A Phonetician's View of Signal Generation for Speech Synthesis

Eric Keller
LAIP, IMM – Lettres, University of Lausanne, 1015 Lausanne, Switzerland

Abstract. By far most speech science and speech technology is inspired by the premise that every aspect of speech is at least in theory totally predictable. This premise may well be wrong. Totally predictable speech, as found in good current speech synthesis systems, is easily recognized as non-human in origin. A more accurate view may well be a "chaotic" vision of speech which postulates various ranges of probability of an event. In this view, prosodic and phonetic parameters for "attractor sites" would be close to their statistically predicted values, while further removed from such sites, the parameters could be more variable. We have new evidence to support this view. Speakers place vowel onsets for "strong" vowels at highly predictable places in a phrase, but show much more variation with respect to the placement of "weak" vowels. Strength or weakness is determined automatically from the acoustic signal, and strong vowel onsets are likely to correlate with locations of both motoric and perceptual saliency. From this perspective, the speech motor control system's responsibility would be to closely approximate expected articulatory events at attractor sites, and to exert lesser control over articulatory events in-between attractor sites. The perceptual system would be preferentially oriented towards events at attractor sites. Between attractor sites, events would be attended to less strongly. Implementation of "motivated variation" for weak vowel onsets is likely to lead to more human-like speech synthesis.

1. A Personal View: Speech Synthesis is Still not Human Enough

I readily accepted the organizers' kind invitation to speak to this audience, since it is a welcome opportunity to present my view of a few issues in speech technology. I didn't need to look very far to find a topic; in fact, I looked no further than the considerable ambiguity I feel whenever I listen to synthesized speech, other people's as well as our own. I feel this way about *any* synthesized speech, no matter what its form of production, whether it is pieced-together pre-recorded speech, or whether it is speech synthesized from diphones or from smaller pieces of text. While I have great respect and admiration for work that is increasingly well-done and that deals well with the technical difficulties, I inevitably I come to the conclusion that this recording will still not "pass" (be accepted as) as human speech.

For example one recent morning, I found myself chuckling along with some teenagers as they were mimicking the predictable intonations and the hacked pauses of the train announcements in Switzerland: "Voie 2, le train pour Nyon || Genève || Genève–Aéroport || entre en gare..." There is no pretence here. Anyone can hear that this is not a live human announcer. But wouldn't it be nice if it *did* sound like human speech?

The case is paradoxically more annoying in the case of a US-English example from a major European provider of multilingual speech synthesis that uses AT&T's Natural Voice, reputedly the best speech synthesis for English today[1]. Among professionals of speech synthesis, there is wide agreement that the sound quality is technically excellent. Still, if one listens to this voice for any length of time, one easily tires of the repetitiousness of its sweet, mellow voice. The intonations are always the same, and the voice quality is always just as syrupy. This is annoying because with a first-class professional speech synthesis system, there is at least some hope that it could pass as human speech.

With so much dissatisfaction, could I do better? Is it technically possible to reach the level where somewhat more extended synthesized passages of speech can pass as human speech? In

[1] See the US-English example found at http://www.loquendo.com/en/demos/demo_tts.htm.

Lausanne, we've built systems for French, German, and classical Latin, and it was tempting to find out. I thus ran a short pilot project. I recorded as best as I could a short publicity spot with just five sentences[2] and presented the result to some native speakers of French. I used these three components:

- LAIPTTS-F, our 1990's vintage French speech synthesis system
- Mimic 3.1, a low-cost translator program that derives lip and head movements from the acoustics of a speech recording
- Poser 6.0, currently the best low-cost facial animation system

Please take a minute to judge the result yourself. It helps if you speak French, but it's not required. Ask the following questions: Is the speech rhythm natural? Are the head movements natural? Are the facial expressions human-like? *And:* what keeps us from accepting this as human-generated speech? Do you think this fellow could pass on television?

