TYPE Brief Research Report
PUBLISHED 26 September 2022
DOI 10.3389/fcomm.2022.994162

Check for updates

OPEN ACCESS

EDITED BY
Claudia Lehmann,
University of Bremen, Germany

REVIEWED BY
Marcel Schlechtweg,
University of Oldenburg, Germany
Oliver Niebuhr,
University of Southern
Denmark, Denmark

*CORRESPONDENCE
Axel G. Ekström
axeleks@kth.se

SPECIALTY SECTION
This article was submitted to
Multimodality of Communication,
a section of the journal
Frontiers in Communication

RECEIVED 14 July 2022
ACCEPTED 06 September 2022
PUBLISHED 26 September 2022

CITATION
Ekström AG, Nirme J and Gärdenfors P
(2022) Motion iconicity in prosody.
*Front. Commun.* 7:994162.
doi: 10.3389/fcomm.2022.994162

# Motion iconicity in prosody

Axel G. Ekström[1]*, Jens Nirme[2] and Peter Gärdenfors[2,3]

[1]KTH Speech, Music, and Hearing, Stockholm, Sweden, [2]Department of Philosophy, Lund University Cognitive Science, Lund University, Lund, Sweden, [3]Paleo-Research Institute, University of Johannesburg, Johannesburg, South Africa

Evidence suggests that human non-verbal speech may be rich in iconicity. Here, we report results from two experiments aimed at testing whether perception of increasing and declining $f_0$ can be iconically mapped onto motion events. We presented a sample of mixed-nationality participants ($N = 118$) with sets of two videos, where one pictured upward movement and the other downward movement. A disyllabic non-sense word prosodically resynthesized as increasing or declining in $f_0$ was presented simultaneously with each video in a pair, and participants were tasked with guessing which of the two videos the word described. Results indicate that prosody is iconically associated with motion, such that motion-prosody congruent pairings were more readily selected than incongruent pairings ($p < 0.033$). However, the effect observed in our sample was primarily driven by selections of words with declining $f_0$. A follow-up experiment with native Turkish speaking participants ($N = 92$) tested for the effect of language-specific metaphor for auditory pitch. Results showed no significant association between prosody and motion. Limitations of the experiment, and some implications for the motor theory of speech perception, and "gestural origins" theories of language evolution, are discussed.

KEYWORDS

**voice perception, gesture, paralinguistics, motor theory of speech perception, evolution of language**

## Introduction

It has been claimed that vocalizations alone lack the potential to express iconic information through their form, whereas gestures provide comparatively rich material for representation (e.g., Hockett, 1978, p. 274; Tomasello, 2010, p. 228). However, accumulating evidence now runs counter to this claim (for an overview see Perlman and Cain, 2014). Beginning with Köhler (1929) mapping of nonsense words "takete" and "baluma" to angular and curvy shapes, respectively – later extended to "kiki" and "bouba" by Ramachandran and Hubbard (2001) – several additional forms of iconicity, defined as statistically observed non-arbitrary analogy between form and meaning, in non-verbal vocalizations have been established. In the present text, we investigate a possible such relationship between prosody – the melody of speech – and movement in space.

Other researchers have demonstrated similar such mapping. For example, Shintel et al. (2006) showed that participants, when using a prescribed phrase, produced utterances increasing in fundamental frequency ($f_0$; of phonation, corresponding to pitch) when describing the upward movement of a dot on a computer screen and declining $f_0$ when describing downward movement; and work by McClave (1998) has documented coordination of $f_0$ and manual gestures in conversation (see Kendon, 2004). Results by Küssner et al. (2014) showed how musical training facilitates similar mapping (where elevation in space was represented by motion-captured hand movements), indicating that such abilities are subject to some degree of plasticity (see overview by Eitan, 2013). Finally, sound-movement iconicity is also regularly employed in animated film soundtracks, such as the music of Carl Stalling, composer for *Looney Toons* short films.

There is a rich history of research on sound symbolism with bearing on this issue, foremost of which is Ohala (1994) "frequency code" hypothesis, positing that higher-frequency signals denote smaller, or submissive vocalizers; and lower-frequency signals denote larger, and dominant ones (Ohala, 1982; Tsai et al., 2010; Pisanski and Rendall, 2011). In everyday speech, variations in $f_0$ are employed to denote various linguistic cues (e.g., end of a speaker's turn in conversation, or emphasis through word stress; e.g., Schegloff, 1998), which have on occasion been interpreted according to the framework of the frequency code hypothesis (Ohala, 1984). Remarkably, however, iconicity of more general prosody has been little studied. Various works touch on the general topic (e.g., Hirst and Hancil, 2013) but does not concern iconicity in the prosody of single words or phrases.

