New Perspectives on Ibibio Speech Rhythm

Dafydd Gibbon

Abstract The goals of documenting and describing languages, whether endangered or widely used, are far-ranging, from preserving the inherited wisdom of the language community to understanding the spectrum of structures and communication events the human mind is capable of dealing with. Part of this spectrum covers the relation of language to other forms of communication, and one property of this subspectrum is the specific rhythm patterns of speech which characterises a language community, along with other regular events in daily life and in culture, such as walking and running, dance and music. The analysis of speech rhythm is placed in a broader semiotic and analytic context than is customary in 'mainstream' linguistics. Two phonetic approaches are discussed in detail, the signal processing spectral analysis method and the annotation based isochrony heuristic method, and the associated procedures are applied to Ibibio (New Benue-Congo) data with the intention of complementing existing studies of Ibibio grammar and phonology. The main emphasis is on rhythm analysis within Modulation Theory. Modulation Theory is a signal processing paradigm, and within this framework Rhythm Formant Theory and its associated Rhythm Formant Analysis (RFA) method are introduced. The modulation theoretic approach contrasts methodologically with the second approach, the irregularity or isochrony heuristic, for deriving an irregularity or variability index of relative isochrony (equal timing) from the timestamps of speech annotations. Results derived from both approaches for the Ibibio data show that rhythm types are not static numbers or patterns which are valid for entire languages, but vary dynamically over time during utterances and between utterances in a corpus. Based on these results, a hierarchical clustering procedure was applied in order to investigate the typological status of Ibibio. The study is designed to be exploratory, with the goal of providing a theoretical basis for future more detailed quantitative studies of the rhythms of Ibibio and its neighbouring languages, and their relation to other languages in Nigeria and beyond.

D. Gibbon (✉)
Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany
e-mail: gibbon@uni-bielefeld.de

M. E. Ekpenyong and I. I. Udoh (eds.), *Current Issues in Descriptive Linguistics and Digital Humanities*, https://doi.org/10.1007/978-981-19-2932-8_34

Keywords Modulation theory · Prosody · Rhythm formant Analysis · Speech · Low frequency spectrum

1 Documenting and Describing Speech Prosody[1]

1.1 Language Documentation and Cultural Identity

The goals of documenting and describing languages, whether endangered or widely used, are far-ranging, from preserving and making available the inherited wisdom of the language community to understanding the range of structures and communication events the human mind is capable of dealing with, and thereby providing a frame of reference for specifying the identity of particular languages and speech communities in relation to each other. This is the task of the description of language in context using linguistic and phonetic methods. Both documentation of the language heritage of a community and linguistic and phonetic analysis of this heritage are essential prerequisites for deepening awareness of identity in a speech community. The present study contributes a preliminary exploratory pilot study on methods for increasing understanding of the Ibibio (New Benue-Congo) language and its neighbours.

Part of the typological spectrum of language variation covers the relation of language to other cultural forms of communication, and one property of this sub-spectrum is *speech prosody*: the rhythms and melodies of speech. Speech in every language and every language variety is characterised by low frequency rhythms with between about 3 beats per second and three seconds per beat, in other words around 1 Hz, from about 0.3 Hz to about 3 Hz. These basic rhythms of speech provide the temporal mould for the melodic patterns of spoken language. It might be thought that in documenting and describing a phonemic and morphemic tone language like Ibibio the main function of prosody would lie primarily in the area of tone, secondarily in the area of intonation, with little attention to rhythm (but cf. Gut et al., 2002).

Focus on speech rhythm, part of the cultural identity of a community, the heartbeat of a society, is the perspective of the present study, and its importance is justified in the following way. Each member of a language community has a repertoire of language varieties—dialect, sociolect, task-oriented register—which determines both rhythm and melody, in partial independence of the words. The use of these varieties may be changed by a speaker at any time, resulting in superimposed long-term variation in rhythm and melody. But there seems to be a general principle which specifies a frequency range around 1 Hz for basic speech rhythms in all languages and cultures, regardless of the actual details of the languages. The rhythms of the syllables and words of speech also fall into the same low frequency range above and

[1] This study is dedicated to my dear friend and distinguished colleague Prof. Eno-Abasi Essien Urua, University of Uyo, Akwa Ibom State, Nigeria, and of course to her outstanding interdisciplinary team of experts in language theory, description, documentation and preservation.

below 1 Hz as dance movements and musical beats, as well as the more mundane rhythms of chewing, heartbeats, walking and running, arm movements, as well as animal and bird calls, and even bouncing objects, swaying branches and other pendulum effects—between a few 'beats' per second and a 'beat' every few seconds. There are many studies of the rhythms of the brain which relate to the rhythms of speech and other physiological rhythms (cf. Poeppel & Assaneo, 2020).

Very much like music, speech combines rhythms and melodies for specific functional effects. Speech prosody is different in detail in every language, depending on syllable and word structures, phrase, sentence and also the interactive and rhetorical discourse structures of the language, as well as attitudinal and emotional functions (Kohler, 2018). One basic function is *metalocutionary*, that is, rhythm conveys information about the structure of the words and phrases in the locution with which it cooccurs. More specifically, the function is *metadeictic*: boundary tones and accents point to the times at which words and phrases occur and thus, semiotically, have an *indexical* function, like co-expressive gestures (McNeill, 2005). In this respect, rhythms are arguably the most basic properties of speech prosody and indeed of speech: the regular beats or low frequency oscillations of speech serve as a framework for the melodies and the locutions, as do the beats of music, which organise the melodic flow, and the steps in dancing, upon which more fluid movements build.

1.2 Objectives and Overview

The present contribution differs from many earlier linguistic phonetic approaches to the analysis and explanation of rhythm by advocating and following a dual strategy with two independent but mutually complementary approaches:

1. an inductive approach with signal analysis of recorded speech with a demodulation model for AM (amplitude modulation) information in terms of the amplitude envelope of speech, and a demodulation model for FM (frequency modulation) information in terms of fundamental frequency (F0) estimation (also known as 'pitch tracking'), and, following demodulation, low frequency spectral analysis of the resulting AM envelope and FM F0 track;
2. a deductive approach with the description of speech regularities and irregularities as a partial account of some properties of rhythm, with pre-analysed linguistic categories which are aligned with speech signals by means of phonetic transcription and annotation in order to validate the linguistic and phonetic significance of the acoustic analysis.

The first approach aims to provide a basis for more detailed quantitative studies and thereby counteract the widely held pessimistic view that rhythm is a purely top-down cognitive construct with no consistent physical correlate. The second provides explicit quantitative empirical grounding for grammatical, semantic and pragmatic studies of rhythm. Phonological and other linguistic studies of rhythm use a qualitative or hermeneutic methodology, based on intuitive impressions of rhythms and their components and are not considered in this study.

The overriding objective of the present study is to complement the detailed analyses of Ibibio grammar by Essien (1990), of Ibibio phonology and prosody by Urua (2000) and of speech and language technology development for Ibibio (cf. Ekpenyong, 2012; Gibbon, 2002; Gibbon & Urua, 2006). The study updates the quantitative analysis of syllable irregularity as a component of rhythm in annotated Ibibio speech (Gut, 2002; Gut et al., 2002; Gut & Gibbon, 2002) with a quantitative signal-based account of Ibibio rhythm. There are two reasons for specifically concentrating on rhythm in Ibibio:

1. analysing speech rhythm provides a methodological and substantive bridge to ethnological studies on music and dance in the cultures of the Ibibio speech community and its neighbours, if required;
2. description of Ibibio as an actively used language provides a point of orientation for capturing typological similarities and differences not only in relation to other languages of South-East Nigeria and Cameroon, but to Nigerian languages in general and beyond.

After the introduction in Sect. 1, Sect. 2 is concerned with the theoretical background and data used in the study, and Sect. 3 applies the bottom-up modulation theoretic signal processing concepts of rhythm spectrum and spectrogram to Ibibio. Section 4 places Ibibio into the context of quantitative phonetic typology in relation to other languages using quantitative distance metrics. In Sect. 5, the top-down linguistic-phonetic methodology of annotation analysis is used as a partial validation method for the signal processing results. Section 6 provides a summary and conclusions together with an outlook for further research using the Modulation Theory paradigm and Rhythm Formant Theory.

2 Rhythm Formant Theory

2.1 Rhythm Formants: AM and FM

In the empirical domain of phonetics three main temporal phases can be distinguished (leaving neural phases out of consideration): production, transmission and perception. There are well-known phonetic models of speech production, for example the source-filter theory of sound generation, and of speech perception, for example the theory of spectral transformation in the cochlea and cognitive filtering. In practical work in phonetics, speech transmission is perhaps the most treated domain, especially with software such as Praat (Boersma, 2001), but, strangely, it is not usually presented with an acoustic theory of transmission, though the nuts and bolts of fundamental frequency, harmonics and phone formants are of course described.

An appropriate framework for acoustic phonetics is provided by Modulation Theory, and in order to bridge this theory gap, modulation theoretic concepts are introduced explicitly in the present contribution. Modulation Theory is the standard engineering framework for speech transmission and reception, which describes the modulation of a carrier signal amplitude or frequency with an information carrying lower frequency signal. Modulation Theory is mainly used in signal processing at radio frequencies, where AM (amplitude modulation) and FM (frequency modulation) broadcasting is well known, but the theory also applies at the audio frequencies of speech.

