Beats for Individual Timing Variation

Eric Keller
IMM, University of Lausanne, Switzerland

Abstract. We subjectively experience humans to speak with a certain regularity – which creates perceived rhythm within speech – at the same time as we expect them to display variation, mostly for emphasis and to satisfy personal preferences. Synthesized speech that does not exhibit these perceptual qualities is often classified as "robotic" and "unnatural". The search for the objective bases of the perceived regularity in speech is old and has produced less than satisfactory results. In 1977, Ilse Lehiste, in an extensive review of the issue of isochrony (acoustic evidence for rhythmicity in speech) came to the conclusion that there were no direct acoustic correlates of rhythmicity [1]. This view, supported by a number of further studies, has formed the consensus for spontaneously produced speech since then. However, Robert Port and his colleagues have in recent years suggested that some parts of perceived regularity may actually be directly dependent on the suddenness and the relative strength of voice onsets (so-called “beats”). This hypothesis was examined here with respect to continuous speech by a series of analyses performed in two languages, and it was found that indeed, beats do provide a minor temporal organizational effect within the speech phrase, but that the effect is so minor that it is of no or only circumscribed value to such applications such as speech synthesis or speech recognition.

Keywords. Speech timing, rhythmicity, voice onsets, beats

Introduction

Human speech has both shared and individual components. The *shared* components of speech make it possible to transmit ideas between speakers, while the *individual* parts identify and distinguish those who are speaking. Scientific analysis of speech has in the past explained primarily the shared parts of language, those that encode systematic linguistic information and are thus held jointly by the various members of a language community. Till a few years ago, speech research had nearly “forgotten” about the individuality of language and speech.

But detailed analyses of speech have recently called much more attention to the individual aspects of speech. Individuals show considerable variation with respect to the verbal material that they use to meet similar communicative challenges (Campbell, this volume). Also, when using a particular expression, individuals employ a certain liberty in how they structure their material in their own individual manner. It may be useful to characterize individual variation in the following simplistic, but initially helpful manner: When information is new and when the communicative situation demands good understanding at first presentation, speakers tend to use hyper-articulated and prototypical forms of speech. On the other hand, when basic agreement between speakers is assumed, or if the communicative situation is more relaxed, or if the speaking party shows low concern about being well understood, hypo-articulated and greatly reduced and individually varied forms of speech tend to predominate.

If one were to record all human speech in all possible combinations of face-to-face, telephoned or televised speech, one would probably find that by far most oral communications are variations of the latter type. This poses considerable challenges to various types of speech analysis. For example one central issue in recent speech recognition work has been the following: “How can relevant information be extracted from strongly divergent forms of speech?” And a well-posed question in current speech synthesis work concerns the types and limits of individual variation that could or should be implemented to obtain realistic and individually coloured forms of speech. Also of interest is the question of whether there might be portions inside stretches of speech with less individual variation, and some others where there is more. If that is so, speech recognition work could initially be directed at the more stable sections of speech, and could return to the more varied portions in subsequent analysis steps.

In this article, we wish to address a subcomponent of these issues, the individual time structure of phrases, and explore the question of more or less stable points in the speech chain. It has been known for some time that phrasal time structure has both shared and individual components. This emerges from prediction experiments where the systematic (information-bearing) timing components of phrases are predicted by common statistical methods. In these experiments, the average timing of the various linguistic units (segments, syllables, minor/major phrases, etc.) is predicted from the components and structures of various grammatical and phonetic contexts (see summary view [2]); this process leads to a fairly adequate modelling of such phenomena such as utterance-final lengthening, word-initial lengthening, polysyllabic shortening, etc. [3]. However, depending on the speaker, the speech material and the predictive algorithm, model adequacy for this type of variation generally shows rates of between 0.7 and 0.9 out of a possible 1.0, leaving a good proportion of the overall variation still unaccounted for. A good part of this remaining variation is likely to be individual variation, while some other parts may be due to “noise” in measuring technique and data analysis.

The present study examines such non-systematic individual variation in the context of the effects of *beat*. “Beat” is the term given by Robert Port, Fred Cummings, Keichi Tajima and various other colleagues to rapid onsets of sonorous information in speech. In their studies, beats were shown to be an important temporal organizing concept for short, repeated sentences [4, 5, 6]. In an experimental setup known as “speech cycling”, subjects had to repeatedly produce short sentences (like “dig for a duck” or “take a pack of cards”) in a time frame set up by two external sounds. After listening to the synthesized sentences once without speaking, they had to repeat the sentences in time with subsequent stimulus sentences. Subjects had to temporally align the first main stress in the sentence (“dig”, “take”) with a first sound, which reoccurred every 1.5 seconds in the study reported in [4]. The second main stress (as in “duck” or “cards”) had to coincide with the second sound. If the second sound was randomly chosen in the range of about one half of the repetition cycle, the second main stress peaked fairly exactly around the mid-point (50%) of the repetition cycle, and did not show the expected random distributions around the middle area of the repetition cycle. Randomly chosen shorter durations between the acoustic signals tended to congregate fairly precisely at about one third of the repetition cycle, while random longer durations peaked at about two thirds of the repetition cycle (Figure 4).

