Part 1: METHODOLOGY

Część 1: METODOLOGIA

Empirical models of tone, rhythm and intonation for the analysis of speech prosody

Empiryczne modele tonu, rytmu i intonacji do analizy prozodii mowy

Daniel Hirst

CNRS Laboratoire Parole et Langage, Aix-en-Provence, France
School of Foreign Languages, Tongji University, Shanghai, China
daniel.hirst@lpl-aix.fr

ABSTRACT

This contribution is a reflection on criteria for an empirical framework for the investigation of speech prosody across different languages and, as far as possible, across different theoretical points of view. It is proposed to define different units, called the *Tonal Unit* [τ] and the *Rhythm Unit* [ρ], as the domains of short-term planning of pitch and timing, respectively, as well as a larger *Intonation Unit* [IU] for longer-term prosodic characteristics like changes of *register* and *tempo*. This makes it possible to compare models and their implementation. A comparison of Jassem's model of rhythm with that of Halliday on a 5.5 hour database of spoken English (Aix-Marsec) largely confirmed several of Jassem's predictions. The same database is now being used to investigate the short term planning of pitch within the Tonal Unit.

STRESZCZENIE

Niniejsza praca opisuje kryteria dotyczące ram empirycznych pozwalających na badanie prozodii mowy w różnych językach i, na ile to możliwe, z różnych teoretycznych punktów widzenia. Proponuje się zdefiniowanie różnych jednostek – Jednostki Tonalnej [τ] oraz Jednostki Rytmu [ρ], odpowiednio jako dziedzin krótkookresowego planowania tonu i rytmu, jak i większej Jednostki Intonacyjnej [JI] dla cech prozodycznych charakteryzujących się dłuższym czasem trwania, takich jak zmiany rejestru i tempa. Dzięki temu możliwe staje się porównywanie modeli i ich implementacji. Porównanie modelu rytmu Jassema z modelem Hallidaya na podstawie pięcioipółgodzinnej bazy mowy angielskiej (Aix-Marsec) w dużym stopniu potwierdza kilka hipotez Jassema. Ta sama baza jest obecnie stosowana do badania krótkookresowego planowania tonu w Jednostce Tonalnej.

1. Introduction

In one of the first attempts to describe the prosody of English, or indeed any language, Steele [1] opens his Chapter 2 with the statement:

> The art of music, whether applied to speaking, singing or dancing, is divided into two great branches, sound and measure, more familiarly called tune and time. Instead of which words, I use (for the most part) the Greek terms of melody and rhythmus, being more significant, as generals, than our vulgar terms. (p. 18)

Over two hundred years later, *Tune* (or *Melody*) and *Time* (or *Rhythm*) are still the essential components of prosodic analysis today, although, following Jassem [2, 3], an adequate empirical model will need a third component, *Intonation*, in order to describe longer term characteristics of both Tune and Time, the shorter term characteristics of which I refer to as *Tone* and *Rhythm*.

The terminology I use is, consequently, essentially that used in Wiktor Jassem's pioneering analysis of English intonation. The only difference at this level is that, in order to make systematic use of the term *Unit*, I use the term *Intonation Unit* for what Jassem, following traditional usage, refers to in [2] as the *Tune* or *Tone Group* and, in [3], as the *Intonation Phrase*. While Jassem's aim was to provide a specific empirical description of the intonation of English, my concern here is to reflect on criteria for an empirical framework for the investigation of speech prosody across languages and, as far as possible, across theoretical points of view.

2. The Structure of Utterances

Language has structure and the way in which we describe this structure depends largely on the variety of linguistic theory we adhere to. A minimal linguistic theory might say that a sentence is made up of words, and that these words are grouped into phrases. Phrases may be made up directly of words, or they may contain smaller phrases. The category *phrase* in other words is a *recursive* category.

We might further say that the word is not an atomic unit but that it is itself made up of one or more morphemes. So we have something like the following hierarchy of syntactic categories:

Categories Sentence > Phrase* > Word > Morpheme

where * indicates a recursive category.

