Ega orature: discourse prosody meets machine learning

Dafydd Gibbon

Bielefeld University

Abstract

Traditional orature meets machine learning in the present exploratory study of spontaneous story-telling in Ega, a putative Western Kwa language of South Central Côte d'Ivoire, with *narrator-responder* interaction and *call-response* song interludes. The empirical basis of this study in digital humanities is the long-term prosody of the story, particularly rhythms, analysed holistically with macrostructural discourse phonetic methods, in relation to sociolinguistic scenario categories. With new quantitative methods, stories up to five minutes long are studied, unlike in phonology and phonetics, which typically deal with very short time domain data of under 1s or up to a few seconds. The main procedure is to apply long-term spectral analysis using the method of Rhythm Formant Analysis (RFA) to uncover long-duration timing regularities in the story-telling exchanges, and to relate the phonetic results to participant roles in story-telling. The overall methodological context is the modulation theory of speech signal properties. The feasibility of the RFA method for typological classification is examined by comparing four traditional stories in Ega with the similar story-telling of two stories in Agni (Anyi), and also with the different register of read-aloud narrative in educated Ivorian French and in Ibibio, also a Niger-Congo language, but not Kwa. For this comparison, quantitative clustering methods of unsupervised machine learning are used. Finally, the utility of discourse phonetics together with shorter domain quantitative analyses and for technological applications in applied digital humanities is discussed.

1. Story-telling in Ega: discourse phonetic analysis and classification

The present exploratory study[1] in digital humanities is concerned with applying new methods in discourse phonetics and unsupervised machine learning algorithms in an investigation of traditional orature (oral narrative and poetry). The starting point is traditional story-telling in Ega (ISO 639-3 *ega*, in the village of Gniguédougou, south of Divo, Côte d'Ivoire, approximately 5.8°N, 5.4°W; cf. Bole-Richard 1983; Connell et al. 2002; Gibbon et al. 2004). Ega, commonly classified as an endangered Western Kwa Niger-Congo language, is spoken in an enclave of about 5000 people in 9 villages, within a larger region where Dida (ISO 639-3 *dic*), a Kru Niger-Congo language (Gibbon 2016), is spoken. There is some dialect variation between villages. The region is located south-west of Divo in the Eastern Kru area in south-central Côte d'Ivoire. The Ega have a mainly horticultural and partly plantation-based economy, and practise traditional hunting techniques by enclosing several hundred square metres with nets, which have the appearance of the canoe trawling nets used by fishermen along the South Coast of Côte d'Ivoire. This suggests a possible migration history of relocation northwards from the coast and earlier westward migration along the coast from the more central Kwa regions.

The objective of the study is exploratory. It explores new methodological configurations with holistic inductive treatment of long-term discourse domains, and is not limited to the application of well-known methods to new data in order to confirm or reject aspects of a prior theory. The

1 This study is dedicated to the memory of the late Gnaoré Marc, chef de village of Gniguédougou, whose pride in his language and willingness to communicate the language to a group of visitors was exemplary.

pragmatic perspective taken is similar to that of a ‘first encounter’ in fieldwork, the ‘humanities’ aspect, coupled with analysis using recent developments in the inductive quantitative phonetics of discourse prosody over long-duration domains, the ‘digital’ aspect. The data set is small, taken opportunistically from a set of stories collected for other purposes, with eight narratives spoken by four speakers. Considering that many linguistic analyses are based on transcriptions of utterances by a single speaker, not infrequently the linguist herself, this procedure is not overly restrictive.

The data are essentially the physical signals of spoken language, coupled with with broad sociolinguistic scenario categories. In such a first encounter scenario there are initially only rudimentary pre-linguistic perceptions of the speech sounds, words and phrases of the very short time domains of linguistic analysis. Perception and understanding tend to be holistic: phonetic properties of complete utterances are recognised, rather than grammatical units, and there is no semantic interpretation beyond basic encounter behaviours. In longer discourse time domains, it is not only the physical aspects of macrostructural prosody, the discourse phonetics, of the language which are evident in the form of the different rhythms and melodies of greeting and welcoming: the music of speech.

Pragmatic scenario properties are also physically discernible by visual and auditory means: speakers are identified by appearance, activities and voice, and their roles, including turn-taking roles in different situative genres, styles or registers, are readily interpreted. The various kinds of first encounter scenario constitute a much-studied topic in sociolinguistics and applied linguistics. The present use of the term relates to a linguist as guest in a village, who initially mainly just hears the rhythms and melodies of the language, and who has none of the detailed cultural and language-specific hermeneutic abilities of the native speaker. Speech-accompanying gestures were dealt with briefly elsewhere (Rossini and Gibbon 2011).

The pragmatic scenario of ritual story-telling, shared by many village communities not only among the Ega but throughout West Africa and beyond, can be briefly characterised as follows:

1. Participants:
    1. *narrator*, a designated role within the village;
    2. *responder*, selected *ad hoc* by the narrator in the current session, or as a fixed role;
    3. *audience*, members of the village community;
    4. *call* initiator and *response* audience in work-song-like interludes.
2. Location:
    central village meeting place.
3. Script:
    1. introduction followed by song with singer-audience alternation;
    2. story with narrator-responder alternation and work-song-like interludes;
    3. closing ‘moral of the story’ and thanks from responder.

The storytelling scenario is shown (with permission) in Figure 1, with the narrator on the right, the responder opposite, and the audience sitting and standing in the background. Seating consists of traditionally constructed chairs. The table is also used for village meetings and was used to position the microphones for stereo recording of the narrator and of the responder with the audience.

The story, which was translated later, is an allegory about a villager walking to a neighbouring village, without knowing that something evil was in store for him. However, a little bird had eavesdropped on people in the neighbouring village and knew the plans. The bird tries to communicate the bad news to the villager in a song, which the narrator sings in neighbouring Dida,

not in Ega. Unfortunately the villager does not understand the bird and continues walking. The dire end of this walk is not narrated but the moral of the story is propounded: learning another language is a necessity for survival, and this is what our visitor [dàvíd] is doing.

Figure 1: Ega storytelling scenario: narrator in the right foreground; responder in the left foreground; audience surrounding in the background (frameshot from Ega_conte2.mpg).

Section 2 introduces the distinction between short and long-duration domains as a basis for macrostructural discourse phonetic studies using the Rhythm Formant Analysis (RFA) method, with an initial illustration of the method using a very short first encounter greeting in Ega. Section 3 applies the method to a sample of traditional spontaneous narrative data in Ega, as described above, which was collected for language documentation purposes in the course of a number of cooperative projects with colleagues in Ivory Coast. Section 4 provides an overview of the modulation theoretic background to macrostructural discourse phonetics. Section 5 is concerned with applying the results of the analysis to discourse typology using unsupervised machine learning methods (clustering) in an exploratory quantitative comparison of four Ega stories with two stories in the closely related language Agni (Anyi, ISO 639-3 *any*) and with two narratives in the different register of read-aloud narrative in educated Ivory Coast French and in Ibibio (ISO 639-3 *ibb*) a Nigerian Niger-Congo language, but Lower Cross, not Kwa. Finally, in Section 6 the arguments and results are summarised, conclusions are drawn and the outlook for further studies is outlined.

2. Macrostructural discourse phonetics

2.1. Short and long duration domains

The present study focuses on timing patterns in discourse, including discourse rhythms. The physical exploration of the longer time domain of discourse requires novel applications of some of the familiar signal processing algorithms of acoustic phonetics to a time domain of the order of minutes rather than the fractions of a second which are occupied by typical acoustic correlates of phonological categories (Gibbon 2021). The waveform of the entire time domain of the five minute story is shown in Figure 2.

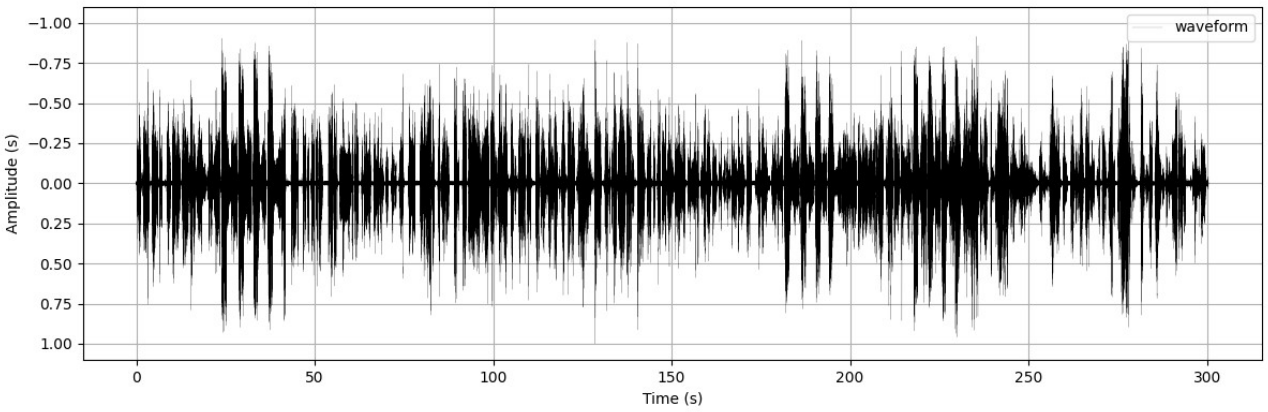

Figure 2: Waveform of the Ega story, including all participant turns, with prominent song interludes visible in the regular spike alternations from 25s, 180s and 220s.

Intuitively, different segments of the waveform can already be identified by visual inspection of the waveform in different temporal segments. For example, the sequence of four conspicuous spikes around 25s-40s is easily interpreted, on listening to the recording, as a song with the structure of the traditional work-song dialogues which later gave rise to some styles of blues and reggae, with a *call-response* sequence which has the narrator in the *call* role and the responder and audience in the *response* role. The same kind of event occurs at the four spikes around 180s-195s and the five spikes around 220s-240s.

