Time, Cohesion, Style: Rhythm Formants in Oral Narrative

Dafydd Gibbon

Faculty of Linguistics and Literature, Bielefeld University, Bielefeld, Germany

gibbon@uni-bielefeld.de

1. Time, Rhythm and Register

Reality is a function of the methods used to observe events in space and time. Based on this pragmatic postulate, and concentrating on the temporal dimension, in the present study a novel signal processing framework is developed in order to analyse the speech rhythm of selected authentic data types, as opposed to the intuited and constructed formal data which are used in the description of 'linguistic rhythm' (advisedly thus named by Liberman and Prince 1977). In the present context, 'authentic' means that the data were not invented for the purpose of scientific study but are, in traditional terms, usage-based data recorded in independently motivated scenarios. The study relies extensively on graph illustrations of acoustic phonetic analyses. Approximately half of the contribution is devoted to theoretical issues, and the other half is concerned with exploratory case studies of rhythm in six different speech registers.

Time is a basic parameter in the analysis of speech utterances, along an extensive time scale. The scale spans the range from phones, with tens of milliseconds, through larger grammatical units of several seconds and discourse linguistic units of several minutes to much longer time spans: the acquisition of first and second language, dialect, and register, then across generations to historical language change and language evolution. One of the shorter span regions on the time scale is that of real-time speech rhythms: perceived regular strong-weak alternations of beats and inter-beat intervals, which appear in the infrasound frequency domain as low-frequency (LF) oscillations between 0.1 Hz and 10 Hz in the long-term spectrum of the speech signal; cf. also Albert and Grice (this volume), Brown (this volume). These oscillations are treated here within the modulation-theoretic framework of *Rhythm Formant Theory*, RFT, for speech stylometry, and RFA, its associated methodology, *Rhythm Formant Analysis* (Gibbon and Li 2019; Gibbon 2021, 2022, 2023).

The study is based on an inhomogeneity assumption: the speech of individuals and communities is not homogeneous, and different registers, styles and genres used by speakers and hearers differ in many properties, including prosody, understood informally as the rhythms and melodies of speech. A null hypothesis would be that all these varieties have the same rhythms. A more realistic set of hypotheses is that speech is more or differently rhythmical in some registers than

others, that differences can be detected, and if rhythms are not found in speech data, then maybe the wrong data have been selected. Conversely, if rhythms are detected in one speech register, they may not necessarily be found in all registers. Results of rhythm analyses depend, for example, on whether selected data are from the formal metalinguistic register of traditional linguistic and phonetic analysis, or whether they are foraged from speech 'in the wild'.

The aim is to investigate whether the interplay of rhythm and function in different kinds of oral narrative can be described and distinguished using the RFT/RFA framework. To this end, the study focuses on small exploratory case studies of a selection of speech registers and starts with a detailed discussion of the theoretical background. The study deals with oral narratives as registers (Section 2), rhythms and their functions (Section 3), approaches to speech rhythm analysis (Section 4), heuristic use of the annotation mining method (Section 5), the RFT/RFA framework (Section 6), RFA analyses of different kinds of oral narrative register (Section 7), comparison of rhythms in different registers by means of unsupervised clustering (Section 8), results and conclusions (Section 9).

2. Oral Narratives as Registers

Six specific registers are analysed in the present study and represent oral narratives of different kinds: toddler dialogue at an early stage in first language acquisition, the narrative genre of African village communities, fluency of reading aloud in English as a second language (L2), a comparison between newsreading and poetry reading in English and a comparison of recitations of different Chinese poetry genres. The prediction is that rhythms in these registers are physically distinguishable and that the differences can be detected with RFT/RFA.

The traditional term 'register' and related terms such as 'genre', 'style' and 'functional style' have been used in too many different ways in the literature to be reviewed here (but cf. Gibbon 1981, 1985). The term is used in the present contribution for family resemblances of text and speech usage in task-oriented contexts such as spontaneous conversation, verbal coordination, story telling, reading aloud, child and mother speech, or the metalinguistic formal register of traditional linguistic data.

The term 'register' is closely related to Wittgenstein's 'language game' (1953:5, §7), referring to language usage in specific contexts such as bricklayers using language as a work tool or children learning a language, which he calls (without negative associations) 'primitive languages':

> Ich will diese Spiele "Sprachspiele" nennen, und von einer primitiven Sprache manchmal als einem *Sprachspiel* reden. (I will call these games "language-games" and will sometimes speak of a primitive language as a language-game.)

This philosophical perspective implies that the language usage of an individual or community can be seen as an inhomogeneous set of overlapping registers, styles, genres, language-games, with virtuoso register-hopping, style-shifting and code-switching between language and speech varieties by community members.

Registers are usually described in terms of specific vocabularies and specific preferences for grammatical and word-formation rules (Biber and Conrad 2019). In spoken registers, but criteria such as clear or fast-speech enunciation, as defined in hyperarticulation and hypoarticulation theory (Lindblom 1990), and speech rhythm and melody are equally relevant (Crystal and Davy 1969).

A basic functional space for registers can be defined. First, *modality* ranges from oral-auditory, unidirectional-multidirectional, through face-to-face conversation to teleglossic (communication at a distance), and to subtypes of teleglossia in many kinds of electronic or other medium. Second, *topic* relates, for example, to domestic or professional, private or public, task-oriented (teaching, carpentry, sport, conversation...). Third, *style* covers language features of formal-informal, polite-impolite communication.

The registers discussed here are in the oral-auditory modality, with varying topics, from the impenetrable conversation of toddlers talking with single-syllable vocabulary on the one hand, to broadcast news or conventional poetry on the other. The data are recordings of authentic natural real-time data, partly from public sources.

3. Rhythms and their Metalocutionary Functions

Rhythms and speech registers have been rare companions in phonological and phonetic studies, while in discourse studies there is a history of interactionist discussion of rhythm in different speech varieties (Brazil 1985; Couper-Kuhlen 1993; Couper-Kuhlen and Selting 2018). From a functional point of view, natural speech rhythms are *metalocutions* with emotional and rhetorical functions, but also with a metalocutionary indexical cohesive function: like head-nodding, finger-pointing and beat gestures (McNeill 1992). Prosodic beats literally *point* (during *utterance time* and at *utterance place*) at constituents of the lexico-syntactic *locution* which they accompany.

From a functional perspective, an inheritance hierarchy of increasing specificity can be defined for directly observable rhythms: *physical rhythms* (ocean waves ,ripples, branches in the breeze), *physiological rhythms* (heartbeats, blinking, brain frequencies), *behavioural rhythms* (walking, chewing), *bonding rhythms* (dancing, handshaking, intimate interaction), *communication rhythms* (gesture, writing, speech), and *speech rhythms.*

Intuitively, rhythms are sequences (sometimes different sequences in parallel as in music) of regular waves and beats in the speech signal at similar intervals in time, where beats and inter-beat intervals are related to stronger and weaker values of some audible parameter ranging around 1 s in duration. When are beat sequences rhythms? An individual beat is not a rhythm, nor is a sequence of two beats, but a sequence of at least three beats permits the two inter-beat intervals to be compared in terms of duration equality, and thus for a rhythm to be identified (Nakamura and Sagisaka 2011). Syllable rhythms alternate between vocalic beats and consonantal inter-beat intervals. Word-level foot beats alternate between stronger syllables as beats and sequences of weaker syllables

as inter-beat intervals. Phrasal level ‘nuclear accent’ beats in major intonation sequences or ‘paratones’ alternate with minor intonation sequences with less prominent nuclear accents. Even longer duration, slower rhythms occur in read-aloud texts and in rhetorical and poetic discourse.

From a physical point of view, speech rhythms are oscillations of parameter values in the acoustic signal (Barbosa 2002), such as the amplitude of speech at a particular frequency, for example 5 Hz for syllables of average duration 200 ms, or about 1.25 Hz for accented words at intervals of about 800 ms, depending on the speech register (see Sections 7 and 8). Rhythms may be considerably longer than this, particularly in carefully crafted speeches or in poetry and song (cf. Chhatwal et al. this volume; Daikoku and Goswami 2022).

A ‘golden fleece’ which has haunted the search for speech rhythm in phonetics for decades is the ideal timing property of *isochrony*, equal timing in a succession of similar events. The isochrony property is not found as absolute duration equality, however, but as a scale of duration similarity of different phonetic event types such as mora, syllable, stress group (Dauer 1983, 1987). Several scales based on descriptive statistics have been proposed as a basis for a rhythm typology of languages (e.g. Grabe and Low 2002). These approaches have been critically discussed by Gibbon and Fernandes (2005), Gibbon (2006) and Arvaniti (2009), among others.

Whichever domain is inspected, rhythms are periodic time functions, i.e. they have a *frequency*. A rhythmic beat has a *magnitude*. Rhythms have a property of *resonance*, the constancy of frequency, and of *bandwidth*, the frequency range within which a varying rhythm remains a rhythm, and they have *persistence* in time: rhythms require at least three component beats and thus at least two inter-beat intervals, as already noted. Speech rhythms may co-occur (Barbosa 2002; Asu and Nolan 2006; Inden et al. 2012) in a hierarchy, as in music (cf. Section 7.2), and may also be shared with other interlocutors when their behaviour is mutually entrained and they adapt to each other (Cummins and Port 1998; Wagner this volume) as in the dialogue case studies in Sections 7.1 and 7.2.