In an extensive theoretical article of these effects, Port motivates this and similar results in chaos terms [6]. He relates these findings to “coordinative gestures”, a theoretical formulation about motor actions that has been applied to speech for a

number of years by scientists from Haskins Laboratory (e.g., Kelso & Tuller [7]). It has been shown in a variety of motor tasks that coordinative cyclic actions tend to subdivide at half-way points, and often as well at other low harmonic frequencies, such as at one third or one quarter of the entire movement. Port notes that with respect to a large number of activities, humans in the presence of other humans show spontaneous emergent coordination, and he argues that this is the effect that manifested itself in his speech cycling data.

One of the curious aspects of this type of coordination is that no physical link is necessary for the production of harmonic subdivisions. For example, Schmidt, Carrillo & Turkey (1990) showed that when two people sit on the edge of a table and swing one of their legs, they found it easiest to swing in phase, somewhat less easy to swing in alternating phase, and most difficult to swing out of phase [8]. This breaks up a perceptually coordinated phase into a harmonic at one half of the phase. In similar fashion, Port argues, speakers will find it easier to make vowel onsets occur at low harmonics of the repetition cycle, such as at halves or thirds of the cycle. Stated in chaos terms, this says that the natural harmonics in the phase constitute "attractors" for the speech behaviour. These are sites where a certain set of behaviour is more likely than in-between "attractors".

Beats can be calculated quite easily (see below), so it was inviting to examine their effect on timing in speech phrases that were more typical of continuous speech than Port's simple, memorized and repeated phrases. We had some carefully-recorded and transcribed corpora of read French and English texts in our laboratory, so we posed the following question: "Given the documented time structuring effect of beats within the sentence, do beats have an effect on individual temporal variation in speech read aloud from text?" By examining the effects of beat on various portions of speech produced in similar phonetic manner by the various speakers, we could examine the potential organizing effects of beats on speech in the two languages.

We did not have any firm expectations about the results of this experiment. On the one hand we estimated that beats might indeed reduce the variation between the speakers, at least to some degree, since there should be a natural tendency to form coordinate structures at common beat locations. Given the results from Port's laboratory, we expected that the locations of strongly beat-marked sound transitions would show somewhat less inter- and intra-subject variation than un-marked or weakly marked sound transitions.

On the other hand, it was not certain that our experimental conditions would permit beats to fully develop their coordinating effects on speech. Continuous speech, spoken freely or read aloud as in these experiments, is quite different from the memorized and repeated utterance of a just a few words or syllables. Phrases – however we wish to define them – are of different duration and syllabic structure, and there may not be enough similarity between phrases for a phase-locked pattern to evolve. The two corpora were examined successively during the summers of 2005 (French) and 2006 (English) to clarify these issues and to probe for evidence of the presence of beats. We shall present the method and the results separately for the two corpora.

1. Study 1: French Inter-Subject Experiment

1.1. Method

Task and Segmentation. Nine student and faculty members (2 F, 7 M) of the University of Lausanne recorded the same reading-aloud task of a paragraph of 218 words (11 sentences) in quiet surroundings. The recordings were performed at 96 kHz mono, were normalized against the peak volume in the file, and downsampled to 16 kHz using anti-aliasing software (Pristine Sounds 2000 Pro). Sentences varied considerably in duration and in underlying phrase and syllable structure (Appendix 1). The same text was also read by our French speech synthesis system ("subject 10") and its output was subjected to the same types of analysis as were the human sentences.

As a reference object comparable to Port's and colleagues' repetition cycle we elected the phrase, defined as an acoustically identified, prosodically unified group of temporally and grammatically and semantically coherent linguistic materials (see Appendix 2). All temporal positions for beats and segment transitions were calculated in terms of proportions of the containing phrase. In the material in this experiment, it was appropriate to consider that an on- or offset of a phrase had occurred if any of these three conditions was encountered: (a) punctuation, such as a period, interrogation mark, comma or suspension mark in the input text, (b) the coincidence of a pause of 50-150 ms and an intonational reset, or (c) any pause in excess of 150 ms. When unvoiced plosives initiated the phrase, the phrase was considered to start with the portion of the sound considered audible to both speaker or listener, which was generally the onset of the burst (and not the silent period preceding the burst). Likewise, the end of phrases with slowly waning amplitude was judged to have occurred at a point where the signal-to-noise ratio no longer permitted the audible distinction of the speech signal.

