Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Speech rhythms: learning to discriminate speech styles

Dafydd Gibbon

Bielefeld University, Germany
gibbon@uni-bielefeld.de

Abstract

This study addresses the role of continuous speech rhythms in characterising speech styles. The assumption is that speech rhythms are wave-like oscillations with frequencies below 10 Hz and that the oscillations can be detected in a holistic approach to analysing low frequency spectral peaks in the amplitude and frequency modulations of speech. Superimposed rhythms at different frequencies are taken to vary within formant-like low spectral frequency zones, depending on speech style. Further, it is assumed that speech rhythms are not simply cognitive epiphenomena but are physically sufficiently distinct to characterise speech styles in the signal. The study combines signal processing by rhythm formant analysis (RFA) and basic unsupervised machine learning (ML) with detailed interpretation and explanation of the phonetic ML results by comparison with linguistically annotated data. The data comprise two styles from the Aix-MARSEC English speech style database. Results confirm the relation of continuous rhythms to speech styles.

1. Introduction

Rhythm is an emergent holistic property of the temporal patterning of speech signals. In phonology and linguistic phonetics, speech rhythms are seen as correlates of sequences of morae, syllables, words and phrases, which vary from language to language [8]. This qualitative concept of speech rhythm is atomistic, however, and fails to capture the 'beats' of continuous rhythms and rhythm variation. Further, it would be a category mistake to atomise description of continuous rhythms into the usual 'features' of duration, frequency and intensity, for three reasons: first, frequency and intensity patterns are related time functions, while durations are components of the time axis and thus conceptually different; second, these components are projected into the signal from phonological abstractions rather than inductively generalised.

The fundamental claim is that in local domains rhythm is a function of morphology, and thus of linguistic typology (cf. [1], [13], [21]) but in global domains a function of speech style [14]. The study combines methodological exploration with confirmatory testing in investigating continuous long-term rhythms in different speech styles in a single language, using samples from the Aix-MARSEC database of English speech styles [3]. Complete discourse events, typically 60s long, are analysed holistically by spectral analysis of the whole event providing LF spectra of the amplitude modulation (AM) and the frequency modulation (FM) of the signal as well as LF spectrograms with sequences of spectra over 2s frames.

The present study combines signal processing, in the form of spectral analysis, and basic unsupervised machine learning (ML), in the form of clustering algorithms, with detailed 'phonological interpretation' (as opposed to the 'phonetic interpretation' of traditional phonology) based on annotated data. A key concept is *rhythm formant*: by this is meant an LF formants (<5Hz), i.e. a frequency zone in the LF spectrum which related to linguistic units such as syllables, words or phrases, and whose spectral distribution and variation characterises different speech styles. LF rhythm formants (<5Hz) differ acoustically from high frequency (HF) phone formants (>300Hz) only in frequency; their status in speech production and perception is different.

The variation of continuous rhythms in this sense is studied in many disciplines, from oceanography, cardiology and musicology to clinical phonetics and language typology (cf. [6], [18], [23], [24], [26]). This modulation-theoretic approach has been applied to speech with different spectral analysis techniques (absolute Hilbert transform or rectification and low-pass filtering, with rhythmogram, empirical mode decomposition, Fourier transform) in many studies (e.g. [1], [4], [7], [9], [10], [12], [13], [17], [18], [22], [23], [24], [25], [27], [28], [29], [30], [31], [32]). The present method adds the concepts of *rhythm formant* and *low frequency spectrogram*.

2. Data and Method

2.1. Data

Data are taken from the Aix-MARSEC database of English speech styles [3], selecting the first ten of each of the database categories *News Broadcasts* (category B, 1 female, 3 male), and *Poetry* (H, 1 female, non-rhyming; 1 male, rhyming) as being clear cases of prosodically distinct styles. The categories were intuitively determined, not based on explicit discourse analytic models. Two of the reading-aloud styles were preferred, partly as clear cases with more restricted variables than spontaneous dialogue, but partly because of the intrinsic value of reading styles as social and cultural skills.

