Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Phonetics 31 (2003) 619–626
Journal of Phonetics
www.elsevier.com/locate/phonetics

Patterns of timing in the acquisition, perception, and production of speech

Christopher T. Kello*

Department of Psychology, George Mason University, Fairfax, VA 22030, USA

Received 17 September 2002; received in revised form 13 February 2003; accepted 17 February 2003

Abstract

In this commentary, the bodily and cognitive bases of timing in speech are discussed. Timing is viewed as playing a supportive role for the acquisition of speech, and for the cognitive processes involved in speech comprehension and production. The patterns of timing in speech are viewed as emerging from the coordinated dynamics of the body, and from the coordinated dynamics between behavior and cognition. These coordinated dynamics are viewed as existing both within and between individuals, and on multiple time scales.

1. Introduction

All languages code at least some amount of linguistic information in the ordering of events in the speech stream. The fact that order information is fundamental to speech has been underscored by recent studies showing that both infants and adults can pick up on statistical regularities in the ordering of events (syllables, in this case) in streams of novel speech stimuli (Saffran, Aslin, & Newport, 1996). Of course, order is not the only kind of temporal information that is important to speech. Linguistic information is also coded by *patterns* in the timing of events in the speech stream, i.e., regularities in the quantities of real time that occur between speech events.

In this volume, Anne Christophe and Robert Port and their colleagues demonstrated the importance of patterns of timing in the perception (Christophe, Gout, Peperkamp, & Morgan, 2003) and production (Port & Kaipainen, 2003) of speech. In this commentary, the Christophe and Port studies serve as points of departure for rumination on the bases of timing in speech.

*Tel.: +1-703-993-1744; fax: +1-703-993-1330.
E-mail address: ckello@gmu.edu (C.T. Kello).

0095-4470/03/$ - see front matter
doi:10.1016/S0095-4470(03)00017-2

Attention is given in particular to some of the bodily and cognitive factors that may play a role in the ability to use the timing of speech events as a channel for speech communication.

A key hypothesis that emerges from the bodily/cognitive approach is that a common set of mechanisms exist in the brain to support functions of both speech perception and production (Werker, 1993; Plaut & Kello, 1999; Hickok, 2001). This hypothesis is motivated by the fact that much of the knowledge and skill learned for speech perception is also useful for speech production (and vice versa), and that an effective and efficient (if not optimal) way to share learning across modalities is to share mechanisms across modalities. Thus, some of the mechanisms that process the timing of events in speech perception should also be responsible for the timing of events in speech production. In this commentary, the "shared mechanisms" hypothesis helped to guide the search for the bases of timing in speech acquisition, perception, and production.

2. The supportive role of timing in speech acquisition

There are many challenges to learning a spoken language, and a major aim of speech research is to identify the factors that help the infant language learner overcome those challenges. One such factor may be patterns in the timing of speech events. In other words, do language learners rely on the timing of speech events to "bootstrap" the process of speech acquisition, and if so, how?

One useful approach to this question is to consider that one of the biggest challenges to learning a spoken language is to segment the continuous speech stream into words. Christophe et al. (2003) asked whether infants or adults can use the perceptual cues to phonological phrase boundaries in order to help overcome this challenge. Phonological phrase boundaries are cued mainly by durations of the pre- and post-boundary syllables, and syllable durations are determined by the timing of speech events, e.g., the onsets and offsets of voicing. Therefore, by testing both infants and adults, Christophe et al. addressed the more general question of whether the timing of speech events may play a role in bootstrapping the process of speech acquisition.

