RHYTHM FORMANTS OF STORY READING IN STANDARD MANDARIN

Dafydd Gibbon

Abstract: Rhythm Formant Theory (RFT), a modulation-theoretic approach to the physical modelling of speech rhythm, is described and applied in an exploratory analysis of the rhetorical rhythms of read-aloud Mandarin Chinese translations of the IPA benchmark text *The North Wind and the Sun*. Rhythm Formant Analysis (RFA), a methodology for empirically investigating Rhythm Formant Theory without prior annotation of the speech signal, is presented in some detail, with the aim of studying rhythm variation in larger units throughout longer texts, rather than restricting analysis to words, phrases and sentences. A test case of read-aloud narratives was investigated, with the null hypothesis that male and female readers do not differ in rhetorical reading strategies. RFA was used to generate vectors of low frequency (LF) variation in spectrograms, for analysis with hierarchical clustering methods. The clustering indicates that the null hypothesis was falsified and rhetorical differences between female and male speakers were tentatively confirmed. Ongoing work includes the analysis of linguistic factors underlying LF variation. In the conclusion, RFT is placed into a more general framework of a Speech Modulation Frequency Scale of modulation types.

Keywords: Rhythm Formant Theory, speech rhythm, rhythm formant, rhythm spectrogram, modulation-theory.

1. SPEECH RHYTHM

The goal of the present study is to demonstrate core aspects of the relation between rhythmic patterns in speech, in terms of amplitude modulation (related to sonority curves) and frequency modulation (related to tones, pitch accents, intonations), and from an empirical and descriptive rather than an intuitive and structural point of view.

There have been many linguistic and phonetic studies of the rhythmic patterns of speech in the past, in three main paradigms, in addition to rhetorical and pedagogical accounts:

1. A phonological paradigm with taxonomic and generative sub-paradigms, in which rhythm is an abstract structure consisting of relations between stronger and weaker abstract stress values, sometimes expressed numerically (cf. Pike [38]; Jassem [25]; Abercrombie [1], [2]; Chomsky, Halle and Lukoff [8]; Chomsky and Halle [9]; Liberman and Prince [32]; Selkirk [42]).
2. A linguistic-phonetic paradigm, in which abstract units such as voiced phone sequences or syllables are annotated with time-stamps and interval durations from the speech signal, using software tools. The annotations are used for deriving speech rate, pause patterns, degrees of isochrony (equal timing), using descriptive statistics for distinguishing between languages in terms of duration irregularities (cf. Lehiste [30]; Roach [40]; Jassem et al. [26]; Scott et al. [41]; Ramus et al. [39]; Low et al. [33]; Asu and Nolan [4]; Li and Yin, [31]; Yu and Gibbon [47]).
3. A modulation-theoretic paradigm in which rhythms are modelled as low frequency (LF) oscillations in the speech signal, with signal processing applications in both speech production and speech perception models (Dogil and Braun [15]; Todd et al. [45]; Cummins and Port [12]; Cummins and Schmidthuber [11]; O'Dell and Nieminen [36]; Foote and Uchihashi [16]; Barbosa [5]; Galves et al. [17]; Lee and Todd [29]; Tilsen

and Johnson [44]; Heřmanský [23]; Ludusan et al. [34]; Barbosa and da Silva [6]; Inden et al. [24]; Tilsen and Arvaniti [43]; He and Dellwo [22]; Varnet et al. [46]; Gibbon [19]; Gibbon and Li [20]; Kallio et al. [27]; Ludusan and Wagner [35]).

The first two approaches are deductive, based on intuition-based linguistic categories, often in a search for correlations in the speech signal or in physiology. The phonological paradigm has shown how rhythms relate to language structure, but does not address the issue of physical empirical grounding. The linguistic-phonetic paradigm does address physical grounding, and has provided heuristics for distinguishing languages in terms of differences in timing irregularity (partial isochrony, timing similarity). But it has often been pointed out that this approach does not in fact describe rhythm in the usual sense of regular beats or oscillation. These approaches also fail to identify rhythm because their use of descriptive statistical methods introduces a methodological problem: these methods apply to sets of static data, not to the dynamic time series data of speech signals which characterise physical properties of rhythm. Time series data require different methods. The second paradigm has been critically discussed by Dauer [13], Barry et al. [7], Gibbon [18], Gut [21]; Arvaniti [3] and Kohler [28].

The modulation-theoretic paradigm, on the other hand, is inductive and provides a physical grounding which complements the other two paradigms. A related early account of rhythm as 'waves' rather than partially isochronous units was given by Pike [37]. Physical utterances are the starting point in this paradigm. Utterances are analysed in terms of component modulation frequencies, mainly amplitude modulation (AM) in the spectral region below about 10 Hz, leading to generalisations about rhythms as timing patterns at different frequencies. The present study extends this modulation-theoretic paradigm by (a) permitting analysis of long texts (rather than words and phrases) (b) including frequency modulation (FM) in the domain of investigation, (c) introducing the concept of *rhythm formant* (RF), as opposed to *phone formant* (PF) to interpret high magnitude LF spectral peaks, and (d) extending LF spectral analysis to the *low frequency rhythm spectrogram* in order to analyse rhythm variation through longer speech instances.

2. METHODOLOGY

2.1 Basics of Rhythm Formant Analysis

The purpose of this and the following subsections is to introduce the *Rhythm Formant Analysis* (RFA) methodology with 'clear case' data, for the purpose of understanding the *Rhythm Formant Theory* (RFT) of AM and FM information in the speech signal.

