1

Talking Fast: The Use of Speech Rate as Iconic Gesture

MARCUS PERLMAN

In the recently popular country song "Our Song," Taylor Swift recalls a memory of talking to her high school boyfriend on the phone. She sings,

> When we're on the phone and you talk *reeeeal slow*,
> because it's late, and your mama don't know.

As Swift sings the phrase "real slow", she performs an interesting effect with her voice. Whereas the other words in these two lines last for an eighth or a sixteenth note, the word "real" is extended for a quarter note. It is, in fact, the only instance in the song in which the second to last word, opposed to the last, is the longest in the line. The listener is given the clear impression that Swift is using her voice to iconically depict the slow temporal quality of the phone conversation.

In another popular country song, Garth Brooks uses the pitch of his voice to create a similar sort of iconic vocal effect when he sings, "I've got friends in *low* places." Prior to "low," the line moves between small intervals: two major seconds and a minor third. However, as he sings the word "low," Brooks pushes the lower end of his vocal range by descending a much larger interval of a major sixth, thus creating an iconic emphasis on the sense of lowness he is expressing. It is also interesting in this example to note the metaphorical extension of the spatial word "low" to the target domain of social class. This metaphor is expressed linguistically and mirrored with the iconic drop in pitch, similar to how people often make metaphorical extensions with gestures (McNeill, 2005).

Casual observation reveals that the use of iconic voice is not isolated to country music, and indeed, seems to be a common occurrence within everyday, ordinary conversation. For example, it is easy to imagine a speaker lengthening vowel duration when describing a movie that was *sooo looong*, or perhaps deepening her voice when describing the movements of a large, lumbering animal. Yet, although scholars occasionally make reference to these phenomena (particularly as they relate to gesture, e.g., Emmorey, 1999; Liddell, 2003; Okrent, 2002), they have been the subject of little empirically study, especially regarding their occurrence within naturalistic discourse. Thus, the present study aimed to document whether people regularly produce iconic vocalizations when they talk, and additionally, to explore how these vocalizations are integrated within an utterance.

Guiding the study, iconic vocal behaviors are understood within a modality-independent framework of gesture, which is established in the first part of this paper (Emmorey, 1999; McNeill, 1992; 2005; Okrent, 2002). Then, after reviewing the limited empirical study on "vocal" gesture, I report findings that show that people frequently vary their speech rate in correspondence with the speed of an event they are describing. These findings are discussed with respect to implications for understanding how people integrate iconic speech rate within their linguistic utterance, as well as the cognitive processes that drive the production of iconic speech rate during communication. On this second point, I put forward the hypothesis that vocal gestures arise as a physical manifestation of the sensorimotor-based simulations that people enact as they communicate (cf. Hostetter & Alibali, 2008, which considers exclusively the action-based simulations that create manual gestures). I suggest that people do not just use their voice as a means to channel arbitrarily symbolic language; but rather, vocalizations share a fundamental role along with manual gesture in structuring the embodied basis for how people actually understand and think about imagery-laden concepts such as speed.

1 Vocal Gesture

The sort of iconic vocal behaviors described above share many essential similarities to manual iconic and metaphoric gesture (Okrent, 2002). To illustrate this point, it is useful to consider them with respect to some of the prototypical properties of gesture as adapted from McNeill (1992, 2005). However, in making this comparison, it is critical at the outset to keep two principles in mind (see also Okrent). First, it is important to note that the properties represent the prototypical gestural end of a continuum, with prototypical linguistic properties on the other end. Thus, we are not looking to

make absolute distinctions, but rather, wish to show that iconic vocal phenomena can be aptly treated within this framework and generally exhibit qualities that closely match those of gesture. The second principle to keep in mind is that the set of properties that distinguish gesture from more linguistic forms of communication are, in theory, wholly independent of modality. Just as language can manifest in sign, gesture can manifest in vocalizations.

McNeill (1992: 12) demonstrates the properties of gesture with the following example of a man describing a scene from a comic book story in which a character bends a tree back to the ground. The speaker describes the event:

And he *bends it way back*

As he utters the words "bends it way back," the speaker gestures by using his right hand to demonstrate grabbing the tree, and then pulling his hand back, from high and in front of his body to back and down near his shoulder.

