

Analog acoustic expression in speech communication

Hadas Shintel^{a,b,*}, Howard C. Nusbaum^{a,b}, Arika Okrent^a

^a Department of Psychology, The University of Chicago, 5848 S. University Ave Chicago, IL 60637, USA

^b Center for Cognitive and Social Neuroscience, Department of Psychology, The University of Chicago, 5848 S. University Ave Chicago, IL 60637, USA

Received 9 November 2005; revision received 8 March 2006

Abstract

We present the first experimental evidence of a phenomenon in speech communication we call “analog acoustic expression.” Speech is generally thought of as conveying information in two distinct ways: discrete linguistic-symbolic units such as words and sentences represent linguistic meaning, and continuous prosodic forms convey information about the speaker’s emotion and attitude, intended syntactic structure, or discourse structure. However, there is a third and different channel by which speakers can express meaning in speech: acoustic dimensions of speech can be continuously and analogically modified to convey information about events in the world that is meaningful to listeners even when it is different from the linguistic message. This analog acoustic expression provides an independent and direct means of communicating referential information. In three experiments, we show that speakers can use analog acoustic expression to convey information about observed events, and that listeners can understand the information conveyed exclusively through that signal.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Language production; Spoken language comprehension; Prosody

There is general agreement regarding the properties that make language special as a communication system (cf. Hockett, 1960). One of the important defining characteristics of language is that utterances are constructed using conventional combinations of discrete symbols that have an essentially arbitrary relationship to meaning (see also Saussure, 1959). Meaning is not generally reflected in the sound patterns of words. For example, neither *DOG* nor *CHIEN* has a transparent relationship to their referent. Although the physical transition between different speech sounds can be gradual and continuous, the basic units of speech are perceived categorically as belonging to one phonetic category, rather than

continuously as a graded signal that can lie between different phonetic categories (Liberman, Harris, Hoffman, & Griffith, 1957). However, communication can also be based on the use of non-symbolic analog (or iconic) signs (cf. Deacon, 1997; Pierce, 1932). In contrast to symbolic signs, analog signs consist of non-digital continuous signals in which the pattern properties of the signal correspond in some way to the information conveyed by the signal. Examples for the use of this kind of analogical mapping between the intended message and pattern properties of the signal can be found in animal communication (Templeton, Greene, & Davis, 2005; Von Frisch, 1967) as well as in the gestures that accompany speech (Goldin-Meadow, 1999; McNeill, 1992).

Although prosodic properties of speech such as intonation or rhythm vary continuously, they have not been conceptualized as providing the kind of analogical map-

* Corresponding author.

E-mail address: hadas@uchicago.edu (H. Shintel).

ping found in non-linguistic communication systems. Prosody is not viewed as directly conveying information about objects and events in the world. Research on prosody has mainly focused on the information it conveys about the speaker's internal state (e.g., emotion, Banse & Scherer, 1996; Cosmides, 1983; Scherer, Ladd, & Silverman, 1984; or attitude, Bryant & Fox Tree, 2002; Rockwell, 2000), intended syntactic structure (e.g., Beach, 1991; Carlson, Clifton, & Frazier, 2001; Ferreira, 1993; Snedeker & Trueswell, 2003; Steedman, 1991), or discourse structure and function (e.g. Birch & Clifton, 1995; Bock & Mazzella, 1983; Dahan, Tanenhaus, & Chambers, 2002; Pierrehumbert & Hirschberg, 1990; Terken & Nootboom, 1987). Any function that prosody has for modifying the linguistic message and for conveying semantic and referential information has generally been regarded as deriving from its role in communicating these other kinds of information. For example, prosodic information can facilitate reference resolution due to the role of prosodic stress in marking information that is new in the discourse context (Dahan et al., 2002). Similarly, an intonation contour can convey information about the propositional attitudes of the speaker, such as uncertainty (Hirschberg & Pierrehumbert, 1986), and information about the speaker's metacognitive states can be used in reference resolution (Barr, 2001). However, the analogical variation of acoustic properties has not generally been viewed as directly mapping conceptual information onto speech or as directly conveying information about external referents. In fact, discussions of the role of speech in conveying semantic and referential information have focused on the conversion of the continuous speech signal into perceived discrete linguistic units, rather than on the role the analogical properties of the continuous speech signal may have in conveying such information.

Although rarely discussed in the psycholinguistic literature (but see Okrent, 2002), casual observation suggests there are situations in which speakers make use of more sign-like analog acoustic expressions in speech that map meaning onto speech sound patterns in a continuous way. For example saying "Baseball should be *morelikethis* [spoken rapidly and run together] and less...like...this [drawn out with deliberate pauses]" demonstrates that baseball is typically a slow game and should be faster. This message depends on continuous variation of acoustic properties rather than on the specific choice of words and linguistic structure.

This indicates that speakers can sometimes vary acoustic properties to express meaning in speech, but such observations often suggest a special performance or special circumstances (such as using exaggerated prosody while addressing infants) rather than normal communication. If, on the other hand, analog expression is a routine part of speech, even with no intent to dramatize a description, this would suggest a basic and impor-

tant dimension of communication has generally been overlooked by psycholinguistic research. Moreover, this means of expression may link speech to non-linguistic communication in a fundamental way suggesting greater evolutionary continuity than has generally been inferred on the basis of more traditional consideration of linguistic structure alone (cf. Hauser, 1997).