Recordings were manually segmented by experienced segmenters for both segmental and phrase duration using the Praat software[1], and were spot-checked by two other segmenters. After training [9], segmenters agreed on segmentation placement within an average of about 5-10 ms.

Beat detection. An adapted version of the method used by Port and colleagues for identifying beats was applied. It consists of the following steps:

Low-frequency information in speech signal: (a) Filter low-frequency information: Obtain the sound file and bandpass it with cutoff frequencies of 800 and 200 Hz[2]. This filtering step can be performed very satisfactorily with Praat's pass Hann band filter. (b) Obtain intensity: Take the intensity curve of the filtered signal (with Praat: 100 ms limit).

Smoothing and peak location: (a) Take a spline of the intensity curve. The spline's tension is initially adjusted in such a fashion that peaks can be measured on the basis of *continuous* vowel onset intensity rises (no double-peaks measured from double-hump rises). In our version of the spline algorithm, the tension setting was 0.001 for French (0.0001 for emphatic English speech [second experiment]). Once set for a given speech

[1] *Praat: Doing Phonetics by Computer*, phonetics analysis software, www.praat.org.

[2] We implemented the limits of 200-800 Hz reported on by Port in 2003 [6], rather than the 300-2000 Hz limits used by Cummings and Port in 1998 [4].

condition, the spline tension was not changed any more (1 setting for French, 1 for English). (b) Locate peaks in the derivative of the splined intensity curve (in ms) and identify potential segment transitions.

Data cleaning: Only beats situated within 25 ms of a segment transitions were scored. The 25 ms limit was arrived at by examining the error fall-off of beats situated close to segment transitions. Also several beats often "competed" for a segment transition. If two or more beats were identified for the same segment transition, the stronger beat was taken.

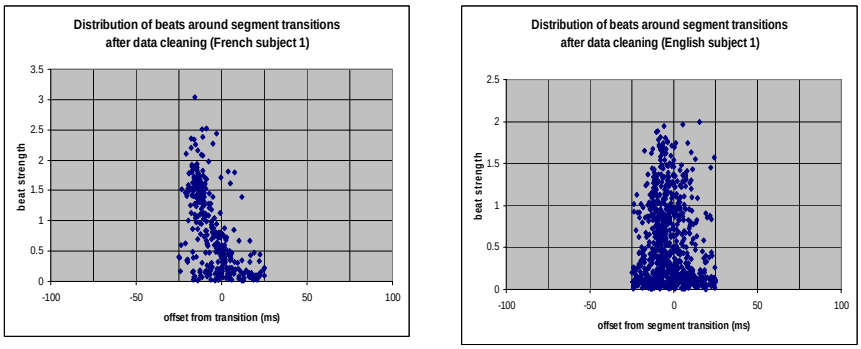
Strong and weak beats: Beats vary considerably, and it was argued that the "stronger" beats (those with greater derivative amplitude, i.e., those with more rapid and more definite onsets of low-frequency information) might have more temporal structuring effect than the "weaker" beats. For this reason, beats were initially divided in such a manner that about half the beats lay above average strength ("strong beats") and half below average strength ("weak beats"). Because of the non-linear distribution of beat strengths, average strength lies considerably below the halfway point of the distance between minimum and maximum amplitude.

The Hypothesis: A time structuring hypothesis was formulated on the basis of Port's hypothesis for beats. If a number of speakers read aloud the same paragraph, we expected under a generalized hypothesis for beats that variation in timing should be lower at strongly beat-marked transitions than at other places in the speech chain.

The Test: To test this hypothesis, beat-marked segment transitions situated between >0.0 and <1.0 of the containing phrase were compared for the 9 subjects. If two subjects used the *same phrase* in their reading, and if they used the same *beat-marked segment clusters* (the same *triphones*), the segment transition positions were included in the comparison. It is to be noted that points 0.0 and 1.0 were excluded from the test. Many phrases begin with some form of vocalic onset, and a coincidence on point 0.0 is therefore not meaningful for the test. Also, all phrases end in point 1.0, and this point is therefore excluded as well. Only points between 0.0 and 1.0 for which different speakers could indeed show different locations were compared in the test.

