DOI: 10.2478/linpo-2014-0004

Duration and speed of speech events: A selection of methods

Dafydd Gibbon[1], **Katarzyna Klessa**[2] **& Jolanta Bachan**[2]

[1] Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, gibbon@uni-bielefeld.de

[2] Institute of Linguistics Adam Mickiewicz University in Poznań, klessa@amu.edu.pl, jbachan@amu.edu.pl

Abstract: Dafydd Gibbon, Katarzyna Klessa & Jolanta Bachan. *Duration and speed of speech events: A selection of methods*. The Poznań Society for the Advancement of the Arts and Sciences. PL ISSN 0079-4740, ISBN 978-83-7654-384-0, pp. 59–83

The study of speech timing, i.e. the duration and speed or tempo of speech events, has increased in importance over the past twenty years, in particular in connection with increased demands for accuracy, intelligibility and naturalness in speech technology, with applications in language teaching and testing, and with the study of speech timing patterns in language typology. However, the methods used in such studies are very diverse, and so far there is no accessible overview of these methods. Since the field is too broad for us to provide an exhaustive account, we have made two choices: first, to provide a framework of paradigmatic (classificatory), syntagmatic (compositional) and functional (discourse-oriented) dimensions for duration analysis; and second, to provide worked examples of a selection of methods associated primarily with these three dimensions. Some of the methods which are covered are established state-of-the-art approaches (e.g. the paradigmatic *Classification and Regression Trees*, CART, analysis), others are discussed in a critical light (e.g. so-called ‘rhythm metrics’). A set of syntagmatic approaches applies to the tokenisation and tree parsing of duration hierarchies, based on speech annotations, and a functional approach describes duration distributions with sociolinguistic variables. Several of the methods are supported by a new web-based software tool for analysing annotated speech data, the *Time Group Analyser*.

Keywords: speech timing, Polish, English, speech technology

1. Objectives and topic overview

The present contribution concentrates on a selection of methods for analysing speech timing in English and Polish. The unifying principle is not so much extensive data analysis or historical review, but rather methodological, looking at speech timing from three points of view: paradigmatic or classificatory, syntagmatic or structure-building, and functional in discourse contexts.

Unauthenticated
Download Date | 12/13/15 8:16 AM

Preferred methods have varied considerably over time, partly in dependence on available statistical, formal and technological techniques. For example, in 1950s linguistic phonetics, Jassem & Abercrombie analysed structural relations between phonemes, syllables, feet and rhythm. By contrast, in 1960s quantitative phonetics, Lehiste, Jassem and others concentrated on isochrony (equal unit timing) in relation to words, syllables and phonemes, while in the 1970s psycholinguistics introduced perceptual experiments with timing and sentence complexity. In the 1980s and 1990s, work in speech technology by Campbell and in computational phonology by Bird led to statistical and logical models of speech timing. Subsequently, continuing to the present day, rhythm modelling in comparative phonetics and formal oscillator models, and the analysis of large corpora with CART (*Classification and Regression Trees*) methods, as well as quantitative applications to L2 learning, have emerged. What has emerged, at this global level of discussion, is the very large number of degrees of freedom manifested in speech timing, including properties of different phone types, phone positions in syllables, syllable positions in words, and word positions in relation to boundary types in parallel syntactic and intonational phrase structures (cf. Campbell 1992), as well as pause distribution and functionality (cf. Dechert & Raupach 1980). In this brief review, only a few immediately relevant trends are selected.

A major influence in the investigation of speech timing models has been the need for predictive models of segment duration in speech technology, particularly speech synthesis. The earliest models were rule-based, and used a combination of linguistic and phonetic analysis to create sets of segment duration rules for English. In an early model (Klatt 1976) each segment is attributed an inherent duration, and is shortened or lengthened by a context-dependent percentage value, subject to a specified minimum duration. The contexts included pre-pausal final lengthening, non-final shortening, non-initial vowel shortening, non-stressed sound shortening, and vowel lengthening before voiced consonants. The model was successfully applied in speech systems such as *Klattalk* and *DECtalk* (Klatt 1987). Rule-based duration models were also created for many other languages, e.g. for French (O'Shaughnessy 1984), German (Portele et al. 1990) and Hungarian (Olaszy 2002). In the development of rule-based models, linguistic knowledge, experience and intuition dominate over extensive quantitative analysis of actual corpora, and both the rules and their parameters are defined with a sequential, (semi-)manual trial and error approach. Corpus-based models focus more on variation and constancy in large collections of data, though they necessarily also involve linguistic information.

Studies of English timing have been well documented (cf. contributions to Gibbon et al. 2012). For Polish, initial significant results on speech timing were achieved several decades ago (e.g. Richter 1973; 1974; 1987; Jassem et al. 1981), with investigation of relations between logatoms and linguistic features, the influence on segment duration of position in accent units, and distinctions between duration classes, many of the studies focusing on isochrony and its limits. The methods used included rhythm structure modelling with logatoms and linguistic features, a power function relating segmental duration and the number of syllables within an accent unit, and regression models for isochrony in the NRU (Narrow Rhythm Unit) vs. number of syllables in rhythm units (cf. Jassem et al. 1981), finding that the greatest tendency to isochrony was present in the Narrow Rhythm Unit.

Unauthenticated
Download Date | 12/13/15 8:16 AM

Variable speech rate, and its effect on phone durations, vowel formant patterns and syllable structure, poses another challenge for speech timing models (Łobacz 1976a, b; Zee 2002; Cummins 1999; Crystal 1969). Findings include dependency on type of speaker, type of sentence, position in the phrase, and asymmetry of distance between tempi: slow to normal is greater than normal to fast. The optimal fast rate was calculated to be almost double that of the optimal slow rate.

In the present contribution, a small selection of current methods for the investigation of speech timing is brought into focus, with particular emphasis on syllable duration patterning. We take the pragmatic position that theory and its empirical grounding are heavily influenced by available methods, techniques, procedures and tools, and consequently we do not concentrate on linguistic or cognitive theories of rhythm. The methods we select focus mainly on the computational treatment of large corpora.

The Polish and English data used include the following:

1. Analysis of 'authentic data', i.e. speech which is not elicited for the specific purpose of analysis.

2. Analysis of a well-defined data set via perceptual judgments by selected subjects.

3. Linguistic corpus analysis and functional interpretation of temporal properties of speech in relation to features of discourse, specifically focusing on gender differences.

2. A paradigmatic perspective on contextual factors

2.1. Cart analysis

More recently, technological advances have resulted in the use of techniques based on universal statistical tools such as CART (*Classification and Regression Trees*, first introduced by Breiman et al. 1984), cluster analysis (e.g. Everitt et al. 2011) and neural networks (used for duration analysis e.g. by Vainio 2001), with data obtained from large (and very large) corpora of continuous speech. However, it needs to be mentioned that although corpus-based models often guarantee, for instance, better naturalness of synthesised speech, and thus are strongly preferred in many practical applications, rule-based models are also still present. They can be developed with the support of available statistical techniques, but without costly speech corpora, and have thus found applications in situations where it is more important to achieve speech characterised by high speed while still retaining intelligibility and relative correctness (over "naturalness") (e.g. Moos & Trouvain 2007; Moers et al. 2010). In fact, nowadays it is often the case that the two approaches overlap, and careful linguistic feature extraction is usually an important stage preceding the actual statistical processing. Linguistic knowledge may be used not only at the data preparation stage, but also in the modelling process itself (van Santen 1993; Möbius & van Santen 1996).

Studies vary in the choice of the unit used as the base for segmental duration modelling. Frequently, the phone is used as the unit, though Campbell's model (1992) analyses phone duration as dependent on syllable properties. The huge number of combinatory possibilities for units in natural speech generate a large space of coarticulation and other inter-unit effects (van Santen 1993): unnatural distortion results at concatenation points which do not capture these effects, even if a TTS (text-to-speech) system otherwise works well.

Unauthenticated
Download Date | 12/13/15 8:16 AM

A related challenge for acoustic inventory design and acoustic modelling is the high rate of occurrence of rare events, the so called LNRE problem (Large Number of Rare Events; Möbius 2001). A compromise between database size and sufficient coverage of unit combinations can be reached by optimising the contents of the database, e.g. using greedy set covering algorithms, i.e. heuristic approximation based on locally optimal choices (Buchsbaum & van Santen 1997) and by manipulating the size of units used for unit selection. Non-uniform unit selection has been reported to result in a good quality of synthesised speech for many languages (e.g. King et al. 1997): selecting longer concatenation units is expected to result in a smaller number of glitches at concatenation points, and a more natural sound. However, for highly inflecting languages (e.g. Polish, Turkish, Arabic) it is especially challenging to use larger concatenation units, because a very large number of inflected forms in these units would be required.

Setting unit selection preferences by means of cost functions and penalties influenced by constraints from structures at different levels is another strategy for improving duration models. In a duration model developed for the Polish BOSS synthesiser (Szymański et al. 2011) the best results of perception tests as regards the quality of synthesized speech were achieved when the system's unit selection algorithm was set up to use phone level units as the basis with a duration model containing features from both segmental and suprasegmental levels of utterance structure (Klessa et al. 2007). Thus, although the unit selection algorithm is phone-based only, information from different levels of utterance structure is provided.

