VIABLE SIGNAL PERIODICITIES IN SPEECH RHYTHM

Jessica Campbell, Dani Byrd, Louis Goldstein

University of Southern California, United States
jac95339@usc.edu, dbyrd@usc.edu, louisgol@usc.edu

ABSTRACT

Despite well-known rhythmicities in spoken language, neither strict nor absolute isochrony is present; linguistic structures are not located at equal intervals. An alternative periodicity-based approach to linguistic rhythm instead considers quasicyclic or oscillatory properties of the speech signal—in articulation or acoustics—over extended stretches of speech. Additionally, while periodicity in speech acoustics has been implicated in neural entrainment, it is unknown if periodicity in the articulatory domain is also present in a way that could promote intelligibility. In the present study, the stability of an articulatory spatiotemporal modulation function calculated from kinematic point tracking data and a correlated acoustic modulation function calculated from MFCCs is compared to the stability of amplitude envelope modulation, which is shown to reflect vowel onsets and lexical stresses. Results demonstrate that both types of modulation functions are comparable in stability to the amplitude envelope, indicating their potential viability for speech perception and/or production processing.

Keywords: Rhythm, Modulation, Amplitude Envelope, Articulation, Speech Timing

1. INTRODUCTION

While language exhibits rhythmic characteristics, strict or absolute isochrony is not present. An analytic approach that leverages the stability of quasi-cyclic or oscillatory properties of the speech signal—either in production or acoustics—over longer stretches of speech can provide a means for defining rhythm as durational regularity. Such a global (i.e., long timescale) approach to modulation can be calculated irrespective of language, speaking task, or speaker and may be influenced by or reflect linguistic structures [7], [9], [17], [14]. We call such an approach a "periodicity approach."

Models of rhythm incorporating a periodicity approach vary regarding which specific oscillatory properties they index. In the acoustic domain, these oscillatory properties tend to match frequencies at which neural entrainment occurs [12], [1], [8], [14], [9] and thus may aid in speech perception [5], [16], [3]. The most common periodicity approach uses the fluctuating speech acoustic amplitude envelope (e.g., [17], [14], [11], [4]). In Tilsen and Arvaniti's [17] acoustic amplitude mode decomposition method, a series of filters is applied to the speech signal to extract the quasi-cyclic fluctuations of amplitude during the speech stream. The amplitude envelope can then be decomposed using a sifting process into a series of *intrinsic mode functions*, which reflect amplitude oscillations at increasingly longer timescales [17]. The first (i.e., shortest timescale) of these functions roughly corresponds to the syllable-level timescale, and the second to the stress-foot-level time scale; however, Tilsen and Arvaniti [17, p. 631] establish this correspondence only qualitatively, stating that they "observ[ed] it to hold true in the majority of cases…examined by inspection."

Oganian and Chang [14], in contrast, calculate the derivative of the positive amplitude envelope, so that instead of fluctuations in amplitude, the derived global signal captures fluctuation in instantaneous amplitude change. The instants of peak change in amplitude, they argue, are likely involved in speech perception, because they found that these peak "events" produce time-locked response in the human superior temporal gyrus [14]. Periodicity approaches' relevance to neural events may aid in their interpretability in terms of the cognitive activities involved in speech production and perception (e.g., [7]).

While speech amplitude envelope methods of tracking global periodicity capture oscillations of the acoustic signal, another method for indexing speech rhythm is feasible—the spatiotemporal modulation function approach [6]. This approach can capture rhythmicity in either an articulatory or an acoustic signal, quantifying the oscillatory properties of the rate of global change in the vocal tract or in the acoustic spectrum over time. We speculate additionally that neural responses can entrain to the modulation function (in addition to or alternatively to the amplitude envelope), which would potentially implicate the modulation signal in production, perception, and learning processes. That said, for entrainment to be possible, the modulation pulse frequency must fall in an appropriate and sufficiently stable frequency range across and within speakers. Specifically, for viability in linguistic neurocognitive processes, the appropriate frequency range would be expected to be in or near the theta band range, 4-8Hz [15]. In sum, temporal stability in articulatory and

concomitant acoustic modulation pulse rate is predicted across speakers such that the average frequency of modulation is expected to be resistant to variation, as other periodicity approaches have been found to be. The present study uses articulatory and acoustic modulation data to assess this prediction.