4. Approaches to Speech Rhythm Analysis

The study of speech rhythm dates back to antiquity and rhythm has traditionally been seen as a poetic or rhetorical device. Since the mid-20th century phonological accounts of rhythms applied a metaphorical concept of structure as rhythm (‘linguistic rhythm’, Liberman and Prince 1977) using intuited and constructed data, along with other categories which are labelled with metaphors from poetics (e.g. ‘metrical phonology’, ‘trochaic’, ‘iambic’, ‘foot’). In poetics itself, distinctions are made between *metrical frameworks*, such as the iambic pentameter, on the one hand, and *grammatical stress patterns* on the other, and between these structural concepts and *performed rhythms* in poetry readings; cf. also rhetorical rhythms in public speeches (Gibbon and Li 2019).

The main approaches to speech rhythm in acoustic phonetics are listed here in approximate historical chronological order of appearance, as context for the present approach:

1. *Qualitative linguistic and applied linguistic models*, typically related to pronunciation teaching, from Sweet (1908) through Pike (1945), Jassem (1952) to Abercrombie (1967) and many textbooks; cf. Gibbon (1976) for an overview of these traditional approaches.
2. *Qualitative algebraic models* in universalist theories, from Chomsky, Halle and Lukoff (1956) through metrical theories originating with Liberman and Prince (1977) and the prosodic hierarchy of Selkirk (1984) to search-theoretic optimality theories originating with Prince and Smolensky (1993).
3. *Annotation mining* in descriptive phonetics with hybrid qualitative and quantitative signal-symbol mappings based on annotated speech, from Lehiste (1970) and Jassem et al. (1984) to Asu and Nolan (2006) and many others; cf. Section 5 below.
4. *Modulation theoretic analysis* in experimental and clinical acoustic phonetics, with concepts derived from radio engineering: demodulation and spectral analysis of meaningful information signals from the acoustic speech signal, from Ohala (1992), Todd and Brown (1994), Traunmüller (1994), Greenberg and Kingsbury (1997) to Barbosa (2002), Tilsen and Johnson (2008) and several later studies; cf. the overviews included in Gibbon (2021), Braun (this volume), Greenberg (this volume) and Section 6 below.

Annotation mining and RFT/RFA are described in the following two sections and annotation mining is used in two of the case studies in Section 7 and Section 8 as a heuristic source of predictions for further analysis with RFT/RFA.

5. Annotation Mining: Theory and Practice

The earliest and most popular family of methods for measuring rhythms in the physical speech signal is annotation mining, which inspects the duration relations between speech units such as vocalic or consonantal segments, syllables, feet, etc. The assumption that rhythm is solely a function of speech unit durations is an oversimplification, however, since the prominent beats or waves of a rhythm may also involve other parameters such as pitch patterns (Lehiste 1970).

Annotation-mining has a broader and a narrower sense. In the broader sense, annotation mining has six steps. The first step includes *recording* the speech signal and *visualising* properties such as the waveform (oscillogram), the F0 estimation (‘pitch’ track), the intensity curve and the spectrogram. The next step is *labelling* (*segmentation* and *classification*) of the speech signal by the assignment of categorial linguistic labels to intervals or points in the signal by close listening and by inspection of the display. The *annotations* are triples <*startpoint, endpoint, label*> for intervals, or <*midpoint, label*> pairs for points. In any given annotation sequence (*tier*), labels denote units of a particular linguistic category: phonetic (e.g. phones, syllables, larger units), structural (e.g.

words, phrases) or functional (e.g. focus, question, parenthesis). The parallel annotation tiers implicitly define time-synchronised mappings between tiers.

An example of phonetic annotation with parallel tiers using the Praat phonetic workbench software (Boersma 2001) is shown in Figure 1 (cf. Tracy and Gibbon 2023 for data description). The annotations are in parallel horizontal tiers, with annotation segments marked by vertical lines. From top to bottom: the tiers are of syllables, of syllables without hesitation particles, of sentence transcripts and of coordinating conjunctions and hesitation phenomena.

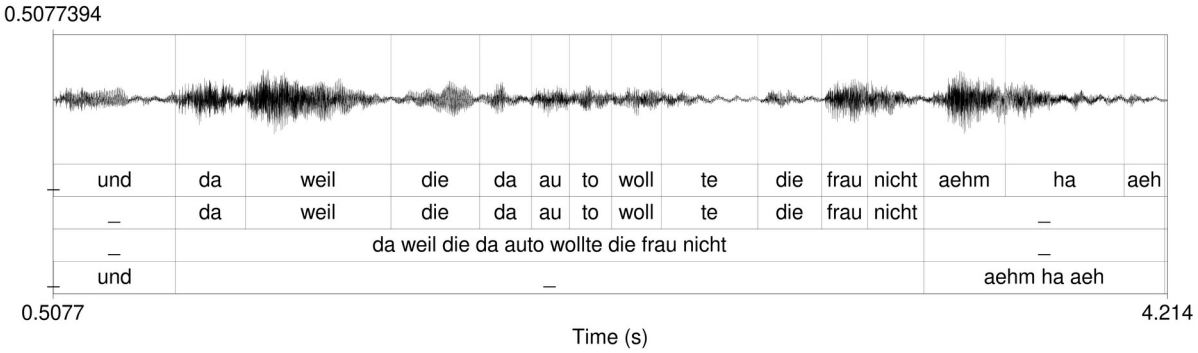

FIGURE 1. ANNOTATION WITH THE PRAAT PHONETIC WORKBENCH SOFTWARE (GERMAN SPONTANEOUS REPORT, ILLUSTRATING HESITATIONS).

The final step, *annotation mining* in the narrow sense (Gibbon and Fernandes 2005) is statistical analysis of sequences of annotated interval durations and their relations, often in an attempt to discover the isochronicity (degree of isochrony) of the sequence. Visual inspection of similarly spaced labels in Figure 1 gives an initial indication of possibly rhythmical sections in the recording.

Annotation mining traditionally involves descriptive statistics such as the *mean* together with dispersion measures (*standard deviation*, *coefficient of variation*), to provide an index of duration regularity (cf. the overview and comparison in Gibbon and Fernandes 2005). These methods are useful, but problematic as rhythm measures for several reasons in addition to concentration on the duration parameter alone: (1) descriptive statistics applies to static populations, not to dynamic time functions such as the speech signal; (2) taking squared or absolute values ignores the key alternation property of rhythms and thus the same index may refer to alternating or non-alternating sequences; (3) the 'rhythm metrics' are not metrics in the mathematical sense: they do not compare vectors of length *n* in an *n*-dimensional metric space (the triangle inequality criterion).

The *Pairwise Variability* (PVI) metrics are an exception and also overcome the disadvantage of unsuitability for time series. However, they retain the second disadvantage of ignoring alternations by using absolute differences. The PVI metrics also introduce a further asumption of binarity: the subtraction operation subtraction implies that rhythms are binary. This may be true on average (Nolan and Jeon 2014) but in reality three or more neighbouring units may be involved, as in the 'triple time' of *Everly Blenkinsop worried a lot about allergies*. The heuristic is saved by the *de facto* preponderance of binary rhythms.

The PVI metrics apply to sequences of interval durations and have non-normalised ('raw') and normalised versions, the *rPVI* and the *nPVI* (Grabe and Low 2002; Asu and Nolan 2006):

$$rPVI = \left(\sum \left| d_k - d_{k+1} \right| \right) / (m-1) \qquad nPVI = 100 \times \left(\sum \frac{\left| d_k - d_{k+1} \right|}{(d_k + d_{k+1})/2} \right) / (m-1)$$

Formally, the PVI measures are metrics: they are derivable directly from Manhattan Distance and Normalised Manhattan Distance (Canberra Distance), respectively, which are known metrics. The PVI metrics measure the average 'next-door-neighbour distance' between adjacent durations d_k, d_{k+1} of neighbouring intervals in the annotation. A duration sequence $d_1,\ldots, d_m$ is essentially treated as two overlapping vectors, $d_1, \ldots, d_{m-1}$ and $d_2, \ldots, d_m$, and the element-wise absolute difference (distance) between these two vectors is calculated. Manhattan Distance, Normalised Manhattan Distance and similar distance metrics yield comparable results to the PVI metrics.

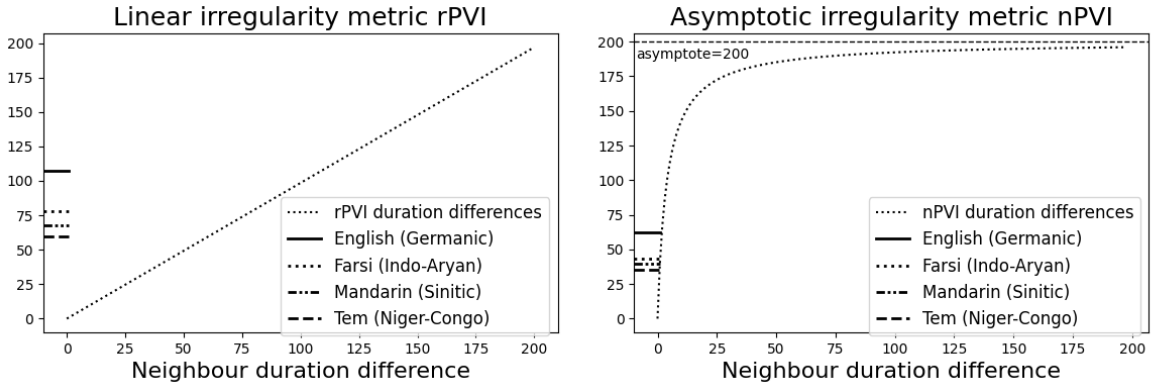

FIGURE 2. THE RPVI AND NPVI FOR DIFFERENT LANGUAGES, SHOWING LINEAR AND NON-LINEAR PROPERTIES FOR THE TWO METRICS.

