The rhythms of rhythm

Dafydd Gibbon
Universität Bielefeld, Bielefeld, Germany & Jinan University, Guangzhou, China
gibbon@uni-bielefeld.de

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, with the aim of developing a unified theory of speech rhythms. Current methodologies in the field are first discussed, then the choice of data is motivated and the MODULATION-THEORETIC rhythm spectrum and rhythm spectrogram approach is applied to the 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 English with Mandarin Chinese. The LF rhythm spectrogram is introduced in order to recover temporal information about long-term rhythms, and to investigate the configurative function of rhythm. The trajectory of highest magnitude frequencies through the component spectra of the LF spectrogram is extracted and applied in classifying readings in different languages and individual speaking styles using distance-based hierarchical clustering, 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 RFT 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 since the development of electronic and computational analysis. 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 physics, psychology, medicine, linguistics and phonetics or musicology are concerned, from rhetorical rhythms to prominence relations between syllables. Each approach has addressed the problem of the empirical grounding of the elusive concept of rhythm in different ways, providing different pieces of the overall puzzle, metaphorically: *Language as Particle, Wave and Field* (Pike 1959).

The various different concepts of rhythm in the literature depend partly on the data selected, but mainly on the intradisciplinary and transdisciplinary methods used and the

Journal of the International Phonetic Association (2023) 53/1

doi:10.1017/S0025100321000086 First published online 16 August 2021

questions asked, for example whether rhythms are determined by discourse, by grammar or by constraints of production and perception, whether the mora, the syllable or the foot is the unit counted, whether rhythms are abstract, physical or both, whether physical rhythms are beats or waves, whether rhythm intervals are equally timed (isochronous), what the frequencies of different speech rhythms are, or how speech rhythms relate to neural patterns. A brief overview of five linguistic and phonetic paradigms of rhythm description and analysis is provided in Section 1.3.

The methodology proposed in the present study starts with current MODULATION-THEORETIC 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. The general objective is to provide a unified approach to describing and comparing rhythms in different prosodic domains. This understanding of speech rhythms is close to the common-sense understanding of rhythm as regularly occurring waves and beats. The complexities of the multiple rhythms of natural speech are, however, better understood as regular oscillations with specifiable low frequencies below about 10 Hz, deriving from neural patterns of resonance which drive the phonatory and articulatory muscles, and depending on physiological constraints, on the lexical and grammatical typology of the language, on rhetorical style and on idiosyncratic properties of speakers.

In contrast to many qualitative studies of rhythm description in the TIME DOMAIN of speech, which are sceptical about finding physical correlates, modulation-theoretic approaches since Todd & Brown (1994), Traunmüller (1994) and Cummins, Gers & Schmidhuber (1999) take a more optimistic position and use bottom–up signal processing approaches in the FREQUENCY DOMAIN. On a meta-level, qualitative choices of formal procedures are also made in these approaches, of course. In the bottom–up modulation-theoretic approaches, rhythms are not primarily conceived as regular durations in the time domain, but rather as spectral properties in the frequency domain with spectral analysis of the amplitude envelope of speech, generally as regular oscillations in given frequency zones, identifiable as magnitude peaks in the low frequency (LF) spectrum of the speech signal. ISOCHRONY (equal timing of intervals) is then a secondary *a fortiori* consequence of oscillation, not a primary category. Qualitative and quantitative time-domain approaches and quantitative frequency domain approaches each have their justification in respect of different properties of rhythms (Kohler 2009).

The following subsections provide a brief overview of a new approach within the modulation-theoretic paradigm, RHYTHM FORMANT THEORY (RFT) and its associated methodology RHYTHM FORMANT ANALYSIS (RFA), and a detailed account of relevant previous studies of speech rhythms. Section 2 describes the data and methods used in the present study. Section 3 addresses the relation of amplitude modulation (AM) and frequency modulation (FM) in speech rhythm with reference to British English and Mandarin Chinese. Section 4 examines signal–symbol association in the morphophonological domain using both annotation-based and direct 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 technique of distance-based clustering 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, and concluding that for long utterances the method relates more to speaking style than to language typology. In Section 7, the results are summarised, conclusions are drawn, and prospects for further development are outlined. The intent of the study is to explore the potential of RFT and RFA in developing an integrated framework for rhythm analysis at this stage, rather than to deploy *ad hoc* method combinations in a more conservative confirmatory approach.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

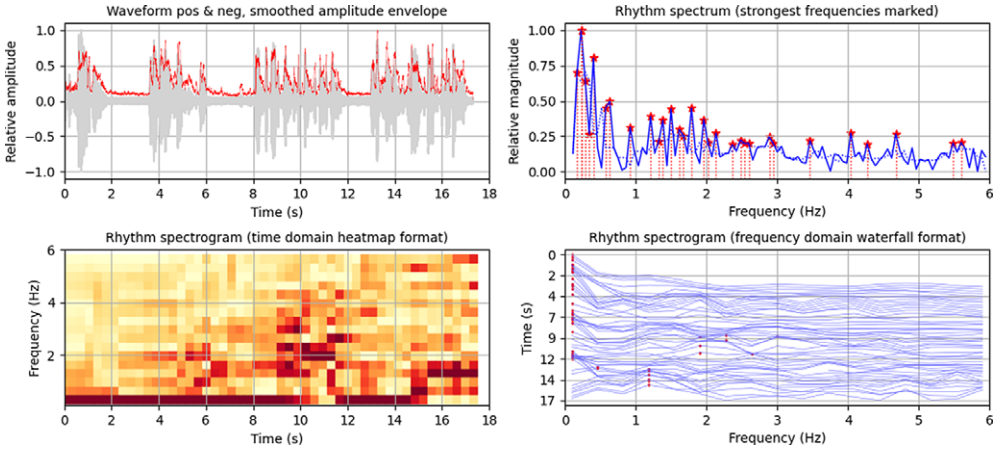

Figure 1 (Colour online) RFT AM analysis of a segment of Martin Luther King's 1963 speech: 'I have a dream that one day on the red hills of Jordan the sons of former slaves and the sons of former slave-owners will be able to sit down together at the table of brotherhood . . .'. Upper left: waveform and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

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. 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 German. The emphasis is on analysing details of rhythms as speech style features in utterance tokens from individual speakers in specific genres, rather than in claiming validity for the typology of entire languages on the basis of sparse data, as in some 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) envelope of the speech signal, related to the syllable sonority outline of the waveform, and the frequency modulation (FM) envelope of the signal, related to perceived pitch contours.

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 (the terminology is justified in Section 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 & Li 2019; Ludusan & Wagner 2020), and not only with identification of a rhythm but with simultaneous rhythms and rhythm variation over time. The RFT frequency domain definition of rhythm as oscillation implies that rhythms, whether AM or FM, have two main properties, SPECTRAL FREQUENCY and SPECTRAL MAGNITUDE, and that isochrony follows *a fortiori* from the concept of oscillation.

Figure 1 shows four components of an RFT analysis of low frequency (LF) rhetorical RHYTHMS AS OSCILLATIONS, in the first 17 seconds of the well-known and widely

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

available 1963 speech of Martin Luther King, *I Have a Dream*.[1] The positive AM ENVELOPE superimposed on the waveform (Figure 1, upper left) shows an episodic pattern with pauses separating speech, with lower TEXT:PAUSE RATIO in the first half.

In the AM LF SPECTRUM (Figure 1, upper right), the long-term spectral frequencies over the entire data stretch are shown, with the highest magnitude spectral frequencies marked, a salient frequency zone below about 0.7 Hz (corresponding to an average phrasal rhythm unit duration of about 1.7 s) and a salient frequency zone around 1.5 Hz (corresponding to an average word rhythm unit duration of about 0.7 s). The dotted overlay represents smoothing with a moving median window.

In Figure 1, lower right, the AM LF RHYTHM SPECTROGRAM adds a time dimension as a sequence of spectra, from top to bottom, in an overlapping moving 3 s window in WATERFALL FORMAT. Figure 1, lower left, shows the same AM LF rhythm spectrogram but with time from left to right in the more familiar HEATMAP FORMAT. The waterfall and heatmap formats are both derived from the same numerical matrix, but have different heuristic values as visualisations of long-term rhythms. Along the time axis in each spectrogram format, very low frequencies in the first half reflect the low text:pause ratio, a discourse rhythm. Conversely, the higher text:pause ratio in the second half reflects rhythms of information structure and grammatical patterns. Rhythm variation as a rhetorical strategy is clearly reflected in the spectrograms.

In the waterfall spectrogram format, the highest magnitude spectral frequency is marked for each component spectrum. The sequence of these frequencies constitutes the RHYTHM FORMANT TRAJECTORY. This rhythm formant trajectory represents the highly variable LF spectral frequencies through time, which are not captured in earlier analyses of the global, atemporal (time-free, sometimes referred to as anachronic, anachronous) LF spectrum. The spectrogram representations show that the distinct rhythm formants in the spectrum (Figure 1, top right) are not necessarily simultaneous, as the atemporal spectrum may be taken to imply, but are distributed in time. The heatmap spectrogram representation shows clearer details of the temporal dynamics of rhetorical speech rhythms, in particular a rhetorical RHYTHM CYCLE, the 'rhythms of rhythm', spaced at about 7 s with clear divisions in the highest magnitude frequency zone which relate to the distribution of text and pauses in the waveform.

The FREQUENCY DOMAIN and TIME DOMAIN LF spectrogram representations involved in RFT each involve a very different 'mindset' in observation and interpretation practices from 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).

To summarise, RFT extends the spectral analysis approach with new concepts:

1. LOW FREQUENCY ZONES of spectral peaks (Figure 1, top right, marked with vertical lines topped with stars and interpreted as LF RHYTHM FORMANTS, by analogy with the mathematically related high frequency (HF) PHONE FORMANTS.
2. LOW FREQUENCY SPECTROGRAM (cf. the waterfall and heatmap spectrogram formats in Figure 1, bottom right and bottom left, respectively), interpreted as a RHYTHM SPECTROGRAM.
3. LOW FREQUENCY FORMANT TRAJECTORY (dot markers in Figure 1, bottom right, cf. also Figure 9 below, bottom left), the RHYTHM FORMANT TRAJECTORY of highest magnitude peaks in the component spectra of the LF spectrogram.

[1] The use of this 'clear case' example was suggested by Dr. Xuewei Lin, Jinan University, Guangzhou, China.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

4. FM ENVELOPE (fundamental frequency, F0, estimation, 'pitch' track) of the speech signal, associated with tones, pitch accents and intonation, introduced in parallel to the AM ENVELOPE of earlier studies, which is associated with the sonority curve postulated in phonology.

RFT is agnostic with regard to theories of rhythm production and perception, unlike several earlier modulation-theoretic approaches. RFT is open to interpretation in these fields, but is more oriented towards developing a unified rhythm theory and extending previous practical work in the field, for example linguistic and phonetic description and explanation (Gibbon & Li 2019), practical applications in individual speech diagnosis (Liss, LeGendre & Lotto 2010, Carbonell et al. 2015), small group language testing (Gibbon & Lin 2020, Wayland & Nozawa 2020), or speech technology applications (LeGendre et al. 2009).