Most native speakers I consulted had mixed reactions to this passage. Most were reasonably satisfied with the technical quality of the speech and the video, but they all agreed that it could still not pass as human speech. Here are some of the comments:

- The speech is too regular, too artificially perfect.
- The head and eyebrow movements are too regular and too predictable.
- The eyes show no movements.

The comments fuelled suspicions that the dissatisfaction may well be grounded in the *excessive regularity* of the synthesized speech, especially with respect to speech rhythm. Human speech shows contractions and expansions, accelerations and decelerations. This virtual fellow's speech rhythm is fine and regular; phonemes all have the expected length and the phrase lengths all fall nicely into place, just as predicted by the statistical model of our synthesizer[3]. In fact, everything is just too perfect. What our speech synthesis may well need to acquire is *irregularity*.

2. Regularity and Irregularity in Speech Rhythm

Of course, one cannot simply increase the variability of phonetic or prosodic parameters by some random function. Beyond a certain point, that would simply decrease the intelligibility of the synthetic speech[4]. Instead, we must find out how humans vary speech timing, and then attempt to implement similar schemes in synthesis. Where in speech can we liberally contract and expand, where can we accelerate, decelerate, or even stop unexpectedly? And on the other hand, where can we *not* vary the rhythm, where must utterances be produced just at the right moment, so as to be perfectly well understood the first time?

In recent years, a set of experiments performed by Robert Port and his colleagues has generated considerable interest [3, 12], and these experiments that may well be of relevance here. They suggest that at least in regular, repeated speech, vowel onsets tend to drift towards specific places in the sentence, which perceptually lends them a particular "beat". This "beat" in turn aids in localizing and perceiving the message correctly by focusing the listener's attention to these sites.

It is important to examine the concepts developed by Port and colleagues a bit closer. Both concepts and empirical evidence are of considerable interest, since they provide some of the strongest evidence that rhythmicity may in fact be directly correlated with measurable events in the acoustic signal. While phoneticians have been saying for well over half a century

[2] http://mypage.bluewin.ch/ekeller00/trabajos/trabajos.html.
[3] A linear prediction model comparable to one presented in [4].
[4] See for example the unpublished experiment mentioned in [13], p. 9. Informal experimentation with our own synthesizer provided similarly discouraging results.

that speech is rhythmic at least to some degree, few have been able to adduce any systematic psycholinguistic or acoustic evidence in support of the notion. In fact in 1977, Lehiste[6], 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, and that the common sense of regularly repeating patterns of speech is probably a perceptual construct based on a variety of acoustic, grammatical and semantic markers. This view has in many different forms formed the general consensus on the issue since then (for some recent data, see [7]). So who is right, Port or Lehiste? Let us begin by recalling exactly what experiments by Port and colleagues show.

3. Experiments by Port and Colleagues

In an experiment published by Cummings and Port in 1998 [3], subjects were asked to repeat 4-syllable phrases like "dig for a duck" in rhythm with two metronome beats, where the occurrence of the first beat was stable and that of the second beat was variable. Subjects were asked to make the first syllable coincide with the first, stable beat, such as "*dig* for a duck, *dig* for a duck, *dig* for a duck…" This set up a repetition cycle for the phrase. But at the same time, subjects also had to make the last, stressed syllable coincide with the second beat which was placed randomly in the time between 20% and 80% of the repetition cycle. So sometimes, the phrase had to be squeezed considerably, and sometimes, it was stretched out quite a bit.

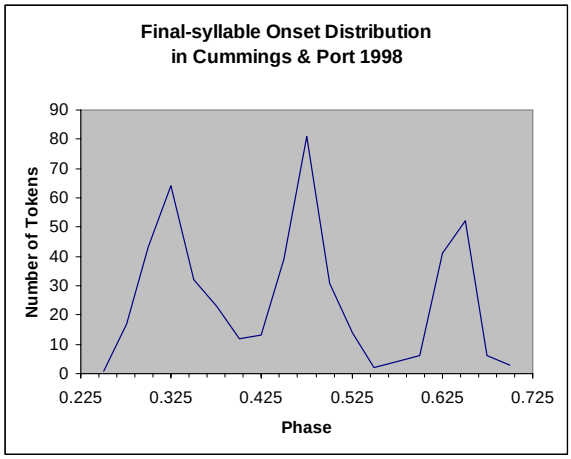

Figure 1. Final-syllable onset distribution in terms of a phase angle in the repetition cycle set up by repeating a 4-syllable phrase like *Dig for a duck* in time with two metronome tones. The first tone marked the onset of the phrase and the second tone varied randomly from 0.20 to 0.80 of the phrase's duration. (Recreated from Fig. 1 in [10]).

