LOQUENS 7(2)
December 2020, e074
eISSN 2386-2637
https://doi.org/10.3989/loquens.2020.074

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

Lei He[1], Yu Zhang[1]

[1]Department of Computational Linguistics, University of Zurich, Switzerland
lei.he@uzh.ch ORCID: https://orcid.org/0000-0002-9552-9075
yu.zhang@uzh.ch ORCID: https://orcid.org/0000-0002-0865-7897

**Submitted:** 17/12/20; **Accepted:** 18/05/21; **Published on line:** 28/12/21

**Citation / Cómo citar este artículo:** Lei He, Yu Zhang (2020). Characterizing speech rhythm using spectral coherence between jaw displacement and speech temporal envelope. *Loquens*, *7*(2), e074. https://doi.org/10.3989/loquens.2020.074

**ABSTRACT:** 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). 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 carried by both acoustic and articulatory domains. Two English corpora (mngu0 and MOCHA-TIMIT) were used 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.

**Keywords:** speech rhythm, spectral coherence, temporal envelope, jaw displacement

**RESUMEN:** *Caracterización del ritmo del habla usando la coherencia espectral entre el desplazamiento de la mandíbula y la envolvente temporal del habla.*— Se piensa que las frecuencias de modulación más bajas en la envolvente temporal (ENV) de la señal acústica constituyen la columna vertebral rítmica del habla, facilitando su comprensión a nivel de enlaces neuronales en términos de los rangos δ y θ (estos rangos son comparables fonéticamente a los rangos de pie métrico y silábicos). La mandíbula funciona como un articulador que regula la abertura de la boca de una manera cuasi cíclica, lo que se corresponde, como una consecuencia física, con las modulaciones de baja frecuencia. Este artículo describe un método para examinar el papel conjunto de la oscilación de la mandíbula y de la envolvente ENV en la producción del ritmo del habla utilizando la coherencia espectral. Las potencias relativas en las bandas de frecuencia correspondientes a las oscilaciones δ y θ en la coherencia (indicadas respectivamente como %δ y %θ) se cuantificaron como un posible modo de revelar la cantidad de ritmicidad concomitante a nivel de pie métrico y de sílaba que los dominios acústicos y articulatorios comportan. Para someter a prueba esta idea, en este estudio se analizaron dos corpus en inglés (mngu0 y MOCHA-TIMIT). Para un primer análisis, se realizó una regresión de %δ y %θ en función de la duración del enunciado. Los resultados mostraron que los grados de ritmicidad del pie y de la sílaba son diferentes y dependen de la longitud del enunciado.

**Palabras clave:** ritmo del habla, coherencia espectral, envolvente temporal, desplazamiento de la mandíbula.

## 1. INTRODUCTION

This paper characterizes speech rhythm in terms of the spectral coherence between jaw oscillations and speech temporal envelopes (ENV, henceforth). Two frequency bands in the coherence spectrum covering the neuronal δ- and θ-rates were particularly analyzed in terms of their relative contributions to the entire coherence power. These bands have been claimed to correspond to the foot- and syllable-timescales in speech and have been demonstrated to play a crucial role in neurological speech processing via brainwave-to-ENV entrainment (e.g. Doelling, Arnal, Ghitza, & Poeppel, 2014; Ghitza, 2017; Poeppel & Assaneo, 2020). This paper reports an initial analysis on the relationships between relative powers of the δ- and θ-bands in their coherence and utterance length using two English corpora: mngu0 (Richmond, Hoole, & King, 2011) and MOCHA-TIMIT (Wrench, 1999).

Historically, phoneticians described the rhythm of world languages in terms of intuitive isochronous units: stress-timed vs. syllable-timed rhythm[1] (or metaphorically, Morse code vs. machine gun rhythm[2]) (e.g. Abercrombie, 1967; Jones, 1922; Lloyd James, 1940; Pike, 1945). Failed attempts to corroborate strict isochrony instrumentally (e.g. Bertrán, 1999; Dauer, 1983; Roach, 1982; Wenk & Wioland, 1982) motivated researchers to search for acoustic correlates of different rhythmicities with regard to durational variability of different phonetic intervals, i.e. the rhythm metrics (e.g. Dellwo, 2006, 2009; Grabe & Low 2002; Ramus, Nespor, & Mehler, 1999). Meanwhile, alternative approaches to speech rhythm have also been postulated focusing on different yet interrelated physical properties in the signal: (i) prominence (e.g. Cichocki, Selouani, & Perreault, 2014; Fuchs, 2016; He, 2012, 2018; He & Dellwo, 2016; Lee & Todd, 2004), (ii) phase or power analyses of the modulation envelope (e.g. Lancia, Krasovitsky, & Stuntebeck, 2019; Leong, Stone, Turner, & Goswami, 2014; Tilsen & Arvaniti, 2013; Tilsen & Johnson, 2008) and (iii) coupling strength between feet and syllables (e.g. Barbosa, 2002; Cummins & Port, 1998; Eriksson, 1991; O'Dell & Nieminen, 1999).[3] In terms of phonological theorization, the metrical grid can be constructed based on intuitive assessment of prominent values, exhibiting the rhythmic skeleton of an utterance (e.g. Liberman & Prince, 1977; Nespor & Vogel, 1986; Selkirk, 1980).