According to McNeill, one way gestures are distinguished from linguistic elements is by their structural form and their relation to other gestures. This is exhibited by three basic properties. Gestures are global, meaning that their form is primarily interpreted as a whole, rather than as a combination of morphological parts. In the example above, the parts of the gesture—for instance, the grab and the pull-back—are only interpretable after understanding the gesture's full form. In contrast, the linguistic meaning of the event requires the combination of morphological elements. Gestures are also synthetic—a single gesture can convey many meanings. The single gesture above expresses a meaning for which language uses four words to express. And finally, gestures are noncombinatoric and thus cannot be combined together to create a larger hierarchical meaning. Whereas the words above combine into a verb phrase, the parts of the gesture do not comprise any such organizational structure.

A second way gestures are distinguished from linguistic communication relates to how they express meaning. Unlike linguistic elements, gestures are not conventionalized and have no standards of form. Instead they are expressive by their iconic formal similarity to their referent. In our example, for instance, the speaker's gesture appears to imitate the act of bending the tree back. Related to this point, gestures exhibit meaningful gradience in their form. The speaker could have made subtle changes to the meaning of his gesture by making the movement the other way across his body, or faster, or at a sharper angle, etc.

A final defining set of properties has to do with gesture's relationship to speech. Gestures tend to be semantically coexpressive with speech and depend on context to be meaningful. Without the words to contextualize it, one would be hard-pressed to make sense of the speaker's bending-back gesture. McNeill also observes that the semantically coexpressive elements of gesture and speech tend to be synchronous in time; the bending-back gesture above was produced just as the speaker articulated the semantically-related words "bends it way back."

Together these modality-independent qualities characterize prototypical gesture. Now, for comparison, let us consider an example of the iconic use of speech rate. The following utterance was spoken to a person in an adjacent room to explain the source of a loud bang—a sound caused by a screen door blowing open by the wind and then slamming shut after the drop in pressure.

> The door did that thing where it *openns reeeally sloowly and then slamsshut*.

In this example, the speaker first slows and then increases his rate of speech to depict a change in the speed contour of the door-slamming event. The first verb phrase "opens really slowly" is stretched in duration to depict the slow manner in which the screen door was blown open. In contrast, the subsequent verb phrase "slams shut" is contracted to depict the high velocity with which the door slammed close, causing the bang that was heard in the next room.

So how does this use of speech rate measure up against the first set of form-related gestural characteristics? It appears to be global in that its meaning comes from tracing the holistic time contour of the event and is not built from any combination of parts. It is also synthetic—the single speech rate contour conveys multiple meanings including manners of movement and degrees of speed, whereas the linguistic message requires several morphological elements to express this information. Third, it is noncombinatorial, lacking any kind of hierarchical organizational structure.

With respect to the ways gesture expresses meaning, this use of speech rate also exhibits qualitative similarities. It is relatively unconventional, without any apparent regularities in use or form (but see Okrent, 2002). It is also iconic and gradient. Meaning is conveyed by a graded contrast in speech rate and correspondence with the temporal contour of the described event, and not by categorical interpretations of 'slow' and 'fast'.

Lastly, this use of iconic speech rate is clearly semantically coexpressive and temporally synchronous with speech, and furthermore depends on

the context of the speech to be meaningful. The speech rate modulations occur precisely as the speaker is describing each related component in the temporal sequence of the event, and would not be interpretable without the synchronous speech to contextualize them.

Because of these essential similarities to McNeill's notion of gesture, in the remainder of this paper, I refer to iconic vocal phenomena such as those described above, without qualification, as a fully fledged form of gesture.

1.1 Empirical Study of Vocal Gesture

While scholars have occasionally made reference to vocal gesture as exemplifying the principle of modality independence when considering the status of gestural versus linguistic communication (E.g., Emmorey, 1999; Liddell, 2003; and see Okrent, 2002 for more detailed treatment), actual direct empirical investigation has been limited, amounting to only a handful of studies by Shintel and colleagues.