The discussion of iconicity carries significant weight for research on the evolutionary origins of spoken language (Perlman and Cain, 2014). Following from "gestural origins" theories of language evolution (e.g., Corballis, 2002), Gentilucci and Corballis (2006, p. 951) have speculated that "if spoken language had evolved from a manual system, one would expect a continuation of more analog representation, perhaps with increasing pitch to indicate climbing motion, rapid speech to indicate fast motion, and so on. The fact that spoken languages are almost entirely dependent on discrete recombinant representations suggests, according to Talmy, that language arose through the vocal and auditory channel." We disagree with this assessment. The study of cross-modal mapping is relevant for the question of how interactions between gestural and vocal communication evolved in ancestral primates and humans. Modern humans readily pantomime gesturally, i.e., represent a sequence of events iconically with gestures, but this ability is extremely limited in other species. Humans also pantomime vocally. We hypothesize that vocal pantomimes may have evolved from gestural ones (where gesture includes articulation and vocal production; Liberman and Mattingly, 1985; Browman and Goldstein, 1990; Shayan et al., 2011).

The research question of the present paper is whether increasing and declining $f_0$ maps iconically onto motion events that involve upward and downward movement, respectively. In our main experiment, we presented participants with two short videos, where one involved an ascending object, and the other a descending one; presented simultaneously with each video was a disyllabic non-sense word, pitch resynthesized with increasing or declining $f_0$ trajectory. Participants were tasked with selecting which of the two videos the word described. Results of the experiment tentatively indicate that prosody is indeed iconically associated with motion, primarily driven by selection of words with declining $f_0$. Further, pitch discrimination ability, did not significantly impact results.

Because not all languages use the "high-low" metaphor to describe polarities of pitch perception, an alternative interpretation is that aspects of a listener's language – here, own-language metaphor for auditory pitch – influence object perception and pitch perception. Such "Sapir-Whorf" effects (Sapir, 1929; Whorf, 1956) are highly controversial in the literature (Pullum, 1991; Pinker, 2007; Deutscher, 2010; McWhorter, 2014) but minor effects have proven resilient in limited settings such as color perception, such that speakers of Russian – possessing words for more shades of blue compared to English – more readily perceive minor differences between blue color stimuli, compared with native English speakers (e.g., Boroditsky, 2009). Comparably, languages such as Turkish predominantly use a "thick-thin" metaphor (Shayan et al., 2011) to describe pitch; a Sapir-Whorf hypothesis applied to motion-iconic mapping thus suggests that native Turkish speakers less readily map $f_0$ trajectories to congruent visuospatial relationships. We conducted a follow-up experiment with native Turkish speakers, observing no overall significant effect of iconicity on selections. However, a significant positive relationship between selection congruency and accuracy of pitch discrimination (a separate task), which had not been observed in the original study, was observed in the follow-up.

# Methods

## Auditory stimuli

20 disyllabic non-sense words were constructed to assess the hypothesis. Words were constructed avoiding unvoiced consonants mid word, which interfered with linear pitch resynthesis in early testing (see Supplementary Table 1). A female speaker was then recruited and recorded at a sampling rate of 44,100 Hz speaking all words, using a Sennheiser ME 2 sub-miniature clip-on microphone, SK100 G4 bodypack transmitter, and EK100 G4 camera portable wireless receiver.

All stimuli were pitch-synthesized (using Praat; Boersma, 2006) in two iterations; with (1) $f_0$ declining from 250 Hz to 125 Hz; and (2) $f_0$ increasing from 125 Hz to 250 Hz,

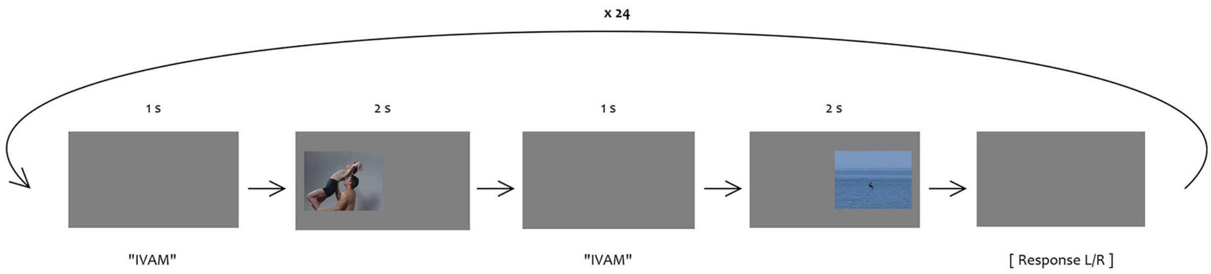

FIGURE 1
Estimated means and 95% confidence intervals of selection congruency related to factors $f_0$ direction **(left)** and pitch discrimination accuracy **(right)** from analysis of the main experiment.

respectively, the difference in Hz between beginning and end values corresponding to 12 semitones (one octave). Stimuli were then selected based on criteria of naturalness before implementation in the experiment proper. Resynthesized utterances were deemed unnatural if containing unvoiced segments painful to the ear; all co-authors were required to agree on a set of 12 × 2 stimuli (the same words, in both iterations), which were then implemented in the experiment.