2. METHODS

Readings of the "Grandfather Passage" [2] by nine speakers were sourced from the X-Ray Microbeam (XRMB) Corpus [19]. Two analyses were conducted. The first compares the frequencies and temporal stability of the articulatory and acoustic modulation pulses to those of peaks in the differentiated amplitude envelope. The second analysis explores linguistic drivers of this periodicity by comparing the frequencies of acoustic vowel and stressed vowel onsets to the first and second intrinsic mode functions of the amplitude envelope, respectively.

2.1. Derived variables

The articulatory modulation function was calculated to index instantaneous change of the global articulatory posture over time as captured by pellet positioning on the articulators, and the acoustic modulation function was calculated to index instantaneous spectral change over time as captured by MFCCs. These calculations were conducted following [6], which also used XRMB data. Specifically, the articulatory modulation function was calculated as the squared Euclidean distance between the 14-dimensional vector of XRMB marker positions at successive time samples. To calculate the acoustic modulation function, the difference across successive frames of the first 13 MFCC parameters were squared and summed. Each function was then smoothed using a low-pass 12 Hz-cutoff, ninth-order Butterworth filter. "Pulses" are defined as peaks in a modulation function and were calculated by identifying local maxima of the function.

The acoustic amplitude envelope was calculated following [17] for each speech analysis interval (see below). Specifically, the acoustic signal was bandpass filtered using a fourth order Butterworth filter with cutoffs [400, 4000] Hz to determine vocalic energy, then lowpass filtered with a fourth order 10Hz Butterworth filter. The function was then normalized by subtracting the mean and dividing by the maximum of the function and windowed using a Tukey window. To make the signal comparable to the modulation functions, the amplitude envelope was differentiated and then filtered again with a fifth order 12Hz lowpass Butterworth filter. To evaluate the correlation between the intrinsic mode functions and driving linguistic structures posited by [17], the non-differentiated function was sifted using Matlab's *emd* function into two intrinsic mode functions (S. Tilsen, personal communication, February 6, 2022), which were also filtered with a fifth order 12 Hz lowpass filter. Peaks in the differentiated amplitude envelope, as well as the first two intrinsic functions of the non-differentiated amplitude envelope, were designated as local maxima in the function.

Means and variability in frequency were calculated for both articulatory and acoustic modulation pulses and peaks in the differentiated amplitude envelope for each analysis interval (defined below). The reciprocal of the periods between timepoints provided frequencies for each timeseries. The mean frequencies of the first and second intrinsic mode functions of the amplitude envelope and of acoustic vowel onsets and stressed vowel onsets (defined below) were likewise calculated.

2.2. Analysis intervals & defining linguistic timepoints

Specific windows of speech analysis were determined based on phrasal groupings by identifying potential phrase boundaries in the passage. However, since not all speakers produced a pause at every one of these boundaries, only boundaries occurring with a pause greater than 100ms were used for segmenting the passage into analysis intervals. (Pauses due to disfluencies were not used to define analysis intervals.)

To determine the timepoints of linguistic structures, each of the passage's syllables in the acoustic signal was annotated as having either primary stress ("stressed") or non-primary stress ("unstressed"), consulting the CMU Pronouncing Dictionary [18] in cases of ambiguity. The acoustic onset of the vowel in each syllable was selected as the syllable's timepoint of occurrence.

While phrase boundaries were used to demarcate analysis intervals, some participants produced (ungrammatical) filled or silent pauses interior to phrases as defined here. Pulse and peak intervals during pauses longer than 100ms (as marked by the Penn Force Aligner [20]) were excluded in data cleaning, and vowel onset and stressed vowel onset interval measures spanning such pauses were also excluded.