The present studies were carried out to test whether analog acoustic expression is spontaneously used as a natural part of speech and whether the information conveyed by analog acoustic expression is understood by listeners. In Experiment 1, we evaluated whether speakers spontaneously use analog acoustic expression when describing vertical direction of motion by mapping fundamental frequency (high vs. low) to direction of motion (up vs. down). In Experiment 2 and 3, we evaluated whether speakers use analog acoustic expression to convey information that is independent of the information conveyed in words, and whether listeners can understand information conveyed exclusively through analog acoustic expression.

Experiment 1

Previous research has shown that people associate variation in the pitch of a tone with vertical location (e.g., Bernstein & Edelstein, 1971; Melara & O'Brien, 1987). Classification of a target stimulus was facilitated by the presentation of a congruent-frequency sound (high location and high-frequency tone) and inhibited by the presentation of an incongruent-frequency sound. A similar cross-modal effect has been demonstrated between pitch and the words *high* and *low* (Melara & Marks, 1990a, 1990b). Thus, the fundamental frequency (the acoustic correlate of pitch perception) of a talker's voice can provide a means for expressing direction of motion. However, although such audio-visual cross-modal mappings have been shown to influence the processing of perceptual information, it is not known whether these mappings are functional in conveying semantic or referential information in everyday linguistic communication. Moreover, words such as *UP* or *DOWN* are readily available for speakers when describing an upward or a downward motion. Given the availability of the lexical items *UP* and *DOWN* for describing motion, will speakers spontaneously use this auditory-visual cross-modal mapping and express direction of motion by varying vocal pitch according to direction? In other words, when a propositional message describes a specific aspect of an event, will speakers spontaneously use analog acoustic expression to express the same aspect?

We instructed participants to describe the direction of motion of an animated dot by saying *It is going up* or *It is going down* (the Animation condition). A second group of participants were instructed to read the

corresponding sentences to assess whether any analog acoustic expression was strictly a function of the physical motion observed or more broadly a means of expressing semantic information (the Sentence condition). The fundamental frequency (F_0) of the final word of each sentence was measured (Talkin, 1995). If speakers use analog acoustic variation even when they make use of the lexical items *UP* and *DOWN*, the F_0 of *up* should be higher than the F_0 of *down*, over and above the intrinsic F_0^1 difference between them. Moreover, if this modulation reflects the meaning of the terms *UP* and *DOWN* and the semantics of their contrastive use rather than being driven just by an analog mapping between a visual stimulus and a sound response, simply reading the sentences could produce the same effects as describing observed motion.

Method

Participants

Sixty seven University of Chicago students participated in the experiment. There were 24 participants in each of the test conditions (Animation vs. Sentence) and 19 additional participants in the control conditions. All participants had native fluency in American English and no reported history of speech or hearing disorders. Participants were paid for their participation.

Materials

Test stimuli for the Animation condition consisted of animations of a dot moving upward or downward on the computer screen. In each animation a dot appeared in the center of a 22 × 22 cm display frame. It stayed at the center of the display for 2 s, and then moved to the top or the bottom edge of the display over the next 2 s. Additional control stimuli consisted of animations of a dot moving left or right on the screen. These animations were the same as the test animations except for the dot's direction of motion. Test stimuli for the Sentence condition consisted of the written sentences *It is going up* or *It is going down* presented on the screen.

¹ Vowels may differ, other things being equal, in the fundamental frequency (or F_0) with which they are produced (cf. Peterson & Lehiste, 1960). For example high vowels (such as /i/ or /u/) tend to have a higher F_0 than low vowels (such as /a/). These differences may reflect an automatic consequence of articulation (e.g., Whalen, Levitt, Hsiao, & Smorodinsky, 1995) or a deliberate enhancement of the phonetic quality differences between vowels (e.g., Diehl, 1991), but they are not associated with a message-dependent prosodic variation.

Design and procedure

Test conditions

In the Animation condition, participants saw 12 trials of each animation (up and down), presented in random order. Trials were separated by 2 s. Participants were instructed to describe the dot's direction by saying either *It is going up* or *It is going down* as soon as they determined which direction the dot was moving. In the Sentence condition, participants saw the 12 tokens of each written sentence (*It is going up* and *It is going down*) presented in random order on the computer screen. Each trial began with 2 s of blank display. The sentence appeared in the center of the screen and stayed there for 2 s. Trials were separated by 2 s. Participants were instructed to read the sentence aloud as soon as they saw it.

Control conditions

To control for potential effects of the contrastive experimental paradigm on variation in F_0 height, 12 additional participants participated in a control condition in which they saw an animation of a dot moving left or right on the computer screen and were instructed to describe the dot's direction by saying either *It is going left* or *It is going right*. In all other respects the procedure for this condition was identical to the procedure for the up/down Animation condition. Finally, 7 additional participants participated in a baseline condition testing the effect of the different local phonetic contexts of /p/ vs. /n/ in the words *up* vs. *down*. In principle, the velopharyngeal opening that occurs in the production of a nasal consonant following a vowel should not affect F_0 for the preceding vowel in American English. This supralaryngeal opening only changes the resonant characteristics of the vocal tract and therefore does not directly interact with vocal fold vibration (cf. Stevens, 1999; pp. 190–193 and 303–322 for the acoustic effects of nasalization on speech). However, to investigate the effect of the differing local phonetic context on F_0 , participants were instructed to read the nonsense words *bup* and *bown*, embedded in the carrier phrase *Please say the word ____*. Participants saw 12 tokens of each carrier phrase, in addition to 24 filler items in which two other nonsense words were embedded in the context of the same carrier phrase (this was done to prevent participants from noticing the similarity of *bup* and *bown* to the words *up* and *down* which may have led to an effect of the semantics of the original words on the production of the nonsense words).