The LF amplitude modulation (LFAM) of the speech signal is related to the sonority curve, a complex epiphenomenon resulting from the coordinated sequencing of consonants and vowels and from the emergent distribution of high frequency (HF) *phone formants* in the speech signal spectrum, which are due to resonances in the articulatory tract.

The LF modulation of the fundamental frequency (F0) of the speech signal (LFFM) is the direct correlate of tones, pitch accents and intonation.

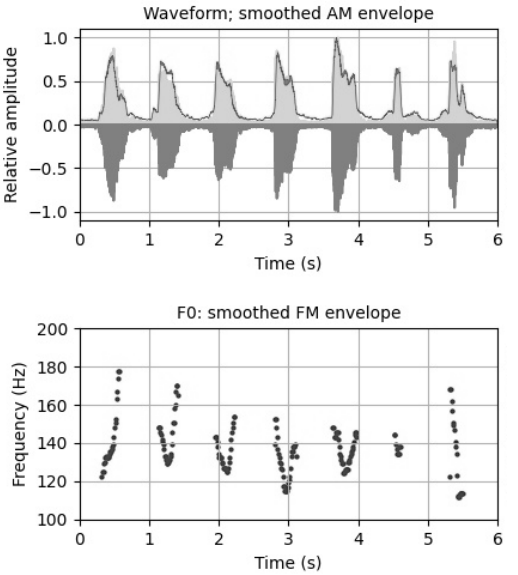

Figure 1: AM and FM envelopes for English counting one to seven (adult male speaker).

Figure 1 shows the amplitude and frequency modulation of a counting sequence from one to seven in Standard British English. The AM component is shown by outlining the positive

amplitudes of the signal, shown in the upper half of Figure 1, i.e. the *amplitude modulation envelope* (*AM envelope*). The FM component is provided by the fundamental frequency estimation function (informally: 'pitch track'), shown as the *frequency modulation envelope* (*FM envelope*) in the lower half of Figure 1. The FM envelope is extracted with the *Average Magnitude Difference Function*, is which is similar to autocorrelation but uses window subtraction rather than multiplication.

The AM envelope corresponds roughly to the sonority curve of weak consonantal syllable margins which frame stronger vowel nuclei, resulting in an alternating or oscillating pattern. The sonority curve also delineates larger units than the syllable, for example phrases or 'paragraph-sized' utterance segments with changing amplitudes.

The differences between the AM envelopes of each numeral in the sequence are rather small (except for 'six', a short vowel flanked by voiceless consonants), and it is obvious where the rhythmic effect originates: the syllable-sized patterns are very similar and occur in similar time-slots.

Rhythms have frequencies. This obvious fact is ignored in many earlier analyses. The main frequency of the AM envelope of the numerals in this sequence can be calculated informally by observing that 7 numerals divided by 6 seconds yields a speech rate of 1.167 numerals per second, a frequency of 1.167 Hz with an average unit duration or beat as the inverse of the frequency: 6/7 = 0.857 s, 857 ms. The overall amplitude envelope pattern also contains lower frequencies from longer term patterns, as well as higher frequencies due to shorter syllable components.

The AM and FM envelopes have the same main frequency, due to alignment of the stress-pitch accents with the numerals, but the detailed pattern of the FM envelope is quite different, being determined by three main factors:

1. Initial boundary tone, which involves a very high rising F0 in this sequence.
2. Similar local fall-rise F0 patterns in the subsequent numerals (corresponding to the 'head' of the intonation unit, in traditional terminology), with a stress-pitch accent sequence (SPAS) constraint on the head which creates expectations of up-coming structures (Dilley [14]). Each local F0 pattern contributes to higher frequency rhythms, with still higher frequencies determined by consonantal perturbations of the local F0 pattern at syllable margins.
3. Final boundary tone, which starts higher and finishes lower than the previous local patterns, and contributes to a much lower phrasal frequency pattern.

The AM and FM envelopes represent modulations of the basic carrier signal which is generated either in the larynx or by obstruent 'noise', and are demodulated in order to recover the rhythmic information generated in speech production for input to speech recognition and perception. Four assumptions are involved in the demodulation procedure:

1. Rhythms can be discovered by spectral analysis of the AM and FM envelope modulation by using low-pass filtering and the Fourier Transform, and visualised in the LF spectrum over the entire utterance.
2. Spectral analyses of the AM envelope and the FM envelope are similar in revealing high magnitude LF spectral zones, interpreted as *rhythm formants*. The AM and FM frequency zones differ in detail, because of differences between syllable sonority patterns and pitch accent patterns.
3. Languages with different prosodic typology show different pitch patterns. Specifically, lexical and morphological tone languages show more varied LFFM spectral structure than languages with no SPAS constraint, because lexical tones are arbitrary phonemes and not constrained to be similar.
4. Variations in rhythm over time can be shown by frequency spectrograms of either AM or FM envelopes.

The claims #1 to #3 are discussed in the following subsections. The AM part of claim #4 is treated in Section 3 in the context of the distance-based clustering of reading aloud by Mandarin Chinese female and male readers.

2.2 Low frequency spectral analysis

The assumptions outlined in the previous subsection can be taken as predictions and investigated operationally by extracting AM and FM envelopes from the speech signal and

applying a Fast Fourier Transform (FFT) to the envelopes to obtain the LF spectra. Figure 2 shows the LF spectra (1...5 Hz) which result from applying the FFT with appropriate low-pass filtering to the AM and FM envelopes. The text boxes in the two graphs show the 5 highest magnitude frequencies in each spectrum.