There were two notable results. First it was found that subjects tended to have the *vowel onset* of the last stressed syllable coincide with the metronome beat. This was interesting, because there is a considerable literature that suggests that subjects tend to pay particular attention to this area of the syllable, an area known as the "p-centre" (for *perceptual centre*, e.g., [8, 9, 15]), which is the place where a syllable is perceived to occur. It is also in accord with previous studies that found that speakers tend to have vowel onsets co-occur with other events. Allen (1972, 1975, cited in [10]) for example found that English speakers will align vowel onsets of stressed words with their own finger taps. Vowel onsets, particularly of stressed syllables, thus have a special status in speech.

The second, fairly surprising finding was the distribution of responses to the second variable beat. Theoretically, it was expected that following the randomly placed stimulus, subjects would place final vowel onsets anywhere in the 20-80% range of the repetition cycle,

giving also a roughly random distribution for the responses. However the actual distributions were heavily peaked in three areas: roughly at one third, two thirds and one half of the repetition cycle (Fig.1).

In a recent paper, Port motivates this and similar results in chaos terms [10]. With respect to a large number of activities, humans in the presence of other humans show spontaneous, emergent coordination. No physical link is necessary. For example, Schmidt, Carello & Turvey (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[5]. 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 harmonics of the repetition cycle, such as at one half or at 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 behaviour is more likely than in-between attractors.

4. Does Port's Model Generalize to Normal Speech?

Of course, "normal speech" is quite different from the repeated utterance of 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 generalization of Port's model to normal speech is thus much in question. However, the notion of beats has intuitive appeal, and it may well provide a solution to the question of where variability is more or less likely. I propose to take Port's chaotic concept of speech one step further by rephrasing the arguments in terms of regularity and variability.

In this view, strong vowel onsets could well provide an underlying, regular structure for speech which at the surface can only partially translate into surface events, due to various linguistic, psycholinguistic, or sociolinguistic constraints such turn-taking, phrase structure, or syllable structure. *However the degree of variability in the surface events depends on the strength of the vowel onsets.* Close to strong vowel onsets (*attractor sites*), the variability would be small, while further removed from these sites, the variability could be much greater.

This proposal is testable. If a number of speakers read aloud the same paragraph, we expect less inter-subject variability for the placement of strong vowel onsets than for the placement of weak vowel onsets, always assuming that placement is expressed as a fraction of the containing utterance phrase.

We used French speakers for our test, which set up an additional difficulty. In French, strong and weak vowels are not distinguished on phonological grounds, as are stressed and un-/less stressed vowels in English or accented and un-/less accented vowels in such languages as German or Spanish. And yet, we would expect French to be subject to the same rules of variability as other languages. This means that we have to distinguish strong and weak vowel onsets solely on the basis of the empirical acoustic criteria. For this, we shall follow largely the method established up by Port and colleagues.

[5] More recent systematic investigations by Bingham have confirmed the stability of purely perceptual coupling for 0-degree phase movements and, to somewhat lesser degree, for 180-degree phase movements [14].

5. A Test with Nine French Speakers

Task and Segmentation. Nine members (2 F, 7 M) of the University of Lausanne recorded the same reading-aloud task of a paragraph of 218 words[6] in quiet surroundings. The recordings were performed at 96 kHz mono, normalized against the peak volume in the file, and downsampled to 16 kHz. Sentences varied considerably in duration and in underlying phrase and syllable structure. The phrase was taken as the reference duration (comparable to Port's repetition cycle). An on- or offset of a phrase was considered to have 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, and (c) a pause in excess of 150 ms. Recordings were manually segmented by experienced segmenters for both segmental and phrase duration using the Praat software, and were spot-checked by two other segmenters.