In the first of these studies, Shintel, Nusbaum, and Okrent (2006) investigated the use of speech rate and pitch to express information about speed and verticality. In one experiment, participants watched an animated dot move either left or right at different speeds across a computer screen, and were instructed to use the sentences *It is going left* or *It is going right* to describe its motion. It was found that the mean utterance duration was significantly shorter for fast animations than for slow animations. Moreover, when these utterances were replayed for listeners, it was found that listeners, presumably using information conveyed by speech rate, were significantly better than chance at guessing whether the speaker had watched a fast or slow animation.

In a second complementary experiment, Shintel and colleagues looked at whether people use fundamental pitch to convey information about verticality. Participants watched an animated dot move up or down and were instructed to describe its direction by saying *It is going up* or *It is going down*. The findings revealed that people used a higher fundamental frequency for upward compared to downward-moving dots. According to Shintel and colleagues, these results provide evidence for what they call "analog acoustic expressions", an analog prosodic feature of speech that communicates propositional information about external referents.

Two additional comprehension studies focused specifically on how the speech rate phenomenon is understood by listeners. In the first, participants listened to sentences spoken at varying speaking rates followed by a visually presented picture of an object, and had to make speeded judgments of whether the pictured object had been mentioned in the sentence (Shintel

& Nusbaum, 2007). The critical manipulation was whether the mentioned pictures were in motion or at rest. For example, a participant might hear the sentence "The horse is brown" followed, in the matching condition, by a picture of a brown horse that was either running or standing still. Shintel and Nusbaum found that participants were faster to verify that the pictured object had been in the sentence when the motion implied by the picture matched the spoken rate of the sentence—a fast rate with a moving horse, a slow rate with a stationary horse, for instance. Based on this observation, the authors conclude that people integrate speech rate information into an online perceptually-based simulation that constitutes the meaning of the sentence.

In a second follow-up study, Shintel and Nusbaum (2008) looked at whether speech rate influences listeners independently of context, or alternatively, whether the effect occurs most strongly when it is contextually relevant. Participants read short scenarios ending either with or without an implication of urgency, and after each, they heard a recorded instruction to press a key, spoken at a fast or slow rate. The results showed that participants were faster to respond to the fast key press instruction specifically after having read the scenarios implying urgency. The finding suggests that speech rate influences listeners not through some automatic cross-modal speed-matching process, but instead functions like other communicative aspects of prosody, such as those related to syntax or discourse.

2 Study

Though interesting, it is unclear how the work of Shintel and colleagues relates to the use of vocal gesture within natural discourse, which is the focus of many manual gesture studies. Thus the present study aimed to systematically document and analyze the use of vocal gesture within a relatively unconstrained speech production task. The goal was first to see whether people would regularly produce vocal gestures within natural discourse, and second, to gain some understanding of how these gestures are integrated with the spoken language of an utterance. Participants watched video clips of various fast or slow-paced events and described them to an experimenter. The descriptions were recorded and later analyzed for the occurrence of speech rate effects that matched the speed of the described event. It was predicted that people would talk fast—i.e., increase their speech rate—when describing the fast events, and talk slow when describing the slow ones. More specifically, it was predicted that iconic speech rate would function like manual gesture with respect to how it is integrated

semantically and temporally within an utterance. Since manual gestures tend to be semantically linked to concurrent speech, vocal speech rate gestures ought to function the same way. Based on this reasoning, speech rate gestures related to the speed of an event were predicted to cooccur with spoken linguistic expressions of speed, such as during the production of speed-related adverbial phrases.

2.1 Method

2.1.1 Participants

45 University of California, Santa Cruz undergraduates participated in the study in exchange for course credit. (However, as will be described below, data was used only from those 25 participants who made explicit mention of speed in their descriptions.) They were all native speakers of English.