Applied to a sentence like:

(1) They expected her election in September.

we could represent the constituent structure as a familiar syntactic tree as shown in Figure 1.

While different linguists may disagree about the exact form of this tree, and even more about the names to give to the nodes of the tree, something like this tree seems a reasonable approximation to an informal representation of the structure of this sentence.

There is today, though, largely a consensus view that syntactic structure is not the whole story, but that the *sounds* of an utterance are structured in a way which is to some extent independent of syntax.

Again, there is a fair consensus that words can be defined as a sequence of *phonemes* and that these phonemes are grouped into syllables. The difficulty begins, though, when we try to relate these phonological categories, phoneme and syllable, to the syntactic categories.

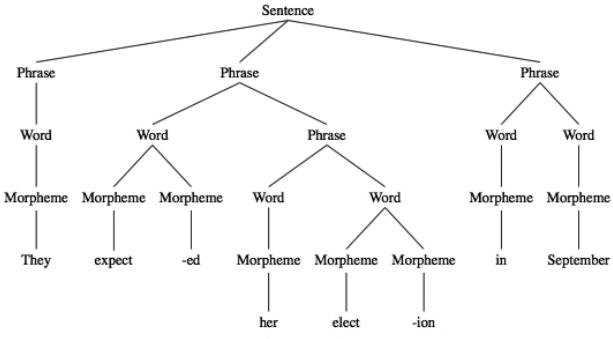

Figure 1: Syntactic tree for example (1).

If we take, for example, the French sentence:

(2) Il est en or. (*It's made of gold*)

we can transcribe this as

(3) /i.lɛ.tɑ̃.nɔʁ/

Like (2), which contains four words, (3) contains four syllables, the boundaries of which are indicated in the transcription by dots. Not one of the four syllables, however, corresponds exactly to one of the four words in the orthographic representation. The reason for this is that French makes regular use of liaison and enchaînement (linking) so that there is a strong tendency to 'resyllabify' words whenever possible, to favour open syllable structure.

If we look above the level of the syllable, the situation is no better. Many descriptions of (British) English intonation and rhythm make use, following [4] and [5], of a unit called the *foot*, a concept originally proposed by Steele [1] under the name of *cadence* or *bar*, and which is obviously derived from musical and/or poetical notation. Since the term *foot* has also been used as a theoretical construct in metrical and autosegmental phonology with a different interpretation, I shall use the term *stress foot* here for the unit proposed for the description of intonation. This unit can be defined for speech as a sequence of syllables beginning with an accented syllable or with a silent beat at the beginning of a sentence, and continuing up to (but not including) the next accented syllable or silence. With this definition, we can represent example (1) as:

(4) ‖ˆ they ex- ‖ pected her e- ‖ lection in Sep- ‖ tember. ‖

where the symbol '‖' represents the foot boundary and 'ˆ' a silent beat. As can be seen in this example, there can be a considerable mismatch between the level of syntactic words and that of stress feet.

Higher level phonological categories present similar or even worse problems. As we saw above, it is generally considered that syntactic phrases are recursive syntactic

categories whereas phonological phrases do not show the same type of recursive structure.

Since the relationship between phonological structure and syntactic structure seems far from direct, the solution adopted in most descriptions of intonation (particularly but not exclusively that of British English) has been to assume that the two types of structure are independent levels of representation and to suppose that there must exist some sort of *mapping rules* to explain how one type of structure is related to the other, even though the nature of these mapping rules is often far from explicit. For one explicit formulation of such a mapping rule cf. [6], also summarized in [7].

By contrast with the categories of syntactic structure, then, in models of this type the structure is flatter with the following categories:

Categories Intonational Phrase > Stress foot > Syllable > Phoneme

and assuming that, unlike for syntax, there are no recursive categories. See, however, [8] (section 6.3) for a more extensive discussion of arguments for recursive phonological categories in the form of compound prosodic domains.