1.2. Results

Some initial observations. As Port and colleagues indicate, beat peaks cluster somewhat to the left of transitions towards what are generally voiced segments. Most peaks (about ¾ of sound transitions) marked vowel or semi-vowel onsets. The remaining ¼ generally marked the onsets of voiced consonants (/n, m, l, r, d/). Beats tend to be anticipatory. Also, subjects show more coherent behaviour with respect to stronger than to weaker beats (figures 1 and 2). In addition, peak clusters in some subjects show an incline to the left (more anticipation for the stronger peaks). Some relationship with fluidity of speech was observed in the 13 speakers seen in the two corpora: the more fluid a speaker appeared to the ear, the more incline was shown. On the basis of these initial observations, it was considered important to systematically distinguish results obtained from stronger and weaker beats in this study. Furthermore, inter-subject variation for beat-marked sound transitions was examined for all inter-subject comparisons (Figure 3). Absolute inter-subject differences between beat-marked phrase points averaged between about 2% and 4% of phrase, but notable variation was observed.



Figures 1 and 2. Beats measured for individual speakers in the two corpora. The zero-point represents the transition point between two speech segments (typically the onset of a vowel or a voiced consonant).

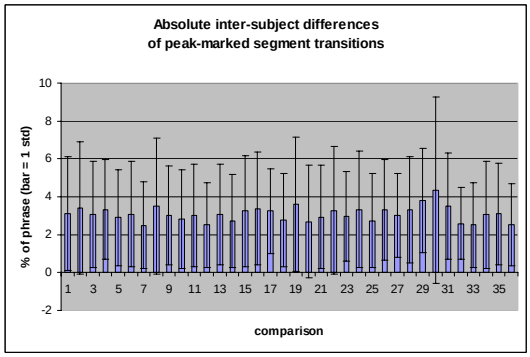


Figure 3. 36 absolute inter-subject differences obtained by comparing transitions for all 9 voices [9*8/2=36 pairs]) at positions > 0% and <100% in phrase as marked by beat peaks. In these positions within the phrase, average agreement between subjects ranged between about 2% and 4%, but the one-standard deviations were generally well above 1%, suggesting considerable variation within the data.

Onset distribution. To examine the relationship between beats and temporal position, the distribution of beats within the phrase was examined[3]. Beats were divided into approximately equal numbers of strong and weak amplitude exemplars ("strong" and "weak" peaks). In this analysis, only vowel-onset information was examined (no voiced consonant onsets). Combined data from the nine human speakers were placed into the same positional bins that had been created for the first figure by Cummings and Port, 1998 (as reported by Port in 2003) (Figures 4 and 5). The number of tokens was similar between the two studies (Our study: N_{strong} = 703, N_{weak} = 671, N_{total} = 1374, Cummings & Port N = 484).

Minor peaking was found at 50% of the phrase for both strong and weak vowel onsets (Figure 5). However, other peaks of similar sizes were found in other parts of the phrase which were not evident harmonic positions. We thus conclude that for our

[3] These results were first reported in [10].

French-language readers, “harmonic positional attraction” was barely a factor in the determination of sites where beats were recorded in continuous (read) speech.

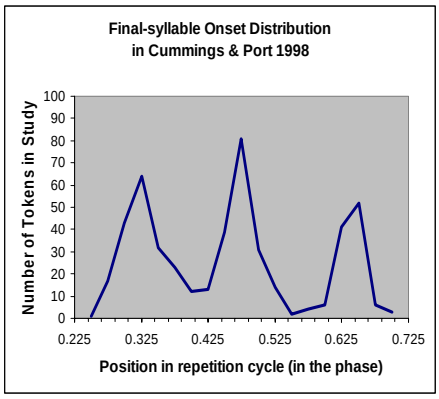


Figure 4. Data recreated from Figure 1 in Port (2003) [6]. Final-syllable onset distribution in terms of a phase angle in the repetition cycle created by repeating a 4-syllable phrase like “Dig for a duck” in time with two metronome tones. The first tone marked the onset of the phrase and the second tone varied randomly from 0.20 to 0.80 of the phrase’s duration.

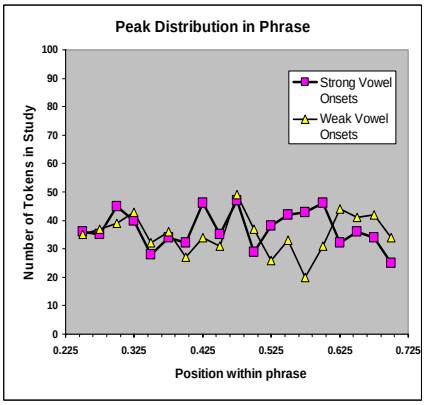


Figure 5. Data from our experiment 1. Although there is a minor peak close to the 50% mark in the phase for both strong and weak vowel onsets, the hypothesis that beats have a strong “attracting force” for temporal structures in continuous speech (reading aloud) did not receive strong support.