## Visual stimuli

The visual stimuli consisted of video clips showing different examples of movements with either downward or upward trajectories. The set of videos were selected in two stages. In the first, we searched an online resource for free-stock videos (www.pexels.com), using verbs associated with ascending or descending movement (e.g., "raise", "dive"). 48 videos (24 for each movement direction) were then edited to 2.0s duration (isolating a single movement in a consistent direction), 25 frames per second and 800 × 600 pixel resolution. All authors agreed on 24 videos (12 for each of the two directions of movement) from the larger set using consistency and clarity of movement within each clip, and variation over the set of clips, as criteria. Guiding principles were that clips should not obviously resemble each other, and clearly display the trajectory of movement (see Supplementary Table 2).

## Participants

135 Participants were recruited using the *Prolific* platform (Prolific.co) with no limitations on eligibility. The most frequent self-reported first languages of the participants were Polish (19%), English (16%), and Portuguese (13%), followed by a group of Bantu languages (13%). 17 participants aborted participation and were excluded, resulting in a sample size of ($N$ = 118) (54 female, age $M = 26.82$, SD = 8.04).

In a follow-up experiment, an additional 113 participants, who had specified Turkish as their first language in their Prolific user profiles, were recruited. Nineteen participants aborted their participation, resulting in a sample size of ($N = 92$) (49 female, age $M = 29.33$, SD = 5.82).

Participants were compensated an hourly average of £6.50, depending on completion time.

## Experimental platform

The experiment was constructed in PsychoPy (psychopy.org), adapted to be convertible to JavaScript and uploaded to PsychoPy's integrated online platform Pavlovia (pavlovia.org).

## Procedure

Participants were instructed that they were to take part in an experiment about pairing verbs with one of two available videos. Before starting the main task, participants entered their age, sex, and first language/s. They were then instructed about the main task. In the instructions, spoken verbs were claimed to be in Amondawa, spoken by the Amazonian tribe of the same name, indigenous to Western Brazil; this to minimize risk of any presupposition from participants' prior knowledge of any given language, culture, or nationality. Two practice trials were then presented, using English verbs.

In each trial of the experiment proper, one video depicting downward movement and one upward movement were presented; one video was presented on the left-hand side of the screen, which played to completion and disappeared, before the other was presented on the right-hand side of the screen. Both videos in a trial were accompanied by the same non-sense word. The order of non-sense words over trials were pseudorandomized, and word-video combinations within a trial were pseudo-randomized with the constraint that each video appeared once as congruent with a word, and once as incongruent with a different word, throughout the experiment. Participants responded by indicating the position of the video –

left or right – they selected as more likely to be described by the nonsense word, by pressing the corresponding arrow key (see Figure 1). In total, 24 trials were presented (12 videos, presented twice each; 12 words, presented once as descending and once and ascending).

After conclusion of the main task, participants were asked whether they had noticed any pattern in the sequence of words, and/or – prompted as a subsequent question – in the sequence of video pairs. These questions were included as a check on participants' awareness of the pattern under investigation ($f_0$-movement congruence). Finally, participants were presented with a pitch discrimination task in which they made explicit selection of the $f_0$ direction of the same 12 nonsense words as in the main task by pressing the "up" or "down" arrow keys, after first hearing explicit examples of increasing and declining $f_0$. The order of non-sense words was again randomized.

## Data analysis

Data were analyzed in R (R Core Team, 2020). Mixed effects linear regression models were performed using the *lme4* package (Bates et al., 2014), and figures generated using the *effects* package (Kuznetsova et al., 2017). Main task and pitch discrimination data were aggregated (mean values) per participant and $f_0$ direction (crossed factors). Significance was interpreted by 95% confidence intervals.

The rate (relative chance level) of selecting the *congruent* video (where movement in the video was congruous with that of the $f_0$ direction of the accompanying word stimulus) was modeled as a sum of the fixed factors: $f_0$ *direction* (centered in the middle of the two levels *down* and *up)* and *pitch discrimination accuracy* (centered on the global mean) and a random factor representing participant *ID*. The intercept of the model thus represents the overall congruency of the selections relative the chance level (50%).

Differences in pitch discrimination accuracy between ascending and descending pitch were analyzed by paired samples $t$-tests.

# Results

## Main experiment

One participant was excluded from data analysis due to responses indicating awareness of the task. Results of the pitch discrimination task (performed after the main task) indicated that participants were able to correctly classify $f_0$ direction above chance level for words both declining ($M = 0.704$, $SD = 0.222$) and increasing in $f_0$ ($M = 0.726$, $SD = 0.201$), with no significant difference in success rate between the two (paired samples $t$-test $t = -1.178$, $p = 0.241$).