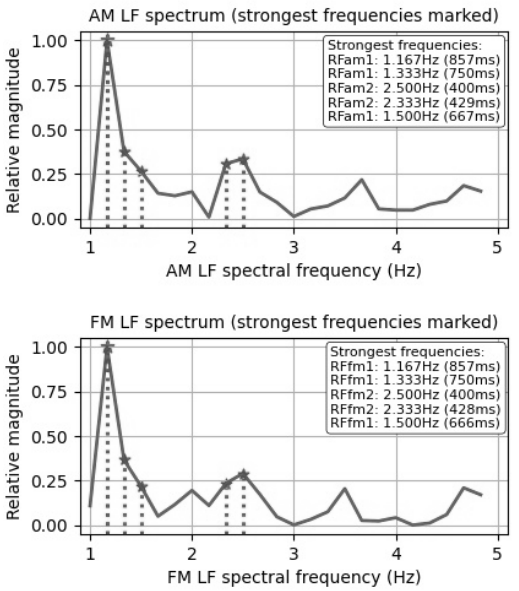

Figure 2: LF FFT spectra of AM (upper) and FM (lower) envelopes of the English counting sequence.

Evidently, the AM and FM spectra agree closely even without correlation calculation: in each case there is an AM high magnitude frequency zone around 1.3 Hz, peaking at 1.167 Hz. This zone is a rhythm formant corresponding to the frequency of numerals in the counting sequence, which thus have an average duration of about 750 ms. The result agrees roughly with the manual calculation.

A second rhythm formant occurs at around 2.42 Hz, corresponding to an average period of 413 ms. It is no accident that the frequency region of the second formant is twice that of the first formant: the AM and FM envelopes both show a binary syllable structure. The AM numeral envelopes have a secondary peak due to the voiced codas (or nasal syllable in 'seven'). In the FM envelopes a binary structure also dominates due to the fall and rise components of the pitch contours.

The term 'formant' is defined acoustically in exactly the same way for these LF rhythm formants (RF) as it is for HF phone formants (PF). In the speech production domain, the LF patterns are, of course, not resonances of the vocal tract, but neural resonances or oscillations which 'conspire' to drive the muscle patterns of articulation. In the speech perception domain, the functions differ: perception of an LF resonance is rhythm, perception of an HF resonance is sound quality.

2.3 FM envelope: tone vs. stress-pitch

The prediction that FM envelope rhythm differs in tone and stress-pitch accent languages can be demonstrated by comparing the English counting sequence with a Mandarin counting sequence, in Pinyin transcription *yī èr sān sì wǔ liù qī*, with high, fall, high, fall, fall-rise (dipping), fall and high lexical tones (i.e. tones 1, 4, 1, 4, 3, 4, 1). The null hypothesis is that the two languages do not differ in relevant ways.

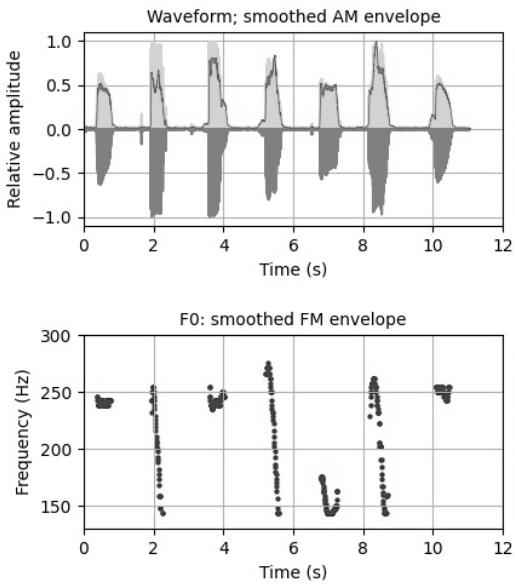

Figure 3: AM and FM envelopes for Mandarin counting sequence one to seven (female).

An example of counting from 1 to 7 in Mandarin Chinese by a female subject is shown in Figure 3. Informal observation shows that average F0 is higher, as expected. The individually determined speech rate of the Mandarin speaker is also considerably slower than the speech rate of the English speaker: an informal count shows 7 numerals in just under

11 seconds, i.e. around 7/11 = 0.636 items/s, 0.636 Hz, averaging 1.571 s per numeral.

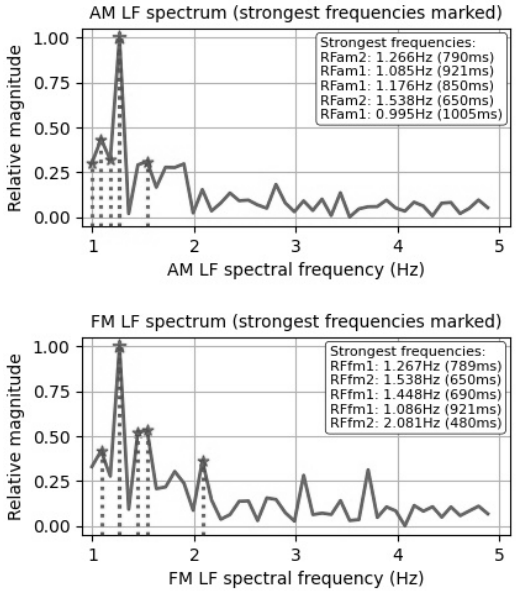

Figure 4: AM and FM LF spectra for Mandarin counting sequence one to seven (adult female).

The overall AM envelope pattern is similar to the English pattern, with expected lower LF formant frequencies and some differences related to the variation in Mandarin syllable structure. Measurements yield a somewhat variable first LF formant of about 1.085 Hz and an average unit duration of about 921 ms for both the AM envelope rhythm. These values are expected because of the slower tempo than in the English counting, and reflect the informal observation of differences in numeral rate and duration. A more precise measurement would need a higher spectral resolution than was possible with these recordings.