The CART statistical method of analysis is based on two kinds of tree techniques for solving the tasks of (1) classifying objects (for categorical variables) and (2) predicting the actual values of a feature (continuous variables). In the case of segmental duration modelling, both types of tasks are highly useful, due to the fact that duration models need to be based on various types of (often interacting and interdependent) variables. The target task of creating a duration model (and predicting durations) can be solved using, for example, nominal categorical variables (such as the type of vowel, place or manner of articulation), numerical variables (the length of a syllable, word or foot containing the sound in question expressed in time units or as a number of component sub-units), and also ordinal categorical variables (the position of a sound within a higher structure, e.g. a syllable or a word). Generally, the aim is to define a set of logical *if-then* split conditions that allow prediction or classification of cases. Example conditions for duration prediction might include instances such as: is this the sound /a/? – if yes, then is the sound position within the syllable structure "onset"? – if not, then is the sound's manner of articulation "fricative"? etc.

Among other things, CART-based models surely owe their popularity to the availability of easy automatic construction of the models (e.g. King et al. 2003). However, although the tree building procedures are automated, the input for CART still depends on corpus data, so it is crucial to provide high-quality annotations and to define features whose values will be derivable from the data. The influence of the features is usually analysed in several stages during model development: separately for individual features or for small subsets of a larger feature set (using various statistical methods such as analysis of variance or correlations between factors) and with the use of the whole set of features. The *wagon* CART building programme (King et al. 2003), for example, offers an automated stepwise option that incrementally finds features that contribute most to the predicted variable within a specific

Unauthenticated
Download Date | 12/13/15 8:16 AM

feature set. The feature set is treated as a whole, and the correlation of particular features is expressed as a cumulative correlation, i.e. the features are ranked in a way that the most contributive feature is treated as the best, and the correlation of each of the subsequent features is increased by a number depending on their percent contribution to the overall mean correlation of the feature set. This provides the possibility of observing the impact of the inclusion of particular features on the overall result of the developed feature set. For instance, in the Polish BOSS duration model obtained with such a CART prediction procedure, the context information for phone duration is provided for the phone in question and for three adjoining left and right context sounds. The features in the final set relate to the current phone identity, its manner/place of articulation, presence of voice, and sound position as regards higher-level units. The correlation obtained with the final 57-element set of features was 0.8 (with RMSE at 15.4, and Error at 11.3451).

2.2. Measuring and perceiving speech rate

When speaking of speech rate or tempo, the fundamental question is the definition of what actually is meant by the terms and how the accepted acoustic or articulatory measures are related to human perception of speech rate. In order to address these questions, it might be helpful to mention at least several concepts and definitions. First, there is the distinction between the *objective* (actually realised and measurable/quantifiable) and *subjective* speech tempo (depending on individual judgment, referring to either intended or perceived tempo). Then there are the notions corresponding to the time span under consideration, according to which speech rate can be seen as *global* / *long-term* (related to the whole uttered text, sentence, individual characteristics of a person's speaking style) or *local* / *short-term* (local variations of tempo within the uttered text). A related issue will be the multi-directional relationships between the global and local rates, both as regards acoustic measurements and perception-based rate judgments (cf. Wagner & Windmann 2011). Another dichotomy comes from the distinction of *gross* (including pauses) and *net* (excluding pauses) speech tempo. Respecting the *gross* vs. *net* distinction may be especially important when dealing with longer-term speech rate, in terms of acoustics or the perceptual assessments of speaking rate as a characteristic of a longer stretch of speech or of a person's speaking style.

Speech rate can be understood and thus measured in various ways depending on the accepted definitions and prospective application of the measurement results. The same refers to the choice of the base unit and the interval for calculations (syllables, speech sounds, morphemes, words or even sentences per unit of time in milliseconds, seconds or minutes). Łobacz (1976b) points out issues related to the ease of discerning limits of the units and the desired unambiguity of their borders, and on the other hand the questions of reductions, omissions or transpositions of segments in different realisations of the same text. Word-based rate measures are in some cases preferred (e.g. Syrdal et al. 2012) due to the ease of distinguishing words in transcripts. However, the apparent ease might not always be borne out in reality, especially in the case of comparative studies. When comparing measurements based on words or phones with models constructed for automatic speech recognition, the results achieved with phone-based rate measures were significantly better than those achieved with rates calculated using words as the basic units (Siegler & Stern 1995), due especially

Unauthenticated
Download Date | 12/13/15 8:16 AM

to differences in word lengths or structures. An example of a problematic word element in Polish might be the case of non-syllabic prepositions *z*, *w* (pronounced /z/ or /s/ and /v/ or /f/, respectively, depending on the presence of voice in the directly following context). In fact, for technical applications, these prepositions are often treated not as independent units but as parts of subsequent syllabic words; in this way the pronoun becomes merged with the neighbouring word. Such a solution was chosen for the Polish BOSS synthesiser (Demenko et al. 2010). Pfitzinger (1996) compared automatic estimations of speech rate using local phone rate versus syllable rate, and claimed that although both of these measures gave significant and similar results, they were not identical and thus the overall speech rate measure should be treated as a combination of the two types of measures rather than as any of them separately.

Regardless of significant rate variation across particular utterances produced by a speaker, the speaker's overall rate can be viewed as his/her individual characteristic. As an acoustic correlate, the "individual" speech rate can be treated as the mean rate per unit of time (probably differing for a particular type of speech: read, spontaneous, affective, etc.). The perceived rate of speech may be characterised by a range of cues (e.g. pausing schemes, articulation), and the weight attributed to these cues by the listeners can depend on various factors such as variability of the cues in the signal (Grosjean & Lass 1977). In the perception domain, listeners also somehow compensate for the local speech rate variations or use them as cues to formulate a general impression of overall rate. Although the task of assessing the speech tempo often appears to be a quite intuitive and easy task for a listener, the exact manner of compensation and specific ways of using the cues are not obvious. Speech rate perception can be affected by both the intended and the realised rate (Koreman 2006), as dependent on the actually perceived speech signals as well as on the listener's previous knowledge and their own speaking habits. Thus yet another complicating issue is the subjective nature of listeners' judgments.

In most studies, speech rates are grouped into two, three or sometimes five categories (fast-slow, fast-neutral-slow, medium-fast/slow, etc.). However, as was observed by Łobacz (1976b), the "distances" between the nominal categories of speech rate are not symmetrically distributed around the *neutral* speech rate, which might suggest a need for verification of the categorisation. A possible starting point might be made by using a more sophisticated or a continuous rating scale for speech tempo assessment (Treiblmaier & Filzmoser 2009; Arnold et al. (2011): on prominence rating scales).

In the next two subsections the results of speech rate measurements and perceptual assessment in Polish read speech are discussed. The main goal of the first part of the study (Section 2.3) is to inspect selected quantitative measures of speech rate expressed in sounds, syllables and words, and make a comparison with dialogues. For the second part (Section 2.4), it is aimed to compare the measurements with perceptual judgments of global tempo, and to investigate the assessments obtained with the use of a continuous rating scale. As the text material, Aesop's Fable *The North Wind and the Sun* (for Polish transcription see Jassem 2003) is used. The recordings come from the Paralingua corpus (Klessa et al. 2013) and were realised according to two scenarios: the speakers were asked: (1) to read the text neutrally using their habitual reading style, and (2) to read the text as if they were pushed for time but still needed to read the text in an understandable way. For the present study, the

Unauthenticated
Download Date | 12/13/15 8:16 AM

recordings of 6 speakers reading the text twice (according to each of the two scenarios) are used, as well as, additionally, the recordings of 2 more speakers reading the text only once (one speaker in scenario 1, and one in scenario 2).

2.3. Some quantitative measures of tempo

Figure 1 shows the obtained measurements of mean syllable, sound, and word rates per second as well as the total number and mean duration of pauses produced by six speakers (for better comparability of tendencies, some values were scaled as shown in the legend; the figure depicts results only for 12 out of 14 speakers, i.e. those who read the text twice, but the numbers in the text are given for all participants). As can be seen, the mean values of the syllable and word rate tend to differ in a similar way across speakers, while the mean sound rate differences show more inter-speaker variability. It should be noted here that the very high correlation (see also Table 1 for numbers) between the syllable rate and word rate might be partly explained by the repetitive occurrence of several monosyllabic words, most of them containing complex consonant clusters (e.g. wiatr /vjatr/ or płaszcz /pwaSt^S/). Since the text is not phonetically balanced and the number of speakers is limited, the obtained rates ought to be treated only as rough estimates of tempo in Polish read speech. The results generally confirm the figures reported by Łobacz (1976b), where fast phone rate was found to lie between 14.3 and 16.2 and normal phone rate between 11.8 and 14.6. In the present results, the overall means for all speakers were 16.69 and 13.66 for fast and normal intended speech respectively. The fast tempo was significantly higher than the maximum in the study of Łobacz in the case of three speakers, and in the case of normal tempo all speakers used rates within the respective range, except for speaker H, who spoke the fastest overall.