The task is first, to identify participants in a coarsely characterised scenario in order to relate these to temporal properties of the signal throughout the story and second, to compare the temporal properties of different stories. The physical properties of the story are interpretatively grounded by association with the scenario. Questions need to be answered about long-term variation in short-term properties of speech, such as syllable durations and tones or pitch accents, for example as speech rate acceleration and deceleration, or as regular changes of rhythm, the rhythms of rhythm, in the stories and other task-oriented communication processes which are of interest for ethnolinguistic and sociolinguistic characteristics of orature as well as orature historians in digital humanities.

In phonetics and linguistics the focus tends to be on behaviourally tiny events generated by repeated activities of the brain and actions of the muscles of the mouth and throat which take place in a time range from about 0.05s (rate: 20/s or 20Hz) to about 5s (0.2Hz). Sentences and short dialogue exchanges may take 5s or less, phrases and basic rhythm beats about 1s, and around half a second for words and pitch accents, while about one quarter of a second is characteristic of syllables and tones, down to one twentieth of a second for short phones and consonantal perturbations of

phone formants and of syllable prosody. The duration of the story under consideration is almost exactly 300s.

In linguistic phonetics, this time restriction to domains between one second divided by five and one second multiplied by five is also constrained by the limits of available analysis and display software: up to ten seconds, for example, is a typical setting for phonetic visualisation and workbench software, beyond which few dare to venture. Much shorter intervals with temporal resolution down to 50ms are also common. There are good reasons for studying tiny events and their repetitions in the word and phrase phonetic domain, as well as in neural domains (Poeppel and Assaneo 2020): this domain is at the lower speed limit of human experience, of the same order of magnitude as heartbeats, blinks, breaths, chewing, copulation, walking and running, dancing, clapping and musical beats, as well as many bird and animal calls and actions.

There is a linguistic life beyond these narrow borders, shown in research by conversational analysts and proponents of interactional linguistics (Couper-Kuhlen and Selting 2018), by ethnographers with a concern for literature and orature, by students of rhetoric and public performance, and by performers of more sophisticated dance figures than one normally sees on local dance floors. The present study aims precisely at providing a dual grounding for long-range speech episodes by examining the relation between scenario categories and the prosodic dynamics of story readings.

This concern raises a more general epistemological query about the extent to which qualitative methods of the humanities can be related to the quantitative methods of natural sciences and engineering. Linguists traditionally use qualitative methodologies to study speech by ‘reducing languages to writing’ (Pike 1947) in the form of transcriptions, while phoneticians traditionally use quantitative methodologies for the study of speech signals, with a modicum of qualitative input for identifying the categories whose phonetic properties are to be investigated. The present study of discourse prosody demonstrates one way of combining qualitative and quantitative methods.

Many definitions of ‘prosody’ have been proposed in the past, which do not need to be recapitulated here. For present purposes, this intuitive explicandum, covering both form and function, is sufficient:

Prosody is the rhythm and melody of speech.

As already noted, it is a category mistake to atomise prosody into three autonomous compartments of ‘duration’, ‘frequency’ and ‘intensity’ and then to search for interdependencies: intensity (or, more simply: amplitude) and frequency are quite simply time functions in their own right. Durations and sequences of durations are quite different, being components of the time axis for the two time functions, and identifiable by low-frequency (LF) spectral analysis of the two functions. Speech rhythms have specific properties as perceived acoustic regularities in time, such as the repetition of regular patterns with durations under 1s in a trajectory along some physical parameter or simultaneous physical parameters.

The possibility of finding a physical identity for speech rhythms has often been doubted in qualitative studies in linguistics and linguistic phonetics. This, however, was largely due to the use of inadequate descriptive statistical methods which treated constituents of utterances as if they were a static population, not time functions over events (Gibbon 2021).

Appropriate methods of modelling speech as a time function have been developed over the past quarter of a century (Todd and Brown 1994; Tilsen and Johnson 1998; Traunmüller 1994; Gibbon

2018, 2021 and many others). In this approach, speech rhythms are constituted by regular variations in the overall amplitude contour of syllable sequences. Regular variations in the fundamental frequency of these sequences also play a role. Basic rhythmical beats emerge from alternations of low-amplitude consonants with high-amplitude vowels in syllables, and by alternations of longer, higher-amplitude syllables with shorter, lower-amplitude syllables, as well as in slower regular rhythms of rhythm over longer stretches of speech.

Quantitative methods inevitably presuppose a minimum of qualitative modelling conventions, whether these involve a prior decision to look at the moon or a decision to analyse stories. In the present context it is the semiotic properties of stories which figure in the pretheoretical qualitative modelling conventions.

2.2. Initial illustration of the method

Annotation

Before the quantitative signal processing analysis, the first step taken is a traditional phonetic analysis with qualitative annotation of an example utterance using the Praat phonetic workbench (Boersma 2002), followed by a brief quantitative analysis. A short example is shown in Figure 3, a multitier annotation of a first encounter greeting by the Ega speaker who was the main Ega project contact in the fieldwork project. The annotation was made with later 'hindsight knowledge' of the language, and is saved in a Praat *TextGrid* format file.

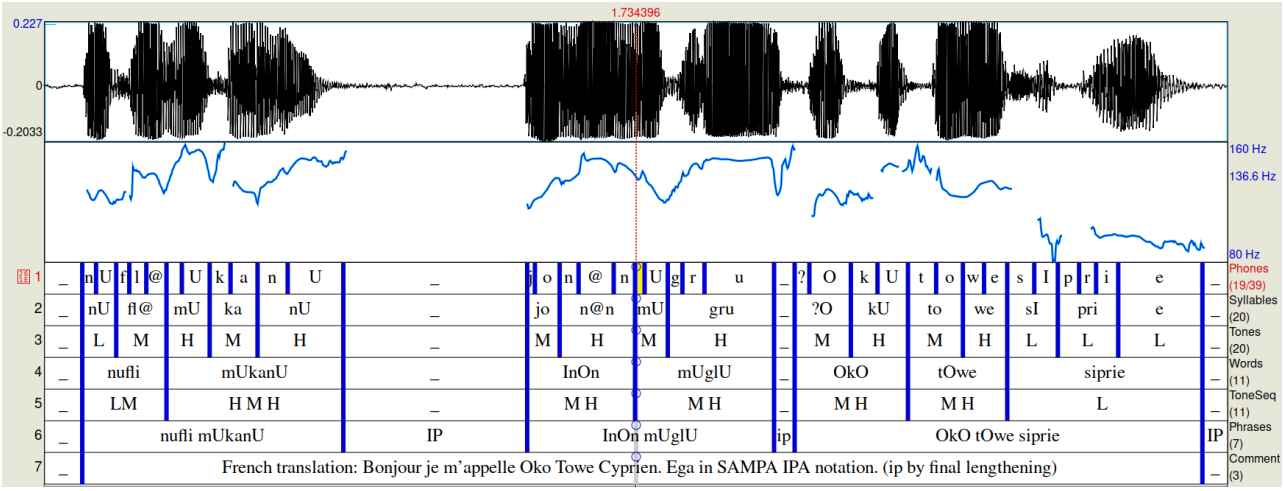

Figure 3: Annotation with Praat of a self-introduction by an Ega speaker with tiers phone, syllable, tone, word, word tone sequences, prosodic phrase and comment.

The Praat annotation screenshot shows *phone*, *syllable*, *tone*, *word*, *word tone sequence*, *prosodic phrase* and *comment* tiers. The phrase tier is segmented into two major prosodic phrases, the second consisting of two minor prosodic phrases. The first minor prosodic phrase is marked mainly by final lengthening with no pause, so the border is taken to begin within the lengthened final vowel. The second minor prosodic phrase terminates at the final pause and is subsumed by the IP boundary. The fundamental frequency (F0) gives rise to a pitch impression of a tone sequence with automatic tonal downstep and tone terracing. Although the phonological abbreviations 'IP/ip' (Intonation Phrase, intermediate phrase) are used in the annotation for brevity, the traditional terminology of major and minor prosodic phrases is preferred in the present study: Ega is a terraced tone language with tonal downdrift and tonal upsweep, not intonation in the conventional sense, so 'intonation phrase' is an inappropriate term here.

Measurements and descriptive analysis

After the qualitative annotation step, an initial quantitative analysis is made of data taken from the TextGrid annotation file. The numbers of this single example have illustrative methodological value only, and do not necessarily represent generalisable properties of Ega (though this is not discounted). Basic measurements and descriptive statistics are shown in Table 1.

Table 1: Quantitative syllable properties in the illustrative Ega greeting.

Syllable: measurements	
Attributes	*Values*
n:	16
total duration (ms):	2684
min (ms):	95
max (ms):	311
duration range (ms):	216

Syllable: descriptive statistics	
Attributes	*Values*
mean (ms):	167.75
mean syllable rate (Hz):	5.96
slope:	4.92
SD (ms):	59.47
nPVI:	42

The standard deviation is fairly low, as is the *nPVI* (normalised pairwise variability index, the averaged normalised difference between durations of neighbouring items), indicating a tendency towards syllable timing (English, for example, typically has an *nPVI* around 65, Chinese around 35), also shown by the mean syllable rate (the inverse of the mean) of nearly 200ms (nearly 6/s, 6Hz) . The positive slope shows deceleration: shorter syllables at the start, longer at the end.