Representing the Intonational Phrase by IP, the stress foot by Σ and the syllable by σ, we can give the phonological tree in Figure 2 for our example (1).

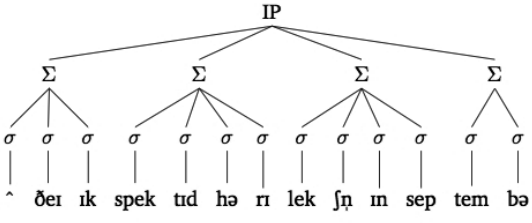

Figure 2: Phonological tree for example (1).

Since Halliday [5], most of the systematic descriptions of British English intonation (e.g. [9, 10, 11], up to and including [12]) have used, or implied, a framework similar to that which we saw in the preceding section. The same phonological unit, the stress foot, is used in these studies to describe the (short-term) domains of both melody and rhythm.

Just 60 years ago, however, Jassem [2] suggested that we need *distinct* units to represent tonal and rhythmic structure. As I mentioned above, Jassem describes the longer term units of intonation with a unit which he calls the *tune* or *tone group*. For the shorter units, he follows the practice of studies of tone languages and cites with approval [13], who says:

> In Chinese, for example, the syllable is universally accepted as the tone-unit, for the reason that practically every syllable of the language can mean different things according to the way it is intoned... In Panjabi and Lahuda the tone-units are practically all disyllabic. In English and practically all other European languages the tone-unit is neither the syllable, nor even the word, but a phrase consisting of one or more words. (p 124).

Jassem adopts the term *Tonal Unit* rather than *Tone Unit*, presumably to avoid confusion between what he calls the *Tone Segment* in English and *lexical tones* in tone languages. It also helps to avoid confusion between this unit of prosodic structure and the longer term unit which he refers to as the *Tone Group*.

In fact, Jassem's definition of the *Tonal Unit*, given by a list of five different types of such units [2:pp49–50] is precisely equivalent to the definition given above (section 2) of what Abercrombie, twelve years later, was to call the (*stress*) foot.

Unlike Abercrombie and Halliday, who use the same unit to describe both melody and rhythm, Jassem makes a clear distinction between the *Tonal Unit*, which is conceived of as the domain of occurrence of local pitch movements in English, and the *Narrow Rhythm Unit* and the *Anacrusis*, conceived of as the domain of segmental timing.

After noting that he does not consider the word "to be either a phonetic or a phonological unit" [2:p38], Jassem defines the *Narrow Rhythm Unit* (*NRU*) as a sequence of syllables, the first of which is rhythmically strong and the last of which is followed by a rhythmical juncture. He distinguishes this from other sequences of syllables which are "pronounced extremely rapidly" [2:p40] and which he calls the *Anacrusis*.

The difference between the two is illustrated by a minimal pair, taken from [14]:

(5) a. Summer dresses b. Some addresses

In this example Jassem notes that the although the phonemes and stresses are identical there is a subtle difference of rhythm in the two, the first syllable of (5a) being shorter than that of (5b). whereas the second syllable is longer in (5a) than in (5b). He attributes this difference to the fact that first two syllables of (5a) constitute a single *NRU*, whereas in (5b) the first syllable constitutes a *NRU* on its own and the second syllable constitutes an *Anacrusis*. He proposed to represent this in the phonetic transcription by the simple device of a space after each *NRU* as in:

(6) a. /ˈsʌmə ˈdresiz/ b. /ˈsʌm əˈdresiz/

Here the spaces neatly correspond to the spaces in the orthographic transcription but this is not always the case. Another example, from [15]:

(7) a. Take Greater London b. Take Grey to London

could be transcribed, using Jassem's proposal, as follows:

(8) a. /ˈteɪk ˈgreɪtə ˈlʌndn/ b. /ˈteɪk ˈgreɪ təˈlʌndn/

where the spaces are no longer identical to the orthographic version.

