3. Baptista, B.O.: An analysis of errors of Brazilians in the placement of English word stress. Master's thesis, Postgraduate Program on Linguistics, Federal University of Santa Catarina, Brazil (1981)
4. Blanc, J.M., Dominey, P.F.: Identification of prosodic attitudes by a temporal recurrent network. Cognitive Brain Research 17, 693–699 (2003)
5. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (2011), `http://www.praat.org`, version 5.2.10 (retrieved January 11, 2011)
6. Cummins, F., Gers, F., Schmidhuber, J.: Language identification from prosody without explicit features. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH 1999), pp. 305–308 (1999)
7. Deller, J., Hansen, J.H.L., Proakis, J.: Discrete-Time Processing of Speech Signals. John Wiley & Sons, Chichester (2000)
8. Elman, J.L.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
9. Farkas, I., Crocker, M.W.: Syntactic systematicity in sentence processing with a recurrent self-organizing network. Neurocomputing 71, 1172–1179 (2008)
10. Gauthier, B., Shi, R., Xu, Y.: Learning prosodic focus from continuous speech input: A neural network exploration. Language Learning and Development 5, 94–114 (2009)
11. Kaznatcheev, A.: A connectionist study on the interplay of nouns and pronouns in personal pronoun acquisition. Cognitive Computation 2, 280–284 (2010)
12. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer, Heidelberg (2001)
13. Li, P., Farkas, I., MacWhinney, B.: Early lexical development in a self organizing neural network. Neural Networks 17, 1345–1362 (2004)
14. Li, P., Zhao, X., MacWhinney, B.: Dynamic self-organization and early lexical development in children. Cognitive Science 31, 581–612 (2007)
15. Mairs, J.L.: Stress assignment in interlanguage phonology: an analysis of the stress system of spanish speakers learning english. In: Gass, M., Schatcther, J. (eds.) Linguistic Perspectives on Second Language Acquisition, Cambridge University Press, Cambridge, USA (1989)

16. McClelland, J.L.: The place of modeling in cognitive science. Topics in Cognitive Science 1, 11–38 (2009)
17. McClelland, J.L., Botvinick, M.M., Noelle, D.C., Plaut, D.C., Rogers, T.T., Seidenberg, M.S., Smith, L.B.: Letting structure emerge: connectionist and dynamical systems approaches to cognition. Trends in Cognitive Sciences 14, 348–356 (2010)
18. Miikkulainen, R.: Dyslexic and category-specific aphasic impairments in a self organizing feature map model of the lexicon. Brain and Language 59, 334–366 (1997)
19. Miikkulainen, R., Kiran, S.: Modeling the bilingual lexicon of an individual subject. In: Príncipe, J.C., Miikkulainen, R. (eds.) WSOM 2009. LNCS, vol. 5629, pp. 191–199. Springer, Heidelberg (2009)
20. Poveda, J., Vellido, A.: Neural network models for language acquisition: A brief survey. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 1346–1357. Springer, Heidelberg (2006)
21. Silva, A.C.C.: The production and perception of word stress in minimal pairs of the english language by brazilian learners. Master's thesis, Postgraduate Program on Linguistics, Federal University of Ceará, Brazil (in Portuguese) (2005)
22. Ultsch, A., Siemon, H.P.: Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the International Neural Network Conference (ICNN 1990), pp. 305–308. Kluwer Academic Publishers, Dordrecht (1990)
23. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11(3), 586–600 (2000)