Modulation theory was arguably first applied to speech modelling by Dudley (1939). The present approach stands in the line of modulation theoretic analyses of rhythm, with concepts such as *rhythmogram*, *rhythm spectrum* and *complex rhythm composition* (pioneered by Todd et al., 1994; Cummins & Port, 1998; O'Dell & Nieminen, 1999; Barbosa, 2002; Galves et al., 2002; Tilsen & Johnson, 2008 and applied by numerous scholars in later studies). A somewhat different approach was taken by Traunmüller (1994), with a functional definition of the carrier and its modulations. The modulation theoretic approach is extensively discussed by Gibbon (2021) with the novel concepts of *rhythm spectrogram* and *rhythm formant*, in the framework of Rhythm Formant Theory (RFT) and its associated method of Rhythm Formant Analysis (RFA). Rhythm formants are high magnitude low frequency zones in the spectra of the amplitude and frequency modulation of the speech signal. The terminology may appear unusual, but this acoustic definition of 'formant' is exactly the same as the acoustic definition used for high frequency phone formants. The physiological bases in speech production are different, of course.

The modulation theoretic approach to speech transmission accounts for the way in which audio frequency information is transported by a speech signal, which is in principle also the way in which audio frequency information and data packets are transported by a radio frequency signal over mobile phones. There are many kinds of signal modulation, but the two which are most relevant to the present discussion are *amplitude modulation* (AM) and *frequency modulation* (FM). In AM, the amplitude of the *carrier signal* is varied by the amplitude of the *modulation signal* or *information signal*, for example by the changing amplitudes of the consonants and vowels which form the sonority curve of syllables and words. In FM, the frequency of the carrier signal is varied by the amplitude of the modulation signal, for example by tones, pitch accents and intonation.

For acoustic rhythm analysis, the task is to demodulate the low frequency (LF) components of the modulating signal as rhythm formants, with frequencies associated with the repetition rate of syllables, words, phrases or longer units in the case of AM, and of tones, pitch accents and longer prosodic and discourse units in the case of FM. The demodulation process recovers the audio information from the speech carrier signal. In order to identify the detailed information in the demodulated signal, spectral analysis is applied to both the AM envelope and the FM contour of the entire utterance with the Fast Fourier Transform (FFT), and from the resulting spectrum to extract the low frequencies, $f < 10$ Hz, which constitute rhythms.

2.2 *Rhythm in Modulation Theory*

A general definition of Formant Theory states that a formant is a narrow high magnitude frequency zone in the spectrum of a signal, whether speech or music, and that a formant defines specific identifiable characteristics of a speech signal or the tone of a musical instrument. A narrow high magnitude frequency zone with this property is interpreted as a *formant*. High frequency zones above about 300 Hz represent the formants of speech sounds, particularly resonant harmonic sounds such as vowels, and are generated by shapes of the oral and nasal tracts. Low frequency zones below about 10 Hz are interpreted as rhythm formants. Rhythm formants are generated by relatively regular sequences of syllables (containing alternating consonants and vowels), and of words (containing alternating strong, maybe stressed syllables), and also of longer units with alternating strengths, such as phrases, sentences and texts. The low frequency formants can be identified by analysing either the AM modulations or the FM modulations of speech, or both.

The low frequency modulations of speech, both in AM and FM, can be placed in a more general modulation theoretic account of speech, with a Speech Modulation Scale (SMS) in which low frequency rhythm modulation is distinguished from mid-frequency tone and intonation modulation and from high frequency harmonic and phone formant modulation (Fig. 1).

2.2.1 Rhythm Formants and Amplitude Modulation

The AM properties of speech rhythm depend primarily on the phonological and morphological typology of the language: on the phonological complexity of syllables and words, and the morphological complexity of words.

Figure 2 shows the AM envelope, superimposed on the waveform, for Ibibio, Mandarin Chinese and English. Visual inspection shows the considerable differences between disyllabic patterns in Ibibio after ‘kèèd’ but monosyllabic patterns in Chinese, and both have less complex syllables than English (most conspicuously for ‘six’ and a disyllable ‘seven’). It is intuitively clear that in terms of syllable

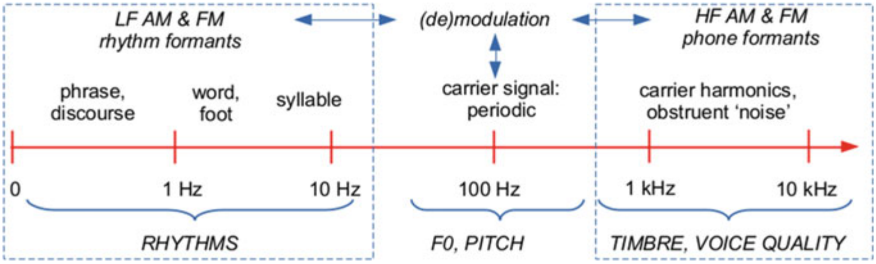

Fig. 1 Speech modulation scale (SMS)

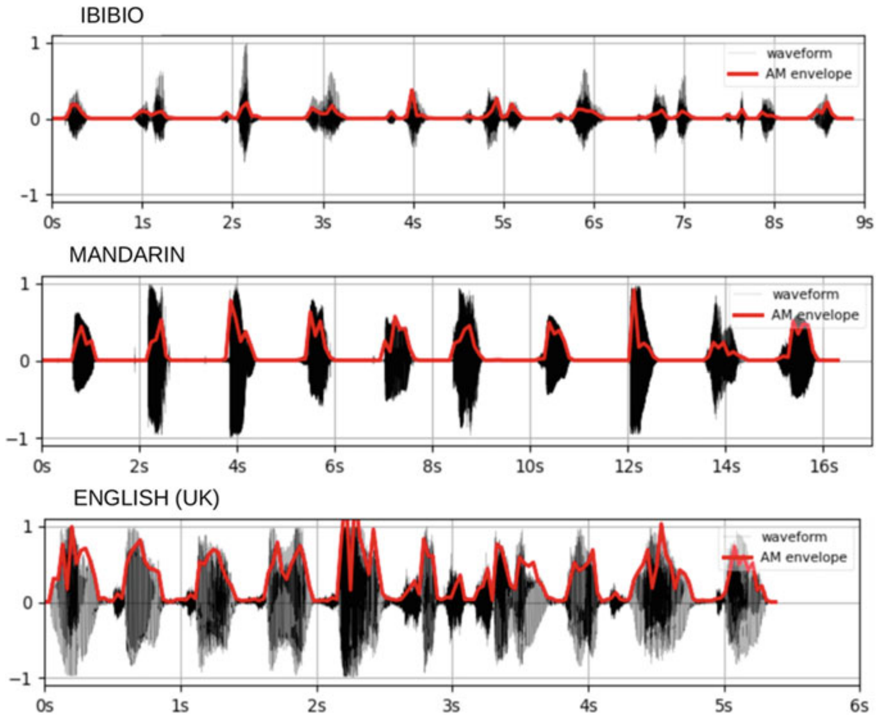

Fig. 2 AM envelope contour superimposed on waveform, counting from one to ten in Ibibio (kèèd, ìbà, ìtá, ìnààñ, ìtíòn, ìtíòkèèd, ìtíàbà., ìtíàìtá, ùsúkkèèd, dùòp), Mandarin (yī, èr, sān, sì, wǔ, liù, qī, bā, jiǔ, shí) and English. X: time, Y: amplitude

patterning in these examples, Ibibio is more complex than Chinese (in having disyllabic words) but less complex than English (mainly in having less complex syllables). It may be predicted, therefore, that quantitative analysis of differences will reflect this pattern.

More generally, there are several AM rhythm strata, which can be related fairly directly, though not deterministically, to the prosodic hierarchy of prosodic phonologies. The relation is not deterministic because rhythms vary with speech style (formality) and speech register (task orientation) as well as with speaker-specific idiosyncrasies (Arvaniti, 2009; Gut, 2012).

The first rhythm stratum is syllable-based, and is constituted by rhythm zones around 0.3 Hz: the amplitude of a syllable is lower at consonantal boundaries than at the vocalic centre, an alternating pattern of around three beats per second, which underlies the most basic kind of speech rhythm.

The second rhythm stratum is word-based, and is constituted by rhythm zones around 1 Hz, mainly in languages with inflected and derived words, in which some syllables in the word are stronger (longer, less reduced) than others, and some of these are stressed (positions in the word which in the default case are marked by pitch accents). Languages with affixal inflection, like Ibibio and English or German, have rhythm patterns which are different from the rhythms of non-affixal languages like Chinese, due to the greater complexity of syllables, together with derivational and inflecting morphology and affix reduction, as in English (as in ‘restrained’) or German (a in ‘beschränkt’, limited) as well as different melodic patterns.

The third and further rhythm strata are determined by the information structure of phrases, sentences, texts, and by dialogue interactions.

From a typological point of view, the position of Ibibio may be triangulated with reference to morphological properties of these languages (Essien, 1990; Urua, 2000):

1. Ibibio has prefixal and suffixal morphology like English and German with potential for syllable reduction, but also tonal morphology, unlike English, German and Chinese, as well as contrastive lexical tone like Chinese. The tonal morphology is inflectional, partly also as a constituent in compound word formation.
2. Languages such as English and German have suffixal inflectional morphology, prefixal and suffixal derivational morphology, as well as compounding; lexical prosody in these languages is restricted to stress positions which are sometimes contrastive, and serve as anchor points for pitch accents whose shape is determined by non-lexical factors.
3. Languages like Chinese have compounding and contrastive lexical tone, but no affixal inflectional and derivational morphology and no morphological tone.

Ibibio therefore occupies an intermediate position between languages with affixal morphology (and only traces of prosodic morphology, for instance in the stress system of English and German), on the one hand, and languages with no affixal morphology but contrastive lexical tone like Chinese, on the other hand. Most of the present study deals with the comparison of rhythms in Ibibio speech with each of these two types of language, based on the prediction that Ibibio will be closer to English and German than to Chinese in terms of AM rhythms. Rhythm also relates to melody, and therefore the present study compares not only rhythmic AM but also rhythmic FM properties of Ibibio with other languages.

2.2.2 Rhythm Formants and Frequency Modulation

Languages differ considerably in their use of lexical and phrasal frequency modulation. To illustrate the typological differences in FM patterning, the FM demodulation (F0 estimation) of recordings from Ibibio, Mandarin Chinese and English is shown in Fig. 3. The data consist of spontaneous counting from one to ten in each language; the data are not read aloud, but counting is a highly regularised register in any language, and to some extent comparable with read speech.