2.2.2 Materials

The video stimuli consisted of trimmed video clips selected from the website http://www.youtube.com. Ten event themes were chosen (including *dog running*, *man juggling*, *couple dancing*, *girl bouncing on a ball, robot walking, man smoking, fish swimming, man eating, car driving, and man/woman boxing*), and videos were selected in pairs, with one item of the pair showing a fast version of the theme, and the other item a slow version. For example, in one fast video, a greyhound was shown racing around the beach with a ball in its mouth, whereas in the corresponding slow video, a large, shaggy dog was shown clumsily trotting along the beach. The clips were each edited to be within 10 to15 seconds in length, and then combined together into a single long movie of all 20 clips. Between each clip, 15 seconds of black screen was inserted to provide time for participants to describe the previous clip. The video clips were combined together into one of four counterbalanced orders, which were presented in equal proportions to participants. Videos of the same event theme were spaced maximally apart and the clips alternated in speed through the movie. (Despite the alternation between fast and slow videos, participants were generally unaware that speed was a factor of the study. This was indicated in postexperiment interviews, as well as evidenced by the fact that only about 20% of participants' descriptions made explicit mention to speed.) All four movies began first with two practice clips.

2.2.3 Procedure

Participants were seated by an experimenter at a table in front of a 16" laptop computer on which they watched the movie. They were told that they would watch and describe a series of video clips. To make the interaction more purposeful and natural, the experimenter explained that she had her own task to do, which was to check off each video from a list as it was described by the participant. The experimenter said that she was unfamiliar with the videos and would check off each video based on the participant's description.

After seating the participant, the experimenter sat down across the table a few feet away and gave spoken instructions. Seating position was selected so that the participant could easily make eye contact and talk to the experimenter without being obstructed by the computer screen. When the participant verified that she understood the task, she was told to click on the play symbol with the mouse, and the movie began (the movies were played with Windows Media Player). Throughout the movie, the experimenter acted out her end of the task, actively listening and, after each description, checking off an item from her list. The descriptions were recorded with a flat table microphone which was connected to a digital recorder and placed inconspicuously on the table about two feet in front of the participant.

2.2 Analysis and Results

Below are some representative descriptions produced by participants during the task. As will be described, analysis focused only on descriptions making explicit reference to speed, and so this is reflected in the examples:

> Umm some guy hitting a punching bag like really fast. Like it looks like in a boxing gym or something.

> This clip showed a fish, in what I can only assume to be like the natural environment not an aquarium, and it was swimming very slowly, and it was it wasn't a tropical fish like before it was- I don't know how to describe it.

> Uhh, someone was sitting at a table and there was some type of robot slowly walking across, but you couldn't see the person's face.

> Someone driving very fast in a suburban area taking a left turn, and then driving down a side street until making a stop.

The quantitative analysis consisted of two primary comparisons: A comparison between the overall spoken rate of descriptions for fast versus slow videos (measured in syllables per second), and a more specific analysis comparing the spoken rate of speed adverbial phrases (measured more precisely in phones per second). As discussed above, this second speed adverbial analysis was motivated by previous empirical findings that gestures tend to be semantically and temporally coexpressive with speech (McNeill, 1992; 2005).

All speech rate measurements, including the determination of speech onset and offset boundaries and the durations between them, were made using Praat phonetic analysis software. Statistical tests in both analyses consisted of within-subject paired t-tests comparing speech rates between the fast and slow video conditions.

2.2.1 Overall Description Rates

Analysis of overall description rates was conducted only on descriptions which made explicit mention of speed, thus assuring that speed was a salient aspect of how the participant perceived the event. Additionally, since the analysis was within-subject, only the descriptions from participants who referred to speed in both the fast and slow conditions could be used. This resulted in data from 25 participants, a total of 142 descriptions, 75 for fast videos and 67 for slow. Each participant produced a mean of 3.00 fast descriptions and 2.68 slow descriptions, with standard deviations of 1.58 and 1.68 respectively. All 20 video clips were represented in the descriptions.

Speech rates were measured in syllables per second. Utterance duration was calculated from the onset of any vocalization related to the description until the point of silence when the participant was finished. Mean description length was 7.83 seconds with a standard deviation of 2.58, with no difference in utterance duration between conditions. Syllables were counted based on written transcriptions of the descriptions.

