USING SPECTRAL COHERENCE BETWEEN TEMPORAL ENVELOPE AND MOUTH MOVEMENTS TO REVEAL RHYTHMIC REGULARITIES IN SPEECH

Lei He
[a] Department of Computational Linguistics, University of Zurich, Switzerland
[b] Department of Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich, Switzerland
lei.he@uzh.ch

ABSTRACT

This paper features a method to characterize the rhythmicity in speech from the temporal envelope and mouth opening-closing movements using spectral coherence [He (2022) *J. Acoust. Soc. Am*. **152**, 567–579]. Such a coherence reveals the joint regularities in both domains of signals. For illustrations, recordings from both L1 and L2 English utterances are used for comparisons.

The recordings utilized for this Prosody Visualization Challenge (PVC-2023) have been chosen from the EMA-MAE corpus [1],[1] which is openly accessible. As such, there are no ethical concerns associated with their current scientific and not-for-profit usage.

Keywords: speech rhythm, spectral coherence, temporal envelope, mouth opening-closing movements

1. INTRODUCTION

Rhythm is a ubiquitous phenomenon in nature. The rhythmicity in speech has been a central interest to phoneticians, speech scientists, psychologists, and neuroscientists. Different approaches to speech rhythm have been proposed over the decades (see [2] for a general overview). Here, I follow the "frame/content" perspective of speech rhythm put forward by MacNeilage [3].

The crux of this theory is under the backdrop of speech evolution: Speech production is evolved from pre-existing cyclical jaw movements in ancestral primates in the form of lip-smacking. Such cyclical lip movements were important visual-facial gestures in non-human primate communications [4]. In the course of human evolution, the coupling between mouth opening-closing cycles gradually emerged. As a result, the sonority of speech typically waxes and wanes following the mouth movement cycles. Such opening-closing alternations are organized into syllable-sized unites corresponding to the amplitude modulations, which constitute the rhythmic frames [3]. It is thus evident that the mouth opening-closing movements are pivotal to speech rhythm. In the previous Prosody Visualization Challenge (ICPhS 2019 in Melbourne), Erickson et al. [4] has presented how jaw movement patterns are indicative of prosodic differences across languages.

Here, I combine both mouth opening-closing movements and the temporal envelope to reveal the rhythmic frame using spectral coherence. Spectral coherence is a Fourier transform-based signal processing technique that captures the connectivity between both domains of signals. By nature, the coherence is nothing but a power spectrum, whose magnitude shows the correlations of both signals in the frequency domain.

2. SIGNAL PROCESSING STEPS

2.1. Extracting the temporal envelope

The audio signal (typically in the format of WAV) is bandpass filtered between 700 and 1,300 Hz to keep the vocalic energy. Then the filtered signal is full-wave rectified and further lowpass filtered at 20 Hz (This can also be achieved by re-sampling the rectified signal to 40 Hz. Thus, high frequencies not responsible for rhythmicity are naturally excluded by the Nyquist theorem).

Here, I use **ENV** to notate the temporal envelope time series.

2.2. Extracting the mouth opening-closing time series

Depending on the equipment used, mouth opening-closing functions can be extracted, for example, using a particular piece of hardware [5], or counting the number of pixels within the lip contour from a video clip [6].

Here, the mouth opening-closing time series is extracted from the electromagnetic articulography data that include the *x-y-z* coordinates from the sensors attached to the upper and lower lips in the midsagittal plane and the lip corner. The area of the triangle formulated by these three sensors is used to approximate the mouth opening magnitude at each sampled instant.[2] The whole area time series is re-sampled to 40 Hz, ad modum § 2.1 above.

I use **Om** to notate the mouth opening-closing time series.

2.3. Calculating the spectral coherence from ENV and Om

(1) The DC-biases in both **ENV** and **Om** are removed, and the Tukey window (α = .1) is used to taper the edges [7].

(2) Both signals are zero-padded to the nearest $2^{\wedge}N$ ($N \in \mathbb{Z}^{+}$), and fast Fourier transformed to yield their spectra, $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$.

(3) The cross-spectrum of both $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$ is calculated (i.e., taking the product of the Fourier coefficients of $\mathbf{FFT_{ENV}}$ and the complex conjugate of the Fourier coefficients of $\mathbf{FFT_{Om}}$). The cross-spectrum is then normalized to the product of the amplitudes of both $\mathbf{FFT_{ENV}}$ and $\mathbf{FFT_{Om}}$ to yield the spectral coherence of both **ENV** and **Om**.

The coherence illustrates the magnitudes of common periodicities in both **ENV** and **Om**. It is intrinsically a spectrum, and can thus be converted to the time domain via the inverse Fourier transform if desired.