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

A number of overviews of methods used in rhythm studies are available, including Adams (1979), Dauer (1983), Jassem, Hill & Witten (1984), Gibbon (2006), Arvaniti (2009), Gut (2012), Wayland & Nozawa (2020). White & Malisz (2020) provide a particularly comprehensive survey of the main qualitative and quantitative phonetic approaches to rhythm analysis. It is nevertheless useful to note five main paradigms as a brief general orientation:

1. QUALITATIVE FUNCTIONAL ANALYSIS, often with strong influences from music, dance, poetry and rhetoric, which can be traced back to Aristotle and Cicero and which continues in discourse analysis and interactional linguistics (Brazil, Coulthard & Johns 1980, Couper-Kuhlen & Auer 1991, Couper-Kuhlen & Selting 2018).
2. QUALITATIVE LINGUISTIC MODELS, which often have a pedagogical background, from Sweet's (1908) stress-syllable timing distinction, Jones (1909, 1918) and Palmer's (1924) tonetic structures, through Pike's (1945) stress numerals to Jassem's (1952) rhythm hierarchy and the metrical feet of Abercrombie (1967); cf. the partial overview by Gibbon (1976).
3. QUALITATIVE ALGEBRAIC MODELS OF STRESS HIERARCHIES, from Chomsky, Halle & Lukoff (1956) through Chomsky & Halle (1968) and the metrical theory of Liberman & Prince (1977) to Selkirk (1984) and optimality theories (Prince & Smolensky 2004), sometimes accompanied by quantitative studies of phonetic correlates.
4. SYMBOL–SIGNAL INTERFACE ANALYSIS, which combines qualitative with quantitative methods and relates 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 (Lehiste 1970, Jassem 1952, Roach 1982, Jassem et al. 1984, Scott Isard & de Boysson-Bardies 1985, Ramus, Nespor & Mehler 1999, Grabe & Low 2002, Asu & Nolan 2006, Li, Yin & Zu 2006, Wagner 2007, Dellwo 2010, Yu & Gibbon 2015, Dihingia & Sarmah 2020).
5. QUANTITATIVE MODULATION-THEORETIC SIGNAL PROCESSING, with production and perception models which represent low frequency components of the speech signal, ranging from the rhythmograms of Todd & Brown (1994) and Ludusan, Origlia & Cutugno (2011) through the coupled oscillators of Cummins & Port (1998), Barbosa (2002), Malisz et al. (2016) and the sonority patterns of Galves et al. (2002) and Fuchs & Wunder (2015), to the low frequency envelope spectrum of Tilsen & Johnson (2008), the cubic spline approximation approach of Tilsen & Arvaniti (2013), and the long-term LF spectrum approach of Gibbon (2018, 2019).

There are many more studies in each paradigm, but the items listed here are appropriate representatives.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

1.4 Linguistic–phonetic interface analysis in the time domain

A standard procedure in many quantitative phonetic analyses has been the top–down manual or (semi-)automatic ANNOTATION METHOD of signal–symbol interface analysis by 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, recorded as timestamps paired with transcriptions. The signal annotation (labelling, markup) technique was originally developed an the 1970s and 1980s for statistically trained automatic speech recognisers and speech archiving, and has acquired many other uses: illustration in qualitative and formal linguistic studies, quantitative analysis of interval durations in descriptive phonetics.

In the symbol–symbol interface approach, descriptive statistical techniques variously known as IRREGULARITY METRICS, ISOCHRONY METRICS, INTERVAL METRICS or RHYTHM METRICS are applied to the timestamps in the signal–symbol interface method in order to capture regularities and irregularities of durations among units in the annotation which may be ascribed to rhythmic and arhythmic segments of speech utterances, in particular in search of an isochrony (equality of duration or timing) property. 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 are filtered through the annotator's perception of segmentation and classification in the speech signal (as also noted by Tilsen & Johnson 2008, Liss et al. 2010).

One class of irregularity metric in the search for isochrony relates directly or indirectly to descriptive statistics in the form of global variance or standard deviation, i.e. dispersion of absolute or squared values from the mean, 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 offer a useful HEURISTIC METHOD for initial analysis, but have no theoretical basis: they are suitable for describing static populations, but not sequences, as they ignore interval ordering and positive–negative difference alternation and thus destroy the alternation property of rhythms. Nor do they capture dynamic temporal rhythm variation or relate to linguistic properties of rhythms such as left and right headedness which have been described in phonological studies, as noted by Varnet et al. (2017).

The *Pairwise Variability Index* (*PVI*), with 'raw' (*rPVI*) and 'normalised' (*nPVI*) versions (Grabe & Low 2002, and many other studies), was introduced to reduce the effect of changing speech rate in utterances, but is also inherently more suitable for describing irregularity in time-ordered sequences than the variance-related earlier studies. Asu & 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 relate formally to standard binary DISTANCE METRICS, which compare pairs of numerical vectors. The *rPVI* relates to Manhattan Distance[2] and the *nPVI* relates to Canberra Distance (normalised Manhattan Distance). In *PVI* distance measurement, the vectors V_1 and V_2 of annotated interval durations are not independent, as is generally the case with distance metrics, but are subvectors of the same vector of annotated interval durations $V = <d_1, \ldots, d_m>$:

$$V_1 =< d_1, \ldots, d_{m-1} >,\ V_2 =< d_2, \ldots, d_m >$$

[2] Manhattan Distance (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 diagonal 'as the crow flies' distance.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

The relation $nPVI(V_1,V_2)$ thus defines a binary 'next-door-neighbour' distance, as the 'pairwise' property implies, which has been represented by Wagner (2007) as a two-dimensional scatter plot. The metrics thus embody the simplifying assumption that, formally, rhythm is binary rather than ternary or with longer component sequences (cf. Gibbon 2003). Nolan & Jeon (2014: 3) weaken the binarity assumption to a heuristic assumption of 'sufficient predominance of strong–weak alternation'. Averaging absolute differences means that the *PVI* metrics share the same issues already noted for the variance-based irregularity metrics in failing to model rhythm as such. The *PVI* metrics represent the most widely used signal–symbol interface method, and have also been applied in other disciplines, for example as a heuristic for comparing musical styles (Daniele 2017). The *PVI* metrics have been reviewed exhaustively by Barry et al. (2003), Gibbon (2003, 2006), Tortel & Hirst (2008), Arvaniti (2009), Kohler (2009), Gut (2012), Condit-Schultz (2019) and White & Malisz (2020).

1.5 Frequency domain and time domain approximation approaches

The modulation-theoretic approaches to rhythm analysis have their origins in speech engineering and speech pathology. They are based on direct bottom–up spectral analysis of LF oscillations in the speech signal, without prior annotation or reference to linguistically defined units. These approaches measure frequencies of oscillations in speech with a variety of signal processing methods, including band pass filtering, low-pass smoothing, Hilbert Transform, Fourier Transform or wavelet analysis, and are sometimes conceived as perception models. Very early ideas were developed by Dudley (1939) and Potter, Kopp & Green (1947), noted by Liberman (2013), and later by Dogil & Braun (1988). Contributions in a variety of domains were made by Todd & Brown (1994), Cummins et al. (1999), Foote & Uchihashi (2001), Galves et al. (2002), Lee & Todd (2004), Tilsen & Johnson (2008), Cumming (2010), Heřmanský (2010), Liss et al. (2010), Ludusan et al. (2011), Tilsen & Arvaniti (2013), Liberman (2013), Leong et al. (2014), Fuchs & Wunder (2015) He & Dellwo (2016), Gibbon (2018), Gibbon & Li (2019), Wayland & Nozawa (2020), Ludusan & Wagner (2020).

Models of speech production based on the same formal foundation, in principle, but with COUPLED OSCILLATORS to handle mutual relations between different low frequency rhythms, have also been developed (Cummins & Port 1998, O'Dell & Nieminen 1999, Barbosa 2002, 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 in principle be construed as formal models of abstract structural rhythm cycles, not unlike the coupled oscillator models.

Concepts related to the rhythm spectrum and rhythm spectrogram approach taken here were developed by Todd & Brown (1994) and Ludusan et al. (2011) with the RHYTHMOGRAM, by Ioannides & Sargasyan (2012), based on an amplitude-sensitive and frequency-sensitive auditory hair cell demodulation algorithm with pre- and post-demodulation filtering, and by Foote & Uchihashi (2001) with the BEAT SPECTRUM and BEAT SPECTROGRAM for identifying rhythm variation over time in music (cf. also Brown, Pfordresher & Chow 2017).

Tilsen & Arvaniti (2013) used two methods: a frequency domain AM ENVELOPE method and, as an alternative to FFT analysis, a heuristic EMPIRICAL MODE DECOMPOSITION (EMD) method for establishing variance of instantaneous frequencies on different time scales which are thought to be present in time series measurements. The method combines several distinct empirical techniques: separate peak-picking and cubic spline interpolation of positive peaks and negative peaks (essentially low-pass filtering), subtraction of the mean of the two smoothed curves from the original input and iterating with the same procedure, terminating when there are no longer zero crossings in the mean curves. The EMD, with application of a Hilbert transform to the output, was applied to spontaneous and read data genres (referred to there as elicitation methods), with the aim of finding metrics for degrees of rhythmicity

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

and distinguishing between languages and genres, with plausible results for syllable-sized and stress-group-sized periodicities. The combination of EMD and Hilbert Transform is sometimes known as the Hilbert-Huang Transform, HHT (Huang et al. 1998), which lacks a principled basis in physics or mathematics, unlike the FFT, but with exploratory value for dealing with additive and potentially independent signal mixtures of unknown origin in arbitrary domains such as epidemic modelling and oceanography, and has considerable potential for wider deployment in speech research.

The present study differs from the Tilsen & Arvaniti (2013) study in a number of respects: by testing on larger data sets; by not concentrating only on frequency ranges associated with syllables and stress groups; by, in particular, not excluding pauses or frequencies below 1 Hz; by using data segments orders of magnitude longer than maximally 2.5 s chunks; by introducing the FM envelope (F0 estimation, 'pitch' track) and analysing the FM envelope with the same methods as the AM envelope.

In contrast to the HHT method, in order to include time-domain information the RFA method introduces the LF rhythm spectrogram, using a moving FFT window, from which the highest magnitude frequency trajectories are then extracted. Using the spectrogram as a time domain approximation model, very long-term RHYTHMS OF RHYTHM are identified, and used as input to basic unsupervised machine learning techniques of hierarchical clustering in order to discriminate tokens of different language varieties and speech styles.

A practical 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, Tilsen & Arvaniti 2013, 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. In an early application of envelope analysis in phonetics, Dogil & Braun (1988) used the AM envelope in the form of 'intensity tracing' in order to detect 'pivots' (fast intensity changes) for edge detection in speech segmentation.

In the 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 & Li (2019) and Ludusan & Wagner (2020). FM has indeed been widely studied, but in the contexts of perceived prominence (cf. Malisz & Wagner 2012, Suni et al. 2017, Kallio et al. 2020 for recent treatments) or of the form and function of tone, pitch accent and intonation.

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 rhythm analysis studies (e.g. Grabe & Low 2002, Tilsen & Arvaniti 2013). In order to facilitate validation in future studies, open access recorded data from the Edinburgh *The North Wind and*

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

the Sun corpus[3] (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, one is Northern British English and one is received pronunciation with minor Scottish influence. Of the seven German recordings, five are Swiss German and two 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 Martin Luther King's 1963 speech *I Have a Dream* were examined (cf. Section 1.2 above).

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 languages with different lexical and phrasal 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 artificial 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 order to minimise the range of 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 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 phases of the RFA study for spectrogram calculation and hierarchical classification, and for distance-based hierarchical clustering of the LF formant trajectories.