Analysis

Speech was recorded with a SHURE SM94 microphone onto digital audiotape. Utterances were digitized at a sampling rate of 44.100 kHz with 16-bit resolution.

F_0 measurements were extracted using an autocorrelation pitch extraction algorithm (Talkin, 1995) with a 75 ms autocorrelation window. For each response sentence we displayed a waveform and a spectrogram of the utterance, and marked the target word (*up* or *down*) using auditory and visual cues. The beginning of *down* was marked at the burst for /d/ and the beginning of *up* was marked at the onset of voicing if there was a pause after *going* and after the final nasal zero in *going* if there was no pause. The end of the word was marked at the first zero F_0 measurement at the end of the utterance. Any zero measurements remaining at the beginning or in the middle of each marked token were cut out in order not to skew the measurement mean.

Results and discussion

Data from one participant in the *bup–bown* control condition were discarded because she did not pronounce *bown* to rhyme with *down*.

For both the Animation and the Sentence conditions, the mean F_0 of *up* (155.4 and 158.3 Hz, respectively) was higher than *down* (149.4 and 148.2 Hz). An ANOVA with Condition (Animation vs. Sentence) as a between-subjects factor and Direction (*up* vs. *down*) as a within-subjects factor revealed a significant effect of Direction, $F(1, 46) = 6.7, p < .02, MSE = 1554.3$. There was no significant effect of Condition or significant Condition by Direction interaction ($F < 1$). The final /n/ in *down* was included in the F_0 computation because we reasoned that the F_0 of any voiced segment represents relevant information. To make sure the results are not due to the F_0 comparison between a vowel in *up* and a vowel-consonant in *down* we reanalyzed the data from 12 randomly selected participants (6 in each Animation and Sentence condition), this time without including the final /n/ in the F_0 computation. Mean F_0 of *down* for these participants was 149.6 Hz when the final /n/ was included in the computation compared to 144.3 Hz when the /n/ was excluded from the computation. A t test conducted on the results for this sub-sample showed that the mean F_0 of *down* differed significantly from the mean F_0 of *up* (164.8 Hz) both when the /n/ was included in the F_0 computation ($t(11) = 2.49, p = .03$) and when it was not included in the F_0 computation ($t(11) = 2.85, p < .02$). These results show that the F_0 difference between *up* and *down* is not due to the inclusion of the /n/ in the F_0 analysis. In fact, the F_0 difference between these lexical items was even bigger when the final /n/ in *down* was excluded from the F_0 computation.

The fundamental frequency difference found here can be compared with prior research indicating that the F_0 of these vowels in neutral lexical context (/hVd/) differs only by about 2.2 Hz (based on the F_0 values for vowels

reported in Peterson & Barney, 1952; and weighted by the intrinsic duration of vowels reported by Peterson & Lehiste, 1960). By contrast, the control sentences, *It is going left* and *It is going right* revealed no significant F_0 differences ($F < 1$) between *left* (156.1 Hz) and *right* (156.4 Hz)², although the intrinsic F_0 difference between the vowels in these words is 3.39 Hz (Peterson & Barney, 1952; Peterson & Lehiste, 1960). This suggests that the F_0 difference observed for *up* and *down* reflects the semantic contrast between these lexical items, rather than just the contrastive nature of the experimental paradigm. In addition, the *bup–bown* control condition revealed no significant F_0 difference. The mean F_0 for *bup* was 130.9 Hz, while the mean F_0 for *bown* was 129.9 Hz when the final /n/ was included in the analysis and 131.4 Hz when the final /n/ was excluded from the analysis. In both analyses, the F_0 difference between the two nonsense words was not significant ($p > .9$).

Although talkers were not instructed to act out or emphasize their speech, when describing motion the contrast in conveyed direction of motion was analogically mapped to a difference in the fundamental frequency of the talker's voice, the acoustic property underlying the perception of pitch height. Interestingly, speakers analogically varied fundamental frequency both when they described an actual visual motion as well as when they read a sentence describing motion. This pattern of results suggests that the F_0 modulation does not reflect just a simple analog mapping between the visual physical motion and the physical sound emitted in response, but a mapping of the meaning of 'up' and 'down' onto the physical speech signal. Furthermore, several recent accounts suggest that language comprehension involves perceptual simulation of the described events (Barsalou, 1999). Specifically, findings suggest that the comprehension of sentences describing movement involves perceptual simulation of the movement described by the sentence (Kaschak et al., 2005; Zwaan, Madden, Yaxley, & Aveyard, 2004). If people routinely activate perceptual representations of motion in reading sentences describing motion, we would indeed expect a similar pattern of results both when people describe a visual motion and when people read a sentence describing such motion. The lack of a significant Condition by Direction interaction in this study is thus consistent with these findings.