The irregularity measures have been successfully used as heuristics to show systematic regularity differences between formal register data in different languages. The graphs in Figure 2 illustrate properties of the two metrics, measured with story readings in different languages,. The two metrics yield the same ordering and demonstrate that with both PVI variants the languages which are considered to tend toward so-called syllable timing, like Mandarin, Tem (ISO 639-3 *kdh*), Farsi, have considerably lower irregularity values than English, with so-called word, foot or stress timing. In Section 7.1 and Section 8.2 the PVI metrics are used as heuristic sources of predictions for RFT/RFA measurements.

6. RHYTHM FORMANTS: THEORY AND PRACTICE

6.1. SPEECH MODULATION THEORY

The approaches which enable analysis of the regularly alternating properties of real-time rhythms in authentic data, such as *frequency*, *magnitude*, *resonance*, *bandwidth* and *persistence*, are applications to speech of a signal processing theory which is as old as radio: *Speech Modulation Theory*, SMT (Ohala 1992;

Traunmüller 1994; Todd and Brown 1994; Cummins and Port 1998; Odell and Nieminen 1999; Barbosa 2002; Galves et al. 2002; Tilsen and Johnson 2008; Inden et al. 2012; Tilsen and Arvaniti 2013; Gibbon 2021, 2022, 2023; Frota et al. 2022; Braun this volume; Greenberg this volume). Neighbouring disciplines, particularly neurology and neurolinguistics, have applied similar methods (Meyer 2018; Boulenger this volume). The RFT/RFA framework is a further development of SMT, introducing the phonetic concept *rhythm formant* and using the semantic concept of *metalocution* for the indexical cohesion-marking functions of rhythm.

In SMT the speech signal is modelled as a *carrier wave* which is modulated by *information signals*: *frequency modulation* (FM) relating to intonation, pitch accent and tone and *amplitude modulation* (AM) relating to the sonority curve shaped by phonotactics, word formation, grammar and patterns of discourse (Ohala 1992; Galves et al. 2002). Simplifying the speech production process, the carrier wave is the complex sawtooth-like wave generated in the larynx, with a fundamental frequency and a series of harmonics (overtones) as multiples of the fundamental frequency. The carrier can be imagined as a monotone "Ah!" As in radio frequency technology, the carrier can be represented in stylised form (*A*: amplitude, *f*: frequency, *t*: time; phase is not included):

FM (variable phonation): $S_{fm} = A_{carrier} \cos\left(2\pi\left(f_{carrier} + f_{fm}\right)t\right)$

AM (variable oral-nasal filter) & FM: $S_{amfm} = A_{am} S_{fm}$

Both the FM and the AM information signals are normalised relative to the frequency and the amplitude of the carrier before modulation. The AM and FM frequencies which are relevant for speech rhythms, between about 10 Hz and 0.1 Hz, are between about 10 and 1000 times lower than the carrier frequency.

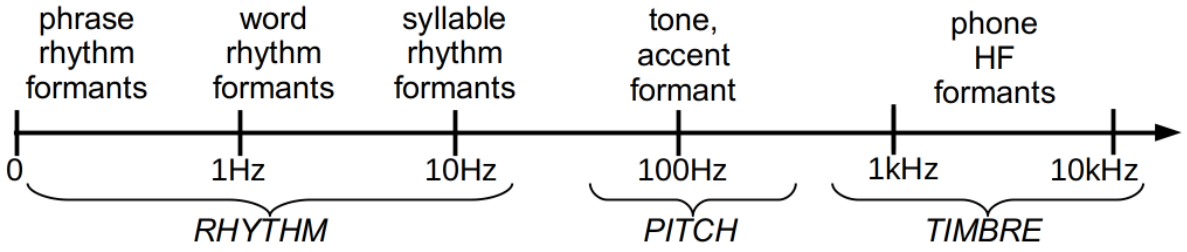

FIGURE 3. SPEECH MODULATION FREQUENCY SCALE.

Figure 3 shows the three main frequency zones which enter into FM and AM speech signals from a modulation-theoretic perspective. The carrier signal is in the central area of the logarithmic scale between about 70 Hz and 500 Hz (depending on sex, age, conventions and individual factors). The harmonics are amplitude modulated as high-frequency (HF) carriers for phone formants.

The term *rhythm formant* is used for low frequency (LF) magnitude peaks in the spectrum, by analogy with high frequency (HF) phone formants, which are also defined acoustically as magnitude peaks in the spectrum. Definitions of LF formants and HF formants differ when based on production and perception rather than transmission of speech.

6.2. Demodulation in the RFT/RTA Framework

In speech analysis (and in speech perception) the FM and AM components of the composite carrier wave are demodulated in order to extract the signals representing structural and semantic information. Low-pass filtered FM demodulation corresponds to F0 estimation ('pitch' tracking) in conventional terminology and the resulting F0 track is interpreted in terms of tones, pitch accents and intonations. Demodulated low-pass filtered AM approximates to the sonority curve of phonology and provides the acoustic grounding for speech rhythms. In the present study the FM signal is only discussed in passing (Section 8.3). *Rhythm Formant Theory*, RFT, adds adds the following postulate to SMT:

> Magnitude peaks in the spectra of the low-frequency FM and AM information signals function as *rhythm formants* determined by utterance categories from phone and syllable to longer discourse units, and indicate properties of rhythms in the frequency domain: *frequency* (comparable with speech rate in the time domain), *magnitude* (how prominent the beats are), *resonance* (constancy of frequency), *bandwidth* (the frequency range covered by the rhythm) and *persistence* (the duration of the rhythmic sequence).

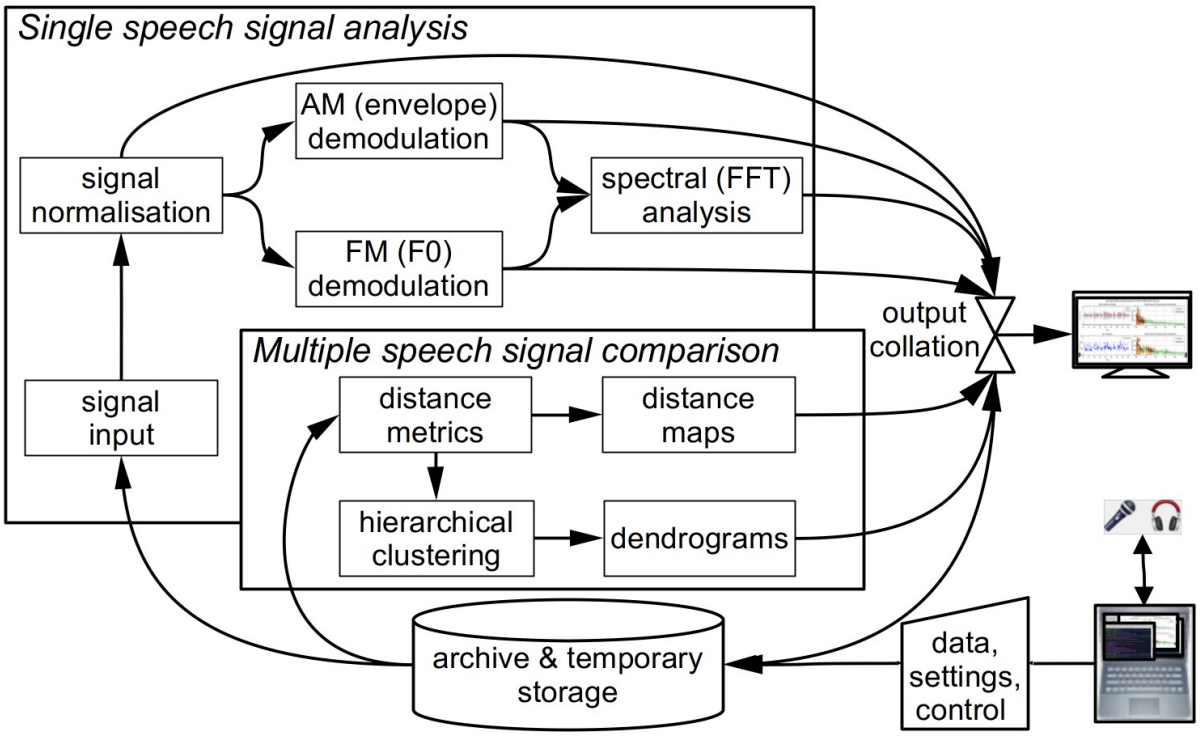

Figure 4. Rhythm Formant Analysis data-flow.

Rhythm Formant Analysis, RFA, is the methodology associated with RFT. Figure 4 illustrates the data-flow: inputs and outputs are stored; inputs are demodulated in order to estimate the FM (F0) and AM information signals; FFT is applied and sets of spectral properties are analysed and compared.

For FM demodulation (cf. Section 8.3) a modified time domain algorithm is used, AMDF (*Average Magnitude Difference Function*) with preprocessing and post-filtering. AM demodulation is performed by obtaining the absolute values of the

low-pass-filtered waveforms and smoothing of the resulting amplitude envelope (see figures in Section 7 and Section 8). The absolute Hilbert transform (He and Dellwo 2016) and other techniques are also used for AM demodulation.

In RFA the demodulated FM and AM signals are analysed holistically by applying a Fast Fourier Transform (FFT) to the entire recording, or to a selected long stretch of the recording, with transform windows of several seconds. The *LF spectrum* shows *frequency* × *magnitude*, with no temporal information.

To regain temporal information, a low-frequency *spectrogram* is calculated with overlapping spectral slices (*time* × *frequency* × *magnitude*). The spectrogram allows detection of rhythm *resonance*, *bandwidth* and *persistence* as well as frequency. Several property vectors and variance values, for example the ten most prominent well-defined peaks or the entire low-frequency spectrum (Section 7) are then compared using unsupervised cluster analysis (Section 8).