2.2. Method: modulation theoretic analysis

The speech signal is analysed as a carrier wave [32] with superimposed LF information-carrying FM (relating to tones, pitch accents and intonation), and AM (relating to sonority cycles). The AM and FM information signals are extracted and analysed in three steps: AM and FM demodulation, LF spectrum and spectrogram transforms, and comparison of AM and FM spectral properties of different data items using basic unsupervised machine learning (clustering, cf. [13]).

2.2.1. Demodulation and spectral analysis

The following systematic RFA (rhythm formant analysis [13]) procedure is followed (implementation in Python3 with the standard libraries NumPy, SciPy and MatPlotLib):

10.21437/SpeechProsody.2022-62

1. Demodulation of the modulated signal in the time domain:
    1. AM: amplitude demodulation (envelope extraction by full-wave rectification and low-pass filtering);
    2. FM: frequency demodulation (normalised fundamental frequency, F0 estimation with an AMDF algorithm);
2. Rhythm detection by spectral analysis of the demodulated AM and FM signal envelopes in the frequency domain:
    1. LF spectrum (0Hz...10Hz by FFT):
        1. AM: long-term FFT of the whole AM envelope;
        2. FM: long-term FFT of the whole FM envelope;
    2. LF spectrogram (0Hz...10Hz, in 2s frames with extraction of a trajectory through the highest magnitude frequency in each frame, 'rhythm formant track'):
        1. AM: LF formant track of the AM envelope;
        2. FM: LF formant track of the FM envelope.

The examples illustrated in Figure 1 (news) and Figure 2 (poetry) show demodulation of both AM and FM envelopes in the time domain (left panels). The AM envelope is extracted by full-wave rectification (absolute signal values) and LF smoothing of the 60s long signal. The FM envelope is extracted by AMDF (Average Magnitude Difference Function). LF spectra (here <3Hz) are obtained by FFT from the demodulated AM and FM signals (utterance duration and thus also FFT frame duration in this example is 58.34s).

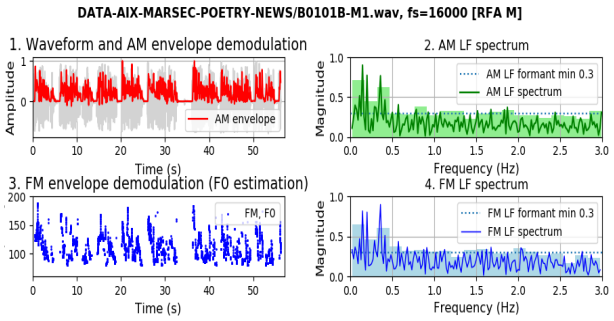

Figure 1: Waveform with AM and FM envelopes (left), and LF spectra (0...3 Hz) for the envelopes (news).

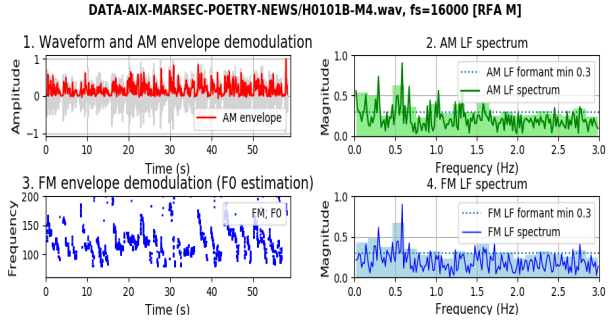

Figure 2: Waveform with AM and FM envelopes (left), and LF spectra (0...3 Hz) for the envelopes (poetry).