The results show that even though speakers were not instructed to act out or emphasize their speech, when reading a sentence describing motion or describing an

² There are large speaker differences in F_0 . Because different speakers participated in the *up–down* condition and in the *left–right* condition, the absolute F_0 values cannot be compared. The critical point concerns the F_0 differences, rather than their absolute values. The results of the *bup–bown* control condition should be likewise interpreted.

actual motion, speakers increased or decreased the fundamental frequency of their speech and created an analog mapping between vocal fundamental frequency and the conveyed direction of motion. These results demonstrate that speakers naturally use analog acoustic expression when talking, even when there is no intent to dramatize a description. Speakers used analog expression even though the motion information they intended to convey could have been conveyed simply by using the appropriate lexical items. This suggests that acoustic variation can be used to emphasize analogically the meaning of a lexical item, particularly when used as semantically contrastive. By exploiting non-linguistic cross-modal mappings within the domain of linguistic communication, speakers can extend the range of possibilities that is available for expressing information in speech beyond those that are afforded by the propositional structure alone.

Experiment 2

Although speakers can use analog variation of acoustic properties to express the same information conveyed by lexical items in a sentence, can analog expression convey information that is independent of the propositional content or does it function just to modulate propositional meaning? If analog acoustic expression is a channel of communication that is truly different from the linguistic-propositional channel, speakers should be able to express information that is different from the information conveyed in words and sentences. Experiment 2 tested whether speakers can indeed use this channel to express information that is independent of the propositional content of the utterance and, importantly, to see whether listeners are sensitive to the message conveyed exclusively through analog acoustic expression. If analog acoustic expression serves a communicative function, listeners should be able to understand the information it conveys, even when this information is not conveyed by any of the lexical items used by speakers.

We instructed five speakers to describe a moving animated dot using the sentences *It is going left* or *It is going right*. In each animation, a dot moved left or right on the screen at different speeds. Speakers were only instructed to describe the dots' direction. However, in principle, speakers could also convey speed information through analog acoustic expression, most obviously by varying speech rate to indicate relative dot speed. To test whether listeners are sensitive to the information conveyed by analog acoustic expression, we presented all recorded utterances to a group of listeners and asked them to judge the speed of the dots observed by speakers. If speakers do indeed convey speed information solely through the use of analog acoustic expression

and listeners understand this information, they should be significantly better than chance at guessing the speed of motion from the utterances.

Method

Participants

Five speakers and twelve listeners participated in the experiment. Data from one speaker were not used due to considerable clipping and noise in the recording. All participants were University of Chicago students with native fluency in English and no reported history of speech or hearing disorders. Participants were paid for their participation.

Materials and analysis

The experimental materials consisted of 36 animations of a dot moving left or right (horizontally or diagonally) on a computer screen at two different speeds. Both diagonal and horizontal movement were used because we did not want speakers to be aware of the purpose of the experiment. By including another dimension differentiating between the displays we hoped to make the speed contrast slightly less salient (subsequent post-experimental questioning revealed that speakers were not aware of the purpose of the experiment). In each animation a black dot appeared in the center of a 22 × 22 cm display frame and then moved to an endpoint at the edge of the frame. The velocity of the dot was kept constant within each of the Speed conditions (11 cm per second for the fast dots and 3.9 cm per second for the slow dots), resulting in different duration of motion for horizontally and diagonally moving dots. Fast dots moved to the endpoint over the course of 1 s (for horizontally moving dots) or 1.33 s (for diagonally moving dots). Slow dots moved to the endpoint over the course of 2.833 s (for horizontally moving dots) or 4 s (for diagonally moving dots). Out of the 36 animations, there were 3 tokens in each speed (fast/slow) × direction (left/right) × orientation (diagonally up/diagonally down/horizontally) combination.

Speech was recorded using a SHURE SM94 microphone onto digital audiotape and then digitized at a sampling rate of 44.1 kHz with 16-bit resolution. Utterance duration was measured from the first clear glottal pulse until the end of aspiration in the final /t/ in "left" or "right." Measurements were done by the first author who was blind to the Speed condition. Utterances were edited into separate sound files beginning with the onset (first glottal pulse) of each utterance. A total of 144 utterances served as the stimuli in the comprehension part of the experiment.

Design and procedure

Speakers watched the 36 animations, presented in random order. Each animation was followed by a blank display for 4 s before proceeding to the next trial. Speakers were asked to describe the direction of the dot's motion by saying either *It is going left* or *It is going right*. Speakers were told they did not need to make any other response and that the experiment would progress to the next trial by itself.

For the comprehension part of the experiment, listeners were presented with the 144 spoken utterances produced by the speakers. Listeners saw a demonstration of 4 (2 fast and 2 slow) dot animations in random order. Following the demonstration, listeners heard the sentences produced by all 4 speakers, for a total of 144 sentences. Sentences produced by each speaker were divided into three 12-sentences blocks. The order of the sentences within each block corresponded to the (random) order in which the sentences were produced. Blocks were presented in a random order. Listeners were instructed to guess whether the talker saw a fast or a slow moving dot and to respond as quickly and as accurately as possible by pressing a response key (marked F or S) with their dominant hand. There was no mention of speech rate or of any cues they could or should use in making their decision. The next trial started only after subjects responded.