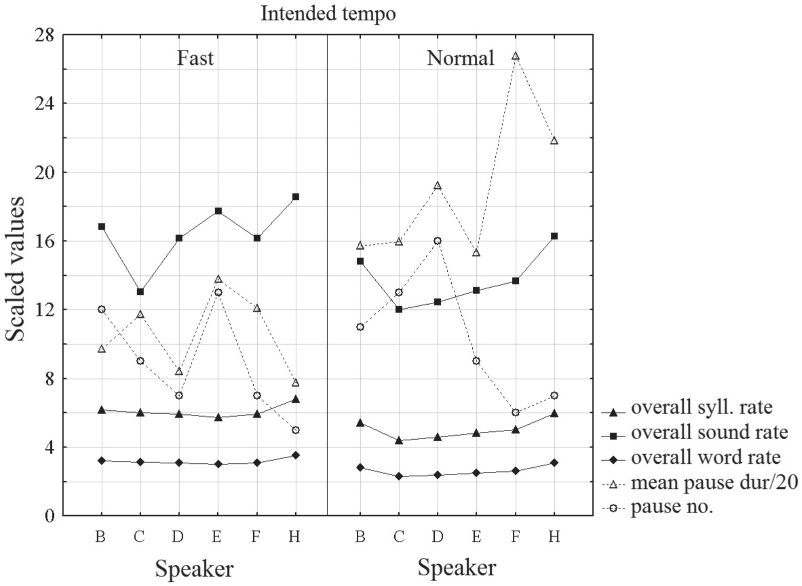


Figure 1: Mean values of: syllable, sound, and word rates, number and duration of pauses for six speakers in fast and normal intended speech tempo

Unauthenticated
Download Date | 12/13/15 8:16 AM

The number of pauses appeared to be slightly higher in normal intended tempo (the mean number of pauses was 8.43 for fast speech and 10.71 for normal), apart from the results for two speakers (one of whom was speaker H, who had the fastest rate of all). On the other hand, in these two cases, the mean duration of pauses was significantly higher than in the case of any other speaker, which might suggest a kind of compensation for the lower number of pauses by pause lengthening. The mean durations of pauses differ significantly in fast and normal rates, being consistently (and not surprisingly) higher in the latter (206.24 / 373.69). In the case of speaker E, the smallest differentiation as regards the number and length of pauses can be observed in the two intended rates, however at the same time for this speaker the largest difference between mean phone rates was noted (4.62) which in turn might show that this speaker's preference was to differentiate rates by altering articulation rate rather than pausing schemes.

In order to further examine the timing properties of fast and slow read speech, selected global measures of timing were performed for the above material with the TGA tool (Section 3 and Gibbon 2013) and compared with the results obtained from six Polish dialogues (details in Section 4). The results are based on 148 interpausal time groups for read speech and 390 groups for dialogues. As can be seen in Figure 2, apart from the expected difference in overall durations between the fast speakers and the remaining ones, the largest discrepancies can be observed in the overall and mean slopes between read and conversational speech, which might be regarded as a confirmation of the tendency reported in Section 3.2 below, i.e. the slope being a potential style or genre marker. However, this observation requires further verification, especially due to the speaker-related differences in slopes in dialogues. Another discrepancy can be seen in the tendencies for SD measures, especially the overall SD, which appears to differ for each of the three datasets.

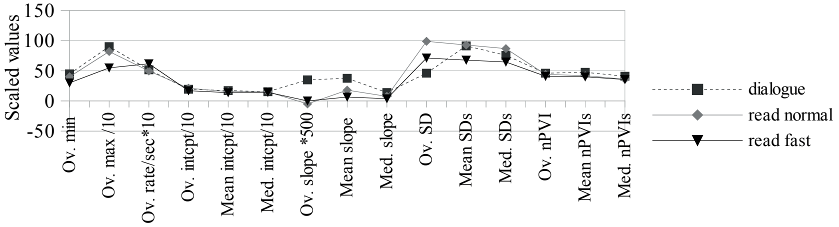


Figure 2: Comparison of selected quantitative measures of timing in read speech (fast and normal rates) and dialogues

2.4. Perception-based subjective assessment of tempo

The same recordings of read speech were used in a perception test in which 23 listeners (students of the same linguistics department) were asked to perceptually assess the speech rates of the speakers. During the test, the signals were played in a random order to each subject individually via headphones. Participants were instructed to listen to each of the recordings, and were also allowed to replay the recording or its fragments. Subjects were presented with the rating scale and the method of rating before they started to listen. After listening, the task was to mark their own subjective judgment of the speaker's overall speech tempo on a continuous scale without any number or scale given (only *min-max*

Unauthenticated
Download Date | 12/13/15 8:16 AM

markers at the ends of the scale). It was emphasized that the task was not to compare rates between particular recordings, but rather to express one's personal judgment or impression. After marking an answer, the subject could modify it (as many times as desired) but only until she/he proceeded to a recording of another speaker – it was not possible to alter the ratings afterwards.

Table 1: Correlation table for overall rates expressed in syllables, sounds, words per second, mean of perceptual ratings, number of pauses and mean duration of pauses

Coefficients in bold significant with *p* < .05000	Overall syllable rate	Overall sound rate	Overall word rate	Mean of ratings	Pause number	Mean pause dur.
overall syllable rate	1.000000	**0.873802**	**0.999914**	**0.957529**	**–0.627509**	**–0.635721**
overall sound rate	**0.873802**	1.000000	**0.874916**	**0.827639**	–0.493676	**–0.539552**
overall word rate	**0.999914**	**0.874916**	1.000000	**0.958162**	**–0.628586**	**–0.637595**
mean of ratings	**0.957529**	**0.827639**	**0.958162**	1.000000	**–0.605941**	**–0.715032**
pause number	**–0.627509**	–0.493676	**–0.628586**	**–0.605941**	1.000000	0.163704
mean pause duration	**–0.635721**	**–0.539552**	**–0.637595**	**–0.715032**	0.163704	1.000000

The results of the perception test showed that the listeners' judgments of speech rate were generally in line with the speakers' intentions (all recordings intended as fast obtained mean ratings above the general mean, and conversely, all "normal" ones were given rates below the overall mean). Table 1 presents correlations between the mean rate expressed in syllables, sounds, and words, and also the mean of perceptual ratings, pause number and mean pause durations. The perceptual ratings were found to be highly positively correlated with the overall syllable rate and overall word rate (corr. above 0.95). The correlation with the phone rate is also positive and statistically significant, but slightly weaker. As was already mentioned above in Section 2.3, there is a very high correlation between word and syllable rates, thus at this stage it is not conclusive whether the listeners based their judgments more on word or syllable rate cues. The negative correlation of ratings with the number and duration of pauses is also significant, with pause duration being a little more influential (–0.71) than the pause number.

In order to examine the outcome of using the continuous rating scale in the perception experiment, a tree diagram (Figure 3) was produced as a result of cluster analysis performed with Statistica software. Partitions of the results visualised on such a tree diagram can be achieved by cutting the tree at a specific height (*y*-axis value). In the search for methods of attaining the optimal cutting level, several approaches have been developed (cf. e.g. Everitt et al. 2011: 95–96). Considering the standard agglomerative clustering, the division should be made at a height "such that clusters below that height are distant from each other by at least that amount", thus informally suggesting the number of clusters. In Figure 3, two main clusters of judgments can be distinguished (cutting at a distance of ca. 60); however, the optimal clustering might be expected with cutting either at an agglomeration distance of 30 (thus giving 3 categories of speech rate) or at a distance of 10 (resulting in 5 categories).

Unauthenticated
Download Date | 12/13/15 8:16 AM

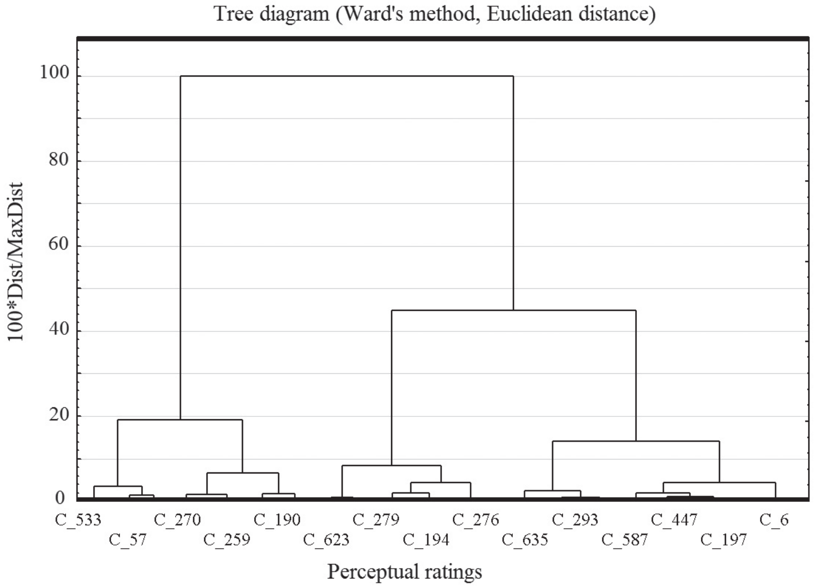


Figure 3: Cluster analysis of the perception test results: agglomeration tree diagram

For the two hypothesised groupings, a *k*-means clustering was performed to look at the means of ratings grouped into 3 or 5 clusters of greatest possible distinction. The results are given in Table 2. All distances of means between clusters are significant, and range from 13.6 to 18.78 for the 5-cluster grouping, while for the 3-cluster grouping the difference in means between cl.1 and cl.2 (29.93) was slightly higher than between cl.2 and cl.3 (25.02). The results in Table 2 show means for clusters ordered according to the tree diagram (and not the rating values). This finding might tentatively be considered to contribute to the discussion initiated by Łobacz (1976a: 178–179), who found that speakers tended to differentiate more between slow and normal rates than between normal and fast rates (the extremely fast tempo being limited by physiological factors). However, the clustering results presented here are preliminary and need to be examined in more detail, especially as regards the qualitative validity of the grouping.

