ITERATIVE VOCAL CHARADES: THE EMERGENCE OF CONVENTIONS IN VOCAL COMMUNICATION

MARCUS PERLMAN, RICK DALE
Cognitive and Information Sciences, University of California, Merced
Merced, CA, 95343, USA

GARY LUPYAN
Psychology, University of Wisconsin, Madison
Madison, WI, 53706, USA

Evidence suggests that signed languages emerge from communication through spontaneously created, motivated gestures. Yet it is often argued that the vocal modality does not afford this same opportunity, and thus it is reasoned that language must have evolved from manual gestures. This paper presents findings from an iterative vocal charades game that shows that under some circumstances, nonverbal vocalizations can convey sufficiently precise information without prior conventionalization to ground the emergence of a spoken communication system.

1. Introduction

Where do languages come from? One common approach to this question seeks to understand the evolution of the cognitive processes that enable the creation and use of language, as a singular, cognitive abstraction. A complementary approach focuses on the processes by which specific language systems may have originated and evolved. This latter approach endeavors to explain the origin of the conventional forms and grammatical constructions that are realized across the world's languages. In signed languages and more rudimentary home sign systems, these creative processes have occasionally been directly observed in action (Goldin-Meadow & Feldman, 1977; Sandler, et al., 2005). In such cases, observations point to the production of iconic and deictic – that is, *motivated* – gestures in the creation of signed linguistic forms and grammatical constructions (Armstrong & Wilcox, 2007). These and other findings raise the intriguing possibility that over the course of interactions and transfer across generations, spontaneous motivated gestures become conventionalized and

organized into grammatical systems as duality of patterning emerges naturally over time. The same opportunity for direct observation is absent from more conventional spoken languages. Indeed, the study of language evolution is inherently handicapped by our extremely limited ability to observe spoken languages being created *de novo*.

It has often been argued that vocalizations do not have the same potential for motivated mappings as visible gesturing or drawing (Armstrong & Wilcox, 2007; Tomasello, 2008). For example, the linguist Charles Hockett reasoned that, "When a representation of some four-dimensional hunk of life has to be compressed into the single dimension of speech, most iconicity is necessarily squeezed out. In one-dimensional projection, an elephant is indistinguishable from a woodshed. Speech perforce is largely arbitrary" (1978: 274). This reasoning is vividly illustrated in a thought experiment by Tomasello (2008). He asks the reader to imagine two groups of children, each isolated on their own island. The children are well cared for, but lack a language model. One group is constrained to communicate only through bodily gestures, and the other only through the vocal modality. The crucial question concerns whether each group would develop a language and how that would happen. Tomasello concludes that gestures – by virtue of their potential for deixis, iconicity and mimesis – would weigh in favor of a gestural origin in the evolution of language. He then notes, in contrast, that it is difficult to imagine the children "inventing on their own vocalizations to refer the attention or imagination of others to the world in meaningful ways … And so the issue of conventionalizing already meaningful communicative acts never arises" (p. 228).

This leaves us with a puzzle: While there is a plausible account of progression from motivated/iconic to unmotivated/arbitrary mappings in the gestural domain, no such account is presently available for the vocal modality. In this work we present evidence that, counter to Hockett and Tomasello, nonverbal vocalizations can, under some circumstances, convey sufficiently precise information without prior conventionalization. This raises the possibility that as with sign languages, conventionalized spoken language may emerge from an initially motivated system. Our goal here is to offer empirical evidence, using an experimental semiotics paradigm (Garrod & Galantucci, 2010), that the vocal medium can serve as such a motivated medium.