Degree of variation according to strength of beat. As indicated above (Figures 1 and 2), some differences were observed with respect to the temporal position of strong and weak beats. In particular, strong beats showed a more coherent temporal position than weak beats with respect to the onset of the succeeding sound segment. If beats exert an organizational “attractor effect” within the phrase, it was possible that it could be more reliably demonstrated for stronger than for weaker beats.

In this analysis, similarity in the 36 inter-subject comparisons was examined for different degrees of beat strength. In Figure 6, the upper line shows the effects of beats below average beat strength (“weak beats), and the lower line shows those for beats above average beat strength (“strong beats”). Average beat strength was found to be 3.266. The degree of cut-off was varied along the x-axis in tenths of the average strength. At the leftmost position, data were included from all beats. In this position, only minor differences in inter-subject agreement on temporal position in the phrase were noted. But as we excluded more data and compared stronger and stronger with weaker and weaker beats, substantially more positional stability emerged for temporal positions marked by strong beats. Differences reached the statistical significance level of $p < .05$ at condition 7, and the greatest difference (condition 8) was different at $p = 0.0175$ (Paired t-Test, Mean of Paired Differences = 0.462, t = 2.49 w/36 df). At the extreme position at the right, it was found that average temporal differences between subjects were at 2.9% for strong beats, and at 3.4% for weak beats.

This average difference of one half a percent must be put in perspective. As seen in Figure 3, there was considerable variation in the agreements on beat-marked positions of segment transitions (average s.d. well over 1%). Even though the half percent average difference is statistically significant and is of great theoretical importance

when evaluating the work by Port and colleagues, the practical value of using beats to identify high-agreement areas in speech remains circumscribed, because the effect is generally lost in the high signal-to-noise ratio established by the beat effect vs. average difference between the individuals. We conclude from this examination that although segment transitions with strong transitions between unvoiced and voiced sounds tended to be in somewhat more similar position in the phrase than those with weaker transitions, the effect was minor and regrettably not of notable practical value for applied purposes like speech recognition or speech synthesis.

This conclusion is also supported by an examination of the relationship of beats and position in the speech synthesis rendition of the passage (Figure 7). Strong beats did not show the same temporal organizing effect for synthesized speech as it had in human speakers. There was no decrease in variability with increased beat (points 12, 13, 14), and the human-synthesis variability was greater (3.4 - 4.6%). This suggests once more that the "beat attractor" phenomenon described here is a real human phenomenon, even though its practical effect in continuous speech is probably minor.

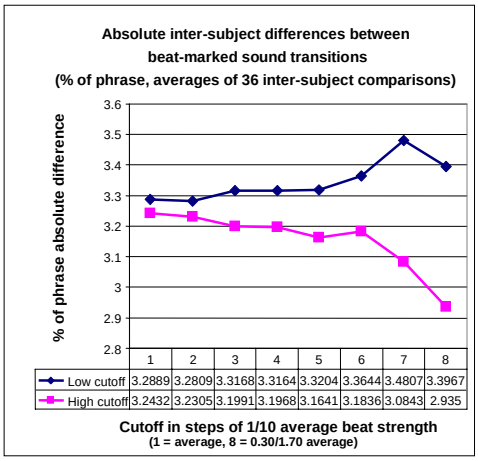


Figure 6. Comparisons between human speakers, study 1, French. On the average for these human speakers, strongly beat-marked positions show 0.5% more agreement than weakly beat-marked positions. The pattern is not evident for the speech synthesis system shown in Figure 7. In the two graphs, the data cutoffs move from average cutoff (left) to high/low cutoffs (right). Specifically, at the *left extreme*, inter-subject variation is calculated for all parts of data above and below the beat average, and at the *right extreme*, variation is calculated only for very high and very low beats, and the intermediate beat data is eliminated. At the extreme right position, average positional differences between subjects were 2.9% for strong beats, and at 3.4% for weak beats.

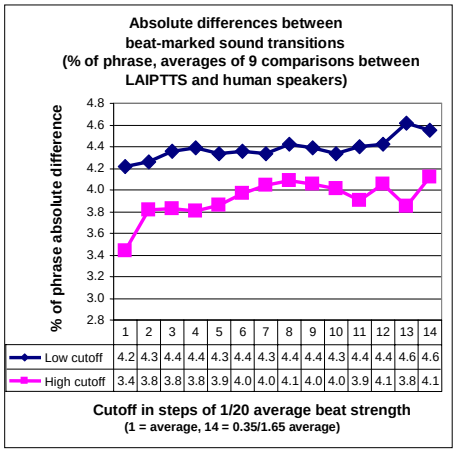


Figure 7. Comparisons between human speakers and the speech synthesis system, study 1, French. In contrast to human speakers (Figure 6), the output of the speech synthesis system did not show any reduction in variance at stronger beat locations.