How did rhythm evolve in speech? From a Darwinian perspective, MacNeilage (1998) held that the rhythmicity in speech evolved from pre-existing cyclical jaw movements in ancestral primates. These movements were found to be important visuofacial gestures in extant non-human primate communications (Ghazanfar, Chandrasekaran, & Morrill, 2010). It is believed that the coupling between jaw cycles and vocalization arose in the course of human evolution: the sonority of speech typically waxes and wanes with mouth opening and closing gestures (Ghazanfar et al., 2010; MacNeilage, 1998; Morrill, Paukner, Ferrari, & Ghazanfar, 2012). Such opening-closing alternations are temporally organized into syllable-sized units corresponding to the ENV modulations, which constitute the rhythmic "frames"; the open and closed phases are filled with vocalic and consonantal "contents" — the frame/content theory of speech evolution (MacNeilage, 1998). By calculating ENV spectra (e.g. Tilsen & Johnson 2008) or the syllable intensity variability (e.g. He, 2018), the characteristics related to the rhythmic "frames" can be revealed; by calculating the durational variability of vocalic and consonantal intervals (e.g. Grabe & Low, 2002; Ramus et al., 1999), the characteristics related to the rhythmic "contents" may be evaluated.

Speech rhythm is not evolutionarily redundant; it is functional in the neurological processing of the speech signal. The recurring oscillations in the ENV – which supposedly reflect the rhythmic frames – facilitate the brain to parse the incoming speech signal for comprehension. It has been demonstrated that the δ-oscillation (.5–3 Hz, corresponding to foot/stress rates) and θ-oscillation (3–9 Hz, corresponding to syllable rates) in the auditory cortex entrain to the speech ENV at these modulation rates (Doelling et al., 2014; Ghitza, 2017; Giraud & Poeppel, 2012; Strauß & Schwartz, 2017). These slow neuronal oscillations formulated a temporal window structure whereby the auditory cortex tracks the speech signal at the foot and syllable rates. Within such longer temporal windows, information encoded in finer timescales (e.g. phonemes up to ~40 Hz, corresponding to the γ-oscillation) can then be processed to achieve comprehension (Doelling et al., 2014; Giraud & Poeppel, 2012).

The motor knowledge of speech production is arguably indispensable in the neurological processing of speech signals (Strauß & Schwartz, 2017). The jaw performs the role of a carrier articulator responsible for lower modulation frequencies that may correspond to the rhythmic frames (Strauß & Schwartz, 2017) to which the slower neuronal oscillations can be phase-locked, not only in the auditory cortex, but also in the visual cortex (Park, Kayser, Thut, & Gross, 2016). Seeing the speaker's mouth movements facilitates the listener to understand speech, particularly in adverse conditions with excessive noise (Park et al., 2016[4]). The mouth movements help the listener visually access the rhythmic structure, like visual scaffolding. Therefore, the jaw

[1] Quintessential "stress-timed" languages include the Germanic languages, and "syllable-timed" languages, the Romance languages.

[2] Arthur Lloyd James illustrated the "Morse code" rhythm of English to a foreign student whose native language was Sinhalese in a historical film *48 Paddington Street* archived by British Pathé (URL: https://www.britishpathe.com/video/48-paddington-street/, accessed 2 January 2020). Lloyd James' patronizing manner in the film is least appreciated though.

[3] The crux of all these approaches is the consensus of revealing different rhythmicities through different forms of variability in the speech signal. Variability can either be quantified via different physical quantities, such as duration (e.g. Dellwo, 2006, 2009; Grabe & Low, 2002; Ramus et al., 1999), intensity (e.g. Cichocki et al., 2014; Fuchs, 2016; He, 2012, 2018; He & Dellwo, 2016), and a mixture of various parameters (e.g. Lee & Todd, 2004); or evaluated through the coordination between prosodic hierarchies, such as the phase difference between syllable- and stress/word-level timescales (e.g. Lancia et al., 2019; Leong et al., 2014), the power of recurring frequencies in the ENV (e.g. Tilsen & Arvaniti, 2013; Tilsen & Johnson, 2008), and a linear relationship between syllable and feet durations (e.g. Barbosa, 2002; Eriksson, 1991; O'Dell & Nieminen, 1999).