7. Exploratory Case Studies of Spoken Registers

7.1. An Interactive Protodialogue Register: Talking Twin Babies

A well-known YouTube meme is 'Talking Twin Babies', showing video recordings of the 'communicative babbling' of 17-month-old American twins in a kitchen, holding a prosodically very fluent-sounding conversation with each other, using only iterations of the single syllable 'da'.[1] The children are apparently imitating conversations between older children or adults. The overall duration of the selected dialogue is 112.41 s. In the present context the dialogue is a minimal interactive speech register, and a true 'language game'; cf. also Daikoku and Goswami (2022); Kalashnikova et al. (this volume) on infant speech registers.

The dialogue grammar is iterative and has two iteration levels: the utterance cycle enclosing the cycle with the syllable "da". Iteration cycles are easily modelled in a finite state machine, which requires only linear processing time and finite working memory, a realistic assumption. This contrasts with recursive grammars, which, an unrealistic assumption, in principle require non-linear processing time and non-finite working memory, though they are often used as a convenience. The finite state grammar also relates easily to rhythms as beat iterations; cf. Pierrehumbert (1980), whose finite state intonation grammar can also be interpreted as a rhythm machine. The grammar is rendered here as a regular expression: a disjunction of at least one utterance by a twin of at least one "da" syllable:

$$(da_{twina}{}^{+} \mid da_{twinb}{}^{+})^{+}$$

Annotation mining is used to predict values for possible confirmation in the follow-up RFA analysis, cf. Figure 5 (top). Table 1 lists a selection of descriptive statistics[2] for syllable and interpausal unit (IPU) annotation. There are 147

[1] https://www.youtube.com/watch?v=Iih0Z2IblUQ

[2] Annotation mining of Praat TextGrid files uses the TGA (Time Group Analysis) online application: http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

syllables with mean syllable duration 346.38 ms, a relatively slow rate of about 2.89 syll/s. For comparison, syllable rates in adult reading aloud and conversation in the Aix-MARSEC database (Auran et al. 2004) were measured as reference values, finding tempo variation between about 4 syll/s for religious readings and poetry readings to almost 6 syll/s for radio news. The nPVI metric for syllables yields an average 'next-door neighbour distance' of 23, a highly regular pattern, in contrast to values near 40 for Standard Mandarin (often said to have syllable timing) and near 50 for English (often said to have foot, word or stress group timing). The intra-IPU duration slope is 0.401, indicating tempo deceleration.

TABLE 1. TWIN TODDLER SYLLABLE DURATIONS (IN MILLISECONDS).

Syllable durations (N: 197)				*Interpausal Unit (IPU) durations (N: 37)*			
min:	147	max:	1033	min:	751	max:	7265
total:	68237	range:	886	total:	113419	range:	6514
mean:	346.38	mean rate/s:	2.89	mean:	3065.38	mean rate/s	0.33
median:	325.0	median rates/s:	3.08	median:	2782.0	median rate/s	0.36
intercept:	307.088	slope:	0.401	intercept:	3281.105	slope:	-11.985
std:	111.937	coeff var (%):	32.316	std:	1490.889	coeff var (%):	48.636
nPVI:	23	rPVI:	85	nPVI:	53	rPVI	1582

Each IPU is measured with the following pause. The mean IPU duration of 3.065 s and the 0.33 IPU/sec rate, with an nPVI distance of 53, is quite irregular. The slope of -11.985 for IPU duration sequences is negative, indicating shorter IPUs as the utterance proceeds. The toddler syllable sequences are synchronised with dance-like arm and leg movements.

The mean syllable rate of 2.89 syll/s suggests that a spectral magnitude peak at around 2.89 Hz will be found as a syllable rhythm formant. Similarly, the IPU rate of 0.33 IPU/sec suggests that a spectral magnitude peak of around 0.33 Hz will be found as an IPU rhythm formant. These values are taken as predictions for the RFA analysis.

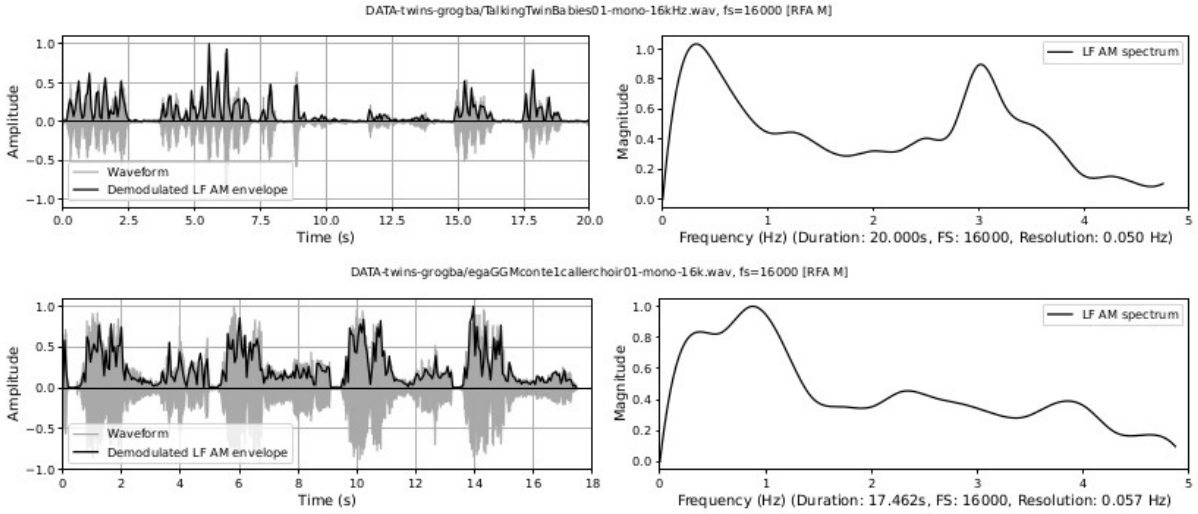

FIGURE 5. DIALOGUE REGISTERS: TOP, 20 S, TODDLERS (SECTION 7.1); BOTTOM, 18S CALLER-CHOIR EXCHANGE (SECTION 7.2).

Figure 5 (top) visualises the RFA measurements of the twin toddler dialogue. The left panel shows the waveform (grey values) and the demodulated amplitude envelope (dark positive values). The right subpanel shows a representation of the LF spectrum, smoothed with a moving median filter and spline interpolation. The spectrum of the toddler dialogue shows two very well-defined peaks, one at 0.3 Hz (IPU rhythm formant) and one at 3 Hz (syllable rhythm formant), as predicted by the annotation mining, showing superposed IPU and syllable rhythms as metalocutionary pointers to the dual patterning of the dialogue grammar. The extreme regularity is generated jointly by both toddlers and can thus be seen as evidence for rhythm entrainment (Cummins and Port 1998; Inden et al. 2012; Rathcke et al. 2021; Wagner this volume) in natural speech.

7.2. A Poetic Interactive Dialogue Register: Ega Orature

7.2.1. Structure and Spectrum

Part of an interactive orature session is analysed: a story in Ega, an endangered Niger-Congo language with agglutinative tonal morphology, spoken in South Central Ivory Coast (ISO 639-3 *ega*; Gibbon 2023). The selected session segment consists of a chanted caller-choir exchange (adjacency pairs) between the narrator and the audience. The dialogue grammar for the orature session as a whole can be modelled as a finite state machine (as already noted, an appropriate formalism for rhythm iteration), here in transition network format (Figure 6) with four iterating cycles: *narrative-pause*; *narrative-pause-backchannel-pause*; *call-response chant*; overall *narrative-chant*.

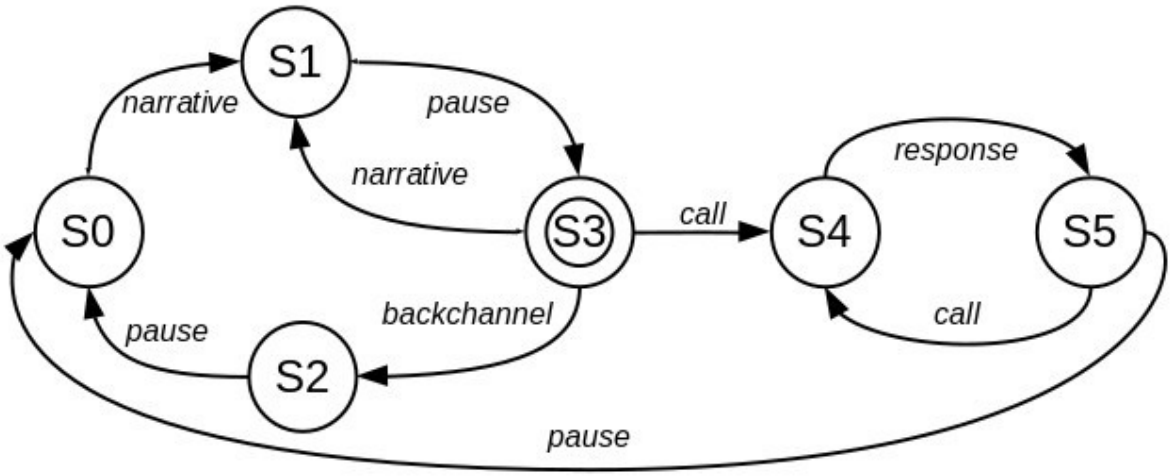

Figure 6. State machine (finite transition network) representing Ega orature dialogue grammar.

7.2.2. Resonance, Persistence: the Low-Frequency Spectrogram

Figure 5 (bottom) shows the waveform, amplitude envelope and LF spectrum of one of the chant exchanges. The peaks below 1 Hz, at 0.25 Hz and an octave higher at 0.5 Hz reflect the overall two cycle levels of caller-choir and caller and choir, respectively. The example illustrates metalocutionary cohesive functions of multiple rhythms as markers of locutionary patterns.