Jassem's notation can even make a distinction which is not made in normal orthography. The sentence:

(9) He bought her chocolates.

can be interpreted in two ways, depending on whether 'her' is taken to be an indirect object (i.e. = He bought chocolates for her) or whether it is a possessive determiner (i.e. = He bought the chocolates she was selling). Jassem's model predicts that in the first of these 'her' will be in the same *Narrow Rhythm Unit* as the verb 'bought' but that in the second it will be part of the *Anacrusis*, so that the two interpretations would be transcribed respectively:

(10) a. /hi'bɔ:thə 'tʃɒkləts/ b. /hi'bɔ:t hə'tʃɒkləts/

The definition of the *Narrow Rhythm Unit*, then, is a unit beginning with an *accented syllable* and ending before a *rhythmic juncture*. The *rhythmic juncture* corresponds in the majority of cases to the following word boundary except in the case of enclitics which are assimilated to the preceding *NRU*.

Jassem's model for rhythm also has a higher level unit called the *Total Rhythm Unit*, which consists of an optional *Anacrusis* followed by a *Narrow Rhythm Unit*. In (6), (8) and (10), for example, the *Total Rhythm Units* are the sequences which are separated by the spaces.

In the next section, I build on Jassem's idea of specific units for *Rhythm*, *Tone* and *Intonation* with a view to developing a relatively theory-independent and language-independent empirical framework for the study of prosodic structure.

3. An empirical framework for the study of prosodic structure

Klatt [16], in his review of twenty years of research on speech synthesis, came to the pessimistic conclusion that:

> One of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena. (p. 760)

What is true for speech timing is also true for the study of tonal phenomena. As we saw above, Halliday used the same unit, the stress foot, to describe both pitch and rhythm whereas in Jassem's model these are dealt with by assuming different units for rhythm and for tone.

An empirical framework for the study of prosodic typology will obviously need a way to test which size unit is the most effective in modeling our data. In ongoing work [17] on a prosody editor, designed specifically for linguists to test models of prosody, I consequently take a deliberately agnostic view on what corresponds to the rhythm unit and what corresponds to the tonal unit. Instead I *define* the *Rhythm Unit* [ρ] and the *Tonal Unit* [τ] as the domains of interpretation of short term planning of timing and pitch respectively and at the same time the *Intonation Unit* [IU] is *defined* as the domain of longer term interpretation of pitch and tone via changes in register and tempo.

This means that we can then formulate the differences between two hypotheses more precisely by saying, for example, that for Jassem, ρ corresponds to his *Narrow Rhythm Unit* and his *Anacrusis*, whereas for Halliday it corresponds to the *Stress Foot*. In my representation of Jassem's model of rhythm (Figure 3) I do not implement his *Total Rhythm Unit* as part of the rhythmic structure since as far as I am aware this unit

has no specific properties other than those which can be attributed to those of the *NRU* and the *Anacrusis*.

In the representation of the tonal domains, a further difference in representation is that instead of implementing the *Anacrusis* as a single unit, I treat each sequence of phonemes delimited by either a stress or a word boundary as a separate *Rhythm Unit*. This is purely a formal difference with no loss of information since Jassem's *Anacrusis* can now be defined as a sequence of *Rhythm Units* not containing a stress, and the *Total Rhythm Unit*, as we saw above, correponds to an optional *Anacrusis* followed by a *Rhythm Unit* containing a stressed syllable. For Jassem, τ corresponds to Jassem's *Tonal Unit* which, as we saw, is identical to Abercrombie's and Halliday's *Stress Foot*.

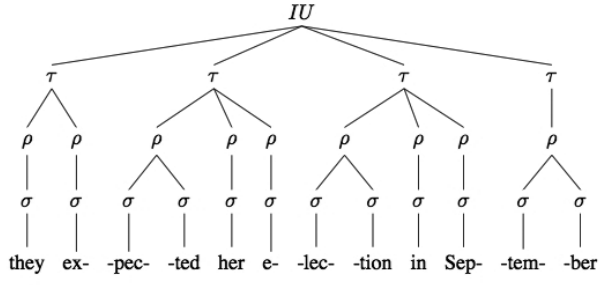

Figure 3: Representation of (1) in my formulation of Jassem's model.