Table 2: Results of k-means analysis for 3 and 5 clusters (cl.) of rate assessments

No. of clusters	Mean for cl.1	Mean for cl.2	Mean for cl.3	Mean for cl.4	Mean for cl.5
5	49.7791	33.4365	63.3864	82.1709	14.9900
3	23.5961	53.5356	78.5576	N/A	N/A

Unauthenticated
Download Date | 12/13/15 8:16 AM

3. Syntagmatic aspects: time types, linearity, alternation, hierarchy

3.1. Time types: a framework for defining contextual factors

A theoretical framework, *Time Type* theory (Gibbon 1992; 2006), was developed for distinguishing between formal types of temporal structure: *Categorial Time* (e.g. duration as distinctive feature), *Relational Time* (e.g. parallel relations between different phonetic or phonological properties such as intonation and phrases or syllables and tone, or co-articulating phonetic features), and *Fuzzy Time*, that is, quantitative statistically measurable properties of speech signals. Time Type Theory was designed to provide a framework for linguistic and phonetic speech timing studies: Categorial Time as the distinctive feature 'long-short', Relational Time as isochrony, rhythmic alternation and hierarchical timing relations, and Fuzzy Time as the statistically accessible domain of speech signal measurements. The following discussion first addresses the quantitative linear models, known as 'rhythm metrics', at the Fuzzy Time level, followed the inter-level relations between the three Time Type levels. Time Type theory was applied by Carson-Berndsen (1998) in a computational linguistic approach to automatic speech recognition.

3.2. Linear models

Gibbon et al. (2005) and Gibbon (2006) regard rhythm as an epiphenomenon determined by many linguistic and cognitive factors, but abstract a number of properties for the structural component of an epiphenomenal approach. The *Base Unit* of a rhythm (or other timing relation) is pattern, generally a syllable or a foot (accented syllable plus unaccented syllables) consisting of a finite trajectory through an n-dimensional parameter space (pitch, duration patterns, segmental patterns in syllables, etc.). Sequences of Base Units are related by *Alternation*, i.e. dynamic traversal through at least two positions in the Base Unit parameter space (e.g. high-low pitch, CV syllable structure, long-short or strong-weak syllable patterns). The Base Unit sequences with Alternation must enter into an *Iteration* relation, i.e. the alternating base pattern must repeat with at least two occurrences. Finally, for a rhythm to be identified, the Base Units in a sequence with Alternation and Iteration must enter into an additional relation of *Isochrony*, i.e. the Base Units must be equal in length. Rhythmic Base Units are rarely exactly equal in length at the Fuzzy Time level, but are subject to *fuzzy isochrony* ('*sloppy isochrony*'): Base Unit durations are measured on a scale from more-or-less equal to more-or-less unequal, but may nevertheless be interpreted perceptually, and explained cognitively, as isochronous, within specifiable difference thresholds.

Several quantitative linear models have been proposed for speech timing. Table 3 summarises three of the most well-known models, which specifically address the topic of isochrony in presumed foot-timed languages, together with the extent to which the models fulfil the necessary conditions on rhythm. The methods using Linear Models are corpus-based, inductive, *a posteriori* procedures which start with input from annotated speech data, and extract time-stamps, differences between time-stamps (i.e. unit interval durations), and differences between durations (i.e. deceleration and acceleration of intervals).

Unauthenticated
Download Date | 12/13/15 8:16 AM

Table 3: Quantitative linear rhythm models (Scott et al. 1986; Roach 1982; Low et al. 2001)

Model		Description	Constraint fulfilment	
PIM	$\sum \left\| \log \frac{I_i}{I_j} \right\|$	Sum of the ratios of each foot to each other foot (the log function reduces the impact of longer feet).	*Basic unit:* *Alternation:* *Iteration:* *Isochrony:*	foot no no yes
PFD	$\frac{100 \times \sum \left\| MFL - len(foot_i) \right\|}{n \times MFL}$, $MFL = \frac{\sum \left\| foot_i \right\|}{n}$	Sum of absolute (unsigned) differences of each foot from mean, divided by the mean foot length (%, max = 100%).	*Basic unit:* *Alternation:* *Iteration:* *Isochrony:*	foot no no yes
nPVI	$100 \times \sum \left\| \frac{d_k - d_{k-1}}{(d_k + d_{k+1})/2} \right\| / (m-1)$	Mean absolute (unsigned) difference between neighbours (normalised by division by mean length of neighbours); scale from 0 to asymptote of 200.	*Basic unit:* *Alternation:* *Iteration:* *Isochrony:*	vocalic seq no yes yes

Gibbon et al. (2005) showed that there is a strong correlation between each of these measures when applied to syllables, and between these measures and standard deviation of syllable durations, and rejected claims that these linear models are models of rhythm, on the grounds that they do not account for rhythmic alternation (cf. also Gut 2012) because they operate on absolute (unsigned) duration differences.

In the remainder of this section the results of using a new tool, the Time Group Analyser (TGA: Gibbon 2013), to investigate syllable durational properties, some of them novel, using the ‘Syllables’ annotation tier of the Aix-MARSEC corpus of English (Auran et al. 2004), are reported. Six of the eleven genre categories represented in the Aix-MARSEC corpus were selected on the grounds of greater similarity of informally defined speech styles: A (‘Commentary’), B (‘News broadcast’), C (‘Lecture aimed at general audience’), D (‘Lecture aimed at restricted audience’), F (‘Magazine-style reporting’), K (‘Propaganda’). The functionally less similar five categories E (‘Religious broadcast including liturgy’), G (‘Fiction’), H (‘Poetry’), J (‘Dialogue) and M (‘Miscellaneous’) were not dealt with.

The following procedure was used:

1. Annotations in each genre category were analysed separately.

2. The annotations were divided into pause-delimited (inter-pause, interpausal) syllable groups.

3. For each genre, overall values for duration maximum, mean, range, intercept, slope, standard deviation and nPVI were automatically calculated with the TGA.

4. Values for all sequences were displayed together on a line graph in order to permit direct ‘eyeballing’ of similarities and differences between measures and between genres (further correlations were not investigated in this study).

Unauthenticated
Download Date | 12/13/15 8:16 AM

The results of the quantitative analysis of the genre categories are visualised in Figure 4. Some results are scaled (see legend of Figure 4) in order to create a visually interpretable combined display of values for each measure and each genre.

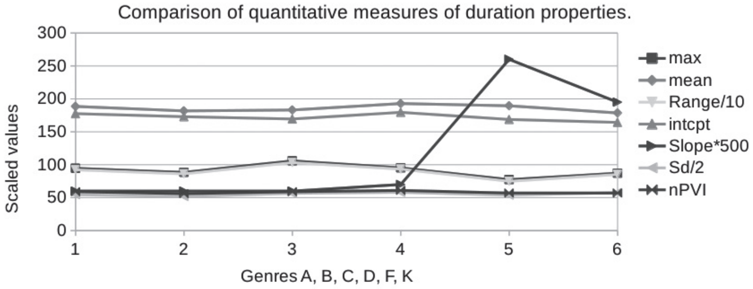


Figure 4: Comparison of quantitative measures in six Aix-MARSEC genre categories

Predictably, high correlations hold between mean and intercept, between SD and nPVI, and between range and maximum. The interesting parameter is slope: each case shows deceleration, i.e. average increase in duration over the pause-defined segment. The slopes for genre categories A, B, C and D (news broadcast and lectures) are close together, while the more informal, audience-directed genres F and K (magazine and propaganda) show much larger deceleration. This result suggests a phonostylistic effect, with syllable slope patterning over pause-delimited segments as a contribution to speech style, which needs further investigation in terms of speech rate, as well as more precise sociolinguistic specification of genre categories.

3.3. Alternation models

The second relevant property of speech timing is alternation. The Linear Models fail because they lack this alternation detection property. One approach to characterising alternation in speech timing is the Oscillator Model, incorporating quantitative measures of rhythm as oscillations in perceptions of relative rhythmicity (cf. Barbosa 2009; Inden et al. 2012). The present approach using the TGA tool takes a more opportunistic approach, and retains the essential unit interval duration difference property of the Linear Models (referred to here as *ΔD*), extracted in the same way from speech signal annotations, but also has an alternation detection property. Unlike in the Oscillator Models, instead of attempting to characterise ‘always on’ oscillators, the interval duration differences are tokenised into discrete units (increase, decrease and equality of duration), and a distributional analysis of the frequencies of these interval duration tokens is made, following familiar computational procedures from corpus linguistics.

The initial output of the Alternation Model is a stream of *ΔD* tokens: for this conversion, minimal duration changes are defined by means of an adjustable local threshold, typically around 50 ms, and changes below this threshold count as equal duration (currently thresholds are investigated manually; no algorithmic optimising search is performed). The *ΔD* tokens are represented as symbols: equality (‘=’), acceleration (‘/’) and deceleration (‘\’).

Unauthenticated
Download Date | 12/13/15 8:16 AM

Threshold-determined equality will be referred to as *fuzzy isochrony* or *sloppy isochrony*. To some extent, the procedure parallels for duration some of the stylisation procedures used in considerations of pitch: for the analysis of pitch into discrete entities (e.g. ’t Hart et al. 1990; Auran et al. 2004).

Second, in order to identify alternating, isochronous or random duration tendencies, frequencies of token digrams, trigrams, quadgrams and quingrams are measured.

In view of the methodological emphasis of the present contribution, the token *n*-gram frequency analysis procedure is illustrated using a single monologue file *a0102B.TextGrid* from the Aix-MARSEC corpus. The results are shown in Table 4. The table shows the first five ranks for frequencies of digram, trigram, quadgram and quingram *ΔD* token patterns at local threshold settings of 0 ms, 20 ms, 40 ms, 60 ms and 80 ms. Figures given are percentages and, in parentheses, absolute numbers.

Inspection of the rows in Table 4 shows that the threshold values 0 and 20 lead to almost identical results for all of the top three *ΔD* token pattern ranks, indicating a prevalence of alternations, and only rare threshold-determined equality. At 40 ms the situation starts to change, with more equalities, and from a threshold of 60 ms there is increasingly a preponderance of equalities. Informally, these results indicate a source of evidence for a limit of around 50 ms on the contribution of duration differences to the identification of isochrony in this English text.