The atemporality deficit of the spectrum is remedied by using a low-frequency spectrogram (Figure 7) to show the resonance and persistence of rhythms (cf. Todd and Brown 1994; Greenberg and Kingsbury 1997). The spectrogram frequency and temporal resolutions are low because of the Küpfmüller time-frequency uncertainty principle $\Delta t \Delta f \geq c$ (meaning that time windows and frequency ranges cannot both be arbitrarily reduced) and therefore a long FFT window is needed in order to capture the low frequencies. The low temporal resolution is partly compensated for by overlapping the spectral slices.

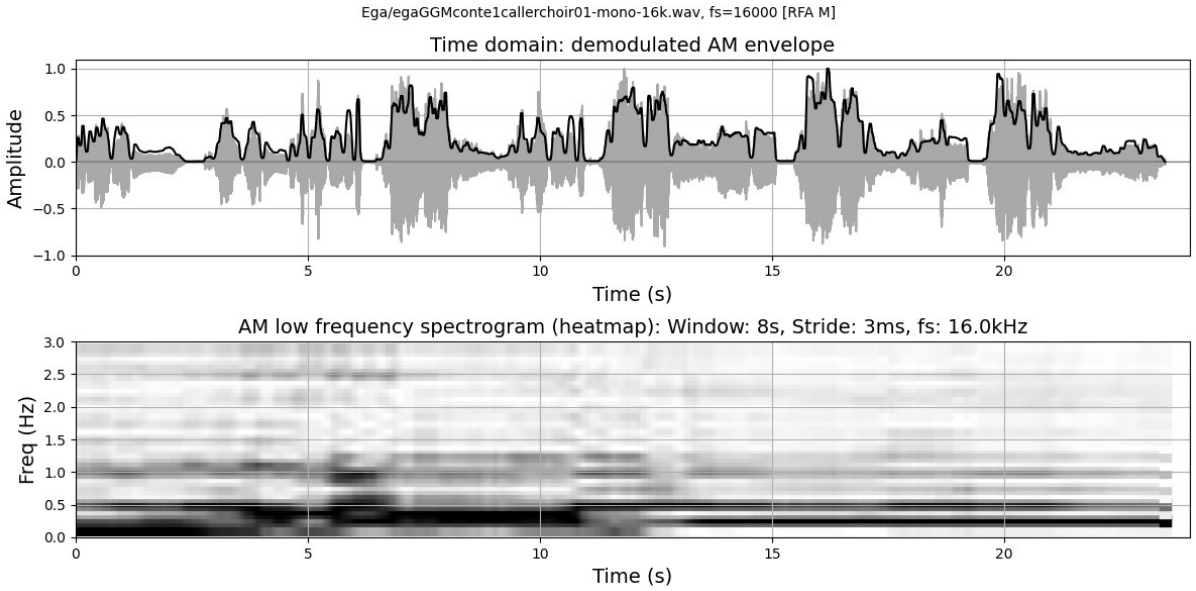

FIGURE 7. LOW-FREQUENCY SPECTROGRAM: FIRST CHANT SECTION OF THE ORATURE SESSION.

The chant rhythms appear in the spectrogram as *rhythm formant bars* an octave apart, at 0.25 Hz and 0.5 Hz, as in the spectrum, starting at about 12 s. In addition to frequency, rhythm bars have temporal properties of resonance and persistence and point to dialogue sections, showing the metalocutionary cohesive function of discourse rhythms. Since the rhythm formants are jointly supplied by audience members, together with their shared metalocutionary function they can be seen as further evidence for natural speech-song entrainment.

7.3. A PRACTICAL REGISTER: L2-READING ALOUD

The text prompt is the IPA benchmark text, an English translation of Aesop's fable *The North Wind and the Sun*, which has been used in previous rhythm studies to compare language varieties (e.g. Tilsen and Arvaniti 2013; Gibbon 2021). For present purposes, reading aloud in a second language (L2) is regarded as a different register from reading aloud in the first language (L1).

There have been many descriptions of L2 fluency in the time domain in relation to the rhythms of partially automatised production skills: syllable rate and reduction, mean run (IPU) duration, filled and unfilled pause ratio, and expert ratings (cf. overviews in Thomson 2015, Trouvain and Braun 2020). The use of frequency domain spectral parameters for L2 fluency assessment was introduced by Lin and Gibbon (2019; 2023), showing that RFA can be used to distinguish

between three speaker types: British native speakers as readers, a fluent Chinese speaker of L2 English, and a class of intermediate level L2 students of English.

Visual inspection of the LF spectra in Figure 8 reveals conspicuous differences between speaker types. The top panel (male British native speaker, reading duration 40 s) shows a clear sentence rate at 0.3 Hz and a foot (pitch accent) rate between 2 Hz and 3 Hz. The middle panel (accented though fluent female Chinese university teacher of L2 English, reading duration 60 s) shows IPU peaks at 0.2 Hz and 0.8 Hz, and foot peaks around 1.5 Hz, indicating less regular IPUs and a slower foot rate than the native speaker. The bottom panel (intermediate level male Chinese student of English, Tongji corpus, Yu 2013, reading duration 55 s) shows a very scattered distribution of peaks, a possible indicator of uncertainty and lower fluency. The comparison indicates that metalocutionary cohesion marking can be a component of fluency evaluation, and that RFT/RFA analysis can help to identify these markers.

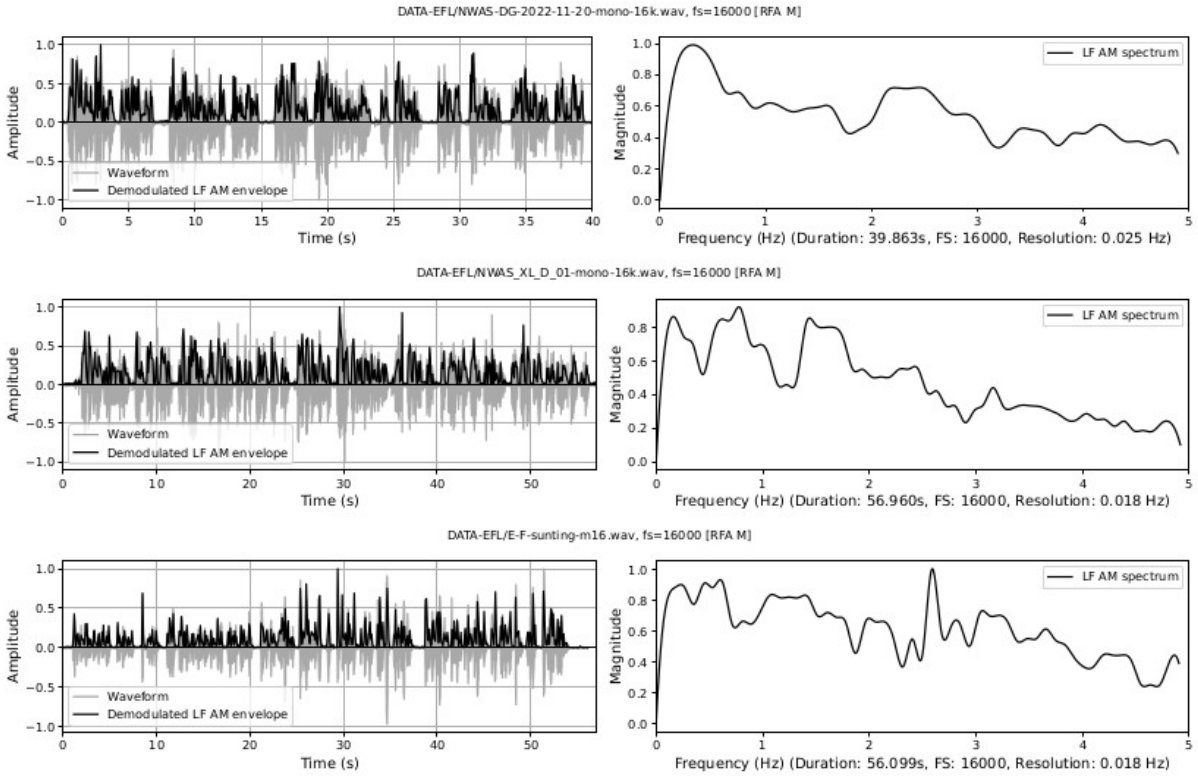

FIGURE 8. ENGLISH: TOP, L1 MALE SOUTH-EASTERN BRITISH ENGLISH; MIDDLE: CHINESE L2 FEMALE (FLUENT); BOTTOM, CHINESE L2 MALE (LESS FLUENT).

8. STYLOMETRIC RHYTHM COMPARISON WITH RFT/RFA

8.1. COMPARISON METHODS

Having analysed different kinds of data, RFA results from different data samples can be compared. RFA output is a set of vectors of spectral parameters which are relevant for rhythm analysis. The values include frequency and amplitude envelopes; spectra and trajectories of highest-magnitude frequencies in the spectral slices of spectrograms; FM and AM magnitude peaks; vector variance.

Unsupervised clustering algorithms are used to compare rhythms of different speakers as an alternative to classical difference testing. In Section 8.2 *k*-means clustering is shown as a scatter plot with marked clusters. Vector pairs can also be compared using distance metrics and display of the resulting distance network (8.3). A third method is to use the values from a distance network together with a clustering criterion to calculate a hierarchical dendrogram (Section 8.4).