Results and discussion

To determine if speakers used utterance duration to convey speed we analyzed the duration of the recorded utterances. As predicted, mean utterance duration was shorter for spoken descriptions of fast- (920 ± 31 ms SEM) compared to slow-animations (1047 ± 32 ms). As production data were collected from only a small number of speakers, we analyzed the duration by items, averaging across the four speakers. Results showed the difference in duration between the two Speed conditions was highly significant, $t(34) = -4.84$, $p < .0001$. Unsurprisingly given the small number of subjects, the difference between fast- and slow-animation sentences did not reach statistical significance in an analysis by subjects, $p < .17$; however the difference was in the predicted direction for three of the four speakers averaged over all their productions. It appears that speakers spoke faster or slower, analogically mapping dot speed to articulation speed, independently of the use of the words *right* and *left* for direction of dot motion. The variation in utterance duration is consistent with the claim that speakers expressed speed information by manipulating speech rate, although speakers may have mapped speed of motion onto other acoustic parameters, for

example, by speaking louder or softer to indicate fast or slow motion, respectively.

Listeners were significantly better than chance in classifying dot speed from the sound of the speech alone (mean accuracy 62.9%, ranging from 58.3 to 67.4% across listeners, $t(11) = 14.12$ $p < .0001$). It is important to note that judgments were carried out on all utterances produced by speakers. We did not cull out any utterances based on possible speech errors or subjective impressions. Moreover, the lexical content of all utterances referred exclusively to direction of motion (left or right). In addition, correctly classified sentences differed acoustically on the basis of duration compared to incorrectly classified sentences. Utterances correctly classified by 75% or more of the listeners (63 out of 144 sentences) were significantly shorter for fast (772 ± 32 ms SEM) than for slow animations (1183 ± 45 ms), $t(61) = -7.07$ $p < .0001$. However for utterances incorrectly classified by 75% or more of the listeners (19 out of 144) the reverse pattern emerged: Utterances were longer for fast (1243 ± 63 ms SEM) compared to slow animations (751 ± 80 ms), $t(17) = 3.78$ $p = .0015$. No difference was found for utterances classified at chance level (14 out of 144; mean fast 904 ± 87 ms SEM, mean slow 875 ± 86 ms, $t(12) = .225$ $p = .83$). Again, these results are consistent with the idea that duration was the acoustic cue used to express speed information.

These results show that speakers convey information about the speed of motion by analogically mapping speech rate to speed of motion. Furthermore, listeners can use the information conveyed exclusively through analog acoustic expression, even though we did not explicitly tell listeners to use speaking rate as a cue for speed and they were free to rely on any property of the speech signal. As in the first experiment, it is important to emphasize that speakers conveyed this information spontaneously; the task did not call for any reference or attention to the speed of motion. Speakers may not have attended to speed consistently and moreover speakers did not know their utterances would be played to listeners prior to completing the experiment. Furthermore, post-experimental questioning revealed that speakers were not aware of the purpose of the experiment or that it tested the relation between motion speed and articulation speed, and thus were not intentionally trying to convey speed information.

Experiment 3

One potential problem with the production part of Experiment 2 is that the speed of object motion in the animation presented to speakers was confounded with the duration of the animations. Because the dots in all of the displays travelled the same distance, the time it took the dots to reach the endpoint differed for

fast-moving and for slow-moving dots. Fast-moving dots remained on the screen for 1–1.33 s, while slow-moving dots remained on the screen for 2.833–4 s. Speakers may have increased their speech rate while describing fast-moving dots because they were trying to produce the descriptions before the dot disappeared. Although speakers were not instructed to finish speaking before the end of the animation, and although each animation was followed by at least 2 s of blank display before the next animation appeared on the screen, the use of the present tense (*It is moving...*) may have led speakers to believe the description should be completed while the dots were actually in motion. In this case, the variation in speech rate would reflect task demands specific to the experimental paradigm rather than analog expression of speed of motion that is a natural part of speech.

To control for the potential effect of animation duration on speakers' speech rate we presented speakers with animations of dots moving continuously on the screen in a fast or a slow speed for a duration of 3 s. This duration was held constant across the two Speed conditions (fast and slow). Thus, although the speed of motion of the dots was different in the different displays, the duration of motion was the same for all displays. Moreover, for both conditions the duration of motion was considerably longer than the mean duration of utterances in Experiment 2 so that speakers would have enough time to complete the descriptions before the dots disappeared. Speakers were instructed to describe the direction of motion of the dots by saying *They are going left* or *They are going right*. If the variation in speakers' speech rate resulted from speakers analogically matching their speech rate to the speed of motion of the described event, rather than from the specific task demands of Experiment 2, then utterances describing fast-moving dots should be significantly shorter than utterances describing slow-moving dots, even though animation duration is constant.

Method

Participants

Five speakers participated in the experiment. All participants were University of Chicago students with native fluency in English and no reported history of speech or hearing disorders. Participants were paid for their participation.