Each speaker chose to count at a different tempo (with durations of 9s, 16 s, 6 s) so the time axis is linearly scaled to permit direct visual comparison, as in Fig. 3. It is easy to acquire an initial impression of melodic differences between the languages, even if one does not know the languages.

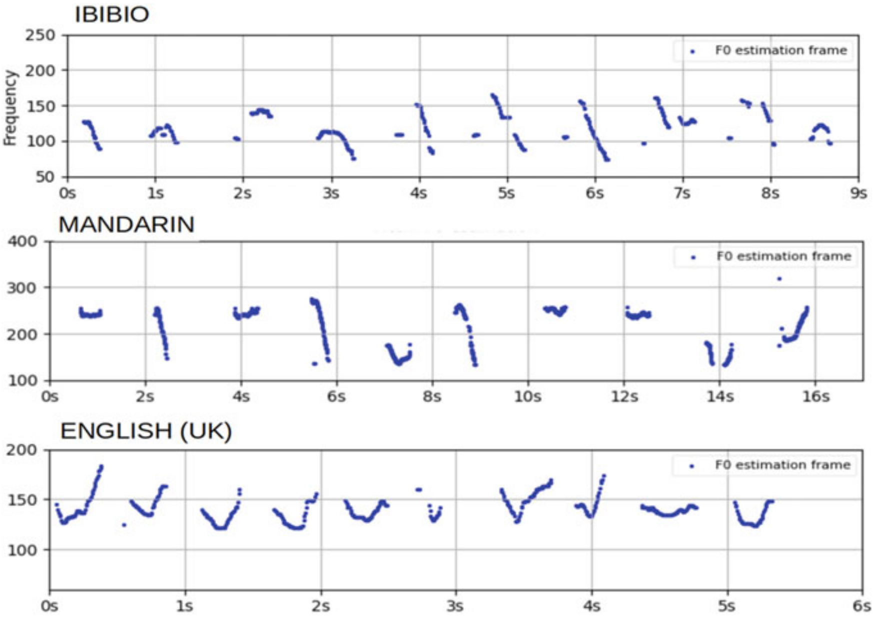

Fig. 3 F0 (fundamental frequency), counting from one to ten in Ibibio (kèèd, ìbà, ìtá, ìnààñ, ìtíòn, ìtíòkèèd, ìtíàbà., ìtíàìtá, ùsúkkèèd, dùòp), Mandarin (yī, èr, sān, sì, wǔ, liù, qī, bā, jiǔ, shí) and English. X: time, Y: frequency (Hz)

Visual inspection shows that in English, in contrast to Chinese and Ibibio, the 'tones' in the sequence tend to remain the same throughout: they are actually pitch accent markers of stress positions in the sequence (Liberman, 1975; Jassem & Gibbon, 1980; Pierrehumbert, 1980), and are therefore referred to here as *stress-pitch accents*. The actual shapes of the pitch accents vary in different dialects and registers and with different discourse functions.

Although superficially like tones, the English stress-pitch accents do not distinguish lexical words (except in contrastively focussed contexts) and can therefore in principle vary freely, independently of lexical constraints. In actuality, however, they do not vary freely but follow a pattern of rhythmical iteration, following a path of least resistance: once a pitch pattern has been selected, the same pitch pattern is repeated throughout the sequence with minor variation in detail.

These sequences of repeated stress-pitch accents have been characterised for a over century in English pronunciation textbooks as the 'body' of an intonation group, and more recently they have been functionally explained as creating an expectation to be resolved by the end of the group (Dilley, 1997). Another explanation could also be formulated as an economy or simplicity constraint: *Don't change the stress-pitch accent pattern unless necessary*.

The pattern could also be treated as being based on a generalised OC (obligatory contour) constraint: the pattern is assigned once to the whole body of the contour, and spreads as required as syllables appear later. In optimality theoretic terms, this would be a kind of sequential *faithfulness constraint*. This major difference between English and both Ibibio and Mandarin has major consequences for the contribution of speech melody to speech rhythm. The tones of Ibibio and Mandarin, on the other hand, tend to be variable from word to word and do not follow this constraint.

This is not the whole story, however: a typologically different morphological feature of Ibibio is involved, which distinguishes Ibibio from both Chinese and English. Ibibio counting is partly based on a quinary (base five) numeration system, unlike the decimal systems of English and Chinese, or older European duodecimal ('dozen', i.e. 12; 'gross', i.e. 12 × 12, 144) and vigesimal ('score', i.e. 20, e.g. 'three score and ten' for 70, the biblical life-span). This means that in Ibibio, the numbers above five are morphologically complex, sharing the same prefix, derived from the word for five. The patterns for numbers above five in Ibibio indeed share an iterative tonal property, to some extent resembling English, but for an entirely different reason: the iterative effect is due to similarity in morphological composition.

The triangulation of prosodic patterning between different languages is discussed further in the following sections. The next section outlines the field of rhythm analysis in frequency domain and time domain contexts in preparation for this discussion.

3 Rhythm in Ibibio—AM, FM and Rhythm Formants

3.1 Data Types and Overview

Two types of Ibibio data are used, in addition to the counting data. One data type consists of recordings of readings in Ibibio (ISO 639-3 *ibb*), which were made during a cooperative research and development project for text-to-speech synthesis in Ibibio, between the universities of Uyo, Nigeria and Bielefeld, Germany. The second type is a reading of the Ibibio translation *Áfîm Édèré yè Útín* of the traditional Greek fable attributed to Aesop, *The North Wind and the Sun*, published together with an audio recording in the *Illustrations of the IPA* series (Urua, 2004). Both data types were recorded in reading aloud scenarios, which may not be as varied as conversational scenarios, but have their own justification as necessary communication registers in education, business, politics and religion, and are in any case sufficiently complex to be of scientific interest in their own right. The Ibibio fable reading is supplemented with an opportunistic selection of readings of translations of the fable into other languages for language comparison: four Scottish English readings, six readings by a bilingual, three in English and three in German, and five readings in Mandarin Chinese.

3.2 The Low Frequency Rhythm Spectrum

Figure 4 shows a visualisation of low frequency spectral properties, in an analysis of the reading of the Ibibio sentence *ke ŋkpɔ nte isua tɔsɪn keet je ikie usʌkkeet je anaaŋ je efɪteta sɪm isua ekeboijoke do ɔnɔ mbon ŋwo ke itie ɔkɔk*. The upper left panel shows the waveform with superimposed positive AM envelope; the upper right panel shows the low frequency spectral analysis of the envelope contour. The lower left panel shows the FM (F0) contour, and the lower right panel shows the low frequency spectral analysis of the FM contour.

The rhythm formants are detected by identifying magnitude peaks in different frequency zones in the spectrum; for this purpose, the first 15 frequency peaks are marked by vertical lines, and used in the later analysis.

Visual inspection of the waveform and application of a fuzzy rule of thumb reveals that approximately 44 syllable-sized units, divided by approximately 11 s yield 4 syll/s, i.e., 4 Hz, with the durations of these units therefore averaging about 200 ms. If the phrases are counted, a phrasal rhythm can also be tentatively identified: 6 phrases divided by 11 s yields approximately 0.5 phrases/sec, i.e., 0.5 Hz, or one phrase every 2 s. A more exact analysis can be performed by annotation, i.e., assigning timestamps and transcription labels to the signal, for example with Praat (Boersma, 2001).

The AM envelope spectrum (Fig. 4, upper right), confirms these rough rule of thumb calculations. There are three main groups which cluster along the scale, indicating strata in a rhythm hierarchy, not a single rhythm:

1. between 3.8 and 4.6 Hz, corresponding to the syllable rate;
2. between 2.3 and 3.2 Hz, along with frequencies around 2 Hz, corresponding to word rates;
3. at a distance from these, frequencies around 0.5 Hz, corresponding to the iteration rates of longer units such as phrases and sentences.

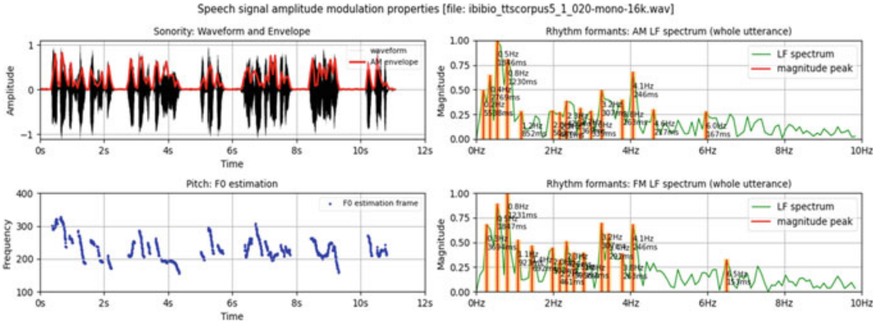

Fig. 4 'ke ŋkpɔ nte isua tɔsɪn keet je ikie usʌkkeet je anaaŋ je efɪteta sɪm isua ekeboijoke do ɔnɔ mbon ŋwo ke itie ɔkɔk'. Upper left: waveform and AM envelope. Upper right: long-term spectrum of the AM envelope. Lower left: F0 estimation or FM envelope. Lower right: long-term spectrum of the FM envelope

These frequency zones are interpreted as rhythm formants. There is some fuzziness in identifying the frequency zones: human speech is not produced by clockwork, and there are many individual and social factors which introduce variation into the rhythms.

Typologically, the LF spectrum shape in Ibibio resembles that of English and German and contrasts sharply with that of Mandarin Chinese (Gibbon, 2021), indicating the presence of prosodic feet or stress groups in addition to syllable tone, corresponding to the morphophonological properties of Ibibio which were outlined in Sect. 1.