The phonotactic diversity of the Mandarin syllable structures is reflected in the different spectral patterns between 1.5 Hz and 2 Hz, due to the absence of the SPAS constraint.

In addition to the difference between the English and Mandarin spectral patterns, the variety of Mandarin syllable patterns and lexical tone patterns, with absence of the SPAS constraint, leads to differences between the Mandarin AM and FM spectra which are not observable for English: the English LFFM and LFAM rhythm formants (Figure 2) are relatively compact and resemble each other, while the Mandarin LFFM and LFAM rhythm formants (Figure 4) are more diffuse.

Although there are general similarities, the conspicuous differences do not support the null hypothesis of strong similarity between the rhythm formant analyses of the English and Mandarin counting sequences.

2.4 Data

One major empirical innovation of RFT and its RFA methodology is that in contrast to earlier approaches they permit the analysis of rhythm variation over entire narratives. The data used in the small-scale study in the following section consist of a read-aloud translated narrative in Pǔtōnghuà, Standard Mandarin Chinese. The narrative is the IPA benchmark fable, *The North Wind and the Sun*, spoken by 5 female and 5 male native speakers, from the Tongji University Yu corpus. The data were analysed in a preliminary study, and are further investigated here with hierarchical classification, a basic variety of unsupervised machine learning, using vectors derived from the RFA LF spectrogram.

Table 1: Durations of narratives.

	Durations of narratives (seconds)					**Mean**
F	53.30	38.74	40.90	45.41	53.90	46.45
M	40.50	43.00	54.24	60.93	40.60	47.85

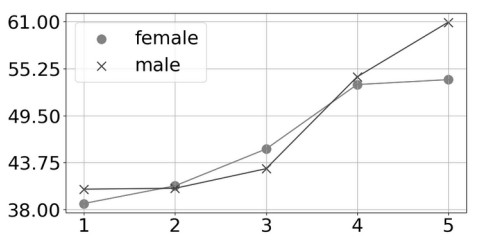

Figure 5: Distribution of recording durations of female and male Mandarin story readers.

The narratives differ considerably in duration, from 38.74 s to 60.93 s, whereby the female and male narratives share roughly similar range and distributions (cf. Table 1 and Figure 5), with female readings on average very slightly faster.

It might therefore be predicted that on the basis of duration, and thus of utterance rate, or indeed of F0 height alone, a null hypothesis of

no difference between female and male speakers can be confirmed. But a more interesting and potentially easily falsifiable null hypothesis is that a modulation-theoretic analysis will find no difference in rhythm between female and male readers.

To investigate this hypothesis an inductive search for hierarchical clusters is undertaken, rather than conventionally looking for flat clusters or sets of correlations. The criterion for clustering is least distance (or: least difference, similarity) between rhythm variation vectors over the LF spectrograms of the readings. The goodness of clustering criterion *G* is the sum of the largest male and female reader cluster sizes, divided by the size of the data set.

3. RESULTS

3.1 Envelope extraction

For the extraction of the AM envelope, the absolute Hilbert Transform is often cited (cf. [22]). However, in this study the waveform is rectified (converted to absolute values) and low-pass filtered, which achieves the same result (Figure 6).

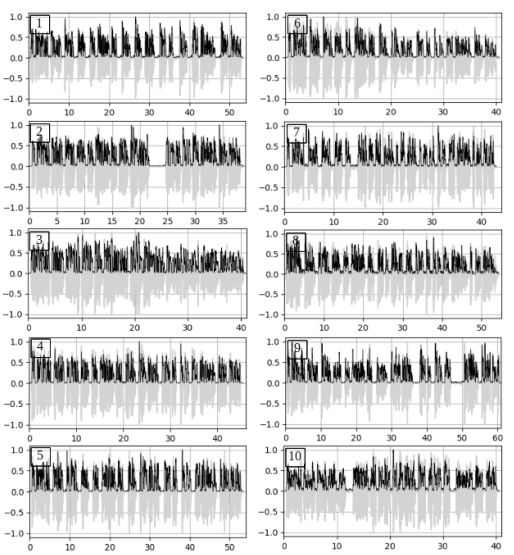

Figure 6: AM waveforms. F (left) and M (right), same order in all graphs.

3.2 Spectral analysis

Figure 7 shows the female (left) and male (right) results, each ordered alphabetically.