[4] Park et al. (2016) did not examine the jaw movements *per se*, but the size of mouth opening. Although other factors such as the lip rounding or protrusion also affect the mouth aperture, the principal determinant of the mouth area is the jaw oscillation. It is sensible to include the whole mouth in the visual stimuli as it resembles personal communications more naturally. However, to characterize the spectro-temporal features of the jaw movements, it is more appropriate to measure the kinematics of the jaw free from the interference of other articulators.

as a carrier articulator plays an important role in both production and perception of speech rhythm; the temporal windows facilitating the neuronal entrainment to the speech ENV must be discoverable in the jaw oscillation as well. However, the roles of the jaw and ENV have been disjointedly studied: The jaw displacement has been shown to well explain the metrical structure of the utterance (Erickson, Suemitsu, Shibuya, & Tiede, 2012; Erickson & Kawahara, 2016; Huang & Erickson, 2019). The ENV has been extensively investigated in terms of its recurring patterns (He, 2018; Tilsen & Arvaniti, 2013; Tilsen & Johnson, 2008) and synchronizations between different modulation rates (Cummins & Port, 1998; Lancia et al., 2019; Leong et al., 2014).

We thus propose to characterize speech rhythm using the spectral coherence between the jaw oscillation and speech ENV (hereinafter, JAW-ENV coherence). A spectral coherence is a Fourier transform-based representation that quantifies common periodicities in two signals. It evaluates the correlation between these two signals in the frequency domain, hence its advantage over assessing simple correlations in the time domain.[5] A similar approach has been attempted, though, by calculating the coherence between the ENV and mouth opening size in terms of the number of pixels shrouded by the lip contour or the inter-lip distance (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009); however, the roles of the jaw elevation and depression and peripheral lip gestures could not be disentangled thereof (also see footnote 4). This study, instead, examined the sole role of the jaw movement.

Since the lower frequency components pertaining to the δ- and θ-oscillations are crucial for the neurological speech processing in the auditory, visual and motor cortices, the jaw oscillation and the ENV should be coherent in these frequency ranges. The degree of such coherence is measurable in terms of the percentage of the spectral integral bounded by the δ- or θ-band cutoffs out of the entire spectral integral of jaw-env coherence (notated as %δ and %θ, see Eq. (1) in §2.3). These two measures capture the relative amount of power shared by the jaw oscillation and ENV in terms of regularities at the frequency bands corresponding to the neuronal δ- and θ-samplings. Moreover, %δ and %θ are analyzed as a function of the utterance length (§3), because the rhythmic structure is more likely to evolve into a more complex pattern over time (a 5-sec utterance would intuitively have a more complex rhythmic structure than a 1-sec utterance “Hello!” which contains a single iamb). It is expected that higher %δ is associated with longer utterances, because more sizeable prosodic boundaries (including foot-sized timescales) may be included; for an utterance with higher %δ, a smaller %θ is expected because the total power of jaw-env coherence is fixed, and determined by the joint temporal amplitudes of both jaw oscillation and ENV (in reference to Parseval’s theorem of energy conservation).

## 2. METHOD

### 2.1 The corpora

The mngu0 (Richmond et al., 2011) 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. For both corpora, the electromagnetic articulograph (Carstens AG500 for mngu0 and AG100 for MOCHA-TIMIT) was used to record the kinematic trajectories of various articulators (with 200 Hz or 500 Hz temporal resolutions) together with the audio speech signal (16-bit @ 16 kHz). All kinematic data were head-corrected and translated to a new Cartesian coordinate system in the midsagittal plane. Sensor histories data from the lower incisor were used for the jaw movements for the study.

### 2.2. Calculating JAW-ENV coherence

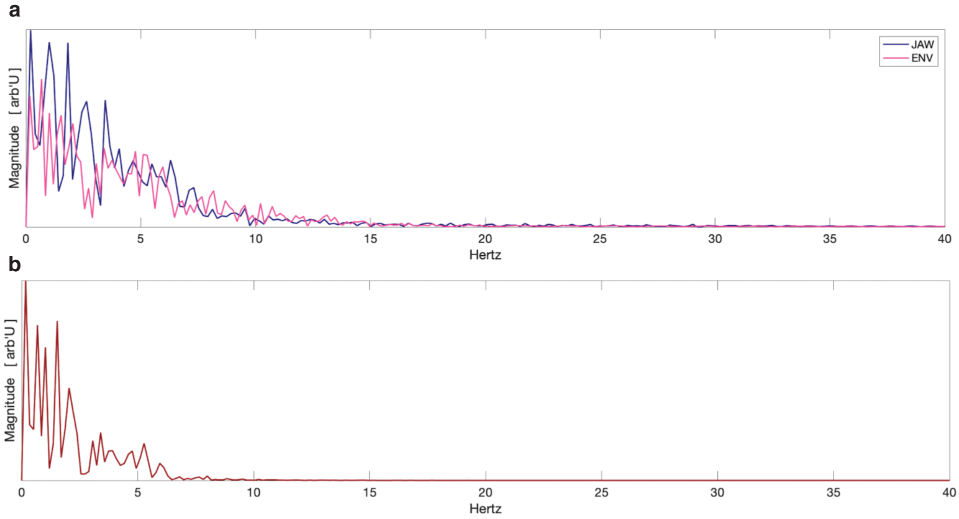
Jaw-env coherences were calculated following three steps using Matlab® R2018b:

(i) Obtaining the spectra of the jaw oscillation functions (the matrix $\mathbf{FFT}_{\mathrm{JAW}}$) for each utterance. First, the jaw oscillation time series were estimated as the Euclidean distances of the lower incisor coordinates to zero (the vector $\mathbf{d}_{\mathrm{JAW}}$). To obtain $\mathbf{FFT}_{\mathrm{JAW}}$, a 512-point fast Fourier transform was applied to $\mathbf{d}_{\mathrm{JAW}}$ which had been offset-removed, down-sampled to 80 Hz, cosine-tapered ($\alpha = .1$), and zero-padded. The magnitude of $\mathbf{FFT}_{\mathrm{JAW}}$ was then linearly normalized in 1 arbitrary unit (arb’U, henceforth).

(ii) Obtaining the spectrum of the speech ENV (the matrix $\mathbf{FFT}_{\mathrm{ENV}}$). First, a “beat” detection filter (Cummins & Port, 1998; Tilsen & Johnson, 2008) was applied to the speech signal (first-order Butterworth, center frequency = 1,000 Hz, bandwidth = 300 Hz) to keep the vocalic energy while removing the glottal energy and obstruent noise. Then, the filtered signal was full-wave rectified and further bandpass filtered (fourth-order Butterworth, center frequency = 5 Hz, bandwidth = 10 Hz) to obtain the ENV. To obtain $\mathbf{FFT}_{\mathrm{ENV}}$, the ENV was offset-removed, down-sampled to 80 Hz, cosine-tapered ($\alpha = .1$), zero-padded, and supplied to a 512-point fast Fourier transform. The magnitude of $\mathbf{FFT}_{\mathrm{ENV}}$ was then linearly normalized in 1 arb’U. The object obtained this way is called the beat histogram in music information retrieval (Lykartsis & Lerch, 2015).

[5] In fact, simple correlations for time-series data are problematic in general with spuriously high correlation coefficients.

**Figura 1:** The spectra of the jaw oscillation and the speech ENV (a); the JAW-ENV coherence calculated from the spectra of the jaw oscillation and the speech ENV (b).



(iii) The jaw-env coherence (the matrix $\mathbf{COH}_{\text{JAW-ENV}}$) was calculated as the Hermitian inner product[6] of the Fourier coefficients in $\mathbf{FFT}_{\text{JAW}}$ and $\mathbf{FFT}_{\text{ENV}}$ normalized to the individual power of $\mathbf{FFT}_{\text{JAW}}$ and $\mathbf{FFT}_{\text{ENV}}$ (a code snippet in Cohen, 2017 was applied); negative frequencies were neglected. Figure 1 shows an example of calculating the jaw-env coherence from the spectra of jaw oscillation and the speech ENV. This process computes the common periodicities in two signals by evaluating the correlation between these two signals in the frequency domain.

### 2.3. Calculating %δ and %θ in jaw-env coherence

Eq. (1) illustrates the conceptual calculations of %δ and %θ — 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 $C$(f) ($f_{\text{Nyq}}$ = 40 Hz). The Nyquist frequency ($f_{\text{Nyq}}$) of 40 Hz was arbitrarily chosen at the upper γ-band boundary responsible for processing phonemes and smaller features. Empirically, the frequency granularity ($df$) is equal to 2 × $f_{\text{Nyq}}$ (40 Hz) ÷ FFT points (512) = .16 Hz. Because of the frequency discretization, the coherence function $C$(f) is effectively the matrix $\mathbf{COH}_{\text{JAW-ENV}}$. The integrals (approximated using Riemann sums) can be calculated through iterations at the step size of $df$ in $\mathbf{COH}_{\text{JAW-ENV.}}$

$$\%\delta \text{ or } \%\theta = \frac{\int_{f_1}^{f_2} C(f)\,df}{\int_{0}^{f_{\text{Nyq}}} C(f)\,df} \times 100$$

## 3. DATA ANALYSES AND RESULTS[7]

For the mngu0 data, simple linear regressions between utterance length and %δ and %θ were performed using R. The utterance duration was right skewed, hence was natural log transformed. Table 1 and Figure 2 illustrate the results: %δ increased as utterance duration increased, whereas %θ decreased as utterance length increased, conforming to the expectation.

The MOCHA-TIMIT data were subsequently analyzed to examine whether consistent results would be obtained. Random-slope models were fitted by maximum likelihood (response variables: %δ and %θ; random effects: speaker and utterance; fixed effect: utterance length) using R{*lme4*, v1.1–21} (Bates, Mächler, Bolker, & Walker, 2015). The significance of the slope estimate and between-speaker variability were tested in particular (see Table 2 and Figure 3): 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.