Before describing our empirical study, we begin by noting recent evidence showing that people in fact do spontaneously produce motivated vocal "gestures" in their speech, at least within some limited domains. For example, speakers iconically modulate their speech rate when describing fast or slow events (Perlman, 2010; Perlman, et al., submitted; Shintel, et al., 2006). They

also produce iconic modulations in their pitch when referring to events happening in low or high vertical space (Clark, et al., in press; Shintel et al., 2006), or when describing large or small entities (Perlman, et al., submitted). A study investigating iconic prosody in infant directed speech found that adult speakers reliably modulated their prosody along several acoustic parameters when expressing antonymic pairs of meanings happy/sad, hot/cold, big/small, tall/short, yummy/yucky, and strong/weak (Nygaard, et al., 2009). Even greater potential for iconicity in the vocal modality is exhibited by research in sound symbolism, onomatopoeia, ideophones, mimetics, and related phenomena (Dingemanse, 2012; Hinton, et al., 1994; Lupyan & Casasanto, submitted; Nuckolls, 1999; Perniss, et al., 2010). This research from across the world's languages finds iconicity in conventional spoken forms relating to a range of meanings including shape, manner of motion, texture, size, brightness, distance, and temporal aspect. Yet the processes by which such iconic forms are created and conventionalized in spoken language have never been observed.

Experimental semiotics offers new empirical techniques to address the origin and development of conventional communication systems. These methods enable a sort of set up of Tomasello's thought experiment within the laboratory. While prior work in experimental semiotics has examined the emergence of conventions and linguistic structure with strings of letters (Kirby, et al., 2008), picture drawings (Garrod, et al., 2007), and with the sounds produced by a slide whistle (Verhoef, et al., 2012), it has not investigated the emergence of conventions in vocal communication. Thus the present experiment used an iterative vocal charades game to examine the development of a conventional vocal lexicon, and particularly the role of iconicity in this process.

2. Method

2.1. *Participants*

Participants were ten pairs of undergraduates enrolled in Psychology courses at the University of California, Santa Cruz. They received course credit in exchange for their participation.

2.2. *Materials*

The game was played with index cards. Printed on each card was one of eighteen different words. The words included nine pairs of semantic antonyms: *attractive/ugly, bad/good, big/small, down/up, far/near, fast/slow, few/many, long/short*, and *rough/smooth*.

2.3. *Design and Procedure*

Participants were told they would be playing a game of "vocal charades". In this game, they would take turns vocalizing the meaning of words on index cards, as their partner tried to guess the word. No actual words or body movements were allowed. After the instructions, each player was given twelve shuffled cards, which included six antonym pairs (one word per card, in random order). Each player held three pairs of words that were uniquely held by just that player, and also three pairs that were shared and held by both players.

The game was played over ten rounds. For each round, the first player vocalized each of his or her twelve cards. Then the microphones were switched, and the second player took a turn with all of their cards. Players shuffled their respective deck of cards before each round. A turn lasted for up to ten seconds, monitored by the experimenter. During this time, players were permitted to make multiple sounds and guesses. The turn ended either when the time was up, or with a correct guess, which was immediately noted by the vocalizer. If no correct guess was made, then the experimenter called time at ten seconds, and the vocalizer shared the correct word with her partner.

The players were audio recorded as they played the game. A lapel microphone was attached to the vocalizer and a flat boundary microphone was placed in front of the guesser. Both microphones were plugged into a digital recorder.

2.4. *Analysis*

The speech analysis software Praat was used to measure the acoustic properties of each sound, including its mean, minimum and maximum pitch (Hz), duration (s), intensity (dB), and harmonics to noise ratio (Boersma, 2001). First, the boundaries of each individual sound produced during a turn were marked on a Praat text grid. These boundaries were determined by a combination of listening and viewing a spectrogram of the sound. Additionally, each guess and the time it occurred was noted on the text grid. The sound files and resulting text grids were then fed into a Praat script that computed the value of each acoustic measurement.

3. Results

Measures of performance improved across rounds of the game. The number of guesses per turn decreased over rounds as the correct answer was guessed with fewer attempts (t = -2.44, p < .05; see Figure 1). Also, the number of sounds

produced by the vocalizer decreased over rounds, as fewer sounds were needed for the partner to guess the correct answer (t = -5.04, p < .001; see Figure 1).