1.3. Conclusion for Study 1

In this experiment with French speech read aloud, we were able to demonstrate that strong transitions from unvoiced to voiced segments imposed some weak temporal organization effects. At points of unvoiced stop-vowel transitions for example, subjects tended to agree more on the exact placement for initiating the post-stop vowel than in other positions in the phrase. Although theoretically interesting, plausible, and demonstrated to be statistically significant in a fine-combed analysis, the practical effects were not excessive, since the reduction in subject-to-subject variation was not very great. It is interesting to note that the increase in temporal similarity with greater beat strength was not found in the speech-synthesized data.

This first study left some questions open, notably (a) whether the temporal organization effect would be greater if we examined material from another language, particularly a language with strong stress patterns, and (b) whether we would obtain compatible effects in a second and theoretically independent verification of the coordinative gestures hypothesis. We thus decided to examine another database which permits these verifications, a database for *English* where in addition to *inter*-subject verifications we also had *intra*-subject comparisons. This study will be presented next.

2. Study 2: English Intra- and Inter-Subject Experiment

2.1. Method

Corpus and segmentation. A speech database was examined for English with a well-known speaker, Winston Churchill, recorded in (a) a prepared war speech presented under emotionally charged circumstances to free country representatives and Dominion high commissioners in London on June 12th, 1941 (the “St-James Palace

speech"), (b) a post-war BBC re-recording of the same speech performed under less straining circumstances (probably in 1947). Furthermore (c), we made a re-recording in 2003 of the same speech with another UK-English speaker (initials PW), who is currently of the same general age group as Winston Churchill was in 1941, and who speaks the same dialect variant (British Received Speech). The analyses of the historical recordings were performed with the permission of the recording firm.

The historical recordings presented some minor difficulties in the analysis, due to elevated noise levels (e.g., judgment of initial /h/), echoes in the original environments (e.g., some minor difficulties of establishing phrase duration; where does a terminal /n/ end and where does the echo begin?), and due to microphone reverberation. Also, some slight slurring of speech was noted in some of the BBC re-recordings, suggesting a few local effects of inebriation. Churchill's liberal use of alcoholic beverages is well-attested, particularly for the post-war period [12]. Manual segmentation was performed by an experienced phonetician trained in the LAIP segmentation procedure [9] to mark for segments and phrases in the three recordings. A complete segmental verification was performed by the author.

Calculation of voice onset peaks. Peaks more than 25 ms from the peak-marked transition were removed prior to automatic beat analysis. Comparable phrases were identified by automatic comparisons of phonetic transcriptions in the three re-recordings, and beats were identified for each segment-to-segment transition by the procedure indicated in Study 1. If there were several candidate beats relating to the same transition, the stronger beat was chosen. The speech contained 980 comparable tokens of beat-marked segment transitions.

2.2. Results

Initial observations. As in study 1, peaks clustered around segment transitions, but no incline was observed in the data.

Inter-subject and intra-subject variation. For both analyses, similar patterns of positional precision were identified as in Study 1. For both inter- and intra-subject variation, agreement on position in phrase increased as stronger beat exemplars were chosen. It was found that (a) the *inter-subject* similarity between Churchill and PW was greater than the average French inter-subject similarity (1.5% average difference at point 18 for Churchill-PW vs. 2.9% for the comparable French analysis) and (b) that curiously, Churchill shows **less** *intra-subject* (3.5%) than *inter-subject* (1.5%) similarity (Figures 7 and 8). In other words, the difference in *style of speech* between the two Winston Churchill recordings introduced greater positional differences in the phrase than the differences between the *two speakers*. Similar overall variability in the data was shown as in Study 1.

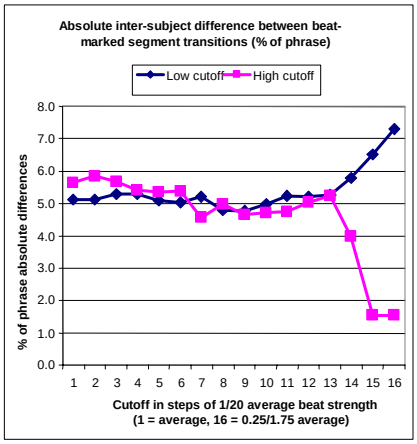


Figure 8. Study 2, English, ***inter***-subject differences between positions of beat-marked segment transitions in the phrase. The tendencies are similar to those found in the French study, but are more pronounced. As in Figure 6, the data cutoffs move from average (left) to high/low (right).