## 4. DISCUSSION

This paper introduced a method to characterize speech rhythm using spectral coherence between jaw oscillation and the speech ENV, i.e. the jaw-env coherence. It provides a spectro-temporal representation of the common periodicities in both signals. Two frequency bands corresponding to the brain δ- and θ-oscillations were analyzed in terms of the percentage of power accounted for by these two bands in jaw-env coherence, i.e. %δ and %θ. In general, utterance length was found to be a significant

[6] The Hermitian inner product of two signals (in this case, the jaw oscillation and speech ENV) is simply the multiplications of the Fourier coefficients of the first signal and the complex conjugates (sign change of the imaginary part) of the Fourier coefficients of the second one. It reveals the covariance between the two signals in the frequency domain.

[7] A more stringent α-level (= .01) was chosen in statistical analyses to reduce the chance of false positive findings.

**Table 1:** Results of linear regression analyses for the mngu0 data.

| Model (Y~X) | F-test of overall significance | | t-test of estimated slope | | |
|---|---|---|---|---|---|
| | F(DoFs) | p | β | 99% CI | \|t\| |
| %δ ~ ln(utterance duration) | 881.5(1,592) | ≪ .01 | 30.66 | 28.00, 33.32 | > 2.576 |
| %θ ~ ln(utterance duration) | 545.5(1,592) | ≪ .01 | –24.44 | –27.14, –21.75 | > 2.576 |

**Figura 2:** 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. Log durationvis-à-vis linear duration at abscissa tick marks in both subplots: .75 ln(sec) ⇌ 2.12 sec, 1.0 ln(sec) ⇌ 2.72 sec, 1.25 ln(sec) ⇌ 3.49 sec, 1.5 ln(sec) ⇌ 4.48 sec, 1.75 ln(sec) ⇌ 5.75 sec, and 2.0 ln(sec) ⇌ 7.39 sec.

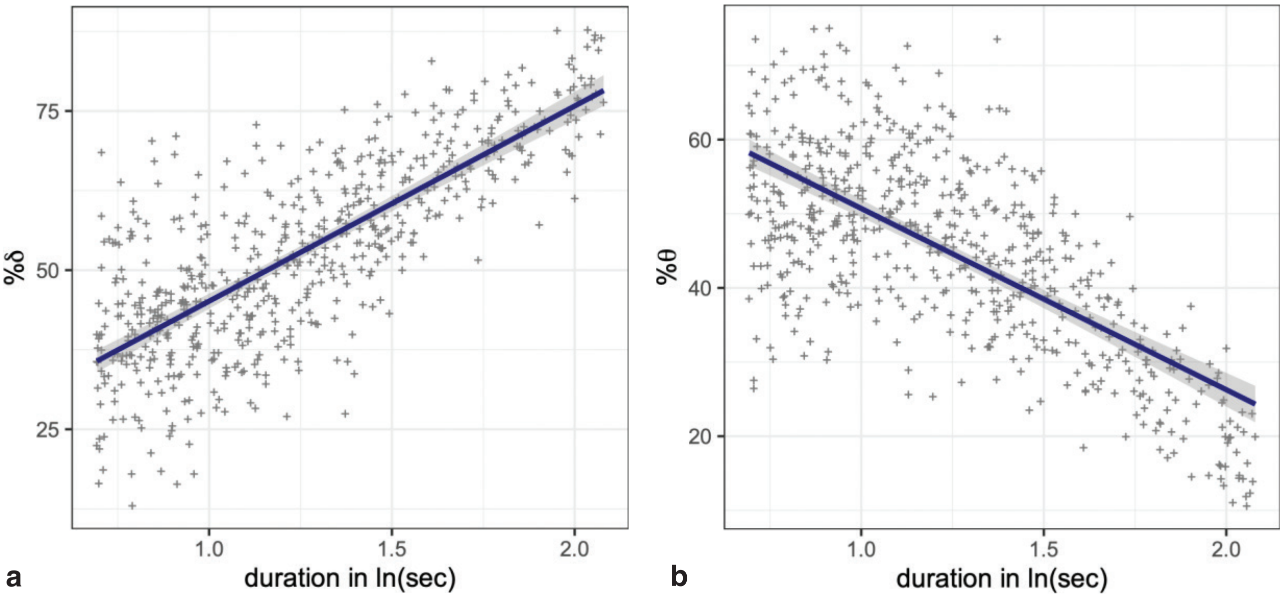


**Table 2:** Results of random-slope models for the MOCHA-TIMIT data.

| Response variable | Fixed effect: utterance length | | | Random effect: speaker [a] | | | |
|---|---|---|---|---|---|---|---|
| | β | 99% CI | \|t\| | AIC (full; reduced) | –LogLik (full; reduced) | χ2(DoF) | p |
| %δ | 8.38 | 5.49, 11.27 | > 2.576 | 10121; 10510 | 5248.8; 5051.5 | 394.47(3) | ≪ .01 |
| %θ | –2.89 | –5.16, –.62 | > 2.576 | 10038; 10363 | 5009.7; 5175.7 | 331.92(3) | ≪ .01 |