2.2 The modulation-theoretic framework and RFA methodology

Discovery and description of relations between physical rhythm and properties of language and speech within a unified framework are the core goals of RFT and the RFA methodology. Part of this is the discovery of signal–symbol relations, the association of physical sound with linguistic categories, one of the central goals of phonetics. In the present study, methods from 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).

In addition to the signal processing algorithms described below, pairwise distance metrics are used for (a) explication of the *PVI* irregularity metrics as a next-door-neighbour distance metric (Section 1.4), and, together with hierarchical clustering, (b) identification of rhythm

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

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

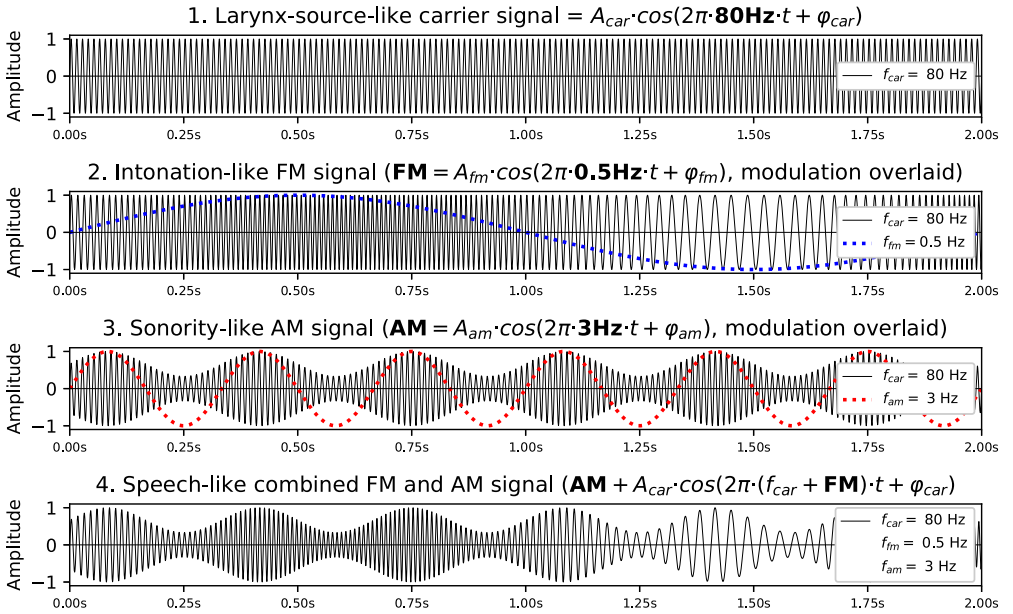

Figure 2 (Colour online) Amplitude and frequency modulation of the speech carrier signal.

formants in the LF spectrum (Section 4.1) and (c) classification of readings in different languages and language varieties on the basis of rhythm variation over time in low frequency rhythm trajectories (Sections 5 and 6).

In the present modulation theoretic approach to rhythm analysis, the key concepts are the CARRIER SIGNAL and the FM and AM MODULATION SIGNALS (Oppenheim, Willsky & Young 1983). Modulations superimpose information-carrying signals on the higher frequency carrier signal. The terms FM and AM have the same meanings in the present context as in FM and AM radio broadcasts, but applied to audio frequencies. Audio frequency modulation of phone formants in higher frequency harmonics of the modulated carrier signal is not considered.

There are many signal modulation modes and several of these modes are relevant for the study of speech. The modes which figure in the present study are frequency modulation (FM) and amplitude modulation (AM), illustrated in Figure 2 in a simplified synthesis perspective which ignores phonation properties and interaction between FM and AM production. The principle is that with FM the frequency of a source signal in the larynx (Figure 2, graph 1) is modulated (Figure 2, graph 2), driven by tone, pitch accent and intonation. With AM the amplitude of the resulting FM signal is modulated (Figure 2, graphs 3 and 4), driven by the coupling of syllable, foot and phrase sonority patterns. The output is 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.

The FM and AM modes have their general signal processing meanings. Simplifying, since the speech source signal is more complex, for a basic sinusoid carrier signal, $S_{car} = A_{car}\cos(2\pi f_{car}t+\varphi_{car})$ and for a basic sinusoid modulation signal $S_{mod} = A_{mod}\cos(2\pi f_{mod}t+\varphi_{mod})$:

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 f_{mod} (in this case: f_{modfm}) values at source;
2. in AM the component A_{car} (amplitude) of S_{car} is modified (for vowel and consonant sequences) by multiplication with the corresponding A_{mod} (in this case: A_{modam}) values of the modulation signal.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

For present purposes, the phase φ of carrier and modulation signals is not considered; however, in some respects, such as speech tempo change, LF FM can be considered to be LF PM, phase modulation.

The terms 'FM' and 'AM' tend not to be discussed as such in phonetics, unlike in audio and radio engineering.[4] In phonetics, the distinction is usually made in terms of source-filter models of speech phonation and articulation, respectively, which tend to be treated as the two entirely different domains of 'prosodic' and 'segmental' phonetics and phonology. Modulation theory is explicitly introduced here for a diametrically opposite reason: in order to be able to focus on similarities rather than differences between AM and FM as factors in speech rhythms within a unified framework.

2.3 RFA procedure

The signal processing model illustrated in Figure 1 is in a sense minimalist, in that a number of filtering and normalisation procedures used in previous studies are not represented. Some of these are also omitted in the present study since they do not make a substantive empirical difference. The procedure covers eight signal processing stages, described in Sections 2.3.1–2.3.8, followed by a summary (see also Figure 1 and figures in Sections 4–6 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 optionally further downsampled.

2.3.2 AM demodulation (envelope extraction)

The positive signal envelope is extracted by rectification of the signal, i.e. by taking positive values of the signal (the superimposed top line in Figure 1, top left), and smoothing the peak sequence. This yields a smoothed curve or envelope which outlines the positive amplitude extremes of the signal (or the intensity curve, here with 'intensity' meaning squared amplitude, and the pivot parser 'intensity track' of Dogil & Braun 1988). The envelope can then be further analysed as a modulation signal in its own right.

More generally, the amplitude envelope is extracted with the absolute values of the Hilbert Transform (cf. He & Dellwo 2016) together with low pass filtering. In the present RTA implementation, however, AM envelope extraction is by band-pass filtering, followed by full-wave rectification (taking absolute values of the signal), followed by low-pass filtering, the 'crystal set' demodulation principle. AM demodulation can be briefly summarised as follows:

$$S_{modam} = lowpass(absolute(bandpass(S_{AM})))$$

Two AM signals with different time window domains are discussed: (a) relating to the shorter-term sonority curve of speech sounds which characterises syllable, word and phrase patterns, in time domains from centiseconds to seconds, and (b) relating to long-term rhythm changes, including pausing, for emotional or rhetorical purposes, in discourse contexts over time domains of tens of seconds or more.

[4] There are passing general references to the concept of amplitude modulation in some phonetics textbooks, but neither to the concept of frequency modulation nor to modulation theory as such.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

2.3.3 FM demodulation

Fundamental frequency patterns are extracted from the signal to form the FM envelope (cf. Section 3 below). FM is the physical modulation mode used for the linguistic information conveyed by lexical tone, pitch accent and intonation. There are many F0 estimation ('pitch' tracking) algorithms; the algorithm used in the present study is a time-domain AVERAGE MAGNITUDE DIFFERENCE FUNCTION (AMDF) algorithm (Krause 1984), related to auto-correlation, which searches for regular period durations and converts them to frequencies. Low-pass filtering and centre and peak clipping are used, then AMDF application, then moving median window smoothing. Discontinuities in the FM envelope caused by voiceless consonants and pauses are treated as spectrally relevant segments of the signal, normalised to median F0 (fundamental frequency) for spectral analysis, but normalised to zero for the figures. FM demodulation can be briefly summarised as:

$$S_{modfm} = lowpass(AMDF(bandpass(S_{FM})))$$

2.3.4 LF spectrum analysis

The Fast Fourier Transform is applied to the demodulated AM and FM envelopes, resulting in a spectrum from which a long-term LF segment below 10 Hz is extracted, for each modulation mode. A flexible window with cosine edges (Tukey window) is applied prior to FFT application. This step is related to the rhythm spectrum approach of Tilsen & Johnson (2008), except that a bandwidth restriction to reduce F0 influence is not included (cf. also Wayland & Nozawa 2020). The long-term signal-length FFT window is mathematically the same as but empirically different from the 10 ms or so short-term window length of pitch extraction and phone formant analysis. The spectrum is atemporal and has frequency and magnitude properties, but no time properties, and therefore also contains no information about dynamic rhythm changes.

2.3.5 LF rhythm formant identification

The spectral frequency zones around magnitude peaks in the LF spectrum (Figure 1) are identified as communicatively relevant rhythm formants (cf. Sections 4 and 5 below). An LF spectral frequency zone is effectively a subsequence of the spectrum and therefore also contains no timing information about dynamic rhythm changes. The rhythm formant concept is indirectly related to the spectral bands intrinsic mode functions of Tilsen & Arvaniti (2013).

2.3.6 LF spectrogram analysis

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

2.3.7 Rhythm formant trajectory identification

The purpose of identifying the rhythm formant trajectory is to provide a source of dynamic temporal rhythm variation for further analysis. The highest magnitude peak in each component spectrum of the spectrogram matrix is identified and a vector pair through time,

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

consisting of the magnitudes and frequencies of these peaks is extracted (cf. the marker dots in the waterfall spectrogram format in Figure 1 and the trajectory graphs in the figures in Sections 4–6). More complex vectors are planned as classifier input.

2.3.8 Hierarchical speech variety classification

The purpose of hierarchical speech variety classification with distance-based dendrograms is to investigate rhythmical similarities between speech tokens in different speech styles and language varieties. The vectors extracted from the rhythm spectrograms of different data items are compared in a novel procedure, using competitive evaluation of a set of combinations of standard distance metrics and standard hierarchical clustering operations. Hierarchical clustering is used in preference to the more usual flat clustering 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 outlined in the previous subsections extend previous AM LF spectrum approaches (a) to include FM spectrum analysis, (b) to include the time variation dimension of LF spectrogram analysis for both AM and FM. For this purpose, the new concepts of rhythm formant, rhythm spectrogram and rhythm formant trajectory are introduced. The unsupervised machine learning technique of distance-based hierarchical clustering is used to classify the speech recordings on the basis of the highest magnitude trajectories through the spectrogram.

2.4 Terminology: Why 'rhythm formant'?

In acoustics, whether in music or in speech processing, the term 'formant' refers (a) to a frequency zone of higher magnitude in the spectrum of harmonics of a periodic sonorant sound, or of non-harmonic noise, which is independent of the fundamental frequency and which shapes the acoustic characteristics of a musical instrument or of speech production, and (b) to properties of the mechanism which shapes this frequency zone, either in a musical instrument or in speech phonation and articulation.

In phonetics the term 'formant' customarily refers to a region of high frequency (HF) resonance in the spectral region above several hundred Hertz and ranging to several thousand Hertz, which characterises speech sounds (Stevens 1998). In RFT, the acoustic definition of 'formant' is generalised to apply also to LF rhythm formants,[5] as low frequency higher magnitude spectral zones. The acoustic differences between HF phone formants and LF rhythm formants are (a) their frequency ranges, (b) the substrate of the formant. The substrate of HF phone formants is the fundamental frequency harmonic series or non-harmonic obstruent 'noise', while the substrate of LF rhythm formants is the sonority pattern which emerges from consonant–vowel sequences and the melodic patterns of tone, accent and intonation. Their functionalities in speech communication and in speech production and perception also differ but this does not affect the acoustic definition.

