SPEECH TIMING: APPROACHES TO SPEECH RHYTHM

Eric Keller[1] and Robert Port[2] (Session Organizers)

[1] University of Lausanne, [2] Indiana University
eric.keller@unil.ch, port@indiana.edu

ABSTRACT

In recent years, a number of authors have suggested various oscillator-based mechanisms to account for rhythmicity. This session brings together a number of researchers who have proposed and/or examined these proposals in detail with respect to a number of languages (English, Japanese, Brazilian Portuguese and French).

1. INTRODUCTION TO SESSION

The temporal aspects of phonetic structure pose varied challenges to phonetic science. Part of the puzzle has been solved for some time, since it is now well established that the time occupied by speech components reflects approximately the number of segmental elements to be produced, as well as their membership in various grammatical and lexical structures. This has permitted the creation of powerful predictive algorithms for speech synthesis and speech recognition; predicted times typically show correlations of .75 to .85 with observed times. But it has long been known that the segments do not capture all that is important about speech timing.

Indeed, researchers have for many years suspected that speech timing contains more systematic information that gives rise to the elusive perceptual impression of "speech rhythm". The exact moments for producing phonetically important elements might obey some larger and possibly context- and language-independent logic. In 1977, Ilse Lehiste, in an extensive review of the issue of isochrony (evidence for rhythmicity in speech) came to the conclusion that there were no direct acoustic correlates of rhythmicity. In this view, rhythmicity was a perceptual gestalt that emerges from the combination and complex interaction of a great number of acoustic and/or motor parameters. This view, supported by a number of further studies, has essentially formed the consensus for spontaneously produced speech since then.

2. THIS SESSION

In recent years, a number of authors have suggested results that have challenged this established notion. Robert Port and colleagues have shown that speakers appear to structure their utterances as a harmonic fraction of time available to produce sequences of whole phrases. By inserting their finding into the chaos-based "coordinative structures" rationale [1, 2], they provided an important conceptual link. Coordinative structures (also known as "synergies") are complex control structures produced by a large dynamical system. The crucial element identifying a coordinative structure is its independent and self-organizing nature. By information and activation paths that leave few external and measurable traces, the system attains its goals and objectives. An important property is that perturbations to the system are absorbed by it and are rapidly corrected or compensated for via internal adjustments. These concepts were applied to speech first by Scott Kelso [3] and then by Port [2] who showed how some very simple and widespread metrical patterns (like 2-beat or 3-beat waltz-like patterns) can be modelled by pulse-generating oscillators at integer-ratio frequencies that influence each other's phase and firing rate.

Because of its rapid and internally regulated dynamics, convincing demonstrations of coordinative structures in speech require the isolation of very simple systems. Luckily, there turns out to be a uniquely salient speech event that allows simplification of the speech signal to a series of discrete pulses, the vowel onsets. Several papers here emphasize that vowel onsets or voice onsets, (also called `perceptual centres' or `P-centres') as focal events in speech whose distribution in time can be attracted to oscillatory systems of various kinds.

The papers in this special session address different aspects of the issue of *how speech is globally structured in real time.* Keller's previous

research on timing and beats verified that certain features of speech do show a tendency to organize themselves into harmonic behavioural clusters, even when continuous speech was examined (in contrast to the repetition experiments by Port and colleagues). Specifically, strong vowel onsets or "beats" showed some tendency to cluster around the halfway point of coherent speech phrases (often called "spurts" in conversational analysis). What's more, strong beats were places of temporal structuring. They acted as "anchor spots" within the phrase when different speakers read the same text aloud, since at strong beat locations, speakers were in greater temporal agreement than at weak beat locations.

What made these beats so special? In his paper, Keller argues that voice onsets emerge from the combination of neurologically anticipated and actually occurring events. Because beats are perceptually salient and require precise articulatory coordination between glottal structures and supralaryngeal vocal tract, they seem to be used as temporal demarcation points in the speech chain. Also, the repeated and temporally constrained neuronal expectancy of beats may contribute to setting up pulsing patterns for oscillatory behaviours in speech.

The paper by Yukari Hirata and Connor Forbes provides further support for the importance of voice onsets. In an analysis of Japanese disyllables as measured from *one voice onset to the next voice onset*, the authors were able to show surprisingly clearly structured mora durations in geminates vs. singleton production, which calls into question the traditional mora definition for Japanese in terms of C and V segments. Their analysis used vowel onset interval durations divided by the local mean mora duration and supported the previous claims about the mora by Port and Brady and colleagues [4, 5].