There are a number of consequences to be drawn from this analysis in terms of further clarifications which are needed, but which are not within the scope of the present contribution:

1. The 50 ms limit itself is very likely an indication of a structurally relevant boundary. However, this can only be verified by examination of the linguistic constructions associated with the *ΔD* token patterns.

Table 4: Stylised duration difference token patterns for Aix-MARSEC files with A initial. Tokens: \ (increasing), / (decreasing), = (equal), + (initial pausal unit boundary), # (final pausal unit boundary)

		LT = 0		LT = 20		LT = 40		LT = 60		LT = 80	
Unit	Rank	Count	Pattern	Count	Pattern	Count	Pattern	Count	Pattern	Count	Pattern
2-gram	1.	24% (65)	/\	20% (55)	/\	15% (41)	/\	17% (46)	==	24% (64)	==
	2.	23% (61)	\/	18% (48)	\/	13% (34)	\/	11% (29)	=\	11% (29)	=\
	3.	13% (36)	\\	9% (24)	\#	9% (24)	\=	10% (26)	/=	10% (26)	\=
3-gram	1.	17% (39)	\/\	13% (31)	\/\	9% (21)	\/\	8% (20)	===	12% (29)	===
	2.	13% (31)	/\/	10% (23)	/\/	7% (17)	/\/	6% (13)	==\	8% (18)	==\
	3.	9% (21)	/\\	6% (13)	/\\	5% (11)	=/\	5% (12)	\/=	6% (15)	\==
4-gram	1.	10% (20)	\/\/	7% (14)	\/\/	5% (10)	/\/\	4% (8)	====	5% (11)	===\
	2.	9% (18)	/\/\	7% (14)	/\/\	4% (9)	\/\/	3% (7)	===\	5% (11)	====
	3.	5% (11)	\/\\	4% (8)	=\/\	3% (7)	=\/\	3% (7)	==/\	4% (9)	\===
5-gram	1.	6% (10)	\/\/\	5% (9)	\/\/\	4% (6)	\/\/\	3% (5)	==/\/	4% (6)	====\
	2.	5% (9)	/\/\/	4% (7)	/\/\/	3% (5)	\=/=\	3% (5)	+====	3% (5)	=\===
	3.	5% (8)	\/\//	3% (5)	/\/\\	2% (4)	/\/\\	2% (4)	====\	3% (5)	+====

Unauthenticated
Download Date | 12/13/15 8:16 AM

2. One fundamental problem of the so-called rhythm metrics is that they can identify degrees of isochrony, but in the direction of non-isochrony the values become less and less meaningful, since they do not distinguish between alternating and random sequences. The *ΔD Analysis* procedure outlines a path forward in this respect.

3. Another fundamental problem of the so-called rhythm metrics is that they do not employ thresholds, but indiscriminately incorporate all duration differences, however small.

There are a number of open issues with the *ΔD Analysis* procedure, which are currently under investigation, concerning automatic threshold optimisation, numerical weighting of *ΔD* tokens, further numerical evaluation of the *ΔD n*-gram distributions to induce a 'rhythm grammar', and, not least, alignment of *ΔD* token patterns with grammatical patterns in order to determine the significance of *ΔD* thresholds.

However, the general conclusion is that this novel method provides one interesting way forward for identifying the essential alternation properties of rhythm, and thereby correcting a core weakness of so-called rhythm metrics which ignore alternation.

3.4. Hierarchical models

The two best-known hierarchical models of speech timing are those of Jassem & Abercrombie (cf. discussion in Gibbon et al. 2012) for English, which identify the 'rhythm group' or 'foot' as a basic unit with syllable components. These models have become standard models for providing frameworks for statistical analyses. The Jassem model identifies two units, the *Narrow Rhythm Unit*, *NRU*, which starts with a stressed syllable and continues (optionally) with unstressed syllables until the next clear word boundary, and the (optional) *Anacrusis*, *ANA*, a sequence of unstressed syllables from a clear word boundary to the beginning of the next NRU. The Jassem model claims that the ANA and the NRU differ in their timing properties: each NRU in a sequence tends towards equal length (conditioned by the number of syllabic and phonemic constituents it contains), while the ANA tends to be faster, less stressed, and less constrained towards isochrony. A sequence of *ANA* and *NRU*, bounded left and right by clear word boundaries, constitutes a *Total Rhythm Unit*, *TRU*. The Abercrombie model, on the other hand, postulates only the foot, defined in a similar way to Jassem's NRU, and introduces the concept of the 'silent beat', which relates indirectly to Jassem's ANA. Both models are candidates for a rhythm theory, since the claims embody a clear Base Unit (the foot), Alternation (stressed-unstressed syllables), Iteration (foot sequences), and Isochrony (tendency to equal NRU or foot timing). Jassem et al. (1984) demonstrated the quantitative validity of the Jassem model; investigations of the Abercrombie model have been less successful, which has in turn led to pessimism about finding quantitative rhythm correlates in the speech signal.

Campbell (1992) investigated hierarchical structures in speech timing from several perspectives, including the dependence on phone durations on syllable properties, and at a higher level the relation of syllable durations to prosodic structure (using Break Indices marking different levels in a hierarchy of boundaries between phonological and prosodic units) and grammatical structures. He found a number of tendencies: syllable durations tend to shorten in proportion to the hierarchical depth of a preceding grammatical phrase structure boundary, and lengthen in proportion to the hierarchical depth of either a following

Unauthenticated
Download Date | 12/13/15 8:16 AM

grammatical phrase structure boundary, depth of grammatical embedding, or a following prosodic boundary in terms of Break Indices (cf. Figure 16.12 in Campbell 1992).

The present approach to hierarchical modelling introduces the notion of *Time Tree Induction*, which, like the Linear Model and Alternation Model approaches, is a data-driven *a posteriori* approach, in contrast to approaches which start with *a priori* models, such as linguistically motivated Prosodic Hierarchy trees. In this sense, the Time Tree Induction approach builds on the Linear Model and Alternation Model approaches, and extends Campbell's duration-hierarchy correlation model. A first attempt to compare *a posteriori* duration hierarchies to *a priori* grammatical hierarchies was made by Gibbon (2003; 2006).

Like Alternation Model analysis, *TTI* is also determined by relations between accelerating or decelerating tokens, except that, in contrast to discrete token sequence analysis, the numerical durations are used for tree induction. Currently the induction algorithm uses either deceleration relations or acceleration relations, but not both. The following rules define binary decelerating (short-long) trees, for example:

(1) A syllable s_i is a tree constituent.

In a tree constituent sequence $S = < s_i , s_j >$, if $dur(s_i) < dur(s_i)$, then S is a tree constituent with the duration label $dur(s_j)$.A bottom-up algorithm applies the rules until no more applications are possible. Trees with other structures emerge, depending on several factors: (1) how '=' is dealt with (e.g. as '>=' or 'not >'), (2) with '>' (acceleration) instead of '<' (deceleration), and (3) whether a right-left or left-right schedule together with early or late recursive closure is used to implement the grouping criterion.

The following illustration of the procedure uses the duration-annotated sequence for one inter-pausal group which was extracted automatically from the monologue file *a0102B.TextGrid* from the Aix-MARSEC corpus:

'mO::160 'nju:z:330 @:60 'baUt:150 D@:100 're:160 vr @n:210 'sVn:290 'mjVN:290 ,mu:n:500

A left-right recursive algorithm applies the specified *ΔD* criterion to the current and following input-level annotation durations to create a binary subtree; if the criterion fails, a stack of previously constructed subtree constituents is examined in order to create larger subtrees, and if this fails, the bottom-up search for a new subtree restarts. (Note that an alternative algorithm which processes the stack immediately after successful input-level construction may lead to different results.) The TGA tool computes the derivation step by step (cf. Table 5). The automatically generated output of the implementation is a parsed tree-bracketing (which is visualised as a tree graph in Figure 5):

(('mO: 'nju:z) ((((@ 'baUt) ((D@ 're) vr@n)) 'sVn) ('mjVN 'mu:n)))

Table 5: Time tree derivation

1. 160 330 60 150 100 160 210 290 290 50
2. (160 330) 60 150 100 160 210 290 290 500
3. (160 330) (60 150) 100 160 210 290 290 500
4. (160 330) (60 150) (100 160) 210 290 290 500
5. (160 330) ((60 150) (100 160)) 210 290 290 500
7. (160 330) ((60 150) ((100 160) 210)) 290 290 500
8. (160 330) (((60 150) ((100 160) 210)) 290) 290 500
9. (160 330) (((60 150) ((100 160) 210)) 290) (290 500)
10. (160 330) ((((60 150) ((100 160) 210)) 290) (290 500))
11. ((160 330) ((((60 150) ((100 160) 210)) 290) (290 500)))

Unauthenticated
Download Date | 12/13/15 8:16 AM

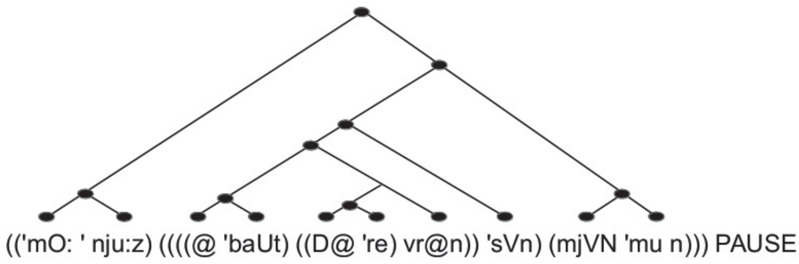

(('mO: ' nju:z) ((((@ 'baUt) ((D@ 're) vr@n)) 'sVn) (mjVN 'mu n))) PAUSE

Figure 5: Time Tree parse with the ΔD iambic criterion

Comparison of the Time Tree with grammatical units reveals six correspondences (given in orthography, for readability): '*more news*', '*about*', '*the Reverend*', '*about the Reverend*', '*about the Reverend Sun Mun Moon*', and, non-trivially, the whole inter-pause unit '*more news about the Reverend Sun Mun Moon*'. Two sequences do not correspond exactly to grammatical units: '*the Re*', '*the Reverend Sun*', of which the sequence *'the Re*' can be analysed as ANA in the Jassem timing model, followed by a more prominent '*verend*'. A tree-comparison algorithm has been used to determine the degree of similarity between Time Trees and grammatical trees (Gibbon 2003; 2006). Experiments with an acceleration condition yield a largely right-branching structure which does not yield any correspondences with grammatical or other plausible units beyond suffixed words. The *ΔD* relations are not necessarily related to rhythm, though symmetries in the tree may provide clues to rhythmic patterns. However, grammatical structure, not rhythm, is at issue at this point.