In one experiment, Christophe et al. (2003) showed that French-speaking adults were faster to detect target words in sentences with ambiguities that straddled a phonological phrase boundary compared with ambiguities that did not straddle such a boundary. For example, participants were faster to detect the French word *chat* in a sentence like "…, [le gros chat] [grimpait aux arbres.]" compared with a sentence like "[Le livre] [racontait l'histoire] [d'un chat grincheux]…" (brackets denote phonological phrases). Both sentences contain a potential segmentation ambiguity because both *chat* and *chagrin* are words in French. Therefore, one might expect the *chagrin* option to interfere with the detection of *chat*, relative to a control condition in which no such ambiguity exists. The detection times collected by Christophe and her colleagues suggested that the *chagrin* option does not cause interference when the two syllables are divided by a phonological phrase boundary, but does when there is no such division. In another experiment, infants showed a similar sensitivity to phonological phrase boundaries in a variant of the conditioned head-turning paradigm. When considered in the context of other similar studies (e.g., Kemler-Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989; Christophe, Mehler, & Sebastián-Gallés, 2001), the experiments conducted by Christophe et al. (2003) provide some initial support for the idea that the timing of speech events can be used by a language learner to bootstrap the ability to segment words in the speech stream.

To further question how timing might play a supportive role in the acquisition of speech, it is useful to consider that events in the speech stream can recur in somewhat regular periods (on multiple time scales), and that periodicity is perhaps the simplest and most basic kind of timing possible. Given these premises, it stands to reason that periodicity in the timing of speech events may serve as a foundation upon which processes of speech perception and production can develop. In fact, with respect to speech production, MacNeilage and his colleagues have argued for the idea that periodicity in the mandibular oscillations of babbling infants serves as the syllabic frame upon which segmental gestures can be overlaid (i.e., the *frame/content theory*; Davis & MacNeilage, 1995; MacNeilage, 1998). The ability to generate mandibular oscillations presumably has its evolutionary roots in the chewing of food for ingestion, and controlling the rate of oscillation is a matter of timing. The frame/content theory holds that these "old" mechanisms of oscillatory timing are used for the purpose of producing speech. This is one way that periodicity might support the acquisition of speech.

It is important to note that, while periodicity may play a supportive role in the acquisition of speech, this does not necessarily entail that adult speech must exhibit periodicity (although the existence of periodicity in adult speech has been argued for, e.g., stressed syllables may tend to occur in quasi-regular intervals in stress-timed languages; Pike, 1943). A more circumspect statement would be to say that periodicity may or may not occur in any given utterance, but periodicity is always intrinsic to the dynamics of the speech apparatus.

In their conclusion, Port and Kaipainen (2003) made exactly this point about periodicity in speech, but they reached it from a different angle. Their study was motivated by the dynamics of coupled oscillators. In a variety of instances in which two periodic movements are coupled in some way (e.g., the wagging of two fingers (Kelso, 1984) the swinging of two legs of two different people (Schmidt, Carello, & Turvey, 1990) clapping to the beat of a metronome; Kelso, DelColle, & Schöner, 1990), researchers have consistently found that the relative phase of those movements tends towards two values (i.e., *attractors*), one at 0 (*inphase*, i.e., synchronized) and one at 0.5 (*antiphase*, i.e., syncopated). Moreover, the antiphase attractor can weaken and eventually disappear as the frequency of the periodic movements increases. Haken, Kelso, and Bunz (1985) developed an elegant and influential model of these oscillatory dynamics (the HKB model), and the model makes some very general predictions about the nature of transitions as the antiphase attractor goes in and out of existence.

Port and Kaipainen (2003) tested whether the model could be applied to speech production. They conducted two experiments in which adult English speakers were asked to repeatedly pronounce the syllable /da/ such that the phase of pronunciations was either synchronized or syncopated with the phase of a metronome. Syllable repetition can be considered as one oscillatory movement, and the metronome can be considered as a second oscillatory movement. The task was designed to couple these oscillators such that their phase relationship was either inphase or antiphase. The metronome was manipulated to test two predictions made by the HKB model (not discussed here in the interest of space), and results were consistent with these predictions.

The Port and Kaipainen (2003) study does not directly demonstrate that oscillatory dynamics underlie adult speech production, nor does it directly demonstrate that oscillatory dynamics play a supportive role in the acquisition of speech. What the study does show is that the dynamics described by the HKB model are at least latent in speech production, and under the proper

conditions, they can be observed directly. These latent dynamics become more relevant to speech acquisition in light of the fact that other periodic movements of the body have been shown to exhibit the dynamics described by the HKB model. Oscillatory movements are common throughout human behavior, and when two such oscillators couple together, evidence suggests that the emergent dynamics will conform to the HKB model. More generally, evidence is mounting to suggest that coupled oscillators provide just one example of how the coordinated dynamics of human movements follow consistent, lawful patterns (Kelso, 1995). The lawful and ubiquitous nature of these dynamic patterns makes them prime candidates to play a supportive, foundational role in the acquisition of speech. Exactly how they might play such a role is currently an open question.