Analysis showed that descriptions of fast events were spoken at a faster rate than descriptions of slow events, with means of 4.26 syl/sec and 3.91 syl/sec and standard deviations of 0.92 and 0.62, respectively. A paired t-test showed this to be a significant difference ($t(24) = 2.27$, $p = 0.033$).

2.2.2 Speed Adverbial Rates

This analysis focused on adverbial phrases which made reference to speed following the verb. Thus the analysis included adverbials in utterances like "He was throwing punches at a bag *really fast*" or "This was a black dog just kind of running *slowly* around on the beach." Comparatives such as

"faster" or "more slowly" were not included. Again, within-subject analysis required that participants produce speed adverbials in both conditions for their data to be included. This resulted in data from 18 participants and a total of 94 descriptions, with 58 for fast videos and 36 for slow. Individual participants produced a mean of 3.21 fast adverbials and 2.05 slow adverbials with standard deviations of 1.51 and 1.13. Again, all 20 video clips were represented in the descriptions.

Because adverbial phrases generally consisted of only a few syllables, a more precise measurement of phones per second was used. So, for example, the phrase "super fast" was counted as 9 phones. In addition, to compensate for the high frequency with which participants produced the root word "slow" and its relatively long open syllable, the /o/ in these cases was counted as 1.5 phones. Thus, the phrase "really slowly" would be counted as containing 10.5 phones. Adverbial phrase boundaries were determined using phonetic cues such as the onset and offset of voicing, frication, and formant transitions.

The analysis found that fast adverbials were spoken significantly faster than slow adverbials, with means of 12.86 phones/sec and 10.54 phones/sec and standard deviations of 2.48 and 2.10, respectively (paired t-test, $t(17) = 3.19$, $p < 0.01$).

3 Discussion

The first aim of the study was to gain a general sense of how frequently people would produce vocal gestures when describing the fast and slow events shown in the videos. Specifically, how often would people increase or decrease their rate of speech when describing fast versus slow events? The second aim was to investigate more closely *how* people manipulate their speech rate to gesture about speed. For example, how are vocal gestures integrated within the spoken language of an utterance? Two relevant findings were made. First, people appear to modulate their overall rate of speech in correspondence with video speed throughout a particular description. Second, they also specifically alter the spoken rate of adverbial phrases used to describe the speed of an event, such as when they say "real slow" or "very quickly".

To interpret these two findings, it is first necessary to consider whether the difference in overall description rate is driven by the difference in adverbial rate, or whether they are two separate effects. Given the relatively short duration of the speed adverbial phrases in comparison to the overall duration of the full descriptions, it is unlikely that this second effect is the basis for the first. (The adverbial effect size amounts to roughly 0.15

seconds in comparison to a mean utterance duration of 7.83 seconds.) Thus, it appears that people are producing two different forms of speech rate gestures throughout their descriptions.

A second important question is whether these differences in speech rate between fast and slow descriptions are large enough to be communicatively meaningful. A useful starting point is to consider the just noticeable difference in speech tempo, estimated at 5% (Quené, 2007). In comparison, the effect sizes found in this study were 8.6% for overall utterance rate and 20.0% for speed adverbial rate—differences that certainly exceed the minimum criteria to be perceptually noticeable.

3.1 The Coordination of Vocal Gesture with Speech

In principle, as described above, gesture is characterized by a set of qualities that are independent of modality (McNeill, 1992; 2005; Okrent, 2002), and indeed we have seen that people appear to gesture quite naturally vocally. Yet clearly, there exist certain physical constraints when coordinating vocal gestures with speech that do not apply to the coordination of manual gestures. With manual gesture, the hands are free to function independently as the vocal apparatus articulates speech. Thus, though a gesture is constrained by the conceptualizing processes shared with linguistic communication, it is physically unfettered. In contrast, vocal gestures must be coordinated with and actually physically integrated with the articulatory movements involved in producing the linguistic stream of speech. How does this integration happen?

The finding in this study that people produce two different forms of speech rate gestures offers some initial ground for understanding this integration process. In one case, people modulate their speech rate through a relatively general and wide portion of their utterance. They appear to demonstrate the overall pace of the event with their rate of speech, more-or-less independently of the detailed semantics being directly expressed concurrently in the linguistic signal. In comparison, the second effect consists of a more specific change in the spoken rate of semantically-related speed adverbials. Here the gesture and linguistic semantics are tightly integrated together as the expression of speed is simultaneously expressed through both gesture and language.