3. VISUALIZATION

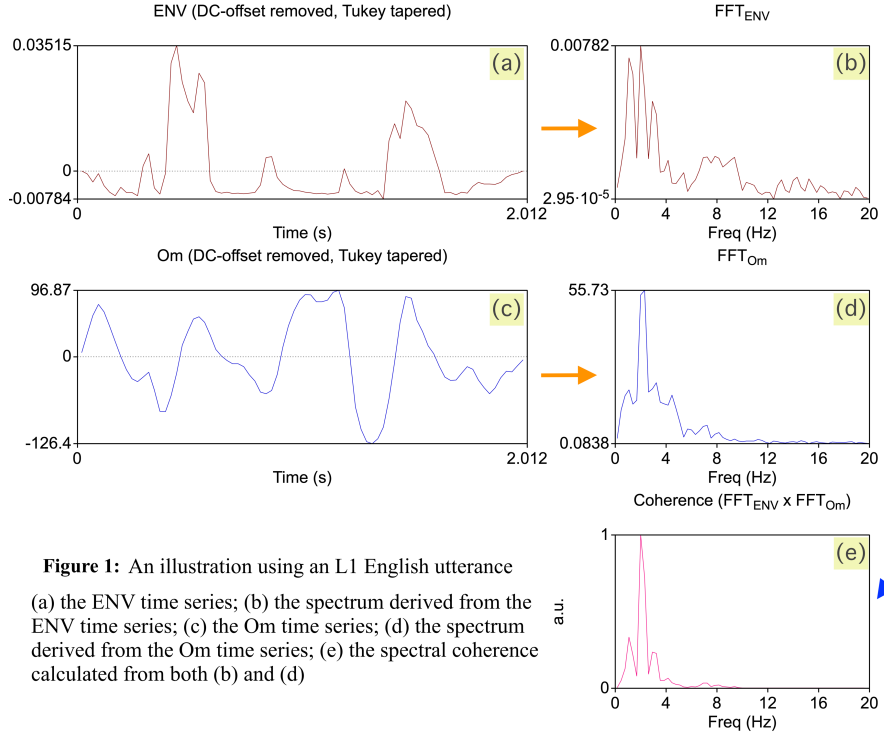

Figure 1: An illustration using an L1 English utterance

(a) the ENV time series; (b) the spectrum derived from the ENV time series; (c) the Om time series; (d) the spectrum derived from the Om time series; (e) the spectral coherence calculated from both (b) and (d)

The materials for the visualization are accessible online in a repository: https://osf.io/vq2jn/ The materials include a readme, the original waveforms, extracted **ENV** and **Om** time series, plots, and the main Praat scripts to do major signal processing steps. This repository may be updated with more materials in the future. All updates will be documented in the readme.

Figure 1 illustrates the steps used to calculate the coherence from an L1 English utterance; figure 2 illustrates the steps used to calculate the coherence from an L2 English utterance, and is only accessible in the repository due to the page limit. An initial inspection of the two coherences in both plots suggests that the joint regularities occur at a frequency of ~2 Hz for the L1 utterance (Fig.1); whereas for the L2 utterance, multiple regularities coexist (Fig. 2).

4. ACKNOWLEDGEMENTS

Swiss National Science Foundation (Grant № PZ00P1_193328) / University of Zurich Forschungskredit (Grant № FK–20–078).

5. REFERENCES

[1] A. Ji, J. J. Berry, and M. T. Johnson, 2014. The electromagnetic articulography Mandarin accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In: Proc. ICASSP, Florence (May 4–9, 2014), pp. 7769–7773.

[2] L. He, 2022. Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am*. 152, 567–579.

[3] P. F. MacNeilage, 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci*. 21, 499–511.

[4] D. Erickson, T. Huang, C. Menezes, and S. Kawahara, 2019. Using jaw movement patterns to visualize prosody. Poster presented for the Prosody Visualization Challenge (PVC-2019) at the 19th ICPhS, Melbourne (August 7–8, 2019).

[5] D. Erickson, O. Niebuhr, W. Gu, T. Huang, and P. Geng, 2020. The MARRYS cap: A new method for analyzing and teaching the importance of jaw movements in speech production. In: Proc. 12th International Seminar on Speech Production (Haskins Press, New Haven), pp. 48–51.

[6] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol*. 5, e1000436.

[7] L. He, 2022. tukey_window.praat (A Praat script to apply the Tukey window to a signal). URL: https://osf.io/6r839

[1] The EMA-MAE dataset is based on work supported by the National Science Foundation of the United States under Grant #IIS-1142826 to Marquette University, which support does not constitute an endorsement.

[2] The formula below is used to calculate the area of this triangle. In the formula, *x, y,* and *z* refer to the coordinates from sensors attached to the upper, lower lips and the lip corner (vis-à-vis the subscripts UL, LL and LC):

$$\frac{1}{2}\sqrt{\begin{vmatrix} y_{UL} & z_{UL} & 1 \\ y_{LL} & z_{LL} & 1 \\ y_{LC} & z_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} z_{UL} & x_{UL} & 1 \\ z_{LL} & x_{LL} & 1 \\ z_{LC} & x_{LC} & 1 \end{vmatrix}^2 + \begin{vmatrix} x_{UL} & y_{UL} & 1 \\ x_{LL} & y_{LL} & 1 \\ x_{LC} & y_{LC} & 1 \end{vmatrix}^2}.$$