Using the annotation I propose, then, we can formulate a representation of example (1) in Jassem's model as in Figure 3 whereas in Halliday's model the representation would be as in Figure 4.

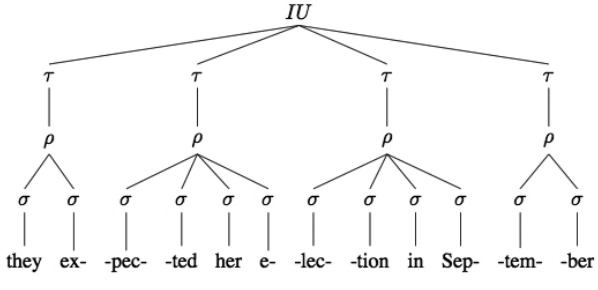

Figure 4: Representation of (1) in my formulation of Halliday's model.

Note that with this formulation of the different models, it is possible to represent both the rhythmic and the tonal structure of an utterance in the same tree, and using the same formal apparatus to describe two different models makes it possible to test the predictions of the models empirically. In the editor described in [17] it is possible to implement representations as in Figure 3 or Figure 4 and to synthesise the output in order to compare the different representations either informally or to create stimuli to be used in more formal perceptual tests.

4. Looking for the units of phonological representation

The formal framework described above is designed to be as language-independent and as theory-independent as possible. In order to apply the framework to a specific language it is not enough, obviously, to simply test a few examples, although this may be a useful heuristic to suggest relevant parameters for future analysis. In this final section I report some results and some ongoing work on the application of the framework to British English using a large database of spoken English.

4.1. Aix-Marsec: a database for British English prosody

The Spoken English Corpus [18] consists of about five and a half hours of recordings of British English speech in a number of different speaking styles from 68 speakers, recorded in the 1980's. The corpus was later converted to machine-readable form (hence the name *Marsec*) and subsequently [19] transcribed phonetically and the transcriptions aligned with the speech signal in the form of Praat TextGrids [20].

Since then, the corpus, now the Aix-Marsec database, has been the object of a number of enrichments including re-constituting the original recordings from the earlier segmented version[1], and adding annotation tiers for *Phonemes*, *Syllables*, *Rhythm Units* (coded as *ANA* or *NRU*), *Tonal Units* (coded as *Anacrusis* or *Foot*), *Words*, *Intonation Units* and *Speakers* [21] as well as the original manual prosodic annotation in the form of *Tonal Stress Marks* (*TSM*). Figure 5 shows a TextGrid for a fragment from file A01 of the corpus.

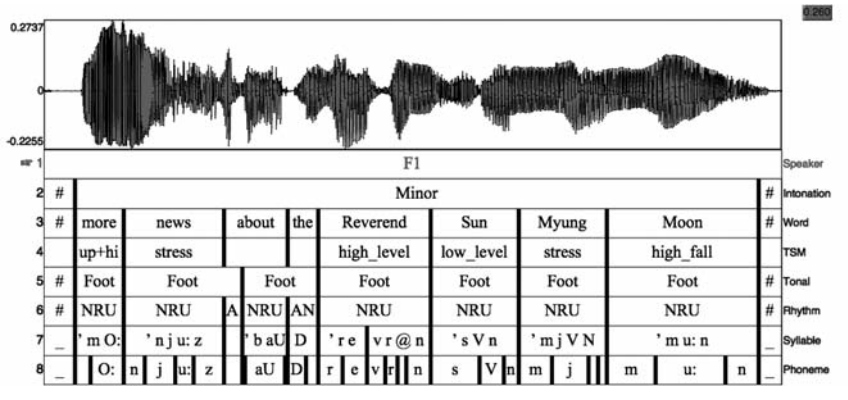

Figure 5: A TextGrid for an extract from the file A01 from the Aix-Marsec database corresponding to the utterance: "More news about the Reverend Sun Myung Moon."

