Rhythms of rhythm

Dafydd Gibbon (2020-10-06, rev. 2021-01-10)

Abstract

The low frequency (LF) spectral analysis or 'rhythm spectrum' approach to the quantitative analysis and comparison of speech rhythms is extended beyond syllable or word rhythms to 'rhetorical rhythms' in read-aloud narratives, in a selection of exploratory scenarios. Current methodologies in the field are first discussed, then the choice of data is motivated and the rhythm spectrum approach is applied on modulation-theoretic grounds to both amplitude modulation (AM) and frequency modulation (FM) of speech. New concepts of *rhythm formant*, *rhythm spectrogram* and *rhythm formant trajectory* are introduced in the *Rhythm Formant Theory* (RFT) framework with its associated methodology *Rhythm Formant Analysis* (RFA) in order to capture second order regularities in the temporal variation of rhythms. The interaction of AM and FM rhythm factors is explored, contrasting LF spectral features of the accent similarity constraint in English with those of arbitrary sequences of lexical tones in Mandarin Chinese. The LF rhythm spectrogram is introduced in order to recover temporal information about long-term rhythms, and to investigate signal-to-symbol mapping between LF spectrogram patterns and morphophonological patterns. The trajectory of highest magnitude frequencies through the component spectra of the LF spectrogram is extracted and applied in classifying readings in different languages using distance-based hierarchical clustering, following the practice in dialectometry and stylometry, but with speech signals, not texts, and the existence of long-term second order 'rhythms of rhythm' in long narratives is shown. In the conclusion, pointers are given to the extension of this exploratory rhythms of rhythm approach for future quantitative confirmatory investigations.

1 Speech rhythms

1.1 Time domain and frequency domain methods

Speech rhythms have been a field of enquiry since antiquity, yet there are still many open questions. The topic is multi-faceted, and has been addressed with many different methods in several neighbouring disciplines, particularly in the past hundred years. Nevertheless, the full potential even of more recently developed methods is only just being worked out. Several different concepts of rhythm have been and are used, depending on which branches of

psychology, medicine, linguistics and phonetics or musicology are concerned, from rhetorical rhythms to prominence relations between syllables, and each approach has addressed the problem of the empirical grounding of the elusive concept of rhythm in different ways, each providing a piece of the overall puzzle.

The methodology proposed in the present study starts with current methods for investigating short-term rhythms in physical signals and expands them to enable analysis of 'rhythms of rhythm', the dynamics of long-term rhythm variation, for example in story-telling or speeches. This understanding of speech rhythms is close to the common-sense understanding of rhythm as regularly occurring beats, but the complexities of the multiple rhythms of natural speech are better understood as regular oscillations with specifiable low frequencies below about 10 Hz, deriving from the neural patterns of resonance which drive the articulatory and phonatory muscles. Utterances vary in their rhythms, depending on the lexical and grammatical typology of the language, on rhetorical style and on idiosyncratic features.

The various different concepts of rhythm in the literature depend partly on the data selected, but mainly on the intradisciplinary and transdisciplinary methods used. Some approaches such as conversation analysis are based on qualitative methods, with rhythm seen as hermeneutically understood temporal regularity (Couper-Kuhlen and Auer 1991). Phonological approaches combine qualitative methods with top-down formal methods, defining 'abstract rhythms' as structures with paired positions labelled 'strong' and 'weak' since Chomsky et al. (1956), extended with hierarchical (Selkirk 1984) and linearising (Liberman and Prince 1977) constraints. Since Jassem et al. (1984), in many quantitative phonetic studies a top-down approach has also been taken, using manual or (semi-)automatic annotation of the speech signal by pairing timestamp measurements with previously identified locutions, in a search for regular, possibly isochronous (equal duration) units such as syllables or feet. In spite of the motivating goal of identifying the physical empirical grounding of speech rhythms as isochronous (equally timed) utterance segments, physical isochrony has generally been regarded sceptically in these approaches.

Taking a strictly physical approach, with automatic bottom-up signal processing approaches to phonetics and speech pathology, a more optimistic position on the physical grounding and automatic analysis of rhythm has been taken since Todd and Brown (1994) and Cummins et al. (1999). Unavoidably, on a meta-level, qualitative choices of formal

procedures are also made in these approaches, of course. In the bottom-up approaches, rhythms are not primarily conceived as regular durations in the time domain, but rather as spectral properties in the frequency domain, generally as regular oscillations at given frequencies, identifiable as magnitude peaks in the low frequency (LF) spectrum of the speech signal, with isochrony as a secondary *a fortiori* consequence of oscillation. The qualitative and quantitative, time-domain and frequency domain approaches each have their justification, but with each providing only part of the overall picture (Kohler 2009).

The following subsections provide a brief overview of a new approach, Rhythm Formant Theory (RFT) and its associated methodology (RFA), and then a more detailed account of relevant previous studies of speech rhythms. Section 2 describes the data and methods used in the present study along with Rhythm Formant Theory and its associated methodology, Rhythm Formant Analysis. Section 3 addresses the relation of amplitude modulation (AM) and frequency modulation (FM) in speech rhythm with reference to English and Mandarin Chinese. Section 4 examines signal-symbol association using both annotation-based and spectral analysis methods. Section 5 describes an exploratory experiment using RFA with readings of a narrative by a bilingual speaker with high proficiency in the two languages German and English, and introduces a basic unsupervised machine learning distance-clustering technique to classify the readings based on spectral variation in the low frequency spectrogram, before exploring the limits of this technique in comparisons of larger, more inhomogeneous sets of readings in Section 6. In Section 7, the results are summarised, conclusions are drawn, and prospects for further development are outlined.

1.2 Basics of RFT

In the present study, an exploratory frequency domain approach to the study of rhythm is developed, where ‘exploratory’ means a novel theory-guided initial study of an empirical domain, without necessarily performing full statistical confirmatory analyses. The domain of rhetorical rhythm is shared with discourse analysis, and grammatical structure also plays a role. Long-term dynamic changes in physical speech rhythms are examined, on the one hand in relation to structural properties of locutionary units such as syllables and words, on the other hand as long-term rhetorical ‘rhythms of rhythm’ patterns in story readings in English and Mandarin Chinese. The emphasis is on analysing details of rhythms in utterance tokens from individual speakers in specific genres, as in earlier studies, rather than in claiming validity for the typology of entire languages, as in more recent studies of rhythm.

A new bottom-up definition of the physical properties of speech rhythms is proposed, based on the modulation theoretic perspective of signal processing:

> *Speech rhythms are fairly regular oscillations below about 10 Hz which modulate the speech source carrier signal and are detectable in spectral analysis as magnitude peaks in the LF spectrum of both the amplitude modulation (AM) of the speech signal, related to the syllable outline of the waveform, and the frequency modulation (FM) of the signal, related to fundamental frequency (F0) or perceived pitch contours of the carrier signal.*

The restriction 'fairly regular' means that the oscillations are not based on a precision clock but fall into approximate ranges. The plural 'rhythms' implies that different speech rhythms coexist. The frequency zones around magnitude peaks in the spectrum constitute *rhythm formants*, lending the name *Rhythm Formant Theory* (*RFT*) to this approach (the terminology is justified in subsection 2.4). Crucially, RFT is concerned not only with the contribution of AM to speech rhythm, but also with the contribution of FM (cf. also Varnet et al. 2017; Gibbon 2018, 2019; Gibbon and Li 2019; Ludusan and Wagner 2020), and not only with identification of a rhythm but with rhythm variation over time. The RFT definition of rhythm implies *a fortiori* that rhythms have two main properties: *frequency* and *magnitude*, and concepts such as isochrony in speech timing are derived from these more fundamental concepts. Figure 1 shows four components of an RFT analysis of low frequency rhetorical rhythms as oscillations, in the first 17 seconds of the well-known and widely available 1963 speech of Martin Luther King, *I have a dream*.

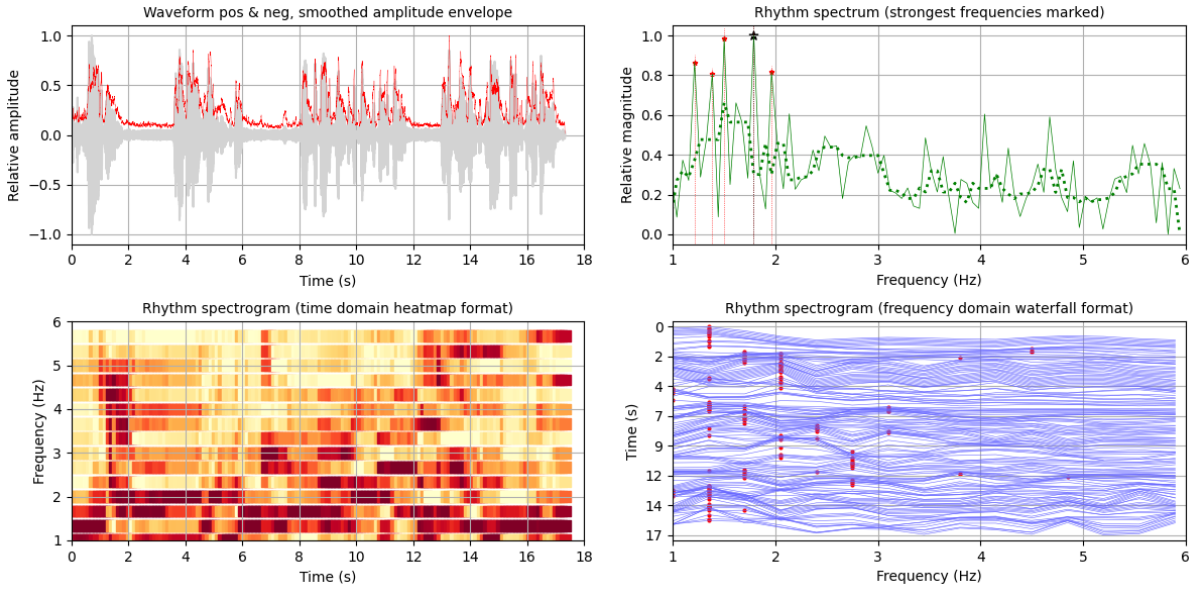


Figure 1: RFT AM analysis of a segment of Martin Luther King's 1963 speech: "I have a dream that one day on the red hills of Jordan the sons of former slaves and the sons of former slave-owners will be able to sit down together at the table of brotherhood...". Upper left: waveform and amplitude envelope. Upper right: long-term LF spectrum. Lower right: waterfall format long-term spectrogram. lower left: heatmap format long-term spectrogram.

The analysis steps are, in clockwise direction: upper left, the positive *AM envelope* (superimposed on the *waveform*); upper right, the AM LF *spectrum* (5 highest magnitude spectral frequencies marked, the dot sequence represents smoothing with a median window); lower right, the AM LF *rhythm spectrogram* in a *waterfall format*; lower left, the same AM LF *rhythm spectrogram* but in traditional *heatmap format*. The waterfall and heatmap formats are both derived directly from the same numerical matrix but have quite different heuristic values as visualisations.

In the waterfall spectrogram format, the highest magnitude spectral frequency is marked for each component spectrum, with the sequence of these frequencies constituting the *rhythm formant trajectory*. This trajectory is very variable, and demonstrates the variability of LF spectral frequencies through time. The spectrogram representation (lower right) shows that the high magnitude frequencies in the spectrum (top right) are not simultaneous, as the atemporal spectrum may be taken to imply, but are distributed in time. The heatmap spectrogram representation shows more detail about the temporal dynamics of rhetorical speech rhythms, in particular a rhetorical *rhythm cycle*, spaced at about 7 s, the 'rhythms of rhythm' with three clear peaks in the highest magnitude frequencies. Frequencies below 1 Hz are not included because they are better represented as second order long-term temporal

variations of short-term rhythms in the spectrogram, *rhythms of rhythm*, than as frequencies in the atemporal long-term spectrum (cf. Sections 3, 4, 5, 6).

The frequency and time domain LF spectrogram representations involved in RFT each involve a very different 'mindset' in observational practices from the annotation-based time domain analyses of interval durations. The frequency domain approach suggests different questions, analyses, modelling conventions, visualisations and answers, even though frequencies and durations are closely related (i.e. $f = 1/T$, where f stands for frequency and T represents Δt, the period, i.e. the duration of a cycle).

RFT extends the basic spectral analysis approach with the following new concepts:

1. the low frequency (LF) zone of spectral peaks in the LF spectrum (cf. top right graph in Figure 1, the vertical dotted lines topped with stars), interpreted as rhythm formant, by analogy with the mathematically similar high frequency (HF) formant or phone formant in the high frequency (HF) regions of the spectrum, which distinguishes phones (speech sounds);
2. the LF spectrogram (cf. the waterfall and heatmap spectrogram formats in Figure 1, bottom right and bottom left), interpreted as rhythm spectrogram;
3. the LF formant trajectory (not shown in Figure 1; cf. Figure 9 bottom left), interpreted as rhythm formant trajectory of the highest magnitude peaks in the component spectra of the LF spectrogram.

RFT is also applied to the *FM envelope* of the speech signal, the fundamental frequency (F0) track which is associated with tones, pitch accents and intonation, in order to examine the role of F0 in the prosodic typology of speech rhythms (cf. also Varnet et al. 2017; Gibbon 2018, 2019; Gibbon and Li 2019; Ludusan and Wagner 2020). Other previously mentioned spectral analysis approaches to the study of rhythm have concentrated on the *AM envelope* of a speech signal, associated with the sonority curves of syllables, words and longer units.