Clearly, in view of the number of degrees of freedom depending on the selected duration difference criterion and parse schedules, further levels of automation are required in order to search the space of relations between Time Trees and grammatical structures.

Finally, the genre under consideration ('Commentary' by a female speaker) represents a somewhat formal, rehearsed style, where prosody-grammar correspondences may be expected. It is not only duration and grammatical structure which are likely to correlate, but also semantically and pragmatically motivated constrastive and emphatic structures, while on the phonetic side pitch patterning will also be involved, as well as effects of intrinsic phone duration on syllable duration and hence on the duration trees. These complexities require extensive further research.

4. Functional interpretation of timing in dialogue

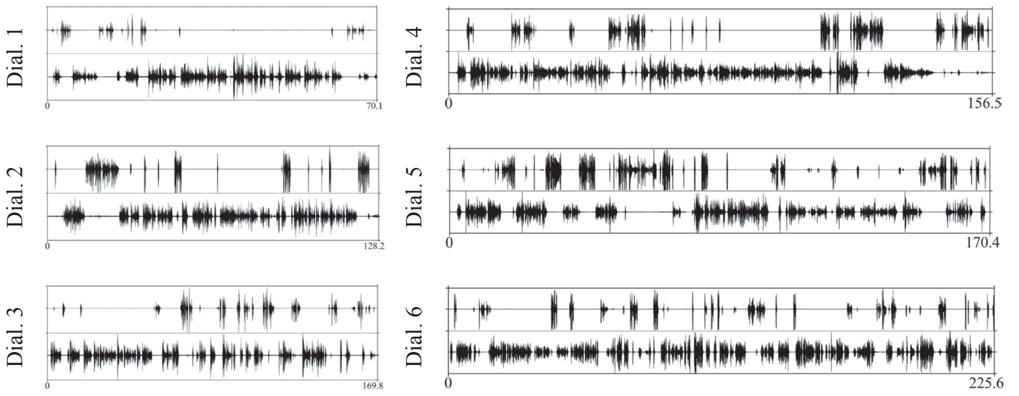
Speech timing functions at several levels in dialogue: in turn-taking (relative length of turns, gaps and overlaps between turns), and within turns (pauses, prominence patterns and hierarchical rhythm structures). To investigate sociophonetic timing in dialogue in connection with the phonetic alignment or non-alignment of participants, a scenario was designed in which misunderstandings are elicited: speaker A has a caller role and gives instructions to speaker B, in a call-centre role, about how to get from a hospital to a person with a heart attack. Because speakers' maps differed a little, misunderstandings occurred and the speakers had to negotiate the route in order to finish the task (Bachan 2011). The dialogues were

Unauthenticated
Download Date | 12/13/15 8:16 AM

conducted in Polish between Polish native speakers, and recorded in stressful conditions between people who did not know each other. Six dialogues (total duration 15 min 20 sec, three male, three female) were recorded, annotated at syllable level and analysed using descriptive statistical methods.

The following discussion addresses the specific questions of whether there are gender or role differences in stressful dialogues, and which speech timing models perform better than others in this task. The oscillograms in Table 6 illustrate the turn-taking activity of the dialogues: speaker B does not have a simple listener role, but gave a lot of feedback to speaker A about whether the instructions were understood. The upper and lower oscillograms show the speech of speaker B (call-centre) and speaker A (caller) respectively.

Table 6: Oscillograms of the female (left) and male (right) dialogues



Initial analysis of the temporal turn organisation showed that the female B speakers speak less than male B speakers, giving less feedback and enquiring less about the correct route. Deeper analysis of the dialogues showed that speech in female dialogues hardly overlaps, this occurring only when female speaker B misunderstood an instruction and speaker A interrupted speaker B to clarify. Different kinds of turn-taking occur. In Dialogue 2 speaker B gave belated positive feedback: speaker A gave speaker B time to provide positive feedback, both speakers were silent for a few seconds, then when speaker A continued, speaker B provided feedback to the previous instructions, perhaps due to speaker B's initially being silent while concentrating on marking the route on the map.

Male dialogues were much more lively and interactive, and their turn timing shows three phases: initial, medial and final. Initially, their speech overlaps in the greeting and introductory part of the dialogue (e.g. arranging what the task is and where to start). The male B speakers gave brief positive feedback, and their utterances were much longer when they were asking for information or providing information about understanding instructions or about where they were moving on the map. Although initially the speakers' speech overlapped, regardless of the function of the turn (positive feedback, information providing), in the course of the task, in the medial phase, the speakers tended to align, with speaker A waiting for speaker B to give positive feedback (no speech over-

Unauthenticated
Download Date | 12/13/15 8:16 AM

lap), before continuing with a further instruction. Also, when the B speakers were asking questions, speaker A waited until the question finished before answering. As with female speakers, overlaps happened when the instructions were misinterpreted by speaker B, and speaker A had to interrupt to clarify the route. In the final phase after the dialogue, when participants had accomplished the task, they took leave of each other, and their goodbye utterances again overlapped.

4.1. Quantitative analysis of dialogue

For quantitative analysis of the dialogue the TGA tool was used, with further evaluation as necessary. The annotations of silent pauses, speaker noises, intrusive noises, and laughter were treated as pauses. A set of different measures based on syllable timing within inter-pause groups was selected and investigated:

1. Overall timing properties: for each speaker, overall duration, minimum and maximum syllable lengths, syllable/sec speech rate.

2. Global tendencies: for each speaker, overall median, mean and normalised pairwise variability index (*nPVI*), i.e. mean differences between adjacent syllable pairs, normalised by dividing the difference by the mean of the pair.

Figure 6 presents the mean and median duration of syllables and the standard deviation. The overall mean durations vary within a dialogue (the exceptions are dialogue 1 and 2), whereas the overall median duration values are more similar. The standard deviation is very high, indicating a broad range of variation between very short (e.g. in fast speech) or very long (e.g. filled pauses and hesitations).

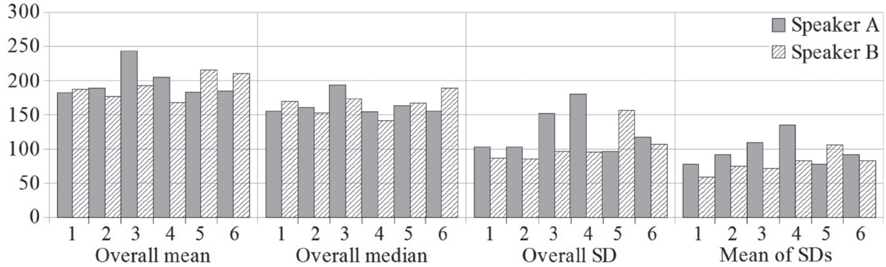


Figure 6: Mean and median duration of syllables and standard deviation in six dialogues

The *nPVI* values are presented in Figure 7. The overall *nPVI* values for all the dialogue pairs are almost the same – an exception is dialogue 5 (speaker A: 39, speaker B: 46), with smaller *nPVI* for female speakers and higher for male speakers. Across the dialogues, mean and median *nPVI* values are more diverse, but between interlocutors they tend to be more similar, indicating phonetic alignment of speakers within a dialogue.

The detailed results of analysis of the six dialogues are presented in Table 7. The analysis confirms the impression that both speakers were active in the dialogue: Comparison of the ‘Valid Time Groups’ shows that one of the speakers, here speaker A, spoke much more than speaker B.