Signal processing

Qualitative methods are empirically deductive: previous experience has led to expectations: hypotheses which are compared with the speech signal in the process of transcription or annotation. The quantitative signal processing method, on the other hand, is formally deductive, using mathematical relations and operations, but empirically inductive from the phonetic point of view as there are no prior qualitative phonetic categories, only measurements of values of physical parameters. The signal processing method of spectral analysis leads to a frequency domain representation of the rhythms of the recording, and proceeds in the following steps:

1. Extraction of the waveform from the recording.
2. Annotation with Praat.
3. Visualisation of the duration relations in interpausal units.
4. Holistic extraction of the *amplitude modulation track* (AM track), also *amplitude modulation envelope*, from the entire waveform.
5. Low frequency (LF) spectral analysis of the AM track, showing the LF spectrum below 10Hz.
6. Identification of high-magnitude LF spectral regions (rhythm formants).
7. Validation of the rhythm formants with descriptive statistics.

Visualisation

Steps 1 and 2 in this procedure were already shown. Steps 3, 4, 5, 6 and 7 above are shown in Figure 4. The top panel of Figure 4, the event duration panel, visualises step 3 in two ways: first by the length of a horizontal bar for each utterance, second by the position of the horizontal bar on the *y*-axis. The upper mid panel represents the waveform, and the lower mid panel represents step 4, the *AM track*, the phonetic correlate of the *sonority curve* which is postulated in many phonological analyses as a function of the alternation between consonantal strength of syllable margins and vocalic strength of syllable centres.

The lowest panel of Figure 4 visualises a different kind of information: a frequency domain representation of the entire AM track as an LF spectrum, a signal analysis procedure designed to identify rhythms in the speech signal. That is, the demodulated AM information track is treated as a separate signal, and analysed in its own right, yielding a LF AM spectrum, in order to find the strongest frequencies in the signal modulation. The ten highest-magnitude frequencies, highlighted with stars and annotated with frequencies and mean 'beat' interval durations, occur in frequency zones which are interpreted as rhythm formants because they correspond to frequencies of syllable, word, phrase rates, etc. The correspondence can be validated by comparison with annotation results.

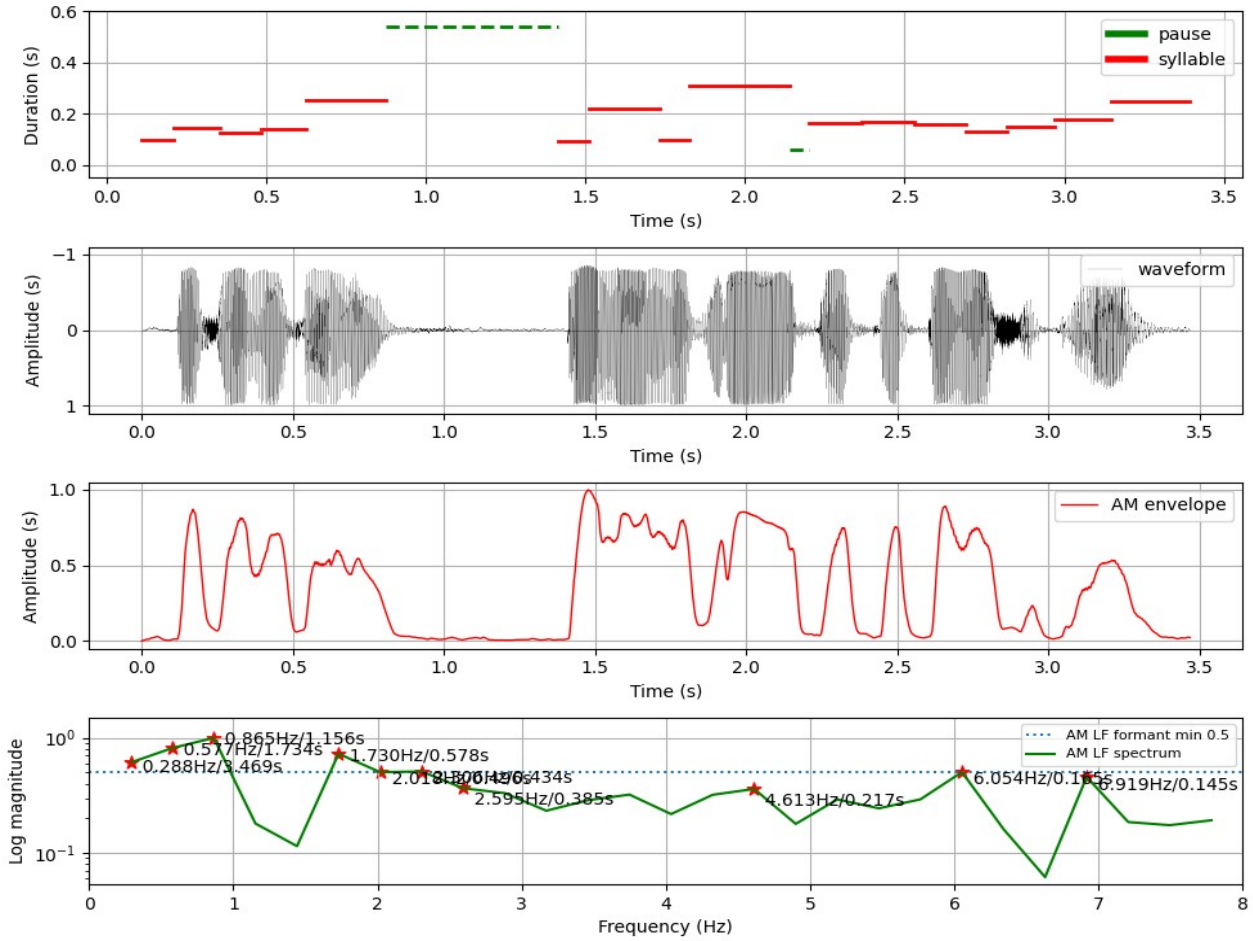

Figure 4: Time and frequency domain of Ega greeting: duration distributions (top); waveform (upper mid), amplitude track (lower mid), AM LF spectrum (lowest).

The spectral analysis method is entirely independent of the annotation process. Whereas annotation is an extremely labour-intensive and relatively slow qualitative process (often several hundred times the real time duration of the recording being annotated), the quantitative signal processing is accomplished in a few seconds, sometimes less than one second, depending on the length of the signal and the computing environment.

Interpretation

Inspection of the lower panel of Figure 4 and comparison with the annotation shows that a number of rhythm formants can be found and identified: at approximately 0.3Hz to 0.9Hz (i.e. 3.3s to 1.1s mean duration), relating to the major prosodic phrase, 1.7Hz to 2.6Hz (i.e. 0.59s to 0.38s mean duration), relating to the minor prosodic phrases, and between 4.6Hz and 6.9Hz (217ms and

145ms), relating to words or syllables. The frequency zones are ‘fuzzy’: the human voice is not clockwork. In the following section, these steps are applied to the entire duration of the Ega story.

3. Qualitative and quantitative analysis

3.1. Syntagmatic and paradigmatic properties of participant categories

The initial qualitative stage of the story analysis consists of characterisation of the situational context and annotation of the recording with discourse role labels, using the Praat phonetic workbench (Boersma 2002): *narrator*, *responder*, *other*, *song* (divided into *call* by the narrator and polyphonic *response* by responder and audience), and pauses >0.2s. Orality and the systematic role of dialogue and song are underestimated in much recent literature on African folk narration. Lô et al. (2020), for example, develop a computational machine learning model of the structure of the narrator contribution to West African stories, but they treat it as text, not orature, ignoring the *narrator-responder* dialogue and song process, though a simplified account of the audience role is given with reference to Berry and Spears (1991): “The role of the audience is to interrupt throughout the tale by making remarks, correcting the storyteller, or by singing songs” (Lô et al. 2020:4). Ninan et al. (2016) also model the narrative text only, but provide more detail about song and its oral function for the rhythm of the interchange: “The contents of these songs usually convey requests, reasons, petitions, vital instructions or information[...]. There are terms in the song that do not serve any semantic function but only to serve the purpose of creating a rhythm.”

The *call-response* structure of the singing and the audience role in the scenarios which are dealt with in the present study are more sophisticated than this, however, and can be characterised by the cyclic linear dialogue grammar shown in Figure 5 and describedd in Table 2, which applies to all the stories analysed in the following sections.

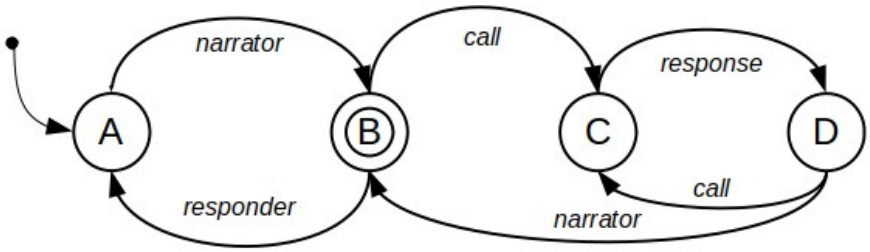

Figure 5: Discourse grammar of narrator-responder stories with call-response interludes. Initial state: A. Final state: B.

Table 2: Description of the dialogue grammar of the stories.

Start:	state A (arrow from black dot)
End:	state B (double circle)
A	proceed with the narrator to state B
B	either end or return to state A with the responder and continue the loop with the narrator
	or continue to state C with the call
C	with the response to choices at state D
D	either another call back to state C
	or move with the narrator to the choices at state B

The cycles in the grammar underlie the rhythms of rhythm discussed in the present study. This long-term cyclical structure is not discussed further but within this syntagmatic framework the thesis is proposed that utterances with more compact distributions (i.e. higher kurtosis or peak-tail relation) are candidates for ritualised utterances, while utterances with greater distributional spread are more likely to be extemporised. The initial prediction, therefore, is that the *backchannel utterances* by *responder* and *response* by the audience are distributionally more compact than the *narrator* and the *call* utterances, which can be interpreted as more strongly ritualised or stereotyped.