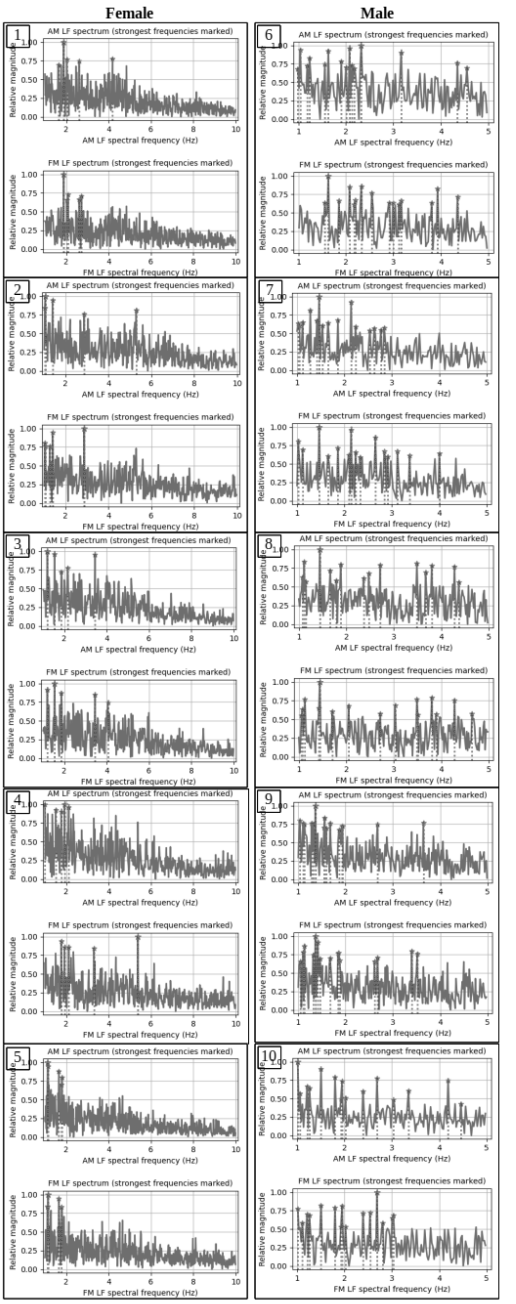

Figure 7: LFAM spectra, peak frequencies marked. F (left) and M (right), same order in all graphs.

Visual examination of the spectra shows conspicuous differences between the spectral shapes and in the distribution of the highest magnitude frequencies. In the spectra of the narratives of female readers, the highest magnitude frequencies tend to cluster around the lower part of the LF spectrum, while in the spectra of the narratives of the male readers, the highest magnitude frequencies tend to be distributed more widely across the LF band. The bandwidth differences suggest that the female readers tend have more regular rhythms, while the male readers tend to have more varied rhythms.

The purpose of presenting spectral analyses of all the readings is not to examine them in detail by 'eyeballing', but to gain an impression of the 'gestalt' of the spectrum in each case. The spectra are purely in the frequency domain and contain no information about time, and thus no information about temporal variation through the narration. Temporal information is added by the *rhythm spectrogram* analysis described in the following section.

3.3 Rhythm variation in time

Intuitively, one simple way of distinguishing male from female readings is to determine the height and range of F0 patterning. The present task is more interesting from a phonetic and linguistic perspective: to determine whether the temporal 'rhetorical' variation of rhythm within the narrative aligns with the gender distinction. Other factors such as practice, social role, age, dialect, factors conditioning pitch height and range, utterance rate, etc. could be considered, but these are matters for future study.

To enable the study of rhythm variation, the next stage in the combined frequency and time domain analysis of rhythm is to generate *LF spectrograms* (cf. Figure 8). Again, the purpose of presenting all the spectrograms is not for detailed analysis, but to gain initial visual impressions with which general tendencies can be observed.

The spectrograms in these examples consist of the low frequency end of the vertically oriented spectra resulting from FFTs of overlapping windows, following standard spectrographic conventions. The episodic structure of the narrative readings is shown in the vertically oriented dark clusters showing simultaneous rhythms, contrasting with intervening stronger horizontal tendencies indicating more homogeneous rhythm patterns. Rhythms vary, but according to this criterion there are also coherent rhythmic stretches.

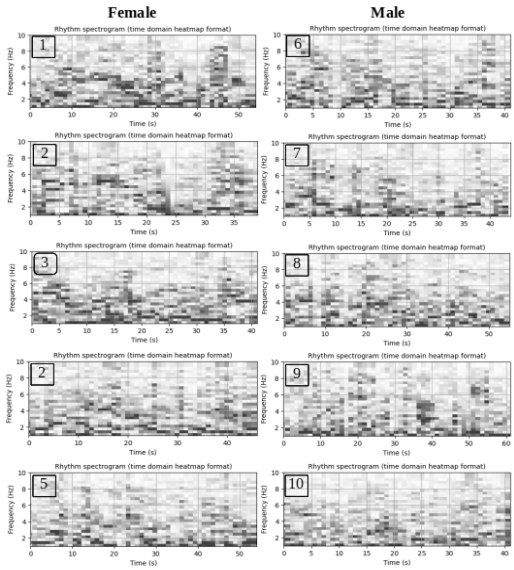

Figure 8: LFAM rhythm spectrograms (1 Hz to 10 Hz) in heatmap format. Same order in all graphs. Highest to lowest magnitude: dark to light colour.

The vectors for comparison in hierarchical clustering are extracted from the highest magnitude points in each component spectrum of the spectrogram:

1. It is assumed that the time domain factor of utterance rate will affect the rhythm analysis; however, isolation of this factor by time normalisation is a topic for future study.
2. It is assumed that modulation frequencies in the AM envelope, the physical analogue of the abstract phonological 'sonority curve', constitute the main physical property which contributes to speech rhythm.
3. It is assumed that the combined frequency and time domain variation factor $\Delta F0$ in the FM envelope is relevant for rhythm analysis, but that other frequency domain factors such as height and range are not. FM rhythm variation is not dealt with in this study.
4. An FFT is applied in a 2 s moving window with an overlap step parameter, in order to derive a 3-dimensional spectrogram.

5. A rhythm variation vector consisting of the highest magnitude frequency (HMF) in each window in the LF spectrogram is constructed as input for distance-based hierarchical clustering algorithms.
6. The HMF vector is a vector of pairs of magnitude and frequency, which can be separated for further analysis into a magnitude vector and a frequency vector.