Unlike many previous studies based on spectral analysis, which have developed biological and psychoacoustic models of rhythm, RFT is agnostic with regard to theories of rhythm production and perception. RFT is open to interpretation in these fields, but is more oriented towards further developing previous practical work in the field, for example linguistic and phonetic description and explanation (Gibbon and Li 2019), practical applications in individual speech diagnosis (Liss et al. 2010; Carbonell et al. 2015), small group language

testing (Lin and Gibbon 2020; Wayland and Nozawa 2020), and speech technology applications (LeGendre et al. 2009).

1.3 Qualitative and quantitative, top-down and bottom-up approaches

This is not the place for a comprehensive history of stress, accent and rhythm studies, but numerous critical overviews of the methods used in rhythm studies are available, including Adams (1979), Dauer (1983), Jassem et al. (1984), Gibbon (2006), Arvaniti (2009), Gut (2012), Wayland and Nozawa (2020). White and Malisz (2020) provide a particularly comprehensive survey of the main formal and quantitative phonetic approaches to rhythm analysis. The present study is more concerned with phrasal and discourse rhythms, but it is nevertheless useful to note the five main paradigms as general orientation:

1. qualitative and functional or rhetorical analysis, often with strong influences from music, dance and poetry, which can be traced back to Aristotle and Cicero and continues in conversational analysis (Brazil et al. 1980; Couper-Kuhlen and Auer 1991; Couper-Kuhlen and Selting 2018);
2. linguistic models with a pedagogical background, from Sweet's stress-syllable timing distinction (1908), Jones (1909), Jones (1918) and Palmer's tonetic structures (1924), through Pike's stress numerals (1945) to Jassem's rhythm hierarchy (1952) and the metrical feet of Abercrombie (1967), partial overview by Gibbon (1976);
3. algebraic models of recursive stress hierarchies from Chomsky et al. (1956) through Chomsky and Halle (1968) and Liberman and Prince (1977), to Selkirk (1984) and to optimality theories (Prince and Smolensky 2004);
4. a linguistic-phonetic interface paradigm relating linguistic units to intervals in the speech signal by annotating them with timestamps, often in an attempt to find 'rhythm classes' of languages in terms of irregularity of timing, for example Lehiste 1970, Jassem 1952, Roach 1982, Jassem et al. 1984, Scott et al. 1985, Ramus et al. 1999, Low et al. 2000, Asu and Nolan 2006, Li and Yin 2006, Wagner 2007, Dellwo 2010, Yu and Gibbon 2015, Dihingia and Sarmah 2020;
5. a modulation-theoretic signal-processing paradigm, with production and perception models which represent low frequency components of the speech signal, from the rhythmograms of Todd et al. (1994) and Ludusan et al. (2011) through the coupled oscillators of Cummins and Port (1998), Barbosa (2002), Malisz et al. (2016) and the

sonority patterns of Galves et al. (2002), to the low frequency envelope spectrum of Tilsen and Johnson (2008), Tilsen and Arvaniti (2013), Gibbon (2018).

There are many more studies in each paradigm, but the items here are appropriate representatives for investigations on, mainly, English, though many studies have typically made typological comparisons between languages.. Relevant aspects are discussed further in the following subsections.

A standard procedure in many quantitative phonetic analyses has been the top-down manual or (semi-)automatic annotation method referred to above, of measuring and recording the alignment of linguistically defined event tokens (vocalic and consonantal segments, syllables, and words or feet) with points or intervals in the speech signal as timestamps paired with transcriptions. Descriptive statistical techniques variously known as irregularity metrics, isochrony metrics, interval metrics or rhythm metrics are applied to the timestamps in order to capture regularities and irregularities of durations in the speech signal which may be ascribed to rhythmic and arhythmic segments of speech utterances, in particular in search of an *isochrony* property (equal timing of units in a sequence). Despite the quantitative properties of annotation-based analyses, a qualitative component of human judgment is also involved: the studies are not 'acoustic' in the signal processing sense but filtered through the annotator's perception of the speech signal (as also noted by Tilsen and Johnson 2008; Liss et al. 2010). The uses of annotated speech signals are many: illustration in qualitative and formal linguistic studies, quantitative analysis of interval durations in descriptive phonetics, training speech recognisers and synthesisers, and text-based archive search.

One class of irregularity metric in the search for isochrony relates directly or indirectly to variance or standard deviation, i.e. differences of unit durations in a sequence or dispersion from the mean unit duration, for example Ramus et al. (1999) on irregularities in consonantal or vocalic intervals and Roach (1982) on percentage deviation in interstress intervals. Similarly, the Irregularity Measure of Scott et al. (1985) compares all durations in a sequence using the absolute value of subtraction of logarithms (in the form of the logarithm of a division). These metrics are actually more suitable for describing unordered sets than sequences, as they ignore the ordering of intervals destroys the alternation property of rhythm, but they offer a heuristic procedure as a starting point. They also do not capture well-known properties of rhythms such as left and right headedness which have been described in phonological studies, as noted by Varnet et al. (2017) and do not distinguish between local

rhythm variation and global tempo variation. Results from these measures have not yielded useful insights about isochrony but have nevertheless been useful heuristic starting points for further study.

The pairwise variability index (PVI), with 'raw' (rPVI) and 'normalised' (nPVI) versions (Low et al. 2000, and many other studies), was introduced to handle the tempo variation issue. Asu and Nolan (2006) noted that the one-dimensional PVI metrics only provide a 'more or less' result for the quantity measured (e.g. syllables), with no information about complementary categories (e.g. stress), and consequently apply PVI metrics in two dimensions, to syllables and feet.

The PVI metrics are unrelated to variance, as each interval duration is compared locally with the immediately following interval duration, and they are inherently more suitable for describing irregularity in time-ordered sequences. The PVI metrics relate formally to standard binary distance metrics for vector comparison. The rPVI relates to Manhattan Distance[1] and the nPVI relates to Canberra Distance, the normalised version of Manhattan Distance. In *PVI* distance measurement, the vectors V_1 and V_2 are not independent, as is generally the case, but subvectors of the same vector of interval durations $V=<d_1, \ldots, d_m>$:

$V_1 = <d_1, \ldots, d_{m-1}>$

$V_2 = <d_2, \ldots, d_m>$.

The binary distance relation $nPVI(V_1,V_2)$ thus defines a binary 'next-door-neighbour' distance, which has been represented by Wagner (2007) as a two-dimensional scatter plot. The distance relation embodies the assumption that, formally, rhythm is binary (cf. Gibbon 2003), expressed in the 'pairwise' concept which is enshrined in the name of the metric. Nolan and Jeon later (2014:3) weaken the binarity assumption to "sufficient predominance of strong-weak alternation". Otherwise, the PVI metrics share the same issues already noted for the variance-based irregularity metrics. In contrast to earlier metrics, the PVI metrics have been widely used in a number of disciplines as a heuristic for comparing language (or music) varieties on the basis of the next-door-neighbour distances in duration annotations. The PVI metrics have been reviewed exhaustively by Barry et al. (2003), Gibbon (2003, 2006), Tortel and Hirst (2008), Arvaniti (2009), Kohler (2009), Gut (2012), White and Malisz (2020).

1 Manhattan Distance, also known as Cityblock Distance or Taxicab Distance, is the 'round the corner' distance between two opposite points of a rectangle, as opposed to Euclidean distance, which is the direct 'as the crow flies' distance.

A second quantitative approach, with origins in signal processing and speech pathology, is the bottom-up spectral analysis method based on direct analysis of LF oscillations in the speech signal without reference to linguistically defined units (except for the empirically identified data units). These approaches measure the frequencies of oscillations in speech with a wide variety of signal processing methods, including band pass filtering, Hilbert Transform, Fourier Transform or wavelet analysis: Todd and Brown (1994); Cummins et al. (1999); Foote and Uchihashi (2001); Galves et al. (2002); Lee and Todd (2004); Tilsen and Johnson (2008); Heřmanský (2010); Liss et al. 2010; Ludusan et al. (2011); Tilsen and Arvaniti (2013); Leong et al. (2014); Fuchs and Wunder 2015; He and Dellwo (2016); Gibbon (2018); Wayland and Nozawa (2020); Ludusan and Wagner (2020), often as models of speech perception (Cumming 2010). Models of speech production based on the same formal foundation, in principle, but with coupled oscillators to handle mutual relations between different low frequencies, have also been developed (cf. O'Dell and Nieminen 1999; Barbosa 2002; Cummins and Port 1998; Inden et al. 2012). Cyclical finite state models of pitch accent sequencing in intonation and of tone sandhi in lexical tone languages (Pierrehumbert 1980; Gibbon 1987; Jansche 1998) can also be construed as models of abstract structural rhythm, not unlike the coupled oscillator models of the oscillator models of O'Dell and Nieminen (1999) or Barbosa (2002).

Related concepts to the rhythm spectrum approach taken here were developed by Todd and Brown (1994) and Ludusan et al. (2011) with the *rhythmogram*, by Ioannides and Sargasyan (2012), based on an amplitude-sensitive and frequency-sensitive auditory hair cell demodulation algorithm with pre- and post-demodulation filtering, and by Foote and Uchihashi (2001) with the *beat spectrum* and *beat spectrogram* for identifying rhythm variation over time in music (cf. also Brown et al. 2017).

The overall guiding interest in the spectral analysis paradigm has often been the diagnosis of rhythms in the speech of individual speakers or speaker types for the purpose of medical diagnosis and therapy tracking, rather than in language typology, though analyses with a typological goal have also been made (Liss et al. 2010; Varnet et al. 2017). As a matter of interest, disco lights, as well as the popular music detection application Shazam (Wang 2003), also use varieties of this technique.

All of the approaches mentioned so far, whether based on annotation or on spectral analysis, examine properties of the amplitude modulation of the speech signal, with

demodulation and spectral analysis of the AM envelope (meaning the smoothed outline of the extreme values of a signal). In an early application of this technique in phonetics, Dogil and Braun (1988) used the AM envelope, in the form of 'intensity tracing', to detect 'pivots' (fast amplitude changes) for edge detection in speech segmentation.

In the present RFT approach the same methods are applied to both AM and FM, from rhythm formants through the rhythm spectrogram to the rhythm formant trajectory, creating common ground for the direct comparison of the contributions of AM and FM to the rhythmicality of speech.The role of AM demodulation in speech rhythm has been studied relatively frequently, as discussed above. However, only a few studies relate to FM in rhythm modelling, including Cumming (2010), Varnet et al. (2017), Gibbon (2018, 2019), Gibbon and Li (2019) and Ludusan and Wagner (2020). FM has also been very widely studied, but in terms of the fundamental frequency (F0) contours associated with form and function tone, pitch accent and intonation and with perceived prominence (cf. Malisz and Wagner 2012; Suni et al. 2017; Kallio et al. 2020 for recent treatments).

2 Method

2.1 Data

The main data type used in this study is read-aloud narrative. Although dialogue data are often considered more 'natural' and 'authentic' and may have greater intrinsic interest for many purposes, the activity of reading aloud has high cultural value in activities ranging from bedtime stories for children through newsreading and lecturing to assistive support for visually challenged readers, and understanding this genre is of interest in itself. The selected narratives are English and German translations of the IPA benchmark fable attributed to Aesop, *The North Wind and the Sun*, recordings of which have also figured in earlier spectral analysis studies (e.g. Tilsen and Arvaniti 2013). In order to facilitate validation in future studies, open access recorded data from the Edinburgh *The North Wind and the Sun* corpus[2] (Abercrombie 2013) are used in the present study. The corpus was recorded in the period 1951-1978, and contains 87 WAV audio files with recordings of translations into 99 different languages and language varieties, from Arabic to Yorùbá. In general there is one recording of the narrative per file, but the Scottish English file, for example, contains 12 recordings of readings in different accents. Of the 14 readings in English, 12 are Scottish English, 1 is

2 https://datashare.is.ed.ac.uk/handle/10283/387?show=full

Northern British English and 1 is received pronunciation with minor Scottish influence. Of the 7 German recordings, 5 are Swiss German and 2 are Standard Northern German. One reading by each person was recorded except for a small number of recordings by bilinguals.

For fine-grained comparison of more than one reading per person, additional recordings of *The North Wind and the Sun* and *Nordwind und Sonne* were recorded with a female adult bilingual speaker of English and German, with pronunciation features: (a) slight southern Rhineland accent in German; (b) British and North American pronunciation elements in English. The reader also has extensive experience in lecturing and literary readings.

For initial illustration of some aspects of the methodology, the first 17 seconds of an open access recording of of Martin Luther King's 1963 speech "I have a dream" were examined (cf. Subsection 1.2).

For 'clear case' illustration of the RFT methodology, recordings of the genre of rhythmic counting aloud were made, following established practice in many earlier studies. Two sequences, both in English, are used: first a short sequence of counting from one to seven for illustrating general principles, then a longer sequence of counting from one to thirty for investigating morphophonological factors in rhythmic sequences. A counting sequence spoken by a female speaker of Mandarin Chinese was also recorded for comparison of FM properties in a language with different prosodic typology. The phrase-initial and phrase-terminal properties of counting aloud are not considered here; the focus of attention is on the phrase-internal sequence of counted numerals. In addition, a formally defined calibration signal of 200 Hz, amplitude modulated at 5 Hz, modulation depth 50%, was synthesised in order to illustrate formal aspects of the analysis procedure.