Unauthenticated
Download Date | 12/13/15 8:16 AM

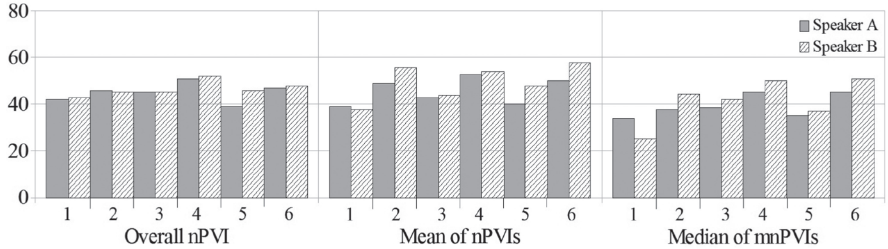


Figure 7: nPVI values for six dialogues

Table 7: Results of quantitative analysis of six dialogues

	Female dialogues							Male dialogues						
	Dialogue 1		Dialogue 2		Dialogue 3		**Mean**	Dialogue 4		Dialogue 5		Dialogue 6		**Mean**
Duration:	70.1		128.2		169.8		–	156.5		170.4		225.6		–
Speaker:	A	B	A	B	A	B	–	A	B	A	B	A	B	–
Age:	27	25	23	31	21	28	–	19	28	30	29	22	25	–
Overall duration	44357	10311	77365	25898	94915	37816	48443	107892	42018	100320	58364	144105	47121	83303
Overall min	42	55	48	59	31	39	45.67	25	44	62	41	39	54	44.17
Overall max	710	442	769	535	1002	607	677.5	1680	594	930	1577	1218	754	1125.5
Valid Time Groups	21	9	31	12	38	30	23.5	44	28	41	30	72	34	41.50
Overall rate/sec	5.48	5.33	5.29	5.64	4.11	5.18	5.17	4.88	5.93	5.43	4.64	5.39	4.75	5.17
Overall slope	0.18	0.65	0	0.07	0.09	0.16	0.19	–0.12	–0.12	0	0.08	0	–0.11	–0.05
Mean of slopes	24.11	33.67	29.29	75.01	22.38	43.57	38.01	14.1	40.45	7.88	67.13	35.32	57.88	37.13
Median of slopes	10.07	25.17	9.63	28.24	2.66	20.25	16	0.98	19.53	0.5	21.85	8.25	20.08	11.87

Clear gender differences are indicated by two variables. First, ‘Overall duration’ shows that female B speakers were silent about 66% of the time; male dialogues were longer; male B speakers spoke more, about 40% of the time; female and male speech rates are equal (5.17 syll/sec), but females in a dialogue had more similar speech rates except in Dialogue 3, while male speakers varied more in speech rate. Second, ‘Overall slope’ shows that in female dialogues, for female B speakers (instruction followers) the slope is steeper than for A speakers, which means that the B speakers slowed down their speech during an utterance. Male speaker slope values are less steep and even negative, suggesting that male speakers sometimes increased their speech tempo during an utterance. Overall slope values for male speakers are more similar in each pair, but ‘Mean of Slopes’ and ‘Median of Slopes’ for female and male dialogues show that speakers in the A and B dialogue roles differ considerably.

4.2. Comparison of female vs. male dialogues

In Figure 8 various measurements of the syllable duration, standard deviation and *nPVI* index in dialogues between female and male speakers A and B are presented. The overall mean and median of syllable durations for each group differ a great deal, which suggests that there are many extreme values (either very short syllables in fast speech or long syllables, i.e. hesitations and filled pauses).

Unauthenticated
Download Date | 12/13/15 8:16 AM

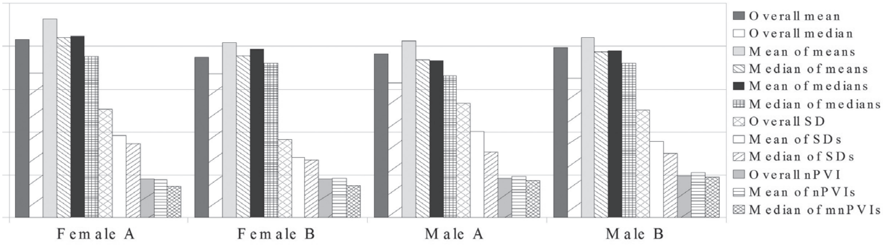


Figure 8: Measurements of syllable durations, standard deviation and nPVI index

Table 8: Quantitative results of the analysis of speech of female A and B and male A and B speakers

	Female A	Female B	Male A	Male B
Overall duration	216638	74024	352318	147378
Overall min	31	39	25	41
Overall max	1002	607	1680	1577
Valid Time Groups	90	51	157	93
Overall rate/sec	4.81	5.36	5.25	5.04
Components: Global Tendencies				
Overall mean	207.91	186.46	190.65	198.36
Overall median	168.5	168	157	163
Overall *nPVI*	45	45	46	49
Overall intercept	162.62	173.54	206.15	171.67
Overall SD	126.93	91.44	133.34	125.62
Overall slope	0.09	0.07	–0.01	0.07

Table 8 shows the summary analysis of the dialogues between female and male speakers A and B. The *nPVI* values (i.e. the overall mean and median) are almost the same for the female speakers, while male values diverge. The values of standard deviation are higher for A speakers, probably due to their changing their speaking style or speed from very fast speech when giving instructions to very slow hesitating speech and filled pauses when they could not find correct words to express themselves. The overall intercepts for B speakers are very similar, while the values for A speakers are quite different. However, when looking at the mean and median of the intercepts, the results of female and male A speakers are similar, as well as the results of female and male B speakers. The overall slope values for female speakers are very close, while the male values differ, even being negative for A speakers.

Table 9 shows the results of a summary comparison between A speakers and B speakers, as well as between female speakers and male speakers. The results show that A speakers spoke much more than B speakers, and also the male speakers spoke more than females. The overall minimum value is the smallest for male A speakers – caused probably by fast speech. The overall rate is similar, but the values for female speakers are the smallest. A similarity is seen between the overall mean and median values between speakers A and B, while the difference is larger between female and male speech. In all cases, females' syllable dura-

Unauthenticated
Download Date | 12/13/15 8:16 AM

Table 9: Summary table: speakers A vs. speakers B and female speakers vs. male speakers

	Speakers A	Speakers B	Females	Males
Overall duration:	568 956	221 402	290 662	499 696
Overall min	25	39	31	25
Overall max	1680	1577	1002	1680
Valid Time Groups	247	144	141	250
Overall rate/sec	5.08	5.15	4.95	5.19
Components: Global Tendencies				
Overall mean	196.87	194.21	201.99	192.86
Overall median	161	164	168	159
Overall *nPVI*	45	47	45	47
Mean of *nPVIs*	47	50	45	50
Median of *mnPVIs*	41	44.5	38	44
Overall intercept	209.25	173.21	164.41	203.28
Overall SD	131.33	115.02	118.6	131.22
Overall slope	0	0.04	0.05	0
Mean of slopes	23.28	52.84	33.87	34.34
Median of slopes	4.94	20.78	10	7.93

tions are the longest. The mean and median *nPVIs* values differ less between speakers A and B, while the difference is larger between female and male speakers. The mean and median value of the slope is the smallest for A speakers, indicating that their speech was fast and speeding up towards the end of the utterance. Standard deviation is high for all analysed groups of speakers.

4.3. Conclusions

The temporal structure of dialogues indicated a clear difference between female and male dialogues. Female dialogues were shorter, and the speakers' speech did not overlap much, apart from the misunderstandings and hesitations, while male speakers interacted a lot, interrupting each other, but finally also accommodating and reducing speech overlap. Such a difference may be caused not only by the female-male differences, but also by the specific nature of the task. It is suspected that males felt more comfortable when giving directions on how to get to the place and also in following instructions about turning left or right. The dialogue strategies differed between females and males. While females did not interrupt each other during speaking, males provided a lot of feedback and interrupted each other. However, in the course of the dialogues, the male speakers aligned their behaviours and did not start talking before the other speaker finished. In general, the B speakers slowed down in the course of their utterances, as shown by the slope high values, whereas the slope of A speakers was much smaller, even being negative overall for male A speakers. The standard deviations for all speakers were high, indicating that the speech was vivid and spontaneous.

Unauthenticated
Download Date | 12/13/15 8:16 AM

5. Summary and outlook

Both the study of the literature and the original research reported in this study reveal a wide variety of fruitful methodologies which have been and are continuing to be deployed in the study of speech timing. On the one hand, the complexity of identifying valid timing paradigmatic properties by means of contextual factors is made very clear by the *Classification and Regression Trees* (CART) studies. On the other, the need to examine the syntagmatic structures of linearity, alternation and hierarchy has also been demonstrated. Finally, the options for interpreting duration patterning at the discourse level from a functional point of view are clear.

The results of the various timing analysis methods can be used in various application scenarios. One very common scenario, which cannot be dealt with here, lies in the computational support of foreign language learning proficiency testing by objective comparison of duration properties of native speaker and foreign language speaker timing patterns. An open question concerns the possible potential of using the results of perception-based studies as a support for characterising long-term features of speech and speakers. These are ongoing research fields. Another scenario, to which the present study is closely related, is speech technology and dialogue system design. It is not only the paradigmatic and syntagmatic properties of timing patterns that are useful in this scenario, but also the sociolinguistic patterns which emerge from dialogue corpus study. The female-male differences showed that different dialogue strategies could be implemented in a dialogue system when interacting with females or males, though much further sociolinguistic research on the reasons for these differences is necessary, and it would not be advisable to apply these descriptive results without careful consideration of these reasons.

References

Arnold, Denis & Wagner, Petra & Möbius, Bernd. 2011. Evaluating different rating scales for obtaining judgments of syllable prominence from naive listeners. In *Proceedings of XVIIth International Congress of Phonetic Sciences*, 253–255. Hong Kong.

Auran, Cyril & Bouzon, Caroline & Hirst, Daniel. 2004. The Aix-MARSEC project: an evolutive database of spoken English. In Bel, Bernard & Marlien, Isabelle (eds.), *Proceedings of the Second International Conference on Speech Prosody*, 561–564. Nara, Japan.

Bachan, Jolanta. 2011. *Communicative alignment of synthetic speech*. Poznań: Adam Mickiewicz University in Poznań. (Doctoral dissertation.)

Barbosa, Plinio. 2009. Measuring speech rhythm variation in an oscillator-based framework. In *Proceedings of Interspeech 2009.* Brighton: International Speech Communication Association.

Breiman, Leo & Friedman, Jerome & Olshen, R. A. & Stone, Charles. 1984. *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Buchsbaum, Adam & van Santen L., Jan P. H. 1997. Methods for Optimal Text Selection. In *Proceedings 5th Euro. Conf. on Speech Communication and Technology*, Vol 2, 553–556. Rhodes, Greece.