The quantitative stage of the analysis starts with the temporal properties of the events in the story sequence, represented in Figure 6 as a box and whisker plot, which enables a systematic comparison of duration distributions of the label types. For each category, the rectangular box represents the region between the first and third quartiles, with the median as an inside bar and the mean as a dot. Outliers are represented with circles, with one extremely long outlier in the case of the category *call* (in one of the song intervals). To the left of each box is a vertical bar representing one standard deviation above and below the mean and to the left of the bar is a further bar representing the mean (short mid horizontal bar), the 95% confidence interval (upper and lower horizontal bars), and the standard error of the mean (horizontal bars between upper and lower confidence interval bars and the mean). To the right of the box are black dots representing durations of each data point.

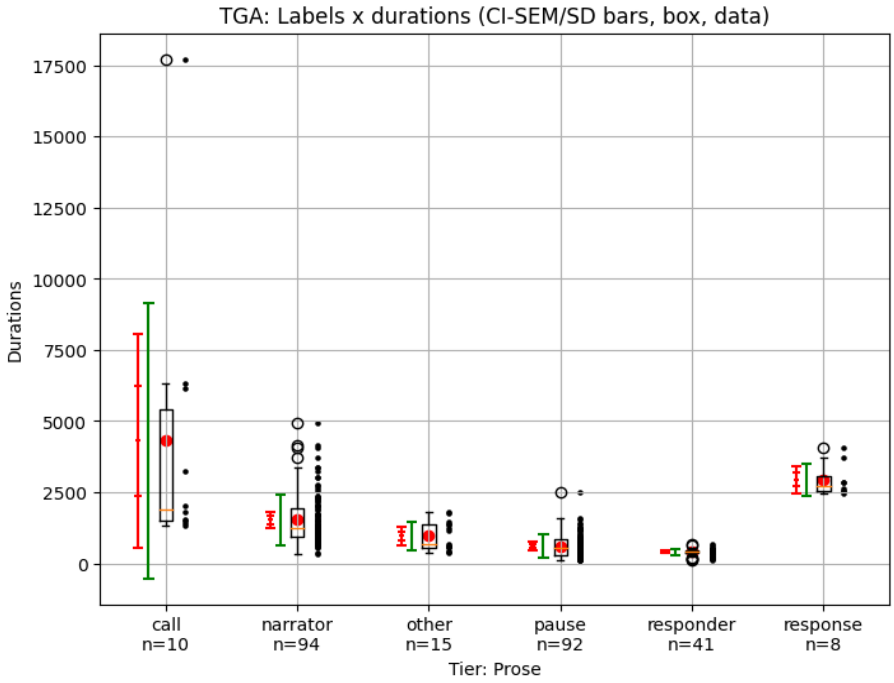

Figure 6: Box and whisker plot of event distributions in annotation of Ega story.

Table 3: Descriptive statistics for utterance categories in the Ega story (durations in ms).

Label	n	min	max	range	mean	median	SD
narrator	94	339	4918	4579	1535	1237	897
other	15	357	1813	1456	962	665	497
response	8	2441	4044	1603	2948	2725	562
responder	41	97	677	580	405	417	118
call	10	1342	17721	16379	4305	1899	4824
Total:	168						

From Table 3 and the boxplot in Figure 6 the following properties of the utterance categories can be inferred as intuitively plausible explicanda for later analysis:

1. The *narrator* utterances and the *responder* backchannel utterances are significantly different, as the familiar rule of thumb of non-overlapping boxplots indicates. The responder typically briefly interjects the ritual response [sɛsɛ] with either HH level or HL falling pitch, meaning roughly: "Mhm!" or "Aha!".
2. The same applies to the *call* category, which is distinct from both the *narrator* and the backchannel *responder* utterances.
3. The distribution of the *narrator* utterances (*n*=94, mean duration 1.535s), which are interpausal units, not necessarily complete turns, is not at all compact (i.e. has low kurtosis), indicated also by the relative values of standard deviation, 897 and standard error, 146; the distribution is also skewed with longer utterances in the tail.
4. The *responder* utterances (*n*=41, mean 0.405s) on the other hand are much shorter and very compact, i.e. have a high kurtosis value (cf. standard deviation, 118).
5. The *call* utterances (*n*=10, mean 4.305s) are less compact, but they are fewer in number; two compact groups are shown in the graph (SD 4824).
6. The *response* utterances (*n*=8, mean 2.948s) are even fewer in number, but are very compact (SD 562) in relation to the *call* utterances.

The expectation that the *responder* and *response* utterances are more compact and putatively more ritualised, while *narrator* and *call* are not, is supported by the quantitative result. This result is likely to be a useful guide to prioritising future studies of genre, style and register, and also for identifying lexicalised or idiomatic components of the story interaction.

3.2. Time and frequency domain properties of the utterance categories

The descriptive statistical results are systematic and useful to a certain extent, but they do not reflect the dynamics of the story as a complex interactive event in a long-term temporal domain. In order to visualise the sequential temporal properties of the annotation for the entire five minutes of the story, the technique illustrated in Figure 4 was used. Durations of the scenario parameters are shown in the top panel of Figure 7. The interpretation of Figure 7 is exactly the same as that of Figure 4, the long-duration domain representation of the Ega greeting, except that the duration is almost ten times as long.

The upper panel represents more categories than those in the example analysis.

1. Each event is rendered in sequence along the *x*-axis of the diagram as a horizontal bar, with colours or shapes for each category. The *x*-axis represents the mid-positions of each event in time, over a total length of 5 minutes (300s).
2. The length of the event interval is represented by the *y*-axis of the diagram, which has a logarithmic scale in order to accommodate the outlier events more easily, and also in the lengths of the bars.
3. There is no temporal overlap between events (with two exceptions, in which the *responder* slightly overlaps with the *narrator*).

The immediately obvious properties of the sequence are:

1. The durations of the *call* component of the song sequences: the long-duration non-narrative events are easily identifiable as three song segments of the narration, in which the narrator leads the song and the audience responds.

2. The brevity of *responder* feedback and the extreme brevity of the pauses.
3. A tendency to somewhat cyclical (shorter)-longer-shorter interpausal unit length patterns in the contributions of the narrator.

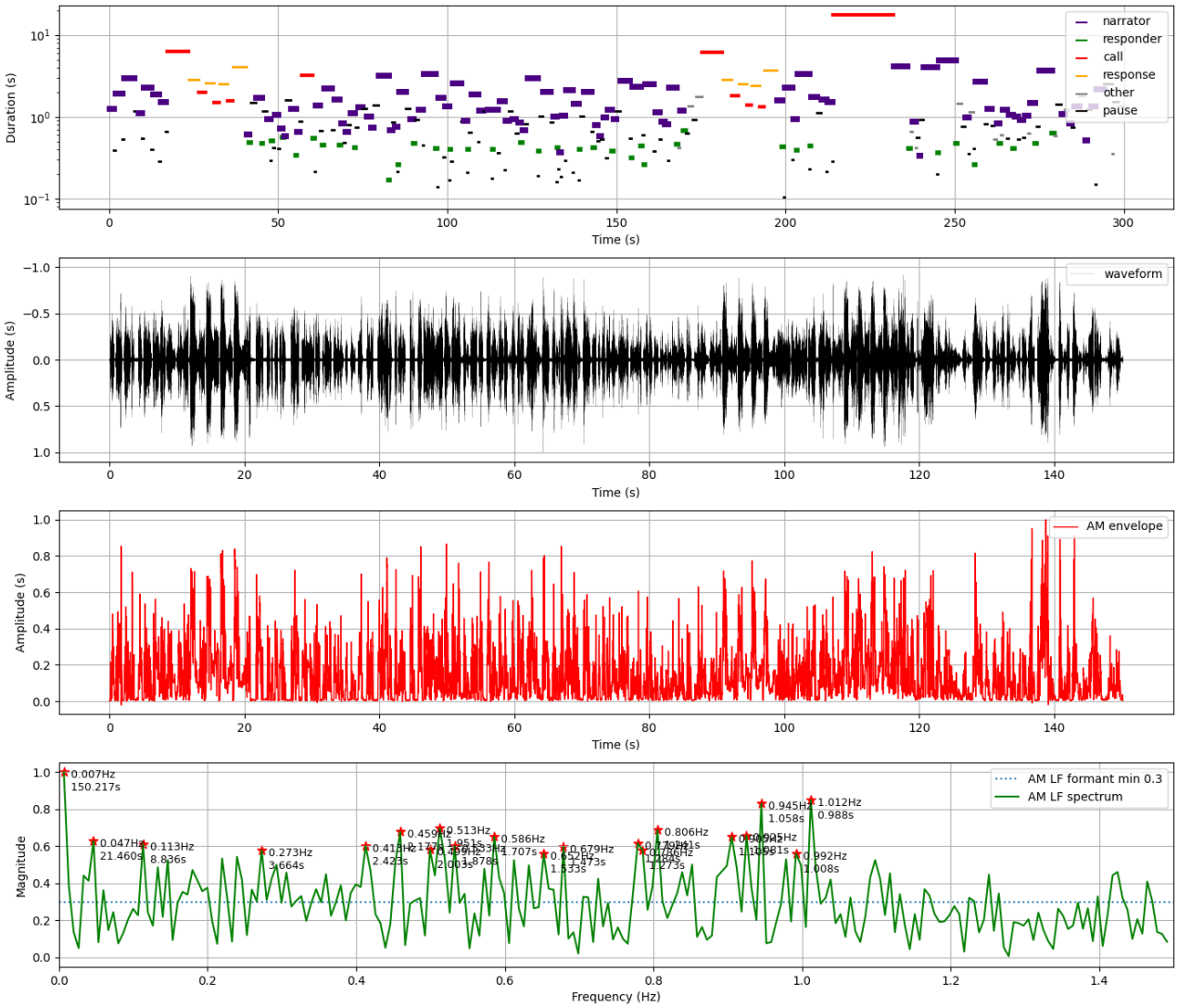

Figure 7: Time and frequency domain representations of Ega story: duration distributions (top: indigo=narrator, red=call, orange=response, green=responder, black=pause); waveform (upper mid), amplitude track (lower mid), AM LF spectrum (lowest).