Questions relating to more specific details of how vocal gestures take form and are integrated with speech remain to be addressed experimentally. For example, with speed adverbials, specifically which phonetic segments are subject to lengthening or shortening? Informal investigation of the "really" speed adverbial constructions produced in this study (e.g., "really quickly", "really slowly", etc.) reveals this to be quite variable. For

example, in some instances there was extension or contraction on the vowel of "really" without any apparent change to the speed adverb. In other instances the effect was on the speed adverb with little effect on "really." This was demonstrated in one case by a speaker describing a girl bouncing slowly on Yoga ball:

> There's another girl bouncing on a ball but really *ssssloowly*…

Interestingly, in this utterance, the word "slowly" is extended both across the initial frication of the /s/ as well as across the vowel. In fact, this sort of idiosyncratic variability is exactly in line with the characteristics of gesture described by McNeill. Yet, in this instance we also see evidence of the physical constraints on coordination as a consequence of shared expression through a single modality. A particular vocal gesture is, by necessity, formally constrained as it is incorporated with the phonetic segments of the utterance.

Particularly with respect to the speed adverbial effect, we might also ask whether vocal gestures are marked in any special way, such as with a pause or a quotation device such as "like", to indicate a transition in communicative style (cf. Clark & Gerrig, 1990). For at least some of the data from this study, this appears to be the case. Consider the following description of a man eating a cheeseburger at a fast food restaurant.

> This one's a guy at McDonalds. He's eating a cheeseburger, like, *reallyfast*. He just like downs the whole thing.

After the first "like" the speaker leaves a short pause before saliently speaking the phrase "really fast" at an increased rate. In a description by a different speaker, only pauses were used:

> This one's like a man playing fetch with his dog on the beach and the dog is running around, *superfast,* with a red ball in its mouth.

Here the short pauses separating the speed adverbial phrase appear to mark the vocal gesture.

We may also gain insight from researchers of sign language, who have noted an analogous problem for signers. For example, Emmorey (1999) observes that, as signers' hands are involved in producing a linguistic utterance, there are constraints on bimanual coordination and motor resources that prevent signing on one hand while gesturing with the other. According to Emmorey, one way that signers gesture within these constraints is by alternating sign and gesture, using what Clark (1996) refers

to as component gestures. In this process, a signer might temporarily break from signing, then gesture, and then continue back to signing, with the gestural meaning incorporated into the utterance. (Clark and Gerrig (1990) also offer examples of this sort of alternation between speech and manual gesture.) Though strictly speaking, component gestures, due to their independence from concurrent language, fall towards the linguistic end of the gesture-language continuum, it is nevertheless worthwhile to consider these model cases of how same-modality constraints are worked out during communication. For example, it is interesting to note the similarities between the use of component gestures and the phenomenon described above in which the speaker uses pauses to mark vocal gestures within speech.

It is additionally interesting to observe that sign languages commonly integrate communication with more gesture-like properties into the linguistic signing system, such as with the use of classifier constructions of American Sign Language (Emmorey & Herzig, 2003; Liddell, 2003). In these constructions, handshape takes on the categorical, conventional form associated with a linguistic morpheme, yet their locations within signing space are gradient like gesture. A similar mixed system dictates the use of pronouns in ASL: Morphological handshape distinguishes case, orientation distinguishes person, and movement distinguishes number, yet these signs exhibit gradience in how they are directed towards their referent within signing space. A spoken parallel may be seen in the combination of a morphological word form with gradient gesture-like qualities manifesting in the prosody with which it is articulated.