Several relevant prosodic properties can already be seen by subjective inspection of the visualisations in Figure 1 and Figure 2: the time domain panels show temporal grouping patterns in the readings; these are particularly clear in the downtrending F0 patterns. The AM and FM LF spectra in the right-hand panels are similar in the spectrum segment below 1Hz, less so above 1Hz. High magnitude frequency peaks tend to group into the formant-like frequency zones which are interpreted as rhythm formants.

The main rhythm formants of the two readings are similar in shape, but at different frequencies. The formant vectors are analysed with two unsupervised machine learning algorithms.

3. Results

3.1. From waveform to LF trajectory

Figure 3 shows the waveform with superimposed AM envelope (panel 1), the AM LF spectrogram in traditional heatmap format (panel 2) and the rhythm formant track through the spectrogram (panel 3), in each case overlaid with the rectified waveform. The FM envelope (F0 track, panel 4), the LF spectrogram (panel 5) and FM rhythm formant track (panel 6) are shown in the lower three panels.

The rhythm formant tracks visualise the constantly changing AM and FM rhythm frequencies. To clarify: the LF FM rhythm formant track in panel 6 should not be confused with the F0 itself, which is shown in panel 4.

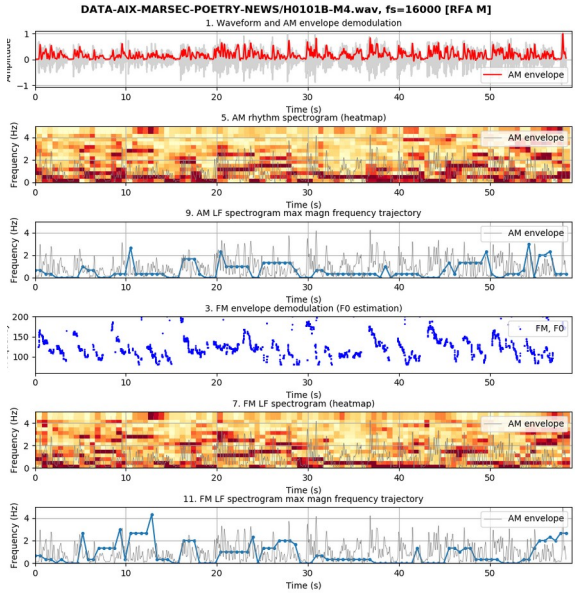

Figure 3: (top to bottom) AM envelope extraction; AM LF spectrogram; AM LF highest magnitude frequency track; FM envelope; FM LF spectrogram; FM highest magnitude frequency track (poetry reading).

3.2. ML: k-means, hierarchical clustering

The twenty data items were compared using two basic unsupervised machine learning (ML) methods: *k-means* clustering and hierarchical clustering. The reasons for using unsupervised rather than supervised ML are simple: the data are too sparse for supervised training on 'big data'; also, unsupervised learning is essentially posterior and inductive, avoiding top-down prior training except for parameter choice.

Two-dimensional *k-means* clustering was used with pairwise combinations of the variances of the five available vectors: *F0 estimation*, *AM spectrum*, *FM spectrum*, *AM spectrogram formant track*, *FM spectrogram formant track*. Variance is used partly as a dimension reduction measure, partly to normalise the parameters. The best separation of styles was given by *AM spectrogram formant track variance* in combination with the *FM spectrogram formant track*

variance or the *F0 estimation variance* (Figure 4 and Figure 5): only reading 15 (H, poetry, male) was an outlier in B (newsreading) territory. The A*M spectrogram formant track variance* was clearly the dominant discriminating factor.

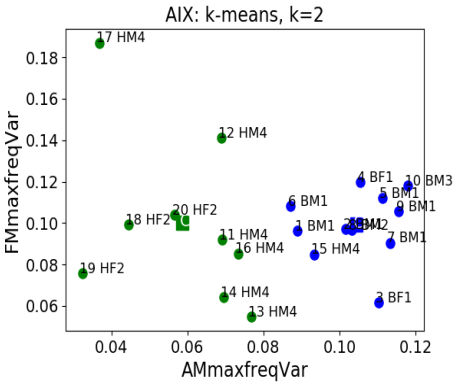

Figure 4: AM spectrogram formant variance ✕ FM spectrogram formant variance. B news, H poetry; gender F, M.

