Characterizing speech rhythm using spectral coherence between jaw displacement and speech temporal envelope

Lei He
Department of Computational Linguistics, University of Zurich,
lei.he@uzh.ch

Introduction

Lower modulation rates in the temporal envelope (ENV) of the acoustic signal are believed to be the rhythmic backbone in speech, facilitating speech comprehension in terms of neuronal entrainments at δ- and θ-rates (these rates are comparable to the foot- and syllable-rates phonetically) [e.g. 1–3]. The jaw plays the role of a carrier articulator regulating mouth opening in a quasi-cyclical way, which correspond to the low-frequency modulations as a physical consequence. This paper describes a method to examine the joint roles of jaw oscillation and ENV in realizing speech rhythm using spectral coherence. Relative powers in the frequency bands corresponding to the δ-and θ-oscillations in the coherence (respectively notated as **%δ** and **%θ**) were quantified as one possible way of revealing the amount of concomitant foot- and syllable-level rhythmicities borne by both acoustic and articulatory domains. This idea was illustrated using two English corpora (mngu0 and MOCHA-TIMIT) [4, 5] for the proof of concept. %δ and %θ were regressed on utterance duration for an initial analysis. Results showed that the degrees of foot- and syllable-sized rhythmicities are different and are contingent upon the utterance length.

Method

The mngu0 contains one male English speaker producing over 1,000 utterances, amongst which 594 in the duration range of 2–8 sec were chosen for the present study. The 2-sec cutoff allowed at least one cycle of the lowest δ frequency (.5 Hz) to be included; the 8-sec cutoff excluded sentences with medial pauses. The MOCHA-TIMIT (Wrench 1999) contains three English speakers (1f, coded as “fsew0”; 2m, coded as “maps0” and “msak0”) producing the same set of 460 sentences. Altogether 5 sentences shorter than 2 sec were excluded. All utterances were shorter than 6 sec. The EMA data were collected in the meanwhile; of particular interest to this study were the articulatory trajectories of the lower incisor.

Jaw displacements were parameterized as the Euclidean distance of the lower incisor coordinates in the mid-sagittal plane. The ENV were extracted using the full-wave rectification and low-pass filtering. The Jaw-ENV coherence spectra were calculated as the Hermitian inner product of the Fourier coefficients in the FFT of jaw displacement and the FFT of the ENV normalized to the individual power of both FFTs.

The **%δ** and **%θ** were calculated as the percentage of the spectral integral bounded by the δ-band cutoffs ($f_1 = .5$ Hz, $f_2 = 3$ Hz) or θ-band cutoffs ($f_1 = 3$ Hz, $f_2 = 9$ Hz) over the entire spectral integral of the coherence function ($f_{Nyq} = 40$ Hz).

Results

For the mngu0 data, linear regressions between utterance length and %δ and %θ were performed. The utterance duration was right skewed, hence was natural log transformed. %δ increased as utterance duration increased, whereas %θ decreased as utterance length increased (Figure 1).

For the MOCHA-TIMIT data, Random-slope models were fitted by maximum likelihood (response variables: %δ and %θ; random effects: speaker and utterance; fixed effect: utterance length). The significance of the slope estimate and between-speaker variability were tested in particular (Figure 2): in general, a positive slope estimate was found significant between %δ and utterance length, and a negative slope

estimate was found significant between %θ and utterance length. Moreover, individual differences were significant at the same time.

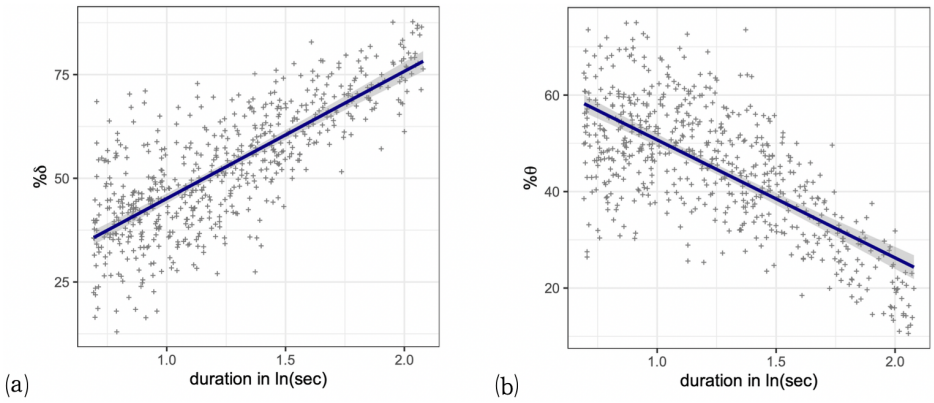

Figure 1: Regression lines and the 99% confidence intervals (shaded areas) superimposed over the scatterplots showing the relationships between %δ and log utterance duration (a), and %θ and log utterance duration (b) in the mngu0 corpus.

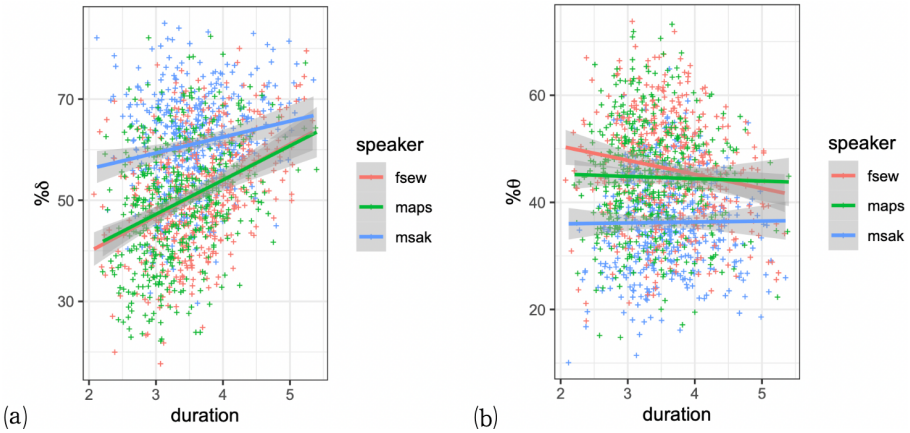

Figure 2: Regression lines and the 99% confidence intervals (shaded areas) superimposed over the scatterplots showing the relationships between %δ and utterance duration (in sec) (a), and %θ and utterance duration (b) for each of the three speakers in the MOCHA-TIMIT corpus.

References

[1] Doelling, Keith B., Luc H. Arnal, Oded Ghitza & David Poeppel. 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85(2). 761–768.

[2] Ghitza, Oded. 2017. Acoustic-driven delta rhythms as prosodic markers. Language, *Cognition and Neuroscience* 32(5). 545–561.

[3] Poeppel, David & M. Florencia Assaneo. 2020. Speech rhythm and their neural foundations. *Nature Reviews Neuroscience*. 21(6). 322–334.

[4] Richmond, Korin, Phil Hoole & Simon King. 2011. Announcing the electromagnetic articulography (Day 1) subset of the mngu0 articulatory corpus. In *Proceedings of INTERSPEECH 2011*, 1505–1508. Florence, Italy.

[5] Wrench, Alan. 1999. MOCHA MultiCHannel Articulatory database: English (MOCHA-TIMIT). http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html