4.2. Data on rhythm from the Aix-Marsec database

Hirst & Bouzon [22] found that, as predicted by Jassem but contrary to Halliday's model, word boundaries *do* play an important role in the rhythmic structure of English. Strong negative correlations were found between the duration of a segment and the

[1] In order to economize computer memory, the original recordings had been cut up into segments of about two minutes each with an overlap of a few seconds.

number of phonemes in the stress-foot, in the *NRU* and in the word, but no similar effect was found either in the syllable or in the *Anacrusis*. Moreover the correlation was greater for the *NRU* than either the stress-foot or the word, thus confirming Jassem's predictions.

The most surprising result was the further confirmation of Jassem's prediction that, once we know whether a given phone belongs to an *Anacrusis* or to a *NRU*, and in the latter case once we know the number of phones in that unit, the fact that the phone occurs in a stressed or an unstressed syllable has no specific effect on its duration.

One of the major factors influencing phoneme duration is the identity of the phoneme itself. To neutralise this, following Campbell [23], the z-score of the phoneme duration was used instead of the raw duration. A second well known effect is that of final lengthening, which in [22] was shown to effect in particular the final 3 phonemes of an intonation unit. In order to neutralise final lengthening, the last three phonemes of each intonation unit were excluded from the analysis.

An analysis of variance [24] revealed a highly significant effect of both number of phonemes in the *NRU* and position of the phoneme within the *NRU*. When phonemes were coded as *NRU Initial*, *Medial* and *Final*, analysis of variance once again revealed highly significant differences ($p<2.2e-16$) between the three positions, with mean values of z-score as shown in Figure 7.

For *NRU* initial phonemes, the size of the *NRU* was non-significant ($F(1, 23309) = 2.5$ $p = 0.1522$). For *NRU* medial phonemes, also, analysis of variance on index of phoneme within the *NRU* was also non-significant.

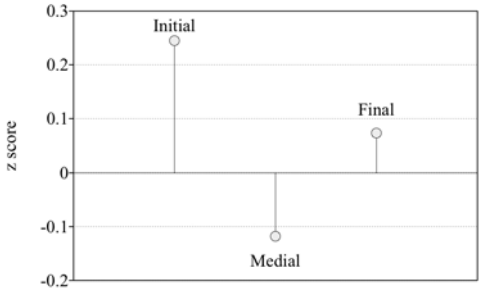

Figure 6: Mean values of z-score for initial, medial and final phonemes within the *NRU*.

The model suggested by this data is simply a lengthening of the initial and final phoneme of each *NRU*. Since each *NRU* contains one initial and one final phoneme this could explain why, as noted in earlier work [25], the lengthening appears to be uniform across the *NRU* regardless of the number of phonemes it contains. It remains to be seen the way in which this *NRU* initial and final lengthening interacts with the lengthening of the *NRU* as a whole, as well as with the final lengthening observed in *Intonation Unit* final position.

4.3. From data to models

In order to test this type of interaction, the segmental information has now been converted into tabular form with one line for each phoneme and columns corresponding to

recording, speaker, phoneme, word and duration. For each phoneme, its position in the utterance is encoded by thirty binary variables indicating for each unit whether it is initial or final in each higher-level unit, for the hierarchy: *Phoneme* (*p*) < *Rhythm Unit* (*r*) < *Tonal Unit* (*t*) < *Word* (*w*) < *Intonation Unit* (*i*). For example *pis* = 1 when the phoneme is initial in the syllable and 0 otherwise, while *sfw* = 1 indicates that the syllable is final in the word.

This data is now being used to build and test predictive models for rhythm. At the same time, the raw fundamental frequency has been modelled using the Momel algorithm [26]. This modelling, using the technique described in [17] will be used to test whether, as suggested by both Jassem [2] and Halliday [5] the *Tonal Unit/Stress foot*, is indeed the most appropriate unit to model the timing of the pitch targets.

