A tool for exploring building blocks in speech and animal vocalizations

Yannick Jadoul[*1,2] and Bart de Boer[2]

[*]Corresponding Author: Yannick.Jadoul@uniroma1.it
[1]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy
[2]Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium

An important characteristic of human language is that it is based on combinatorial structure: it combines a small set of (mostly meaningless) building blocks into an unlimited set of (meaningful) utterances (de Boer, Sandler, & Kirby, 2012), which then in turn can get recombined into even bigger ones. Similar structure can be found in animal vocalizations (e.g., Kershenbaum et al., 2014) but the evolutionary link between say, primate vocalizations and human speech is not clear yet. Here we argue that to address this question we need tools to identify building blocks directly from the signal, but that these tools should be as little biased by linguistic notions as possible. Current analyses sometimes start from notions such as vowels (e.g., Fitch, de Boer, Mathur, & Ghazanfar, 2016; Boë et al., 2017) or consonants (e.g., Lameira, Maddieson, & Zuberbühler, 2014). Many other approaches assume a full segmentation of the input signal, and with no overlap or gaps between segments (e.g., Kreuk, Keshet, & Adi, 2020). We propose a tool, inspired by the field of data mining (frequent pattern mining; Aggarwal, Bhuiyan, & Hasan, 2014) to identify building blocks of speech directly from the signal, without assuming notions such as consonants, vowels, or syllables.

Identifying building blocks directly from a signal is hard, due to real signals being continuous and noisy. Even under ideal circumstances (single subject and low noise) no two occurrences of the same utterance are identical. We therefore adopt a number of techniques from speech recognition and combine them with frequent sequence mining to automatically derive candidate building blocks.

In order to demonstrate and evaluate our methods, we recorded a dataset with three-syllable nonsense words. Each CV-syllable consists of one of three consonants (/b/, /d/, /g/) and one of three vowels (/a/, /i/, /u/), resulting in 9 different syllables combined into $9^3 = 729$ highly structured nonsense words. All 729 words were read by a single speaker and recorded in a low-noise environment.

Frequent sequence mining algorithms can identify frequently occurring patterns in large sets of sequences. As they typically operate on symbolic sequences, a first step is to extract feature vectors from the acoustic data and cluster these in a set of discrete categories. Following standard speech processing procedure,

we extracted 13 mel-frequency ceptral coefficients (MFCCs) using Parselmouth (Jadoul, Thompson, & de Boer, 2018; Boersma & Weenink, 2021) and used k-means to cluster the normalized MFCC vectors into 24 clusters. It should be noted that MFCCs are based on properties of human hearing, so they may need to be replaced by an appropriate model when analysing other species – based on the properties of that species' perception. This can easily be accommodated in our system.

The large number of patterns found by a frequent pattern mining algorithm need to be filtered before they can be interpreted as building blocks. On our small data set, running the CM-SPAM algorithm (Fournier-Viger, Gomariz, Campos, & Thomas, 2014; Fournier-Viger et al., 2016) already results in 7177 patterns that occur in more than 10% of input sequences. To filter these patterns to a manageable number of building blocks, we incrementally selected patterns which together cover more and more of the input sequences (Figure 1). The patterns we recover represent vowel formant patterns and formant transitions related to consonants.

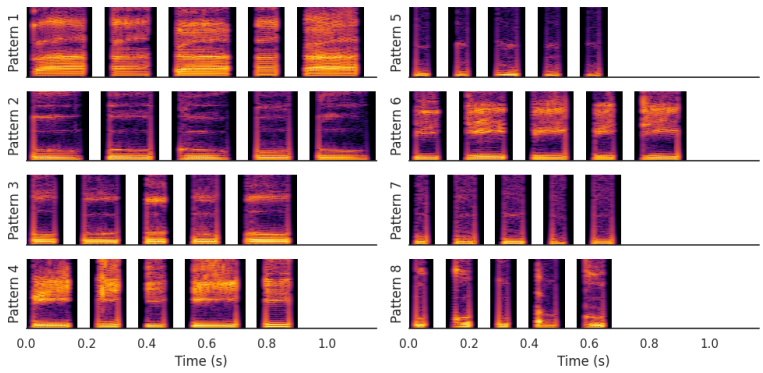

Figure 1. Audio fragments matched by the 8 most important building blocks show vowels and formant transitions representing consonants. The spectrograms' frequency ranges from 0 to 5000 Hz.

The current dataset is highly structured and has less noise than most real-world data, so it remains to be seen how our technique performs in less ideal circumstances. However, this preliminary result shows that our very general technique can identify relevant (i.e. in this case clearly related to the vowels and consonants in the original data) building blocks from a real signal, without assuming linguistic notions. Therefore it seems a promising approach for analysing the combinatorial structure in animal vocalizations, which could assist in further investigations into how the ability to use such structure has evolved.

References

Aggarwal, C. C., Bhuiyan, M. A., & Hasan, M. A. (2014). *Frequent pattern mining algorithms: A survey.* Springer.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer [Computer program].* Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`.

Boë, L.-J., Berthommier, F., Legou, T., Captier, G., Kemp, C., Sawallis, T. R., Becker, Y., Rey, A., & Fagot, J. (2017). Evidence of a vocalic proto-system in the baboon (papio papio) suggests pre-hominin speech precursors. *12*(1), e0169321.

de Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *4*(4), 251–259.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *2*(12).

Fournier-Viger, P., Gomariz, A., Campos, M., & Thomas, R. (2014). Fast vertical mining of sequential patterns using co-occurrence information. In *Advances in knowledge discovery and data mining: 18th pacific-asia conference, pakdd 2014, tainan, taiwan, may 13-16, 2014. proceedings, part i 18* (pp. 40–52).

Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016). The spmf open-source data mining library version 2. In *Machine learning and knowledge discovery in databases: European conference, ecml pkdd 2016, riva del garda, italy, september 19-23, 2016, proceedings, part iii 16* (pp. 36–40).

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15.

Kershenbaum, A., Bowles, A. E., Freeberg, T. M., Jin, D. Z., Lameira, A. R., & Bohn, K. (2014). Animal vocal sequences: not the markov chains we thought they were. *281*(1792), 20141370.

Kreuk, F., Keshet, J., & Adi, Y. (2020). Self-supervised contrastive learning for unsupervised phoneme segmentation. *arXiv preprint arXiv:2007.13465*.

Lameira, A. R., Maddieson, I., & Zuberbühler, K. (2014). Primate feedstock for the evolution of consonants. *18*(2), 60–62.