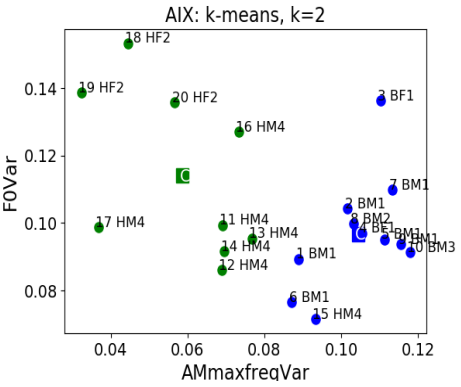

Figure 5: AM spectrogram formant track variance ✕ F0 estimation variance.

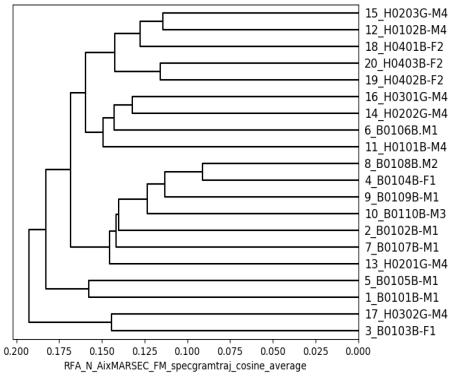

Figure 6: Example of dendrogram for two registers, (B, newsreading, H, poetry reading) with Cosine Distance and average linkage.

Induction of the intuitively perceived style differences using well-defined acoustic prosodic parameters is thus achieved in this small trial. Newsreading clustered more closely than poetry reading, with higher AM spectrogram track variance and less FM (F0) formant track variance. The FM variance in poetry reading and its greater overall variability relates to a livelier impression than for newsreading. Gender has only a small effect.

The same five vectors were used in hierarchical clustering analyses using combinations of 5 distance metrics (Canberra, Chebyshev, Cosine, Euclidean, Manhattan) and 4 linkage criteria (average, Vorhees, single, weighted) applied to the 5 prosodic vectors (100 combinations). Objective assessment of of results was by cluster overlap count. The best result was achieved for the *AM spectrogram formant track* variance, which had already been shown to be effective in the *k-means* variance analysis. The best distance metric was Cosine Distance with average based linkage (Figure 6), expressing orientations of the two styles in the formant track vector space.

4. Discussion: interpreting ML

Which empirical facts underlie the inductive result? As a first step in the task of understanding ML, the spectra, which in effect compress the three-dimensional spectrogram into a two-dimensional summary, are interpreted in detail.

Figure 1, Figure 2 and Figure 3 illustrate spectral analyses of the two styles. Very low frequencies below 1Hz in the AM and FM demodulation spectra are present in each style, though at different frequencies in the different styles: newsreading has two neighbouring main rhythm formants at 0.2Hz and 0.4Hz, relating to interpausal unit (IPU) frequency, while the poetry reading rate in this frequency zone is higher, at 0.3Hz and 0.6Hz. The poetry reading also has clear formants at 1.0Hz, 1.6Hz and 2.1Hz, while the newsreading has a more diffuse distribution of spectral frequencies in this region.

The relatively clear rhythm formants, i.e. zones of neighbouring peak frequencies in Figures 1 and 2 and the spectrogram patterns in Figure 3, suggest an acoustic prosodic hierarchy of rhythms.

This is a first step in understanding inductive ML applied to timing in phonetics. The next step is *phonological interpretation* of the phonetic results (in contrast to *phonetic interpretation* in top-down approaches), in order to enquire how the prosodic AM LF spectrum relates to a prosodic default hierarchy of syllable, word and phrase (or rather IPU, interpausal unit). For this purpose, the top-down analysis and annotation techniques of linguistic phonetics are suitable ([2], [11], [16], [19], [20], [21], [33]).