There is also evidence that the sounds became increasingly conventionalized over the course of the game. First, the sounds became shorter, as the average duration of each vocalization decreased over rounds (t = -7.18, p < .001 See Figure 1). The sounds also appeared to stabilize, as the correlation between the acoustic properties of the sounds from round to round increased from the beginning to the final rounds (see Figure 2). This indicates that by later rounds, participants were generally using more stable acoustic strategies to communicate.

Figure 1. *Left*: Average vocalization duration (s) by round. *Right*: Number of guesses by round.

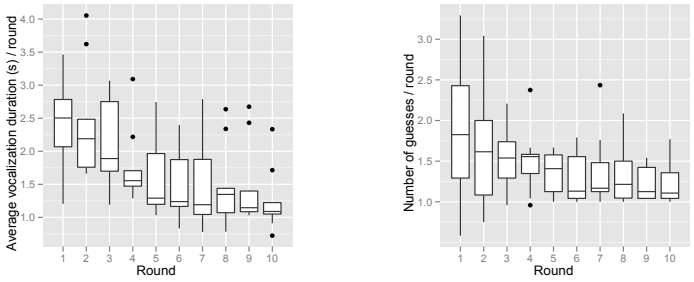
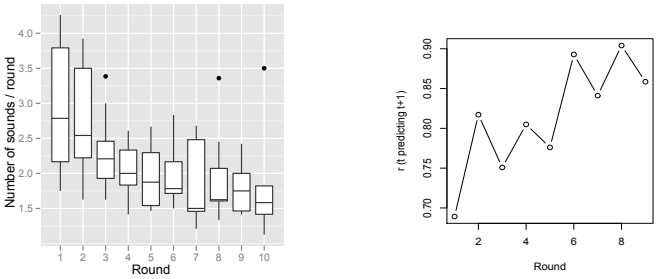


Figure 2 *Left*: Number of sounds by round. *Right*: The correlation between the first principal component (rendered from all 5 acoustic measures) at round t and round t + 1. By the final round, the correlation with the first component from the previous component approaches 1.



A glance at the acoustic properties that distinguish word pairs such as smooth/rough and fast/slow immediately reveals expected patterns of iconicity. Some of these are shown in Figure. For example, the HNR measures for rough and ugly are lower than for smooth and attractive; the sound duration and total duration for fast and few are lower than for slow and many. In order to test whether acoustic values significantly separate categories, we ran a series of

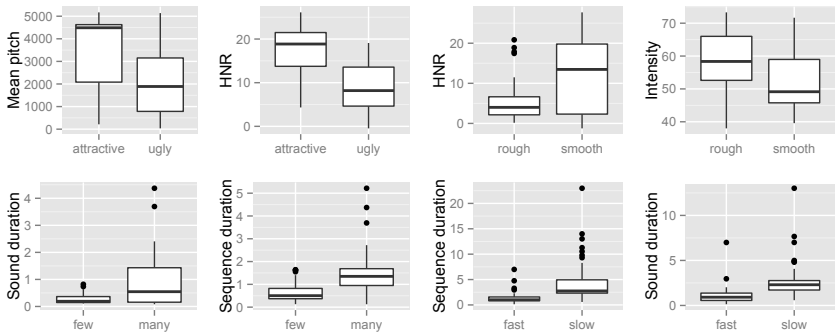
logistic mixed-effects models. Models predicted word status (e.g., rough vs. smooth) by using the acoustics as predictors (intercepts for participant pairs were used as random effects). In each of these models (for each word pair), coefficients reflect the extent to which a given acoustic measure predicts one versus the other word, while controlling for all acoustic variables simultaneously. For target word pairs, *all* models significantly predicted one word vs. its partner, even using a conservative cutoff of $p = .001$. For example, mean pitch (log-odds = -.74) and HNR (log-odds = -2.6) both predict ugly vs. attractive (p's < .001). Up is also distinguished from down with acoustic measures of duration (log-odds = -.85) pitch change (log-odds = 0.68) and intensity (log-odds = .80; p's < .001). Space limitations restrict detailed description of model coefficients, but all models showed this kind of acoustic separation.