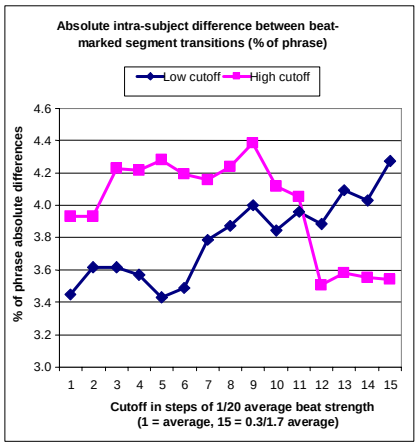


Figure 9. Study 2, English, ***intra***-subject differences. Again, stronger beats found with high cutoffs are related to greater positional agreements within the phrase.

Relationships with non-phonetic parameters. A variety of further examinations were undertaken to examine possible relationships with non-phonetic parameters (data reported in [11]). For example, earlier speech productions in the phrase were expected in the psychologically stressed speech condition, while delayed productions in conditions under less psychological stress. The relative timing of peak-marked transitions was thus compared between the original (stressed) and the re-reading (less stressed) of the same speech. However, no significant pattern was observed. Also, no relationship was found between specific phonetic transitions and anterior and posterior beat-marked position. Finally, beats were found to relate to nothing else than the type

of phonetic transition between unvoiced and voiced status; for example, they showed no relationship with grammatical class.

3. Conclusion

Support was found in corpora from two languages with different accentual patterns for the temporal organizing effects of beats, in that strong beats were correlated with reduced temporal inter- and intra-subject variation in the phrase. The variation effects noted for French, where accents are difficult to define and to measure, were also found in the English corpus, where there is widespread agreement on the placement of accentual ("stress") patterns. Furthermore, this effect was absent in synthesized speech, which suggests in turn that the absence of beat patterns may to some degree be responsible for the "robotic" impressions that speech synthesis often evokes in human listeners.

A number of questions remain open. First of all, it is possible that the elusive "rhythmic effects" frequently perceived in continuous language may indeed be related to the beat effects that were found by Port and colleagues and that were again demonstrated in this study. However, the experimental demonstration that this is indeed so has yet to be made. For example, a listener experiment could be imagined in which perceived rhythmicity is related to degree of inter- / intra-subject variation and the presence of beat phenomena.

Also, this study raises some interesting new questions concerning the relationship between beat and accent. Although accentual patterns are quite different for French and English, the similar patterns of beat dependency shown here raise the question of the exact relationship between beats and accent marking in these and other languages. A question of further interest is the issue of how beat effects may be related to known systematic timing effects such as utterance-final lengthening, word-initial lengthening, polysyllabic shortening, etc.

Finally, the differences between natural and synthetic speech could be more systematically investigated. In speech synthesis, durations are obtained with a multidimensional average calculation of elements determing segment and syllable duration, word-central reduction, phrase-final lengthening, etc [2]. Since increased temporal organizing effects were documented here for beat positions, individual variation could be reduced in beat positions (i.e., one would apply a strict multidimensional average calculation there), while in non-beat positions, individual variation would be given more latitude (i.e., average computations for those places could be varied more freely). One would have to investigate with detailed data sets whether that approach improves the temporal calculations for speech synthesis applications.

4. Appendix 1: The text for Study 1

« Pour fêter ses vingt-cinq ans, le Festival d'automne invitait, en 1996, tous ceux qui, au cours des années, façonnèrent sa légende, dans la musique, le théâtre, la danse et les arts, à l'initiative de Michel Guy. L'édition 1997 s'inscrit dans une autre logique, en forme de questions : où trouver la nouveauté ? Où puiser, à l'approche de cette fin du siècle, de quoi nourrir la réflexion et l'imaginaire ? Comment rendre compte des mouvements qui traversent le monde improbable

d'aujourd'hui ? En allant là où le soleil se lève : au Japon. Sous la direction d'Alain Crombecque, le Festival d'automne consacre une part majeure de son programme à ce pays à double face, où la splendeur du kabuki, du nô et du bunraku, polie par des siècles de tradition, côtoie les quêtes formelles menées au théâtre. Autre invitée de marque : l'Egypte, la folie de ses nuits, ses chants millénaires, les transes soufies et les ballades amoureuses du delta du Nil. Bien sûr, il y a des retrouvailles dans le programme du Festival. Il y a aussi des "curiosités" - indispensables rendez-vous insolites proposés cette année par Jérôme Nicolin ou Christian Boltanski. Il y a enfin, qui donne son sens à l'ensemble, le désir de trouver en l'art une "voûte de lumière" sous laquelle poser son regard. »

5. Appendix 2: Definition of a phrase

Phrases are difficult to define, and yet the precision, validity and reproducibility of the current experiments depend on the use of precise definitions. Here are the guidelines used in the current set of experiments.