The lower mid panel of Figure 7 shows the AM track with the regular peaks which were already observed for the waveform of the story.

The lowest panel of Figure 7 visualises the long-term LF spectrum, in the frequency domain, and shows 20 marked frequencies in high-magnitude LF zones, grouped into the fuzzy regions which are interpreted as rhythm formants. There are several rhythm formants which are relevant for discourse structure (i.e. 1s duration or more), which can be visually inspected in the interpausal unit pattern panel and the waveform and AM track pattern of Figure 7:

1. 150s, i.e. 2min30s (0.007Hz rhythm), up to the middle high-magnitude *call-response* singing;
2. 21s (0.048Hz rhythm), involving the length of periods involving singing, and some other amplitude changes;
3. 1.7s to 2.2s (0.59Hz to 0.45Hz), approximately matching some *narrator* interpausal units.

4. 1s to 1.2s (1Hz to 0.83Hz), matching other *narrator* interpausal units as well as *other* contributions.

The discourse category-to-event interval matching is not exact and there is also a great deal of variation even in the more regular contributions to the story, as shown in Table 3. Nor would the rhythms of telling a story be expected to be clockwork-regular. However, the long term AM spectrum indicates useful regions for further investigation of discourse functions using either conversational analysis or other pragmatic, ethnolinguistic and sociolinguistic methods for long-duration sequences, or grammatical methods for short-duration sequences.

4. Modulation theoretic background

Information is conveyed by speech in two main ways: by amplitude modulation (AM) or frequency modulation (FM) of a higher-frequqncy carrier wave. Modulation theory has been familiar in the context of FM and AM radio signals for over a century. In principle, the only difference between modulation of radio signals and modulation of audio signals, apart from the different medium, lies in their frequencies: in the radio domain above 100kHz through several GHz, and in the audio domain below about 10Hz. The carrier-modulation distinction of modulation theory is related in part to the source-filter theory of speech production (Traunmüller 1994).

The carrier wave consists either of the consonantal noise of obstruents (which will be ignored), or vocalic harmonic sounds whose source is the larynx, at frequencies from about 50Hz for very deep male voices to about 600Hz for infants. The carrier wave is not a sine wave, but complex, having an approximately ‘sawtooth’ form which is composed of many integer multiples (harmonics or overtones) of the carrier frequency.

Frequency modulation (FM) of the carrier (and therefore also of its harmonics) in speech is the most fundamental kind of modulation and conveys information about tones, pitch accents and intonations: the information signal varies at a much lower frequency than the carrier wave, lower than about 10Hz (corresponding to event intervals longer than about 100ms). The slowly changing information signal is added at source to the carrier signal, whose frequency then changes as the information signal changes, yielding the FM signal. The FM procedure necessarily produces parallel modulations of the harmonics of the carrier wave.

A highly idealised model (in reality sawtooth-like, not sinusoid) of features of speech modulation which are relevant for the present study can be formulated as follows:

1. A higher-frequency sinusoid carrier signal, $S_{car} = A_{car} \cos(2\pi f_{car} (t+\varphi_{car}))$, f_{car}>50Hz
2. A lower frequency information signal for FM, $S_{fm} = A_{fm} \cos(2\pi f_{fm} (t+\varphi_{fm}))$, $f_{fm} << f_{car}$
3. A lower information signal for AM, $S_{am} = A_{am} \cos(2\pi f_{am} (t+\varphi_{am}))$, $f_{am} << f_{car}$
4. Addition of the FM signal to the f_{car} component of the carrier signal at each point in time, resulting in a modulated carrier signal with frequencies which vary slowly around the original carrier frequency.
5. Multiplication of the resulting FM signal with the scaled A_{car} component of the FM signal resulting from stage 4 at each point in time, leading to amplitude changes in the FM-modulated carrier signal which vary slowly around the amplitude of the original carrier signal.

Tht is, amplitude modulation (AM) in speech is essentially by a multiplication operation which filters the FM signal with different configurations of the oral tract, also in a time domain lower than 10Hz, by means of the alternating sequence of low-amplitude consonantal events (silent in the case

of unvoiced stops), and high-amplitude vocalic events. The different configurations of the vocal tract modulate the amplitudes of the harmonics of the carrier in complex ways, producing the spectral patterns with formants, that is, the regions of harmonics with higher magnitudes which characterise vocalic events.

FM, that is, the modulated F0, results in the varying perceived pitch of tones, pitch accents and intonations, and of sequences of these, which relate functionally to phrase configuration marking, turn-taking, rhetorical and emotional characteristics of speech in story-telling, argumentation, reporting and other kinds of spoken interaction. AM, i.e. filtering of the FM signal, results in varying perceived volume of the signal, which relates functionally to syllables and sequences of syllables in morphemes, words, phrases, and larger units. In lexical tone and lexical pitch accent languages, FM and AM overlap: FM also contributes to distinguishing syllables, and in languages with morphological tone FM also has morphemic meanings.

The task of macrostructural discourse phonetics, in modulation theoretic terms, is to demodulate the complex AM and FM signal into its two modulation components and perform spectral analysis on the demodulated AM and FM tracks, i.e. the LF AM track of the signal is demodulated into the AM information signal and the LF FM track of the signal is demodulated into the FM information signal. For this purpose, the Rhythm Formant Analysis (RFA) method is introduced, which analyses long-term acoustic properties of the signal by means of AM and FM demodulation and then spectral analysis of each of the demodulated AM and FM information signals.

AM demodulation in speech takes place in two different time domains:

1. Short-duration phone-related domains between 5 and 20ms, with high-frequency spectral analysis up to about 3kHz of the amplitude variations of the harmonics of the carrier wave, leading to identification of the high-frequency regions of amplitude variation of the carrier harmonics (high-frequency phone formants) which characterise phones and syllables.
2. Long-duration rhythm-related domains above about 2s (in fact arbitrarily longer), with LF spectral analysis showing frequency regions of AM variation (LF rhythm formants) which characterise rhythms related to syllables, words and phrases.

FM demodulation in speech takes place in similar domains to the long-duration domains of AM:

1. Estimation of the frequency of the carrier wave (F0 estimation, ‘pitch’ tracking) in time windows around 5ms to 20ms which are long enough to match the lowest expected frequencies and short enough to avoid matching the second harmonic of the carrier wave frequency, common errors in F0 estimation algorithms.
2. LF spectral analysis of the estimated F0 in order to determine the variations which are due to syllable-sized tones and pitch accents and the variations which are due to the longer intonations.

In the present study, the time window used for the LF spectrum of both AM and FM is the duration of the entire story: 300s, resulting in a spectrum which shows rhythms at different frequencies of syllables, words or phrases during the entire story, as in the LF spectra shown in the lowest panels of Figure 4 and Figure 7 and in the top two right-hand panels of Figure 8. Since both AM and FM have frequency components which relate to the same units of language, in addition to their own properties, it is not surprising that the AM and FM LF spectra are often very similar.

The LF spectra provide a useful picture of rhythmic properties averaged over the whole signal, the rhythms of rhythm. However, they do not show the dynamics of rhythm change during the story. For this, a LF *spectrogram* is required. To calculate the LF spectrogram, a window of 2s is defined,

and a spectrum is calculated in 2s segments progressing through the signal until the entire signal has been analysed. Very long-term FM and frequency domain properties are not generally investigated (an exception: Huber 1988), but are focussed in the present study.

Figure 8 shows the vectors derived by the RFA method which are used to compare different stories, illustrated with the first 8s of the story:

1. The top row shows the waveform with the superimposed AM track, and the LF spectrum of the AM track below 3Hz.
2. The second row shows the FM track (F0 estimation, 'pitch' track), calculated with a custom AMDF (Average Magnitude Difference Function) algorithm, and the LF spectrum of the FM track below 3Hz.
3. The third row shows the AM LF spectrogram in 2s steps and on the right (not used in the analysis) a rhythm formant hierarchy based on the AM spectrum.
4. The fourth row shows the FM LF spectrogram in 2s steps, and on the right (not used in the analysis) a rhythm formant hierarchy based on the FM spectrum.