3.3 Rhythm Variability and the Rhythm Spectrogram

The low frequency spectrum provides the fundamental information about the distribution of the frequency clusters which characterise speech rhythm formants. The spectrum shows rhythms across the whole of the utterance, but it is clear from the spectrum that the magnitude of the rhythm formants, thus the salience of the rhythms, varies. There is a limitation to what can be interpreted in the spectrum, because the spectrum provides a holistic static picture of all frequencies in the entire utterance together, but no temporal information about where (or, more precisely: when) different frequencies occur and change during the utterance.

In order to provide temporal information, it is necessary to extend the modulation theoretic rhythm spectrum toolkit to include a new analytic tool which maps the spectrum to the time domain in smaller steps: the rhythm spectrogram (cf. Fig. 5, lower panel).

Figure 5 shows the waveform and AM envelope in the upper panel, the standard HF spectrogram in the centre panel and the LF spectrogram in the lower panel. The AM envelope is superimposed on the spectrograms in these panels in order to facilitate identification of segments of speech. The HF spectrogram clearly shows the harmonics of F0, as expected, with the formants as strong harmonics while the harmonics which do not carry formants are weak or not present.

The LF spectrogram also shows rhythm formants as dark bars, varying mainly around 4 Hz (roughly: strong syllables), but with faster stretches around 6 Hz, relating to shorter syllables, and at the end a very strong rhythm of less than 1 Hz, relating to syllable groups such as long words or phrases. An HF spectrogram has a far more granular temporal structure, because the individual measurement window is less than 10 ms in duration, while the LF spectrogram has measurement windows of 2 s, with short steps between frames, in order to capture the slower rhythms.

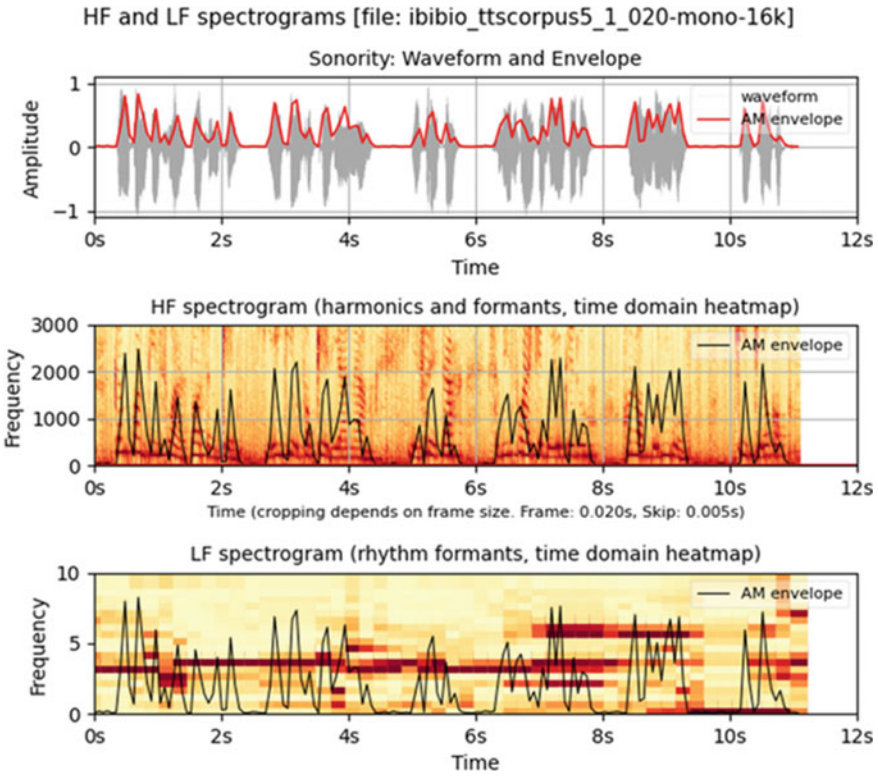

Fig. 5 ‘ke ŋkpɔ nte isua tɔsɪn keet je ikie usʌkkeet je anaaŋ je efɪteta sɪm isua ekeboijoke do ɔnɔ mbon ŋwo ke itie ɔkɔk’. HF and LF spectrograms, with superimposed AM envelopes

Measurement of rhythms and rhythm changes in an utterance is not the end of the story. A number of questions about the function of rhythm changes arise:

1. What are the functional factors underlying the change of rhythms during an utterance?
2. Are changes in rhythm—possibly ‘rhythms of rhythm’—characteristic of particular language types?
3. Do rhythms change with speech style or rhetorical strategy?
4. Are the factors underlying rhetorical or stylistic rhythm changes, and the patterns of rhythm change, shared by all languages?

These are not easy questions to answer; they are tantalising issues which require extensive empirical study, which goes beyond the scope of the present paper and are open to further research, for example in Interactional Linguistics (Couper-Kuhlen and Selting, 2018).

3.4 Rhythm and Narrative Style: Aspects of Story Analysis

A promising perspective on long-term rhythm variation involves the analysis of much longer utterances, for example stories. As a rule, stories have recognisable episodes, and different episodes may involve different rhetorical strategies. Figure 6 provides a global view of the waveform and AM envelope, the FM (F0) pattern and the LF spectrogram for a reading of the entire story *Áfîm Édèré yè Útín* (Urua, 2004), with a total duration of just under 70 s.

Figure 6 shows a number of specific properties of long-term rhythm variation, motivated by narrative patterns, for example:

1. The waveform (upper panel) shows the episodes in the story quite clearly with pauses and interpausal units.
2. The F0 track (mid panel) shows the episodes even more clearly in terms of the downtrends (Connell, 2002), i.e., downdrift or automatic downstep patterns (tone terracing) and the H–L alternations of each phrase.

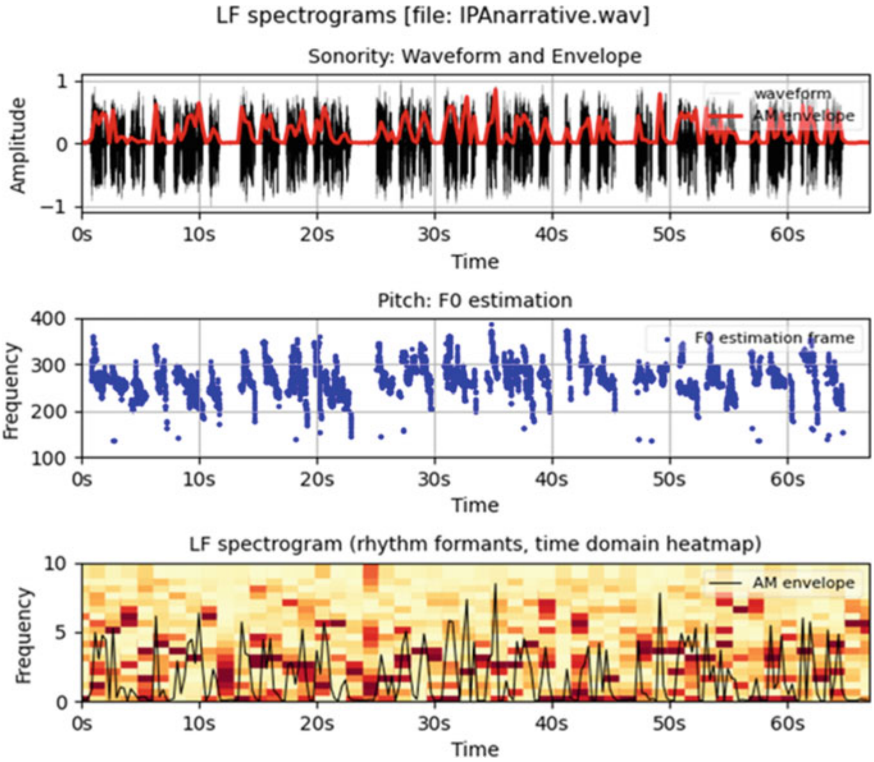

Fig. 6 HF and LF spectrograms of the Ibibio narrative "Áfîm Édèré yè Útín"

3. The F0 estimation track (centre panel) shows short-term frequency changes between syllables, a medium-term downtrend due to terracing (the high-low tone interval is greater than the low–high tone interval), as well as long-term raising and lowering in the course of the utterance, with the higher frequencies continuing from approximately 25 s, corresponding to episodes in the story.
4. The LF spectrogram (lower panel) shows a large number of changes in rhythm frequencies between the different episodes of the story.

Detailed quantitative structural and interpretative functional analyses are needed in order to relate pauses, fundamental frequency patterns and rhythm changes to text episodes. At first glance, there appears to be a regular pattern of terrace iteration, a cycle which represents ‘rhythms of rhythm’ (Gibbon, 2021; Gibbon and Li, 2019).

A semantic-pragmatic interpretation of the temporal prosodic patterns of rhythm variation goes far beyond the scope of the present study, and is the domain of Ibibio text and discourse analysis in terms of narration development, argumentation, speaker attitude and emotion, and audience appeal. Since the story under discussion is a translation into Ibibio from an English translation from an Greek original, the present discussion can only serve as a pointer: the real interest of this method of rhythm analysis will lie in the description of authentic Ibibio stories and other styles, registers and genres.

The following section continues the spectrographic analyses with a comparison of Ibibio with an opportunistic selection of other languages, as a first step in quantitative rhythm typology.

4 Ibibio Rhythm Compared to Rhythm in English and Chinese

4.1 Distance Metrics and Similarity Relations

The detailed information which LF spectral analysis makes available can be deployed to compare Ibibio rhythm with the rhythms of other languages. It was already suggested in Sect. 2 that Ibibio may relate more closely, as a language with agglutinative morphology, to inflecting languages like English and German in terms of AM, but, as a tone language, more closely to Mandarin Chinese in terms of FM. For each of the languages in the corpus described in the introduction, vectors with the 15 highest magnitude frequencies from both the AM and FM spectra were extracted (as shown in Fig. 4) and further investigated. Two methods are used to compare the readings: first, distance metrics and distance networks; second, hierarchical clustering based on the distance network results.