Materials and analysis

The experimental materials consisted of 24 animations of dots moving left or right horizontally on a computer screen at different speeds. In each animation, the dots moved continuously (with five dots present on the

screen at any point in the display) for a duration of 3 s. There were 6 tokens of animation in each Speed condition (fast vs. slow) \times direction (left vs. right) combination. Speech was recorded using a SHURE SM94 microphone onto digital audiotape and then digitized at a sampling rate of 44.1 kHz with 16-bit resolution. Utterance duration was measured by the first author who was blind to the Speed condition. Utterances were edited into separate sound files beginning with the onset (first clear glottal pulse) of each utterance until the end of aspiration in the final /t/ in "left" or "right."

Design and procedure

Each speaker watched all 24 animations. Before each animation started, a fixation point appeared in the middle of the screen for 250 ms. Each animation was followed by 4 s of a blank display before the next trial began. Speakers were asked to describe the direction of the dots' motion in each animation by saying either *They are going left* or *They are going right*. In addition to significantly increasing the duration of the fast-motion displays and equating the animation duration across the two Speed conditions so that speakers would have plenty of time to finish speaking before the dots stopped moving, two extra precautions were taken to reduce the likelihood that variation in speech rate is due to task demands. First, the participants were told to speak naturally. Second, before beginning the experiment, the participant-speakers saw a demonstration of one fast animation and one slow animation to familiarize them with the duration of the displays.

Results and discussion

Two utterances (out of a total of 120 utterances), each produced by a different speaker, were taken out of the analysis because the speakers **were yawning** while producing the descriptions.

Mean utterance duration was 973 ms (± 19 ms SEM) for fast animations and 1034 (± 20 ms SEM) for slow animations. The mean difference between utterances produced for fast animations and slow animations was 61 ms, ranging from 46 to 87 ms. for individual speakers. This difference in duration between the Speed conditions was significant both in the analysis by items, $t(22) = -2.843, p < .01$, as well as in the analysis by subjects, $t(4) = -7.937, p < .001$.

This pattern of results replicates the results of Experiment 2: speakers varied speech rate analogically with the speed of motion of the object they were describing. Utterances produced while watching fast motion differed significantly in duration from utterances produced while watching slow motion, even though the duration of the animation was kept constant across both conditions and

despite the fact that this duration of 3 s was considerably longer than the mean duration of utterances in both conditions. This difference in utterance duration was consistent for all five speakers. These results provide support for the idea that the variation in speech rate reflects a natural part of speech rather than the specific task demands of Experiment 2. When describing a fast or a slow motion, speakers spontaneously varied their speech rate, mapping rate of visual motion to rate of speaking.

General discussion

Speakers can vary the acoustic properties of speech analogically with properties of objects or events in the world. This acoustic variation can make the semantic information conveyed by words more prominent, as shown in Experiment 1 when analog acoustic expression expresses the same meaning expressed in words. Moreover, this acoustic variation can convey information that is independent of the propositional content of the utterance and is expressed exclusively through acoustic properties of speech, as shown in Experiments 2 and 3. Analog acoustic expression serves a communicative function by providing listeners with a “channel” of information over and above the propositional-linguistic content of the utterance. Furthermore, analog acoustic expression may facilitate comprehension by setting up a non-arbitrary mapping between form and meaning adding to the information provided by the arbitrary form-meaning mapping in the linguistic channel.

Importantly, this kind of expression is spontaneously produced even when there is no intention to act out a message or perhaps even to communicate consciously—it appears to be a natural part of speech communication that may be quite broad in use. In the experiments reported here, speakers were not instructed to convey any specific message beyond the information conveyed in words, and no listener was present at the time of production. Although in some cases analog acoustic expression may result from the speaker’s communicative intention to convey a specific message through acoustic variation (the baseball example given earlier may be such an example), the results here are consistent with the idea that analog acoustic expression does not require a specific communicative intention on the part of the speaker.

The processes underlying listeners’ interpretation of analog acoustic expression may also differ in different circumstances. As with production, under some circumstances the speaker’s communicative intention may play a role in the comprehension of analog acoustic expression. Listeners may infer that the speaker intentionally uses acoustic variation to convey specific information and explicitly use the speaker’s inferred communicative intention in interpreting analog acoustic expression.

However, such an inference does not appear to be necessary. In Experiment 2, in post-experimental questioning listeners reported thinking that the speakers’ task was merely to describe the horizontal direction of the dot’s motion and they did not mention anything about trying to convey speed of motion or varying speaking rate to convey such information. Although by no means conclusive, this is consistent with the idea that listeners can interpret analog acoustic expression without attributing a specific communicative intention to the speaker. Listeners may use contextual information to infer that the acoustic variation is not random, and thus can be informative, without inferring it is communicatively intended. Acoustic properties of speech may also influence listeners by priming specific concepts, for example concepts relating to speed or verticality. Such conceptual priming may affect listeners’ representation of the referent objects or events. Although in Experiment 2 listeners were asked to decide explicitly whether speakers observed a fast or a slow dot motion, analog acoustic expression may affect comprehension by unconsciously priming speed-related concepts when listeners are not faced with an explicit decision task or even when they are not consciously aware of the speaker’s use of analog acoustic expression.

