

Unsupervised Syllable-based Behaviors in Phonology

The aim is to model phonotactic contexts in order to discover the syllabic behaviors of English and Italian phonemes. The approach is that of hierarchically unsupervised SOM-based learning. The Self-Organizing Map (SOM) builds a topology, mapping from the high-dimensional space onto map units in such a way that the relative distances between data points are preserved. The process of reducing the dimensionality is essentially the data compression technique known as *vector quantisation*. In addition, the SOM (defined as a sheet-like neural-network array) stores information in such a way that any topological relationships within the training set be maintained. The representations are automatically derived from the metric relationships between the input items; no ‘supervision’ is needed, i.e. no information about input-output relations are *a priori* defined.

Two *corpora* were designed for these preliminary investigations. The *corpora* aim at minimally representing the various syllable types typical of the two languages considered. We had the following number of, respectively, words and syllable types: Italian 83 and 51, English 159 and 78. Needless to say, a more organic and usage-based inspired corpus set would improve the output of the system, by providing the higher-level SOM a more representative and fine-grained phonotactic representation. The present results are thus to be considered exploratory (and the good performance obtained as very promising).

We designed a SOMs first-level, trained with segments as labelled in a contextually sliding window within a given segment chain (Fig. 1). Two different data representations were adopted in the simulations: i) Orthogonal (independent) representation: each phoneme is encoded as a binary vector specifying its natural phonological class (vowel, nasal, stop, etc); ii) Phonological representation: each phoneme is encoded as a binary vector specifying features of place/manner of articulation. This first-level defines the ‘phonotactic level’ of each segment in relation to the adjacent segments (phonotactic environment). This first mapping process provides a preliminary dimensionality reduction of the input data, further computed and reduced in the second-level SOM.

The organization obtained in the ‘phonotactic level’ is used as input for a higher-level SOM (the so-called ‘syllabic SOM’), where the relationships pertaining to the syllabic behavior are spatially emphasized. This final level reflects the organization of the syllable components and thus attributes functional syllabic role to the segments previously mapped in the first-level SOM. Specifically, syllable structures and syllabification processes are modeled by states of activation inside the ‘syllabic SOM’. After the learning phase, the system assigns each phonological segment a spatial location in the ‘syllabic SOM’ on the basis of its phonotactic information. The ‘syllabic SOM’ shows syllabically-motivated entrenchment zones, where emergent clusters reflect the syllabic components (Onset, Nucleus, Coda and ‘ambisyllabic Onset/Coda’). Fig. 2 shows the output for the two languages considered.

It is worth noting that, due to the different phonotactics of English and Italian, the ‘English Syllabic Map’ appears less clearly-defined than the Italian one in terms of spatial cluster organization. This may be explained as a sort of spatial compromise in representing complex instances of ambisyllabicity in English. It is however to be expected that a larger corpus would greatly improve the output.

A generalization process was subsequently attempted. The (spatial) syllabification output was obtained for ‘new words’, i.e. words not present in the training set. The system yielded the syllabification by spatially mapping the new forms segments within the syllable component clusters, thus automatically detecting the implicit syllabic behavior of the new word’s segments. Fig. 2 reports the syllabification process for an ‘unseen word’ as /bɪfɔːr/, ‘before’.

Bibliography

- Pier Marco Bertinetto, Psycholinguistic evidence for syllable geometry: Italian and beyond, in Rennison & Kähnhammer (eds.), *Phonologica 1996. Syllables!?*, Holland Academic Graphics, 1996.
- John Goldsmith, Local Modeling in Phonology, in Davis (ed.), *Connectionism: Theory and Practice*, Oxford, 1992.
- Teuvo Kohonen, *Self-Organizing Maps*, Springer, 1995.
- Bernard Laks, A connectionist account of French syllabification, *Lingua*, 1995, 95:349-372.

