Response to Fred Cummins: Looking for Rhythm in Speech

DAVID HOUSE
KTH, Stockholm: Royal Institute of Technology, Department of Speech, Music & Hearing

ABSTRACT: This commentary briefly reviews three aspects of rhythm in speech. The first concerns the issues of what to measure and how measurements should relate to rhythm's communicative functions. The second relates to how tonal and durational features of speech contribute to the percept of rhythm, noting evidence that indicates such features can be tightly language-specific. The third aspect addressed is how bodily gestures integrate with and enhance the communicative functions of speech rhythm.

Submitted 2012 January 6; accepted 2012 July 13.

KEYWORDS: *rhythm, prominence, grouping, prosody, turn-taking*

IN his contribution, ”Looking for Rhythm in Speech,” Fred Cummins begins his introduction with the observation that rhythm is notoriously difficult to define and means very different things to different people. This is perhaps one of the main reasons why rhythm in speech has proved so difficult to study and why it remains such a fascinating and challenging topic. It is also why we must restrict ourselves to covering only a few sub-topics when dealing with the concept of rhythm in speech. Cummins reviews some of the history of empirical research which has frustratingly and unsuccessfully sought to find isochrony in speech. He then moves on to argue for a more promising and fruitful, holistic approach which embraces rhythm as engaging the entire body during speech. This view of rhythm encompasses bodily gestures as well as gestures of the speech articulators and integrates rhythm into a whole-body activity. I am very sympathetic to this approach and would like here to comment on three related topics. The first concerns what to measure and why, and how these measurements should relate to the communicative functions of rhythm. The second is how tonal and durational aspects of speech contribute to the percept of rhythm. The third is how bodily gestures integrate with and enhance speech rhythm in its communicative function.

Fred Cummins describes what he fittingly terms “the great isochrony safari” as the hunt for measuring and finding regular temporal intervals in speech related to syllables or stresses. Measurements of acoustic regularity are meant to capture the regularity of articulatory dynamics, but as Cummins points out, the motivation for these measurements stems from intuition or at best a perceived regularity, the structure of which seems to differ across languages. This perception of regularity has proved difficult if not impossible to capture and quantify by means of purely temporal measurements leading to a widespread conclusion that rhythm in speech is a perceptual phenomenon rather than an acoustic one (e.g. Kohler, 2009; Niebuhr & Wolf, 2011). This somewhat paradoxical definition of rhythm as based in perception is not very satisfactory alone. I feel that a more useful and constructive definition should take into account the communicative functions of rhythm. This approach concentrates more on the reasons why we have the percept of rhythmical groups in speech and how this facilitates speech processing rather than purely on the metrics and measurements of regularity.

In a linguistic sense, rhythm is part of the prosodic structure of an utterance. The principal linguistic functions of prosody are generally taken to be the structuring of an utterance over and beyond the individual speech segments and thereby the conveyance of prominence and grouping. In other words, prosody transforms the linear progression of speech sounds into a multidimensional stream in which prominence is highlighted against a backdrop of supportive structure. Groups of syllables and words are made to stand out against one another. We can thus approach the perception of rhythmic regularity as one of the major contributing factors to the two functions of prominence and grouping. Prominence on the syllable level can be coupled to recurring patterns of alternation between stronger and weaker syllables, i.e. rhythmic alternation. Stronger syllables are made perceptually more salient and in this way receive added attention. The patterns of prominence themselves can create the percept of grouping and organize the speech stream into phrases, groups or chunks, and in this way further

facilitate perceptual processing. Here we find an interaction between prominence and phrasing which is not altogether well understood. Changes in the general tempo of an utterance, such as acceleration or deceleration can be used to signal phrase boundaries. One example of this is phrase-final lengthening which is one way of typically signaling phrase boundaries while coherence within a phrase, rhythmical uniformity, can also be used to delimit and define a phrase.

Languages differ from one another in the exact ways that they create these percepts of prominence and grouping, but as Cummins points out, seeking to establish two or three mutually exclusive language typologies based on metrics has not been fruitful. Instead we should investigate how the universal principles of prominence and grouping apply to all languages and in what ways languages differ in the manner in which they convey prominence and grouping (c.f. Arvaniti, 2009). A central question is to investigate how we, as speakers of languages, use these principles of prominence and grouping to give structure to an utterance which often reflects its information and grammatical structure.

The notion of rhythm is typically associated most closely with the temporal domain where long and short durational alternation creates the rhythmic percept as related to meter in poetry. It has been demonstrated by numerous experiments, however, that not only temporal patterns but also tonal patterns strongly contribute to the percept of rhythm (Niebuhr, 2009). In House (1990) for example, an experimental paradigm was developed where Swedish listeners were presented with a series of five numbers (such as a telephone number) and asked to say if they perceived the grouping as 2+3 or 3+2 (55-555 or 555-55). In this paradigm, the temporal domain was held constant so that all numbers had the same duration. The tonal configurations of the stimuli were manipulated systematically by introducing various rising and falling patterns to the number series. Strong percepts of rhythmical grouping were obtained by this method. Especially strong grouping percepts were obtained by rising-falling tonal patterns defining a group with a falling tonal movement ending in a low tone marking the end of each group. Percepts of grouping were also created by signals of coherence within each group such as recurrent pitch patterns on each element in the group (e.g. rising on each element in one group contrasting to falling in the other group) or uniformity of pitch levels within each group (e.g. high pitch in one group contrasting to low pitch in the other group).