Prior to analysis, sampling rates were standardised to 16 kHz in order to reduce computing load, and further downsampled within the software for specific LF operations. The Edinburgh recordings contain random spoken metadata information, which was deleted for the study in the interest of minimising scenario variables. Initial and final silences in the original recorded data are of random lengths and were shortened to 0.5 s; the synthetic signal has no silences. Random outlier noise spikes in the data were reduced in amplitude as far as possible without affecting neighbouring signal values, and amplitudes were standardised for internal processing and display purposes. No time normalisation is undertaken except in later stages of the study for correlation and hierarchical classification calculations with LF formant trajectories extracted from LF spectrograms.

2.2 RFT methodology

An important part of phonetic modelling is the discovery of signal-symbol relations, the association of physical sound with linguistic categories. In this study, both bottom-up LF spectral analysis and top-down annotation with prior linguistic categories are combined in order to relate the two domains (Section 4, cf. Figure 8). On the numerical side, distance metrics play three different roles in the study: in addition to the explication of the nPVI irregularity metric as next-door-neighbour distance in connection with signal-symbol association, distance metrics are applied directly, first in the identification of rhythm formants in the LF spectrum, and second in the distance-based clustering of readings in different languages and language varieties, using rhythm formant trajectories.

The present study takes a modulation theoretic approach, in which the key concepts are the *carrier signal* and the FM and AM *modulation signals*, with the same meanings as in radio frequency signal processing. These key concepts are illustrated in Figure 2 from the viewpoint of synthesis. The task of RFA, the RFT methodology, is the inverse: to recover the LF FM and AM modulations from the modulated signal.

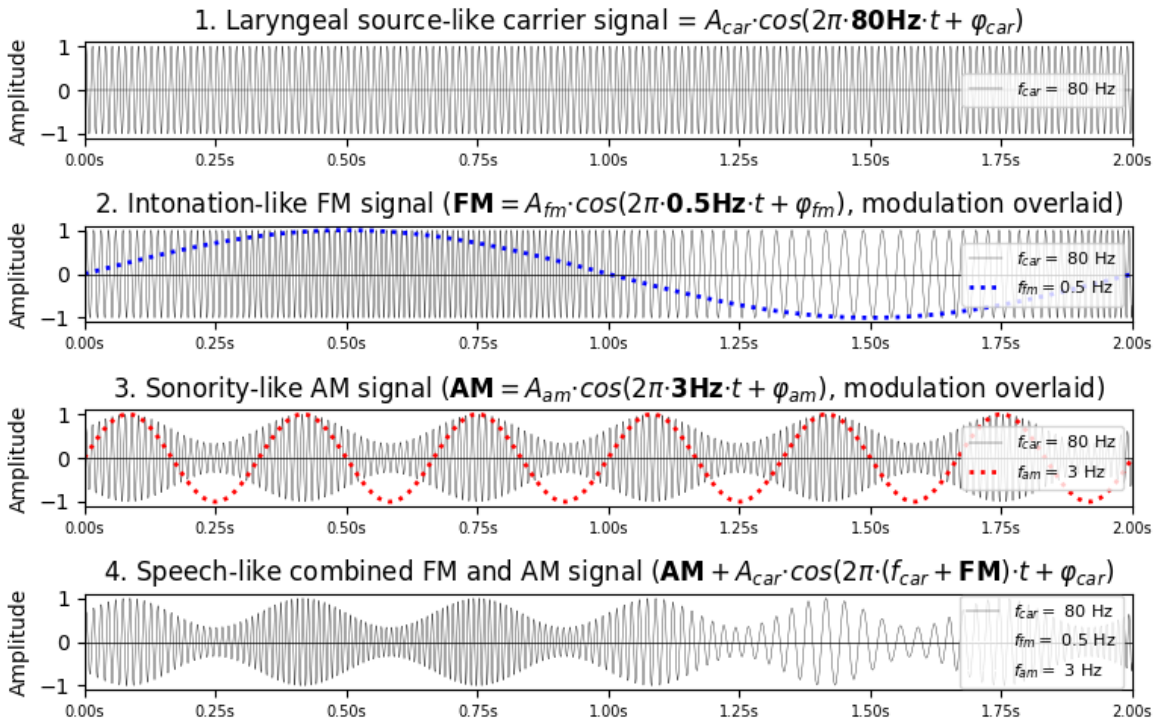


Figure 2: Stylised 'speech-like' model of the principles of modulation with combined FM and AM. At each point in time, the frequency component of the carrier signal (1), here 80 Hz, is changed in frequency by adding the corresponding value of the 0.5 Hz FM modulation signal (2); the frequency modulated signal is changed in amplitude by adding the amplitude value of the 3 Hz AM signal (3); the result is a combined FM and AM signal.

Modulations superimpose information-carrying signals on the carrier signal. There are many kinds of signal modulation, and several kinds are relevant for the study of speech. The two main kinds which figure in the present study are frequency modulation (FM) and amplitude modulation (AM), illustrated in Figure 2. The principle is that with FM the frequency of a source signal (e.g. from the larynx) is modulated, driven by tone, pitch accent and intonation, and with AM the amplitude of the resulting FM signal is modulated, driven by syllable, foot and phrase sonority patterns, resulting in the speech signal, which is both FM and AM (among other modulation properties). The task of RFA is to recover the FM and AM information from the modulated signal.

Both FM and AM have their general signal processing meanings. Simplifying, since the speech source signal is more complex, for a basic sinusoid carrier signal Scar = Acarcos(2πfcart+φcar) and a basic sinusoid modulation signal Smod = Amodcos(2πfmodt+φmod):

1. in FM, the values of f_{car}, the frequency component, are modified (for the functions of lexical tone, pitch accent and in intonation) by addition with the A_{mod} values at source;
2. in AM the component A_{car} (amplitude) of S_{car} is modified (for vowel and consonant sequences) by adding or multiplying with the corresponding A_{mod} values of the modulation signal.

For present purposes, the phase φ of the carrier and modulation signals is not considered.

Unlike in speech acoustics and speech technology, in phonetics the terms ‘AM’ and ‘FM’ are not commonly discussed as such. The distinction is usually made in terms of source-filter models of speech phonation and articulation and the two domains tend to be treated as entirely different in ‘segmental’ versus ‘prosodic’ phonetics and phonology. Modulation theory is introduced here explicitly for a diametrically opposite reason: in order to be able focus on similarities of AM and FM as factors in speech rhythmicality, rather than on differences.

2.3 Procedure

The procedure followed in this study, parts of which are illustrated in Figure 1, is in a sense minimalist, in that a number of filtering and normalisation procedures used in previous studies are omitted since they do not make a substantive empirical difference for the present goal of inducing and comparing long-term physical rhythm patterns. The procedure covers eight signal processing stages (see also Section 5, and Figure 10, Figure 11, Figure 13 and Figure 15 for examples of these stages):

2.3.1 Downsampling

The signal is downsampled to a 16 kHz sampling rate to reduce computation load (common frequencies for audio recording being 44.1 kHz and 48 kHz). For processing the rhythm spectrogram, the signal is further downsampled.

2.3.2 AM demodulation

The positive signal envelope is extracted by rectification of the signal, i.e. by taking positive values of the signal (cf. the top line in Figure 1, top left), and smoothing the peak sequence. This yields the smooth curve or envelope which outlines the positive amplitude extremes of the signal (or, if squared, the intensity curve; cf. Praat, Boersma 2001, and the pivot parser 'intensity track' of Dogil and Braun 1988). The envelope can then be further analysed as a modulation signal in its own right. More generally, the amplitude envelope is extracted, either with the absolute values of the Hilbert Transform (cf. He and Dellwo 2016) together with low pass filtering or, for practical purposes, full-wave rectification (taking the absolute or squared values of the signal) followed by low-pass filtering. Two AM signals are relevant, with different time windows: (a) long-term volume change for emotional or rhetorical purposes in discourse contexts over time domains of several seconds or more, and (b) the oscillating shorter-term sonority curve of speech sounds which characterises syllable, word and phrase patterns, in time domains from centiseconds to seconds.

2.3.3 FM demodulation

The information-bearing fundamental frequency (F0) patterns are extracted from the signal to form the FM envelope. This envelope is also known as the F0 estimate or the F0 track (or, less accurately, pitch track, as 'pitch' refers to a percept, not the acoustic signal). FM is the physical basis for the linguistic information conveyed by lexical tone, pitch accent and intonation. The F0 estimator used in the present study is quite traditional: a modified *Average Magnitude Difference Function* (*AMDF*) algorithm (Krause 1984) with prior low-pass filtering and centre and peak clipping and later moving median window smoothing. AMDF is related to autocorrelation but uses subtraction, which is faster in the RFA implementation than the multiplication involved in calculating correlations. The 'discontinuities' in the FM envelope caused by voiceless consonants and pauses are treated as spectrally relevant segments of the signal, but normalised to zero for the figures and to median for spectral analysis.

2.3.4 LF spectrum analysis

LF long-term spectral analysis is performed on demodulated AM and FM envelopes by scaling to match the envelope extents and applying a Fast Fourier Transform (FFT) in each case. The transform is applied to the entire length of the modulation signal, yielding the LF long-term spectrum of the signal with frequencies <10 Hz. A flexible cosine window (Tukey window) is applied prior to FFT application. This step is related to the rhythm spectrum approach of Tilsen and Johnson (2008) except that a bandwidth restriction to reduce F0 influence is not included, because F0 variation is still represented in the harmonics; cf. also Wayland and Nozawa (2020). The LF spectrum is derived directly from the broad spectrum of the downsampled signal. The long-term signal-length FFT window is quite different from the 10 ms or so short-term window length of pitch extraction and phone formant analysis.

2.3.5 LF rhythm formant identification

High magnitude peaks in the LF spectrum are identified; spectral frequency zones around the highest magnitude peaks are termed rhythm formants. The purpose of this step in the analysis is to identify different communicatively relevant rhythms in the signal. Like the irregularity metrics, the LF spectrum contains no timing information about dynamic rhythm changes. Rhythm formants are related to the spectral bands and intrinsic mode functions of Tilsen and Arvaniti (2013) but are not assumed to relate directly to syllable and stress timings.

2.3.6 LF spectrogram analysis

In order to capture dynamic rhythm changes in possibly hierarchical (Campbell 1992; Sagisaka 2003) second order rhetorical ‘rhythms of rhythm’, a LF long-term spectrogram is created using shorter term moving FFT windows, generally <3 s, which step through the signal and generate a component spectrum at each step. The resulting spectrogram is displayed either in waterfall format (*frequency* × *time*) with magnitude variation indicated spatially by ‘waves’ in each spectrum line, or in spectrogram heatmap format (*time* × *frequency*) with magnitude variation as colour or grey scale differences; cf. also the LF AM *beat spectrogram* of Foote and Uchihashi (2001).

2.3.7 Rhythm formant trajectory identification

The highest magnitude peak in each component spectrum of the spectrogram is identified (cf. the waterfall spectrogram format in Figure 1), and the vector pair consisting of the magnitudes and frequencies of these peaks throughout the spectrogram is identified. The purpose of identifying this vector pair is to obtain distances between vectors as empirical indices of

dynamic timing variability, for use in the basic unsupervised machine-learning technique of distance-based hierarchical clustering.

2.3.8 Hierarchical speech variety classification

A set of readings of the selected narrative is analysed and, instead of extracting variables from the spectrum (Tilsen and Johnson 2008; Liss et al. 2010), frequencies and magnitudes of the rhythm formant trajectory vectors are calculated as component trajectories, introducing temporal information. The trajectories are compared in a novel procedure, using competitive evaluation of a set of combinations of standard distance measures and standard hierarchical clustering operations. Hierarchical clustering is used in preference to the more usual flat clustering (Wayland and Nozawa 2020) because it is more informative about the fine detail of classification. Evaluation is by majority vote: the relevant hierarchical clustering is taken to be the dendrogram shape (cluster tree shape) with the highest number of votes, i.e. shared by the largest number of distance-clustering combinations, regardless of the numerical degrees of similarity expressed by the dendrogram.

2.3.9 Main innovations of the procedure

The steps of the procedure which are outlined in the previous subsections extend previous AM LF spectrum approaches first, to include FM spectrum analysis, and second, to include the time variation dimension of LF spectrogram analysis for both AM and FM using the new concepts of rhythm formant, rhythm spectrogram and rhythm formant trajectory. Sets of rhythm formant trajectories from different speech recordings are classified using the unsupervised machine learning technique of distance-based hierarchical clustering.

2.4 Terminology: why ‘rhythm formant’?

In acoustics, whether in music or in speech processing, the term ‘formant’ refers to a frequency zone of higher magnitude in the spectrum of the harmonics of a periodic sonorant sound, or of a non-harmonic obstruent noise, which is independent of the fundamental frequency and which shapes the acoustic properties of a musical instrument or a speech sound. It may apply also to properties of the mechanism which shapes this frequency zone. In phonetics the term ‘formant’ is used for high frequency phone formants above several hundred Hertz and ranging to several thousand Hertz, which characterise speech sounds. In RFT, the acoustic definition of ‘formant’ is generalised to apply also to rhythm formants,[3] as

3 The term ‘formant’ for LF spectral peak zones was suggested by Dr. Huangmei Liu, Tongji University, Shanghai (personal communication).

low frequency higher magnitude spectral zones. The differences between HF formants or phone formants and LF formants or rhythm formants in modulation theory lie (a) in their frequency ranges, (b) in the superimposition of formants on fundamental frequency harmonics or non-harmonic noise (phone formants) versus superimposition on sonority patterns and (c) in their functionalities in speech communication.

In view of the long tradition of phone formant analysis, the term 'rhythm formant' may appear unusual, perhaps controversial. Nevertheless, the acoustic definition in terms of spectral peak zones is one which also applies in acoustics and in musicology. It may be helpful to locate the concept of rhythm formant in the acoustic frequency space occupied by speech (Figure 3).