2.3. Statistical analysis

To compare the vowel onset timeseries and the first intrinsic mode function of the amplitude envelope as well as the stressed vowel onset timeseries and the second intrinsic mode function, two linear mixed effects models were fitted using the R package *lme4*, predicting frequency by timeseries (vowel onset and IMF1 for the first model; stressed vowel onset and

IMF2 for the second) with random intercepts grouped by speaker. Timeseries in each model was contrast coded, and frequency was centered and rescaled.

To calculate *within* speaker variation of the two modulation pulses and peaks in the differentiated amplitude envelope, the coefficient of variation (standard deviation normalized by mean) was calculated for the frequencies occurring in each speaker's analysis intervals; the mean of these coefficients of variation provided the average variability for each individual speaker. Differences between coefficients of variation of the three timeseries were tested using two modified signed-likelihood ratio tests (MSLRT) for equality of coefficients of variation [10] (R package *cvequality*, Version 0.1.3; [13]). Two tests for each speaker are relevant to our predictions; these compared for each speaker (i) the coefficients of variation for articulatory modulation pulses and the differentiated amplitude envelope and (ii) the coefficients of variation for acoustic modulation pulses and the differentiated amplitude envelope. Bonferroni corrections were applied to account for the 18 pairwise comparisons; p values were therefore considered significant at $p \leq 0.003$.

To calculate *between* speaker variation for the same timeseries, for all speakers the coefficient of variation was calculated across the frequency means of the two modulation functions and the amplitude envelope timeseries. Mixed effects models for the frequencies of the three timeseries were fitted separately on all speakers' pooled data using the R package *lme4*, predicted by random intercepts grouped by speaker. Residuals from each model were squared and divided by the mean frequency of the timeseries with all speakers pooled, and two linear models—the first comparing articulatory modulation pulses and the differentiated amplitude envelope, and the second comparing acoustic modulation pulses and the differentiated amplitude envelope—were fitted to predict residuals by timeseries.

3. RESULTS

The mean frequency of articulatory modulation pulses was 8.18Hz (SD = 2.31), and the mean frequency of acoustic modulation pulses was 9.12 (SD = 3.03). These signals were highly periodic; means for each speaker ranged from 7.76Hz to 8.49Hz (articulatory) and 8.25Hz to 9.66Hz (acoustic). The frequency range of modulation pulses was generally higher than the frequency range of the differentiated amplitude envelope; the envelope mean was 7.19Hz (SD = 2.47), with a range of speaker means from 6.89Hz to 7.65Hz. NB: The auditory cortex is most sensitive to modulations at frequencies between 2 and 8Hz [7]; the mean frequencies of each timeseries falls within or very close to this range.

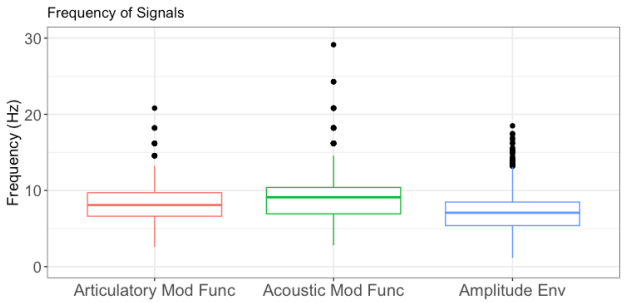

Figure 1: Frequencies of three periodic signals (each significantly different from the others $p < 0.001$). One point (48.5Hz) removed from middle boxplot for space.