Definition. A phrase is an acoustically identified, prosodically unified group of temporally and grammatically and semantically coherent linguistic materials.

Explanation 1. "Acoustically identified, prosodically unified, temporally coherent": the phrase is defined as an uninterrupted acoustic waveform incorporating a unified and uninterrupted temporal and intonational structure.

Explanation 2. "Grammatically and semantically coherent": a single and coherent grammatical and semantic structure is associated with the acoustic structure. Speech interruptions ("uh", etc.) and speech corrections are considered to break up the phrase, resulting in two or more "phrases" for this experiment. Also, acoustic material without evident grammatical and/or semantic structure was eliminated from the experiment.

Questions on the definition of phrase duration. (a) Do ingressive respiration, or clearing the throat prior to the start of phrase form part of the phrase? Decision: no, this is considered part of the planning phase for the phrase. (b) Does the pre-burst part of unvoiced plosives form part of the phrase? Decision: no, we do not generally know how long this pre-burst lasts. Only audible events are part of a commonly defined phrase for both listener and speaker. The phrase is thus considered to start with the burst. (c) Do pauses end a phrase? We distinguished three types of pauses: (i) Short and not interrupting intonational flow; this type of pause does not end the phrase. (ii) Longer and interrupting intonational flow; this type of pause ends the phrase. (iv) Longer pauses. This type of pause ends the phrase. What are the cutoffs for these pauses? That depends on the speaker, but we used 50 ms for (i/ii) and 150 ms for (ii/iii) in the French project.

6. Acknowledgements

We wish to thank Dr. G. Peter Winnington (UNIL, English Department) for re-recording the Churchill passages, Prof. Jürg Schwyter (UNIL, English Department) for help with the Churchill project, and Dr. Brigitte Zellner Keller (UNIL, Psychology) for help with the design, the segmentation and the initial analysis of the French corpus. The transcriptions for the French and Churchill corpora were performed by Lukas Wiget, and were verified by Brigitte Zellner Keller for the French corpus and by the author for the English corpus. We also wish to thank the students and staff at the University of Lausanne who participated in the recordings for the French Corpus. The Winston Churchill recordings were used for the present scientific purposes with the written permission of ARGO, the producers of the Winston Churchill CDs. Work reported here was in part financed by the Swiss Staatssekretariat für Bildung und Forschung SBF under its programme in support of COST projects (COST 277).

7. References

[1] Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5,* 253-263.

[2] Keller, E. & Zellner Keller, B. (2003). How much prosody can you learn from twenty utterances? Linguistik online. 17, 5/03, 57-78. http://www.linguistik-online.de/17_03/kellerZellner.html, ISSN 1615-3014.

[3] White, L. (2002). *English Speech Timing: A Domain and Locus Approach.* Ph.D. Thesis, The University of Edinburgh.

[4] Cummings, F., & Port, R. (1998). Rhythmic constraints on speech timing. *Journal of Phonetics, 26,* 145-171. http://www.asel.udel.edu/icslp/cdrom/vol4/437/a437.pdf

[5] Tajima, K., & Port, R. (2003). Speech rhythm in English and Japanese. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 317-334). Cambridge, UK: Cambridge University Press.

[6] Port, R.F. (2003). Meter and speech. *Journal of Phonetics, 31,* 599-611.

[7] Tuller, B., & Kelso, S. (1990). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance VIII* (pp.429–451). London: Academic Press.

[8] Schmidt, R., Carello, C., & Turvey, M.T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 227-247.

[9] Schwab, S., Keller, E., Zellner, B., Connan, P.-Y., & Siebenhaar, B. (1998). *Conventions de segmentation pour la construction de diphones.* Rapport interne, laboratoire LAIP, Université de Lausanne.

[10] Keller, E. (2005). A phonetician's view of signal generation for speech synthesis. *Studientexte zur Sprachkommunikation, 36.*

[11] Keller, E., *Individual speech rhythm variation within the plosive structure of speech,* Part 1: The beat structure for speech timing, Part 2: Verification of Hypotheses. Nato Advanced Study Institute Advanced Study Institute (Asi) - Summer School " E. R. Caianiello" on The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue, September 2 -12, 2006 - Vietri sul Mare, Italy. Available at http://homepage.sunrise.ch/mysunrise/ekeller00/Kellerdoc.html.

[12] Owen, D. (1996), Diseased, demented, depressed: Serious illness in heads of State, *Quarterly Journal of Medicine, 96*. 325-336. http://qjmed.oxfordjournals.org/cgi/content/full/96/5/325.