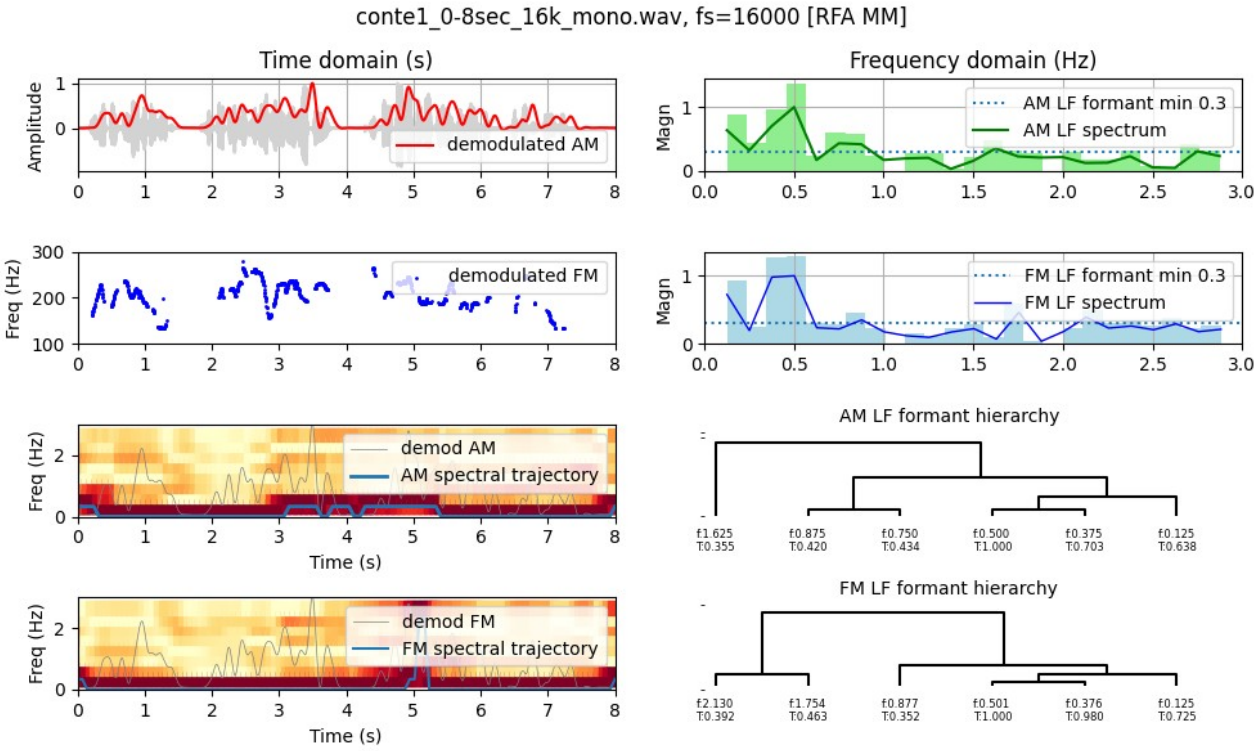

Figure 8: Vectors used in RFA analysis:
time domain (left): AM track, FM track, AM LF spectrogram with top formant trajectory, FM LF spectrogram with top formant trajectory
frequency domain (right): AM LF spectrum, FM lf spectrum, AM formant hierarchy, FM formant hierarchy

It is essential to distinguish:

1. the *FM track (F0 estimation) in the time domain*, with frequencies above 100Hz, on the one hand, and the *FM spectrum in the frequency domain*, on the other, with spectral frequencies below 3Hz in this example;
2. the *FM track in the time domain*, with frequencies above 100Hz and the *highest-magnitude frequency track (top formant track)* in the FM LF spectrogram, also in the time domain, with frequencies below 3Hz in this example.

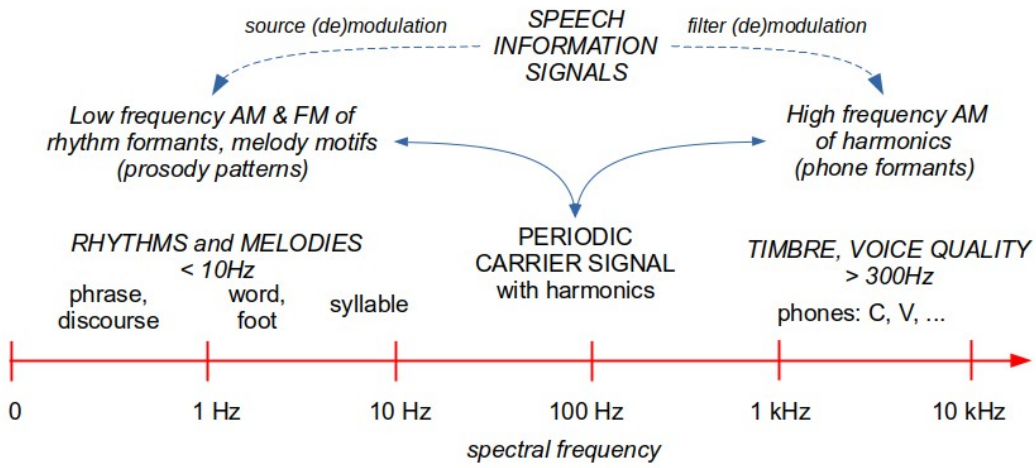

Figure 9: Speech Modulation Scale (SMS): modulation theoretic frequency scale with three frequency regions: prosody (LF), carrier (mid frequency), high frequency (phones)

It is noteworthy that the AM and FM tracks are very different in shape, though timing is closely related and their spectra and spectrograms are somewhat similar but different in detail.

The modulation theoretic model of speech can be summarised as a scale, as shown in Figure 9, which shows a basic division into three spectral regions in the acoustic transmission domain: LF, mid (carrier) frequency and high frequency. The LF domain up to about 10Hz is the area allotted to syllable-oriented FM (tones, pitch accents and sequences of these) and AM (syllable structures and sequences) of the carrier frequency, and is the domain of the present study.

5. Timing and discourse typology

5.1. Interpreting macrostructural discourse phonetics

Having seen an application of the macrostructural discourse phonetic method of long-term spectral analysis, the question of its interpretability and of its uses arises. Indications of its interpretability in terms of a coarsely defined scenario were given in Section 2 and Section 3. More detailed interpretation in terms of information structure, speech acts and dialogue acts, rhetorical features, emotionality and other functional dimensions do not fall within the scope of a first encounter scenario, strictly speaking.

Nevertheless, the idea of a phonetic discourse typology beyond the usual taxonomy of intra-language genres, styles and registers, raises a fascinating question: are macrostructural discourse phonetic properties related more closely to grammatical properties of a given language, or are they specific to a particular culture or language group? Or, on the contrary, are they quite independent of a given culture or language? Or are they in principle idiosyncratic styles?

In order to investigate these questions, a small number of orally presented tales from different Niger-Congo languages were compared using the techniques discussed in the study of Ega stories. The data were originally gathered for discourse and conversation theoretic analysis. The recordings have the following characteristics:

1. Four Ega stories told spontaneously by the same male speaker, but with slightly varying properties such as different relations between narrative and musical content.

2. Two stories told spontaneously in the Indenié dialect of Agni by the same female speaker (also a Western Kwa language); cf. Figure 10.
3. A reading in educated Ivorian French by an Agni speaker (not the speaker of the Agni stories) who was a graduate student at the time.
4. A published story reading (Urua 2004) by a female speaker in Ibibio (ISO 639-3 *ibb*) a Lower Cross language of Nigeria (also Niger-Congo, but not Kwa).

The Ega and Agni community story-tellers share the same story-telling conventions of *narrator-responder-audience* and *call-response* sung interludes. Several contextual variables can be identified which may contribute to differences and similarities between the recordings: spontaneous narration vs. reading (Ega and Agni vs. French and Ibibio); male vs. female (Ega vs. Agni, French and Ibibio); closely related languages (Ega and Agni); distantly related languages (Ega and Agni vs. Ibibio).

Figure 10: Agni storytelling scenario: narrator in the right foreground indicating who is to be the responder; responder in the left foreground; audience in the background background (frameshot from Agni_conte1.mpg, 22s).

The null hypothesis is that no difference will be found. An extreme alternative hypothesis would be that the four data types would separate exactly into four separate clusters, though some overlap might be expected between Ega and Agni (or between Agni and French, due to the French reader's background). A number of pilot analyses were therefore conducted in order to investigate these hypotheses. In fact, intuitively it would not be expected that the data items would show any similarity at all since they are all very complex, particularly in the Ega and Agni cases.

The quantitative output of the spectral analysis method consists of either vectors or summarising statistics over vectors such as mean or variance. The parameters are the LF AM spectrum, the LF FM spectrum and the FM track, as well as the track of highest-magnitude frequencies through the LF spectrograms: the LF AM spectrogram and the LF FM spectrogram.

Lô et al. (2020) demonstrate a supervised (pattern-matching) machine learning approach to West African stories, but they only deal with written versions, not with the orature dialogue and singing. In the present study, the comparisons are made with clustering algorithms (unsupervised machine learning algorithms) based on distance metrics. There are many ways of quantitatively comparing data with these algorithms. The methods chosen here are:

1. the well-known *k-means* clustering algorithm, which groups data specimens according to their Euclidean distance from a pre-defined set of points, *centroids*;
2. distance networks (or distance maps) with selected distance metrics;
3. hierarchical clustering with a selection of distance metrics and clustering linkage techniques.

5.2. Unsupervised machine learning: distance-based clustering

k-means clustering

Very simple derivates of the spectral parameters discussed in Section 4 were used, in particular means and variances of AM spectrum magnitudes and FM spectrum magnitudes, FM (F0), and the highest-magnitude paths through AM and FM LF spectrograms. As expected, different parameter combinations led to different clusterings, not all of which relate plausibly to the contextual variables. Both AM and FM criteria contributed to the cluster formation, though the FM criteria turned out to be more dominant. Two examples with relatively clear clusters were selected.

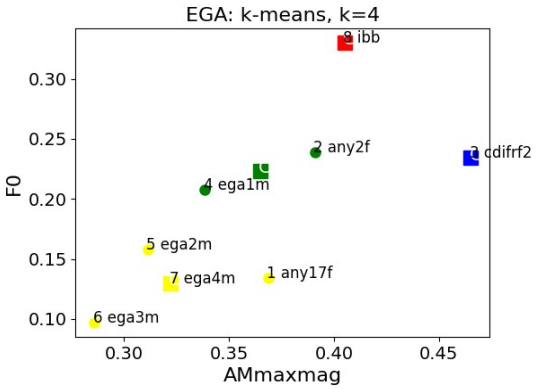

Figure 11: Highest-magnitude path through the AM spectrogram and of F0 (criterion: mean).