The term 'rhythm formant' is likely to appear unusual at first glance, perhaps controversial, so it may be helpful to locate the concept of rhythm formant in a unified acoustic frequency space for the speech spectrum, represented by the Speech Modulation Scale

[5] The generalisation of the term 'formant' to cover LF spectral peak zones was suggested by Dr. Huangmei Liu, Phonetics Department, Tongji University, Shanghai (personal communication).

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

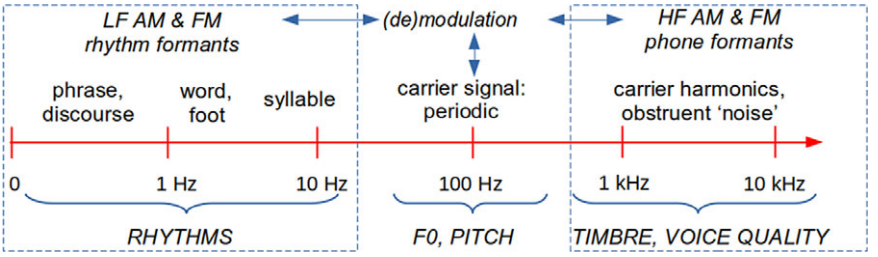

Figure 3 (Colour online) Speech Modulation Scale.

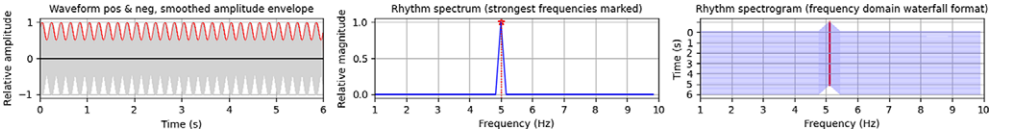

Figure 4 (Colour online) RF analysis of a 5 Hz synthetic sinusoid amplitude modulation of a 200 Hz carrier signal (time is left to right in the left-hand graph, and top to bottom in the waterfall format; frequency is left to right in the centre and right-hand graphs).

(Figure 3). The Speech Modulation Scale is based on a modification of ideas from Cowell's (1930) classic theory of harmonic relations in musicology and shows the places of both HF phone formants and LF rhythm formants, as well as the carrier wave, quite straightforwardly on a logarithmic scale of the frequency ranges used in human speech.

2.5 RFA spectral analysis tool

Empirical study of spectral analysis and its RFT extensions is not possible without a dedicated software tool. For this purpose, a toolset for RFA was developed in Python3, using the standard libraries NumPy for numerical calculation including FFT, MatPlotLib for graphics and SciPy for filter, distance and clustering algorithms.[6] The AMDF algorithm for FM demodulation, together with pre-modulation and post-modulation filtering, was custom designed in 'pure Python' for fundamental frequency estimation over long time stretches. The calculations for this study were carried out under Ubuntu Linux 20.04 with an i7 processor, 32GB RAM and M.2 high speed SSD storage.

Figure 4 shows a synthetic demonstration signal as processed by the RFA toolset. A 200 kHz sinusoid function of 6 s duration was defined with 5 Hz sinusoid amplitude modulation, modulation index 50%, sampling rate 16 kHz, sounding like a regular wave-like stylised voice with a regular fundamental frequency of 200 Hz and a perfectly regular uninterrupted 'syllable sonority' pattern with mean syllable duration 200 ms, mean syllable rate 5 syll/s. The perceptual 'woowoowoo' effect of the synthetic signal is that of a regular wave-like mechanical rhythm, like the 'Wawa pedal' effect for electric guitars.

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 superimposed 5 Hz AM envelope. To make the accuracy check as realistic as possible, the signal was modelled, synthesised and recorded with the Audacity® signal processing tool and analysed in the same way as the authentic speech data.

[6] The information in this paper is sufficient for re-engineering the RFA code, but cf. also the experimental CLI research prototype code: https://github.com/dafyddg/RFA.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

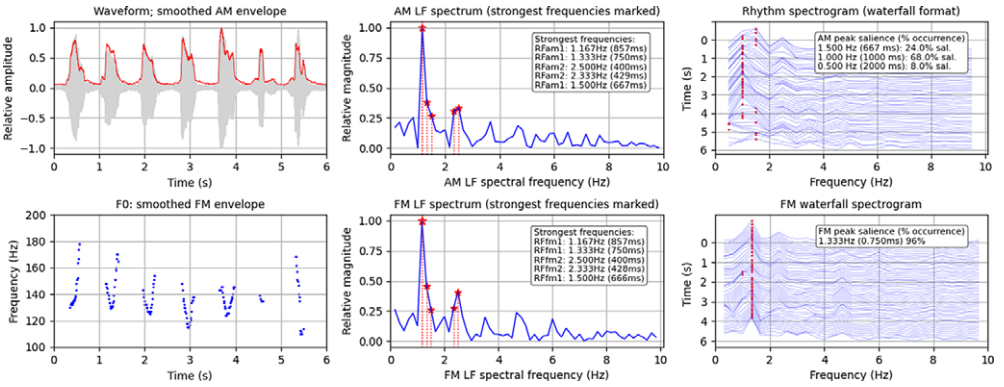

Figure 5 (Colour online) RFT analysis of counting from one to seven (English, adult male); time is left to right in the left-hand graphs, and top to bottom in the waterfall formats; frequency is left to right in the centre and right-hand graphs.

The centre panel of Figure 4 visualises the long-term LF spectrum of the entire 6 s long demodulated AM envelope. The amplitude modulation frequency appears as a high magnitude spectral peak at 5 Hz, marked by a vertical 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 overlapping steps, 60 ms per step) in order to show small gradual rhythm 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 both AM and FM 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 & Li 2019; Ludusan & Wagner 2020). In order to extend this domain, the RFA 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 commonly used rhythmical data type 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, a STRESS–PITCH ACCENT SEQUENCE (SPAS) constraint. 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 in principle to the OBLIGATORY CONTOUR PRINCIPLE (OCP, Leben 1973): an initial pitch pattern assignment spreads rhythmically ('metrically', in phonological terminology) to following non-specified 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 AM and FM domains.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

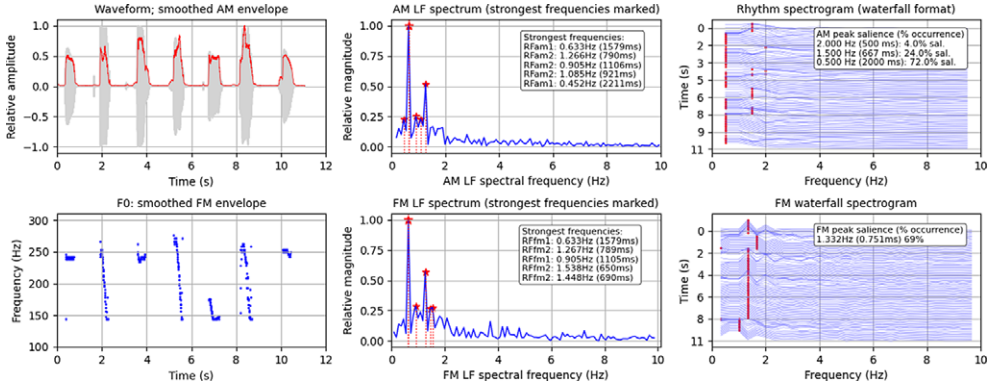

Figure 6 (Colour online) AM and FM rhythm analysis for Mandarin Chinese counting from one to seven.

The English counting sequence follows the SPAS accent sequence similarity constraint for head patterns (Figure 5, lower left). The long-term LF spectrum shows a very close resemblance between the AM spectral pattern and the FM spectral pattern (Figure 5 upper and lower centre). The waterfall spectrogram format (Figure 5 upper and lower right) shows similar temporal dynamics for both AM and FM analyses, with more AM than FM variation, not particularly 'dynamic', of course, in this example: the highest magnitude LF peaks in the rhythm formant trajectory, shown by a dot on each spectral peak, remain fairly constant in the spectrogram region covering the head of the pitch pattern, but differ in the initial 'onset' and final 'nucleus' segments of the signal.

The waveform and the F0 estimation (Figure 5, centre column) both show interesting 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 expected.
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 falling and rising F0 components in this particular accent type.
3. The spectra and spectrograms for both AM and FM show frequencies corresponding to a foot or word repetition rate of around 1.3 Hz (duration mean around 769 ms) and also frequencies reflecting syllable components at 2.5 Hz (400 ms) and 3.7 Hz (270 ms).

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 sequences of binary-structured numerals and binary-structured pitch accents, and to the synchronisation of both domains with words. 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 lexical tones and pitch accents, tone sequences are in principle arbitrary, like other lexical properties. Thus the SPAS constraint does not apply, except for limited tonal sandhi effects, and the prediction is therefore that FM and AM spectral analyses are not as similar in tone languages such as Pǔtōnghuà, Mandarin Chinese, as in English.

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).[7]

[7] The Chinese audio data were kindly provided by Peng Li, M.A., Guangzhou University of Finance.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

Apart from the slower tempo with longer pauses chosen by the Mandarin speaker, the most obvious difference between the FM envelopes of English and Mandarin Chinese is, as predicted, between the SPAS head constraint on local F0 shapes associated with stress–pitch accents in English on the one hand (Figure 5, top left), and the lack of a head constraint on the lexical tones in Mandarin Chinese on the other (Figure 6, bottom left). There is a fortuitous repetition of high level (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 3, the falling Tone 4 and the high level Tone 1. Another difference is the occurrence of salient higher frequency peaks in English above 3 Hz, due to different syllable patterns, which are absent in Mandarin Chinese.

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 (Figure 6, centre). The Mandarin AM and FM spectra are somewhat less similar than the English spectra, both showing very low frequency values below 1 Hz (for interpausal units, average durations around 1.6 s, both measured in the time domain and given by the spectrum) and being similar up to about 1.3 Hz for word-like properties, but showing differences in frequency zones above 2 Hz and especially around 4 Hz (average unit duration 250 ms). Tests to investigate differences of syllable structure and tone structure, also with less regular patterns, are required.

The conclusion is that Mandarin Chinese has a more compact distribution of AM and FM spectral frequencies, and does not have the SPAS FM head pattern of English. 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 to some extent 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 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, morphological tone and a tendency to agglutinative morphology. The roles of variables such as speech genre or register, age, gender, emotionality, in addition to tonal typology, as functional factors in the rhythmic differences, require further investigation.

4 Signal–symbol association in rhythmical speech

4.1 Morphophonology and rhythm: A hypothesis

In a linguistically motivated further step forward in the exploratory use of the RFA toolset, an effect of monosyllabic and polysyllabic morphophonological patterning on long-term rhythmical timing was examined using fairly rapid counting from one to thirty in British English, spoken by an adult male native speaker. The null hypothesis here is that there is no rhythmic difference between these two patterns. The alternative hypothesis is that rhythm has a configurative function and that morphophonological patterns are rhythmically marked. The signal was annotated on syllable and word tiers, with additional tiers containing formant information and interpretative commentary. An RFA spectral analysis is shown in Figure 7.