Campbell, Nick. 1992. *Multi-level timing in speech.* Brighton, UK: University of Sussex (Exp. Psychol). (Doctoral dissertation.)

Carson-Berndsen, Julie. 1998. *Time map phonology: Finite state models and event logics in speech recognition*. Dordrecht: Kluwer Academic Publishers.

Cummins, Fred. 1999. Some lengthening factors in English speech combine additively at most rates. *The Journal of the Acoustical Society of America* 105. 476–480.

Unauthenticated
Download Date | 12/13/15 8:16 AM

Dechert, Hans W. & Raupach, Manfred (eds.), *Temporal Variables in Speech. Studies in Honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Demenko, Grażyna & Klessa, Katarzyna & Szymański, Marcin & Breuer, Stefan & Hess, Wolfgang. 2010. Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *International Journal of Speech Technology* 13(2). 85–99.

Everitt, Brian S. & Landau, Sabine & Leese, Morven & Stahl, Daniel 2011. *Cluster Analysis, 5th Edition*. King's College, London: John Wiley & Sons.

Gibbon, Dafydd. 1992. Prosody, time types, and linguistic design factors in spoken language system architectures. *Proceedings of KONVENS 1992*. 90–99.

Gibbon, Dafydd. 2003. Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of International Congress of Phonetic Sciences* III. Barcelona, 2489–2492.

Gibbon, Dafydd. 2006. Time types and time trees: Prosodic mining and alignment of temporally annotated data. In Sudhoff, Stefan et al. 2006. *Methods in Empirical Prosody Research*, 281–209. Berlin: Walter de Gruyter.

Gibbon, Dafydd. 2013. TGA: a web tool for Time Group Analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (*TRASP*). Aix-en-Provence.

Gibbon, Dafydd & Fernandes, Flaviane Romani. 2005. Annotation-mining for rhythm model comparison in Brazilian Portuguese. *Proceedings of Interspeech 2005*, 3289–3292.

Gibbon, Dafydd & Hirst, Daniel & Campbell, Nick (eds.). 2012. *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem*. *Speech and Language Technology* 14/15. Poznań.

Grosjean, François H. & Lass, Norman J. 1977. Some factors affecting the listener's perception of reading rate in English and French. *Language and Speech* 20(3). 198–208.

Gut, Ulrike. 2012. Rhythm in L2 speech. In Gibbon, Dafydd & Hirst, Daniel & Campbell, Nick (eds.), *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem*. *Speech and Language Technology* 14/15. 105–114. Poznań.

't Hart, Johan & Collier, Rene & Cohen Antonie. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Hirst, Daniel & Di Cristo, Albert (eds.). 1998. *Intonation Systems. A survey of Twenty Languages*. Cambridge: Cambridge University Press.

Inden, Benjamin & Malisz, Zofia & Wagner, Petra, & Wachsmuth, Ipke. 2012. Rapid entrainment to spontaneous speech: A comparison of oscillator models. In Miyake, N. & Peebles, D. & Cooper, R. P. (eds.), *Proceedings of 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Jassem, Wiktor. 2003. IPA: *Polish. Journal of the International Phonetic Association* 33(1). 103–107.

Jassem, Wiktor & Krzyśko, Mirosław & Stolarski, Przemysław. 1981. Regression model of isochrony in speech signal, *IPPT PAN* 33. Warszawa.

Jassem, Wiktor & Hill, David R. & Witten, Ian H. 1984. Isochrony in English speech: its statistical validity and linguistic relevance. In Gibbon, Dafydd & Richter, Helmut (eds.), *Intonation, accent and rhythm. Studies in Discourse Phonology* 8. 203–225.

King, Simon & Portele, Thomas & Höfer, Florian. 1997. Speech synthesis using non-uniform units in the Verbmobil project. *Proceedings Eurospeech* 2. 569–572. Rhodes.

King, Simon & Black, Alan W. & Taylor, Paul & Caley, Richard & Clark, Rob. 2003. Edinburgh Speech Tools. System Documentation Edition 1.2, for 1.2.3 24th Jan 2003. (Retrieved from: http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0 on 27 April 2013).

Klatt, Dennis. H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59. 1208-1221.

Klatt, Dennis. H. 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 88(3). 737–793.

Klessa, Katarzyna & Szymański, Marcin & Breuer, S., & Demenko, Grażyna. 2007. Optimization of Polish segmental duration prediction with CART. In *Proceedings of 6th ISCA Workshop on Speech Synthesis (SSW-6)*. Vol. 1. Bonn.

Klessa, Katarzyna & Wagner, Agnieszka, Oleśkowicz-Popiel, Magdalena & Karpiński, Maciej. 2013. "Paralingua" – a new speech corpus for the studies of paralinguistic features. In Vargas-Sierra, Chelo (ed.), *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013). Procedia – Social and Behavioral Science.* Vol. 95, 48–58.

Unauthenticated
Download Date | 12/13/15 8:16 AM

Koreman, Jacques. 2006. Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America* 119. 582–596.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, Massachusetts–London: M.I.T. Press.

Lehiste, Ilse. 1977. Isochrony reconsidered. *Journal of Phonetics* 5.

Low, Ee Ling & Grabe, Esther & Nolan, Francis. 2001. Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43(4). 377–401.

Łobacz, Piotra. 1976a. Objective and subjective speech tempo in Polish. *Speech Analysis and Synthesis* 4. 173–186.

Łobacz, Piotra. 1976b. Speech rate and vowel formants. *Speech Analysis and Synthesis* 4. 187–218.

Möbius, Bernd & van Santen, Jan P. H. 1996. Modeling segmental duration in German text-to-speech synthesis. *Spoken Language, 1996. Proceedings of ICSLP.* Vol. 4, 2395–2398. Philadelphia, PA: IEEE.

Möbius, Bernd. 2001. Rare events and closed domains: two delicate concepts in speech synthesis. *4th ISCA ITRW on Speech Synthesis.* Perthshire.

Moers, Donata & Jauk, Igor & Möbius, Bernd & Wagner, Petra. 2010. Synthesizing Fast Speech by Implementing Multi-Phone Units in Unit Selection Speech Synthesis. In *Proceedings of 7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-7)*.

Moos, Anja, & Trouvain, Jürgen. 2007. Comprehension of Ultra-Fast Speech–Blind vs. 'Normally Hearing' Persons. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 677–680.

Olaszy, Gábor. 2002. Predicting Hungarian sound durations for continuous speech. *Acta Linguistica Hungarica* 49(3–4). 321–345.

O'Shaughnessy, Douglas. 1984. A multispeaker analysis of duration in read French paragraphs. *Journal of the Acoustical Society of America* 76(6). 1664–1672.

Pfitzinger, Hartmut R. 1996. Two approaches to speech rate estimation. In *Proceedings SST.* Vol. 96, 421–426.

Portele, Thomas & Sendlemeier, Walter & Hess, Wolfgang. 1990. A system for German speech synthesis based on demisyllables, diphones, and suffixes. In *ESCA Workshop on Speech Synthesis Autrans*, 161–164.

Richter, Lutosława. 1973. The duration of Polish vowels. *Speech Analysis and Synthesis* 3. 87–115. Warszawa.

Richter, Lutosława. 1974. Porównanie iloczasu samogłosek polskich wymówionych w logatomach oraz w wyrazach. *Biuletyn Polskiego Towarzystwa Fonetycznego* 32. 173–178.

Richter, Lutosława. 1987. Modelling of the rhythmic structure of utterances in Polish. *Studia Phonetica Posnaniensia* 1. 91–125.

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

Scott, Donia R. & Isard, S. D. & de Boysson-Bardies, Bénédicte. 1986. On the measurement of rhythmic irregularity: a reply to Benguerel. *Journal of Phonetics* 14. 327–330.

Siegler, Matthiew A. & Stern, Richard M. 1995. On the effects of speech rate in large vocabulary speech recognition systems. In *International Conference on Acoustics, Speech, and Signal Processing 1995. ICASSP-95*. Vol. 1, 612–615.

Syrdal, Ann K. & Bunnell, Timothy & Hertz, Susan R. & Mishra, Taniya & Spiegel, Murray & Bickley, Corine & Rekart, Deborah & Makashay, Matthew J. 2012. Text-To-Speech Intelligibility across Speech Rates. In *Proceedings of Interspeech*. Portland, Oregon.

Szymański, Marcin & Klessa, Katarzyna & Breuer, Stefan & Demenko, Grażyna. 2011. Optimization of unit selection speech synthesis. In *Proceedings of XVIIth International Congress of Phonetic Sciences*, 1930–1933. Hong Kong.

Treiblmaier, Horst & Filzmoser, Peter. 2009. *Benefits from using continuous rating scales in online survey research.* Technische Universitt Wien, Forschungsbericht SM-2009-4.

Vainio, Martti. 2001. *Artificial neural network based prosody models for Finnish text-to-speech synthesis.* Helsinki: University of Helsinki. (Doctoral dissertation.)

van Santen, Jan P. H. 1993. Quantitative modeling of segmental duration. In *Proceedings of the workshop on Human Language Technology*, 323–328. Association for Computational Linguistics.

Wagner, Petra & Windmann, Andreas. 2011. The shrinking effects on speech tempo perception. In *Proceedings of XVIIth International Congress of Phonetic Sciences*, 2082–2085. Hong Kong.

Zee, Eric. 2002. The effect of speech rate on the temporal organization of syllable production in cantonese. *Proceedings of Speech Prosody.* Aix-en-Provence.

Unauthenticated
Download Date | 12/13/15 8:16 AM