We have focused here on the role of analog acoustic expression in conveying information about visuo-spatial properties of objects and events in the world. Speakers and listeners can use acoustic analog expression to convey information about visuo-spatial properties by capitalizing on existing audio-visual cross-modal correspondences. Although the present research focused on the cross-modal correspondence between fundamental frequency and vertical location and the correspondence between speed of motion and speed of articulation, speakers and listeners may use other cross-modal correspondences that have been shown to affect non-linguistic perceptual processing. For example the properties of loudness and pitch (the perceptual correlates of the acoustic properties of amplitude and fundamental frequency) have both been shown to be associated with visual properties such as size or brightness (see Marks, 1987). The basis for these cross modal correspondences is not entirely understood. Findings suggest that the perception of some cross modal correspondences develops relatively late and may be based on the co-occurrence of auditory and visual properties in perceptual experience. For example, Marks, Hammeal, and Bornstein (1987) found that the cross modal mapping between pitch and size develops between 9 and 13 years. On the other hand, other cross-modal correspondences, such as the correspondence between pitch and brightness or between loudness and brightness, have an early developmental origin and may reflect intrinsic amodal perceptual similarities, possibly a match in the temporal properties of the underlying neural activity in

the auditory and the visual systems (Marks et al., 1987). There is also evidence suggesting that the mapping between pitch and verticality in physical space is already evident in 11-month-old infants (Wagner, Winner, Cicchetti, and Gardner, 1981). This early emergence suggests that this mapping between pitch and verticality may also be perceptually based, rather than acquired through language or other cultural influences. Still, it is possible that some audio-visual mappings are culturally-specific, or at least are reinforced or shaped in culturally-specific ways. For example, the mapping between pitch and verticality is reflected in language (we describe pitch by using terms such as *high* or *low*) and in written musical notation (in which high notes are placed higher in space). Based on the structure of musical instruments, the terms used to refer to musical pitch, and musicians' gestures in relation to pitch, Ashley (2004) argued that the strong association between pitch and verticality characterizing western cultures is not evident in non-western cultures. If audio-visual cross-modal mappings underlying the use of analog acoustic expressions vary cross-culturally, analog acoustic expression itself may exhibit the same cross-cultural variation. It is possible that speakers and listeners may exploit this communicative channel in conveying more abstract information by metaphorically mapping non-spatial properties onto the spatial domain (cf. Lakoff, 1993; Lakoff & Johnson, 1980). Research has shown that abstract domains can be understood by a mapping to more concrete domains, for example, from the more abstract domain of time to the more concrete domain of space (Boroditsky, 2000). This suggests the possibility of expressing non-spatial information through acoustic analog expression by relying on such metaphoric structuring of non-spatial domains. In addition, speakers may use analog acoustic expression that relies on more complex mapping than the relatively simple audio-visual mappings investigated in the present experiments. For example, a speaker described an automatic handwritten digit recognizer by saying, "It takes *zigz-o-zigs-three-zevn* and turns it into *six-oh-six-three-seven*" meaning a computer digit recognition system takes sloppy handwritten numbers and translates them into a clear form. The speaker used variations in phonetic properties (instead of prosody) to convey a message that is not expressed in the propositional content in itself. Thus analog acoustic expressions may extend beyond the simple prosodic variation we have demonstrated here.

The use of acoustic analog expression is similar to the gestures that accompany speech (McNeill, 1992). In both gestures and acoustic analog expression meaning is conveyed through a non-arbitrary continuous mapping of meaning and form. As with analog expression, the analog continuous mode of representation that gestures provide can be exploited to express either the same meaning represented by discrete symbols in the

accompanying speech or meaning that is not expressed by the linguistic propositional content (see McNeill, 1992). Moreover, as with analog expression, observers are sensitive to information that is conveyed exclusively in gesture (Beattie & Shovelton, 1999; Goldin-Meadow & Sandhofer, 1999; Kelly, Barr, Church, & Lynch, 1999). The similarities between these forms of expression are such that one could refer to this channel in speech as a form of "spoken gesture" (Okrent, 2002).

Language has long been viewed as conveying meaning by using a syntactically structured combination of abstract discrete symbols arbitrarily related to their referents. This form of symbolic representation has distinguished speech from non-linguistic communication systems. However, research has suggested that perceptual-motor representations that are grounded in actual perceptual-motor experience and analogically related to their referents are routinely activated during language comprehension (Barsalou, 1999; Glenberg & Kaschak, 2002; Zwaan, Stanfield, & Yaxley, 2002). For example, comprehension of sentences that implied direction of action interacted with actual response direction, towards or away from the body (Glenberg & Kaschak, 2002). By analogically mapping continuous variation in the referential domain onto continuous variation in speech, analog expression may provide this kind of grounded representation and facilitate comprehension. The use of analog acoustic expression may reflect an evolutionary precursor to the use of discrete linguistic symbols that could ground meaning in the motor system (Rizzolatti & Arbib, 1998). Speakers use analog acoustic expression spontaneously, without instruction or apparent effort, indicating this is a natural dimension of speech that provides a different means of communication that is readily used, but heretofore scientifically overlooked.

Acknowledgments

We thank Anne S. Henly and Dan Margoliash for valuable comments on this work and Nicole Donders for her help in data analysis. We also thank Rolf Zwaan and two anonymous reviewers for helpful criticism and suggestions. The support of the University of Chicago Center for Cognitive and Social Neuroscience and the National Institute of Deafness and Other Communication Disorders of the National Institutes of Health (Grant DC-3378) and the National Institute of Health (Grant P50 MH52384-11) is gratefully acknowledged.