Table 1: Annotation measurements of newsreading and of poetry reading styles (bold: shorter units, faster rates).

	News			Poetry		
	Syll	**Word**	**IPU**	**Syll**	**Word**	**IPU**
n	242	161	16	189	129	22
median (ms)	**162**	**256**	2305	240	363	**1731**
median rate (Hz)	**6.17**	**3.91**	0.43	4.17	2.75	**0.58**
RI (*nPVI*)	51	67	77	53	69	31

Table 1 shows the results of measuring annotation durations (using Praat, [5]) in examples from the two speech styles of newsreading and poetry reading, not as proof of a phonetics-phonology relation, but as illustrations of one

method for determining these relations. The measurements show that newsreading syllable and word rates are faster but the IPU rate is slower. The auditory impression of the speech styles is that the poet reads slowly to convey melancholy at leaving home, however dilapidated it may be. The faster IPU rate in poetry reading is easily explained: metrical line structure constrains shorter IPU durations and thus faster rates than in newsreading, while IPU rates in newsreading correspond to default phrasal and sentential structure, which is, unlike poetry reading, not overridden by metrical structure.

The regularity indices (RI) of the two examples, here represented by the normalised pairwise variability index (*nPVI*), show that in both reading styles syllable indices (51 and 53) and word indices (67 and 69) are close in each case. The word indexes show higher irregularity, which relates to English morphosyntactic structure (short grammatical words, long variable-duration lexical words). The syllable index shows marginally greater regularity and does not take morphological word structure into account. The IP index shows greater regularity in poetry reading, which is relatable to poetic metre constraints which are absent in newsreading.

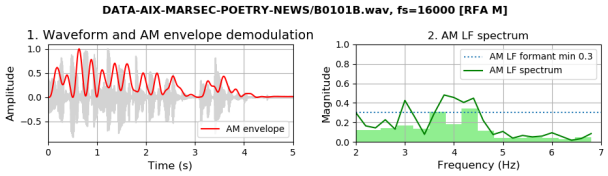

Figure 7: First 5 s of newsreading.

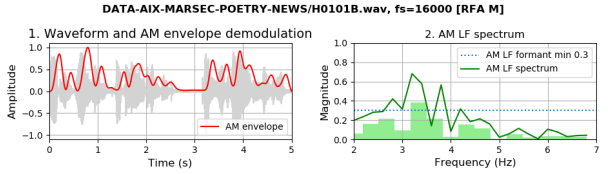

Figure 8: First 5s of poetry reading.

The assumption is that the LF formants in the spectra of these examples relate to the syllable, word and IP rates. Inspection shows that this is very clearly the case for the IPU indices: the IPU rates indicate frequencies of 0.43 Hz and 0.58 Hz for newsreading and poetry reading, respectively. There are also slower frequencies in each style, corresponding to longer episodes than the IPU durations such as verses in poetry or paragraphs in newsreading. The AM spectrum of the poetry reading shows clear LF formants at 1.1Hz, 1.3Hz 1.6Hz, with frequencies which relate partly to half-lines, but also partly to IPU-internal phrase-final lengthening. At the predicted word frequencies of 3.91Hz and 2.75Hz, and syllable frequencies of 6.17Hz and 4.17Hz the long-term spectrum is too diffuse to identify higher frequency formants.