Strong and Weak Vowel Onsets. A slightly adapted version of the method used by Port and colleagues to distinguish between strong and weak vowel onsets was applied. It consists of the following steps:

- Obtain the sound file, low bandpass at 800 Hz and high bandpass at 200 Hz using Praat's pass Hann band filter.
- Take the intensity curve with Praat, 100 ms limit.
- Take a spline of the intensity curve. The tension is empirically adjusted in such a fashion that vowel onset intensity rises are continuous. In our version of the spline, the tension setting is 0.001.
- Take the derivative of the splined intensity curve and locate peaks. Strong peaks correspond to strong vowel onsets, and weak peaks to weak vowel onsets.
- Identify the manually segmented vowel onset at or near the peak. In our sample, segmented vowel onsets were situated on the average 25 ms past the acoustically identified vowel onset peak.
- Iteratively arrive at a criterion value for peak height to distinguish strong and weak vowel onsets. For our experiment, the criterion value was adjusted in such a manner that roughly half of the onsets in the test sample fall below the criterion value and roughly half above it. In other experiments, this criterion could be set in ways that distinguish stressed and unstressed or accented and unaccented vowels.

[6] 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.

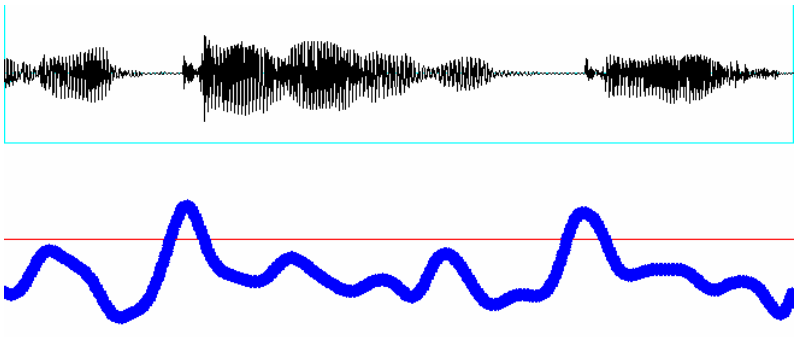

Figure 2. Two strong vowel onsets and several weak vowel onsets in "d'aut***omne*** invit***ait***". The lower graph shows the derivative of the splined intensity curve.

The Hypotheses. Two extensions of the chaos view of speech timing were tested, the harmonicity hypothesis proposed by Port, and the variability hypothesis proposed in this study.

Results 1: The Harmonicity Hypothesis. To test the harmonicity hypothesis, the distribution of vowel onset positions expressed as a fraction of the containing phrase was examined for peakedness at the second, third and fourth harmonic positions (see Fig. 3). Data from all nine speakers were placed into the same bins that had been created for Figure 1 ($N_{strongVO}$=703, N_{weakVO}=671, Total N=1374). Minor peaking was found at 50% of phase for both strong and weak vowel onsets. However, other peaks of same or similar size were found in other parts of the phase which are not evident harmonic positions. We thus conclude that for our French-language readers, harmonicity was barely a factor in the constitution of attractor sites.