8.2. Comparison of newsreading and poetry reading

The data compared in this section are from the Aix-MARSEC database (Auran et al. 2004) and are identified by ID in the figures: newsreadings and poetry readings (Gibbon 2022), with mainly male readers. Excerpts from the selected recordings are as follows (shorter pauses are marked '|' in the newsreading transcript, longer pauses as '||'; rhyming lines are marked '||' and half-lines '|' in the transcript of the poetry reading):

> BBC news extract: *"A thousand people were led to safety | after being trapped by a fire | in the London underground last night. || Many had to walk along the track to the nearest station.||"*

> Poem Eunice, written and read by John Betjeman, first stanza: *"With her latest roses | happily encumbered || Tunbridge Wells Central | takes her from the night, || Sweet second bloomings | frost has faintly umbered || And some double dahlias | waxy red and white.||"*

Table 2. Comparison of news and poetry readings using annotation-based dispersion values for syllables, words and inter-pausal units.

Category	*File*	*n*	*Mean (ms)*	*Rate/s*	*SD*	*CoVx100*	*nPVI*
Syllables	B0101B (news)	242	180.78	5.53	89.763	49.65	51
	H0101B (poetry)	189	257.24	3.89	127.018	49.38	53
Words	B0101B (news)	161	271.73	3.68	141.663	52.13	68
	H0101B (poetry)	129	376.88	2.65	208.762	55.39	69
IPUs	B0101B (news)	16	2734.31	0.37	1655.932	60.56	77
	H0101B (poetry)	22	2209.91	0.45	793.455	35.90	31

The annotation mining results in Table 2 show that syllable and word rates are faster for the newsreading ('B') than for the poetry reading ('H'): 5.53 syll/s vs. 3.89 syll/s and 3.68 word/s vs. 2.65 word/s. The IPU rate for the newsreadings is slower mainly because the utterances are longer than the lines of the poetry readings. Despite these differences, the nPVI values for syllables (51:53) and words (68:69) are almost same in the two cases. The IPUs (77:31) show a striking but expected difference: the newsreading IPUs are very irregular and the poetry reading IPUs are very regular. Standard deviation and coefficient of variation confirm the difference, suggesting that automatic comparison may be possible.

An RFA analysis of 20 s of each recording was made, from 15 s to 35 s into the recording. The results are shown in Figure 9, with the demodulated AM envelopes of selected intervals of the newsreading (upper row) and the poetry-reading (lower row), in the left-hand panels, in the time domain, and the holistic LF spectra of these intervals, in the right-hand panels, in the frequency domain. The demodulated envelopes have very different amplitude distributions.

The LF spectral differences correspond to the annotation-mining results. The left-hand rhythm formant of the newsreading (about 0.3 Hz) corresponds approximately to the mean IPU rate (0.37 IPU/s). The peaks between 3 Hz and 4 Hz approximate to the mean annotated word rate of 3.68 syll/s and the peak at about 5.2 Hz approximates to the mean annotated syllable rate of 5.53 Hz. Different syllable types and small variations in timing account for the bandwidth of frequency variation within the different spectral regions.

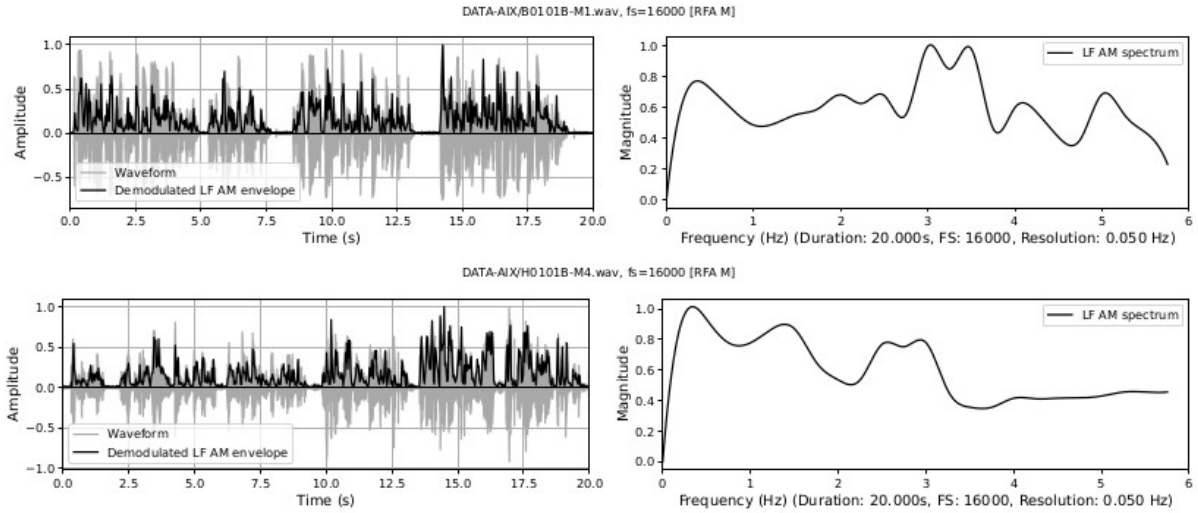

FIGURE 9. RFA DEMODULATION AND LF SPECTRUM OUTPUTS FOR A NEWSREADING (TOP) AND A POETRY READING (BOTTOM).

So far the broad peak at 1.5 Hz and the peak at 5 Hz in the poetry readings are unaccounted for. The situation is quite complicated because the fourfold structure of the poem (syllable, word, half-line, line) does not easily match the three phonetic categories of syllable, word, and IPU. The broad peak at 1.5 Hz may mean somewhat variable half-line durations and the broad peak at 5 Hz may denote weak, unstressed syllables. The low-frequency spectrum of the poetry-reading (lower right) differs: the IPU rate (lines of the poem) was measured at 0.45 Hz and a peak is observed at approximately 0.4 Hz. There are also peaks at 1.5 Hz and 3 Hz, and at just under 4 Hz, and at 5 Hz. The small 4 Hz peaks relate to the 3.89 syll/s measured syllable rate and the 2.65 word/s rate relates to the frequency peak at about 2.7 Hz.

Ten examples each from these two reading registers are compared using *k*-means clustering, *k*=2. The graphs in Figure 10 show the AM spectrogram trajectory on the *x*-axis. The left-hand graph has the FM spectrogram trajectory of highest magnitude frequencies on the *y*-axis and the right-hand graph *y*-axis shows F0 variance (newsreading: ‘B’, poetry reading: ‘H’; male: ‘M’, female: ‘F’; speaker index: numbers; centroids as stars). The *x*-axes show a near-partition between the registers, with just one ‘H’ outlier. The FM spectrogram trajectory (*y*-axis) does not show clear category separation, while F0 variance (relating to linear F0 range) shows higher values for female readers.

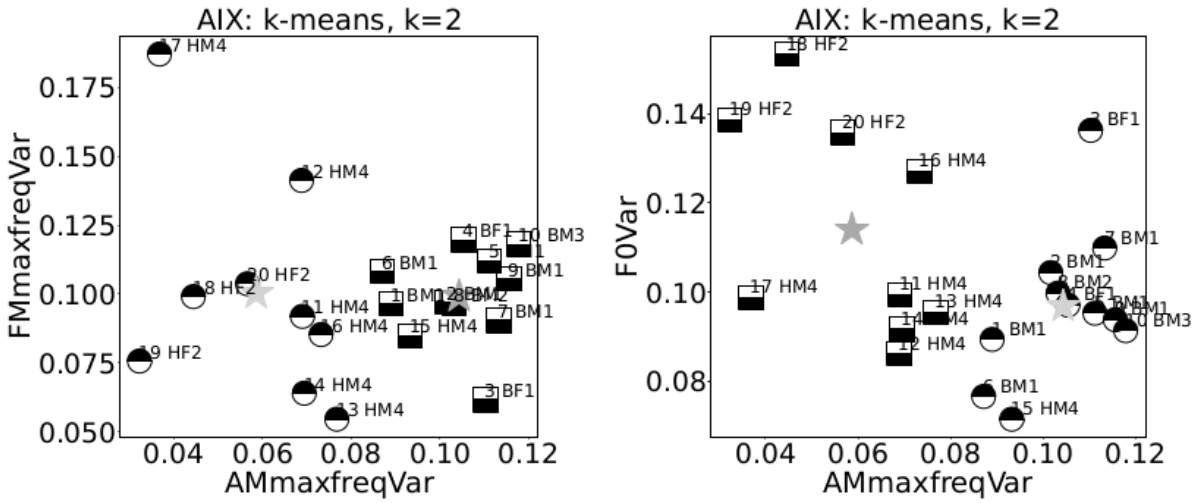

FIGURE 10. COMPARISON OF VARIANCES OF NEWSREADINGS (B) AND POETRY READINGS (H).