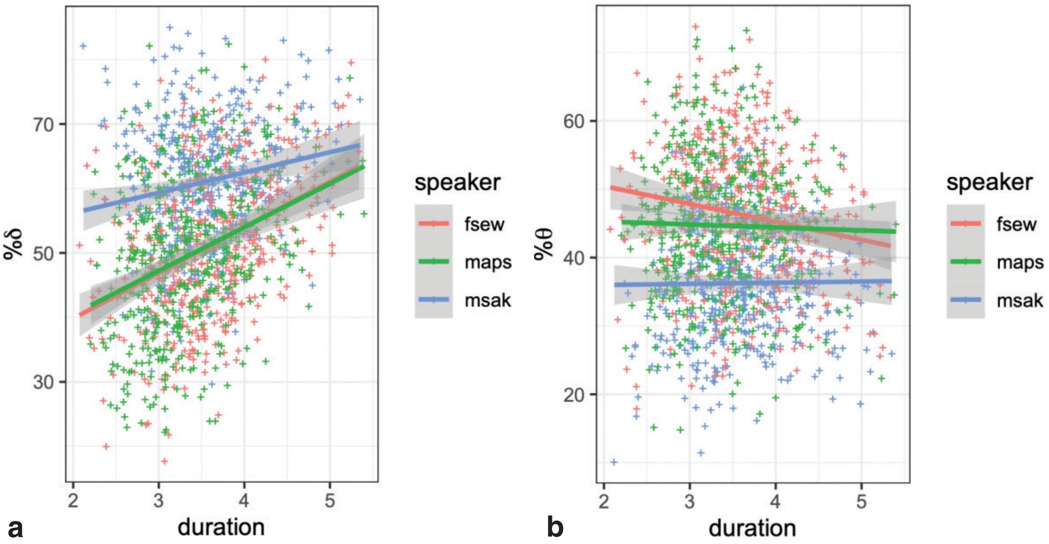
[a] Likelihood ratio test was used to test between-speaker variability between the full model and the speaker-reduced model. The AICs of the full models were smaller than those of the reduced models, suggesting that the full models had better fits. The χ2 values were calculated as the differences between twice the –LogLik of the full and reduced models (the differences of the deviances).

predictor of %δ and %θ, yet individual differences must not be neglected. The findings have several implications:

(i) The jaw oscillation and speech ENV possess strong spectral coherence in the low frequency bands of .5 – 3 Hz and 3 – 9 Hz. This upholds the role of jaw movement and speech ENV in speech rhythmicity. The semi-cyclical jaw movements constantly change the amount of radiated energy corresponding to the lower modulation frequencies in the speech signal, to which the auditory cortex of the listener entrains at the δ- and θ-rates. (Doelling et al., 2014; Ghitza, 2017; Giraud & Poeppel, 2012; Strauß & Schwartz, 2017). The jaw movements also invite neuronal entrainment in the listener's visual cortex (Park et al., 2016). These entrainments play a useful role in speech processing and comprehension.

(ii) Both .5 – 3 Hz and 3 – 9 Hz bands (pertaining to the δ- and θ-rates) are represented in jaw-env coherence, but in different degrees as measured by %δ and %θ. This

**Figura 3:** 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.



suggests that different levels of rhythmicities (including foot-sized and syllable-sized) are present simultaneously but differ in degrees. The amount of regularities at a larger timescale increases as the utterance length increases for all speakers from the two corpora (Figures 2 and 3). It is possible that longer utterances are more likely to contain larger prosodic boundaries or more extreme intonational accents, which would increase the power pertaining to the δ-band. This may have a functional advantage: higher δ-rate regularities would facilitate sensory chunking of a longer utterance under the neuronal δ-sampling (an example of sensory chunking is using temporal groupings when memorizing a series of digits or syllables) (Boucher, Gilbert, & Jemel, 2019). Smaller units pertaining to faster rates (e.g., syllables, phonemes or even phonological features) could be processed within each δ-window.

(iii) Individual differences are conspicuous in %δ and %θ as a function of utterance duration. The amounts of regularities in δ- and θ-bands are inversely proportional for the mngu0 speaker as well as speaker “fsew” in MOCHA-TIMIT (Figures 2 and 3), possibly because more δ power has already taken up the majority of power in jaw-env coherence in longer utterances, leaving little power for the θ-band regularity. For speaker “msak” in MOCHA-TIMIT, δ-band regularities were already prominent even in short sentences (high intercept of “msak” in Figure 3a), leaving little power for syllable-sized frequencies regardless of the utterance length (low intercept and flat slope of “msak” in Figure 3b). Nevertheless, to investigate individual differences fully, it is mandatory to increase the sample size significantly.

(iv) The results may also explain why early phoneticians (e.g. Abercrombie, 1967; Jones, 1922; Lloyd James, 1940; Pike, 1945), despite having undergone rigorous ear training, would still inaccurately describe languages such as English as possessing isochronous feet. Higher %δ may be a strong cue to foot-sized regularity in both jaw oscillation and speech temporal modulation. For all speakers analyzed in the study, a large amount of foot-sized regularity has been found. It is likely that early phoneticians have discerned such foot-sized regularity in English, yet unfortunately described it in absolute terms as “stress-timed.”