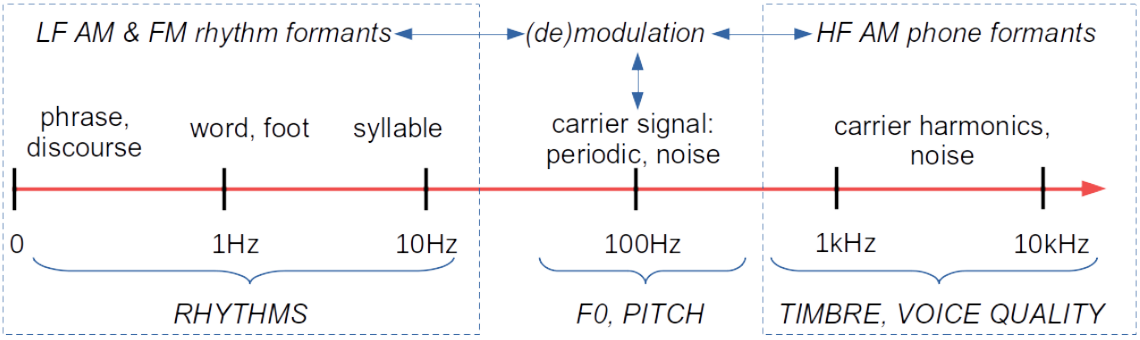


Figure 3: Modulation-theoretic frequency scale.

Figure 3 is based on on a modification of ideas from Cowell's (1930) classic theory of harmonic relations in musicology. The place of both HF phone formants and LF rhythm formants, as well as the carrier wave, can be illustrated straightforwardly on a logarithmic frequency scale of the frequencies used in human speech. At frequencies below 10 Hz, syllables, words and phrases are characterised by slow modulations of both amplitude and frequency (among other properties). If prominent spectral peaks can be identified, then variations are fairly regular and are interpreted as a rhythmical.

2.5 LF spectral analysis tool

Empirical study of spectral analysis and its RFT expansions is not possible without a dedicated software tool. For this purpose, a tool was developed in Python3, using the libraries NumPy for numerical calculation, MatPlotLib for graphics and SciPy for distance and clustering algorithms.[4] The Average Magnitude Difference (AMDF) algorithm for FM demodulation, together with pre-modulation and post-modulation procedures, was custom

4 The Python code is distributed in a research prototype CLI version: https://github.com/dafyddg/RFA

designed in ‘pure Python’. Figure 4 shows a synthetic signal which was generated for demonstrating the RFT-related properties of the tool (compare with Figure 1); cf. the vocoder procedure of Leong et al. (2014).

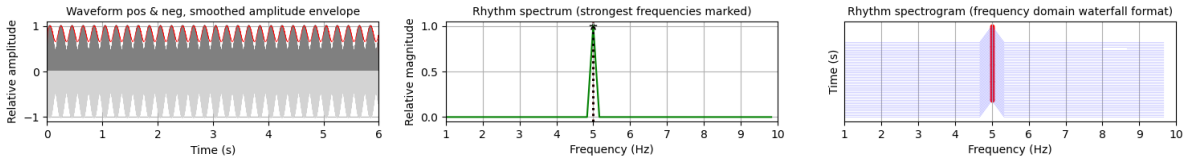


Figure 4: RF analysis of a 5 Hz synthetic sinusoid amplitude modulation of a 200Hz carrier signal (time is left-right in the left hand graph, and top to bottom in the waterfall format; frequency is left to right in the centre and right-hand graphs).

A 200 kHz sinusoid function of 6 s duration was defined with 5 Hz sinusoid amplitude modulation, modulation index 50%, sampling rate 16 kHz. To make the accuracy check as realistic as possible, the signal was not fed directly into the analysis algorithms but modelled, synthesised and recorded with the Audacity® signal processing tool, then analysed in the same way as the recorded authentic speech data. The charactistics of the signal may be thought of as modelling a stylised voice with a perfectly regular fundamental frequency of 200 Hz and a perfectly regular uninterrupted ‘syllable sonority’ pattern with mean syllable duration 200ms (mean syllable rate: 5 syll/s). The perceptual effect of the synthetic signal is that of a highly regular mechanical rhythm. The left-hand panel of Figure 4 shows the 6 s 200 Hz stylised synthetic AM signal with 5 Hz modulation as a double-edged ‘toothcomb’ pattern, which is demodulated by full-wave rectification (converting to absolute amplitude values) and low-pass filtering to recover the 5 Hz AM envelope modulation.

The centre panel of Figure 4 visualises the long-term LF spectrum of the entire 6 s long demodulated signal. The amplitude modulation frequency appears as a high magnitude spectral peak at 5 Hz, marked by a vertical dotted line topped with a star.

The right-hand panel of Figure 4 shows the long-term LF spectrogram representation of the stylised signal in a waterfall format. The spectrogram consists of a sequence of shorter term LF spectra, top to bottom, each spectrum generated by a 3 s moving FFT window (50 steps, 60 ms per step), with long overlaps of 2.94 s in order to show small gradual changes.

3 AM and FM demodulation and spectral analysis

3.1 English stress-pitch accent sequences

It has been a matter of debate to what extent low frequency AM and FM both influence the production and perception of rhythm, and relatively little quantitative analysis has been forthcoming (but cf. Cumming 2010; Varnet et al. 2017; Gibbon 2018, 2019; Gibbon and Li 2019; Ludusan and Wagner 2020). In order to extend this domain, the RFT methodology, having been applied to AM rhythm analysis, was generalised and applied to FM rhythm analysis. First, the case of English is discussed, then Mandarin Chinese.

The variable pitch accents of English, which are associated with abstract word and phrase stress locations, are referred to here as *stress-pitch accents*, to distinguish them from the lexical pitch accents of languages such as Japanese (Poser 1984; Hyman 2009) and from lexical tone. As a 'clear case', the rhythmical genre of counting aloud (1…7, British English, adult male, moderate tempo, recording length 6 s) is analysed.

In an English prenuclear stress-pitch accent sequence, the accents tend to share the same pitch pattern throughout the sequence. This resembles pitch patterns in list concatenations, but the repetitive property of pre-nuclear stress-pitch accents in sequence is more general, and has long been noted in pedagogical textbooks since Jones (1909, 1918) and Palmer (1924), with such sequences being termed 'head' or 'body' of an 'intonation group'. Dilley (1997: 87ff.) proposed an accent sequence similarity constraint for the head pattern, in order to explain such sequential pitch accent patterns as correlates of coherent grammatical patterns and as a means of entraining the attention of listeners to expect pattern changes such as nuclear tones. Abstracted away from the original context, this sequence of like patterns relates to the Obligatory Contour Principle (Leben 1973): an initial tone assignment spreads to following unspecified positions. Figure 5 shows the FM and AM analyses, aligned with AM in the top row and FM in the bottom row, in order to focus on parallels between the AM and FM domains.

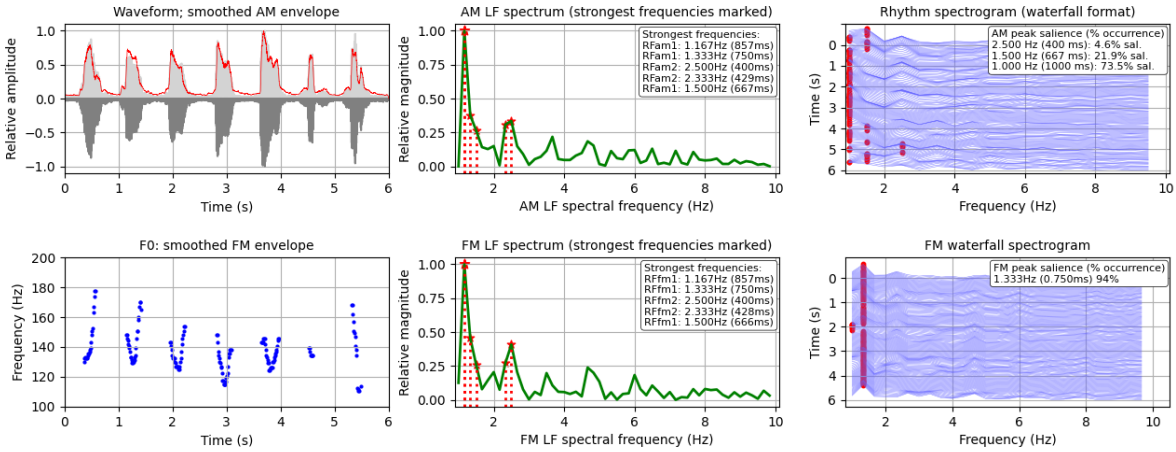


Figure 5: RFT analysis of counting one to seven (English, adult male); time is left-right in the left hand graphs, and top to bottom in the waterfall formats; frequency is left to right in the centre and right-hand graphs.

The English counting sequence evidently follows Dilley's accent sequence similarity constraint for head patterns, and the long-term LF spectrum shows a very close resemblance between the AM spectral pattern and the FM spectral pattern. The waterfall spectrogram format shows similar temporal dynamics for both AM and FM analyses – not particularly 'dynamic', of course, in this highly rhythmical example. The highest magnitude LF peaks in the rhythm formant trajectories, shown by a dot on each spectral peak, remain fairly constant in each component spectrum in the spectrogram.

The waveform and the F0 estimation (Figure 5, centre column) both show interesting local properties. They are clearly different in their local shapes: syllable sonority contours do not have the same shape as stress-pitch accent contours. But there are similarities:

1. Both contours are synchronised with the words in the sequence, as one would expect for monosyllables and for stress-pitch accents.
2. Both local syllable shapes and local FM shapes show a mainly binary pattern, in AM with a secondary peak on the coda (the final consonants, in the case of seven the syllabic [n]), and in FM with a falling and a rising F0 component.
3. The spectra and spectrograms for both AM and FM show frequencies corresponding to a word repetition rate of around 1.3 Hz (estimated duration mean of 769 ms) and also frequencies reflecting syllable components at around 2.5 Hz (400 ms) and 3.7 Hz (270 ms). The 5 highest magnitude spectral frequencies are indicated by dotted vertical lines topped by a star. In the spectrograms the highest magnitude frequency is marked by a single dot; other higher magnitude zones are not marked but appear as smaller 'waves'.

It might be suspected that these similarities are artefactual and that the spectrum is dominated by F0 effects, but this is not so: the similarities are mainly due to the regular sequence of similar pitch accents, to synchronisation of both domains with words and also to local binary similarities. The independence of the two parameters is investigated further in connection with the different F0 patterns of Mandarin Chinese.

3.2 Mandarin Chinese lexical tone sequences

In languages with lexically determined tones and pitch accents, the tonal sequences are in principle arbitrary, like other lexical properties. Consequently the pitch sequencing constraint does not apply, barring limited tonal sandhi effects. It may be predicted that the similarity effect of AM and FM spectral analyses may not be observed to the same extent in such languages, specifically in Mandarin Chinese.

Figure 6 shows the RFT analysis of counting from one to seven in Mandarin Chinese by a female adult, in Pinyin transcription “yī èr sān sì wǔ liù qī”, with high, fall, high, fall, fall-rise (dipping), fall and high lexical tones (i.e. tones 1, 4, 1, 4, 3, 4, 1).

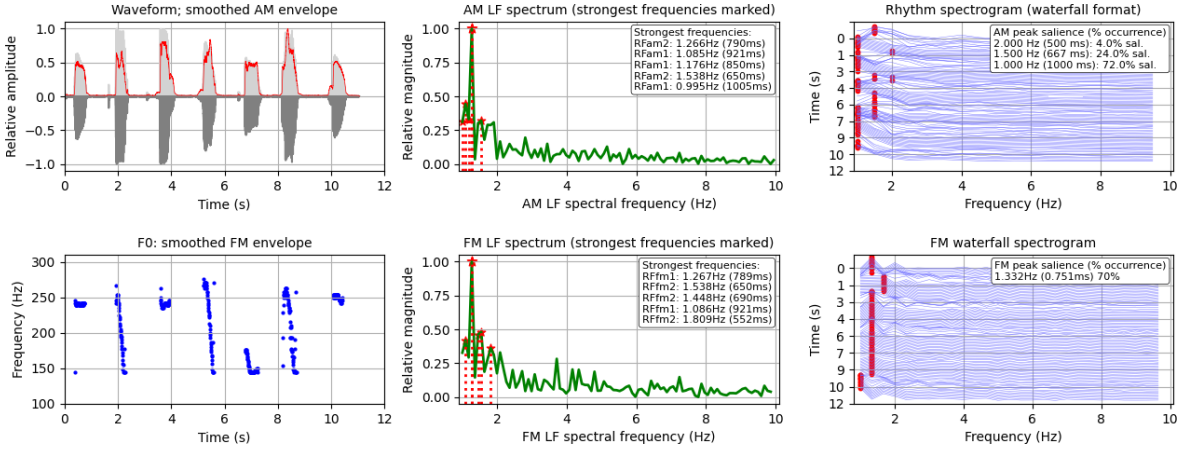


Figure 6: AM and FM rhythm analysis for Mandarin Chinese counting one to seven.

The most obvious difference between the FM envelopes of English and Mandarin Chinese is, as predicted, between the head constraint on local F0 shapes associated with stress-pitch accents in English on the one hand, and the lack of a head constraint on the lexical tones in Mandarin Chinese on the other (Figure 6, leftmost columns). There is a fortuitous repetition of a high (tone 1) and falling (tone 4) pair in the first part of the FM sequence, which no doubt contributes to the word rhythm, but the second part of the sequence is more irregular, with instances of the dipping tone (tone 3), the falling (tone 4) and the high (tone 1).