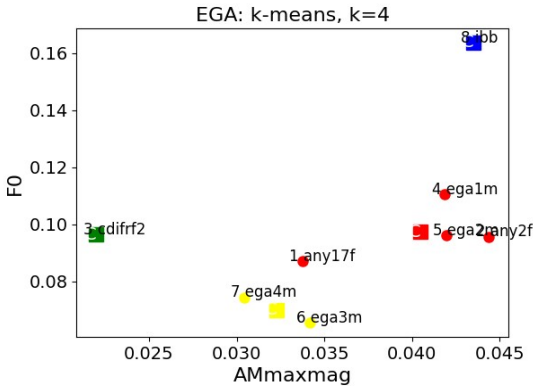

Figure 12: Highest-magnitude path through the AM spectrogram and of F0 (criterion: variance).

In Figure 11 the mean path of highest spectral magnitudes through the spectrogram is compared with the mean F0 for each recording. With these criteria, the two readings are entirely different singleton clusters, while the Ega and Agni stories are clustered together, with three of the four Ega recordings very close together and one Agni recording tending towards the two readings. A similar result is shown Figure 12, which is based on variances in the parameters, not means.

Seen as a simple scatter plot, Figure 11 shows a roughly linear distribution in which the AM maximum magnitude parameter and the F0 parameter have approximately equal influence. The Ibibio recording is separated from the others mainly by F0, and the French reading is separated from the others mainly by AM maximum magnitude. The Agni cases are not completely separated

either by spectrogram highest magnitudes or by mean F0. The Ega cases are also clearly separated by mean AM spectrogram highest magnitudes but overlap with Agni along the mean F0 axis.

Figure 12 shows the same parameters but with variances, not means. Again, seen as a simple scatter plot, Ibibio is clearly distinguished by the variance of F0 and French is differentiated by the variance of AM spectrogram highest magnitudes. Two Ega items are closely related along both dimensions, but both Agni items and the other two Ega items overlap on both dimensions.

Distance networks

Another technique for visualising and comparing clusterings is the *distance network*. A number of different distance metrics were used with the full-length parameter vectors specified in Section 4. The expectation is that Euclidean distance (the 'as the crow flies' distance) is too straightforward, and that Manhattan Distance (Cityblock Distance, Taxicab Distance, Mannheim Distance: the 'round the corners distance') will be more suitable for coping with the sudden value changes in the parameter vectors. The Manhattan Distance network is shown in Figure 13.

The distances are normalised to between 0.0 and 1.0, and a maximum distance cutoff is set to 0.65 to avoid cluttering the graph with distant connections: all nodes are included but there are no distance edges above 0.65. Each node is labelled with a story name, and the edges between the nodes are labelled with the distances: the network is a kind of abstract map.

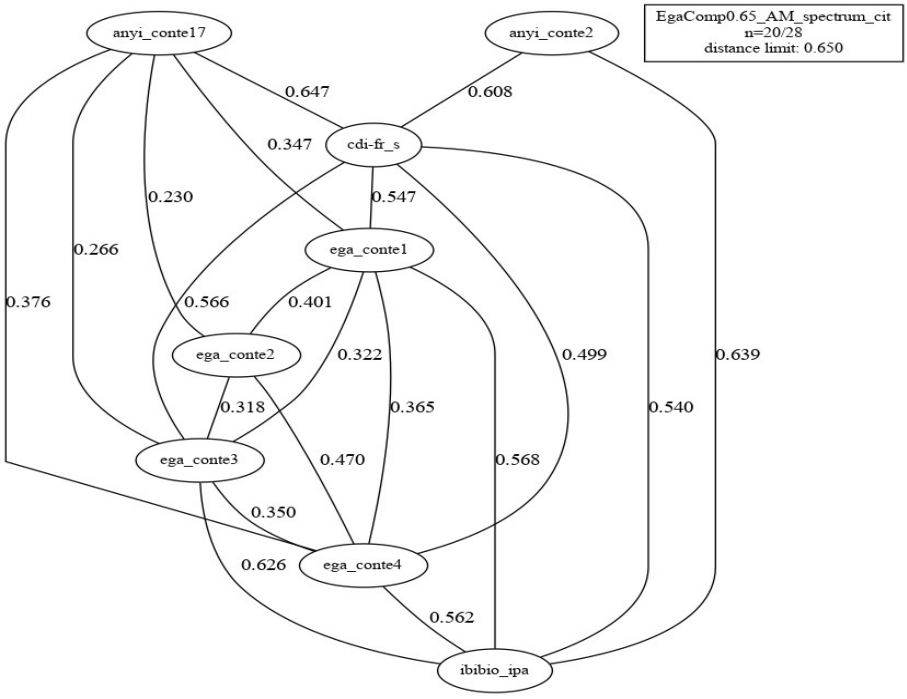

Figure 13: Distance network: AM spectrum, Manhattan Distance.

The distances between nodes are noteworthy:

1. The node representing the Ibibio reading at the bottom is far from its neighbours, with distances of 0.639, 0.626, 0.562, 0.568 and 0.540.
2. The same applies to the node for French reading at top centre, with distances 0.647, 0.608, 0.566, 0.547, 0.540 and 0.499.
3. The two nodes are also relatively far from each other (0.540), possibly motivated by grammatical differences rather than general register features.

4. The distance from the Agni story ‘conte2’ to the two reading nodes is closer than the distance to the other Agni story and to the Ega nodes, which is greater than the cutoff value; this requires explanation.
5. Except for ‘conte2’, the Ega and Agni spontaneous story-telling nodes are closer to each other than to the reading nodes and the Agni ‘conte2’ node, with four edges, all of which have values of 0.470 or less, all less than the edges from the reading nodes.
6. The Ega nodes and one Agni node are not clearly separated; this requires explanation.

That the Agni and Ega nodes are not clearly separated is unsurprising: Agni is uncontroversially a Kwa language and Ega is commonly classified as a Kwa language, and the Agni and Ega communities share traditional spontaneous story-telling genres with *narrator-responder* interaction interspersed with *call-response* singing. In both Figure 11 and Figure 13, the Agni story ‘conte2’ is clearly separated from the other Agni and Ega story-telling sessions, requiring re-examination of the situational variables and the speech variety. Auditory inspection of the ‘conte2’ recording leads to the impression that the story-telling style in this case does in fact have very regular time and frequency domain patterns, like the two recordings of reading. The speech style apparently exercises more influence on similarities and differences than the contrast between the task-oriented register of reading and traditional spontaneous story-telling.

Hierarchical clustering

A different perspective on relationships between the data items is given by hierarchical clustering procedures. In this comparison, the AM and FM spectra are used. Interestingly, Figure 14 and Figure 15 show that for this criterion selection, with Manhattan Distance and the complete linkage clustering criterion, the AM and the FM configurations are the same, though numerical details are slightly different. This is not an accident: both sonority curve and tonal details of the FM curve are linked to syllable patterns. However, this is not the case for all clustering criterion combinations, which focus on different LF signal properties.

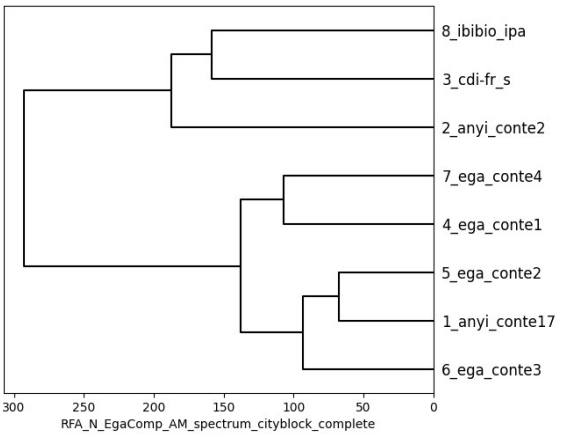

Figure 14: Hierarchical clustering dendrogram: AM spectrum trajectory, with Manhattan Distance and the complete (farthest neighbour) linkage criterion for cluster formation. Note: AM ≈ FM result.

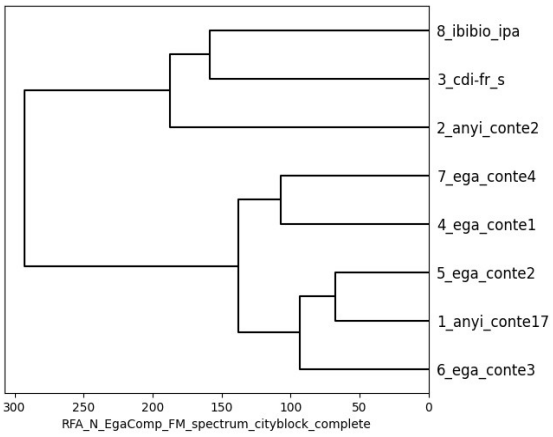

Figure 15: Hierarchical clustering dendrogram: FM spectrum trajectory, with Manhattan Distance and the complete (farthest neighbour) linkage criterion for cluster formation. Note: FM ≈ AM result.

The grouping reflects the results from the distance network analysis: the Agni story ‘conte2’ is more closely attached to the data items in the reading register, Ibibio and French, than to the other Agni and Ega items. The other Ega and Agni items are intertwined.

6. Summary, conclusions, outlook

A novel approach within digital humanities is introduced, with a set of new methods in macrostructural discourse phonetics, Rhythm Formant Analysis (RFA), applied to the Ega language. The aim is primarily to give holistic long-term characterisations of prosodic patterns in traditional spontaneous story-telling sessions. The techniques are designed to handle long-term stretches of speech with durations of more than the standard 10 seconds frequently found in the literature on syllable, word and phrase prosody. These macrostructural discourse phonetic techniques include the visualisation of durations of phonetic events and, applied to the Ega story, they cover *narrator* interpausal segment events, *responder* events, *call* and *response* song events, *pause* segments and contribution segments by audience members.