TABLE 1 Regression analysis of $f_0$-motion mapping congruency (Main experiment).

| | Est. | Std. Er. | df | $t$-value | $p$ |
|---|---|---|---|---|---|
| (Intercept) | 0.02 | 0.01 | 115.51 | 2.16 | 0.033 |
| Direction | −0.04 | 0.02 | 116.47 | −2.17 | 0.032 |
| Pitch discr. | 0.02 | 0.05 | 205.11 | 0.37 | 0.71 |

To investigate our main research question – that prosody non-arbitrarily maps onto motion in space – we analyzed selections from the main task by mixed-effects regression (see "Data Analysis"). The co-efficient of determination of the model (conditional $R^2$, Barton, 2009; Nakagawa et al., 2017) was 0.244. The fitted model estimated a small but significant effect (indicated by the intercept) in the expected direction ($\beta = 0.024$, $t = 2.16$, $p = 0.033$; see Figure 2). The model also estimated a significant effect of $f_0$ *direction* ($\beta = -0.037$, $t = -2.17$, $p = 0.032$), indicating that words with declining $f_0$ were more often congruently selected. The effect of *pitch discrimination accuracy* was non-significant ($\beta = 0.018$, $t = 0.37$, $p = 0.712$; see Table 1). The intercept standard deviation of the random factor *ID* was 0.072.

## Follow-up experiment

To investigate if our findings would replicate in a population limited to native speakers of Turkish (which does not primarily utilize the "high-low" metaphor as descriptive of pitch; see Shayan et al., 2011), we analyzed follow-up experiment data using the same tools/methods as described previously. Of the 92 sampled participants, 9 were excluded, either due to responses indicating awareness of the patterns of increasing or declining $f_0$ and movements in videos ($N = 5$), or not confirming Turkish as one of their first languages when prompted at the start of the experiment ($N = 4$).

Results of the pitch discrimination task again indicated that participants were able to correctly classify the direction $f_0$ above chance level for words both declining ($M = 0.674$, $SD = 0.203$) and increasing ($M = 0.729$, $SD = 0.191$) in $f_0$. However, in this sample we also observed a significant difference in success rate between the two directions (paired-samples $t$-test $t = 2.43$, $p = 0.017$), such that ascending $f_0$ tones were more readily discriminated.

A mixed-effects linear regression model (conditional $R^2 = 0.114$) (Table 2) was then fitted (see "Data Analysis"). The estimated intercept revealed no significant overall congruency of selections above chance level ($\beta = 0.007$, $t = 0.594$, $p = 0.554$; see Figure 3). The model again estimated a significant effect of *pitch direction* ($\beta = -0.066$, $t = -3.124$, $p = 0.002$), such that words with declining $f_0$ were more often congruently selected. The model also estimated a significant positive effect

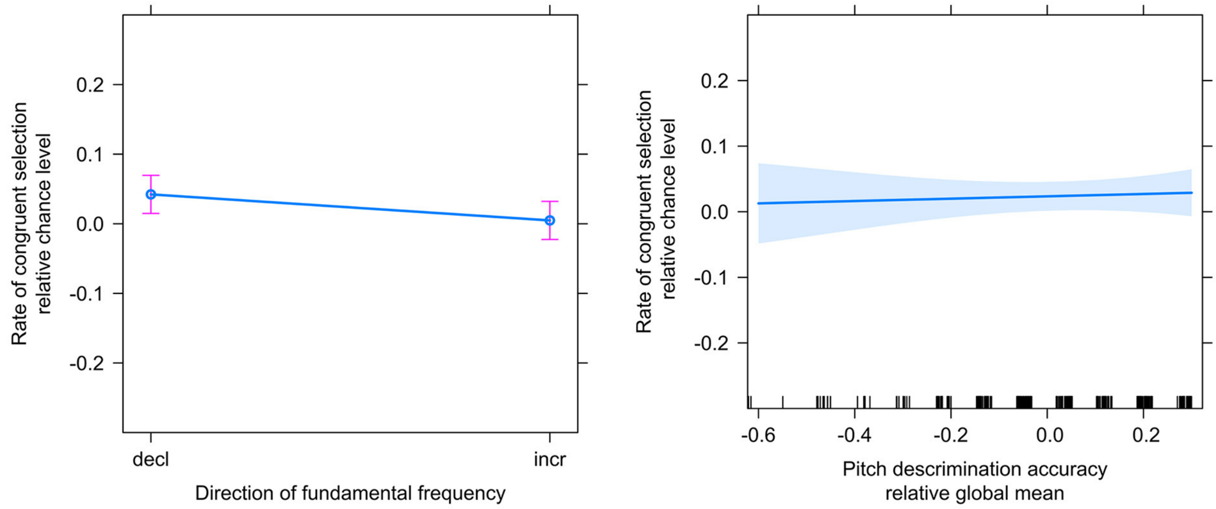

FIGURE 2
Estimated means and 95% confidence intervals of selection congruency related to factors $f_0$ direction **(left)** and pitch discrimination accuracy **(right)** from analysis of the follow-up experiment.

of *pitch discrimination accuracy* ($\beta = 0.130$, $t = 2.351$, $p = 0.020$; see Table 2), indicating that the more accurately individual participants could explicitly classify the $f_0$ movements of the words, the more likely they were to select videos with movement in the congruent direction. The intercept standard deviation of the random factor *ID* was 0.029.