A corollary of the head constraint difference between English and Mandarin lies in the prediction of a discrepancy between the AM spectrum and the FM spectrum in Mandarin (centre columns in each case). In English the AM and FM spectra are strikingly similar, as expected from the head constraint. The Mandarin AM and FM spectra are somewhat less similar, being strikingly similar up to about 1.8 Hz (estimated mean unit duration 555 ms) for word-like properties, but showing differences in the frequency zone above 2 Hz and especially around 4 Hz (250 ms). These differences may be due to syllable-like properties, including local tonal frequency patterns. Tests to investigate the exact sources of these differences in relation to the tones, with less regular tone sequences than these, are a separate and complex task.

The conclusion is that Mandarin Chinese has a less homogeneous distribution of AM and FM spectral frequencies, both due to phonotactic variation and to the arbitrariness of lexical tone patterns, and consequently does not have a head pattern like in English AM and FM analyses. Exploration of the similarities and differences in the examples visualised in Figure 5 and Figure 6 tentatively allow a plausible prediction that the contribution of FM to rhythm varies with the prosodic typology of the language, in contrast to the finding of Varnet et al. (2017), who found no cross-linguistic differences in FM spectra. It will be interesting to see how this prediction works out with other other stress-pitch accent languages such as Dutch and German, and with other types of tone language such as the languages of the Niger-Congo group with terraced tonal sandhi. Additional extensive study is needed in order to establish the role of competing variables such as age, gender or emotionality in addition to tonal typology, as factors in any rhythmic differences which may be found.

4 Signal-symbol association in rhythmical speech

4.1 Morphophonology and rhythm: a hypothesis

In a linguistically motivated further step forward in the exploratory application of RFT, an effect of morphophonological patterning on long-term rhythmical timing was examined using fairly rapid counting from one to thirty in English, spoken by an adult male native speaker. The signal was also annotated on syllable and word tiers, with additional tiers containing formant information and interpretative commentary. An RFT spectral analysis is shown in Figure 7.

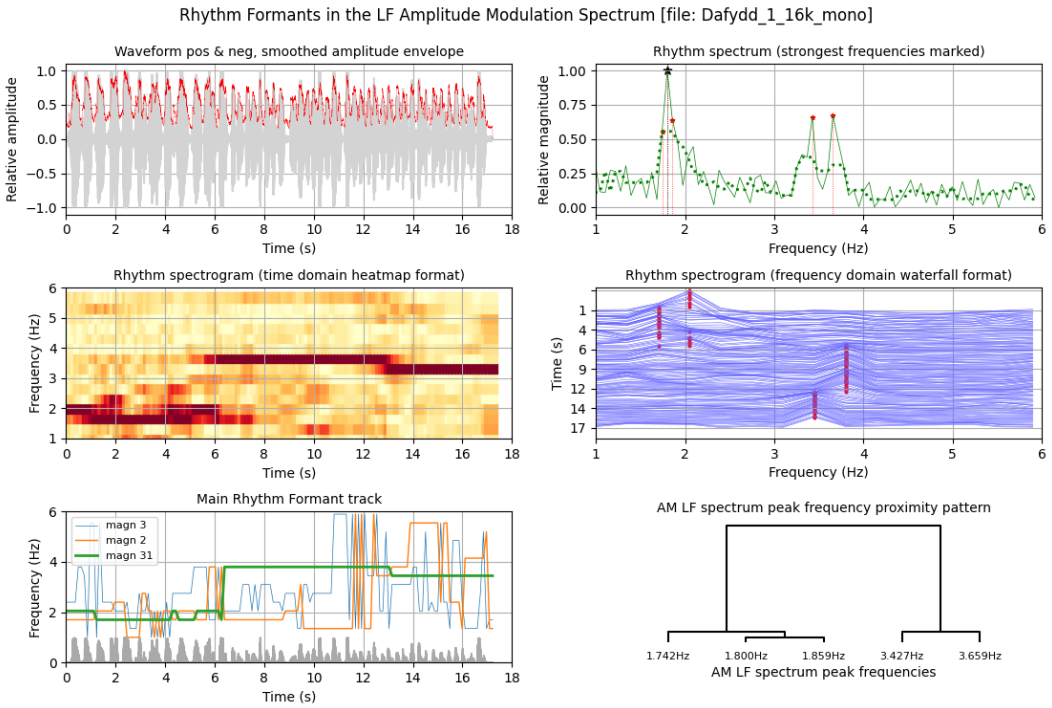


Figure 7: Counting from one to thirty, British English, adult male.

In addition to waveform, envelope, LF AM spectrum, LF AM waterfall and LF AM heatmap spectrograms, a rhythm formant classification dendrogram is shown (Figure 7, bottom right) as well as rhythm formant trajectories through the spectrogram for frequencies with the three highest magnitudes in each component spectrum of the spectrogram and a positive waveform track for location of the transition points (Figure 7 bottom left).

The LF AM spectrum shows two distinct high magnitude frequency zones, interpreted as rhythm formants. The spectrum itself contains no temporal information, which may wrongly be taken to imply that the formants are simultaneous. To fill this gap, the dynamic temporal development of rhythms is visualised in the AM waterfall spectrogram (centre row, right) and in the heatmap format (centre row, left). The spectrograms show that the rhythm formant frequencies overlap slightly, but essentially occur at different times, one up to 6 seconds, the other after 6 seconds, with another change at about 13.5 seconds. The lack of temporal variation information in the long-term rhythm spectrum itself is shared with the irregularity metrics which were discussed earlier, which also abstract away from temporal information. In contrast, the spectrogram visualises the missing information about the temporal distribution of rhythms and reveals a new dimension of information about rhythm.

An initial linguistic hypothesis about the structural factors which influence the interface with rhythm variation is not hard to find: the counting sequence contains morphophonologically distinct utterance segments which are co-extensive with the formant variants. The following subsection examines the temporal properties of these segments. The morphophonologically distinct segments are:

1. 1...10: mainly monomorphemic and monosyllabic words;
2. 11, 12: short transition, monomorphemic but a trisyllabic and a monosyllabic word;
3. 13...20: dimorphemic derivations which are disyllabic except for *seventeen*;
4. 21...30: trimorphemic compounds with initial dimorphemic disyllabic derivations, trisyllabic except for *twenty-seven* and *thirty*.

4.2 Testing the hypothesis

In order to test this initial informal hypothesis, the signal-symbol association was examined more closely by annotating the speech signal shown Figure 7 with time-stamped labels, using the Praat phonetic workbench (Boersma 2001), and analysed using the TGA online tool (Gibbon 2013, Yu and Gibbon 2015). Figure 9 visualises the annotation.

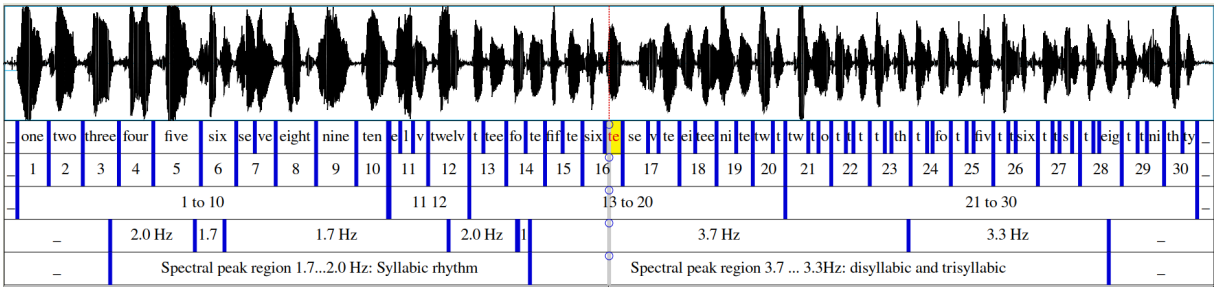


Figure 8: Annotation of the speech signal using the Praat phonetic workbench.

Syllables are annotated in the top tier, then words, then the morphologically characterised sequences one to ten, eleven to twenty and twenty-one to thirty. In the fourth tier from the top, frequencies in the trajectory of highest magnitude peaks are annotated, and the fifth tier contains comments. It was expected on the basis of the spectral analysis that the word or foot rate would be about 1.8 word/s, with an average word or foot duration of about 556 ms and that the syllable rate would be about 3.5 syll/s, with average syllable duration about 286 ms.

Table 1: Descriptive statistics for syllable and word time-stamps ('RF' = 'Rhythm Formant', durations in milliseconds).

	Syllables		*Words*	
n:	61		30	
Total sample count:	16474		16771	
min duration:	78		447	
max duration:	677		808	
Duration range:	599		361	
Duration mean:	270.07		559.03	
Mean syllable rate:	3.7	Approx. 3.7 Hz	1.79	Approx. 1.7 Hz
nPVI:	46	Irregular	12	Regular

Descriptive statistics including the nPVI were extracted from the annotations for both syllable and word tiers (cf. Table 1). The predictions based on the spectrogram are confirmed in the descriptive statistics of the annotation: the syllable rate in the rhythmical counting data is 3.7 syll/s, very close to the measured rhythm formant of 3.5 Hz, with average syllable duration of 270 ms, very close to the predicted 286 ms. The word or foot rate is 1.79 syll/s, close to the rhythm formant at 1.8 Hz, and correspondingly the mean word duration of 559 ms is close to the predicted mean word duration of 556 ms.

The nPVI irregularity metric yields values of 46 for syllables, corresponding to the expected range for English, and 12 for words, indicating much higher regularity for word sequences than for syllables (cf. also Asu and Nolan 2006), and is thus compatible with the RFT result. The RFT analysis provides additional detail about temporal patterning and its variation over time, but there is overall agreement between the two methods.

The informal grammar-rhythm alignment hypothesis is confirmed: the lower frequency rhythm formant trajectory is coextensive with the monomorphemic, predominantly monosyllabic locutionary sequence, the higher peak frequency trajectory is aligned with dimorphemic, mainly disyllabic items, and the lowering in the final third matches trimorphemic, mainly trisyllabic items. The three utterance segments with internally regular rhythms provide additional physical confirmation for the head pattern, in particular for Dilley's accent similarity constraint (1997) that sequences of similar items entrain the attention of the listener in expectation of an upcoming change in the pattern at phrase level, in contrast to the lexical tonal sequences of Mandarin Chinese. Essentially, this change corresponds in the counting example to the transition between the different dominant syllable patterns of the decades 1...10 (monosyllabic), 13...20 (disyllabic) and 21...30 (trisyllabic).

5 Rhythm formant trajectories in language classification

5.1 Classification of readings in two languages by a bilingual speaker

5.1.1 Rhythm formant vector extraction

An initial classification experiment was carried out in order to test the methodology in a different exploratory direction with English and German narrative data, recorded by the bilingual speaker. The instruction was to read the German translation first, three times, then the English version, also three times, in each case as if reading to a child, in order to achieve as close a possible an approximation to authentic data without a full natural scenario but with some control over situational variables. The rhythm formant analysis of these English readings is shown in Figure 9.

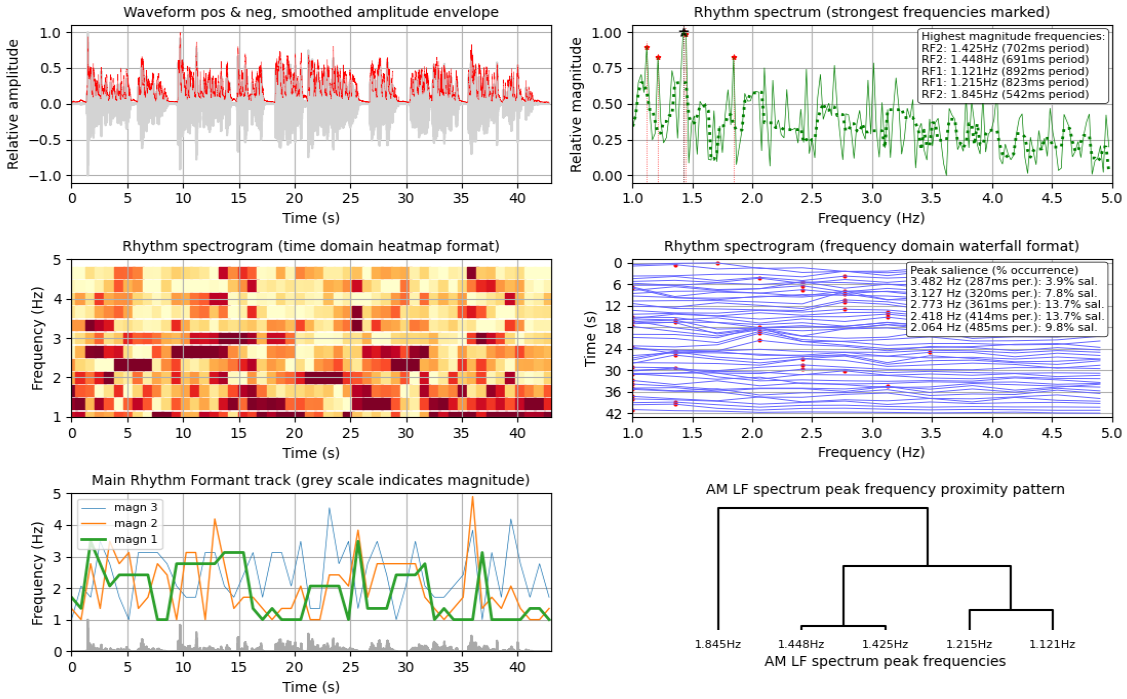


Figure 9: Reading of "The North Wind and the Sun" in English by a female adult bilingual.