In addition to visualisation of phonetic event durations, the analysis of low-frequency (LF) properties of the full interval duration of the story-telling event was investigated by means of applying spectral analysis (Fast Fourier Transform, FFT) to the amplitude modulation (AM) track of the entire story waveform. The highest-magnitude frequencies in the resulting spectrum were characterised as rhythm formants, which relate to the timing patterns of syllables, words and phrases.

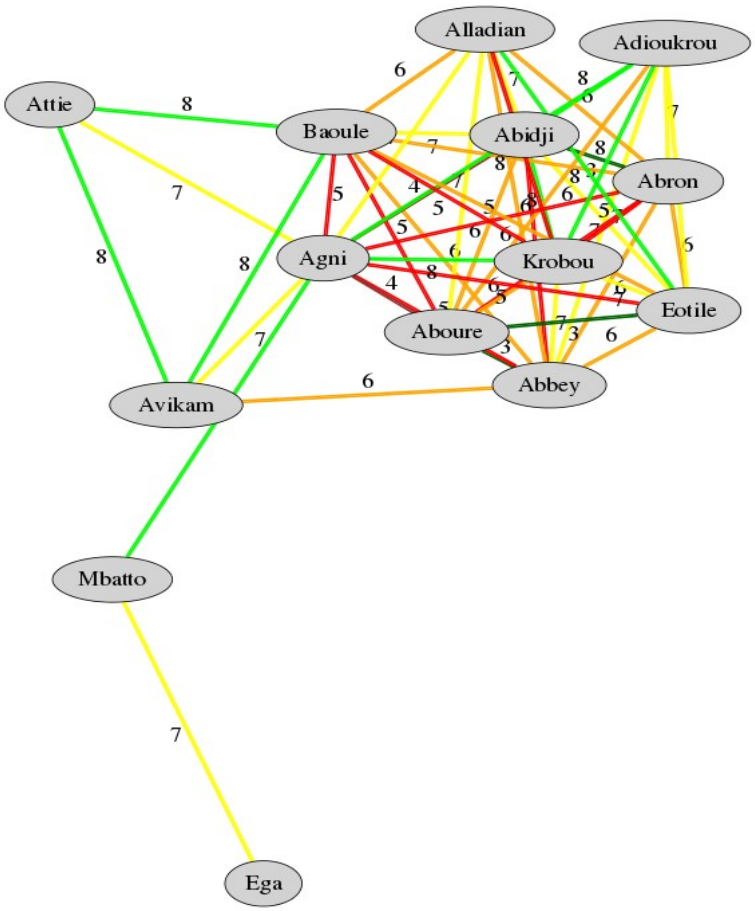

Figure 16: Distance map of Kwa languages of Côte d'Ivoire based on phoneme inventories from Hérault (1983). Map taken from interactive model: http://wwwhomes.uni-bielefeld.de/gibbon/DistGraph/distgraph-kwa.html

A distinction was made between information-carrying AM and FM variations in utterances. The FM properties of speech were not dealt with explicitly in this study. Following an introduction to novel macrostructural discourse phonetic techniques, the techniques were applied in an

unsupervised machine learning analysis to a small opportunistic corpus of similar traditional spontaneous story-telling narratives in a discourse typology study, four in Ega and two in Agni, and reading aloud in Ibibio and educated Ivorian French.

The result is that the Ega and Agni stories, in related languages and with similar story-telling traditions, clustered together, with one exception, and that read-aloud Ibibio and Ivorian French clustered together. The exception, an Agni story whose presentation style resembles reading aloud, in contrast to the other presentations, clustered with the Ibibio and Ivorian French data items. It must be noted, of course, that the study is exploratory in nature, and with such a small data set, no claim of far-reaching generalisations for Ega and Agni and beyond can be made, not unlike phonological studies with small numbers of native speakers.

The combined use of macrostructural discourse phonetic variables and functional variables suggests that interpretations of these macrostructural discourse phonetic typology results could be honed down to finer characterisations in terms of the tiny temporal events which we traditionally study in phonology and phonetics, for example with either the distributions of phones or at a more abstract level of phonological typology. An example would be to compare the discourse analysis results with similar machine learning comparison of lexical inventories.

That the latter is possible is shown by the distance network shown in Figure 16, based on phoneme inventories from the *Atlas des Langues Kwa* (Hérault 1983). Nets of this kind can be used together with macrostructural discourse phonetic typological maps in order to establish the extent to which macrostructural discourse phonetic differences are essentially rhetorical, based on style, register and genre types and their correspondences in different communities, or whether the differences are to some extent determined by the microstructural properties of speech. Distance networks can also be compared with geographical maps in studies of the relation between linguistic typological features and geographical distribution and migration history. It is no coincidence that the distance network in Figure 16 shows Ega to be far away from other Western Kwa languages, nor is it a coincidence that tenacity of customs and areal contact is reflected in the close relations between Agni and Ega in the distance network of Figure 13 and the dendrograms of Figure 14 and Figure 15.

A further dimension which quantitative phonetic modelling of Ega and related languages can offer is technological. The development of text and speech technological tools for local languages, to be used in schools, offices and documentation, such as language-specific keyboard codings, online dictionaries or speech recognition and synthesis, requires *inter alia* the kind of corpus-based documentation, qualitative study and quantitative modelling outlined in the present study and in other work on Niger-Congo languages of West Africa (Gibbon and Urua 2006; Ekpenyong 2012). In this sense, the present approach can be seen as a contribution to Applied Digital Humanities.

7. Acknowledgments

This study would not have been possible without close cooperation over more than four decades with Prof. Firmin Ahoua, Université Houphouet Boigny, Abidjan, and for over twenty years with Dr. Sandrine Adouakou, Université Houphouet Boigny, Abidjan and Universität Bielefeld. A profound debt of gratitude is owed to Father Oko Towe Cyprien for introducing us to Gnaoré Marc and the community of Gniguédougou and the surrounding villages, and to Prof. Eno-Abasi Urua and Dr. Moses Ekpenyong of Uyo University, Nigeria, for close cooperation in a three-country

project with Germany, Côte d'Ivoire and Nigeria which has helped to introduce computational perspectives into fieldwork-based documentation of West African languages.

8. References

Berry, J.; Spears, R. 1991. *West African Folktales*. Evanston, IL: Northwestern University Press.

Bole-Richard, Rémy. 1983. Ega. In: *Atlas des langues Kwa de Côte d'ivoire*, Vol 1. ed. G. Herault. 359-401. Abidjan: Institut de Linguistique Appliquée.

Connell, Bruce, Firmin Ahoua and Dafydd Gibbon. 2002. Illustrations of the IPA: Ega. *Journal of the International Phonetic Association* 32/1, 99-104.

Couper-Kuhlen, Elizabeth & Margret Selting. 2018. *Interactional Linguistics. Studying Language in Social Interaction*. Cambridge: Cambridge University Press.

Ekpenyong, Moses Effiong. 2012. *Speech Synthesis for Ibibio*. Ph.D. dissertation, University of Uyo.

Gibbon, Dafydd. 2016. Legacy language atlas data mining: mapping Kru languages. *Proc. LREC 2016*. Paris: ELDA.

Gibbon, Dafydd. 2018. The Future of Prosody: It's about Time. Keynote. In *9th International Conference on Speech Prosody*.

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

Gibbon, Dafydd & Eno-Abasi Urua. 2006. Morphotonology for TTS in Niger-Congo languages. Proc. 3rd International Conference on Speech Prosody. Dresden: TUD Press.

Gibbon, Dafydd, Catherine Bow, Steven Bird and Baden Hughes. 2004. Securing interpretability: the case of Ega language documentation. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC 2004, Evaluation and Language resources Distribution Agency (ELRA). pp.1369-1372.

Hérault, Georges, ed. 1983. *Atlas des langues kwa de Côte d'Ivoire*. Abidjan: Université d'Abidjan, Institut de linguistique appliquée.

Huber, Dieter. 1988. Aspects of the communicative function of voice in text intonation: constancy and variability in Swedish fundamental frequency contours. Ph.D. dissertation, Department of Computational Linguistics, University of Göteborg.

Lô, Gossa, Victor de Boer and Chris J. van Aart. 2020. Exploring West African Folk Narrative Texts Using Machine Learning *Information*, 11, 236. doi:10.3390/info11050236

Ninan, O. Deborah, George O. Ajíbádé, Odétúnjí Àjàdí Odéjobí. 2016. Appraisal of Computational Model for Yorùbá Folktale Narrative. In: Miller, Ben, Antonio Lieto, Rémi Ronfard, Stephen G. Ware, and Mark A. Finlayson, eds. *7th Workshop on Computational Models of Narrative* (*CMN 2016*), 1-14.

Pike, Kenneth L. 1947. P*honemics: a technique for reducing languages to writing*. Ann Arbor: The University of Michigan Press.

Poeppel, David & Maria Florencia Assaneo. 2020. Speech rhythms and their neural foundations. Nature Reviews Neuroscience 21, pp. 322-334.

Rossini, Nicla and Dafydd Gibbon (2011). Why gesture without speech but not talk without gesture? *Proc. GESPIN 2011*, Bielefeld

Tilsen Samuel & Keith Johnson. 2008. LF Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*. 124 (2): EL34–EL39. 2008. [PubMed: 18681499].

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

Traunmüller, Hartmut. 1994. Conventional, biological, and environmental factors in speech communication: A modulation theory. In: Dufberg, Mats & Olle Engstrand, eds. *Experiments in*

Speech Process. *PERILUS* XVIII. Department of Linguistics, Stockholm University, 1-19. (Also: *Phonetica* 51:170-183, 1994).
Urua, Eno-Abasi Essien. 2004. Ibibio. *Journal of the International Phonetic Association* 34 (1), 105–109.