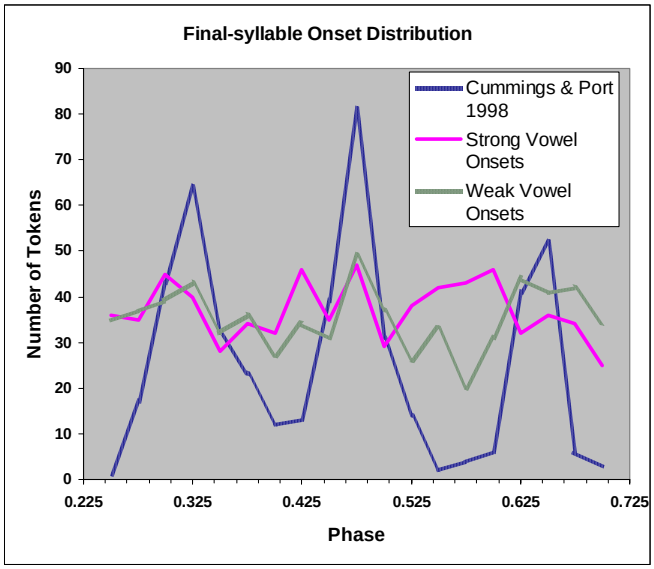

Figure 3. Vowel onset distribution for the two studies. Although there is a minor peak at 50% phase for both strong and weak vowel onsets, there was no overwhelming evidence in favour of the harmonicity hypothesis. Data for the Cummings & Port study were copied from Figure 1.

Results 2: The Variability Hypothesis. To test the variability hypothesis, vowel onset positions expressed as a fraction of the containing phrase were compared between all subjects (see Fig. 4). If two subjects used the same word grouping in their reading, and if they used the same vowels in the same phonetic context, the vowel onset position pair was included in the correlation. For one of the subjects (S8), who had shown more reading-aloud problems than the others, this resulted in less than 10 pairs in both conditions. For the remaining subjects, the number of pairs in the correlation was less critical ($mean_{strongVOpairs}$ = 20.5, $mean_{weakVOpairs}$ = 17.2, total N_{pairs} = 1356, total $N_{strongVOpairs}$ = 738, total $N_{weakVOpairs}$ = 618).

Thirty out of 36 correlations were in support of the variability hypothesis ($\chi 2=16$, df=1, $p<0.000$). Also, correlations for strong vowel onsets showed a much smaller range of values (from 0.972 to 0.788) than those for weak vowel onsets (from 0.988 to -0.042). This can be interpreted as a further indicator of variability. Altogether, these results appear to constitute good support for the variability hypothesis.

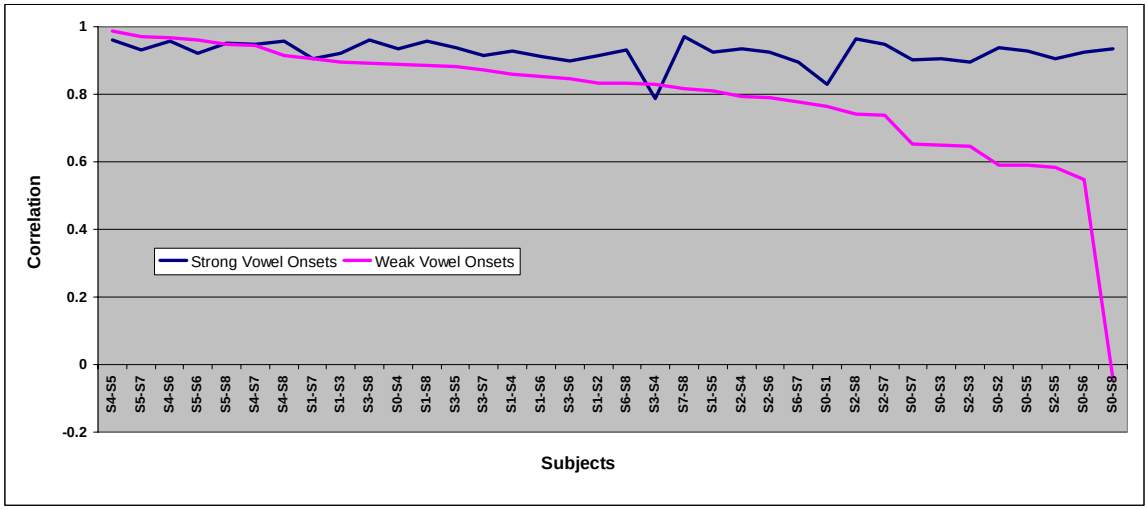

Figure 4. Subject-to-subject correlations for same-vowel onset positions in same-phrase contexts. Subjects showed greater agreement for strong vowel onsets than for weak vowel onsets, constituting good support for the variability hypothesis.