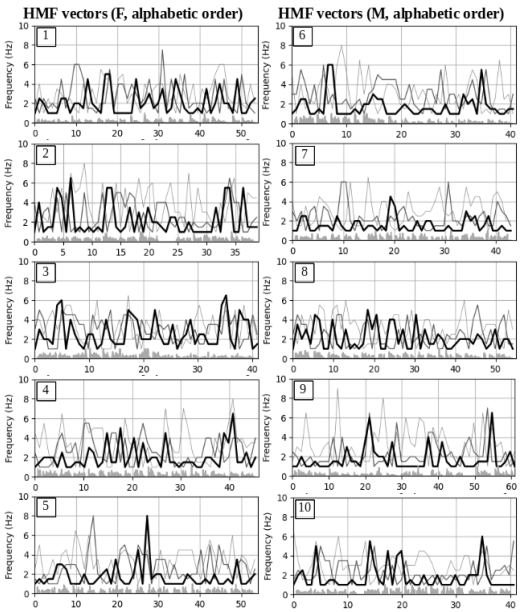

Figure 9: Rhythm formant trajectory variation: female (left), male (left), visually time-normalised. Same order in all graphs.

HMF vector analyses of both AM and FM envelopes can be made, but within the scope of the present study, analysis is restricted to AM results. Figure 9 shows AM frequency trajectories for readings by female readers (left half) and male readers (right half). The three highest magnitude frequencies are represented for each component spectrum in the spectrogram (darker means higher magnitude). Clearly, this can only be the beginning of the exploitation of the rhythm spectrogram as a source of information about rhetorical features of the reader's speech.

The readings are of slightly different lengths, but the graphs are visually scaled to show corresponding relative times across readings. In the following discussion, the graphs are numbered #1 to #5 (F set, left column) and #6 to #10 (M set, right column). A number of similarities are apparent. First, the F patterns #1 to #4 are impressionally similar, and one male pattern (#8 in consecutive numbering) is similar to the F patterns. One F pattern, #5, is distinctly different from the other F patterns, and its central spike has similarities with M patterns #7, #9 and #10. Male patterns #9 and #10 are similar to #6, though #6 is slightly different from the other two. On this basis it might be predicted that F reading #5 would tend to cluster with the M readings, and M reading #8 would tend to cluster with the F readings.

The spectral patterns and the HMF vectors are also influenced by pause patterns (cf. Figure 6), which are clearly somewhat different in each reading. It remains to be seen how these and the vectors relate to the narrative structure of the text. Analysis of text-linguistic and of sociophonetic determinants of the patterns is a task for later specialist text linguistic studies and sociolinguistic experiments. The present study is restricted to acoustic classification.

3.4 Classifying rhythm variation

The task of the hierarchical clustering module in the RFA approach is to determine whether the informal characterisation of differences between male and female readings stands up to algorithmic evaluation.

Given the HMF vector for each reader, whether based on AM or FM envelopes, similarities between the vectors can be determined by using distance metrics, and by hierarchically clustering the vectors and vector clusters based on their pairwise distances. Hierarchical clustering is preferred here over standard flat clustering, since it is more informative and provides a simple but plausible criterion for the goodness of clustering *G*.

For the distance calculation, six well-known distance metrics were compared: Manhattan Distance (also: Cityblock or Taxicab Distance), Canberra Distance (also: normalised Manhattan Distance), Euclidean Distance, Cosine Distance and Chebyshev Distance (also: Chessboard Distance). The Manhattan and Canberra 'round the corner' metrics are intuitively more suitable for comparing somewhat irregular data, while the Euclidean 'as the crow flies' metric takes shortcuts, and appears less suitable. The

Chebyshev metric is intermediate between the two. Cosine Distance shows differences in direction or orientation, not actual distance. The Manhattan and Canberra metrics are assumed to be most relevant for the present task.

Four clustering methods are selected for applying the distance measures: average distance of cluster members (Average Clustering), furthest distance of cluster members (Complete Linkage, Farthest Point, Voorhees Clustering), distance between the closest cluster member (Single Clustering), and minimal variance (Ward Clustering). The evaluation criterion for goodness of clustering G is defined as the size of the largest homogeneous cluster of female readings (F) plus the size of the largest homogeneous cluster of male readings (M), divided by the size of the dataset, yielding a maximum of G=10/10=1.0 in the present data set with perfect separation of F and M instances.

An earlier version of the classifier yielded a number of perfect partitions into five F and five M instances, with G=1 (for example with the Chebyshev-Ward combination); in other cases a right-branching, effectively linear scale emerged with the F instances in one half and the M instances in the other (e.g. Manhattan-Average). However, with revised higher resolution spectrogram analysis the M/F distinction was slightly blurred, but still clear.

The two distance metrics which showed relevant results are the Manhattan (Cityblock) and Canberra metrics, as anticipated. Figure 10 shows sample clustering results, with goodness values G=1, G=0.8, G=0.6, G=0.7 (clockwise starting top left).

The clustering methods yielded varied results but shared some tendencies. For example, members of the M set tended to cluster together, while the F set members are more loosely linked at higher hierarchical ranks. In each case, one of the F set, #5, *wx*, intruded into the M linkage. Examination of the trajectory in Figure 9 reveals a possible reason: as already noted, the spike at about 26 s roughly matches peaks in this region in some M set members. Conversely, #8 in the M set tends to intrude into the F set: inspection shows a less 'spiky' pattern than for the other members of the M set (cf. Figure 9).

Another point worth noting is the preservation of local clusters through each cluster type in some cases, e.g. #2 with #3, and #6 with #10. Further investigation is needed in order to understand the fine structure of the rhythm formant trajectories.