In shorter contexts below 10s these spectral zones are identifiable (selected in Figure 7 and Figure 8). Figures 7 and 8 visualise properties of LF AM spectra in the expected range of word and syllable rhythms. These patterns cannot be identical to the spectral patterns in Figures 1 and 2 because the latter compress long-term ranges (56.18s and 58.34s, respectively) into a single dimension., in contrast to the 5s ranges in Figures 7 and 8. Nevertheless, the spectral range of 2Hz to 7Hz shows the essential properties of the two speech styles in this spectral region: the tendency to faster word rhythms in the newsreading, with a zone of spectral peaks around 4Hz, and slower word rhythms in the poetry reading, with a zone of spectral peaks around 3.5Hz. Long-term syllable timing is too diverse to show clear patterning.

The spectra in Figures 1, 2, 3, 7 and 8 also illustrate the variability of rhythm: spectral peaks occur in groups relating to ranks in a *prosodic default hierarchy*: syllables, words and IP frequencies, (cf. the annotation discussion). An explanation for this variability is overriding by higher ranks of expected lower rank default durations: lexically stressed syllables are longer than lexically unstressed syllables *unless* the unstressed syllable is phrase-final (phrase-final lengthening), other phonotactic conditions being approximately equal (Figure 9).

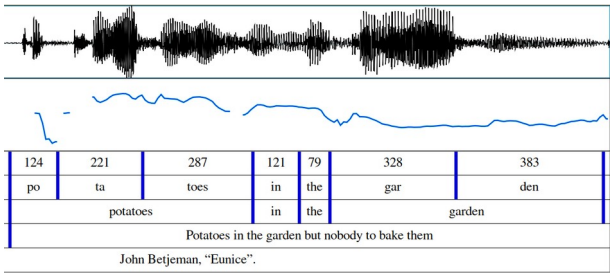

Figure 9: Annotated segment of H101B (times in ms).

Figure 9 (Praat [5] screenshot, data item H101B) illustrates this overriding effect in the prosodic default hierarchy: in the phrasal pattern ((*poTAtoes*)(*in the GARden*)), the lexically unstressed syllables *toes* and *den* would be expected on default lexical grounds to be shorter than the preceding lexically stressed syllables *ta* and *gar*, respectively. However, they are phrase-final, and both post-lexically and acoustically longer than the preceding syllables, having been overridden by phrase-final lengthening.

5. Summary, conclusion and outlook

It was shown that a novel pilot application of basic machine learning techniques (*k-means* and hierarchical clustering) to low frequency (<5 Hz) spectrograms of F0 and of the amplitude envelope over long passages (typically 60s) from the Aix-MARSEC broadcasting database can distinguish plausibly between newsreading and poetry reading styles. Only the reading-aloud register with different speakers and genders was involved, indicating that style timing features were detected independently of speaker or gender. Phonetic explanations of the ML clusterings were argued in terms of differences in variance, for example relating to greater liveliness in the poetry readings.

The sparseness of the available data means that the results have pilot study status, as with the previous studies with related methods which were mentioned in the introduction. It remains to be seen whether long-term rhythmic differences in languages as well as in culture-specific speaking styles [15] can also be detected with this method. Applications to emotion detection, to naturalness assessment in speech synthesis and to the diagnostics and testing of the public speaking proficiency of advanced L2 learners are anticipated.[1]

[1] Many thanks to the anonymous reviewers for constructive suggestions, and to Dr. Xuewei Lin, Jinan University, Guangzhou, for insights on language-dependent typology of poetic rhythm and metre.

6. References

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

[2] Asu, Eva-Liina and Francis Nolan. "Estonian and English rhythm: a twodimensional quantification based on syllables and feet," *Speech Prosody* 3, 2006.

[3] Auran, Cyril, Caroline Bouzon, Daniel Hirst. "The Aix-MARSEC project: an evolutive database of spoken British English," *Speech Prosody* 2. 2004.

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

[5] Boersma, P. "Praat, a system for doing phonetics by computer," *Glot International* 5:9/10, 341-345, 2001.

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

[7] Cummins, Fred, Robert Port. "Rhythmic constraints on stress timing in English," *Journal of Phonetics* 26, pp. 145–171, 1998.