A similar paradigm has recently been developed and used by Cumming (2010) to investigate which domain (temporal or tonal) serves as the primary cue for rhythm across languages. She found that both temporal and tonal variation are strong cues for rhythm, but that the perceptual relationship determining which cue is stronger varies across French, Swiss German and Swiss French. While the two domains were found to be interdependent for listeners of all three languages, the tonal domain (rising pitch) was found to be a more important cue for Swiss German speakers, while the temporal domain was found to be a more important cue for the French speakers (both French and Swiss French).

Another example of language differences in the perception of accentuation can be found in Beaugendre et al. (2001). In a series of experiments using repetitive five-syllable utterances, the timing of the intonational movement was varied so that it gave a percept of accentuation on either the third or fourth syllable. The three middle syllables were exact replicates of each other as to duration and spectral content. A shift in perceived accentuation from the third to the fourth syllable was triggered by a rising intonational movement earlier in the third syllable for French subjects and later in the third syllable for Dutch and Swedish subjects. If we approach different accentuation patterns as giving rise to different rhythmical groupings we see in these examples that the same intonation contour can contribute to different percepts of rhythm depending on the native language of the listener.

Other recent work (e.g. Niebuhr & Wolf, 2011) has demonstrated that not only do temporal and tonal patterns contribute to the percept of rhythm, but that also variation in the intensity of syllables and variation of the power spectrum of the vowels can contribute to rhythm. Taken together, this evidence indicates that as speakers we use a variety of acoustic material to create a percept of rhythm in the perceiver. The way in which this is done varies across languages, with speakers of different languages enlisting different sets of cues, but the universal function of rhythm is to facilitate communication by signaling prominence and grouping, and our measurements should be relevant to and reflect this function.

Although a great deal of effort in measuring rhythm has been in the context of single speaker utterances and monologues, it is most likely in the domain of dialogue that rhythm achieves its highest level of functional importance in speech. Cummins concludes his contribution by viewing rhythm as a social affordance in the joint domain of conversational interaction, where rhythm allows participants to entrain and adapt their rhythmical patterns to each other and where rhythmical patterns are signaled both by acoustic cues and through visual gestures and body movements.

In dialogue, as in monologue, rhythm serves the function of prominence and grouping. However, in dialogue we can approach rhythm both on the micro level (e.g. within the single utterance of one speaker) and on a macro level (e.g. the rhythmic structure of the dialogue as a whole). On the macro level, the grouping of the interaction into speaker turns is a fundamental function of rhythmic alternation (Edlund et al., 2009). While the establishment of rhythmical patterns can be instrumental for

regulating turn-taking, the interruption or disruption of such patterns established in the dialogue can also be important cues for grouping dialogue into turns (e.g. Thòrisson, 2002). This type of disruption, such as a hesitation pause, can also be an important signal on the micro level making, for example, a grammatical statement into a question (House, 2003).

I share Cummins's views that one of the most promising and exciting avenues for research into speech rhythm involves not only the movement of the articulators, but also facial and body gestures and whole body movements. Eyebrow and head movements alone have been shown to serve as strong cues to word prominence when acoustic cues are ambiguous (House, et al. 2001) and can thereby contribute to rhythm on the micro level in the same way as hand and arm gestures. As we are able to increasingly work with larger databases of dialogue we will gain more understanding of how articulator movements are synchronized and organized in relationship to face and body movements, and how these two modalities are perceived. We will also be able to increase our knowledge about how organization and synchronization principles apply between speakers. One way of approaching this is to capture whole body movement during recordings of extended dialogues (Beskow et al. 2010) enabling us to create dialogue participant movement profiles with which we can study in detail the rhythmic structure of a dialogue on the macro level both acoustically and visually.

Fred Cummins has presented a strong case for viewing rhythm as an explanation for our ability to adapt to one another within the patterns of conversational interaction. His orientation is on movement synchrony which facilitates spoken language communication. Discovering how this synchrony relates to different levels of prominence and grouping is without a doubt one of the most exciting areas of research into rhythm in language and speech.

REFERENCES

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, Vol.* 66*,* pp. 46-63.

Beaugendre, F., House, D., & Hermes, D. J. (2001). Accentuation boundaries in Dutch, French and Swedish. *Speech Communication,* London, UK. Vol. 33, pp. 305-318.

Beskow, J., Edlund, J., Granström, B., Gustafson, J., & House, D. (2010). Face-to-face interaction and the KTH Cooking Show. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., & Nijholt, A. (Eds.), *Development of Multimodal Interfaces: Active Listening and Synchrony.* Berlin / Heidelberg: Springer, pp. 157-168.

Cumming, R. (2010). The interdependence of tonal and durational cues in the perception of rhythmic groups. *Phonetica,* Vol. 67, pp. 219-242.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica,* Vol. 66, pp. 15-28.

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009.* Brighton: UK, pp. 2779-2782.

House, D. (1990). *Tonal perception in speech.* Lund: Lund University Press.

House, D. (2003). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. In *Proceedings of ICPhS, XV Intl Conference of Phonetic Sciences.* Barcelona: Spain, pp. 755-758.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of Eurospeech 2001.* Aalborg: Denmark, pp. 387-390.

Kohler, K.J. (2009). Rhythm in speech and language. A new research paradigm. *Phonetica*, Vol. 66, pp. 29-45.

Niebuhr, O. (2009). Fundamental frequency-based rhythm effects on the perception of local syllable prominence. *Phonetica,* Vol. 66, pp. 95-112.

Niebuhr, O., & Wolf, A. (2011). Low and high, short and long by crook or by hook? In *Proceedings of Interspeech 2011.* Florence: Italy, pp. 1869-1872.

Thòrisson, K.R. (2002). Natural turn-taking needs no manual: computational theory and model, from perception to action. In Granström, B., House, D., & Karlsson, I. (Eds.), *Multimodality in language and speech systems.* Dordrecht: Kluwer Academic Publishers, pp. 209-241.