Distance metrics (Gibbon, 2016, 2021) provide a convenient method for measuring the relative similarity or dissimilarity between ordered sets. The prediction is, therefore, that Ibibio is closer to German or English in terms of AM low frequency spectral patterns, but closer to Chinese in terms of FM low frequency spectral patterns.

A specimen analysis using spectral peak magnitude vectors is shown in Table 1. Different distance metrics define slightly different orderings, but all metrics confirm the predicted tendency: AM shows Ibibio tending to be closer to English and German, FM on the other hand tends to show proximity to Mandarin. Since many factors underlie rhythm, the proximity may signify discourse style similarities at least as much as structural typological similarities; this needs clarification in future research.

For example, the Chebyshev distance metric (also known as Chessboard distance), which shares properties of Manhattan distance and Euclidean distance, shows the predicted tendency particularly clearly (cf. Table 1): the nearest readings to the Ibibio reading in terms of AM are the three German readings, then comes a Mandarin Chinese outlier, followed by more English readings; this is not unexpected, since German has a more complex inflectional system than English. Seven of the German and English set are in the closest eight. The remaining Mandarin Chinese readings, along with two English outliers, are furthest from Ibibio.

In terms of FM, the closest is a Scottish English outlier, but otherwise four of the five closest readings are Mandarin Chinese, with one Mandarin Chinese outlier at a greater distance from Ibibio. The results are compatible with the predictions, but the data set is too small to claim confirmation of the predictions.

Table 1 Chebyshev distances ($0 \geq d \leq 1$) between Ibibio and other languages

Chebyshev distance from Ibibio (spectral peak magnitude vectors)			
Distance	AM	Distance	FM
0.548	RT-NWAS-Ger-01	0.488	NW084-EngScot06F
0.608	RT-NWAS-Ger-03	0.505	wangwei_F
0.623	RT-NWAS-Ger-02	0.545	liuqp_F
0.63	liangjj_F	0.552	wuxi_F
0.668	RT-NWAS-Eng-01	0.575	liangjj_F
0.673	RT-NWAS-Eng-02	0.708	NW084-EngScot04F
0.711	NW084-EngScot06F	0.77	RT-NWAS-Ger-03
0.723	NW084-EngScot04F	0.79	NW084-EngScot08F
0.734	wuxi_F	0.794	RT-NWAS-Ger-01
0.766	jiayan_F	0.814	RT-NWAS-Eng-02
0.775	RT-NWAS-Eng-03	0.824	RT-NWAS-Ger-02
0.798	NW084-EngScot08F	0.894	jiayan_F
0.897	liuqp_F	0.904	RT-NWAS-Eng-03
0.897	wangwei_F	0.948	RT-NWAS-Eng-01

4.2 Distance Networks and Hierarchical Clusters

A distance network is an abstract map with links between data objects in which similarities are represented as distances, as in Figs. 7 and 8. For the present data set, the distances were normalised to be between 0 and 1. In Fig. 7 all AM distances closer than 0.6 (minimum distance = 0, maximum distance = 1) are visualised as a distance graph. To visualise all distances would produce an unnecessarily cluttered graph.

The AM low frequency spectrum graph shown in Fig. 7 indicates that Ibibio is closer to German than to Chinese. There are many nodes which are closer together than the distance of Ibibio to its nearest neighbour at the distance level shown in the graph. In particular, the Mandarin Chinese nodes relate particularly closely to each other. The Chebyshev distances for the FM spectral peak magnitude vectors of the readings are shown in Fig. 8. In contrast to the AM visualisation, five items with distances lower than 0.6 are shown, one to the German outlier and four to Mandarin Chinese readings. The prediction that Ibibio is closer to German with respect to the AM rhythms and closer to Chinese with respect to the FM rhythms is fulfilled.

The second method for using the distance information is in a hierarchical clustering procedure: the distance metric results from the similarity network procedure are used together with additional clustering criteria as the basis for inducing a hierarchical classification of the spectral peak magnitude vectors (cf. Fig. 9).

In addition to the different distance metrics, different clustering criteria can be used, for example based on the mean or variance of clusters, or on the closest or furthest members of clusters. The groupings suggested by the distance metrics alone are adjusted according to these definitions of cluster proximity. The clustering given by using Canberra distance together with Vorhees clustering (also known as

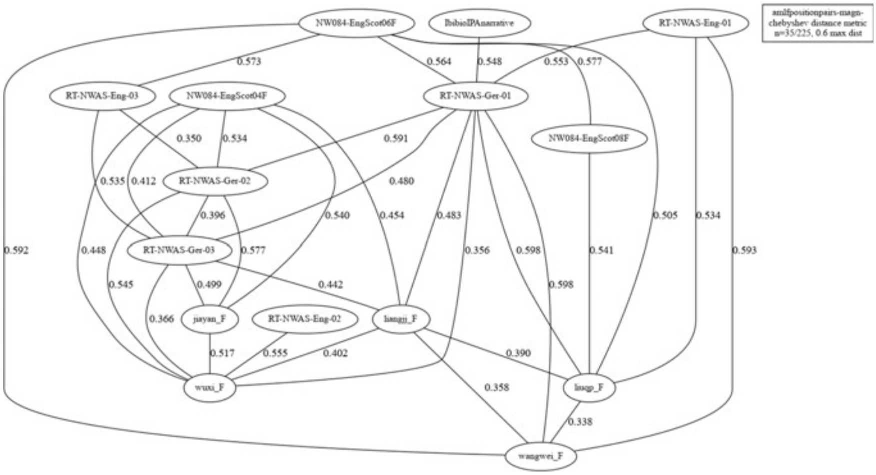

Fig. 7 Distance network with pairwise Chebyshev distances between vectors of the 15 most salient frequencies in the AM spectrum. Maximal distance criterion: 0.6

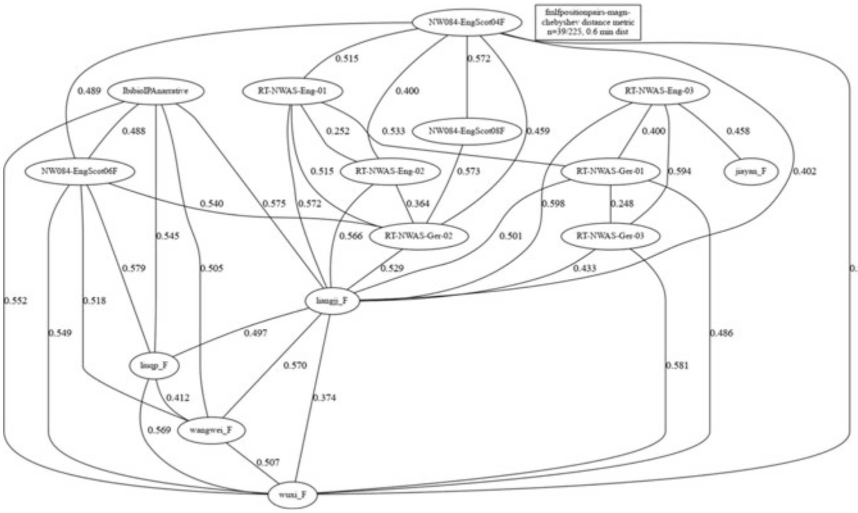

Fig. 8 Distance network with pairwise Chebyshev distances between vectors of the 15 most salient frequencies in the FM spectrum. Maximal distance critrion: 0.6

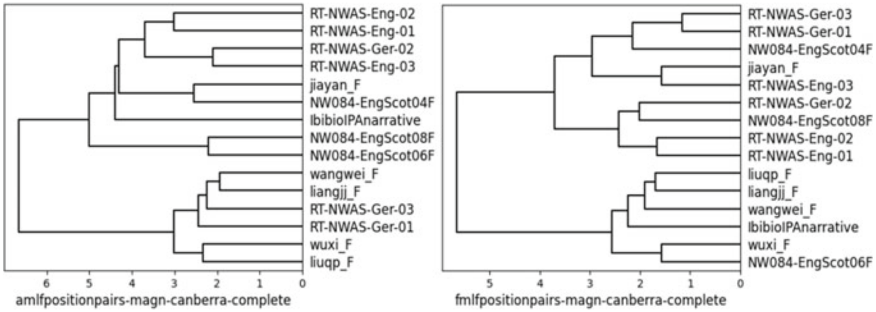

Fig. 9 Hierarchical clustering of readings of "The North Wind and the Sun" in Ibibio, English, German and Mandarin Chinese, with AM and FM LF spectrum vectors

complete or furthest point clustering) also reflects the predictions for different AM and FM clustering, though the details are different in detail from the Chebyshev distance results reported in the preceding discussion.

4.3 Discussion

This inductive method of automatic spectral analysis has a role to play in quantitative prosodic typology. Tentative though these new results may be, the method is innovative and the results suggest that the frequency distribution properties of rhythm formants in the AM and FM spectra can provide a fruitful avenue of further research on the phonetics of prosodic typology, whether for language, dialect, register or speech style analysis.

As already noted, the data set is too small to do more than illustrate typological properties of AM rhythm and FM rhythm in an initial exploratory pilot study. Further, the granularity of these exploratory comparisons is too low to show the detailed linguistic basis for the similarities, which are based on many grammatical and phonological typological factors, or on stylistic and speaker-specific features. The results are nevertheless pointers to fruitful paths for further analysis in order to find out more about long-term Ibibio rhythm in narratives, and for comparison with other related or unrelated languages. For further analysis in linguistic terms, careful and detailed annotation of the recordings with the linguistic categories and functions of Ibibio is required in order to distinguish between structurally and rhetorically motivated factors. As with the use of distance metrics without the additional hierarchical clustering criteria, detailed linguistic analysis is needed in order to distinguish between the different factors which underlie the spectrally measured rhythms.