The LF AM spectrum shows two distinct high magnitude frequency zones (see text box in Figure 2, top right), interpreted as rhythm formants. The spectrum is atemporal, but the spectrograms show that the two frequency zones at different but overlapping times (timestamps are from the annotation shown in Figure 8):

Zone 1. *One* to *ten*, essentially monosyllables, from 0.187 s to 5.482 s, covering the frequencies 1.742 Hz, 1.8 Hz and 1.859 Hz.

Zone 2. *Eleven* to *thirty*, essentially polysyllables, from 5.482 s to 16.978 s, at 3.659 Hz, with a reduction in frequency at 12.5 s (in *twenty-three*) to 3.427 Hz.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

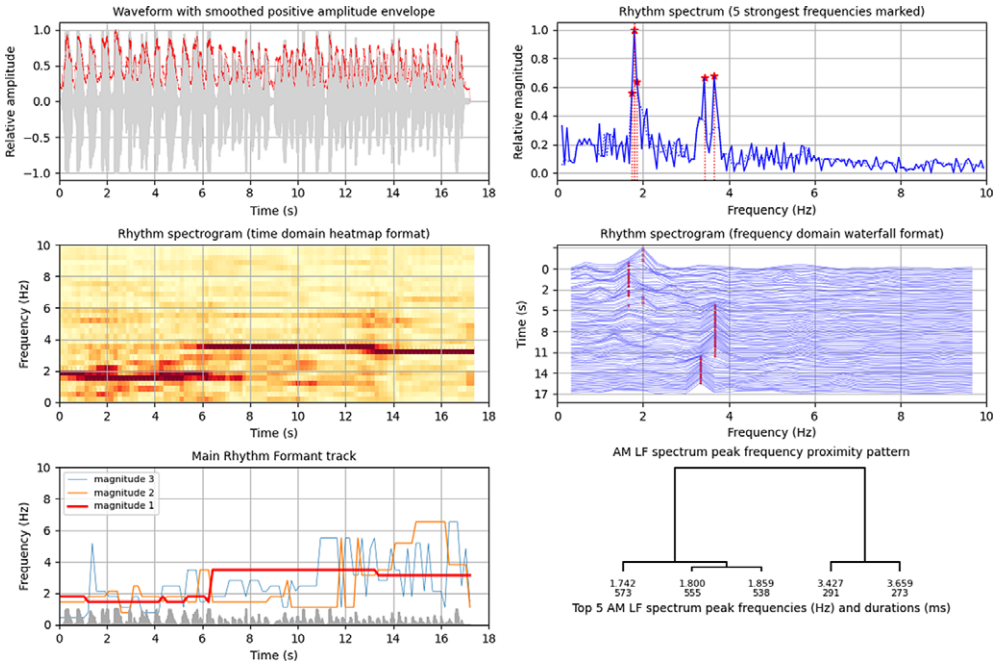

Figure 7 (Colour online) Counting from one to thirty, British English, adult male.

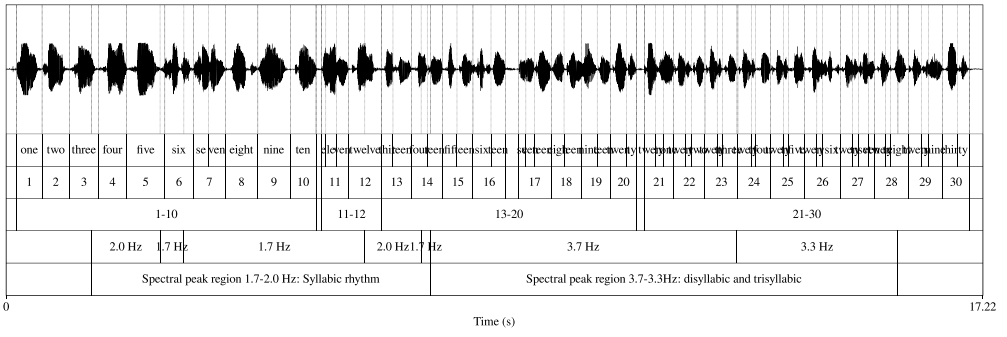

Figure 8 Annotation of the speech signal using the Praat phonetic workbench.

In the transitional area *eleven*, *twelve*, zone 1 and zone 2 overlap for about 2 s, due to the difference between a polysyllable and a monosyllable, each separated from adjacent similar words. Zone 1 reappears weakly at 9.5 s, at a slight pause after *twenty*. The small frequency reduction in zone 2 is from 12.5 s to 13.5 s, covering *twenty-three* and *twenty-four*. While the difference between zone 1 and zone 2 is clear, the differences within zone 1 and zone 2 are imperceptible. Note that the temporal resolution of the spectrograms is necessarily fuzzy due to the overlapping 3 s FFT windows required for capturing rhythm periods up to 1.5 s (frequency zones higher than 0.7 Hz).

The null hypothesis of a random relation between timing and morphophonology is refuted, since there is a clear distinction between the slower monosyllable rhythm in zone 1 and the faster polysyllable rhythm in zone 2. The alternative hypothesis of a configurational function of rhythm appears to be confirmed.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

Table 1 Descriptive statistics for syllable and word time-stamps (durations in milliseconds).

	Syllables		Words	
n:	61		30	
Total sample count:	16474		16771	
Min duration:	78	ms	447	ms
Max duration:	677	ms	808	ms
Duration range:	599	ms	361	ms
Duration mean:	270.07	ms	559.03	ms
Mean syllable rate:	3.7	approx. 3.7 Hz	1.79	approx. 1.7 Hz
nPVI:	46	(irregular)	12	(regular)

4.2 Testing the hypothesis

In order to cross-check the initial informal hypothesis of a rhythm–morphophonology relation with an independent signal–symbol interface method, 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). The annotation file was analysed using the TGA (Time Group Analysis) online tool (Gibbon 2013, Yu & Gibbon 2015). Figure 8 visualises the annotation.

Syllables are annotated in the top tier, then words, then the morphologically characterised sequences 1. . .10, 11. . .20 and 21. . .30. 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.

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: the syllable rate in the rhythmical counting data is 3.7 syll/s, close to the measured rhythm formant of 3.5 Hz, with average syllable duration of 270 ms, close to the predicted 286 ms. The word or foot rate is 1.79 syll/s, near to the rhythm formant at 1.8 Hz, and mean foot or word duration is 559 ms, near the predicted mean 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 & Nolan 2006 and discussion in Section 1.4), and is compatible with the RFA result.

The informal grammar–rhythm alignment alternative 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 rhythm thus has a configurative function for these subsequences. The three utterance segments with internally regular rhythms provide additional physical confirmation for the SPAS constraint on English head patterns.

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 a bilingual

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

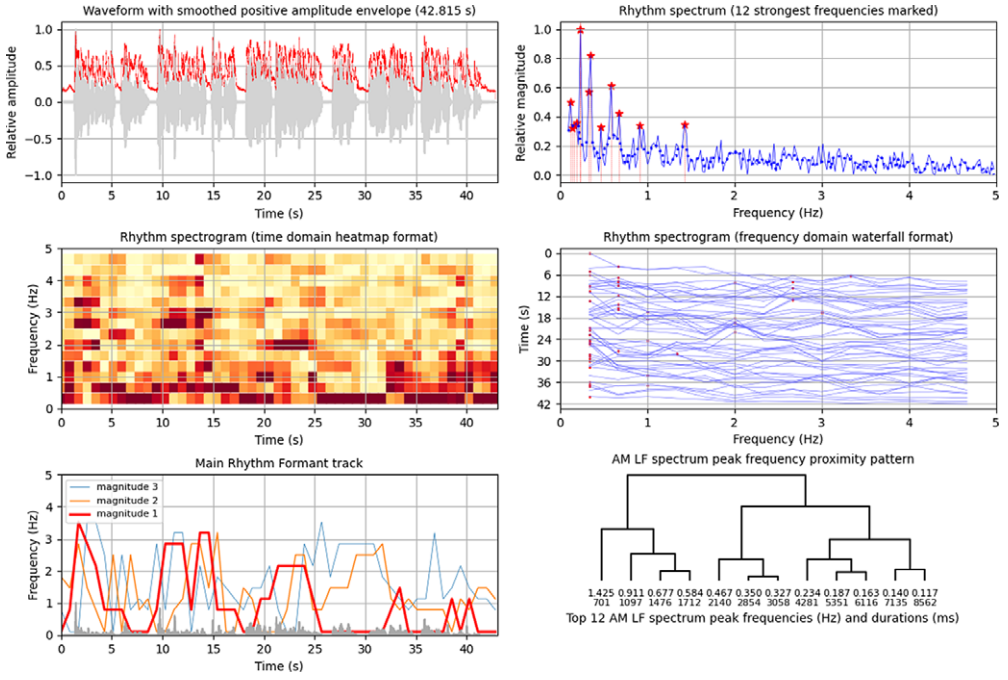

Figure 9 (Colour online) Reading of 'The North Wind and the Sun' in English by a female adult German–English bilingual.

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 as possible an approximation to authentic data without a full natural scenario but with some control over situational variables.

The rhythm formant analysis of English reading 1 is shown in Figure 9. The waveform (Figure 9, top left) shows an episodic structure with intervening pauses. The LF AM spectrum (Figure 9, top right) shows that the most salient frequencies are lower than 1 Hz (period duration greater than 1 s). These very low frequencies are characteristic of the rhythmic patterning of spoken discourse. For example, the most salient frequency at 0.234 Hz (period 4.281 s) implies the presence of a rhythm with 10 cycles during the 42.815 s of the recording. A rough indication of this 4 Hz rhythm may also be observed impressionistically in the interpausal units in the waveform. The 1.425 Hz frequency (702 ms period) indicates foot or word sized units.

The clearest indication of the temporal distribution of discourse-level rhythm patterns is provided by the spectrogram in heatmap format (Figure 9, centre left). The dendrogram (Figure 9, bottom right), based on numerical proximity, shows three main frequency zones which serve as hypotheses for the presence of rhythm formants: 0.12–0.23 Hz, 0.33–0.47 Hz and 0.58–1.42 Hz. Since the formants are characterised by spectral peaks, the highest magnitude peak in each cluster counts as a formant centre hypothesis. The rhythm formant trajectory (Figure 9, bottom left) reflects the salience pattern for the three highest magnitude formant tracks, the highest of which forms the input to the distance-based hierarchical clustering of language varieties.

5.1.2 Distance based hierarchical clustering of AM rhythm formant frequency vectors

Each rhythm formant trajectory is two-dimensional, with SPECTRAL MAGNITUDE and SPECTRAL FREQUENCY pairs, which are extracted as separate vectors. The working hypothesis is

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

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

Symmetrical partitioning		Right-branching scale	
Metric	Cluster votes	Metric	Cluster votes
Canberra	7/7	Canberra	0/7
Chebyshev	2/7	Chebyshev	3/7
Cosine	1/7	Cosine	1/7
Euclidean	0/7	Euclidean	5/7
Manhattan	7/7	Manhattan	0/7
Pearson	2/7	Pearson	1/7
Total:	19/42	Total:	10/42

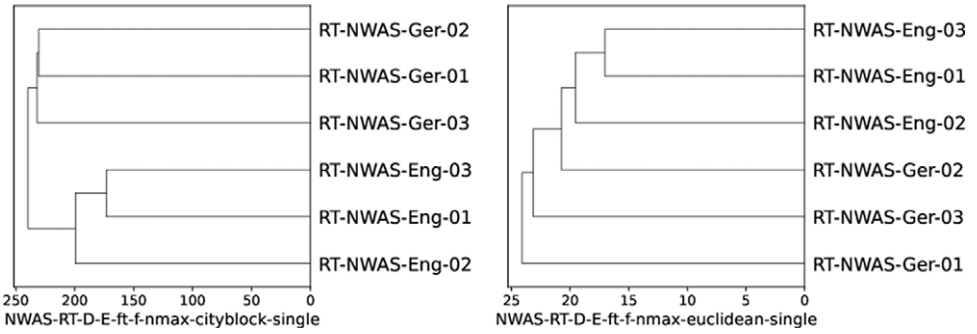

Figure 10 Partitioning shapes with most distance metric votes. Left: symmetrical (19 votes). Right: right-branching (10 votes).

that the English and German data sets are randomly distributed. The alternative hypothesis is that the two data sets are cleanly partitioned.