Finally, a last question that remains to be examined is how vocal gestures, in addition to coordination with speech, are also executed in multimodal coordination with manual and other bodily gestures. One might predict, for example, that as imagistic aspects of a message are sufficiently activated to manifest as vocal gestures within speech, they are also likely to manifest in manual gestural movements (cf. McNeill's (1992, 2005) notion of the growth point). Thus, in the case of iconic speech rate, we might expect to observe faster and slower manual gestures in synchrony with speech rate modulations as the speaker describes the speed of an event. Furthermore, extending this point to an experimental perspective, relevant manual gestures may even be useful to predict when vocal gestures are likely to occur, thereby increasing the power of studies to identify and describe vocal gestures within natural discourse.

3.2 Vocal Gesture as Embodied Simulation

In this final section of discussion, I present a hypothesis for the underlying cognitive process that leads people to produce iconic speech rate gestures

as they describe various fast and slow paced events. This point is particularly interesting because, for the most part, the vocal gestures do not seem to be produced with any special deliberateness. They are not, for instance, simply a product of elaborate vocal pantomimes or part of a purposefully crafted chorus to a Garth Brooks song. On the contrary, they seem most often to manifest as an ordinary part of the process of talking. Just as people are not aware of the extent to which they use their hands to iconically depict the subject of their speech, they seem similarly unaware that they also iconically represent their subject matter with the pattern of their voice.

My proposal is that vocal gesture arises so naturally as a part of spoken communication because it is a physical consequence of how people conceptualize action and other bodily-based domains as they communicate. A growing collection of experimental evidence shows that people make sense of language through the enactment of embodied, sensorimotor-based simulations created during online language processing (Glenberg & Kaschak, 2002; Stanfield & Zwaan, 2001), including metaphorical language (Bergen, 2005; Gibbs, 2006). Building on this, a recent hypothesis suggests that these sorts of sensorimotor simulations also provide the mechanism that leads to the production of representative manual gesture (Hostetter & Alibali, 2008). In their Gesture as Simulated Action framework, Hostetter and Alibali hypothesize that simulations of action-related thoughts lead to activation of neural premotor action states, which then has the potential to spread to motor areas. This spreading activation is realized as the overt action of gesture.

Hostetter and Alibali's exclusive focus on simulated *action* as the basis for representative gesture is questionable, and gesture scholars such as McNeill argue more broadly that sensorimotor imagery, including but not limited to action, forms the semantic basis of gesture. Still, it might be maintained that people activate cortical areas related to bodily movements and postures as part of the process of conceptualizing and communicating about a wide range of imagery-laden domains. Importantly, the findings from this study suggest that the embodied simulations enacted during communication additionally spread to motor areas related to vocalizing. Thus, given the characteristic iconicity of vocal gesture, it follows that the semantics of gestural communication—that is, how we use our bodies to understand and represent actions in the external world--is in part based on actions of the vocal tract, in addition to the hands.

With this perspective in mind, we return again to the two forms of vocal gesture reported in this study. First, recall the finding that people specifically alter their speech rate as they articulate adverbial phrases about the speed of the event. As discussed above, in this case vocal gesture func-

tions similarly to prototypical examples of manual gesture, occurring in semantic and temporal synchrony with concurrent speech. The gesture emerges precisely as the speaker is conceptualizing and communicating about speed as the profiled aspect of her message. In terms of a gesture as embodied simulation framework, the speed-related simulation processes that drive linguistic production of the adverbial phrase simultaneously lead to iconic vocal expression of speed manifesting in the speaker's speech rate.

In contrast, consider the other form of vocal gesture that was found, which differs from the previous more prototypical case. This form consists of a more general increase in speech rate throughout a wider description of the event—an overall shift in the rate at which the description is spoken. In this case, while speed is not actively profiled in language production, it appears that the speaker is more generally engaged with the pace of the action as she proceeds through her description. It may be argued, in this instance, that the speaker's speech rate reflects a backgrounded simulation of the event, as she scans and profiles specific details to highlight with linguistic communication (cf. Langacker, 1990). Though the speaker's attention is not focused enough to produce particular speed-related linguistic constructions and gestures, it is sufficient to more generally influence her overall rate of speech. Thus the wider modulation of speech rate might be considered as a vocal manifestation of simulated background.