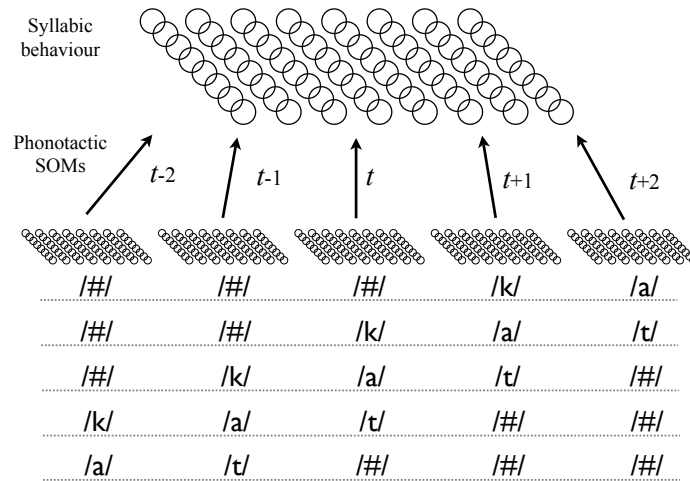


Figure 1: Hierarchically unsupervised SOM-based architecture

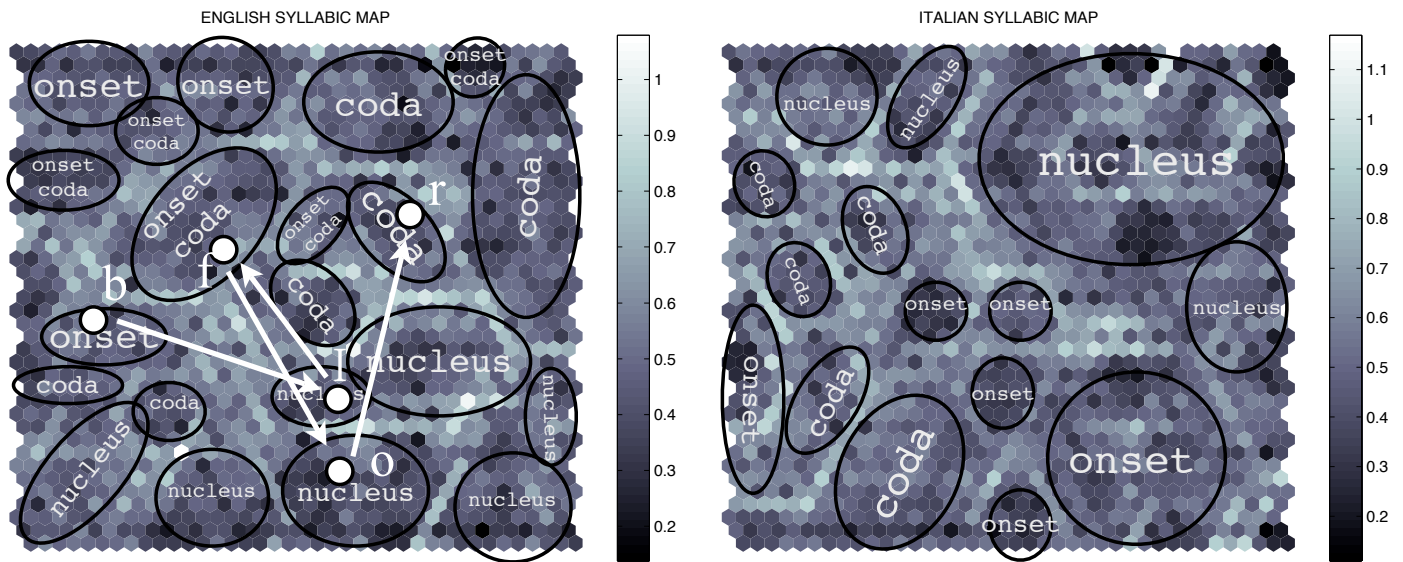


Figure 2: English and Italian ‘syllabic SOM’ in U-matrix visualization. U-matrix (unified distance matrix) represents the distances between reference vectors of neighboring map units. The distance matrix reflects the level of similarity between a neuron and its neighboring neurons. With color scales for representing the distance matrix, we can easily detect clusters, i.e. those neurons tied closely (dark colors in the map). The labeled clusters specify the syllabic behavior of the phonological segments. The syllabification of a new word (as /bɪfəʊr/, ‘before’) is also shown in the English SOM