Plinio Barbosa's results on his two-oscillator model for speech timing illustrate another, larger dimension of a similar set of issues. In previous research, he showed that patterns of V-to-V durations in Brazilian Portuguese could be handled by a two-coupled-oscillator speech rhythm model. Barbosa shows in the current paper that temporal variability can be accounted for within his model by having syntactic and temporal constraints act on at least two temporal organizing principles. The speech events that are coupled to the perceptual system are, again, voicing onsets.

Thus Barbosa's results confirm the trend to focus on the importance of beats or P-centers. He shows that his oscillators function correctly when the pulses are set to the vowel onsets. These become the "carrier components" of speech rhythm production, and then adapt to different temporal objectives established by the syntactic and accentual requirements of a speaker's material.

Brady, Port and Nagao return to the problem of what is regular about the Japanese mora. The authors show that the kind of irregularities found in Japanese moras rule out one kind of analysis scheme, one that models mora periodicity using an ``adaptive oscillator''. This system is a hypothesized cognitive device that would predict the next pulse or beat and then, if the pulse comes early or late, it immediately adjusts its period a little in the appropriate direction. The authors argue that a constantly adapting system like this may keep the system stable when there is a small amount of noise on period durations, but it cannot solve the problem of tracking the vowel onsets that define their version of the Japanese mora since this often involves some pulses that are almost at opposite phase to the predicted beat. This fact forces further model developments so that `discordant' beats can be ignored when necessary.

Along the way, they propose a method to reinterpret measurement of beat intervals in absolute milliseconds into phase angles of the most likely sinusoid to align with the pulse sequence over a neighbourhood of at least several words. The proposed display is a unit circle with pulses represented as vectors at various phase angles. When these vectors are averaged they yield a potentially very useful measure of the consistency of alignment of the pulses.

The final paper by Nick Campbell represents an attempt to understand what two-person conversations are like. Although most people have assumed they involve trading the "floor" from one speaker to the other – like tennis players batting a ball back and forth – his results on large amounts of diadic conversation in Japanese (by both native and non-native speakers) show that speakers actually overlap each other far more than most people thought. Indeed there is so much simultaneous speech that it is difficult to tell from looking at the voice tracks who is the primary speaker over large fractions of a conversation.

Many of these papers encourage looking at the signal for speech timing differently, not in terms of segmental descriptions but in terms of vowel onsets or P-centres. Also, recent work clarifies the notion of an "oscillator," which has been a rather suspect entity in cognitive science, since past searches for the directly responsible neural oscillators for speech have not been successful. The confluence of the work

reported here suggests that models for plausible temporal structures responsible for rhythmic activity are likely to emerge from the self-organizing operation of dynamical systems. These systems apparently tend to orient themselves to salient vowel onsets occurring in the speech chain.

3. CONCLUSION

Most of these contributions focus on various aspects of *vowel onsets* in the temporal structuring of speech. In this manner, research on the temporal aspects of speech may have opened yet another chapter on the central cognitive structuring of human speech. Voice onsets, due to their high perceptual salience, have been shown to play a key role in the temporal organization of syllables. Their phonetic locations have been shown to be crucial for the quality of speech synthesis, and they incorporate many key parameters for speech perception. But more sophisticated dynamical models are also under development and will contribute to our understanding of the real-time structuring of speech. Perhaps we will soon come to a clear understanding of the empirical basis of that elusive feeling of speech rhythmicity.

4. REFERENCES

[1] Port, R., Tajima, K., and Cummins, F. (1999). Speech and rhythmic behavior. In Savelsburgh, G. J. P., van der Maas, H., and van Geert, P. C. L., eds., *The Non-linear Analysis of Developmental Processes.* Elsevier, Amsterdam.

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

[3] Kelso, J.A. (1995) *Dynamic Patterns: The Self-Organization of Brain and Behavior.* (MIT Press, Cambridge, MA)

[4] Port, R. F.. Dalby, J. and O'Dell, M. (1987) Evidence for mora timing in Japanese. *J. Acoust. Soc. Amer*. 81, 1574-1585.

[5] Brady, M. C., Port, R. F., Nagao, K. 2006. Effects of speaking style on the regularity of mora timing in Japanese. *J. Acoust. Soc. Am* 120 (5), 3208.