In the following section, a widely used linguistic phonetic method for looking at more linguistic detail and its relation to rhythm is discussed. The main methodological difference between the signal processing method and the isochrony heuristic method is that the former is inductive and proceeds from the signal data via a sequence of operations to spectral properties, rhythm formants, which can be used to predict the occurrence of linguistic units with these frequencies, while the latter is deductive and proceeds from known linguistic categories in the form of transcription labels paired with time-stamped intervals from the data, which are then used to derive statistical indices from the durations of the intervals.

5 Rhythm in Ibibio—Isochrony and Irregularity

5.1 Isochrony and Irregularity Measures

The second empirical method for the analysis of speech rhythm takes a deductive path based on the interval durations of linguistically pre-defined units (syllable constituents, syllables, words, phrases or longer units). The deductive procedure consists of the following steps (not necessarily in this order in practice):

1. defining linguistic units such as syllables or consonantal and vocalic features (on the basis of previous linguistic analyses);
2. predicting that these units have interval durations in the speech signal (as opposed to abstract features which have no direct correlate in the signal);
3. transcribing the signal in terms of predefined linguistic units;

4. annotating the signal, i.e., assigning time-stamped intervals in the signal to transcribed units (labels) such as consonantal or vocalic segments, syllables, words, phrases, using tools such as the annotation facility of Praat (Boersma, 2001) or automatic segmentation (Loukina et al., 2009);
5. statistical analysis of the time-stamped intervals in the annotation in order to produce an index of irregularity for the annotation;
6. comparison of different languages using the method defined under steps #1 to #5.

Using interval durations of syllables or syllable-like units, different languages have been found to differ in irregularity along open-ended scales starting from zero (signifying isochrony, i.e. equal timing), with so-called syllable timed languages closer to zero and stress-timed or word-timed languages further from zero. Many irregularity measures have been proposed, including variance or standard deviation and similar measures, percentages or standard deviations of consonantal and vocalic intervals: Scott et al. (1985); Roach (1982); Ramus et al. (1999).

A theoretical and practical problem with the descriptive statistical measures is that they were designed for static populations, not for ordered and temporally varying dynamic series with inter-unit dependencies such as those which characterise speech. Variation in speech tempo is an important rhythm parameter in itself, but descriptive statistical measures with a simple index as outcome imply counterfactually that speech tempo is constant in utterances and that durations are independent. A further problem is that their use of absolute (positive) or squared values removes the alternations or oscillations which define speech rhythm. These measures are generally used out of the box as ready-made standards. This is risky, however. The properties of the measures need to be understood. For this purpose, the most popular of these measures will be applied to Ibibio and analysed in detail.

A measure which was designed to separate local beat patterns from varying tempo is a measure of irregularity distance, the *Pairwise Variability Index* (Grabe & Low, 2002; Asu & Nolan, 2006), with a 'raw' (*rPVI*) variant, usually applied to the more regular consonantal intervals, and a 'normalised' (*nPVI*) variant which is usually applied to the more variable vocalic intervals. The two measures are standardly formulated as follows:

$$rPVI(D) = \sum |d_k - d_{k-1}| / (n-1)$$

$$nPVI(D) = \sum \left| \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right| / (n-1)$$

The *rPVI* is the mean of differences between pairs of neighbouring interval durations, assuming rhythm to be reflected in interval durations ($n - 1$ because in a sequence there is one less difference between neighbouring items than there are items). The *nPVI* differs from the *rPVI* in two respects. First, the difference in each pair is normalised by dividing the difference by the mean of the neighbouring pair. The normalisation makes the difference independent of the actual physical duration and comparable with other datasets.

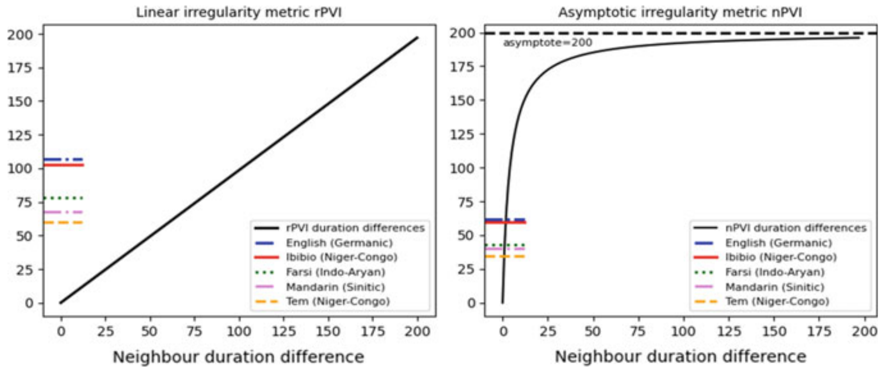

Fig. 10 Properties of rPVI and nPVI, with applications to 5 languages

Second, the normalised mean difference is multiplied by 100 for convenience of reading, misleadingly said to express the result as a percentage on the assumption that the scale is linear. This is not the case: the *rPVI* scale is linear while the *nPVI* is asymptotic (asymptote = 200; for normalisation by sum rather than mean, it would be 100) and both are open-ended, with no maximum which 100% could represent (cf. Fig. 10).

The *nPVI* and *rPVI* results for Ibibio were included in Fig. 10 along with other languages for the purpose of comparison. It is striking that the metrics for Ibibio are very close to those for English, an inflecting language. The presumption (based on these metrics alone) that Ibibio morphophonology may also be inflecting, with a prosodic foot structure, is confirmed by the detailed descriptions of prefixation and suffixation provided by Essien (1990) and Urua (2000), as already indicated in the modulation theoretic study. A point shown by Fig. 10 which is worth further investigation is that Tem (ISO 639-3 *kdh*), also a Niger-Congo language but in the quite different Gur branch, has a much lower *nPVI*, possibly indicating either different morphophonological structure or different prosodic foot structure (Tchagbalé, 2002).

Figure 10 also shows that in practice there is not a great deal of difference between the two measures for a given data set: they arrange the languages under investigation in the same order and create the same overall clusters.

Formally, the *PVI* metrics derive from standard distance metrics: the *rPVI* derives from the Manhattan Distance metric (also known as the Cityblock or Taxicab metric), and the *nPVI* derives from the Normalised Manhattan Distance metric (also known as the Canberra metric):

$$Manhattan(D) = \sum |v_i - w_i|, \quad Canberra(D) = \sum \frac{|v_i - w_i|}{|v_i| + |w_i|}$$

The mechanism involved in interpreting the *PVI* metrics as distance metrics is illustrated in Fig. 11: the duration vector *D* is split into two vectors *V* and *W* by

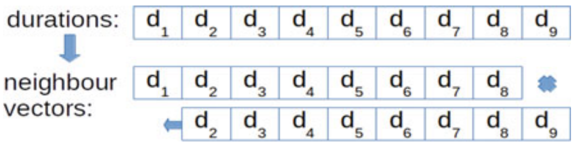

Fig. 11 Duration list copied into two overlapping vectors for distance / similarity analysis

removing the last element of one copy of the vector and shifting a second copy of the vector one place to the left.

Like the *rPVI*, the non-normalised Manhattan Distance metric uses raw measured values, and is thus not so suitable for comparing different data sets. There are two specific differences between the *PVI* metrics and the distance metrics. First, the two vectors in the case of the *PVI* metrics overlap: they are taken from the vector of interval durations D, with $V = {<}d_1, \ldots, d_{n-1}{>}$ and $W = {<}d_2, \ldots, d_n{>}$. Essentially, the 'distance' measured by the *PVI* metrics is therefore an 'average next-door-neighbour distance'.

Second, the distance metrics use the sum of differences between the two vectors, while the *PVI* metrics use the mean. This does not lead to a difference in distance ordering, however.

All the duration measures have in common that they define a binary subtraction relation (and thereby miss unary, tertiary and other non-binary rhythms), and that they abstract away from positive and negative differences by using the absolute difference between neighbouring durations. Thus, they cannot detect rhythmic oscillations, which involve variation of polarity. In fact, the metrics overgeneralise and can actually assign identical indices to alternating and non-alternating sequences. The metrics are thus very far from being an adequate model of speech rhythm. Nevertheless, while they have limitations, the metrics are useful heuristics and can perform a useful job in quantifying irregularites. If the data happen to contain only or mainly binary alternating durations, as suggested by Nolan and Jeon (2014), the measures can be usefully employed to quantify the degree of alternation (but cf. further critique by Arvaniti, 2009; Barry et al., 2003; Gibbon, 2018; Gibbon, 2021; Gut, 2012; Kohler, 2009; Tortel & Hirst, 2008; Wagner, 2007; White & Malisz, 2020).