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 in dendrogram format (synonyms in parentheses):

1. Distance metrics: Manhattan (Mannheim, 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 null hypothesis is random assignment in language variety clustering, the alternative hypothesis is the success criterion of 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 clustering types are given in Figure 10: one is a clear clustering, the other a right-branching pattern which is effectively a linear-ordered scale. Other clustering types were not successful.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

An interesting feature of the data sets is that the German data are more loosely clustered than the English data: in Figure 10 (left), the lengths of the leaf edges for German are longer than for English, and in Figure 10 (right), the English data are lower in the hierarchy than the German data. This may relate to the status of English as the secondary language of the bilingual speaker.

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 results show a total of 29 votes (69%) for full English–German partitioning in contrast to 13 votes (31%) for shapes with varying lower degrees of separation, thereby refuting the null hypothesis of no rhythmical difference between the data sets and confirming the alternative hypothesis of clean partitioning.

The most successful overall clustering criterion was the Ward minimal variance method, with three symmetrical clusterings (based on Canberra, Chebyshev and Manhattan metrics) and one right-branching clustering (based on the Euclidean metric) as well as one mixed symmetrical and right-branching cluster (Pearson). 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 good correlations would be expected separately for English and for German. With Cosine Distance partitioning was not successful.

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: the closely related Manhattan and Canberra 'round the corner' distance metrics appear to be more suitable for the high-dimensional irregular spectral patterns of the present data with quasi-random components, while the Euclidean 'as the crow flies' distance metric takes 'short cuts' through the irregular patterns and tends to be unsuitable for the data. Chebychev distance has intermediate properties. Pearson, unsurprisingly, does not show overall correlation, while Cosine distance shows similarity of orientation in the data space, i.e. direction or angle, rather than similar distance magnitude.

5.2 Rhythm formant trajectories and the 'rhythms of rhythm'

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 spectrograms and the highest magnitude trajectories of all the readings shows that the rhythmic properties of speech vary in the course of the narrative. In addition to well-documented syllable, word and phrase rhythms, these short-term rhythms themselves vary in very long-term higher ranking 'rhythms of rhythm', with which the reader employs rhetorically motivated 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 above.

Figure 11 shows the AM spectral magnitude trajectories of the three English and three German readings 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 needing to detail 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 shown in the dendrograms of Figure 10.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

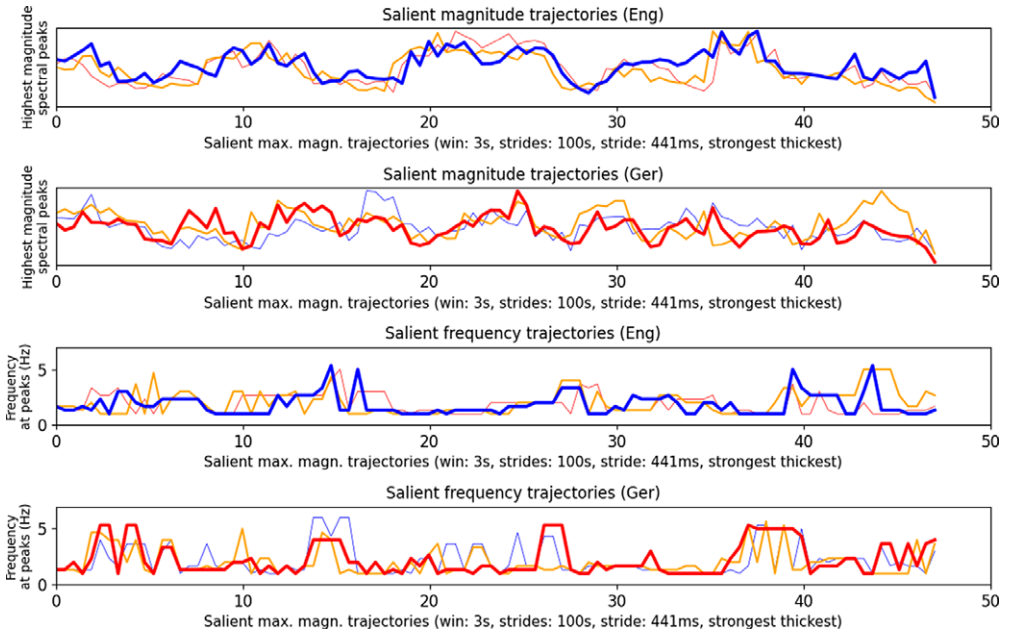

Figure 11 (Colour online) Salient (highest magnitude) trajectory vectors through the rhythm spectrograms. Upper two panels: magnitude vector. Lower two panels: frequency vector.

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 needing to detail 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 0.5 Hz 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 conspicuous property is also in evidence: discourse rhythms with a number of very slow oscillations of different frequencies, as in Figure 1, second order 'rhythms of rhythm', with magnitude trajectories varying fairly regularly in time. One of these rhythms has, for example, approximately 10 s periods (at 10 s, 20 s, 30 s), corresponding to 0.1 Hz, which relates well to the values at the low end of the long-term spectrum (Figure 9). The German discourse rhythm pattern is almost twice as fast: nine crests in the course of 50 s, about 5.6 s per wave, a frequency of about 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, they may derive from asymmetrical bilingualism, they may be 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 the first line: *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

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

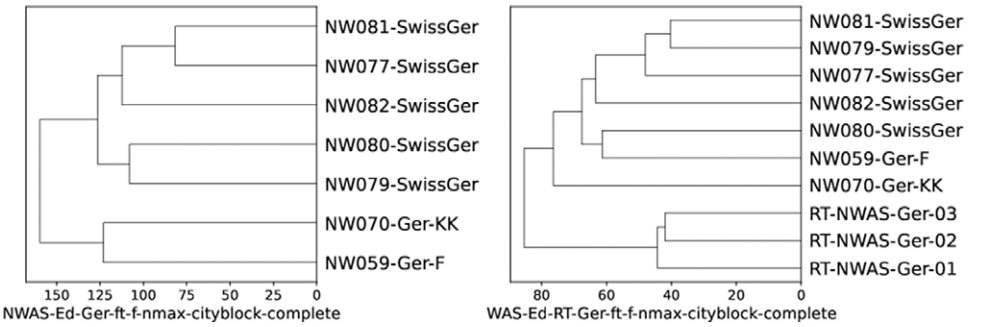

Figure 12 Hierarchical classification of the German readings of 'Nordwind und Sonne' (from the Edinburgh NWAS corpus). Left: Edinburgh German readings only. Right: Edinburgh plus three German readings by a German–English bilingual speaker.

and intonation (cf. Gibbon & 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 in the second sentence, 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, and on the role of FM, is needed than can be provided in the present context. However, this 'rhythms of rhythm' technique of examining the trajectory of the maximum spectral magnitude and frequency through time opens up a new 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 reading in her two high proficiency languages suggest that RFT with the RFA 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 Classification of language varieties

6.1 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.

The first prediction is that the Swiss German speakers are clearly partitioned from the Standard German speakers. The second prediction is that the speakers from the Edinburgh corpus are clearly partitioned from the female English–German bilingual speaker. Results are shown in Figure 12.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

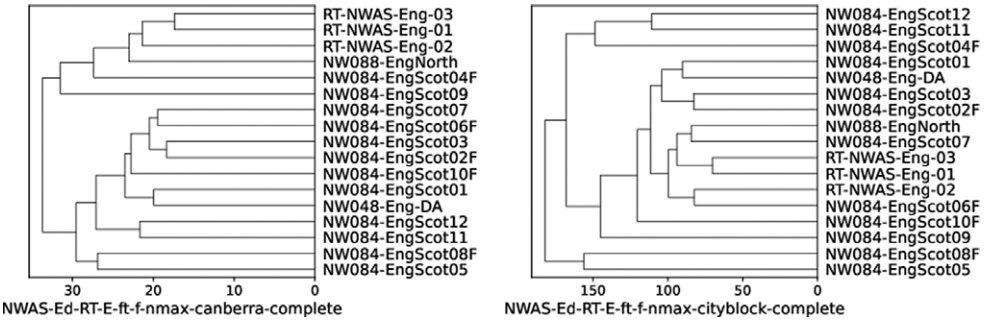

Figure 13 Rhythm formant dendrograms for English readings of 'The North Wind and the Sun': Edinburgh corpus plus three English readings by a German–English bilingual speaker. Left: Manhattan–Vorhees. Right: Canberra–Voorhees.

The left panel shows only German readings from the Edinburgh corpus, the right-hand panel shows results after adding the German readings by the bilingual speaker. The predictions are fulfilled. This exploratory data selection is small, however, though somewhat larger than in some previous studies, and in a subsequent confirmatory study stricter experimental conditions need to be applied.

The results for the English readings from the Edinburgh corpus together with the English readings by the German–English bilingual speaker were classified by the Manhattan–Voorhees pair (Figure 13, left) and the Canberra–Voorhees pair (Figure 13, right), respectively.

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 in both dendrograms, which is perhaps not too unexpected as he lived and worked in Scotland for three decades. Two English readings by the bilingual speaker (Figure 13), which previously clustered together, are also clustered together in both dendrograms. Further discourse analytic and sociolinguistic interpretations may be possible but are not the subject of the present investigation.

6.2 Size and inhomogeneity of the data: A 'stress test'

A 'stress test' with 98 readings of *The North Wind and the Sun* was conducted, a sample size which is several times larger than in previous experiments (e.g. 18 in Grabe & Low 2002, eight in previous sections of the present study, six in Tilsen & Arvaniti 2013). The number of phonetic, linguistic and sociolinguistic variables involved in the analysis of discourse prosody is high, and consequently expectations of perfect partitioning are illusory. As predicted, analyses of all 98 samples of readings in the Edinburgh database show that results become less interpretable as the size and inhomogeneity of the data set increase (cf. Figure 14), though a few plausible small sub-clusters of related languages can be identified *post hoc*, including pairs of readings by speakers of the same language (e.g. Urdu, Swedish, Xhosa, English, North German, Greek, Scottish English), as well as a cluster of four Scottish English speakers.

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 as a whole is an excellent resource for substantive contributions to exploratory research, but results for subsets of the corpus suggest that the RFT method at its present stage of development is likely to be most useful for smaller and more well-defined and homogeneous data sets and for focussing on speaking styles rather than on language typology.