6. Conclusion

This study suggests a solution to the problem of excessive regularity in speech synthesis: introduce variability, but have it affect nearly exclusively weak vowel onsets. In chaos manner, the challenge is to set up a set of attractors, and to chart paths between them.

This in turn opens new questions for the implementation of timing in speech synthesis, the principal one being the following: While the location of attractors can be predicted by traditional statistical linear prediction schemes or by neural networks (see e.g., [4]), it is far from clear how the paths between attractors are to be defined. While it is interesting and instructive to find support for a chaos view of speech structure, further studies are required to establish the psycholinguistic and pragmatic reasons that lead a speaker to deviate from the statistically predicted timing scheme. Only when motivations and degrees of variation are known, implementation of the variation in speech synthesis can be attempted.

Another question is the generality of the present finding and its chaotic interpretation. The chaotic vision of speech, with strong vowel onsets serving as anchor points for temporal (and possibly syllabic and intonational) structure appears to make good sense for languages such as English, German and as we have seen, for French. But what about tone languages such as Chinese or Vietnamese?

It may also prove useful to examine the status of strong vowel onsets in larger context of timing schemes. In this study, strong vowel onsets appear to act as anchor points within larger word groups. Does that make them markers for minor word groups? Previous work in Lausanne has underlined the dependence of the minor group on the grammatical status of the constituting lexical items [5, 6]. What is the relationship between vowel onsets and minor groups?

Finally, it may well be of interest to consider the perceptuo-motor aspects of strong vowel onsets. With some temporal modification due to the influence of phonetic context, vowel onsets correspond to P-centres, the time at which a syllable is perceived to occur.

Given their perceptual saliency in the phonetic chain and the high interspeaker reliability of their placement, it is also likely that they occupy a central position in speech motor programming. It is quite possible that the speech motor control system's responsibility is to closely approximate expected articulatory events at strong vowel onsets, while it can be a bit more lax for other parts of the utterance chain.

References

[1] Allen, G. (1972). The location of rhythmic stress beats in English: An experimental study I. *Language and Speech, 15,* 72-100.

[2] Allen, G. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics, 3,* 75-86.

[3] Cummings, F., & Port, R. (1998). Rhythmic constraints on speech timing. *Journal of Phonetics, 26,* 145-171.

[4] 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.

[5] Keller, E., & Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17, University of York. 53-75.

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

[7] Patel, A.D., Löfqvist, A., & Naito, W. (1999). The acoustics and kinematics of regularly timed speech: A database and method for the study of the P-Center problem. *Proceedings of the XIVth International Congress of Phonetic Sciences*, San Francisco, 1999.

[8] Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics, 17,* 175-192.

[9] Pompino-Marschall, B. (1991). The syllable as a prosodic unit and the so-called P-centre effect. *Forschungsberichte des Instituts f"ur Phonetik und sprachliche Kommunikation der Universität München, 29,* 66-124.

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

[11] 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.

[12] 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.

[13] Tatham, M. and Morton, K. (2002) Computational modelling of speech production: English rhythm. In Angelika Braun and Herbert Masthoff (eds.) *Festschrift for Jens-Peter Koester on the Occasion of his 60th Birthday*. Stuttgart: Steiner, Zeitschrift für Dialectologie und Linguistik Beiheft, #121.

[14] Wilson, A. & Bingham, G.P. (in press). Perceptual coupling in rhythmic movement coordination – Stable perception leads to stable action. *Experimental Brain Research.*

[15] Villing, R., Ward, T., & Timoney, J. (2003). P-centre extraction from speech: The need for a more reliable measure. *Irish Systems and Signals Conference (ISSC),* 2003, Limerick. (Available at http://www.eeng.may.ie/~rvilling/publications/villing2003_pcentre_measure.pdf)

[16] Zellner Keller, B., & Keller, E. (2001). Representing Speech Rhythm. in Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. Eds. *Improvements in Speech Synthesis.* (pp. 154-164). Chichester: John Wiley.