TABLE 2 Regression analysis of $f_0$-motion mapping congruency (Follow-up experiment).

| | Est. | Std. Er. | df | *t*-value | *p* |
|---|---|---|---|---|---|
| (Intercept) | 0.01 | 0.01 | 80.68 | 0.59 | 0.55 |
| Direction | −0.01 | 0.02 | 83.52 | −3.12 | 0.002 |
| Pitch discr. | 0.13 | 0.06 | 138.82 | 2.35 | 0.02 |

## Discussion

The motivation for the present study was the hypothesis that movements of a pantomime can be metaphorically mapped onto a prosodic pattern. For example, it is common to depict a descending object by a vocal sound with declining $f_0$. The basic mapping, then, is from vertical space to $f_0$ such that a higher spatial location maps onto a higher $f_0$, and vice versa. In this manner, prosody maintains at least part of the iconicity of a pantomime. We submit that such a mapping forms a seed that provides many forms of vocalizations with meaning. Results of the main experiment provide some evidence that $f_0$ maps iconically to movement in space, although the overall effect was small, at 2.4% above chance level. Because the experiment task is significantly different to explicit mapping tasks more typically utilized in experiments (Eitan, 2013), such small effects are not surprising. A natural follow-up to findings of the main experiment 3.1 is that, because crossmodal mapping is subject to plasticity through training (Küssner et al., 2014), speakers from tone-accent languages may more readily make such associations. Limitations from sample size prevent us from researching such potential differences; nevertheless, it represents an intriguing suggestion for future work.

Further, the absence of any overall effect in the follow-up experiment with native Turkish speakers also gives some indication that the iconicity effect may depend on exposure to the high-low metaphor, which would support a Sapir-Whorf interpretation. It is, however, worth pointing out the unclear distinction of exposure between the samples: regardless of native language, all participants were sufficiently proficient in English to understand experiment instructions. Moreover, we observed a significant positive relationship between selection congruency and *pitch discrimination accuracy* in the follow-up experiment, indicating that Turkish speakers who did perceive the $f_0$ direction were more likely to make iconically congruent selections – an effect not observed in the experiment proper. It thus remains unclear whether iconic mapping of $f_0$ to movement depends on explicit perception. This question merits further investigation.

Across experiments, the most consistent effect observed was that declining-$f_0$ non-sense words were more likely to be selected congruently than increasing-$f_0$ non-sense words. Ostensibly, declining $f_0$ may be more readily perceived than increasing $f_0$. However, available literature on the topic suggests that rising (and, in general, higher) $f_0$ places greater demands on attention (see e.g., Hsu et al., 2015). Neither do results

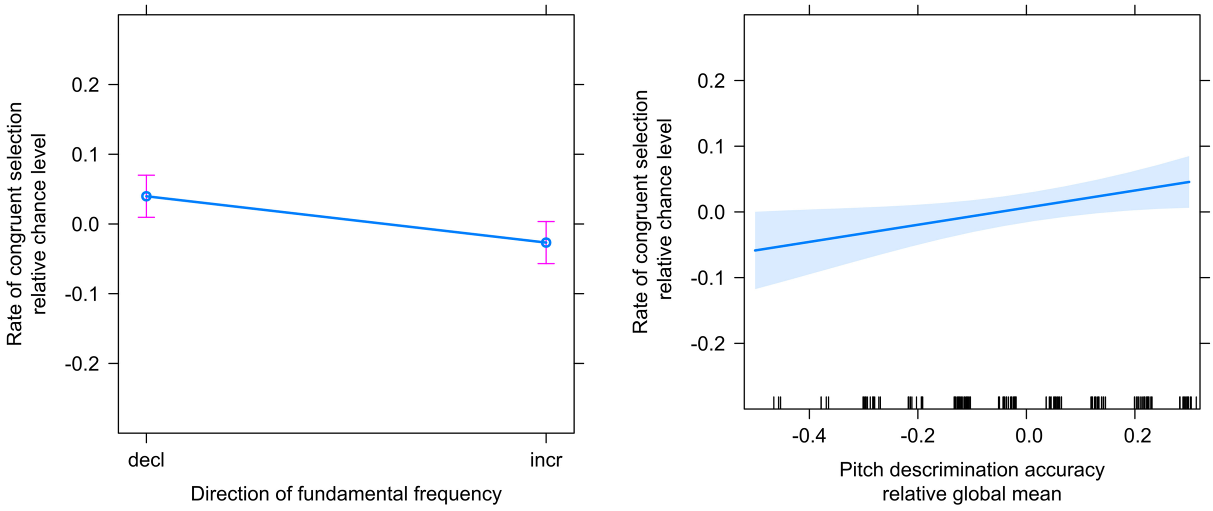

FIGURE 3
Experiment paradigm.

of the pitch discrimination task support such an explanation; rather, among native Turkish speakers, we observed a significant difference in the opposite direction.