3. Cognitive and bodily bases of timing in the perception and production of speech

What are the cognitive and bodily factors that help to create the patterns of timing observed in the speech stream? To begin with, the dynamics intrinsic to the speech apparatus may contribute to the patterns of timing generated in both child and adult speech, in addition to its supportive role that was discussed in the previous section. As mentioned earlier, this was one of the main points in the study by Port and Kaipainen (2003). The dynamic systems approach can also be applied more generally by considering the dynamics that may be intrinsic to the *coordination* of the speech apparatus with other factors at play in the talker and in the language environment.

With respect to the talker, the language and cognitive processes that support the motor output for speech production must overlap in time with the actual movements of the speech apparatus, at least to some degree (i.e., incremental speech production; Kawamoto, Kello, Jones, & Bame, 1998; Roelofs, 1998; Kawamoto, Kello, Higareda, & Vu, 1999). One consequence of this overlap is that language and cognitive processing must be *timed* with movements of the speech apparatus. This timing must be fairly strict to the extent that processing is tightly interleaved with motor output (i.e., highly incremental), and there is evidence showing that processing can be interleaved at least at the level of the syllable (Kello, Plaut, & MacWhinney, 2000; Meyer, Roelofs, & Levelt, 2003). Therefore, it stands to reason that the timing of events in the speech stream is partly determined by the coordinated dynamics between the speech apparatus and the supporting processes of language and cognition. The influence of this coordination on timing is perhaps revealed most clearly when coordination breaks down, i.e., when speech is mis-timed (e.g., stutters, mis-starts) in conjunction with faulty processing (for a review, see Fox Tree, 2000).

Coordination of the speech apparatus with language and cognitive processes happens on a relatively fine-grained time scale. Are there coordinated dynamics on larger time scales that contribute to the creation of patterns in the timing of speech events? One possible answer to this question is based on interactions among talkers in a conversation. In order to generate coherent and effective communication, the speech of one talker must be coordinated with that of the other participants in the conversation, and this coordination is largely an issue of timing. For example, prosodic cues, defined partly by the timing of speech events, may be used to signal when one talker is ready to "pass the baton" onto another talker, or when a talker is ready to change the topic of conversation. More related to the oscillatory dynamics studied by Port and Kaipainen (2003), it has been shown that the speech and gestures among participants in a conversation can become

entrained with each other (Warner, 1992). A similar example outside the domain of speech is the finding that clapping in crowds of people can become self-organized into synchronous waves (Néda, Ravasz, Brechet, Vicsek, & Barabási, 2000). These two examples illustrate how patterns in the timing of speech events might emerge from the coordination of talkers in a conversation. The challenge for future research is to determine whether and how emergent dynamics can contribute to the creation of patterns in the timing of speech.

4. Extra-linguistic functions of timing in the perception and production of speech

The primary function of timing in speech is presumably to support communication by helping to code segmental and suprasegmental information. It is by virtue of this function that timing helps to code linguistic information in the speech stream, but by the same token, timing helps to code nonlinguistic information as well. Types of nonlinguistic information communicated by segmental and suprasegmental information include emotional state (Erickson, Fujimura, & Pardo, 1998), talker identity (Remez, Fellowes, & Rubin, 1997; Nygaard & Pisoni, 1998), and communicative intent (Fernald, 1989).

In addition to its communicative benefits, how else might timing be advantageous to the talker or listener? One answer is directly related to the hypothesized role of timing in talker identity. Talkers may somehow infuse a "temporal signature" into their speech, i.e., patterns in the timing of speech events that are unique to the talker. If so, these patterns may serve not only to identify the talker, but also to distinguish the speech signal emitted by a single talker from the background of signals produced by other talkers and sources of nonspeech sounds.