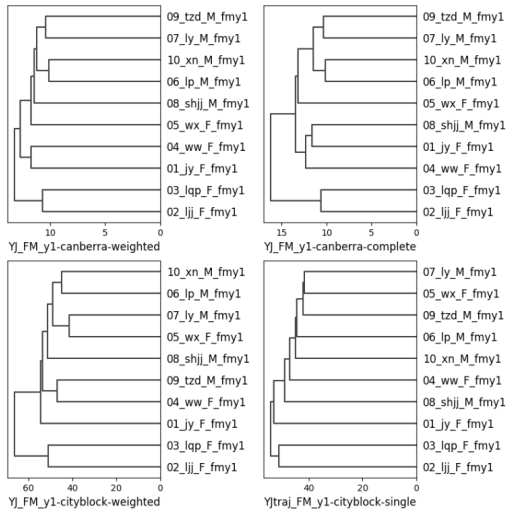

Figure 10: Distance based hierarchical cluster dendrograms for highest magnitude spectrogram frequencies (F set #1 to #5, M set #6 to #10, top to bottom).

Building on the present hierarchical clustering results, it will be interesting in more large-scale studies to examine possible reasons for the partition of female and male readings and for the cluster-internal hierarchies, for example individual styles, conventional gender styles, or physiological differences. Although the separation of the F and M categories is incomplete, it is clear enough to warrant further investigation on the basis of this tentative refutation of the null hypothesis that the M and F sets are indistinguishable.

4. SUMMARY AND OUTLOOK

A modern approach to the study of rhythm, Rhythm Formant Theory, with its Rhythm Formant Analysis methodology, was introduced in terms of modulation theory. RFA was applied to the speech signal in an exploratory analysis with data involving long-term narrative domains, rather than the word or phrase-sized utterances of previous approaches in this paradigm. After a brief account of paradigms of

speech rhythm study in the introduction, the modulation-theoretic background was described, and the specific procedure of modulation-theoretic analysis of both amplitude modulation and frequency modulation of the speech signal was outlined.

Starting with the amplitude modulation (AM) envelope, related to the phonological 'sonority curve', and also the frequency modulation (FM) envelope, the 'pitch' track, low frequency (LF) spectral analysis was motivated and described in a small-scale study of rhythmic variation in narrative readings by ten Mandarin native speakers. A new conceptual orientation was introduced, with the *rhythm formant* and the *rhythm spectrogram* derived from AM or FM envelopes, and the *highest magnitude frequency vector* (HMF vector) through the spectrogram.

The HMF vectors from each reading were subjected to a hierarchical cluster analysis with several distance metrics and clustering methods, which resulted in the tentative falsification of the null hypothesis that there is no difference between the rhythmic patterns used by female and male readers.

Concentration on the frequency property of rhythm in the speech signal involves a different 'mindset' from the traditional approaches which were concerned with manual measurement of duration irregularities of phonologically defined units, missing the resonant or oscillatory property of rhythm frequency.

Taking one more step, the frequency oriented approach to rhythm studies which underlies Rhythm Formant Theory and its methodology, Rhythm Formant Analysis, can be put into a general modulation-theoretic framework of timing patterns, derived originally from Cowell's (1930) classic theory of harmonic relations in musicology, and centring on a logarithmic *Speech Modulation Frequency* (*SMF*) *Scale* (Figure 11).

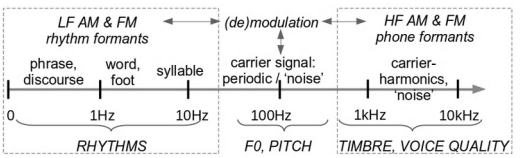

Figure 11: modulation-theoretic frequency scale of carrier and formant types.

This general framework of frequency bands offers a coherently structured model of the 'carrier signals' and 'modulation signals' in speech communication, and provides motivation and orientation for further development of large-scale, machine-learning supported modulation-theoretic studies of speech and language prosody.

5. ACKNOWLEDGMENTS

Debts of gratitude are owed to the referees for insightful comments; to Yu Jue, Liu Huangmei, Ma Qiuwu, Gong Qi, Li Peng and Lin Xuewei for data and discussions on many aspects of Chinese; to Doris Bleiching, Laura Dilley, Alexandra Gibbon, Sascha Griffiths, Ulrike Gut, Rosemarie Tracy and Ratree Wayland for technical and practical discussions and much encouragement; to the late Grzegorz Dogil and the late Wiktor Jassem for their pioneering work and for many discussions on methods and applications of envelope extraction and spectral analysis.

6. REFERENCES

[1] Abercrombie, D. 1965. *Studies in Phonetics and Linguistics*. London: OUP.

[2] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

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

[4] Asu, E-L., F. Nolan. 2006. Estonian and English rhythm: a twodimensional quantification based on syllables and feet. *Proc. Speech Prosody 3*.

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

[6] Barbosa, P., W. da Silva. 2012. A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches. *Proc. Int. Conf. on Computational Processing of the Portuguese Language*, 1-9.

[7] Barry, W.J., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. 2003. Do rhythm measures tell us anything about language type? *Proc. 15th ICPhS*, Barcelona, 2693–2696.

[8] Chomsky, N, M. Halle, F. Lukoff, 1956. On accent and juncture in English. In: M. Halle, H. Lunt, H. Maclean, eds., *For Roman Jakobson.*

Essays on the Occasion of his Sixtieth Birthday, 11th October 1956. Den Haag: Mouton, 65–80.

[9] Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

[10] Cowell, H. 1930. *New Musical Resources*. New York: Alfred A. Knopf Inc.

[11] Cummins, F., F. Gers and J. Schmidhuber. 1999. Language identification from prosody without explicit features. *Proc. Eurospeech*, 371–374.

[12] Cummins, F., R. Port. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26, 145–171.

[13] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11:51–62.