Another possible explanation is that the videos depicting descending movements were generally more likely to be selected. This could be due to being perceived as more plausibly corresponding to a word in an Amazonian language, or simply depicting more everyday actions, compared to more uncommon (ascending) ones. Were this so, overall iconicity in the data would be diminished. Conversely, it may be that declining $f_0$ are in fact eliciting congruent selections while increasing $f_0$ do not, resulting in a preference for videos displaying descending motion. Present results do not allow disambiguation of these possible causal links; however, issues outlined above can straightforwardly be addressed by future work.

## Limitations

It is possible that the nature of the online experiment affected the results. Such experimental settings frequently result in poorer-quality data, resulting from a variety of factors which, in *in-situ* experimental settings, can be controlled by the experimenter. In the context of a listening experiment, for example, such factors include audio equipment, volume settings, and devices (e.g., headphones or speaker) used, both of which will inevitably differ between participants in online settings.

A more significant source of potential error in extrapolating from present results, however, stems from auditory stimuli not being representative of everyday speech prosody, which typically exhibits declination, such that $f_0$ on average tends to decline over the course of longer utterances (Fuchs et al., 2016). Additionally, while ranges similar to those represented by the experimental stimuli are available to most speakers and observed across speakers in natural speech Traunmüller and Eriksson[1], everyday prosody is restricted in comparison and typically does not encompass those same ranges within single-syllable utterances. This being so, however, it is worth noting that Fuchs et al. (2016) found that negative $f_0$ slopes are less variable and shallower for short utterances, more comparable to those implemented in the experiment.

## Implications and future directions

An underlying motivation for experiments presented here is that they may shed light on processes involved in the evolution of language. Several researchers claim that communication was originally mainly gestural (Gentilucci and Corballis, 2006). The question follows, how evolution moved "from hand to mouth" (Corballis, 2002); how did spoken language evolve from pantomime and gesture? [(Corballis, 2014b), p. 185] writes: "Speech can be considered the end point of a conventionalization process in which pantomimic representations are replaced by arbitrary vocal signals". However, this does not answer the question of how vocal sounds initially acquire meaning. Words are indeed conventions, but a conventionalization process must have initial seeds. We argue that the question of whether gesture preceded speech in evolution is poorly stated (cf. Kendon, 2004). Human communication has likely always been a combination of both (though the balance of elements may since have shifted from mainly gestural to mainly vocal). Kendon (2004, p. 165) concludes that "there seems to be no reason to suppose that the

[1] Traunmüller, H., and Eriksson, A. (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. (Unpublished).

development of a capacity for symbolic expression was at first confined to only one modality." Indeed, gesturing and speaking are components of a single process of utterance generation (Kendon, 2004; McNeill, 2012). The appropriate question, then, is how vocalizations of ancestral primates were paired with iconic gestures, becoming a productive and preferably symbolic form of communication (Corballis, 2014a; Goldin-Meadow, 2021).

Prelinguistic infants are sensitive to iconicity of $f_0$ (Dolscheid et al., 2012, 2014) which indicates that infants make iconic $f_0$-spatial associations before they acquire language. Infants of 3–12 months of age associate $f_0$ with vertical position or movement (Walker et al., 2010; Dolscheid et al., 2012; Jeschonek et al., 2013). Further, this ability of infants to match the intonation of utterances with hand gestures facilitates language acquisition (Arbib, 2012; McNeill, 2013). To the knowledge of the authors, no such evidence of ready multimodal association exists for other primates. It is possible, then, that humans' voluntary control of vocal organs (Simonyan and Horwitz, 2011) facilitated mapping between manual movement and the perceptual trajectories of vocal signals.

Animal and human infant vocalization almost exclusively serve the purpose of expressing emotions (Panksepp, 2010), for which prosody is crucial. Even prairie dogs exploit prosodic variation in communication (Slobodchikoff et al., 2009). An example of early-in-life human reliance on prosodic cues was provided by Fernald (1989) study of mothers' use of prosody in interactions with infants (see also Fernald and Simon, 1984), finding that bids for attention were typically associated with increasing $f_0$, while approval was associated with declining $f_0$. She concluded that these prosodic patterns convey emotional content meaningful to infants. Finally, Shayan et al. (2011) speculate that the perceptual and cognitive mapping of $f_0$ as high or low may relate to movements of the larynx in vocalization. Anterior stretching (i.e., raising) of the larynx produces more high-frequency resonances, while posterior stretching (i.e., lowering) produces more low-frequency resonances, which the listener perceives as changes in pitch.