This study has limitations too:

(i) In terms of data variance, all speakers in the MOCHA-TIMIT corpus showed bigger variances than the mngu0 speaker (cf. Figures 2 and 3). This may be due to the data inconsistency issue of the MOCHA-TIMIT corpus. It has been demonstrated that even for a relatively stationary sensor at the velum, a tremendous amount of data inconsistency existed (Richmond, 2009; Richmond et al., 2011). Technical issues with respect to the early generation of the electromagnetic articulograph may be the culprit (Richmond, 2009).

(ii) The two frequency bands analyzed in this study were informed by the low-neuronal oscillations that have been shown to play a key role in the rhythmic parsing in speech processing. Apart from considering these two bands as pertaining to the stress-rate or syllable-rate, further research still needs to be done to assess whether these frequency cutoffs are justifiable in linguistic/phonetic terms.

(iii) The corpora adopted in this study were small in terms of the number of speakers, and only English was analyzed. This reduced the generalizability of this study.

For future research, it is imperative to test the method using more speakers from different languages, including those traditionally labeled as “syllable-timed.” That they have been described as “syllable-timed” may be due to a high degree of syllable-sized cyclicity in jaw oscillations and speech temporal modulations (measurable as high %θ in jaw-env coherence) even in longer sentences. So far, the coherence of the jaw oscillation and ENV has been investigated based on the power spectra. It will also be interesting to explore the coherence based on the phase spectra from multi-domain signals, including acoustic, articulatory and neurological, to further explore their temporal relationships in constituting speech rhythmicity both at the production and perception levels.

## 5. ACKNOWLEDGEMENTS

This study was supported by the Forschungskredit of the University of Zurich (Grant FK-19-069 to YZ and Grant FK-20-078 to LH). It was also benefited from a completed project from the Swiss National Science Foundation (Grant P2ZHP1_178109 to LH). We thank Alejandra Pesantez for her great help in the Spanish abstract.

## REFERENCES

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Barbosa, P. A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2002* (pp. 163–166). Aix-en-Provence, France: Laboratoire Parole et Langage, SProSIG.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bertrán, A. P. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design, 2*, 103–131.

Boucher, V. J., Gilbert, A. C., & Jemel, B. (2019). The role of low-frequency neural oscillations in speech processing: Revising delta entrainment. *Journal of Cognitive Neuroscience, 31*(8), 1205–1215. http://dx.doi.org/10.1162/jocn_a_01410

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology, 5*(7), e1000436. http://dx.doi.org/10.1371/journal.pcbi.1000436

Cichocki, W., Selouani, S.-A., & Perreault, Y. (2014). Measuring rhythm in dialects of New Brunswick French: Is there a role for intensity? *Canadian Acoustics · Acoustique Canadienne, 42*(3), 90–91.

Cohen, M. X. (2017). *Matlab for Brain and Cognitive Scientists*. Cambridge, MA: MIT Press.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*(2), 145–171. http://dx.doi.org/10.1006/jpho.1998.0070

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*, 51–62. http://dx.doi.org/10.1016/S0095-4470(19)30776-4

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In P. Karnowski & I. Szigeti (Eds.), *Sprache und Sprachverarbeitung — Language and language-processing* (Linguistik International 15, pp. 231–241). Frankfurt a/M: Peter Lang.

Dellwo, V. (2009). Choosing the right rate normalization method for measurements of speech rhythm. In S. Schmid, M. Schwarzenbach & D. Studer (Eds.), *La dimensione temporale del parlato: Atti del 5° Convegno Nazionale AISV 2009* (pp. 13–32). Torriana: EDK Editore.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage, 85*(2), 761–768. http://dx.doi.org/10.1016/j.neuroimage.2013.06.035

Erickson, D., & Kawahara, S. (2016). Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistics Vanguard, 2*(1), 20150025. http://dx.doi.org/10.1515/lingvan-2015-0025

Erickson, D., Suemitsu, A., Shibuya, Y., & Tiede, M. (2012). Metrical structure and production of English rhythm. *Phonetica, 69*(3), 180–190. http://dx.doi.org/10.1159/000342417

Eriksson, A. (1991). *Aspects of Swedish Speech Rhythm* (Gothenburg monographs in linguistics 9). Gothenburg: University of Gothenburg Dissertation.

Fuchs, R. (2016). *Speech Rhythm in Varieties of English*. Singapore: Springer.