References

[1] Steele, J. 1779. *Prosodia Rationalis: or an Essay towards Establishing the Melody and Measure of Speech, to be Expressed and Perpetuated by Peculiar Symbols*. London: J. Nichols, 2nd. Edition.

[2] Jassem, W. 1952. *Intonation of Conversational English: (educated Southern British)*. Nakładem Wrocławskiego Towarzystwa Naukowego; skład główny: Dom Książki. [PDF available from the Speech and Language Data Repository: http://sldr.org/sldr000777/en].

[3] Jassem, W. 1999. English stress, accent and intonation revisited. *Speech and Language Technology*, 3, 33–50.

[4] Abercrombie, D. 1964. Syllable quantity and enclitics in English. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, J. L. M. Trim, eds., *In Honour of Daniel Jones*. London: Longmans. 216–222.

[5] Halliday, M.A.K. 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

[6] Hirst, D.J. 1993. Detaching intonational phrases from syntactic structure. *Linguistic Inquiry*, 24 (4), 781–788.

[7] Hirst, D.J. 1998. Intonation in British English. In D.J. Hirst & A. Di Cristo, eds. *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press. Chapter 3, 56–77.

[8] Ladd, D. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[9] Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

[10] Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

[11] Tench, P. 1996. *The Intonation Systems of English*. Cassell.

[12] Wells, J. C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.

[13] Beach, D. 1938. *The Phonetics of the Hottentot Language*. Cambridge, MA: Heffer and sons.

[14] Jassem, W. 1949. indikeiʃn əv spiːtʃ riðm in ðə traːnskripʃn əv edjukeitid sʌðən inglиʃ (Indication of speech rhythm in the transcription of educated Southern English). *Le Maître Phonétique*, III (92), 22–24. [Republished in *Journal of the International Phonetic Association* with the original IPA version and a new orthographic version by D. Hirst].

[15] Scott, N. 1940. Distinctive rhythm. *Le Maître Phonétique* 49, 6–7.

[16] Klatt, D. 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, 82:737–793.

[17] Hirst, D.J. 2012. ProZed: A speech prosody analysis-by-synthesis tool for linguists. In *Proceedings of the 6th International Conference on Speech Prosody*. Shanghai. [Praat plugin available from the Speech and Language Data Repository: http://sldr.org/sldr000778/en].

[18] Knowles G, Williams B, Taylor R (editors). 1996. *A Corpus of Formal British English Speech*. London & New York: Longman.

[19] Auran, C.; Bouzon, C & Hirst, D.J. 2004. The Aix-MARSEC Project: An Evolutive Database of Spoken British English. In *Proceedings of the Second International Conference on Speech Prosody*. Nara, Japan.

[20] Boersma, P & Weenink, D. 1992–2012. Praat, a system for doing phonetics by computer. [version 5.3.10, March 2012]. [available from http://www.praat.org].

[21] Hirst, D.J.; De Looze, C; Auran, C & Bouzon, C. *forthcoming*. Aix-Marsec: a database for the analysis of the prosody of British English. Forthcoming. [Database available from the *Speech and Language Data Repository*: http://sldr.org/sldr000033/en].

[22] Hirst, D.J.; Bouzon, C. 2005. The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). *Proceedings of Interspeech/Eurospeech* 05., Lisbon. 29–32.

[23] Campbell, N. 1992. *Multi-level Timing in Speech*. Ph.D. thesis, University of Sussex.

[24] Hirst, D.J. 2009. The rhythm of texts and the rhythm of utterances: from metrics to models. *Proceedings of Interspeech* 2009, Brighton. 1519–1523.

[25] Eriksson, A. 1991. *Aspects of Swedish Speech Rhythm*. Gothenburg Monographs in Linguistics, 9. Gothenburg University: Department of Linguistics.

[26] Hirst, D.J. 2007. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken. Germany. 1233–1236.