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

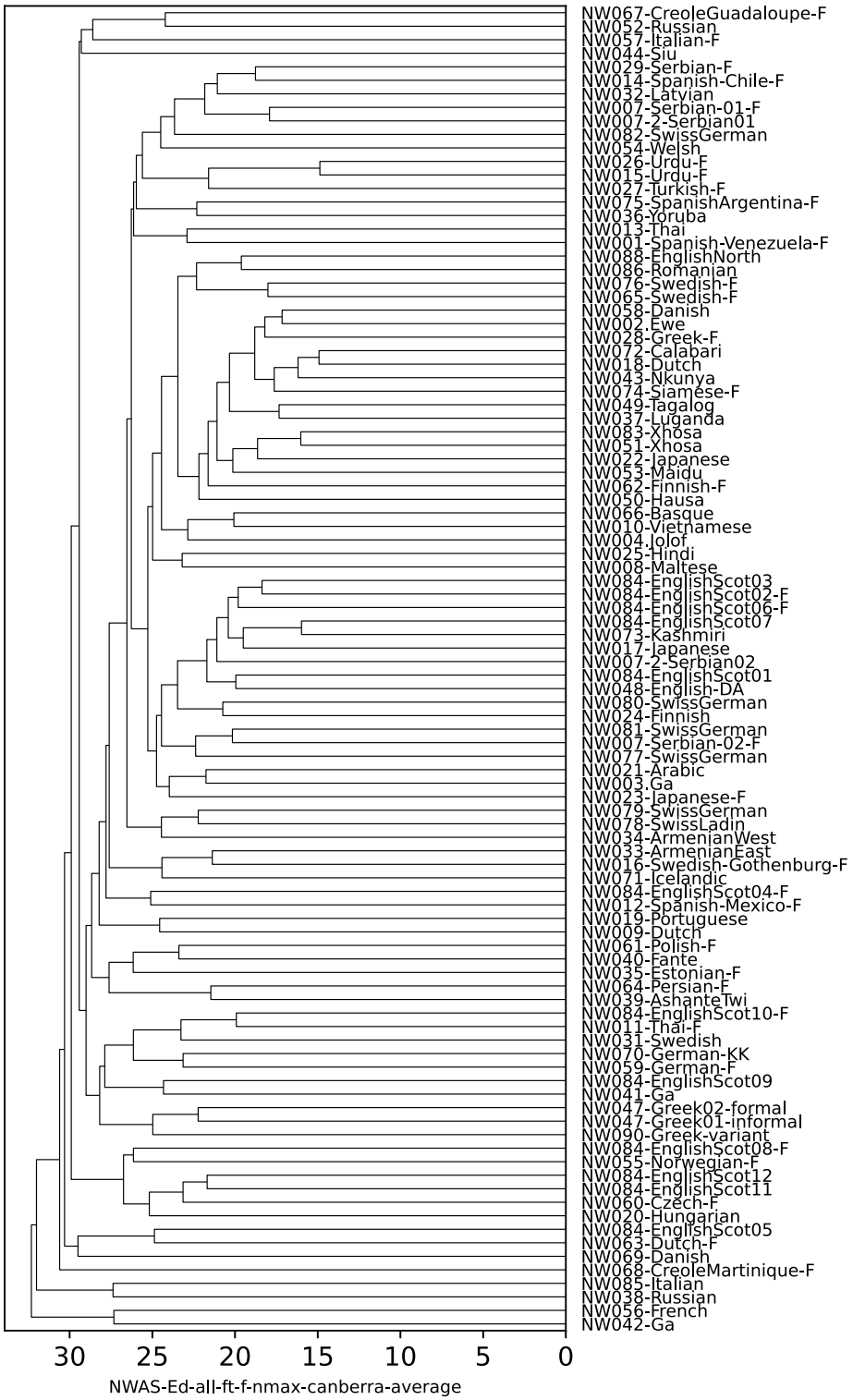

Figure 14 Rhythm trajectory dendrogram for all Edinburgh NWAS corpus readings (N = 98).

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

Results from the 'stress test' provide a pointer to handling larger data sets, for which one or more of the following 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 data inhomogeneity;
3. software optimisation and 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 vector complexity.

Nevertheless, the results for the 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 next phase of development of RFT and RFA is to proceed beyond the exploratory stage of the methodology and to pursue detailed and well-defined confirmatory studies of the large number of variables involved in narrative data, in which rhetorical strategy and style, information structure, grammatical structure, word patterns as well as language and dialect play a role.

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 new unified modulation-theoretic framework, Rhythm Formant Theory (RFT), based on modulation theory and an associated signal processing methodology, Rhythm Formant Analysis (RFA). A number of new concepts were derived from this methodology and used in several independent exploratory investigations not only of amplitude modulation but also of frequency modulation of the speech signal. The method is not restricted to the study of individual cases or statistical prosodic typology but provides pointers to interface issues between phonetic prosody and phonological prosody as a step towards providing abstract language prosody with a previously lacking unified physical empirical grounding in speech prosody.

The central concept of RFT 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 example of a recording of English counting from one to thirty it was found that different morphophonological patterns align with different rhythms, revealing a configurative function of longer term rhythms. It was also shown that there are differences between rhythms of counting patterns in English, in which a similarity constraint on the heads of stress-pitch rhythm groups applies, on the one hand, and Mandarin Chinese on the other, to which the constraint does not apply because of the phonemic arbitrariness and variability 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.

The conclusion is drawn that in order to study the dynamics of rhythm variation, a non-trivial extension of the modulation-theoretic spectral analysis approach to rhythm analysis in the form of RFT and the RFA method opens up fruitful avenues of research, particularly

https://doi.org/10.1017/S0025100321000086 Published online by Cambridge University Press

in discriminating rhythm types as much as in identifying them, focusing on induction from individual discourse tokens, rather than on entire languages with 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 microrhythmic relations between neighbouring syllables which has been the subject of much prior phonetic research (cf. Section 2 above).

A wide range of open issues in the study of speech rhythm remains, for the solution of which the conceptual instrument of RFT and the RFA method may be suitable. One practical issue is the lack of well-defined and available data and tools for enabling the reproduction of results. The present study has attempted to ameliorate this situation, first, by using the readily available Edinburgh multilingual *The North Wind and the Sun* corpus, and, second, by making the research prototype RFA software toolset 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 directly to spectral properties of the speech signal, not to lexicons and texts.

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 and a more sophisticated classification of spoken language varieties, 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 (Rossini & Gibbon 2011), 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 interactions of abstract information structure, grammatical pattern and poetic metre on the one hand and physical speech and poetic performance on the other 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 extensive and unfied modulation-theoretic approach to the empirical grounding of rhythm than was previously available, but is still in need of detailed quantitative confirmatory studies in linguistic and sociolinguistic contexts of rhetorical choice, speaker idiosyncrasy and rhythmic variation in different genres and across languages and language varieties.

Acknowledgements

For detailed and constructive comments and suggestions I am very grateful to two *JIPA* referees. Many stimulating discussions on the topic have helped to sharpen the ideas at various times over the years, and a number have helped with finding out-of-the-way publications: my two ‘prosody brothers’ Daniel Hirst and Nick Campbell, and Jolanta Bachan, Doris Bleiching, Fred Cummins, Grażyna Demenko, Laura Dilley, the late Grzegorz Dogil, Katarzyna Dziubalska-Kołaczyk, Alexandra Gibbon, Sascha Griffiths, Maciej Karpinski, Katarzyna Klessa, Peng Li, Mark Liberman, Xuewei Lin, Huangmei Liu, the late Jack Windsor Lewis, the late Nicla Rossini,

Priyankoo Sarmah, Rosemarie Tracy, Petra Wagner, Ratree Wayland, Laurence White and Jue Yu. Above all, I owe the challenge of research on both structural and spectral patterns of speech to 40 years of detailed discussions with the late Wiktor Jassem, the first fruit of which was Jassem & Gibbon (1980).

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.

Adams, Corinne. 1979. *English speech rhythm and the foreign learner*. Den Haag: Mouton.

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

Asu, Eva-Liina & Francis Nolan. 2006. Estonian and English rhythm: A twodimensional quantification based on syllables and feet. In Rüdiger Hoffmann & Heinz Mixdorff (eds.), *3rd International Conference on Speech Prosody*, paper 229.

Barbosa, Plinio A. 2002. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. In Bel & Marlien (eds.), 163–166.

Barry, William J., Bistra Andreeva, Michela Russo, Snezhina Dimitrova & Tania Kostadinova. 2003. Do rhythm measures tell us anything about language type? In Daniel Recasens, Maria-Josep Solé & Joaquin Romero (eds.), *15th International Congress of Phonetic Sciences* (ICPhS XV), 2693–2696.

Bel, Bernard & Isabel Marlien (eds.). 2002. *1st International Conference on Speech Prosody*, 163–166. Aix-en-Provence: Laboratoire Parole et Langage.

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

Brazil, David, Michael Coulthard & Catherine Johns. 1980. *Discourse intonation and language teaching*. London: Longman.

Brown, Steven, Peter Q. Pfordresher & Ivan Chow. 2017. A musical model of speech rhythm. *Psychomusicology: Music, Mind, and Brain* 27(2), 95–112.

Campbell, W. Nicholas. 1992. *Multi-level speech timing control*. Ph.D. dissertation, University of Sussex.

Carbonell, Kathy M., Rosemary A. Lester, Brad H. Story & Andrew J. Lotto. 2015. Discriminating simulated vocal tremor source using amplitude modulation spectra. *Journal of Voice: Official Journal of the Voice Foundation* 29, 140–147.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Chomsky, Noam, Morris Halle & Fred Lukoff. 1956. On accent and juncture in English. In Halle, Morris, Horace G. Lunt, Hugh McLean & Cornelis H. van Schooneveld (eds.), 1956. *For Roman Jakobson: Essays on the occasion of his sixtieth birthday*, 65–80. The Hague: Mouton & Co.

Condit-Schultz, Nathaniel. 2019. Deconstructing the nPVI: A methodological critique of the normalized Pairwise Variability Index as applied to music. *Music Perception* 36(3), 300–313.

Couper-Kuhlen, Elizabeth & Peter Auer. 1991. On the contextualising function of speech rhythm in conversation: Question–answer sequences. In Jef Verschueren (ed.), *Levels of linguistic adaptation: Selected papers of the International Pragmatics Conference, Antwerp*, 1987, 1–18. Amsterdam: John Benjamins.

Couper-Kuhlen, Elizabeth & 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, University of Cambridge.

Cummins, Fred, Felix Gers & Jürgen Schmidhuber. 1999. Language identification from prosody without explicit features. *Proceedings of The Sixth European Conference on Speech Communication and Technology* (EUROSPEECH '99), Budapest, 371–374. International Speech Communication Association.