References

- Ashley, R. (2004). Musical pitch space across modalities: spatial and other mappings through language and culture. In S. D.

- Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th International Conference on Music Perception and Cognition*. Adelaide, Australia: Causal Productions.
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614–636.
- Barr, D. J. (2001). Trouble in mind: paralinguistic indices of effort and uncertainty in communication. In S. Santi, I. Guaitella, C. Cave, & G. Konopczynski (Eds.), *Oralite et Gestualite: Communication Multimodale, Interaction*. Paris, France: L'Harmattan.
- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.
- Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30(6), 644–663.
- Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18(4), 438–462.
- Bernstein, I., & Edelstein, B. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, 87, 241–247.
- Birch, S., & Clifton, C. Jr., (1995). Focus, accent, and argument structure: effects on language comprehension. *Language and Speech*, 33, 365–391.
- Bock, K., & Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, 11(1), 64–76.
- Borodistky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, 75, 1–28.
- Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol*, 17(2), 99–117.
- Carlson, K., Clifton, C., Jr., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45, 58–81.
- Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology*, 9, 864–881.
- Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2), 292–314.
- Deacon, T. (1997). *The symbolic species: The co-evolution of language and the brain*. London: W.W. Norton & Company.
- Diehl, R. (1991). The role of phonetics within the study of language. *Phonetica*, 48, 120–134.
- Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, 100(2), 233–253.
- Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychological Bulletin & Review*, 9, 558–565.
- Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science*, 3, 419–429.
- Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, 2, 67–74.
- Hauser, M. (1997). *The evolution of communication*. Cambridge, MA: MIT Press.
- Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Conference on Association for Computational Linguistics*. New Jersey: The Association for Computational Linguistics.
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203(3), 88–96.
- Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxely, R. H., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, 94, B79–B89.
- Kelly, S. D., Barr, D. J., Church, B. R., & Lynch, K. (1999). Offering a hand to pragmatic understanding: the role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, 40, 577–592.
- Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought*. Cambridge: Cambridge University Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358–368.
- Marks, L. E. (1987). On cross-modal similarity: auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3), 384–394.
- Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, 52(1), 1–92.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- Melara, R., & Marks, L. (1990a). Dimensional interactions in language processing: Investigating directions and levels of crosstalk. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 539–554.
- Melara, R., & Marks, L. (1990b). Processes underlying dimensional interactions: correspondences between linguistic and nonlinguistic dimensions. *Memory & Cognition*, 18, 477–495.
- Melara, R., & O'Brien, T. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, 116, 323–336.
- Okrent, A. (2002). A modality-independent notion of gesture and how it can help us with the morpheme vs. gesture controversy in sign language linguistics (or at least give us some criteria to work with). In R. Meier, K. Cormier, & D. Quinto (Eds.), *Modality and structure in signed and spoken language*. Cambridge: Cambridge University Press.
- Pierce, C. S. (1932). Division of signs. In C. Hartshorne & P. Weiss (Eds.), *Collected Papers of C.S. Pierce* (Vol. 2). Cambridge, MA: Harvard University Press.
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustic Society of America*, 24, 175–184.
- Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustic Society of America*, 32, 693–703.
- Pierrehumbert, J., & Hirschberg, J. (1990). The meaning if intonational contours in the interpretation of discourse. In

- P. R. Cohen, J. Morgan & M. E. Pollack (Eds.), *Intentions on communication*. Cambridge MA: MIT Press.
- Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neuroscience*, 21, 188–194.
- Rockwell, P. (2000). Lower, slower, louder: vocal cues of sarcasm. *Journal of Psycholinguistic Research*, 29(5), 483–495.
- Saussure, F. de. (1959). *Course in general linguistics*. New York and London: McGraw-Hill.
- Scherer, K., Ladd, R., & Silverman, K. (1984). Vocal cues to speaker affect: testing two models. *Journal of the Acoustical Society of America*, 76, 1346–1356.
- Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: effects of speaker awareness and referential context. *Journal of Memory and Language*, 48(1), 103–130.
- Steedman, M. (1991). Structure and intonation. *Language*, 67(2), 260–296.
- Stevens, K. N. (1999). *Acoustic phonetics*. Cambridge, MA: MIT Press.
- Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 495–518). New York: Elsevier.
- Templeton, C. N., Greene, E., & Davis, K. (2005). Allometry of alarm calls: black-capped Chickadees encode information about predator size. *Science*, 308(5730), 1934–1937.
- Terken, J., & Nootboom, S. G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, 2(3–4), 145–163.
- Von Frisch, K. (1967). *The dance language and orientation of bees*. Cambridge, MA: Harvard University Press.
- Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). Metaphorical mapping in human infants. *Child Development*, 52, 728–731.
- Whalen, D. H., Levitt, A. G., Hsiao, P. L., & Smorodinsky, L. (1995). Intrinsic F0 of vowels in the babbling of 6-, 9-, and 12-month old French- and English-learning infants. *Journal of the Acoustical Society of America*, 97, 2533–2539.
- Zwaan, R. A., Madden, C. J., Yaxley, R. H., & Aveyard, M. E. (2004). Moving words: dynamic representations in language comprehension. *Cognitive Science*, 28, 611–619.
- Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shape of objects. *Psychological Science*, 13, 168–171.