We conducted this same logistic regression modeling on all word pairs. The results of these models suggest there is pervasive acoustic distinctiveness. In all, 154 models were constructed for all pairs, such as up vs. rough, short vs. few, and so on. The results of these models show that for almost all of the 154 pairwise comparisons between words, at least one of the five acoustic measures was a reliable predictor (see Table). Over a third (58/154) of the models had 2 acoustic predictors separating word classes, and about a fifth had 3 (30/154).

Table 1 Breakdown of how many (of 5) acoustic predictors were significant across regression predicting one word vs. another using acoustic variables.

# measures at p < .001	0	1	2	3	4	5
# comparisons	7	55	58	30	3	1

Figure 3 Some exemplary comparisons between the words *attractive/ugly*, *rough/smooth*, *few/many*, and *fast/slow*. The words in these comparisons contrast distinctly along iconic acoustic properties.



4. Discussion

In the context of a vocal charades game, participants showed remarkably robust intuitions for how to vocally express a variety of words without sound-based meanings. Over the course of each game, different players reliably discovered similar iconic sounds for each particular word. In turn, guessers became increasingly proficient at understanding the iconic sound, requiring fewer and shorter sounds to guess correctly with fewer tries.

Many gestural theories of language evolution assume that, compared to manual gestures, the vocal modality is extremely limited in its iconic potential to ground a conventional communication system (Armstrong & Wilcox, 2007; Tomasello, 2008). Our findings to some extent contradict this widely held assumption. Indeed, we find that, even without prior conventionalization, nonverbal vocalizations can convey sufficiently precise information to ground the emergence of spoken conventions. This raises the possibility that as with signed languages, conventionalized spoken language may emerge from an initially motivated system. These findings correspond to growing documentation of iconicity across the world's spoken languages, as well as to experimental findings showing that speakers often produce iconic modulations in their prosody as they talk about different kinds of meanings like fast and slow, up and down, or big and small (Clark et al., in press; Perlman et al., submitted; Perniss, et al., 2010). While these results do not weigh against a multimodal origin of spoken language, they nevertheless suggest that we should not underestimate the iconic potential of the vocal modality in theorizing about language evolution.

References

Armstrong, D.F. & Wilcox, S. (2007) *The gestural origin of language.* New York: Oxford University Press.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341-345.

Clark, N., Perlman, M., and Johansson Falck, M. (in press). Iconic pitch expresses vertical space. In M. Bokrent, B. Dancygier, and J. Hinnell (Eds.), *Language and the creative mind*. Stanford, CA: CSLI.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass, 6,* 654-672.

Galantucci, B., & Garrod, S. (2010). Experimental semiotics: A new approach for studying the emergence and the evolution of human communication.*Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & Macleod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science, 31(6),* 961–987.

Goldin-Meadow, S. & Feldman, H. (1977). The development of language-like communication without a language model. *Science, 197,* 401-403.

Hinton, L., Nichols, J. & Ohala, J.J. (Eds.). (1994). *Sound Symbolism*. Cambridge, UK: Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS. 105(31),* 10681–10686.

Lupyan, G. & Casasanto, D. (Submitted). Meaningless words promote meaningful categorization.

Nuckolls, J. (1999). The case for sound symbolism. *Annual Review of Anthropology, 28,* 225-252.

Perlman, M. (2010). Talking fast: The use of speech rate as iconic gesture. In F. Perrill, V. Tobin, & M. Turner (Eds.), *Meaning, Form, and Body* (pp. 245-262). Stanford, CA: CSLI Publications.

Perlman, M., Clark, N., & Johansson Falck, M. (Submitted). Iconic prosody in story reading.

Perniss , P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology, 1,* 1-14.

Sandler, W., Meier, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences, 102,* 2661-2665.

Shintel, H., Nusbaum, H.C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language, 55*, 167-177.

Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.

Verhoef, T., de Boer, B., & Kirby, S. (2012). Holistic or synthetic protolanguage: Evidence from iterated learning of whistled signals. In Scott-Phillips, T., Tamariz, M., Cartmill, E., A., & Hurford, J. R. (Eds.) The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9) (pp. 39, Hackensack, NJ: World Scientific.