5.2 *Investigation of a Small Ibibio Corpus*

5.2.1 Data, Method and Implementation

The small data set consists of 20 syllable-annotated Ibibio sentences of varying length which were recorded as training data for an Ibibio Text-to-Speech (TTS) system in the early 2000s, in a cooperative research project between the University of Uyo, Nigeria (PIs Eno-Abasi Essien Urua and Moses Ekpenyong),

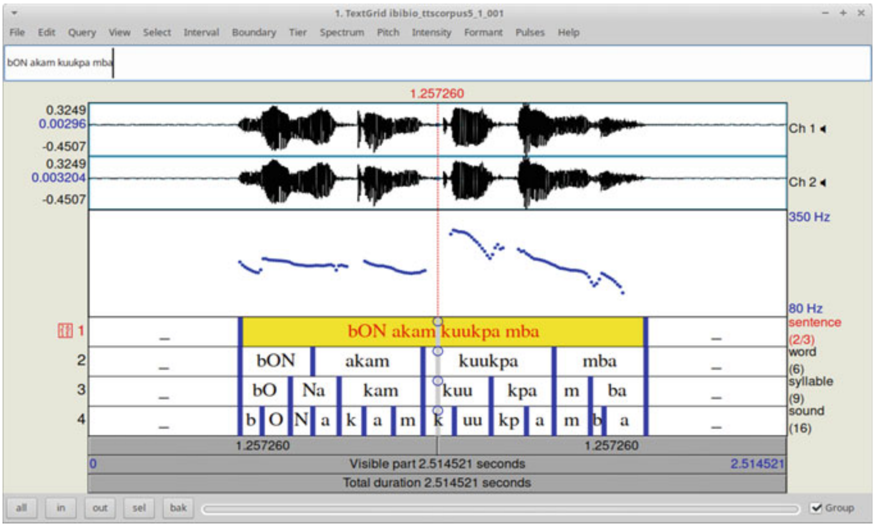
1. TextGrid ibibio_ttscorpus5_1_001
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
bON akam kuukpa mba
1.257260
0.3249
0.00296
-0.4507
0.3249
0.003204
-0.4507
Ch 1
Ch 2
350 Hz
80 Hz
1 _ bON akam kuukpa mba _ sentence (2/3)
2 _ bON akam kuukpa mba _ word (6)
3 _ bO Na kam kuu kpa m ba _ syllable (9)
4 _ b O N a k a m k uu kp a m b a _ sound (16)
1.257260
1.257260
0
Visible part 2.514521 seconds
2.514521
Total duration 2.514521 seconds
all in out sel bak
Group

Fig. 12 Screen shot of Praat edit window for an annotated Ibibio recording

and Bielefeld University, Germany (co-PI Dafydd Gibbon), under the auspices of the Local Languages Speech Technology Initiative (LLSTI). The sentences are extracts from written sources, read by Eno-Abasi Essien Urua and recorded by Peter Walhorn in the audio studio of Bielefeld University.

The recordings were annotated with the X-SAMPA keyboard-friendly encoding of the IPA on sentence, word, syllable and sound (phoneme) tiers using the Praat phonetic workbench software (Boersma, 2001). A sample annotation is visualised in the screenshot of Fig. 12 showing the Praat edit window with sound and annotation visualisations. The upper two panels show the two channels of a stereo recording, the next panel shows a typical downdrifting fundamental frequency panel with a 'reset' for a high tone, and the following panels show the annotation tiers for sentences, words, syllables and phones.

The annotation, with overall duration 2.515 s, is saved as a text file in the Praat TextGrid format (cf. the extract in Table 2). The annotation triples <*label*, *starttime*, *endtime*> are extracted from the TextGrid format and re-formatted as rows in a compact CSV database table format, together with the annotation interval duration, *endtime–starttime* (cf. Table 3).

Using the duration columns for all 20 readings (without the pauses, which are annotated as "_"), the *nPVI* and *rPVI* values are calculated. Very often a spreadsheet is used for this purpose, but in the present context the CSV file was read into a Python script which calculated the two measures along with a median based metric (which reduces the influence of rare extremely long or extremely short durations), and a standard deviation based metric. There are many ways to do this, but this script simply uses the *mean* and *median* functions of the Python NumPy library to calculate the *rPVI* and for the *nPVI* the shifted vectors are created before the index is calculated.

```
rPVI = mean(abs(diff(D)))
shiftedvectors = zip(durations[:−1],durations[1:])
npvi = 100 * mean([abs(x − y)/((x + y)/2) for x, y in shiftedvectors])
```

5.2.2 Results

The results of the data analysis in Table 4 show the syllable count per reading, the mean and median duration differences, and the interval duration standard deviations for each reading. The durations were measured in milliseconds rather than in fractions of a second, in order to bring the raw measured values on which the *rPVI* is based into the same normalised order of magnitude as the *nPVI*.

Figure 13 visualises the relations between different irregularity measures. All measurements show considerable variation in syllable duration differences between utterances, though, unlike the spectrogram method, not utterance internals.

Table 2 Sample syllable annotation in edited TextGrid format with indents removed ("_" indicates silent pause)

Intervals [1]:	Intervals [6]:
xmin = 0	xmin = 1.4490560598640456
xmax = 0.5453873379513452	xmax = 1.6757554993053712
text = "_"	text = "kpa"
intervals [2]:	intervals [7]:
xmin = 0.5453873379513452	xmin = 1.6757554993053712
xmax = 0.72659139435160891	xmax = 1.806514986983156
text = "bO"	text = "m"
intervals [3]:	intervals [8]:
xmin = 0.72659139435160891	xmin = 1.806514986983156
xmax = 0.90204294102488003	xmax = 2.0066090689371419
text = "Na"	text = "ba"
intervals [4]:	intervals [9]:
xmin = 0.90204294102488003	xmin = 2.0066090689371419
xmax = 1.2040497016919862	xmax = 2.5145208333333335
text = "kam"	text = "_"
intervals [5]:	
xmin = 1.2040497016919862	
xmax = 1.4490560598640456	
text = "kuu"	

Table 3 Syllable annotation time-stamps in CSV database table format

Label	Start time	End time	Duration
_	0	0.5453873379513452	0.5453873379513452
bO	0.5453873379513452	0.72659139435160891	0.1812040564002637
Na	0.72659139435160891	0.90204294102488003	0.17545154667327112
kam	0.90204294102488003	1.2040497016919862	0.3020067606671062
kuu	1.2040497016919862	1.4490560598640456	0.2450063581720594
kpa	1.4490560598640456	1.6757554993053712	0.22669943944132553
m	1.6757554993053712	1.806514986983156	0.1307594876777849
_	2.0066090689371419	2.5145208333333335	0.5079117643961917

Both the irregularity measure index and the spectral analysis demonstrate that claims of a single irregularity figure for a language are unrealistic, even for a highly normalised speech activity like reading aloud. The *nPVI* varies in this data set between 30 and 75, average 60; the median version of the metric lies between 19 and 89, average 53 (which is more realistic than the *nPVI* since the role of outliers is reduced with the median), the *rPVI* is between 62 and 169, average 119, and the *SD* of the durations in each reading lies between 20 and 46, average 39. These large variations, and in particular the large *SD* values show that rhythm is unstable and very far from isochronous over long stretches.

The variations are in need of further explanation, partly in terms of tempo, partly in terms of word, grammatical and rhetorical structures. A single overall index for such complex patterning is not enough. The correlation between these values and the overall number of syllables in the sequence, for example, is 0.273 for the *nPVI*, and 0.507 for the *SD*; thus, there is a small tendency for variation to increase with the length of the utterance, an intuitively plausible relation: the longer the reading, the higher the variability. Having established the variability of duration differences both within and between readings, the need arises to explain the variability, even more so than with the modulation theoretic approach.

In summary, an irregularity measure can provide information about the variability of durations within an utterance, but it is a measure of the degree of isochrony or, conversely, of irregularity, and not of rhythm. The dependence of these measures on prior annotation of recordings, in current practice, means that this method, unlike the automatic RFA method, is less efficient to operate with and is not amenable to the processing of large-scale data and comparison of data on a large scale. One way of increasing the granularity of the annotation method can be borrowed from the modulation theoretic method outlined in the present study: the multi-value annotation duration vector itself, rather than the index, can be used in future research as data for processing with distance metrics, distance networks and hierarchical clustering, and then compared with the modulation theoretic results.

Table 4 Basic descriptive statistics from annotation data analysis

Reading	Syll Count	nPVI	rPVI	nPVI median	nPVI SD
1	7	30	62	31	20
2	12	57	81	50	33
3	25	57	107	50	40
4	16	55	91	56	39
5	13	55	96	58	41
6	7	71	132	62	31
7	38	60	126	59	43
8	43	64	133	61	35
9	35	60	132	52	44
10	17	73	121	89	36
11	25	52	91	44	36
12	25	64	144	59	42
13	48	69	169	60	44
14	19	65	134	61	41
15	24	57	104	43	46
16	32	66	167	53	46
17	9	43	67	19	38
18	11	75	142	64	37
19	22	62	118	44	45
20	41	57	162	51	38
Means	**23**	**60**	**119**	**53**	**39**

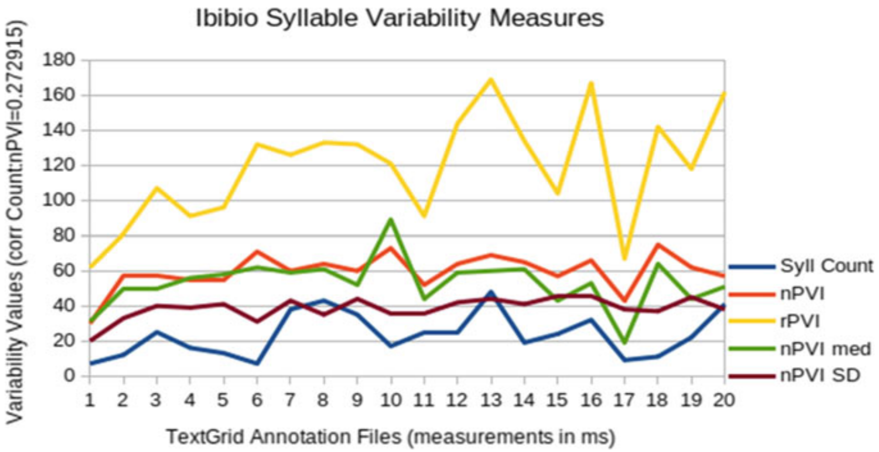

Fig. 13 Variability of nPVI-derived values between readings

6 Conclusion

Speech rhythms in Ibibio were investigated with two current methodologies: an inductive signal-based modulation theoretic approach together with low frequency Rhythm Formant Theory and its Rhythm Formant Analysis methodology,[2] and a deductive annotation-based isochrony heuristic based on the sequence of duration intervals of predefined linguistic units. It was shown with both methods that rhythm in Ibibio cannot be represented meaningfully by a single index, but that multiple rhythms are present both within and across utterances. The concept of 'rhythms of rhythm' was used to indicate that long-term rhythm variation may itself have a regularly varying character.