The claim that anterior and posterior laryngeal stretching facilitates $f_0$ mapping assumes, in accordance with the motor theory of speech perception (Liberman et al., 1967; Liberman and Mattingly, 1985; Galantucci et al., 2006), that speech perception is to some degree contingent on perception of speech gestures. Indeed, a speaker producing high-frequency vs. low-frequency vocalizations can easily identify variable points of resonance in the chest (lower frequencies) and throat (higher frequencies), respectively (see, for example, Zbikowski, 1998; for a discussion, see also Shayan et al., 2011). Finally, laryngeal anterior-posterior stretching in the laryngeal prominence (Adam's apple) is readily observable by attending to speakers' throats – in particular those of male speakers (possessing more visible laryngeal prominences).

## Conclusion

Results here presented provide tentative support for a motor theory-style interpretation of speech prosody perception, while allowing for variation from speaker native language. The present study also presents indications of the ability in humans to non-arbitrarily map prosodic declination to instances of motion in the natural world. Results align with a holistic approach to language evolution. However, further validation of results is required before drawing proper conclusions.

## Data availability statement

The original contributions presented in the study are publicly available. This data can be found at: https://github.com/axel-g-ekstrom/motion_iconicity.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

AE: conceptualization, software, methodology, writing—original draft, and writing—review and editing. JN: software, methodology, investigation, formal analysis, resources, writing—original draft, and writing—review and editing. PG: conceptualization, methodology, supervision, project administration, funding acquisition, writing—original draft, and writing—review and editing. All authors contributed to the article and approved the submitted version.

## Acknowledgments

The authors would like to thank the Swedish Research Council for support to the Linnaeus environment *Thinking in Time: Cognition, Communication and Learning*. The authors gratefully acknowledge Lund University Humanities Lab for providing access to recording equipment and valuable expertise. We thank Melina Tsapos for assistance with recordings.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.994162/full#supplementary-material

## References

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror System Hypothesis, Vol. 16*. Oxford: Oxford University Press.

Barton, K. (2009). MuMIn: multi-model inference. R package version 1.43.17. Available online at: http://r-forge.R-project.Org/projects/mumin/ (accessed April 14, 2020).

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv* 1406, 5823. doi: 10.18637/jss.v067.i01

Boersma, P. (2006). *Praat: doing phonetics by computer*. Available online at: http://www.Praat.Org/ (accessed February 10, 2022).

Boroditsky, L. (2009). "How does our language shape the way we think?," in *What's Next? Dispatches on the Future of Science*, ed M. Brockman (Vintage Press).

Browman, C. P., and Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *J. Phonetics* 18, 299–320. doi: 10.1016/S0095-4470(19)30376-6

Corballis, M. C. (2002). "Did language evolve from manual gestures," in *The Transition to Language,* ed A. Wray (Oxford: Oxford University Press), 161–179.

Corballis, M. C. (2014a). "The word according to Adam," in *From Gesture in Conversation to Visible Action as Utterance: Essays in Honor of Adam Kendon*, eds M. Seyfeddinipur, and M. Gullberg (Amsterdam: John Benjamins Publishing Company), 177–198.

Corballis, M. C. (2014b). The gradual evolution of language. *Hum. Mente J. Philos. Stud.* 7, 39–60.

Deutscher, G. (2010). *Through the Language Glass: Why the World Looks Different in Other Languages*. New York, NY: Metropolitan books.

Dolscheid, S., Hunnius, S., Casasanto, D., and Majid, A. (2012). "The sound of thickness: Prelinguistic infants' associations of space and pitch," in *Proceedings of the 34th annual meeting of the Cognitive Science Society*, eds N. Miyake, D. Peebles, and R. P. Cooper (Austin, TX: Cognitive Science Society), 306–311.

Dolscheid, S., Hunnius, S., Casasanto, D., and Majid, A. (2014). Prelinguistic infants are sensitive to space-pitch associations found across cultures. *Psychol. Sci.* 25, 1256–1261. doi: 10.1177/0956797614528521

Eitan, Z. (2013). "How pitch and loudness shape musical space and motion: new findings and persisting questions," in *The Psychology of Music in Multimedia*, eds S. -L. Tan, A. Cohen, S. Lipscomb, and R. Kendall (Oxford: Oxford University Press), 161–187. doi: 10.1093/acprof:oso/9780199608157.003.0008

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev.* 60, 1497–1510. doi: 10.2307/1130938

Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* 20, 104. doi: 10.1037/0012-1649.20.1.104

Fuchs, S., Reichel, U. D., and Rochet-Capellan, A. (2016). "F0. declination and speech planning in face to face dialogues," in *Proceedings 27th Conference Electronic Speech Signal Processing,* Leipzig (145–152).

Galantucci, B., Fowler, C. A., and Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psycho. Bullet. Rev.* 13, 361–377. doi: 10.3758/BF03193857

Gentilucci, M., and Corballis, M. C. (2006). From manual gesture to speech: a gradual transition. *Neurosci. Biobehav. Rev.* 30, 949–960. doi: 10.1016/j.neubiorev.2006.02.004

Goldin-Meadow, S. (2021). Gesture is an intrinsic part of modern-day human communication and may always have been so. *Oxford Handb. Hum. Symb. Evol.* doi: 10.1093/oxfordhb/9780198813781.013.12

Hirst, D., and Hancil, S. (2013). *Prosody and Iconicity*. Amsterdam: John Benjamins Publishing Company.