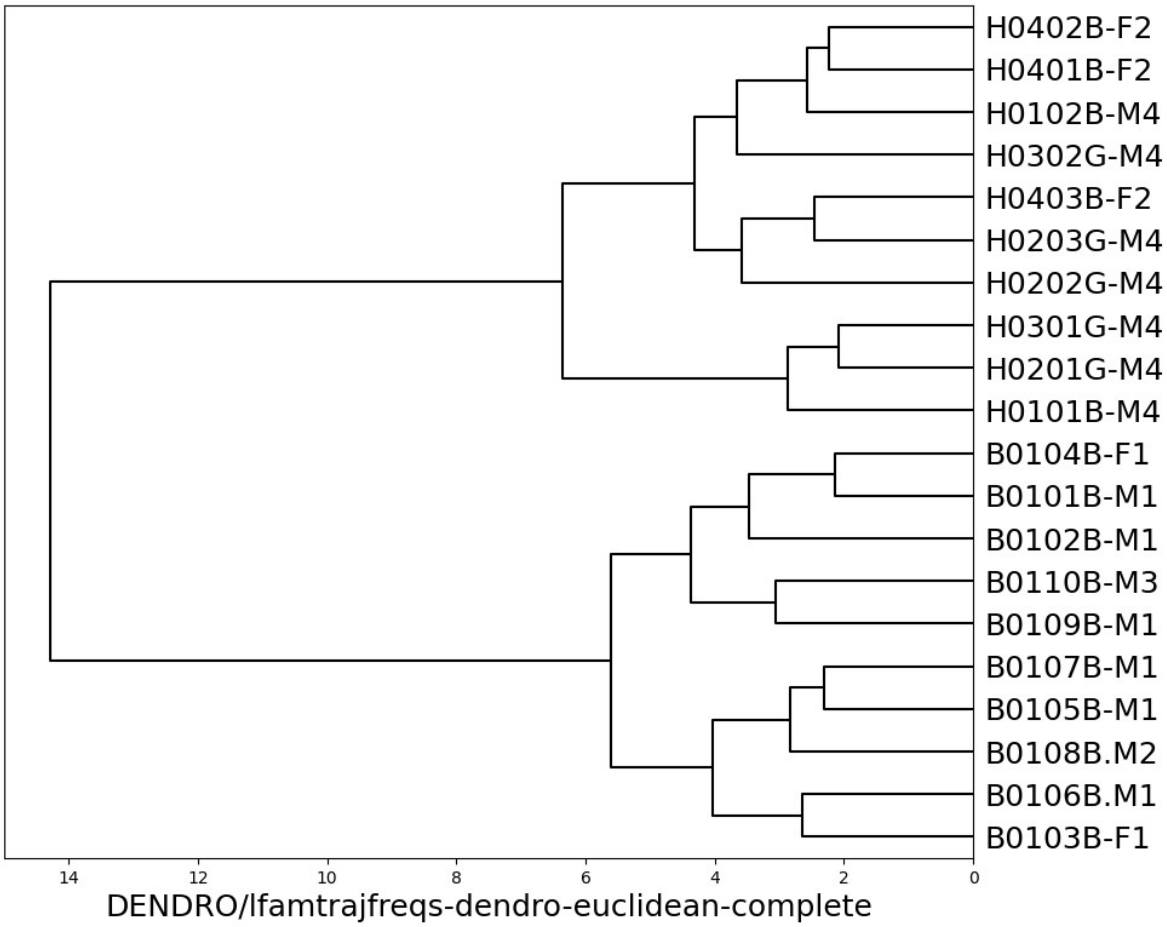

FIGURE 11. HIERARCHICAL CLUSTERING OF NEWSREADING AND POETRY READING (EUCLIDEAN DISTANCE AND FARTHEST NEIGHBOUR CLUSTERING).

When distances between the AM spectrogram trajectories (paths of frequencies with highest magnitude in each spectral slice) are compared with an additional criterion of agglomerative hierarchical clustering, a clear partition between the two registers emerges (cf. Figure 11); the length of the branches indicates the

size of the inter-cluster difference. This result is obtained reliably with different common distance metrics (Chebyshev, Euclidean, Manhattan) and all available clustering criteria (including farthest neighbour, nearest neighbour, mean, median and variance minimisation). It is not obtained with Cosine and Pearson Distance, showing the relevance of absolute difference, not trajectory shape.

8.3. Comparison of Poetry Genres

In a cooperative venture[3] with a specialist in Chinese-English literary translation, two types of poetry were examined, not in contemporary languages, but in a hybrid scenario: Tang dynasty Chinese poetry from the 7th and 8th centuries CE in modern recitations from the early 21st century CE (cf. also Gibbon 2022).

The types of poem to be compared are the 5-character line and 7-character line genres, with 11 poems of each type, including rhythm influences from different conventional tonal patterns. Intuitively it is expected that the rhythms of the two genres differ at the level of line-length rhythms. The demodulated signals were duration-normalised and the spectrum shapes, rather than distance, were compared using a Pearson Distance measure, e.g. 1-abs(*r*). The resulting network is shown in Figure 11. An exact partition was found at a distance limit of 0.47 (range is 0…1): the 5-syllable ('B') poems are on the left in the figure and the 7-syllable ('F') poems are on the right. There is sufficient distance between B and F genres and sufficient intra-genre proximity to yield separate B and F graphs.

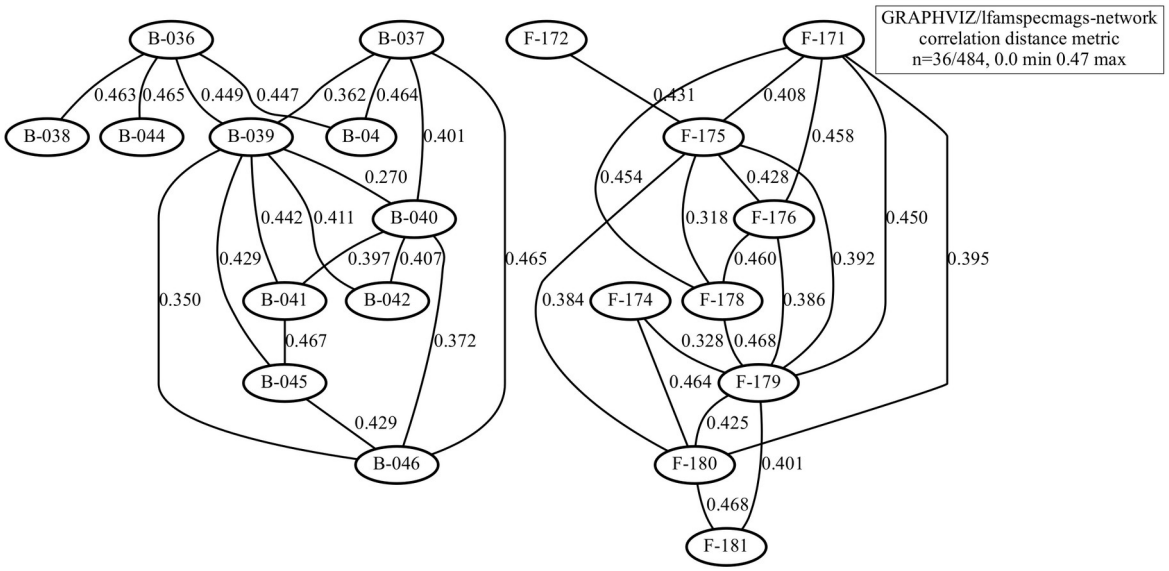

Figure 12. Distance network with modern recitations of two genres of Tang dynasty poetry.

9. Results and Conclusions

In preparation for the rhythm analyses and comparisons, close attention was paid to methodological assumptions about register, rhythm and phonetic analysis as determinants of models of the physical reality of rhythm, in the sense noted at

[3] The data originated in a joint project with Dr. Xuewei Lin, Jinan University, Guangzhou.

the beginning of the Introduction. Exploratory case studies of six different register scenarios and their dynamic rhythm formant properties were carried out, and unsupervised cluster comparisons were conducted with Rhythm Formant Theory and Rhythm Formant Analysis, a recently developed modulation-theoretic signal processing framework, using annotation mining as a heuristic source of hypotheses for these analyses. The analyses demonstrated distinctive rhythms in authentic natural data from spoken registers and their metalocutionary functions as indexical markers of locutionary cohesion.

The case study data sets are very small, so the results, though clear, may not be fully generalisable. Nevertheless, exploratory case studies of this kind are useful sources of hypotheses for future research. Further exact numerical modelling of rhythm formant properties of frequency, magnitude, resonance, bandwidth and persistence remains to be done. The important point to be retained, however, is the proposition that natural real-time rhythms with long-term low-frequency acoustic properties can distinguish between speech registers, styles, or genres.

Open issues for future work concern more detailed rhythm properties which can be found in the natural performance of speech, as in musical performances, like syncopation, attack and decay, or sustain and release. Some of these properties depend on the phonotactics of languages (for example a preponderance of voiceless fricatives as opposed to sonorants may relate to an ‘attack’ category), as in the well-known pair *takete-maluma* (Köhler 1929) with voiceless obstruents in the first word and nasal sonorants in the second. Such properties may be relevant for explaining subjective attractiveness or unpleasantness judgments of rhythms and musicality in typologically different languages.

The results indicate that there are ‘real-time rhythms’ beyond the abstract ‘linguistic rhythm’ domain which can be captured by means of physically grounded empirical analysis, and which have identifiable metalocutionary functionality, marking meaningful cohesive locutions. Applications are anticipated not only in acoustic phonetic speech stylometry, as in the present study, but in speaker, language and register identification and search, including forensic search, pertaining to other categories and dimensions of speech in and beyond the acoustic domain

Summary

Spectral properties of amplitude and frequency modulation of speech are cues to physical correlates of speech rhythms. Low-frequency spectral differences approximate to annotation-mining results and correspondences between rhythm formants in the frequency domain and word, phrase and discourse units can be established. Rhythm comparisons are visualised using the distance networks and hierarchical clustering which characterises text stylometry and dialectometry.

Implications

Spectral analysis shows that the term ‘linguistic rhythm’ for numerical encoding of grammatical structure is far from providing a general rhythm theory with empirical grounding of speech rhythms. Measurable spectral properties of spoken

language relate to linguistic units as well as to the rhetorical and poetic patterns of speech, and also extend stylometric and dialectometric studies into the real-time physical domain of speech. The RFT/RFA framework provides a path towards more detailed investigation of the prosodic correlates of linguistic categories and an acoustic grounding for studies of neural oscillations.

Gains

The novel concept of rhythm formant advances modulation-theoretic comparison of speech rhythms by setting specific spectral properties in relation to linguistic units. Further, the formal semantics of rhythm, defined as metalocutionary indexical pointers to cohesive patterns in the lexico-syntactic utterance patterns, here defined with finite state machines, represents a step toward a formal semantics of prosody. Practical uses in speech classification, self-taught language learning applications and language fluency evaluation are anticipated.