[14] Dilley, L. C. 1997. The Phonetics and Phonology of Tonal Systems. Dissertation MIT.

[15] Dogil, Grzegorz and Gunter Braun. 1988. *The PIVOT Model of Speech Parsing*. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

[16] Foote, J., S. Uchihashi. 2001. The beat spectrum: a new approach to rhythm analysis. *Proc. IEEE International Conference on Multimedia and Expo, Tokyo*.

[17] Galves, A., J. Garcia, D. Duarte, C. Galves. 2002. Sonority as a basis for rhythmic class discrimination. *Proc. Speech Prosody 1*, Aix-en-Provence: Laboratoire Parole et Langage, 323–326.

[18] Gibbon, D. 2006. Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In: Sudhoff, S., D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, J. Schließer, eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, 281–209.

[19] Gibbon, D. 2018. The Future of Prosody: It's about Time. Keynote. *Proc. Speech Prosody 9*. https://www.isca-speech.org/archive/SpeechProsody_2018/pdfs/_Inv-1.pdf

[20] Gibbon, D., Peng Li. 2019. Quantifying and Correlating Rhythm Formants in Speech. *Proc. 3rd International Symposium on Linguistic Patterns in Spontaneous Speech*. Academia Sinica. Taipei, Taiwan.

[21] Gut, U. 2012. Rhythm in L2 Speech. In: Gibbon, D., D. Hirst, N. Campbell, eds. *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem. Special Edition of Speech and Language Technology* 14/15. Poznań: Polish Phonetics Society. pp. 83–94.

[22] He, L., V. Dellwo. 2016. A Praat-Based Algorithm to Extract the Amplitude Envelope and Temporal Fine Structure Using the Hilbert Transform. *Proc. Interspeech, San Francisco*, 530–534.

[23] Heřmanský, Hynek. 2010. History of modulation spectrum in ASR. *Proc. International Conference on Acoustics, Speech and Signal Processing* (ICASSP).

[24] Inden B, Z. Malisz, P. Wagner, Ipke Wachsmuth. 2012. Rapid entrainment to spontaneous speech: A comparison of oscillator models. *Proc. 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 1721–1726.

[25] Jassem, W. 1952. *Intonation of Conversational English (Educated Southern British)*. Wrocław: Wrocławskie Towarzystwo Naukowe.

[26] Jassem, W., D. R. Hill, I. H. Witten. 1984. Isochrony in English Speech: Its Statistical validity and linguistic relevance. In: D. Gibbon, H. Richter eds. *Intonation, Accent and Rhythm. Studies in Discourse Phonology*. Berlin: Walther de Gruyter, pp. 203–225.

[27] Kallio, H., A. Suni, J. Šimko, M. Vainio. 2020. Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics* 80, 1–12.

[28] Kohler, K.. 2009. Editorial: Whither Speech Rhythm Research? *Phonetica* 66: 5–14.

[29] Lee, C. S., N. P. M. Todd. 2004. Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition* 93 (3): 225–54.

[30] Lehiste, I. 1970. *Suprasegmentals*. Cambridge, Mass.: MIT Press.

[31] Li, A., Z. Yin. 2006. A rhythmic analysis on Chinese EFL speech. *Proc. Speech Prosody 3*, Dresden, Germany.

[32] Liberman, M. Y., A. Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249–336.

[33] Low, E. L.; Grabe, E.; Nolan, F. 2000. Quantitative characterisations of speech rhythm: 'syllable-timing' in Singapore English. *Language and Speech* 43: 377-401.

[34] Ludusan, B., A. Origlia, F. Cutugno. 2011. On the use of the rhythmogram for automatic syllabic prominence detection. *Proc. Interspeech*, 2413–2416.

[35] Ludusan, B., P. Wagner. 2020. Speech, laughter and everything in between: A modulation spectrum-based analysis. *Proc. Speech Prosody 10*, 25–28 May 2020, Tokyo, Japan, 995–999.

[36] O'Dell, M. L., T. Nieminen. 1999. Coupled Oscillator Model of Speech Rhythm. *Proc. XIVth International Congress of Phonetic Sciences. San Francisco*, 1075–1078.

[37] Pike K. L. 1962. Practical Phonetics of Rhythm Waves. *Phonetica* 8:9-30.

[38] Pike, K. L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[39] Ramus, F., M. Nespor, J. Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265–292.

[40] Roach, P.. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In: Crystal, David, ed. *Linguistic Controversies: Essays in Linguistic Theory and Practice*. London: Edward Arnold, 73–79.

[41] Scott, D. R., S. D. Isard, B. de Boysson-Bardies. 1985. Perceptual isochrony in English and French. *Journal of Phonetics*, 13:155–162.

[42] Selkirk, E. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: The MIT Press.

[43] Tilsen, S., A. Arvaniti. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America* 134, 628–639.

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

[45] Todd, N. P. McAngus, G. J. Brown. 1994. A computational model of prosody perception. *Proc. International Conference on Spoken Language Processing* (ICLSP-94), 127–130.

[46] Varnet, L., M. Cl. Ortiz-Barajas, R. G. Erra, J. Gervain, C. Lorenzi. 2017. A cross-linguistic study of speech modulation spectra. *Journal of the Acoustical Society of America* 142 (4), 1976–1989.

[47] Yu, J., D. Gibbon, K. Klessa. 2014. Computational annotation-mining of syllable durations in speech varieties. *Speech Prosody* 7.

Dafydd Gibbon is emeritus professor of Linguistics at Bielefeld University, Germany. His main research interests cover computational linguistics, prosody, the documentation of under-resourced languages, and standards and resources for speech technology.