The waveform (top left) shows an episodic structure with intervening pauses and the overall spectrum shows relatively clear peaks up to about 5 Hz, corresponding to interval durations of 200 ms and longer, but the main frequency zone is between 1.1 Hz and 1.4 Hz. One major peak at 1.3 Hz corresponds to intervals with mean duration around 745 ms, which are relatable to word, foot or short phrase units in the locution. Clear patterns which may be relatable to syllables are not so much in evidence.

The figures in the waterfall spectrogram panel (centre right) show the time pattern of the highest magnitude frequencies. The pattern shows considerable variety, except for a short relatively constant sequence at about 1.3 Hz near the 20 sec mark on the time axis. The heatmap spectrogram pattern shows intervals with higher spectral frequencies and intervals with lower spectral frequencies. The rhythm formant trajectory (bottom left) reflects this pattern for the three highest magnitude formant tracks, and shows the relatively constant lower frequency trajectory segment starting at at about 18 s and continuing until about 24 s along the time line of the reading. The lowest frequency trajectory is used in the following classification of the readings.

5.1.2 Hierarchical clustering of AM rhythm formant frequency vectors

Each rhythm formant trajectory is actually two-dimensional: it is based on two properties, *magnitude* and *frequency*, which are extracted as separate trajectories. For the present study, the frequency dimension alone was used. The working hypothesis for the basic unsupervised machine learning technique of distance-based hierarchical classification procedure is that the procedure partitions the data set cleanly into separate English and German readings.

Hierarchical classification was chosen over flat classifiers as the former are potentially more informative than the latter. Rather than following the conventional procedure of selecting a single standard distance metric and clustering criterion combination, a range of such combinations was chosen in order to explore the different properties of the combinations of six standard distance metrics and seven cluster similarity criteria, yielding $6 \times 7 = 42$ classification results:

1. Distance metrics (with alternative names): Manhattan (= Cityblock, Taxicab), Canberra (= Normalised Manhattan), Chebyshev (= Chessboard), Cosine, Euclidean, Pearson.
2. Cluster similarity criteria: linkage by average of cluster members (‘average linkage’), by weighted average, by nearest cluster member (‘single linkage’), by furthest cluster member (‘complete linkage’, Voorhees clustering), by cluster median, by cluster centroid and by minimal variance (Ward clustering).

The evaluation success criterion is full partitioning, with English readings clustered together and German readings clustered together.

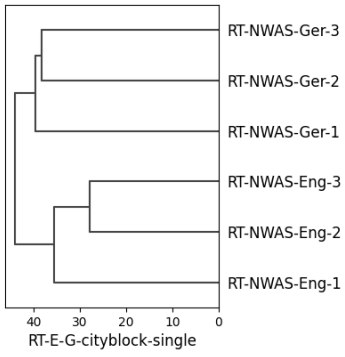
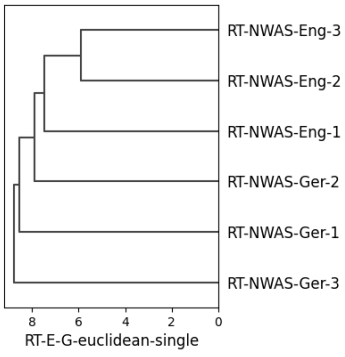
5.1.3 AM Rhythm formant frequency vector results

The votes for the resulting hierarchical clusterings of the rhythm formant frequency vector are shown in Table 2. Examples of the two resulting clear partitioning types are given in Table 3. Other partitioning types were not fully successful.

Table 2: Votes for distance metrics in terms of clustering criteria.

Symmetrical partitioning		*Right-branching scale*	
Metric	*Votes (out of 7)*	*Metric*	*Votes (out of 7)*
Canberra	7	Canberra	0
Chebyshev	2	Chebyshev	3
Cosine	1	Cosine	1
Euclidean	0	Euclidean	5
Manhattan	7	Manhattan	0
Pearson	2	Pearson	1
Total:	19	*Total*:	10

Table 3: Partitioning shapes distance metric votes.

Symmetrical (19 votes)	*Right-branching (14 votes)*
Example: 	*Example:* 

The results show a total of 29 votes (69%) for full English-German partitioning in contrast with 13 partitions (31%) with varying lower degrees of separation. The partitioning falls into two types:

1. *symmetrical clustering* (19/42 votes, 45%) into two main clusters, one for English and one for German, with further internal division;
2. *right-branching clustering* (10/42 votes, 24%), which is effectively flat clustering along a scale, with all English items together at one end of the scale and all German items together at the other end of the scale.

The most successful overall clustering criterion was the Ward minimal variance method, with three symmetrical clusterings (Canberra, Chebyshev and Manhattan) and one right-branching clustering (Euclidean) as well as one mixed symmetrical and right-branching

cluster (Pearson). With Cosine distance the partitioning was not successful. The Voorhees furthest distance method identified four symmetrical partitionings: three similar metrics (Canberra, Chebyshev, Manhattan), and the Pearson metric. The Pearson result is plausible here since a good correlation would be expected for English and for German separately.

The conclusion is that the English and German readings are distinct according to the majority vote for the rhythm formant frequency vector. The different results from different distance metrics are plausible: Manhattan and Canberra distance metrics, which are closely related, are more suitable for the irregular patterns of the present data, while the Euclidean 'as the crow flies' distance metric takes 'short cuts' through the patterns. Chebychev distance has intermediate properties. Pearson, unsurprisingly, does not show overall correlation, while Cosine shows similarity of orientation in the data space, i.e. direction or angle, not similarity of distance.

6 Rhythm formant trajectories and 'rhythms of rhythm'

6.1 Trajectories

The dendrograms discussed in the preceding section show a clear clustering result but they do not reveal the empirical criteria for arriving at the clustering. Closer examination of the trajectories of all the readings which were involved in the classification shows that the rhythmic properties of speech vary in the course of the narrative: in addition to the well-documented syllable, word and phrase rhythms, the rhythms themselves vary in higher ranking 'rhythms of rhythm', with which the narrator employs rhetorically motivated narrative rhythms to focus to greater or lesser extents on dramatising the story. This pattern was already observable in the Martin Luther King speech shown in Figure 1.

Figure 10 shows the AM spectral magnitude trajectories in the top two panels and the FM spectral frequency trajectories in the bottom two panels. The trajectories in each language (English in the top panel, German in the second panel from the top) are not randomly varying, but are evidently, without detailing correlation values, very similar, and show a consistent performance on the part of the speaker. While the English and German curves are each rather consistent, it is equally clear that the English and German curves are very different from each other. It is precisely these similarities and differences which explain the distance-clustering results show in the dendrograms shown in Table 3.

Figure 10: Trajectories through the rhythm spectrogram at highest magnitude frequencies.

Equally interesting are the AM spectral frequency trajectories, in the second panel from the bottom (English highest magnitude frequencies) and in the bottom panel (German highest magnitude frequencies). In each case, the spectral frequency curves match relatively well, but not as closely as the curves in the spectral magnitude graphs in the top two rows. The spectral frequency leaps tend to be more sudden and apparently more discrete than the magnitude changes, with relatively large leaps between frequencies. The determining factor may be the local syllable, word and phrase structure of the co-extensive locutions.

Again without detailing correlation values, it is also clear that the spectral frequency trajectories tend to correlate inversely with the spectral magnitude trajectories: highest magnitudes tend to relate to lowest frequencies, indicating that the lower frequencies between 0.5Hz and 1 Hz bear a heavier burden in conveying rhythmical interest than the higher frequencies.

Looking at the overall pattern of the magnitude curves, another property is also in evidence: a very slow oscillation, as in Figure 1, the second order rhythms of rhythm, with magnitude trajectories varying fairly regularly in time, initially at approximately 10 s intervals (10 s, 20 s, 30 s), at a frequency of about 0.08 Hz. The German pattern is somewhat different: 9 crests in the course of 50 s, about 5.6 s per wave, a frequency of 0.18 Hz.

The sociolinguistic or grammatical reasons for the rhythmic differences between English and German are potentially very diverse. They may be idiosyncratic or rhetorical, based on gender, age or genre, or they may derive from word order and translation choices such as centre-embedding versus right-branching, which can influence phrasing, tempo and intonation.

For example, the German translation has centre-embedding in "Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam." The English translation has right-branching, which is less complex to process and has different effects on phrasing, tempo and intonation (cf. Gibbon and Griffiths 2017): "The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak." But elsewhere the German translation also has right-branching, "Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen" where the English translation has centre-embedding: "They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other."

More detailed examination of the relation between the rhythm variation and its functionality is required than is possible in this context, and the role of FM in this context requires more investigation. However, this 'rhythms of rhythm' technique of examining the trajectory of the maximum spectral magnitude and frequency through time opens up a way of looking at discourse rhythms. The new information which this approach provides is necessarily impossible to attain with LF spectral analysis alone, which is atemporal, or with dispersion indices for annotated durations, which do not account for rhythmic oscillation. The present exploratory discussion of individual differences in performances of a bilingual in reading in her two high proficiency languages suggest that the RFT methodology is potentially relevant for detecting code-switching, for the diagnosis of L2 language learner fluency as well as for classifying pathological speech conditions.

6.2 Classification of language varieties: English and German

The next exploratory step concerns varieties of English (with Scottish English as the largest subgroup) and German (with Swiss German as the largest subgroup) in readings of *The North Wind and the Sun*, from the English and German readings in the Edinburgh corpus, with the addition of readings by the bilingual speaker discussed previously. It is predicted that readings will be partitioned into clusters which correspond to recognisable speaker groups.

First, clusterings for the German readings were calculated using the Manhattan Distance metric combined with the distance-based Voorhees clustering algorithm, which had previously been used successfully with the bilingual reader. The prediction is that speakers of different varieties of German would be partitioned in the hierarchical classification. Results are shown in Figure 11. The left panel shows only readings from the Edinburgh corpus, the right-hand panel shows results after adding the German readings by the bilingual speaker.

The prediction is that the male Swiss German speakers are clearly partitioned from the male Standard German speakers and that both are clearly partitioned from the female English-German bilingual speaker. The prediction is fulfilled (Figure 11), and the more general prediction that partitioning based on rhythm formant trajectories relates to clearly defined speaker groups is again fulfilled. This exploratory data selection is small, however, and in a subsequent confirmatory study stricter experimental conditions need to be applied.

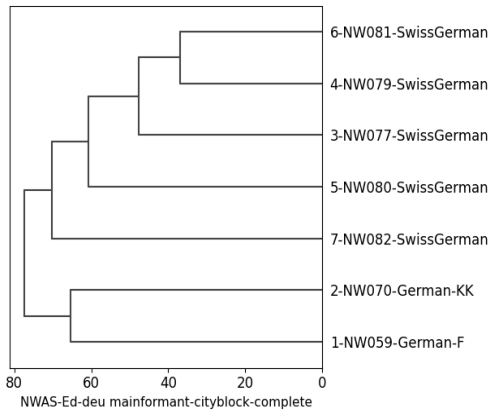


Figure 11: Rhythm Formant dendrogram for the German readings of "Nordwind und Sonne" (from the Edinburgh NWAS corpus and three additional readings).

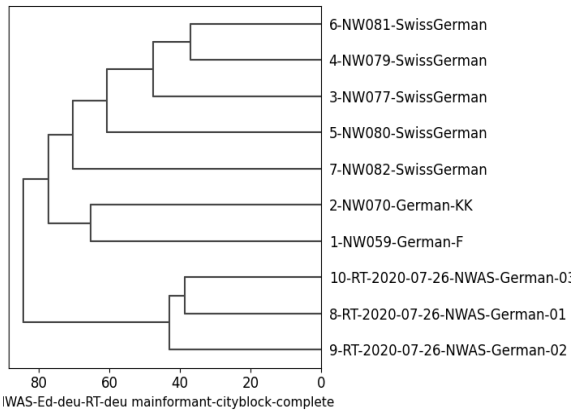


Figure 12: Rhythm Formant dendrogram for the German readings of "Nordwind und Sonne" (from the Edinburgh NWAS corpus and three additional readings).

The results for English are shown in Figure 13 for the readers from the Edinburgh corpus together with the English readings by the German-English bilingual speaker.

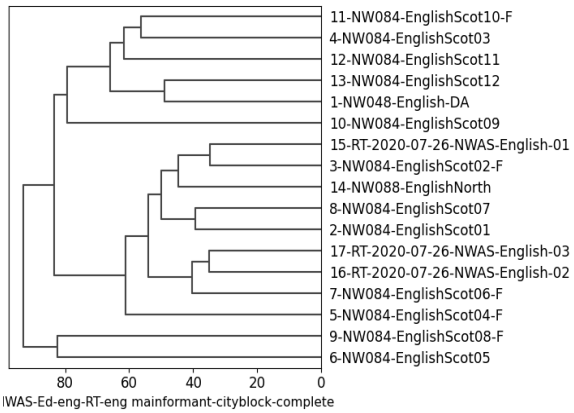


Figure 13: Rhythm Formant dendrograms for the English readings of "The North Wind and the Sun" from the Edinburgh NWAS corpus and three additional English readings by a German-English bilingual speaker. Left: Manhattan-Voorhees distance-clustering, right Canberra-Voorhees distance-clustering.