It is potentially useful to extend the "source separation" use of timing beyond the speech environment, and into the mind of the listener. Processing in the human nervous system is highly parallel in nature, and neural pathways tend to diverge, converge, and then diverge again as they are traced from sensory inputs through sub-cortical areas and into successively higher levels of cortex. These properties of the nervous system have raised a major question in the cognitive neurosciences: how are the objects of perception and cognition segregated and maintained in the nervous system when their dimensions (e.g., frequency or location) are processed in anatomically separate neural pathways, and then recombined at later stages of processing? This is one example of the well-known *binding problem*, and one proposed solution to this problem is that the timing of action potentials (spikes) is somehow synchronized for all the neurons that are processing a given object for a given length of time (e.g., Engel, Konig, & Singer, 1991). If the patterns of timing for a given object are somehow distinguished from the patterns for other objects (e.g., by differences in phase), then this timing information could be used at later stages of processing to bind the sources of information coming in from separate, parallel pathways.

If this proposed solution is correct, then where do these distinguishable patterns of timing come from? One possibility is that the brain somehow assigns each perceptual/cognitive object a unique timing pattern, much like assigning unique identifiers to records in a database. However, if unique patterns of timing exist in the sensory inputs coming from a given object in the environment, then it stands to reason that those patterns of timing could be somehow maintained in the timing of action potentials, albeit through some transformation. On this account, unique patterns of timing in the sensory objects would serve to generate unique patterns of timing in the neural processing

of those objects. The "temporal signatures" emitted by talkers (one class of sensory objects) are hypothesized to be just such unique patterns of timing. In this way, temporal signatures might serve to distinguish talkers in the environment, as well as in the mind and brain of the listener.

Another way to approach the question of how the timing of speech events might benefit the talker or perceiver is to once again consider that periodicity can often be found in the timing of events in the speech stream. Such periodicity is most apparent in singing, but it is also perceptible and measurable in certain types of conversational utterances, such as those in which a list of items is enumerated in English (e.g., imagine saying the following, "I need to buy a loaf of bread, a carton of milk, and a stick of butter"). Everyday experience tells us that it is easier to remember and recall a song with catchy rhythm compared with no rhythm, and there is evidence to support this intuition (Wallace, 1994). Results such as these suggest that periodicity in the timing of speech events may facilitate processes of memory for both the talker and perceiver.

There is also evidence to suggest that periodicity may facilitate the motor processes of speech production (for the developmental bases of this hypothesis, see Section 1). Stutterers can sometimes temporarily reduce or eliminate a stutter by imposing a rhythm on their utterances (Healey, Mallard, & Adams, 1976). Nonstutterers will naturally and unavoidably fall into a rhythm with respect to the timing of vowel onsets when they are instructed to repeat a phrase or sentence (Port, 2002). Moreover, in the experiments reported by Port, rhythms with low-order ratios (1:1, 2:1, 3:2, etc.) were found to be easier and more common than rhythms with high-order ratios (4:3, 5:3, etc.). This finding is consistent with similar studies of human movement in other domains (e.g., finger tapping): the observed ordering of rhythms in terms of ease and commonality obey the *Farey tree* hierarchy (Kelso, 1995). What emerges from these studies is a story similar to the one told for speech acquisition in Section 1 of this commentary. In speech production, as in speech acquisition, dynamics intrinsic to the speech apparatus serve as the basis of timing, and these dynamics can be examined directly under certain experimental manipulations.

5. Concluding remarks

It could be argued that evolution has shaped the nervous system with the overarching purpose of controlling the body. Given that speech is a new behavior relative to evolutionary time, it stands to reason that speech has evolved to leverage the older perceptual, cognitive, and motor mechanisms that control the body. As with all aspects of speech, this logic should apply to both the perception and production of patterns of timing in the speech stream. For example, neuroanatomical and neuropsychological studies indicate that the basal ganglia play a role in controlling the timing and sequencing of most all human movements (Harrington & Haaland, 1998), including speech (Lieberman, 2000). This neural, domain-general basis of timing in speech is consistent with one of the major thrusts of this commentary, that there are coordinated dynamics intrinsic to all human movements, and that patterns of timing in speech are founded upon those dynamics. One of the many issues that remain for future research is to elucidate the role of neural mechanisms such as the basal ganglia in the timing of speech, and more generally, in the dynamics of human movement.