Although the size of the available Ibibio data set is small and the study has more exploratory than quantitative confirmatory features, the claim is that both the modulation theoretic and the isochrony heuristic approaches have complementary potential for further investigation of the prosodic typology of Ibibio and related languages. The problem of sparse data must be solved in future studies. The rhythm spectrogram of the modulation theoretic approach permits precise analysis of the relation between long-term rhythm variation and the episodic structure of oral narrative, and permits comparison of speaking styles within one language or speaking styles in different, related languages.

The potential of the annotation method, which has been extensively and successfully applied in speech technology, has not yet been explored in sufficient depth in the phonetic study of rhythm, but indications of possible extensions of this method were given. Both the modulation theoretic method and the isochrony heuristic method have been described in sufficient detail in the context of typological comparison of different languages to enable the methods to be applied to other languages.

If documentation and description of neighbouring languages to Ibibio is extended to include a well-defined corpus of narratives, the methods described in this exploratory study can be applied systematically in order to investigate not only the quantitative similarity of these languages to each other, but also to provide a foundation for systematic discourse semantic, pragmatic and rhetorical analysis as well as the functions in related cultural domains such as music and dance. Going beyond the immediate neighbours of Ibibio and applied to the languages of Nigeria the methods will permit the creation of a map of the prosodic typology of the major language families in Nigeria and beyond. It is to be hoped that the rich topic of rhythm, the cultural heartbeat of society, will be taken up more extensively not only in linguistic and phonetic investigation of prosody, of which rhythm is a part, but also in in language and culture documentation of music and dance in these speech communities, not only for the sake of scientific understanding but also for understanding and maintaining the cultural heritage of these communities.

[2] Python code and data will be available on GitHub from publication time: https://github.com/dafyddg/RFA.

Acknowledgements First and foremost, I wish to acknowledge the essential contribution of Prof. Eno-Abasi Essien Urua of the University of Uyo to the present study. Eno and her publications not only introduced me to the intricacies of tone patterning and tone assignment in Ibibio and related Lower Cross languages, but also to her efficient style of fieldwork and to her network of colleagues and associates in Nigeria, from whom I was able to learn about many features of Nigerian languages and regional cultures. Furthermore, the Ibibio data for this study are recordings of her voice.

I would also like to acknowledge the contribution of Dr. Moses Effiong Ekpenyong, whose cooperation with Eno, his dissertation applying quantitative information technology methods to Ibibio tone and his work on text-to-speech systems for Ibibio has been a source of encouragement in my own applications of computational linguistic and phonetic methods to Ibibio. Thanks also to Prof. Ulrike Gut for reintroducing me to the study of rhythm and for partnering a number of fruitful research projects, including education and fieldwork research at the University of Uyo, to Mfọn Udọinyang and Okokon Akpan, who kindly helped with the Ibibio data, and to Peng Li and Xuewei Lin (Guangzhou) who kindly helped with the Chinese data.

Acknowledgement for support of travel between Nigeria and Germany is due to the DAAD, and for the vision of Roger Tucker and Ksenia Shalonova of the Local Language Speech Technology Initiative (LLSTI) for applications of speech technology to languages in Africa and Asia.

References

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

Asu, E.-L., & Nolan, F. (2006). Estonian and English rhythm: A twodimensional quantification based on syllables and feet. *Speech Prosody* 3.

Barbosa, P. A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. *Speech Prosody, 1*, 163–166.

Barry, W. J., Andreeva, B., Russo, M., Snezhina Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? In D. Recasens, M.-J. Solé, & J. Romero (Eds.), *15th International Congress of Phonetic Sciences (ICPhS XV)* (pp. 2693–2696).

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

Connell, B. (2002). Downdrift, downstep and declination. In U. Gut, & D. Gibbon (Eds.), *Typology of African prosodic systems*. Bielefeld: University of Bielefeld. http://wwwhomes.unibielefeld.de/gibbon/Dafydd_Gibbon_Publication_PDFs/TAPS/Connell.pdf

Couper-Kuhlen, E., & Selting, M. (2018). *Interactional linguistics. Studying language in social interaction*. Cambridge University Press.

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

Dilley, L. C. (1997). *The phonetics and phonology of tonal systems*. Ph.D. dissertation, MIT.

Dudley, H. (1939). Remaking speech. *Journal of the Acoustical Society of America, 11*(169).

Ekpenyong, M. E. (2012). *Speech synthesis for Ibibio*. Ph.D. dissertation, University of Uyo.

Essien, O. E. (1990). *A grammar of the Ibibio language*. University Press Limited.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In B. Bel & I. Marlien (Eds.), *Speech Prosody* (Vol. 1, pp. 323–326). Laboratoire Parole et Langage.

Gibbon, D. (2002). Computational phonology and the typology of West African tone systems. In: U. Gut & D. Gibbon (Eds.), *Typology of African Prosodic Systems (TAPS)*. Bielefeld Occasional Papers in Typology 1.

Gibbon, D. (2016). Legacy language atlas data mining: Mapping Kru languages. In *Proceedings of the LREC 2016*. ELDA.

Gibbon, D. (2018). The future of prosody: It's about time. Keynote. In *9th International Conference on Speech Prosody*.

Gibbon, D. (2021). The rhythms of rhythm. *Journal of the International Phonetic Association*. First View (Open Access), 1–33. https://doi.org/10.1017/S0025100321000086

Gibbon, D., & Li, P. (2019). Quantifying and correlating rhythm formants in speech. In *3rd International Symposium on Linguistic Patterns in Spontaneous Speech*. Academia Sinica, Taipei, Taiwan.

Gibbon, D., & Urua, E.-A. (2006). Morphotonology for TTS in Niger-Congo languages. In *Proceedings of the 3rd International Conference on Speech Prosody*. TUD Press.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In *Laboratory Phonology 7*. De Gruyter Mouton.

Gut, U. (2012). Rhythm in L2 speech. In D. Gibbon, D. Hirst & N. Campbell (Eds.), *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem. Special Edition of Speech and Language Technology* (Vol. 14/15, pp. 83–94). Polish Phonetics Society.

Gut, U. (2002). Prosodic aspects of Standard Nigerian English. In: U. Gut & D. Gibbon (Eds), *Typology of African prosodic systems* (Vol. 1, pp. 159–165). Bielefeld Occasional Papers in Typology.

Gut, U., & Gibbon, D. (eds.). (2002). *Typology of African Prosodic Systems* (Vol. 1, pp. 159–165). Bielefeld Occasional Papers in Typology. http://wwwhomes.uni-bielefeld.de/Dafydd_Gibbon_Publication_PDFs/TAPS/proceedings.html

Gut, U., Urua, E.-A., Adouakou, S., & Gibbon, D. (2002). Rhythm in West African tone languages: A study of Ibibio, Anyi and Ega. In U. Gut & D. Gibbon (Eds.), *Typology of African prosodic systems*. *Bielefeld occasional papers in typology* (Vol. 1, pp. 159–165). Universität Bielefeld.

Jassem, W., & Gibbon, D. (1980). Re-defining English stress. *Journal of the International Phonetic Association, 10*(1980), 2–16.

Kohler, K. (2009). Editorial: Whither speech rhythm research? *Phonetica, 66*, 5–14.

Kohler, K. J. (2018). *Communicative functions and linguistic forms in speech interaction (Cambridge Studies in Linguistics 156)*. Cambridge University Press.

Liberman, M. Y. (1975). *The intonational system of English*. Ph.D. dissertation, MIT.

Loukina, A., Kochanski, G., Shih, C., Keane, E., & Watson, I. (2009). Rhythm measures with language-independent segmentation. *Interspeech, 2009*, 1531–1534.

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press.

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor *Philosophical transactions of the Royal Society B. Biological Sciences*, 1–11.

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *14th International Congress of Phonetic Sciences (ICPhS XIV)* (pp. 1075–1078).

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Ph.D. dissertation, MIT.

Poeppel, D., & Florencia Assaneo, M. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience, 21*, 322–334.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265–292.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic controversies: Essays in linguistic theory and practice* (pp. 73–79). Edward Arnold.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and French. *Journal of Phonetics, 13*, 155–162.

Samuel, T., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America, 124*(2), EL34–EL39. [PubMed: 18681499].

Tchagbalé, Z. (2002). L'accent tonal du Tem. In: D. Gibbon, U. Gut, E.-A. Urua (Eds.), *TAPS: Typology of African Prosodic Systems*. In: U. Gut & D. Gibbon (Eds.), *Typology of African prosodic systems*. *Bielefeld Occasional Papers in Typology* 1. Universität Bielefeld.

Todd, N. P. M., & Brown, G. J. (1994). A computational model of prosody perception. In *International Conference on Spoken Language Processing (ICLSP-94)* (pp. 127–130).

Tortel, A., & Hirst, D. (2008). Rhythm and rhythmic variation in British English: Subjective and objective evaluation of French and native speakers. In *4th International Conference on Speech Prosody* (pp. 359–262).

Traunmüller, H. (1994). Conventional, biological, and environmental factors in speech communication: A modulation theory. In: M. Dufberg, & O. Engstrand (Eds.), *Experiments in speech process. PERILUS XVIII* (pp. 1–19). Department of Linguistics, Stockholm University. (Also: *Phonetica* 51:170–183, 1994)

Urua, E.-A. E. (2000). *Ibibio phonetics and phonology. CASAS Book Series No. 3*. Centre for Advanced Studies of African Society.

Urua, E.-A. E. (2004). Illustrations of the IPA: Ibibio. *Journal of the International Phonetic Association, 34*(1) (Archived online, including audio recording).

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

White, L., & Malisz, Z. (2020). Speech rhythm and timing. In C. Gussenhoven & A. Chen (Eds.), *The Oxford handbook of language prosody*. Oxford University Press.