Ghazanfar, A. A., Chandrasekaran, C., & Morrill, R. J. (2010). Dynamic, rhythmic facial expressions and the superior temporal sulcus of macaque monkeys: Implications for the evolution of audiovisual speech. *European Journal of Neuroscience, 31*(10), 1807–1817. http://dx.doi.org/10.1111/j.1460-9568.2010.07209.x

Ghitza, O. (2017). Acoustic-driven delta rhythms as prosodic markers. *Language, Cognition and Neuroscience, 32*(5), 545–561. http://dx.doi.org/10.1080/23273798.2016.1232419

Giraud, A-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience 15*(4), 511–517. http://dx.doi.org/10.1038/nn.3063

Grabe, E., & Low, E. L. (2002). Durational variability in speech and rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515–543). Berlin & New York: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110197105.515

He, L. (2012). Syllabic intensity variations as quantification of speech rhythm: Evidence from both L1 and L2. In Q. Ma, H. Ding & D. Hirst (Eds.), *Proceedings of Speech Prosody 2012* (pp. 466–469). Shanghai, China.

He, L. (2018). Development of speech rhythm in first language: The role of syllable intensity variability. *Journal of the Acoustical Society of America, 143*(6), EL463–EL467. http://dx.doi.org/10.1121/1.5042083

He, L., & Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *International Journal of Speech, Language and the Law, 23*(2), 243–273. http://dx.doi.org/10.1558/ijsll.v23i2.30345

Huang, T., & Erickson, D. (2019). Articulation of English "prominence" by L1 (English) and L2 (French) speaker. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS-19)*, paper 134, Melbourne, Australia.

Jones, D. (1922). *An Outline of English Phonetics*. New York: G. E. Stechert & Co.

Lancia, L., Krasovitsky, G., & Stuntebeck, F. (2019). Coordinative patterns underlying cross-linguistic rhythmic differences. *Journal of Phonetics, 72*, 66–80. http://dx.doi.org/10.1016/j.wocn.2018.08.004

Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory "primal sketch" to two multi-language corpora. *Cognition, 93*(3), 225–254. http://dx.doi.org/10.1016/j.cognition.2003.10.012

Leong, V., Stone, M. A., Turner, R. E., & Goswami, U. (2014). A role for amplitude modulation phase relationships in speech rhythm perception. *Journal of the Acoustical Society of America, 136*(1), 366–381. http://dx.doi.org/10.1121/1.4883366

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8*(2), 249–336.

Lloyd James, A. (1940). *Speech Signals in Telephony*. London: Sir I. Pitman.

Lykartsis, A., & Lerch, A. (2015). Beat histogram features for rhythm-based musical genre classification using multiple novelty functions. In P. Svensson & U. Kristiansen (Eds.), *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx)*, paper 42. Trondheim, Norway:

Department of Music and Department of Electronics and Telecommunications. Norwegian University of Science and Technology.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences, 21*(4), 499–511. http://dx.doi.org/10.1017/S0140525X98001265

Morrill, R. J., Paukner, A., Ferrari, P. F., & Ghazanfar, A. A. (2012). Monkey lipsmacking develops like the human speech rhythm. *Developmental Science, 15*(4), 557–568. http://dx.doi.org/10.1111/j.1467-7687.2012.01149.x

Nespor, M., & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS-14)*, pp. 1075–1078. San Francisco, California: University of California.

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife, 5*, e14521. http://dx.doi.org/10.7554/eLife.14521

Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Poeppel, D., & M. Assaneo, M. F. (2020). Speech rhythm and their neural foundations. *Nature Reviews Neuroscience, 21*(6*)*, 322–334. http://dx.doi.org/10.1038/s41583-020-0304-4

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265–292. http://dx.doi.org/10.1016/S0010-0277(99)00058-X

Richmond, K. (2009). Preliminary inversion mapping results with a new EMA corpus. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association* - *INTERSPEECH 2009* (pp. 2835–2838). Brighton, UK.: ISCA Archive, http://www.isca-speech.org/archive/interspeech_2009.

Richmond, K., Hoole, P., & King, S. (2011). Announcing the electromagnetic articulography (Day 1) subset of the mngu0 articulatory corpus. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *Proceedings of the 12th Annual Conference of the International Speech Communication Association* - *INTERSPEECH 2011* (pp. 1505–1508). Florence, Italy: ISCA Archive, http://www.isca-speech.org/archive/interspeech_2011

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (Ed.), *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F. R. Palmer* (pp. 73–79). London: Edwards Arnold.

Selkirk, E. O. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry, 11*(3), 563–605.

Strauß, A., & Schwartz, J-L. (2017). The syllable in the light of motor skills and neural oscillations. *Language, Cognition and Neuroscience, 32*(5), 562–569. http://dx.doi.org/10.1080/23273798.2016.1253852

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America, 134*(1), 628–639. http://dx.doi.org/10.1121/1.4807565

Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America, 124*(2), EL34–EL39. http://dx.doi.org/10.1121/1.2947626

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*(2), 193–216. http://dx.doi.org/10.1016/S0095-4470(19)30957-X.

Wrench, A. (1999). MOCHA MultiCHannel Articulatory database: English (MOCHA-TIMIT). http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (accessed 25 December 2018).