Hockett, C. F. (1978). In search of Jove's brow. *Am. Speech* 53, 243–313. doi: 10.2307/455140

Hsu, C. H., Evans, J. P., and Lee, C. Y. (2015). Brain responses to spoken F0 changes: is H special? *J. Phonetic.* 51, 82–92. doi: 10.1016/j.wocn.2015.02.003

Jeschonek, S., Pauen, S., and Babocsai, L. (2013). Cross-modal mapping of visual and acoustic displays in infants: the effect of dynamic and static components. *Eur. J. Dev. Psychol.* 10, 337–358. doi: 10.1080/17405629.2012.681590

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Köhler, W. (1929). *Gestalt Psychology.* New York, NY: Liveright.

Küssner, M. B., Tidhar, D., Prior, H. M., and Leech-Wilkinson, D. (2014). Musicians are more consistent: gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Front. Psychol.* 5, 789. doi: 10.3389/fpsyg.2014.00789

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Software* 82, 1–26. doi: 10.18637/jss.v082.i13

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431. doi: 10.1037/h0020279

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

McClave, E. (1998). Pitch and manual gestures. *J. Psycholing. Res.* 27, 69–89. doi: 10.1023/A:1023274823974

McNeill, D. (2012). *How Language Began: Gesture and Speech in Human Evolution*. Cambridge: Cambridge University Press.

McNeill, D. (2013). *The Co-evolution of Gesture and Speech, and Downstream Consequences*. Berlin: De Gruyter Mouton.

McWhorter, J. H. (2014). *The Language Hoax: Why the World Looks the Same in Any Language*. Oxford: Oxford University Press.

Nakagawa, S., Johnson, P. C., and Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. Royal Soc. Interf.* 14, 20170213. doi: 10.1098/rsif.2017.0213

Ohala, J. J. (1982). The voice of dominance. *J. Acoust. Soc. Am.* 72, S66–S66. doi: 10.1121/1.2020007

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16. doi: 10.1159/000261706

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. *Sound Symb.* 2, 325–347. doi: 10.1017/CBO9780511751806.022

Panksepp, J. (2010). "Emotional causes and consequences of social-affective vocalization," in *Handbook of Behavioral Neuroscience*, eds J. P. Huston and H. Steiner (Amsterdam: Elsevier), 201–208.

Perlman, M., and Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture* 14, 320–350. doi: 10.1075/gest.14.3.03per

Pinker, S. (2007). *The Stuff of Thought: Language as a Window Into Human Nature*. New York, NY: Penguin.

Pisanski, K., and Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *J. Acoustic. Soc. Am.* 129, 2201–2212. doi: 10.1121/1.3552866

Pullum, G. K. (1991). *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language*. Chicago, IL: University of Chicago Press.

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Available online at: https://www.R-project.org/ (accessed September 20, 2020).

Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *J. Conscious. Stud.* 8, 3–34.

Sapir, E. (1929). The status of linguistics as a science. *Language* 5, 207–214. doi: 10.2307/409588

Schegloff, E. A. (1998). Reflections on studying prosody in talk-in-interaction. *Lang. Speech* 41, 235–263. doi: 10.1177/002383099804100402

Shayan, S., Ozturk, O., and Sicoli, M. A. (2011). The thickness of pitch: Crossmodal metaphors in Farsi, Turkish, and Zapotec. *Senses Soc.* 6, 96–105. doi: 10.2752/174589311X12893982233911

Shintel, H., Nusbaum, H. C., and Okrent, A. (2006). Analog acoustic expression in speech communication. *J. Mem. Lang.* 55, 167–177. doi: 10.1016/j.jml.2006.03.002

Simonyan, K., and Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *Neuroscientist* 17, 197–208. doi: 10.1177/1073858410386727

Slobodchikoff, C. N., Perla, B. S., and Verdolin, J. L. (2009). *Prairie Dogs: Communication and Community in an Animal Society*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2010). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tsai, C. G., Wang, L. C., Wang, S. F., Shau, Y. W., Hsiao, T. Y., Auhagen, W., et al. (2010). Aggressiveness of the growl-like timbre: Acoustic characteristics, musical implications, and biomechanical mechanisms. *Music Percep.* 27, 209–222. doi: 10.1525/mp.2010.27.3.209

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., et al. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychol. Sci.* 21, 21–25. doi: 10.1177/0956797609354734

Whorf, B. L. (1956). "Science and linguistics," in *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, ed J. B. Carroll (Cambridge, MA: MIT Press), 207–219.

Zbikowski, L. M. (1998). Metaphor and music theory: reflections from cognitive science. *Music Theor.* 4, 1–8.