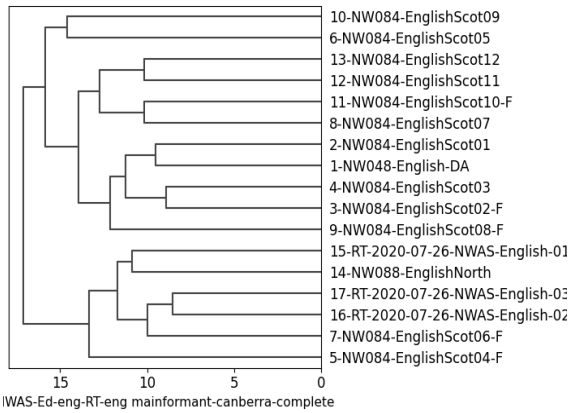


Figure 14: Rhythm Formant dendrograms for the English readings of "The North Wind and the Sun" from the Edinburgh NWAS corpus and three additional English readings by a German-English bilingual speaker. Left: Manhattan-Voorhees distance-clustering, right Canberra-Voorhees distance-clustering.

The results from the Manhattan-Voorhees distance-clustering of the English data are shown in Figure 13. The results are only partly interpretable in terms of the intuitively given reader groups: the largest subclusters are of Scottish English, as expected. The Standard English reader DA clusters with a Scottish group, which is perhaps not too unexpected as he lived and worked for three decades in Scotland. The two English readings by the bilingual speaker (Figure 14), which previously clustered together, are also clustered together, though here together with with a female Scottish reader.

The Canberra Distance (Normalised Manhattan Distance) analysis combined with Voorhees clustering produces a result which corresponds more closely to sociolinguistic predictions: all three readings by the bilingual

The Canberra Distance (Normalised Manhattan Distance) analysis combined withspeaker are in the same third rank cluster: the two most similar readings cluster together, then with a Scottish English female. The outlier reading clusters together with the Northern English male. In the Canberra-Voorhees comparison the Scottish English readings cluster most closely together. Further discourse analytic and sociolinguistic interpretations may be possible but are not the subject of the present investigation.

6.3 Size and inhomogeneity of the data

The number of linguistic and sociolinguistic variables involved in the analysis of discourse prosody is high, and an analysis of the entire Edinburgh database showed that results become

less uninterpretable as the size and inhomogeneity of the data set increase (cf. Figure 15), though there are plausible sub-clusters of related languages.

The readings in the Edinburgh data are opportunistic *données trouvées* and not purpose-designed, apart from the corpus collation goal, which is one reason not to expect rigorous, scalable and broadly generalisable results. The corpus is an excellent resource for substantive contributions to exploratory research and results suggest that the RFT method is likely to be useful for smaller data sets. For larger data sets one or more of these conditions need to be fulfilled:

1. a more closely controlled corpus with well-defined language varieties, genres and styles as well as speaker characteristics;
2. a more complex rhythm formant input vector in order to handle the homogeneity;
3. upscaling of hardware infrastructure to 'big data' calibre in order to handle non-linear growth of time and memory requirements relative to data size and processing complexity.

Nevertheless, the results on more homogeneous data sets indicate that the methodology of exploratory case studies with rhythm formant trajectory analysis is a fruitful starting point for quantitative studies of discourse rhythms. The aim is to proceed beyond the exploratory stage of the methodology to pursue detailed and well-defined confirmatory studies of the large number of variables involved in narrative data of the kind studied here, in which rhetorical strategy, information structure, grammatical structure, word patterns all play a role.

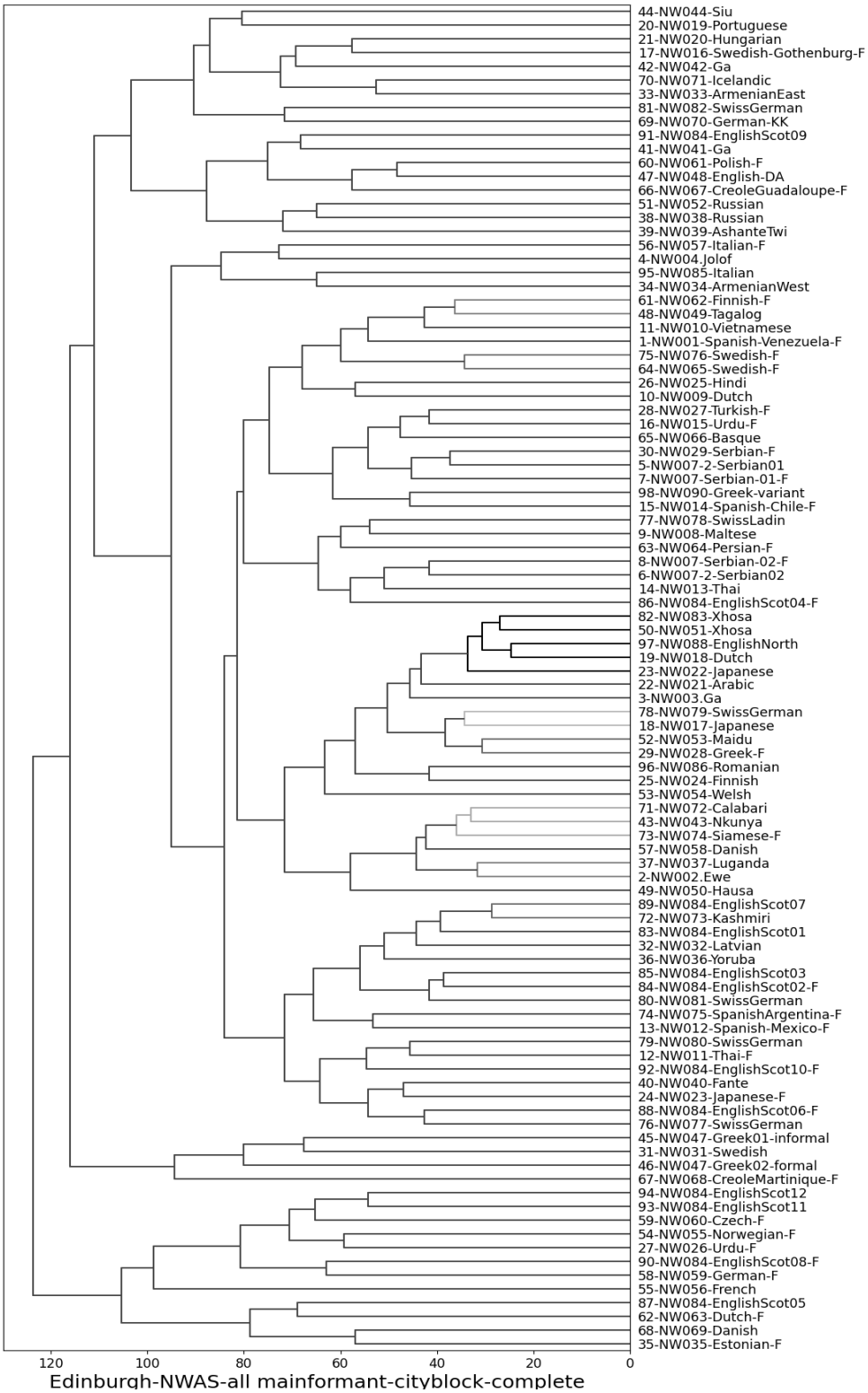


Figure 15: Rhythm trajectory dendrogram for all Edinburgh NWAS corpus readings, N=98.

7 Summary, conclusion and outlook

A platform of diverse methods for low frequency spectral analysis of rhythmic properties of utterances was established by several scholars over the past thirty years or so, and taken as a starting point for a framework, Rhythm Formant Theory (RFT), based on modulation theory and an associated signal processing methodology, Rhythm Formant Analysis (RFA).[5] A number of new concepts were derived from this methodology and used in several exploratory investigations not only of amplitude modulation but also of frequency modulation of the speech signal. The study is not restricted to statistical prosodic typology but provides pointers to interface issues between phonetic prosody and phonological prosody as a step towards providing language prosody with a previously lacking physical empirical grounding in speech prosody, and vice versa.

The central concept of RF is the rhythm formant as a generalisation over frequency zones associated with magnitude peaks in the long-term low frequency rhythm spectrum of speech utterances. The long-term LF spectrum contains no temporal information and therefore cannot help with questions concerning the dynamics of rhythm variability, so the long-term LF spectrogram was introduced to handle this sub-field, using the rhythm formant trajectory, a time function derived from the spectrogram. The dynamic association of different rhythms with morphophonological structures and their temporal alignment was examined using the rhythm formant trajectory: in this exploratory study using a recording of English counting from one to thirty it was tentatively found that different morphophonological patterns in the segments 1-10, 11-20 and 21-30 align with different rhythms. It was also shown that there are differences between rhythms of counting patterns in English, to which Dilley's accent sequence constraint on the heads of stress-pitch language rhythm groups applies, based on pitch similarity, and Mandarin Chinese, to which the constraint does not apply because of the phonemic arbitrariness of lexical tone.

It was also shown that in small data sets of narratives, readings by different readers in different languages can be plausibly classified and that the empirical basis for these classifications can be shown in detail by examining both the magnitude and frequency trajectories derived from the rhythm spectrogram. The current limits of the methodology were illustrated in a classification of the entire Edinburgh fable database.

5 The code and documentation is located on the GitHub developer portal: https://github.com/dafyddg/RFA

The conclusion is drawn that, in order to study the dynamics of rhythm variation, a non-trivial extension of the spectral analysis approaches to rhythm analysis in the form of RFT opens up fruitful avenues of research, particularly in hierarchically discriminating rhythm types as much as in identifying them, focusing on induction from individual discourse tokens, rather than on entire languages with the accompanying problems of overgeneralisation. The exploratory studies described in the present contribution prepare for more extensive confirmatory studies of specific issues such as the identification of specific left-headed or right-headed microrhythms at syllable and foot durations (Leong et al. 2014) and the relation between very low frequency rhythms and the micro-rhythmic relations between neighbouring syllables which has been the subject of much prior phonetic research (cf. Section 2).

A wide range of open issues in the study of speech rhythm remains, for the solution of which the conceptual instrument of RFT may be suitable. One practical issue is the lack of well-defined and available data and tools which to enable results to be reproduced. The present study has attempted to ameliorate this situation, first by using the readily available Edinburgh *The North Wind and the Sun* corpus, and, second, by making the RFT software tool available for open access in the public domain. Another issue which affects the study of rhythm is the compartmentalisation of methods and lack of interchange between different methodological approaches. The present study attempts to surmount this knowledge plateau by incorporating three complementary methodologies: (a) an explicit modulation theoretic account of the bottom-up spectral analysis approach to rhythm modelling, (b) top-down approaches in which the speech signal is aligned with prior defined linguistic categories, and (c) unsupervised machine learning as used in dialectometry and stylometry but with the distance-clustering method applied in RFT not to lexicons and texts but directly to the speech signal.

The immediate issue to be resolved is the scaling up of hierarchical clustering in the RFA methodology to handle larger and more heterogeneous datasets by including more complex rhythm formant input vectors, but also, non-trivially, by further algorithm optimisation and by scaling up the available computational infrastructure to handle the non-linear growth of time and memory requirements of more complex vectors.

There are also more general issues to solve, for example multimodal issues of speech and gesture rhythm, epistemological questions about whether all temporal regularities in the speech signal are to be interpreted as rhythm, to what extent speech rhythms are holistic

epiphenomena, and to what extent speech rhythms are an exact compositional function of physical, structural and semiotic factors in spoken language communication. Issues such as abstract grammatical and poetic metre and physical speech and poetic performance have been addressed in great detail in linguistics and in literary studies, but so far not with methods comparable with those presented in this study.

In summary, the present study provides a more detailed approach to the empirical grounding of rhythm than was previously available, but is still in need of detailed quantitative confirmatory studies of rhetorical choice, speaker idiosyncrasy and rhythmic variation in different genres and across languages and language varieties.

8 Acknowledgments

For careful reading and for detailed and constructive comments and suggestions I am indebted to two JIPA referees. Many discussions on the topic with my two 'prosody brothers' Daniel Hirst and Nick Campbell, and with Jolanta Bachan, Doris Bleiching, Grażyna Demenko, Laura Dilley, Katarzyna Dziubalska-Kołaczyk, Alexandra Gibbon, Katarzyna Klessa, Peng Li, Xuewei Lin, Huangmei Liu, Petra Wagner, Ratree Wayland, and Jue Yu have helped to sharpen the ideas. Above all, I owe the challenge of following up this line of thinking to 40 years of detailed discussions with the late Wiktor Jassem, starting with Jassem and Gibbon (1980).

9 References

Abercrombie, David. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Abercrombie, David. 2013. *The North Wind and the Sun, 1951-1978* [sound]. University of Edinburgh. School of Philosophy, Psychology, and Language Sciences. Department of Linguistics and English Language. https://doi.org/10.7488/ds/157.

Arvaniti, Amalia. 2009. Rhythm, Timing and the Timing of Rhythm. *Phonetica* 66 (1–2): 46–63.

Asu, Eva-Liina and Francis Nolan. 2006. Estonian and English rhythm: a twodimensional quantification based on syllables and feet. In: *Proc. Speech Prosody* 2006.

Barbosa, Plinio A. 2002. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. In: *Proc. Speech Prosody* 1, 163–166.

Barry, W.J., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. 2003. Do rhythm measures tell us anything about language type? In D. Recasens, M.J.Solé, and J. Romero, editors, *Proc. of the 15th International Congress of Phonetic Sciences* (ICPhS 2003), Barcelona, 2693–2696.