Cummins, Fred & Robert Port. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26, 145–171.
Daniele, Joseph. 2017. The "rhythmic fingerprint": An extension of the nPVI to quantify rhythmic influence. *Empirical Musicology Review* 11(2), 243–260.
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*. Ph.D. dissertation, Universität Bonn.
Dihingia, Leena & Priyankoo Sarmah. 2020. Rhythm and speaking rate in Assamese varieties. *10th International Conference on Speech Prosody*, Tokyo, Japan, 561–565.
Dilley, Laura C. 1997. *The phonetics and phonology of Tonal Systems*. Ph.D. dissertation, MIT.
Dogil, Grzegorz & Gunter Braun. 1988. *The PIVOT model of speech parsing*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.
Dudley, Homer. 1939. Remaking speech. *The Journal of the Acoustical Society of America* 11(169).
Foote, Jonathan & Shingo Uchihashi. 2001. The beat spectrum: A new approach to rhythm analysis. Presented at IEEE International Conference on Multimedia and Expo.
Fuchs, Robert & Eva-Maria Wunder. 2015. A sonority-based account of speech rhythm in Chinese learners of English. In Ulrike Gut, 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*, 165–183. Berlin: de Gruyter.
Galves, Antonio, Jesus Garcia, Denise Duarte & Charlotte Galves. 2002. Sonority as a basis for rhythmic class discrimination. In Bel & Marlien (eds.), 323–326.
Gibbon, Dafydd. 1976. *Perspectives of intonation analysis*. Bern: Lang.
Gibbon, Dafydd. 1987. Finite state processing of tone systems. In Bente Maegaard (ed.), *3rd Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen, 291–297.
Gibbon, Dafydd. 2003. Computational modelling of rhythm as alternation, iteration and hierarchy. *15th International Congress of Phonetic Sciences* (ICPhS XV), Barcelona, 2489–2492.
Gibbon, Dafydd. 2006. Time types and time trees: Prosodic mining and alignment of temporally annotated data. In Stefan Sudhoff, Denisa Lenertova, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.), *Methods in empirical prosody research*, 281–209. Berlin: Walter de Gruyter.
Gibbon, Dafydd. 2013. TGA: A web tool for Time Group Analysis. In Brigitte Bigi & Daniel Hirst (eds.), *Tools and resources for the analysis of speech prosody* (TRASP Workshop), 66–69. Aix-en-Provence.
Gibbon, Dafydd. 2018. Keynote: The future of prosody: It's about time. In Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner, Maciej Karpiński & Daniel Śledziński (eds.), *9th International Conference on Speech Prosody*, Poznań, 1–9.
Gibbon, Dafydd. 2019. CRAFT: A multifunction online platform for speech prosody visualisation. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds), *19th International Congress of Phonetic Sciences* (ICPhS XIX), Melbourne, 2956–2960.
Gibbon, Dafydd & Sascha Griffiths. 2017. Multilinear grammar: Ranks and interpretations. *Open Linguistics* 3(1), 265–307.
Gibbon, Dafydd, Daniel Hirst & Nick Campbell (eds.). 2020. *Rhythm, melody and harmony in speech: Studies in honour of Wiktor Jassem*: Special issue of *Speech and Language Technology* 14/15, 83–94. Poznań: Polish Phonetics Society.
Gibbon, Dafydd & Peng Li. 2019. Quantifying and correlating rhythm formants in speech. In Shu-Chuan Tseng (ed.), *3rd International Symposium on Linguistic Patterns in Spontaneous Speech*, 56–61. Taipei: Academia Sinica.
Gibbon, Dafydd & Xuewei Lin. 2020. Rhythm Zone Theory: Speech rhythms are physical after all. In Magdalena Wrembel, Agnieszka Kiełkiewicz-Janowiak & Piotr Gąsiorowski (eds.), *Approaches to the study of sound structure and speech: Interdisciplinary work in honour of Katarzyna Dziubalska-Kołaczyk*, 109–118. London: Routledge.
Gibbon, Dafydd & Helmut Richter (eds.). 1984. *Intonation, accent and rhythm: Studies in discourse phonology*, 203–225. Berlin: Walter de Gruyter.

Grabe, Esther & Ee Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. In Carlos Gussenhoven & Natasha Warner (eds.), *Laboratory Phonology 7*, 515–546. Berlin: De Gruyter Mouton.

Gut, Ulrike 2012. Rhythm in L2 speech. In Gibbon et al. (eds.), 83–94.

He, Lei & Volker Dellwo. 2016. A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform. *Interspeech 2016*, San Francisco, 530–534. ISCA.

Heřmanský, Hynek. 2010. History of modulation spectrum in ASR. *International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 5458–5461.

Huang, Norden E., Zheng Shen, Steven R. Long, Man-Li C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi-Chao Tung & Henry H. Liu. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society, London, A* 454, 903–995.

Hyman, Larry M. 2009. How (not) to do phonological typology: The case of pitch–accent. *Language Sciences* 31, 213–238.

Inden Benjamin, Malisz Zofia, Petra Wagner & Ipke Wachsmuth. 2012. Rapid entrainment to spontaneous speech: A comparison of oscillator models. In Naomi Miyake, David Peebles & Richard P. Cooper (eds.), *34th Annual Conference of the Cognitive Science Society*, 1721–1726. Austin, TX: Cognitive Science Society.

Ioannides, Andreas A. & Armen Sargasyan. 2012. Rhythmogram as a tool for continuous electrographic data analysis. *10th IEEE International Conference on Information Technology and Applications in Biomedicine*, Corfu, 205–211.

Jansche, Martin. 1998. A two-level take on Tianjin Tone. *10th European Summer School in Logic, Language and Information, Student Session*, Saarbrücken, 162–174.

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

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

Jassem, Wiktor, David R. Hill & Ian H. Witten. 1984. Isochrony in English speech: Its statistical validity and linguistic relevance. In Gibbon & Richter (eds.), 203–225.

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, Manfred 1984. Recent developments in speech signal pitch extraction. In Gibbon & Richter (eds.), 243–252.

Leben, William. 1973. *Suprasegmental phonology*. Ph.D. dissertation, MIT.

Lee, Christopher S. & 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), 225–254.

LeGendre, Susan J., Julie M. Liss, Andrew J. Lotto & Rene Utianski. 2009. Talker recognition using envelope modulation spectra. *The Journal of the Acoustical Society of America* 125(5), 2530–2531.

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

Leong, Victoria, Michael A. Stone, Richard E. Turner & Usha Goswami. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America* 136(1), 366–381.

Li, Aijun, Zhigang Yin & Yiqing Zu. 2006. A rhythmic analysis on Chinese EFL speech. In Rödriger Hoffmann & Hansjörg Mixdorff (eds.), *3rd International Conference on Speech Prosody*, Dresden, Germany.

Liberman, Mark. 2013. Speech rhythm in visible speech. Language Log 12/18/2013, https://languagelog.ldc.upenn.edu/nll/?p=9159.

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

Liss, Julie M., Sue LeGendre & Andrew J. Lotto. 2010. Discriminating dysarthria type from envelope modulation spectra. *Journal of Speech, Language and Hearing Research* 53(5), 1246–1255.
Ludusan, Bogdan, Antonio Origlia & Francesco Cutugno. 2011. On the use of the rhythmogram for automatic syllabic prominence detection. *Interspeech 2011*, Florence, 2413–2416.
Ludusan, Bogdan & Petra Wagner. 2020. Speech, laughter and everything in between: A modulation spectrum-based analysis. In Nobuaki Minematsu (ed.), *10th International Conference on Speech Prosody*, Online, 995–999.
Malisz, Zofia, Michael O'Dell, Tommi Nieminen & Petra Wagner. 2016. Perspectives on speech timing: Coupled oscillator modeling of Polish and Finnish. *Phonetica* 73(3–4), 229–255.
Malisz, Zofia & Petra Wagner. 2012. Acoustic-phonetic realisation of Polish syllable prominence: A corpus study of spontaneous speech. In Gibbon et al. (eds.), 105–114.
Nolan, Francis & Hae-Sung Jeon. 2014. Speech rhythm: A metaphor? In Rachel Smith, Tamara Rathcke, Fred Cummins, Katie Overy & Sophie Scott (eds.), *Communicative rhythms in brain and behaviour*: Theme issue of *Philosophical Transactions of the Royal Society, London. B Biological Sciences*, 1–11.
O'Dell, Michael L. & Tommi Nieminen. 1999. Coupled oscillator model of speech Rhythm. *14th International Congress of Phonetic Sciences* (ICPhS XIV), San Francisco, CA, 1075–1078.
Oppenheim, Alan V., Alan S. Willsky & Ian T. Young. 1983. *Signals and systems*. London: Prentice Hall.
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.
Pike, Kenneth L. 1945. *The intonation of American English* (University of Michigan Publications: Linguistics, vol. 1). Ann Arbor, MI: University of Michigan Press.
Pike, Kenneth L. 1959. *Language as particle, wave, and field* (Texas Quarterly 2(2)).
Poser, William J. 1984. *Phonetics and phonology of tone in Japanese*. Ph.D. dissertation, MIT.
Potter, Ralph K., George A. Kopp & Harriet C. Green. 1947. *Visible speech* (Bell Telephone Laboratories Series). New York: D. Van Nostrand.
Prince, Alan & Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell.
Ramus, Franck, Marina Nespor & 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 David Crystal (ed.), *Linguistic controversies: Essays in linguistic theory and practice*, 73–79. London: Edward Arnold.
Rossini, Nicla & Dafydd Gibbon. 2011. Why gesture without speech but not talk without gesture? *GESPIN 2011*, Bielefeld.
Sagisaka, Yoshinori. 2003. Modeling and perception of temporal characteristics in speech. *15th International Congress of Phonetic Sciences* (ICPhS XV), Barcelona.
Scott, Donia R., Stephen D. Isard & Bénédicte de Boysson-Bardies. 1985. Perceptual isochrony in English and French. *Journal of Phonetics* 13, 155–162.
Selkirk, Elizabeth. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.
Stevens, Kenneth N. 1998. *Acoustic phonetics*. Cambridge MA: MIT Press.
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 & Amalia Arvaniti. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America* 134, 628–639.
Tilsen Samuel & Keith Johnson. 2008. Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America* 124(2), EL34–EL39. [PubMed: 18681499]
Todd, Neil P. McAngus & Guy J. Brown. 1994. A computational model of prosody perception. *International Conference on Spoken Language Processing* (ICLSP–94), Yokohama, 127–130.

Tortel, Anne & Daniel Hirst. 2008. Rhythm and rhythmic variation in British English: Subjective and objective evaluation of French and native speakers. In Plínio A. Barbosa, Sandra Madureira & Cesar Reis (eds.), 4th International Conference on Speech Prosody, Campinas, 359–262.

Traunmüller, Hartmut. 1994. Conventional, biological, and environmental factors in speech communication: A modulation theory. In Mats Dufberg & Olle Engstrand (eds.), *PERILUS XVIII: Experiments in Speech Process*, 1–19. Stockholm: Department of Linguistics, Stockholm University. [Also in *Phonetica* 51, 170–183 (1994).]

Varnet, Léo, Maria Clemencia Ortiz-Barajas, Ramón Guevara Erra, Judit Gervain & Christian Lorenzi. 2017. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America* 142(4), 1976–1989.

Wagner, Petra. 2007. Visualizing levels of rhythmic organisation. *16th International Congress of Phonetic Sciences* (ICPhS XVI), Saarbrücken, 1113–1116.

Wang, Avery Li-Chun. 2003. An industrial-strength audio search algorithm. *International Symposium on Music Information Retrieval* (ISMIR), Baltimore, MD.

Wayland, Ratree & Takeshi Nozawa. 2020. Calibrating rhythms in L1 Japanese and Japanese accented English. *Meetings on Acoustics* 178 ASA 39(1), San Diego, CA, 1–13.

White, Laurence & Zofia Malisz. 2020. Speech rhythm and timing. In Carlos Gussenhoven & Aoju Chen (eds.), *The Oxford handbook of language prosody*, 167–182. Oxford: Oxford University Press.

Yu, Jue & Dafydd Gibbon. 2015. How natural is Chinese L2 English prosody? *18th International Congress of Phonetic Sciences* (ICPhS XVIII), Glasgow, 145–149.