4 Conclusion

It is fairly obvious that people occasionally use their voice for iconic communication, as demonstrated, for instance, by the introductory examples taken from popular country music. Yet, despite its commonplace appearance, there has been only limited empirical study directed towards understanding how vocal gesture functions within speech and especially within more natural discourse. In this paper, I have made the case that such study is critical to our understanding of the relationship between speech, gesture, and language and for informing theories of how the body and sensorimotor system contribute to cognitive processes related to conceptualization and communication. The preliminary findings reported here on the use of speech rate suggest that people, as with manual gesture, very naturally produce iconic vocalizations as they talk about imagery-grounded domains such as speed. In his seminal book *Hand and Mind,* David McNeill writes that "Gestures are like thoughts themselves" (1992: 12). Thus, it appears then that just as thoughts are rendered visible by the hands, they are also rendered acoustically by the voice.

References

Bergen, B. 2005. Mental simulation in literal and figurative language understanding. *The Literal and Nonliteral in Language and Thought.* Eds. S. Coulson & B. Lewandowska-Tomaszczyk. 255-280. Berlin: Lang.

Clark, H. 1996. *Using Language.* Cambridge University Press.

Clark, H. & Gerrig, R. 1990. Quotations as demonstrations. *Language 66: 764-805.*

Gibbs, R. W. Jr. 2006. Metaphor interpretation as embodied simulation. *Mind & Language 21: 434-458.*

Glenberg, A. & Kaschak, M. 2002. Grounding Language in Action. *Psychonomic Bulletin & Rewview 9:* 558-565.

Emmorey, K. 1999. Do Signers Gesture? *Gesture, Speech, and Sign,* eds. L. Messing & R. Cambell. 133-159. New York: Oxford University Press.

Emmorey, K. & Herzig, M. 2003. Categorical Versus Gradient Properties of Classifier Constructions in ASL. *Perspectives on Classifier Constructions in Sign Languages,* ed. K. Emmorey. 221-246. Mahwah, NJ: Lawrence Erlbaum Associates.

Hostetter, A. B. & Alibali, M. W. 2008. Visible Embodiment: Gestures as Simulated Action. *Psychonomic Bulletin and Review 15: 495-514.*

Kendon, A. 1988. How gestures can become like words. *Cross-Cultural Perspectives in Nonverbal Communication.* Ed. F. Poyatos, 131-141. Toronto: Hogrefe.

Langacker, R. W. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar.* Berlin/New York: Mouton de Gruyter.

Liddell, S. K. 2003. *Grammar, Gesture, and Meaning in American Sign Language.* Cambridge: Cambridge University Press.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

McNeill, D. 2005. *Gesture and Thought.* Chicago: University of Chicago Press.

Okrent, A. 2002. A Modality-Free Notion of Gesture and How It Can Help Us with the Morpheme Vs. Gesture Question in Sign Language Linguistics. *Modality and Structure in Signed and Spoken Language,* eds. R. P. Meier, K. Kormier, & D. Quinto-Pozos. 175-198. Cambridge: Cambridge University Press.

Quené, H. 2007. On the Just Noticeable Difference for Tempo in Speech. *Journal of Phonetics 35: 353-362.*

Shintel, H. & Nusbaum, H. C. 2007. The Sound of Motion in Spoken Language: Visual Information Conveyed by Acoustic Properties of Speech. *Cognition 105: 681-690.*

Shintel, H. & Nusbaum, H. C. 2008. Moving to the Speed of Sound: Context Modulation of the Effect of Acoustic Properties of Speech. *Cognitive Science 32: 1063-1074.*

Shintel, H., Nusbaum, H. C., & Okrent, A. 2006. Analog Acoustic Expression in Speech Communication. *Journal of Memory and Language 55: 167-177.*

Stanfield, R. A. & Zwaan, R. A. 2001. The Effect of Implied Orientation Derived from Verbal Context on Picture Recognition. *Psychological Science 12:* 153-156.

Acknowledgements

Many thanks to David Garner for contributing his musical expertise. I am also grateful to Ray Gibbs for his very helpful comments on a previous draft of this paper, as well as to two anonymous reviewers for their helpful comments. And finally, Natalie Benitez deserves special mention for her tireless enthusiasm as she listened intently to the same video clip descriptions, over and over and over…