References

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, *31*, this issue.

Christophe, A., Mehler, J., & Sebastián-Gallés, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, *2*, 385–394.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research*, *38*, 1199–1211.

Engel, A. K., Konig, P., & Singer, W. (1991). Direct physiological evidence for scene segmentation by temporal coding. *Proceedings of the National Academy of Sciences*, *88*, 9136–9140.

Erickson, D., Fujimura, O., & Pardo, B. (1998). Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech*, *41*, 399–417.

Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: Is the melody the message? *Child Development*, *60*, 1497–1510.

Fox Tree, J. E. (2000). Coordinating spontaneous talk. In L. Wheeldon (Ed.), *Aspects of language production* (pp. 375–406). Philadelphia, PA: Taylor & Francis.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, *51*, 347–356.

Harrington, D. L., & Haaland, K. Y. (1998). Sequencing and timing operations of the basal ganglia. In D. A. Rosenbaum, & C. E. Collyer (Eds.), *Timing of behavior: Neural, psychological, and computational perspectives* (pp. 35–61). Cambridge, MA: The MIT Press.

Healey, E. C., Mallard, A. R., & Adams, M. R. (1976). Factors contributing to the reduction of stuttering during singing. *Journal of Speech and Hearing Research*, *19*, 475–480.

Hickok, G. (2001). Functional anatomy of speech perception and speech production: Psycholinguistic implications. *Journal of Psycholinguistic Research*, *30*, 225–235.

Kawamoto, A. H., Kello, C. T., Higareda, I., & Vu, J. (1999). Parallel processing and initial phoneme criterion in naming words: Evidence from frequency effects on onset and rime duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 362–381.

Kawamoto, A. H., Kello, C. T., Jones, R. J., & Bame, K. (1998). Initial phoneme versus whole word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 862–885.

Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology: General*, *129*, 340–360.

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, *15*, R1000–R1004.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kelso, J. A. S., DelColle, J., & Schöner, G. (1990). Action-perception as a pattern formation process. In M. Jeannerod (Ed.), *Attention and performance XIII* (pp. 139–169). Hillsdale, NJ: Erlbaum.

Kemler-Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*, 55–68.

Lieberman, P. (2000). *Human language and our reptilian brain: The subcortical bases of speech, syntax, and thought*. Cambridge, MA: Harvard University Press.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–546.

Meyer, A., Roelofs, A., & Levelt, P. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, *48*, 131–147.

Néda, E., Ravasz, Y., Brechet, T., Vicsek, A., & Barabási, L. (2000). Self-organizing processes: The sound of many hands clapping. *Nature*, *403*, 849–850.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, *60*, 355–376.

Pike, K. L. (1943). *Phonetics*. Ann Arbor: University of Michigan Press.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Maweh, NJ: Erlbaum.

Port, R. (2002). Implications of rhythmic discreteness in the production of speech. Presented at the *ISCA workshop on temporal integration in the perception of speech,* Aix-en-Provence, France, 8–10 April 2002.

Port, R., & Kaipainen, M. (2003). Rhythmic production of speech. *Journal of Phonetics*, *31*, this issue.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651–666.

Roelofs, A. (1998). Rightward incrementality in encoding simple phrasal forms in speech production: Verb-particle combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 904–921.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-olds. *Science*, *274*, 1926–1928.

Schmidt, R. C., 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.

Wallace, W. T. (1994). Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1471–1485.

Warner, R. M. (1992). Cyclicity of vocal activity increases during conversation: Support for a nonlinear systems model of dyadic social interaction. *Behavioral Science*, *37*, 128–138.

Werker, J. F. (1993). The contribution of the relation between vocal production and perception to a developing phonological system. *Journal of Phonetics*, *21*, 177–180.