Index terms

annotation mining, cohesion, frequency domain, metalocution, natural rhythm, oral narrative, real-time rhythm, register, rhythm formant, Rhythm Formant Theory, Rhythm Formant Analysis, Speech Modulation Theory, speech rhythm, speech stylometry

10. REFERENCES

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

Albert, A. and Grice, M. (this volume). Rhythm is a timescale.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. Phonetica, 66(1-2):46–63.

Asu, E. L. and Nolan, F. (2006). Estonian and English rhythm: a two-dimensional quantification based on syllables and feet. In Proc. Third International Conference on Speech Prosody.

Auran, C., Caroline Bouzon and Daniel Hirst (2004). The Aix-MARSEC project: an evolutive database of spoken British English. In Proc. Second International conference on Speech Prosody, 561–564.

Barbosa, P. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. In Proc. First International Conference on Speech Prosody, 163–166.

Biber, D. and Conrad, S. (2019). Register, Genre, and Style. Cambridge University Press, Cambridge. Second Edition.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10):341–345.

Boulenger, V. (this volume). Neural tracking of rhythmic information in speech: Where do we stand?

Braun, B. (this volume). Linguistic factors affecting amplitude modulation spectra. In L. Meyer and A. Strauss, editors, Rhythms of Speech and Language: Culture, Cognition, and the Brain. Cambridge University Press, Cambridge.

Brazil, D. (1985). The Communicative Value of Intonation in English. Second edition. Second edition 1997. Cambridge: Cambridge University Press.

Chhatwal, M., Sabetti-Franklin, S. and Vanden Bosch der Nederlanden, C. (this volume). Characterizing rhythmic regularity in speech and song.

Chomsky, N., Halle, M., and Lukoff, F. (1956). On accent and juncture in English. In Halle, M., Lunt, H. G., McLean, H., and van Schooneveld, C. H., editors, For Roman Jakobson. Essays on the Occasion of his Sixtieth Birthday, pages 65–80. Mouton & Co, The Hague.

Couper-Kuhlen, E. (1993). English Speech Rhythm: Form and Function in Everyday Verbal Interaction. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Couper-Kuhlen, E. and Selting, M. (2018). Interactional Linguistics. Studying Language in Social Interaction. Cambridge University Press, Cambridge.

Crystal, D. and Davy, D. (1969). Investigating English Style. Longman, London.

Cummins, F. and Port, R. (1998). Rhythmic constraints on stress timing in English. Journal of Phonetics, 26(2):145–171.

Daikoku, T. and Goswami, U. 2022. Hierarchical amplitude modulation structures and rhythm patterns: Comparing Western musical genres, song, and nature sounds to Babytalk. PLoS One. 2022.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11:51–62.

Dauer, R. M. (1987). Phonetic and phonological components of rhythm. In Proc. Eleventh International Congress of Phonetic Sciences, 447-450.

Frota, S., Vigário, M., Cruz, M., Hohl, F., Braun, B. (2022) Amplitude envelope modulations across languages reflect prosody. In Proc. Eleventh International Conference on Speech Prosody, 688-692.

Galves, A., Garcia, J., Duarte, D., and Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In Proc. First International Conference on Speech Prosody, 323-326.

Gibbon, D. (1976). Perspectives of Intonation Analysis. Lang, Berne.

Gibbon, D. (1981). Idiomaticity and functional variation. a case study of international amateur radio talk. Language and Society, 10:21–42.

Gibbon, D. (1985). Context and variation in two-way radio discourse. Discourse Processes, 8(4), 391–420.

Gibbon, D. (2006). Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In: Sudhoff, Stefan, Denisa Lenertova, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter and Johannes Schließer, editors, Methods in Empirical Prosody Research, pages 281-209 Walter de Gruyter, Berlin.

Gibbon, D. (2021). The rhythms of rhythm. Journal of the International Phonetic Association, First View [online], pages 1–33. [print, 2nd edition] 2023, (1):233-245.

Gibbon, D. (2022). Speech rhythms: Learning to discriminate speech styles. In Proc. Eleventh International Conference on Speech Prosody, 302–306.

Gibbon, D. (2023). Rhythm pattern discovery in Niger-Congo story-telling. Frontiers in Communication 8, pages 1–18. Sec. Psychology of Language; Research Topic: Science, Technology, and Art in the Spoken Expression of Meaning.

Gibbon, D. and Fernandes, F. R. (2005). Annotation-mining for Rhythm Model Comparison in Brazilian Portuguese. In Proc. Interspeech, 3289-3292.

Gibbon, D. and Li, P. (2019). Quantifying and correlating rhythm formants in speech. In Proc. Third International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS), Academia Sinica, Taipei.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In Gussenhoven, C. and Warner, N., editors, Laboratory Phonology 7, pages 515–546, De Gruyter Mouton, Berlin, New York.

Greenberg, S. (this volume). A polychromatic portrait of speech rhythm.

Greenberg, S., and Kingsbury, B. (1997). The modulation spectrogram: in pursuit of an invariant representation of speech. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 3, 1647-1650.

He, L. and Dellwo, V. (2016). A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert Transform. In Proc. Interspeech, 530–534.

Inden, B., Zofia, M., Wagner, P., and Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. In Miyake, N., Peebles, D., and Cooper, R. P., editors, Proc. Thirty-fourth Annual Conference of the Cognitive Science Society, pages 1721–1726, Austin TX. Cognitive Science Society.

Jassem, Wiktor. 1952. Intonation of Conversational English (Educated Southern British). Wrocławskie Towarzystwo Naukowe, Wrocław.

Jassem, W., Hill, D. R., and Witten, I. H. (1984). Isochrony in English speech: Its statistical validity and linguistic relevance. In Gibbon, D. and Richter, H., editors,

Intonation, Accent and Rhythm. Studies in Discourse Phonology, pages 203–225. Walther de Gruyter, Berlin.

Kalashnikova, M., Fernández-Merino, L. and Russo, S. (this volume). Rhythmic structure in cross-modal infant-directed communication.

Köhler, W. (1929). Gestalt Psychology. Horace Liveright, New York.

Lehiste, I. (1970). Suprasegmentals. Cambridge, MA, MIT Press.

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

Lin, X. and Gibbon, D. (2019). Classroom Reading: Speech Assessment from a Phonetic Perspective. In Proceedings of the International Symposium on SLA-based Language Pedagogy, Jinan University, Guangzhou, 312–318.

Lin, X. and Gibbon, D. (2023). Distant rhythms: calculating fluency. In Proc. Twentieth International Congress of Phonetic Sciences. International Phonetics Association, 4219-4223.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, Speech Production and Speech Modelling, pages 403–439. Kluwer Academic Publishers, Dordrecht.

McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. European Journal of Neuroscience, 48(7): 2609-2621.

Nakamura, S. and Sagisaka, Y. (2011). A requirement of texts for evaluation of rhythm in English speech by learners. In Proc. Seventeenth International Congress of Phonetic Sciences, pages 1438–1441.

Nolan, F. and Jeon, H.-S. (2014). Speech rhythm: a metaphor? In Transactions of the Royal Society B, Biological Sciences, pages 1-11.

O'Dell, M. L. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In Proceedings of the 14th International Congress of Phonetic Sciences, pages 1075–1078.

Ohala, J. (1992). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In Papers from the Parasession on the Syllable, pages 319–338, Chicago. Chicago Linguistics Society.

Pierrehumbert, J. (1980). The Phonetics and Phonology of English Intonation. Dissertation, MIT, Cambridge MA.

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

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report no. RuCCS-TR-2.

Rathcke, T., Lin, C., Falk, S. and Dalla Bella, S., (2021). Tapping into linguistic rhythm. *Laboratory Phonology* 12(1).

Selkirk, E. O. (1984). Phonology and Syntax. The Relation between Sound and Structure. MIT Press, Cambridge MA.

Sweet, H. (1908). The Sounds of English. Oxford: Clarendon Press.

Thomson, R. I. (2015). Fluency. In Reed, M. and Levis, J. M., editors, The Handbook of Pronunciation, pages 209–226. Wiley, Hoboken NJ.

Tilsen, S. and 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.

Tilsen, S. and Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. Journal of the Acoustical Society of America, 124(2):34–39.

Todd, N. P. M. and Brown, G. J. (1994). A computational model of prosody perception. In Proc. International Conference on Speech and Language Processing, 127–130.

Tracy, R. & Gibbon, D. (2023). The Beat Goes On: A Case Study of Timing in Heritage German Prosody. In: M. Beißwenger, E. Gredel, L. Lemnitzer, R. Schneider (editors): Korpusgestützte Sprachanalyse. Linguistische Grundlagen, Anwendungen und Analysen. (Studien zur Deutschen Sprache 88), pages 261-283. Tübingen: Narr.

Traunmüller, H. (1994). Conventional, biological, and environmental factors in speech communication: a modulation theory. Phonetica, 51(1-3), 170–183.

Trouvain, J. & Braun, B. (2020). Sentence prosody in a second language. In Gussenhoven, C. & Chen, A., editors, The Oxford Handbook of Language Prosody. Oxford University Press, pages 605-618.

Wagner, P. (this volume). Interaction phonology – rhythmic coordination as scaffold for communicative alignment.

Wittgenstein, L. (1953). Philosophische Untersuchungen. Blackwell, Oxford.

Yu, J. (2013). Timing analysis with the help of SPPAS and TGA tools. In Bigi, B. and Hirst, D., editors, Tools and Resources for the Analysis of Speech Prosody (TRASP), Aix en Provence. Université de Aix en Provence.