[8] Dauer, Rebecca M. "Stress-timing and syllable-timing reanalyzed," *Journal of Phonetics* 11, 51–62, 1983.

[9] Foote, Jonathan and Shingo Uchihashi. "The beat spectrum: a new approach to rhythm analysis," *IEEE International Conference on Multimedia and Expo*, 2001.

[10] Galves, Antonio, Jesus Garcia, Denise Duarte and Charlotte Galves. "Sonority as a basis for rhythmic class discrimination," In Bernard Bel and Isabel Marlien, eds., *Speech Prosody 1*, Aix-en-Provence: Laboratoire Parole et Langage, 323–326, 2002.

[11] Gibbon, Dafydd. "Computational modelling of rhythm as alternation, iteration and hierarchy," *15th International Congress of Phonetic Sciences* (*ICPhS* XV), Barcelona, 2489–2492, 2003.

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

[13] Gibbon, Dafydd. "The Rhythms of Rhythm," Journal of the International Phonetic Association. First View, 1-33, 2021. DOI: 10.1017/S0025100321000086

[14] Gibbon, Dafydd. Rhythm formants of story reading in standard Mandarin. *Chinese Journal of Phonetics* 14, 2020.

[15] Gibbon, Dafydd. Ega orature: discourse prosody meets machine learning. (To appear.)

[16] Gut, Ulrike. "Rhythm in L2 Speech," In Gibbon, Dafydd, Daniel Hirst and Nick Campbell, eds., *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem. Special Edition of Speech and Language Technology* 14/15, 83–94. Poznań: Polish Phonetics Society, 2012.

[17] Hermansky, Hynek. "History of modulation spectrum in ASR," ICASSP 1988.
DOI: 10.1109/ICASSP.2010.5494907

[18] Inbar, Maya, Eitan Grossman and Ayelet N. Landau. "Sequences of Intonation Units form a ~ 1 Hz rhythm." Scientific Reports 10, 15846 (2020).
DOI: doi.org/10.1038/s41598-020-72739-4

[19] Jassem, Wiktor and Dafydd Gibbon. "Re-defining English stress," *Journal of the International Phonetic Association* 10, 1980, 2–16,. 1980.

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

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

[22] Lee, Christopher S. and Neil P. McAngus Todd. "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, 2004.

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

[24] Lewalter, Thorsten und Berndt Lüderitz, eds. *Herzrhythmusstörungen: Diagnostik und Therapie*. Berlin: Springer, 2010.

[25] Malisz, Zofia, Michael O'Dell, Tommi Nieminen and Petra Wagner. "Perspectives on speech timing: coupled oscillator modeling of Polish and Finnish," *Phonetica* 73(3–4), 229–255, 2016.

[26] Michon, Pascal. *Elements of Rhythmology*, Vol. I-V. Paris: Rhuthmos, 2018-2021.

[27] O'Dell, Michael L. and Tommi Nieminen. "Coupled Oscillator Model of Speech Rhythm," *14th International Congress of Phonetic Sciences* (*ICPhS* XIV), 1075–1078, 1999.

[28] Suni, Antti, Juraj Šimko, Daniel Aalto and Martti Vainio. "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech and Language* 45, 123–136, 2017.

[29] Tilsen Samuel and Johnson, Keith. 2008. "Low-frequency Fourier analysis of speech rhythm," *J.Ac.Soc.Am.* 124 (2):EL34–EL39. [PubMed: 18681499]

[30] Tilsen, Samuel and Amalia Arvaniti. "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, 2013.

[31] Todd, N. P. M., Brown, G. J. A computational model of prosody perception. ICSLP 94, 127-130, 1994.

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

[33] White, Laurence and Zofia Malisz. "Speech Rhythm and Timing," In Gussenhoven, Carlos and Aoju Chen (eds.), *The Oxford Handbook of Language Prosody*. Oxford: Oxford University Press, 2020.