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

Brazil, David, Michael Coulthard and Catherine Johns. 1980. *Discourse Intonation and Language Teaching*. London: Longman.
Brown, Steven, Peter Q. Pfordresher and Ivan Chow. 2017. A musical model of speech Rhythm. *Psychomusicology: Music, Mind, and Brain* 2017, Vol. 27, No. 2, 95–112
Campbell, W. Nicholas. 1992. Multi-level Speech Timing Control. Ph.D. dissertation, U. Sussex.
Carbonell KM, Lester RA, Story BH, Lotto AJ. 2015. Discriminating simulated vocal tremor source using amplitude modulation spectra. *Journal of Voice : Official Journal of the Voice Foundation*. 29: 140-7
Chomsky, Noam, Morris Halle and Fred Lukoff. 1956. On accent and juncture in English. In: Halle, Morris, Horace G. Lunt, Hugh McLean and Conrnelis H. van Schooneveld, eds. 1956. *For Roman Jakobson. Essays on the Occasion of his sixtieth Birthday*. The Hague: Mouton & Co.
Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.
Couper-Kuhlen, Elizabeth and Peter Auer. 1991. On the contextualising function of speech rhythm in conversation: question-answer sequences. In: Verschueren, Jef, ed. *Levels of Linguistic Adaptation. Selected papers of the International Pragmatics Conference*. Antwerp, 1987, 1–18.
Couper-Kuhlen, Elizabeth and Margret Selting. 2018. *Interactional Linguistics. Studying Language in Social Interaction*. Cambridge: Cambridge University Press.
Cowell, Henry. 1930. *New Musical Resources*. New York: Alfred A. Knopf Inc.
Cumming, Ruth E. 2010. Speech Rhythm: The language-specific integration of pitch and duration. Ph.D. thesis, U Cambridge.
Cummins, Fred and Robert Port. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26, 145–171.
Cummins, Fred, Felix Gers and Jürgen Schmidhuber. 1999. Language identification from prosody without explicit features. In: *Proc. Eurospeech*, 371–374.
Dauer, Rebecca M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11:51–62.
Dellwo, Volker. 2010. Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. Dissertation, Universität Bonn.
Dihingia, Leena and Priyankoo Sarmah. 2020. Rhythm and Speaking Rate in Assamese Varieties. *Speech Prosody 10*, Tokyo, Japan, 561-565.
Dilley, Laura C. 1997. The Phonetics and Phonology of Tonal Systems. Dissertation MIT.
Dogil, Grzegorz and Gunter Braun. 1988. *The PIVOT Model of Speech Parsing*. Verlag der Österreichischen Akademie der Wissenschaften, Wien.
Foote, Jonathan and Shingo Uchihashi. 2001. The beat spectrum: a new approach to rhythm analysis. In: *Proc. IEEE International Conference on Multimedia and Expo*.
Fuchs, Robert and Eva-Maria Wunder. 2015. A sonority-based account of speech rhythm in Chinese learners of English. In: Gut, Ulrike, Robert Fuchs, Eva-Maria Wunder, eds. *Universal or Diverse Paths to English Phonology? Bridging the Gap between Research on Phonological Acquisition of English as a Second, Third or Foreign Language*. Berlin: de Gruyter, 165–183.
Galves, Antonio, Jesus Garcia, Denise Duarte & Charlotte Galves. 2002. Sonority as a basis for rhythmic class discrimination. In: Bernard Bel & Isabel Marlien, eds. *Proc. Speech Prosody 2002*, Aix-en-Provence: Laboratoire Parole et Langage, 323–326.

Gibbon, Dafydd. 1976. *Perspectives of Intonation Analysis*. Bern: Lang.
Gibbon, Dafydd. 1987. Finite State Processing of Tone Systems. In: *Proc. European Chapter of the Association for Computational Linguistics*, 291–297.
Gibbon, Dafydd. 2003. Computational modelling of rhythm as alternation, iteration and hierarchy. In: *Proc. 15th International Congress of Phonetic Sciences*, Barcelona, 2489–2492.
Gibbon, Dafydd. 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, eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 281–209.
Gibbon, Dafydd 2013. TGA: a web tool for Time Group Analysis. In: Daniel Hirst and Brigitte Bigi, eds. (2013). *Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, 66- 69.
Gibbon, Dafydd. 2018. The Future of Prosody: It's about Time. Keynote. In: *Proc. Speech Prosody 9*. https://www.isca-speech.org/archive/SpeechProsody_2018/pdfs/_Inv-1.pdf
Gibbon, Dafydd. 2019. CRAFT: A Multifunction Online Platform for Speech Prosody Visualisation. *Proc. 19th International Congress of Phonetic Sciences*, Melbourne.
Gibbon, Dafydd and Sascha Griffiths. 2017. Multilinear Grammar: Ranks and Interpretations. *Open Linguistics* 3 (1), 265–307.
Gibbon, Dafydd and Peng Li. 2019. Quantifying and Correlating Rhythm Formants in Speech. In: *Proc. 3rd International Symposium on Linguistic Patterns in Spontaneous Speech*. Academia Sinica. Taipei, Taiwan.
Gibbon, Dafydd and Xuewei Lin (2020). Rhythm Zone Theory: Speech Rhythms are Physical after all. In: Magdalena Wrembel, Agnieszka Kiełkiewicz-Janowiak and Piotr Gąsiorowski, eds. *Approaches to the Study of Sound Structure and Speech. Interdisciplinary Work in Honour of Katarzyna Dziubalska-Kołaczyk*. London: Routledge.
Gut, Ulrike 2012. Rhythm in L2 Speech. In: Gibbon, Dafydd, Daniel Hirst and Nick Campbell, eds. *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*. Special Edition of Speech and Language Technology 14/15. Poznań: Polish Phonetics Society. pp. 83–94.
He, Lei and Volker Dellwo. 2016. A Praat-Based Algorithm to Extract the Amplitude Envelope and Temporal Fine Structure Using the Hilbert Transform. In: *Proc. Interspeech*, San Francisco, 530–534.
Heřmanský, Hynek. 2010. History of modulation spectrum in ASR. In: *Proc. International Conference on Acoustics, Speech and Signal Processing* (ICASSP).
Hyman, Larry M. 2009. How (not) to do phonological typology: the case of pitch-accent. *Language Sciences* 31, 213–238.
Inden B, Malisz Zofia, Petra Wagner, Ipke Wachsmuth. 2012. Rapid entrainment to spontaneous speech: A comparison of oscillator models. In: *Proc. 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 1721–1726.
Ioannides, Andreas A. and Armen Sargasyan. 2012. Rhythmogram as a tool for continuous electrographic data analysis. In: *Proc. of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*., 205–211.
Jansche, Martin (1998). A two-level take on Tianjin Tone. In *Proc. 10th European Summer School in Logic, Language and Information, Student Session*. Saarbrücken, 162-174.

Jassem, Wiktor 1949. indikeiʃn əv spi:tʃ riðm in ðə tra:nskripʃn əv edjukeitid sʌðən ingliʃ. (Indication of speech rhythm in the transcription of Educated Southern English.) In *Le Maitre Phonétique* [The Phonetics Teacher] III/92, 22–24.

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

Jassem, Wiktor, David R. Hill and Ian H. Witten. 1984. Isochrony in English Speech: Its Statistical validity and linguistic relevance. In: Dafydd Gibbon and Helmut Richter eds. *Intonation, Accent and Rhythm. Studies in Discourse Phonology*. Berlin: Walther de Gruyter, pp. 203–225.

Jassem, Wiktor and Dafydd Gibbon. 1980. Re-defining English stress. *Journal of the International Phonetic Association* 10, 1980:2-16.

Jones, Daniel. 1909. *The Pronunciation of English*. Cambridge: Cambridge University Press.

Jones, Daniel. 1918. *An Outline of English Phonetics*. Cambridge: Heffer and Sons.

Kallio, Heini, Antti Suni, Juraj Šimko, Martti Vainio. 2020. Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics* 80, 1–12.

Kohler, Klaus. 2009. Editorial: Whither Speech Rhythm Research? *Phonetica* 66: 5–14.

Krause, M. 1984. Recent developments in speech signal pitch extraction. In: Dafydd Gibbon and Helmut Richter eds. *Intonation, Accent and Rhythm. Studies in Discourse Phonology*. Berlin: Walther de Gruyter, 243–252.

Lee, Christopher S. and Neil P. McAngus Todd. 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): 225–54.

Leben, William (1973), *Suprasegmental Phonology*. PhD dissertation, MIT.

LeGendre, Susan J., Julie .M. Liss, Andrew J. Lotto and Rene Utianski. 2009. "Talker recognition using envelope modulation spectra." *Proc. Acoustical Society of America*, San Antonio, TX.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, Mass.: MIT Press.

Leong, Victoria, Michael A. Stone, Richard E. Turner, and Usha Goswami. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. In: *Journal of the Acoustical Society of America*, 366–381.

Liberman, Mark Y. and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249–336.

Liss, Julie M., Sue LeGendre, and Andrew J. Lotto. 2010. Discriminating Dysarthria Type From Envelope Modulation Spectra. *Journal of Speech, Language and Hearing Research* 53 (5):1246–1255.

Low, Ee Ling, Esther Grabe and Francis Nolan. 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43:4, 377-401.

Ludusan, Bogdan, Antonio Origlia and Francesco Cutugno. 2011. On the use of the rhythmogram for automatic syllabic prominence detection. In: *Proc. Interspeech*, 2413–2416.

Ludusan, Bogdan and Petra Wagner. 2020. Speech, laughter and everything in between: A modulation spectrum-based analysis. In: *Proc. Speech Prosody 10*, 25–28 May 2020, Tokyo, Japan, 995–999.

Malisz, Zofia, Petra Wagner. 2012. Acoustic-Phonetic realisation of Polish syllable prominence: a corpus study of spontaneous speech. In: Gibbon, Dafydd, Daniel Hirst and Nick Campbell, eds. *Rhythm, Melody and Harmony in Speech. Studies in Honour*

of Wiktor Jassem. Special Edition of Speech and Language Technology 14/15. Poznań: Polish Phonetics Society, 105-114.
Malisz, Zofia, Michael O'Dell, Tommi Nieminen, and Petra Wagner. 2016. Perspectives on speech timing: coupled oscillator modeling of Polish and Finnish. *Phonetica*, 73 (3–4), 229–255.
Nolan, Francis and Hae-Sung Jeon. 2014. Speech rhythm: a metaphor? Theme Issue: Smith, Rachel, Tamara Rathcke, Fred Cummins, Katie Overy and Sophie Scott, eds. Communicative Rhythms in Brain and Behaviour. *Philosophical Transactions of the Royal Society B. Biological Sciences*, 1–11.
O'Dell, Michael L. and Tommi Nieminen. 1999. Coupled Oscillator Model of Speech Rhythm. In: *Proc. XIVth International Congress of Phonetic Sciences*. San Francisco, 1075–1078.
Palmer, Harold E. 1924. *English Intonation: With Systematic Exercises*. Cambridge: Heffer.
Pierrehumbert, Janet B. 1980. The Phonology and Phonetics of English Intonation. Ph.D. dissertation, MIT.
Poser, William J. 1984. Phonetics and Phonology of tone in Japanese. Ph.D. dissertation, MIT.
Prince, Alan and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden: Blackwell Publishers.
Ramus, Franck, Marina Nespor and Jaques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292.
Roach, Peter. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In: Crystal, David, ed. *Linguistic Controversies: Essays in Linguistic Theory and Practice*. London: Edward Arnold, 73–79.
Sagisaka, Yoshinori. 2003. Modeling and perception of temporal characteristics in speech. In: *Proc. 15th International Congress of Phonetic Sciences*, Barcelona.
Selkirk, Elizabeth. 1984. *Phonology and syntax The relation between sound and structure*. Cambridge, MA: The MIT Press.
Scott, Donia R., Stephen D. Isard, and Bénédicte de Boysson-Bardies. 1985. Perceptual isochrony in English and French. *Journal of Phonetics*, 13:155–162.
Suni, Antti, Juraj Šimko, Daniel Aalto, Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language* 45: 123–136.
Sweet, Henry. 1908. *The Sounds of English*. Oxford: Clarendon Press.
Tilsen Samuel and Keith Johnson. 2008. Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*. 124 (2): EL34–EL39. 2008. [PubMed: 18681499].
Tilsen, Samuel and Amalia Arvaniti. 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, 628–639.
Todd, Neil P. McAngus and Guy J. Brown. 1994. A computational model of prosody perception. In: *Proc. the International Conference on Spoken Language Processing* (ICLSP-94), 127–130.
Varnet, Léo, Maria Clemencia Ortiz-Barajas, Ramón Guevara Erra, Judit Gervain, and Christian Lorenzi. 2017. A cross-linguistic study of speech modulation spectra. *Journal of the Acoustical Society of America* 142 (4), 1976–1989.
Wagner, Petra. 2007. Visualizing levels of rhythmic organisation. In: *Proc. 16th International Congress of Phonetic Sciences*, Saarbrücken 2007, 1113–1116.

Wang, Avery Li-Chun. 2003. An Industrial-Strength Audio Search Algorithm. In: *Proc. International Symposium on Music Information Retrieval* (ISMIR), Baltimore, MD.
Wayland, Ratree and Takeshi Nozawa. 2020. Calibrating rhythms in L1 Japanese and Japanese accented English. In: *Proc. of Meetings on Acoustics* 178 ASA 39 (1), 1–13.
White, Laurence and Zofia Malisz. 2020. Speech Rhythm and Timing. In: Gussenhoven, Carlos and Aoju Chen, eds. *The Oxford Handbook of Language Prosody*. Oxford: Oxford University Press.
Yu, Jue and Dafydd Gibbon. 2015. How natural is Chinese L2 English Prosody? In: *Proc. 18th International Congress of Phonetic Sciences*, Glasgow, 145–149.