3.1. Linguistic drivers of amplitude oscillation

Before turning to the variability of these signals, we assess the previously alluded to correlation [17] between the frequencies of the first two intrinsic mode functions of the amplitude envelope and linguistic structures. The first and second intrinsic mode function of the (undifferentiated) amplitude envelope were compared respectively to the frequencies of the vowel onsets of each syllable and the vowel onsets of each stressed syllable. These frequencies were expected to be non-distinct for the first intrinsic mode function (IMF1) and vowel onsets and for the second intrinsic mode function (IMF2) and stressed vowel onsets. Our results indicate no difference between the frequencies of IMF1 and vowel onsets. The mean frequency of IMF1 with all speakers pooled was 6.52Hz (SD = 3.15); the mean frequency of vowel onsets was 6.32Hz (SD = 4.49). A linear mixed effects model predicting frequency by timeseries with random intercepts grouped by speaker showed that there was no significant effect of timeseries (β = -0.027, SE(β) = 0.025, $t(1683)$ = -1.07, p = 0.28) Similarly, no difference was found between the frequencies of IMF2 and stressed vowel onsets. With speakers pooled, the mean frequency of IMF2 was 2.64 (SD = 0.89), and the mean frequency of stressed vowel onsets was 2.78 (SD = 1.50). A linear mixed effects model predicting frequency by timeseries with random intercepts grouped by speaker showed that there was no significant effect of timeseries (β = 0.061, SE(β) = 0.038, $t(695.4)$ = 1.59, p = 0.11). Therefore, Tilsen and Arvaniti's [17] qualitative observation that the first two intrinsic mode functions of the amplitude envelope correspond to linguistic structures is generally supported.

3.2. Variability of periodic signals

We turn now to variability of the frequencies of modulation pulses and how it compares to that of the

amplitude peaks, considering this both within and across speakers. The mean coefficient of variation within speakers was 28.1% for articulatory modulation pulses, 31.9% for acoustic modulation pulses, and 34.4% for differentiated amplitude envelope peaks. Figure 2 shows variation within each speaker.

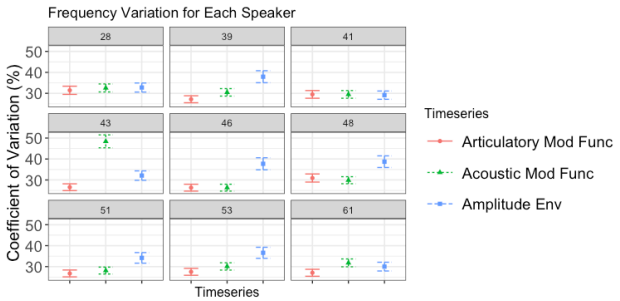

Figure 2: Coefficient of variation of the frequencies of three periodic signals. Error bars represent standard error of coefficient of variation.

Three speakers (speakers 39, 46, & 53) showed significantly lower variation in articulatory modulation pulse frequencies than differentiated amplitude envelope peak frequencies (p = 0.001, 0.0002, and 0.002, respectively); all others were not significantly different. One speaker (speaker 43) showed significantly lower variation in amplitude envelope peak frequencies than in acoustic modulation pulse frequencies, and another (speaker 46) showed significantly greater variation ($p < 0.001$ for these speakers); the rest were not significantly different. In sum, most speakers did not have any significant difference in variation between either type of modulation pulse and peaks in the differentiated amplitude envelope.

To determine between-speaker variability, the mean frequency of each timeseries was calculated for each speaker. The coefficient of variation was then calculated across these nine frequencies for each timeseries. These values were 3.41% for articulatory modulation pulses, 4.97% for acoustic modulation pulses, and 3.30% for differentiated amplitude envelope peaks. Fig. 3 shows variation with all speakers pooled; note that this figure shows higher coefficients of variation than the numbers stated above because it includes within-speaker variation, as well.

Models for between-speaker variability yielded significantly lower variability between speakers for articulatory modulation pulse frequencies than for differentiated amplitude envelope pulse frequencies (β = 0.20, SE(β) = 0.06, t(2476) = 3.56, $p < 0.001$), and no significant difference in variability between speakers for acoustic modulation pulse frequencies than for differentiated amplitude envelope pulse frequencies (β = -0.14, SE(β) = 0.15, t(2631) = -0.95, p = 0.34). The between- and within-speaker variability results indicate the stability of both articulatory and acoustic modulation pulses.

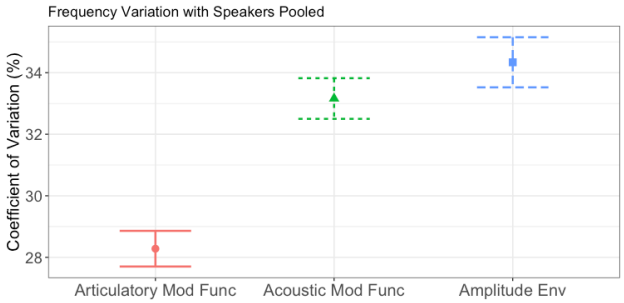

Figure 3: Coefficient of variation of the frequencies of three periodic signals with all speakers pooled. Error bars represent standard error of coefficient of variation.

4. DISCUSSION AND CONCLUSION

The goals of this study were to assess the articulatory and acoustic global modulation functions as an index of rhythmicity in the context of implications for spoken language production and perception processes by comparing them to the speech acoustic amplitude envelope periodicity. Further evaluated were claims that derived functions from the amplitude envelope capture specific levels of linguistic structure.

The modulation pulses were highly periodic in both articulation and acoustics, and their frequencies, like those of the amplitude envelope, fell within or slightly above the range at which the auditory cortex is reported to best entrain with speech [15]. Further, the first two intrinsic mode functions of the amplitude envelope matched the frequency of vowel onsets and stressed vowel onsets, respectively. The frequency of articulatory and acoustic modulation pulses were found to be comparably robust against within-speaker and between-speaker variation when compared with peaks in the differentiated amplitude envelope. The stability of these signals contributes to predictability and could enhance intelligibility and perceptual processing, in addition to foundational internal processes of speech planning and motor control that are thought to require 'binding' of the articulatory and acoustic dynamics [6].

Given that all three periodic signals are extracted from the same speech stream, whether in articulation or acoustics, each may be a window to a common underlying rhythmic phenomenon in speech. Further, given that the amplitude envelope was shown to reflect linguistic structural information of different timescales, the signals may be viable links between abstract linguistic elements and concrete oscillatory acoustic and articulatory patterns in the speech signal. Neural entrainment may then facilitate speech perception by efficiently estimating linguistic structures during listening processes.

5. REFERENCES

[1] Assaneo, M. F., & Poeppel, D. 2018. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, 4(2), 1–9.

[2] Darley, F. L., Aronson, A. E., Brown, J. R. 1975. *Motor Speech Disorders*. W. B. Saunders Co.

[3] Drullman, R., Festen, J. M., & Plomp, R. 1994. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5), 2670–2680.

[4] Frota, S., Vigário, M., Cruz, M., Hohl, F., & Braun, B. 2022. Amplitude envelope modulations across languages reflect prosody. *Speech Prosody 2022*, 688–692).

[5] Ghitza, O., & Greenberg, S. 2009. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66, 113–126.

[6] Goldstein, L. 2019. The role of temporal modulation in sensorimotor interaction. *Frontiers in Psychology*, 10, 1–12.

[7] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. 2003. Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31, 465–485.

[8] Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, 11(12), 1–14.

[9] Keitel, A., Gross, J., & Kayser, C. 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, 16(3), 1–19.

[10] Krishnamoorthy, K., Lee, M. 2014. Improved tests for the equality of normal coefficients of variation. *Computational Statistics*, 29(1-2), 215–232. https://link.springer.com/article/10.1007/s00180-013-0445-2

[11] Leong, V., Stone, M. A., Turner, R. E., & Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America*, 136(1), 366–381.

[12] Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., & Chauvel, P. 2004. Temporal envelope processing in the human left and right auditory cortices. *Cerebral Cortex*, 14(7), 731–740.

[13] Marwick, B., Krishnamoorthy, K. 2019. cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.1.3. Retrieved: https://github.com/benmarwick/cvequality, on 02/13/2022

[14] Oganian, Y., Chang, E. F. 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, 5(11), 1–13.

[15] Poeppel, D., Assaneo, M. F. 2020. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6), 322–334.

[16] Rosen, S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences*, 336(1278), 367–373.

[17] Tilsen, S., Arvaniti, A. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1), 628–639.

[18] Weide, R.L. 1998. The Carnegie Mellon pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[19] Westbury, J. R., Turner, G., Dembowski, J. 1994. *X-ray Microbeam Speech Production Database User's Handbook*. University of Wisconsin.

[20] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5).